ICIW 2014

The Ninth International Conference on Internet and Web Applications and Services

ISBN: 978-1-61208-361-2

July 20 - 24, 2014

Paris, France

ICIW 2014 Editors

Guadaloupe Ortiz, University of Cádiz, Spain
Elena Troubitsyna, Abo Akademi University, Finland
Sergio De Agostino, Sapienza University of Rome, Italy

ICIW 2014

Foreword

The Ninth International Conference on Internet and Web Applications and Services (ICIW 2014), held between July 20-24, 2014, in Paris, France, 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.

Entertainment systems became one of the most business-oriented and challenging area of distributed real-time software applications' and special devices' industry. Developing entertainment systems and applications for a unique user or multiple users requires special platforms and network capabilities.

Particular traffic, QoS/SLA, reliability and high availability are some of the desired features of such systems. Real-time access raises problems of user identity, customized access, and navigation. Particular services such interactive television, car/train/flight games, music and system distribution, and sport entertainment led to ubiquitous systems. These systems use mobile, wearable devices, and wireless technologies.

Interactive game applications require particular methodologies, frameworks, platforms, tools and languages. State-of-the-art games today can embody the most sophisticated technology and the most fully developed applications of programming capabilities available in the public domain.

The impact on millions of users via the proliferation of peer-to-peer (P2P) file sharing networks such as eDonkey, Kazaa and Gnutella was rapidly increasing and seriously influencing business models (online services, cost control) and user behavior (download profile). An important fraction of the Internet traffic belongs to P2P applications.

P2P applications run in the background of user's PCs and enable individual users to act as downloaders, uploaders, file servers, etc. Designing and implementing P2P applications raise particular requirements. On the one hand, there are aspects of programming, data handling, and intensive computing applications; on the other hand, there are problems of special protocol features and networking, fault tolerance, quality of service, and application adaptability.

Additionally, P2P systems require special attention from the security point of view. Trust, reputation, copyrights, and intellectual property are also relevant for P2P applications.

On-line communications frameworks and mechanisms allow distribute the workload, share business process, and handle complex partner profiles. This requires protocols supporting interactivity and real-time metrics.

Collaborative systems based on online communications support collaborative groups and are based on the theory and formalisms for group interactions. Group synergy in cooperative networks includes online gambling, gaming, and children groups, and at a larger scale, B2B and B2P cooperation.

Collaborative systems allow social networks to exist; within groups and between groups there are problems of privacy, identity, anonymity, trust, and confidentiality. Additionally, conflict, delegation, group selection, and communications costs in collaborative groups have to be monitored and managed. Building online social networks requires mechanism on popularity context, persuasion, as well as technologies, techniques, and platforms to support all these paradigms.

Also, the age of information and communication has revolutionized the way companies do business, especially in providing competitive and innovative services. Business processes not only integrates departments and subsidiaries of enterprises but also are extended across organizations and to interact with governments. On the other hand, wireless technologies and peer-to-peer networks enable ubiquitous access to services and information systems with scalability. This results in the removal of barriers of market expansion and new business opportunities as well as threats. In this new globalized and ubiquitous environment, it is of increasing importance to consider legal and social aspects in business activities and information systems that will provide some level of certainty. There is a broad spectrum of vertical domains where legal and social issues influence the design and development of information systems, such as web personalization and protection of users privacy in service provision, intellectual property rights protection when designing and implementing virtual works and multiplayer digital games, copyright protection in collaborative environments, automation of contracting and contract monitoring on the web, protection of privacy in location-based computing, etc.

We take here the opportunity to warmly thank all the members of the ICIW 2014 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 ICIW 2014. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

Also, this event could not have been a reality without the support of many individuals, organizations, and sponsors. We are grateful to the members of the ICIW 2014 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that ICIW 2014 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in the field of Internet and Web applications and services.

We are convinced that the participants found the event useful and communications very open. We hope that Paris, France, provided a pleasant environment during the conference and everyone saved some time to enjoy the charm of the city.

ICIW 2014 Chairs:

ICIW Advisory Committee

Mario Freire, University of Beira Interior, Portugal
Steffen Fries, Siemens AG, Corporate Technology - Munich, Germany
Vagan Terziyan, University of Jyvaskyla, Finland
Mike Wald, University of Southampton, UK
Sergio De Agostino, Sapienza University of Rome, Italy
Kwoting Fang, National Yunlin University of Science & Technology, ROC
Renzo Davoli, University of Bologna, Italy
Gregor Blichmann, Technische Universität Dresden, Germany
Vincent Balat, University Paris Diderot - Inria, France
Ezendu Ariwa, University of Bedfordshire, UK

ICIW Industry/Research Chairs
Giancarlo Bo, Technology and Innovation Consultant- Genova, Italy
Ingo Friese, Deutsche Telekom AG - Berlin, Germany
Sven Graupner, Hewlett-Packard Laboratories - Palo Alto, USA
Alexander Wöhrer, Vienna Science and Technology Fund, Austria
Caterina Senette, Istituto di Informatica e Telematica, Pisa, Italy
Nazif Cihan Tas, Siemens Corporate Research - Princeton, USA
Jani Suomalainen, VTT Technical Research Centre of Finland, Finland
Zhixian Yan, Samsung Research America, USA
Samad Kolahi, Unitec Institute of Technology, New Zealand
ICIW Publicity Chairs
Sven Reissmann, University of Applied Sciences Fulda, Germany
David Gregorczyk, University of Lübeck, Institute of Telematics, Germany

ICIW 2014

COMMITTEE

ICIW Advisory Committee

Mario Freire, University of Beira Interior, Portugal
Steffen Fries, Siemens AG, Corporate Technology - Munich, Germany
Vagan Terziyan, University of Jyvaskyla, Finland
Mike Wald, University of Southampton, UK
Sergio De Agostino, Sapienza University of Rome, Italy
Kwoting Fang, National Yunlin University of Science & Technology, ROC
Renzo Davoli, University of Bologna, Italy
Gregor Blichmann, Technische Universität Dresden, Germany
Vincent Balat, University Paris Diderot - Inria, France
Ezendu Ariwa, University of Bedfordshire, UK

ICIW Industry/Research Chairs

Giancarlo Bo, Technology and Innovation Consultant- Genova, Italy
Ingo Friese, Deutsche Telekom AG - Berlin, Germany
Sven Graupner, Hewlett-Packard Laboratories - Palo Alto, USA
Alexander Wöhrer, Vienna Science and Technology Fund, Austria
Caterina Senette, Istituto di Informatica e Telematica, Pisa, Italy
Nazif Cihan Tas, Siemens Corporate Research - Princeton, USA
Jani Suomalainen, VTT Technical Research Centre of Finland, Finland
Zhixian Yan, Samsung Research America, USA
Samad Kolahi, Unitec Institute of Technology, New Zealand

ICIW Publicity Chairs

Sven Reissmann, University of Applied Sciences Fulda, Germany
David Gregorczyk, University of Lübeck, Institute of Telematics, Germany

ICIW 2014 Technical Program Committee

Charlie Abela, University of Malta, Malta
Dharma P. Agrawal, University of Cincinnati, USA
Mehmet Aktas, Indiana University, USA
Grigore Albeanu, Spiru Haret University - Bucharest, Romania
Nidal AlBeiruti, University of South Wales, UK
Feda AlShahwan, The Public Authority for Applied Education and Training (PAAET), Kuwait
Josephina Antoniou, UCLan Cyprus, Cyprus
Ezendu Ariwa, University of Bedfordshire, 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

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
Peter Teufl, Institute for Applied Information Processing and Communications (IAIK) - Graz University of Technology, Austria
Pierre Tiako, Langston University - Oklahoma, USA
Leonardo Tininini, ISTAT-Italian Institute of Statistics), Italy
Konstantin Todorov, LIRMM / University of Montpellier 2, France
Giovanni Toffetti, IBM Research Haifa, Israel
Orazio Tomarchio, University of Catania, Italy
Victor Manuel Toro Cordoba, University of Los Andes - Bogotá, Colombia
Vicente Traver Salcedo, Universitat Politècnica de València, Spain
Christos Troussas, University of Piraeus, Greece
Nikos Tsirakis, University of Patras, Greece
Pavel Turcinek, Mendel University in Brno, Czech Republic
Samyr Vale, Federal University of Maranhão - UFMA - Brazil
Bert-Jan van Beijnum, University of Twente, Netherlands
Dirk van der Linden, Artesis University College of Antwerp, Belgium
Perla Velasco-Elizondo, Autonomous University of Zacatecas, Mexico
Ivan Velez, Axiomática Inc., 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
R. Zafimiharisoa Stassia, University of Blaise Pascal, France
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

Ebanshu: An Interactivity-aware Blended Virtual Learning Environment

Jingjing Chen, You Li, Xiaowen Chu, Shaohuai Shi, Jianliang Xu
Department of Computer Science
Hong Kong Baptist University
Hong Kong SAR, China
Email: {jjchen, youli, chxw, xujl}@comp.hkbu.edu.hk, shaohuaishi@gmail.com

Tang Tao
Department of Mathematics
Hong Kong Baptist University
Hong Kong SAR, China
Email: ttang@math.hkbu.edu.hk

Lin Cui
Department of Computer Science
Jinan University
Guangzhou, China
Email: tcuilin@jnu.edu.cn

Jingjing Chen, Zhiling Xu
Department of Computer Science
China Jiliang University
Hangzhou, China
Email: {cembcjj, xuzhiling}@cjlu.edu.cn

***Abstract*—Virtual Learning Environment (VLE), as a type of e-Learning platform, is widely used to serve teaching and learning for education in many countries. However, most of the existing systems fail to seamlessly support and monitor the real-time interactivity and collaboration among the learners and instructors in the virtual learning environment. Moreover, instructors are unable to know the real-time learning statuses of learners at distance, which is critical to effective teaching and learning achievement. This paper presents a novel interactivity-aware blended VLE system for synchronous and asynchronous teaching and learning. By using popularity dashboard, instructors can monitor real-time learning statuses of learners. Furthermore, all the teaching activities in the virtual classroom will be automatically recorded as lecture videos for self-directed learning.**

Keywords-e-Learning; interactivity; blended learning; MOOC

I. INTRODUCTION

A. Background

E-Learning is increasingly popular for instructors and learners in universities [8]. Blended learning is a formal e-Learning program in which learners learn at least in part through online delivery of content and instruction with some element of learners control over time, place, path or pace. While still attending traditional instructor-led, face-to-face classroom methods are combined with computer-mediated activities. Many e-Learning tools are available for blended learning. For example, Blackboard [18] and Moodle [19] are widely used in universities and institutions around the world. These tools make the learning and teaching more efficient and productive, but they usually lack effective real-time monitoring of the learning process [14]. In the meantime many universities and institutions planned to construct platforms for Massive Online Open Courses (MOOCs) [5], which lots of learners in or out of campus can much benefit from. The real-time teaching/learning feedback is important for the effectiveness of MOOCs.

B. Challenges

As emphasized in e-Learning theory and practice [3][4][14], effective human interaction is a vital factor for successful e-Learning and teaching. A previous study also showed that beginners of an e-Learning system might feel being isolated from the teachers and other students, because of missing essential interactions components in the system design [10].

Lack of interactivity, usually, leads to terribly negative impact on the outcome of e-Learning. The emotional illiteracy and the feeling of being isolated in e-Learning environments need to be addressed urgently for new generation of e-Learning systems [1][2]. The key solution and objective are to strengthen interactivities among instructors and learners. With well-designed synchronous virtual classrooms and collaborative tools, the negative impact incurred by the lack of interactivity can be significantly reduced. [3][4].

In addition, most of the current VLEs lack monitoring on the interactivity and collaboration activities during the teaching and learning process, especially lack of effective mechanism enabling instructor know learners' learning statues in real-time manner [14]. However, the tracing and monitoring of these activities is important and necessary to analyze the learning behaviors in the VLEs, to provide accurate feedback to instructors for refining the method of teaching [17].

C. Contribution

In this paper, our contributions are three-folded:

1) We propose an interactivity-aware blended e-Learning architecture of *ebanshu* for massive web-based learning. By utilizing the real-time web and HTML5 technologies, the *ebanshu* architecture can enable real-time interactivities and automatic monitoring and recording of the teaching & learning activities, which are very important for improving cognitive engagement that is poorly supported by today's e-Learning systems [14][15].

2) We implement and deploy the *ebanshu* system for MOOCs. The system is available on Internet [9] and has been used by more than 10 universities and institutions in China. More than 30 online courses are hosted by *ebanshu*, serving more than 10,000 users for their daily teaching and learning activities.

3) We conducted extensive evaluations of the *ebanshu* system using the course MATH 7090 in Hong Kong Baptist University, as a case study. The analysis based on the evaluation results indicates that by using the *ebanshu* system, the interactivity in the virtual learning environment can be enhanced significantly.

 ISBN: 978-1-61208-361-2

The remainder of this paper is organized as follows. In Section II, we explain the architecture of *ebanshu*. A math course is used as an example to show the pedagogy design based on *ebanshu* in Section III. Evaluations and data analysis are presented in Section IV. Finally, we conclude the paper and describe some possible future work in Section V.

II. ARCHITECTURE AND FUNCTIONS

A. Design Rationale

The *ebanshu* system is designed as a web-based blended e-Learning platform, with support of full functions for VLE [11]. In particular, it is capable of supporting the monitoring of learning activities, the recording and delivering of massive online open courses. In *ebanshu*, all participants can freely join a virtual classroom with different types of devices. Two different learning models are supported by *ebanshu*: *synchronous learning* and *asynchronous learning*. *Synchronous learning* refers to the case when both instructors and learners are present in the virtual classroom during the teaching/learning process. For *asynchronous learning*, learners are allowed to study the course anytime without an instructor being online.

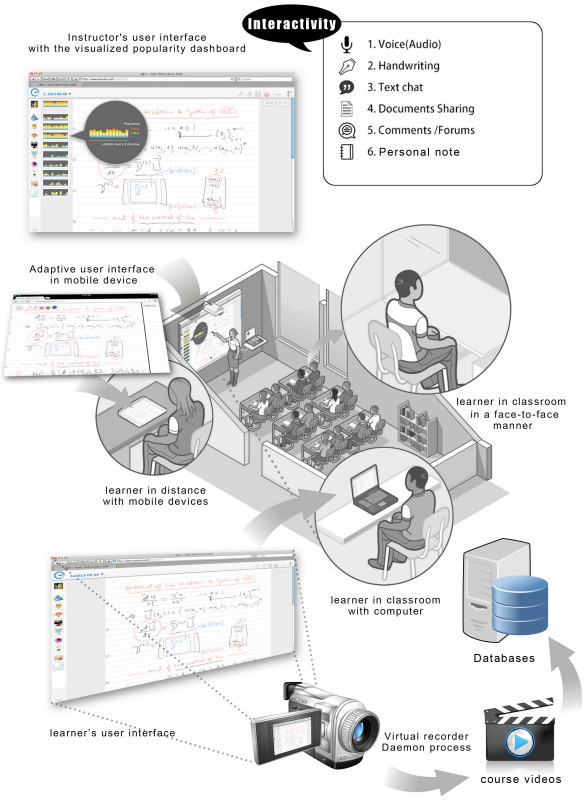


Figure 1. Overview of interactivity-aware blended learning scenario.

The whiteboard is a core module for *ebanshu*. All contents and actions on the whiteboard will be synchronized to all the participants in real-time. During the course, various communication media (e.g., voice, handwriting, text, video) are supported and strongly encouraged. Furthermore, instructors can know learners' real-time learning statuses through a visualized dashboard component. Besides, all the teaching contents and activities took place in the virtual classroom will be automatically recorded and archived in the background to generate course videos for asynchronous learning. A general scenario is depicted in Figure 1.

B. System Components

The proposed *ebanshu* architecture, which is described in Figure 2, includes four major components:

1) *Adaptive User Interface (AUI) component*: AUI offers a responsive web user interface adapting to various devices with consideration of the portability and user experience. Instructors and learners can use *ebanshu* simply through web browsers without installing additional application. The flash player embedded in web browser can be used as voice player and recorder. The HTML5 canvas is used to construct and render the whiteboard. The socket.io client can keep real-time connection with the server side and exchange data from time to time.

2) *Teaching & Learning Support Service (TLSS) component*: TLSS consists of the "virtual classroom" unit and the "course management" unit. The "user" module included in the "virtual classroom" unit serves as the attendance recorder and user manager. It also includes a popularity dashboard, which is updated periodically to show the learners' real-time learning statues. The popularity is defined as the frequency of the interactions (including raw Human-Computer interaction and predefined interactivities: Voice, Handwriting, Text chat, Document Sharing, Comments, Personal note) in ebanshu system. The numerical value of popularity ranges from 0 to 100. Learning status is defined as learner's response toward the teaching activities in ebanshu system, which can be quantitatively presented by the popularity. Instructors can adjust his/her teaching activities according to the popularity dashboard, and identify students who are passive and inactive toward the learning activities, and take actions to get them more involved into the virtual classroom [12][13].

The "Interactivity" module included in the "virtual classroom" offers various interactive and collaborative tools. The host of the classroom can communicate with participants in the virtual classroom by the "Voice" tool. There can be up to 5 participants speaking concurrently in classroom. The host can use the "Handwriting" and "Document sharing" tool to conveniently present and share documents in the whiteboard. Participants can communicate with each other through the "text chat" tool. The "Comments" tool is designed for asynchronous interactions, allowing learners can post questions to instructor about the course even if the instructor is absent. By using the "Personal note" tool, learners can note down personalized notes.

The course homepage included in the "Course" unit consists of syllabus, course forum, etc. Once instructor-led learning session completed, lecture videos would automatically generated by the *virtual recorder*, which was watched in the course homepage.

3) *Real-Time Service Cluster (RTSC)*: RTSC serves as the core middleware in the system, which employs the message notification and distributed computing. These technologies enabled the system capacity of high availability and high

 ISBN: 978-1-61208-361-2

concurrency. RTSC consists of "Socket.io Engine" module, "Document Interpreter" module and "Voice Service" module.

The "Socket.io Engine" module aims to make real-time apps possible in every browser and mobile device, blurring the differences among different transport mechanisms. It utilizes the event-driven and asynchronous model to smoothly handle the high concurrency issue, running websocket protocol and offering the real-time data synchronization service for the system, by which the content and actions in the whiteboard can be synchronized to all participants in real time manner.

The "Document Interpreter" module serves as the powerful web service for archiving and converting uploaded files (e.g., doc, ppt, pdf files) to portable format. It is running as distributed computing clusters with capacity of high availability and high performance.

The "Voice Service" module consists of several media servers running RTMP protocol, which are geographically distributed at different cities and offer reliable voice communication service.

4) *Virtual Recorder*: VR serves as daemon process on the server side, which enables the system support MOOCs. VR can continually record what happened in the virtual classroom. All recorded data will be saved to different data collections in the mongoDB.

MongoDB is an open-source document database, which can offer data storage service with high availability and high performance, we deployed the mongoDB enabling both the "Replica Set" and "Sharding" features, "Replica Set" enabled the mirror access LANs and WANs scale and peace of mind, and the "Sharding" enabled scale horizontally without compromising functionality.

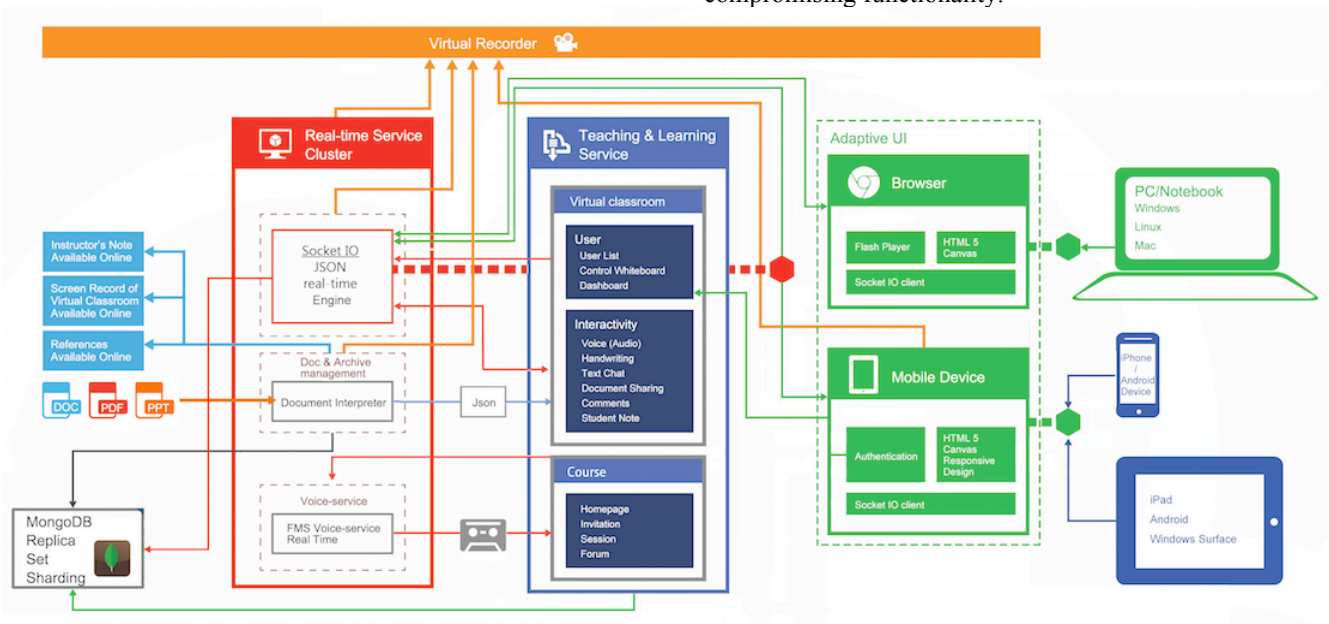


Figure 2. *ebanshu* components architecture.

III. PEDAGOGY DESIGN

In this section, we will take the course MATH7090 - "The high-precision Numeric Calculation" as a case study to show the pedagogy design based on *ebanshu*. MATH7090 is the first online open course offered in Hong Kong Baptist University. Considering the need for both synchronous learning and asynchronous learning, the pedagogical design of the course includes three separate components: Pre-course Learning Activities, Teaching & Learning Activities, and Assessment & Feedback.

A. *Pre-course Learning Activities*

Generally, instructor-led courses are usually organized by sessions, which can be daily or weekly, depending on the duration of the courses and learners' available time. The course MATH7090 was delivered to Science students during Semester 2013-2014-1. The course included 12 sessions, each session lasted two hours. 29 students officially enrolled the course. The course was hosted by the *ebanshu* system, and also was open on the Internet. 53 learners subscribed the course in the *ebanshu* system, including the 29 officially enrolled students and 24 Internet learners.

The course syllabus and *ebanshu* system instructions were automatically sent to subscribers by the system. The course syllabus describes the session topics and learning objectives. The instructions explained how to conduct teaching and learning activities.

Prior to the beginning of each session, Teaching Assistants (TAs) uploaded the teaching materials and references to the virtual classroom in the *ebanshu* system. TAs can join the virtual classroom in advance and perform the teaching assistance for the instructor. The instructor could perform the teaching anywhere, as long as the Internet was available (microphone is required). At the beginning of each session, a remainder letter was automatically sent to learners. They can come to the classroom where they can communicate with the instructor in a face-to-face fashion. They can conveniently join the virtual classroom with mobile devices anywhere.

 ISBN: 978-1-61208-361-2

B. *Teaching & Learning Activities*

During each session, there are many teaching and learning activities in the virtual classroom. Teaching activities are usually initiated and conducted synchronously by instructor. Learning activities consist of the synchronous interactive learning with the instructor and learners in the virtual classroom, and the asynchronous learning with less interaction and collaboration.

The *ebanshu* system offers various synchronous communication tools for instructors to conduct ***synchronous teaching***. The instructor can use the real-time voice and whiteboard to present the lectures, Additionally, the handwriting function is very helpful for presenting complicated questions with much descriptions; and the document sharing function usually is used to present documentations and references.

Learners are encouraged to hand up to (an alert message will be sent to instructor's interface) the instructor if they do not understand some teaching materials. Instructor will handle this case, and further communication with learners. Instructor usually much cares about learners' learning statuses during the teaching, which is crucial for learning outcome. The popularity dashboard described in the Figure 3 is a good assistant for instructor. The instructor can accurately know the learners' learning statuses, and intervene with inactive learners, so that the adjustment of the teaching pace can be conducted timely.

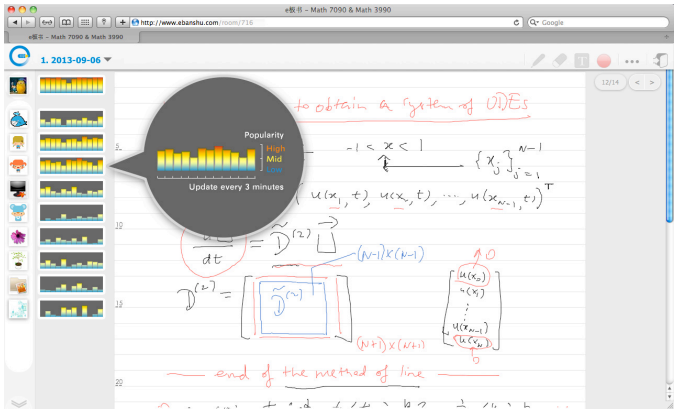


Figure 3. The instructors' user interface with the popularity dashboard.

Each session of the course is composed of a sequence of ***synchronous learning*** activities, which include a range of individual and collaborative activities between instructors and learners.

Learners can view the references and teaching materials, which may include different types of contents, e.g., learning resources (documents and presentations), video and audio contents. They can also take private notes on the whiteboard. Learners' user interface on desktop PC is depicted in Figure 4, and while that on mobile device is depicted in Figure 5. In the virtual classroom, all lectures and materials will be automatically saved and accessible to all participants, which is more convenient and efficient than the traditional classroom learning. Learners can also initiate group discussions and are encouraged to create new virtual classroom for learning proposes. In the background, the system tracks interactivities so that instructors or the TAs can review them afterwards and evaluate learners' involvement and filter out the hard questions in each session.

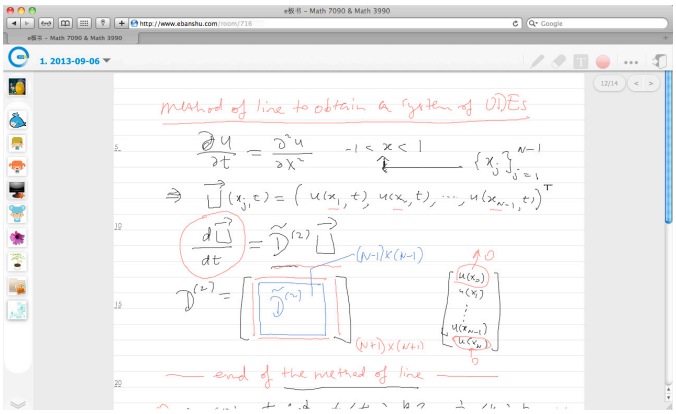


Figure 4. Learners' user interface on desktop PC and laptop.

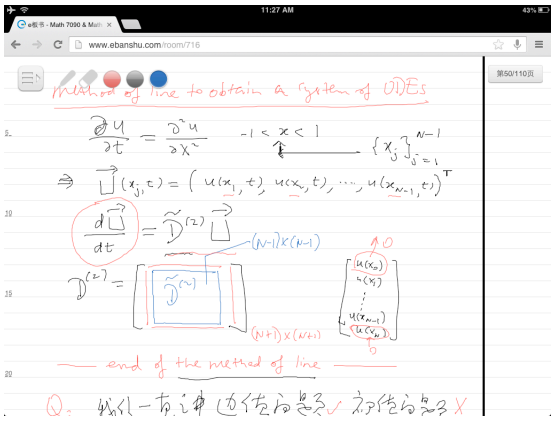


Figure 5. Learners' user interface on mobile device (iPad).

Asynchronous tools are available for learners to conduct ***asynchronous learning***, which includes the assignments requiring detailed response and more time, asynchronous discussions, self study by the learners, who are unavailable for synchronous teaching, and course review. Learners can download the lectures and references, and watch the online course videos, which can be used for self-study and course review. Learners can also leave comments about the course in the instructor's homepage. When the instructor was present, he/she would get the message, allowing to further communication with the learners.

C. *Assessment and Feedback*

In each session, the instructor uploads the assignments in the virtual classroom, which were attached in the whiteboard as separate page. Learners are expected to submit the answer sheets of the assignments in the virtual classroom directly by using uploading function. Finally, the instructor or the TAs can assess the learners' performance and comment on these assignments in the whiteboard. The comments would individually help learners to review the highlighted contents. In order to further enhance the teaching quality through learners' feedback, an evaluation survey was conducted at the last session of the course by utilizing the questionnaires and short talk with learners and instructors. This is a critical component

 ISBN: 978-1-61208-361-2

for e-Learning systems since it allows system designers to improve it over time.

IV. IMPACT EVALUATION

A. Evaluation criteria

According to Graham et al. [16], seven principles (including "Instructors should provide clear guidelines for interaction with students", "Well-designed discussion assignments facilitate meaningful cooperation among students", "Students should present course projects", "Instructors need to provide two types of feedback: information feedback and acknowledgment feedback", "Online courses need deadlines", "Challenging tasks, sample cases, and praise for quality work communicate high expectations", "Allowing students to choose project topics incorporates diverse views into online courses") are helpful to significantly improve learning outcomes. Ebanshu can perfectly meet 4 principles (including "Instructors should provide clear guidelines for interaction with students", "Well-designed discussion assignments facilitate meaningful cooperation among students", "Students should present course projects", "Allowing students to choose project topics incorporates diverse views into online courses") out of these principles [9]. Furthermore, studies show that "interactivity" is the most crucial evaluation criteria for e-Learning systems [7][8][16]. In the following, we conduct a comparative analysis of interactivity from four aspects, which includes "Participants' attendance", "Interactivity by instructor", "Interactivity by learners", and "Popularity", "Records" in Figure 7 and Figure 8 are actions in the whiteboard, which will be saved to database and automatically counted for data analysis.

There are kinds of metrics available to evaluate the performance of online courses in the administrative page of the system, aiming to evaluate and analysis "the way to join the course", we focus on four metrics: "*Device used - Mobile*" describes how many learners use mobile device in each session; "*Device used - PC*" counts how many learners use PC in each session; "*Participants - Synchronous*" records how many learners attend the virtual classroom synchronously; "*Participants - Asynchronous*" records how many learners attend the virtual classroom asynchronously.

Interactivities among the participants help to understand how participants react to the teaching. This can be measured through the real-time tracking and statistical analysis of activities of instructors and learners. For instructors, "*Whiteboard - Handwriting*" describes the usage of the whiteboard-based handwriting tool in each session; "*Whiteboard - Document Sharing*" – describes the usage of the whiteboard-based document sharing. For learners, "*Notes*" records the learners' private note in each session, "*Hands up*" records the learners' handing up during the synchronous virtual classroom in each session, "*Comments*" records the learners' comments on the course in each session, and "*Text chat*" describes the learners' text communication in each session. "Popularity of instructor" will be used to track the change of the average value of the popularity index in each session; "popularity of learner" will be used to track the change of the average value of the popularity index of all the learners in each session.

B. Data Analysis

The following figures show the collaboration and interactivity throughout all the 12 sessions of the course.

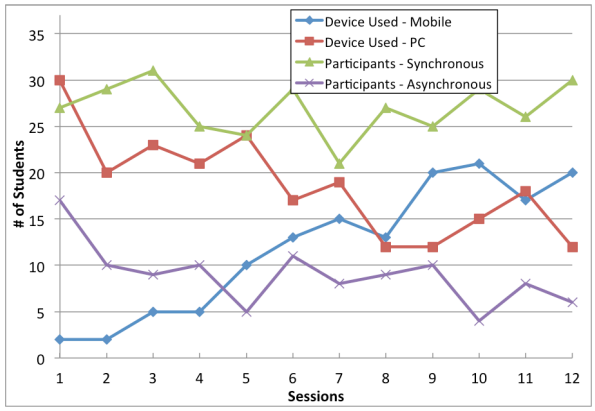


Figure 6. Participants' attendance in virtual classroom throughout the course.

Figure 6 shows that the learners prefer to attend the course by mobile device instead of PC in the end of the course. The instructor should prepare the teaching materials with extra consideration of document size. Most of synchronous participants attended all the 12 session, while asynchronous participants did not attend all sessions. The main reason is probably due to the feeling of being isolated.

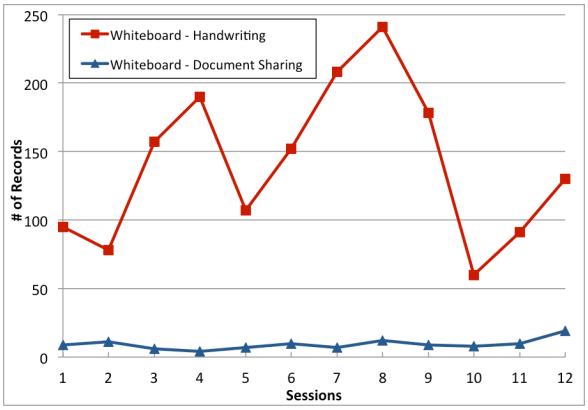


Figure 7. Interactivity by instructors in the virtual classroom in each session.

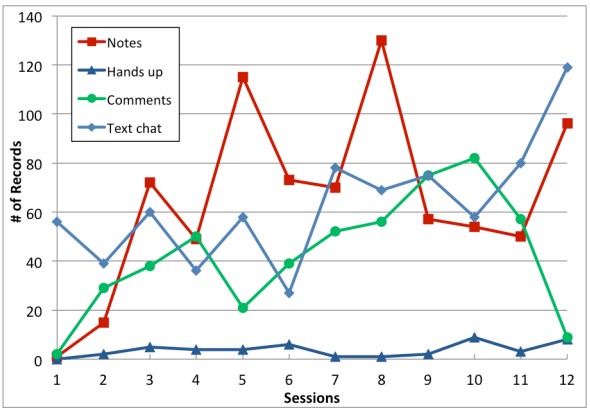


Figure 8. Interactivities by learners in the virtual classroom in each session.

 ISBN: 978-1-61208-361-2

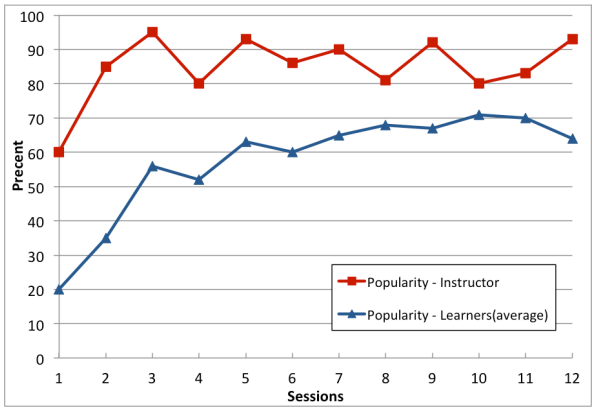


Figure 9. Popularity observation in the virtual classroom throughout the course.

Figure 7 shows that the instructor preferred using the handwriting to deliver teaching, which is very helpful to explain complicated problems. The handwriting is heavily used in Session 8, which is about the "Temporal discretization and FFT" and includes lots of mathematical formulae. The same observation is also found in Figure 6 for the activities of "Note". Figure 8 also shows that note is the most popular tool for learners, followed by text chat, comments, and "hands up". The text chat tool is frequently used and gets the highest record in the last session, which included the course review. In the last session, learners preferred to ask questions by the text chat instead of the "hands up" tool, and the instructor should adjust the pedagogical design to pay more attention to observe the contents of text chat.

Red curve in Figure 9 describes the change of the mean value of instructor's popularity throughout each session, and blue curve describes the change of the mean value of all the learners' popularity throughout each session. Figure 9 shows that the popularity of the instructor was always at a high level throughout the course. The popularity of learners increased over time, which indicates that if the instructor can accurately monitor learners' learning statuses, it is more likely to keep learners in active learning.

V. CONCLUSION AND FUTURE WORK

In this paper, a novel web-based e-earning system - ebanshu was presented. By providing real-time response and supporting mobile devices, *ebanshu* can help facilitate the activities of teaching and learning, improving the teaching efficiency and learning outcomes. In particular, *ebanshu* provides a real-time visualized popularity dashboard for instructors to monitor the learners' statuses in the virtual classroom. The system can also automatically record the instructor-led courses in the background and generate the courses learning materials for asynchronous learning. With *ebanshu*, courses can be easily delivered and made available online as the MOOC, benefiting many off-campus learners out of campus. We have also explored the pedagogical design with the novel e-Learning system and statistically evaluated the impact to the course MATH7090 at Hong Kong Baptist University. The ebanshu has been successfully used in more than 10 universities (including Peking University, Zhejiang University, and Jilin University), hosting more than 30 online courses, and offered teaching/learning service for more than 10,000 users.

The *ebanshu* apps for iOS and Android are under development, which will make offline asynchronous functions available in the future. With the rich performance data and user feedbacks obtained from many universities, the analysis based on the feedback will be an important work in future.

ACKNOWLEDGMENT

This work is supported by Hong Kong Innovation & Technology Fund entitled "Education Data Mining Based Personalized Mathematics Teaching System for Primary and Middle Schools", and the reference number is ITS/166/12FX.

REFERENCES

[1] D'Mello, S., Picard, R., and Graesser, A. "Toward an affect-sensitive AutoTutor." Intelligent Systems, IEEE 22, no. 4, 2007, pp. 53-61.

[2] Bråten, I. and Strømsø, H. I. "Epistemological beliefs, interest, and gender as predictors of Internet-based learning activities." Computers in Human Behavior 22, no. 6, 2007, pp. 1027-1042.

[3] Nedeva, V., Dimova, E., and Dineva, S. "Overcome Disadvantages of E-Learning for Training English as Foreign Language.", 2010.

[4] Kruse K.2004 http://www.e-Learningguru.com/articles/art1_3.htm. [retrieved: 21, January, 2014]

[5] Pisutova, K. "Open education." In Emerging eLearning Technologies & Applications (ICETA), IEEE 10th International Conference on, 2012, pp. 297-300.

[6] Clark R.C., The New Virtual Classroom: Evidence-based Guidelines for Synchronous e-Learning, Pfeiffer, 2007.

[7] Kirkpatrick D.L. and Kirkpatrick J.D. Evaluating Training Programs. The Four Levels. San Francisco: Berrett-Koehler Publishers, 2006

[8] Mahanta, D. and Ahmed, M. "E-Learning Objectives, Methodologies, Tools and its Limitation.", International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol. 2, December 2012, pp. 46-51.

[9] Ebanshu, http://www.ebanshu.com, [retrieved: 20, Febuary, 2014].

[10] Jonas, D., CMMC, N., & Burns, B. "Using e-Learning to Educate Health Professionals in the Management of Children's Pain.", 6th International Conference Creativity & Engegement In Higher Education, 2013.

[11] Fronter, http://com.fronter.info/virtual-learning-environment-lms. [retrieved: 10, Febuary, 2014].

[12] Croft, Nicholas, Alice Dalton, and Marcus Grant. "Overcoming isolation in distance learning: Building a learning community through time and space."Journal for Education in the Built Environment 5, no. 1, 2010, pp. 27-64.

[13] Food and Agriculture Organization of the United Nations, "E-learning methodologies A guide for designing and developing e-Learning courses", FAO, 2012, pp. 59-64.

[14] Belaid-Ajroud, H., Talon, B., and Tnazefti-Kerkeni, I. "Monitoring Activities in an E-Learning 2.0 Environment: A Multi-Agents system." ICIW 2013, The Eighth International Conference on Internet and Web Applications and Services. 2013.

[15] Agarwal, R. and Karahanna, E.. "Time flies when you're having fun: Cognitive absorption and beliefs about information technology usage." MIS quarterly 24, no. 4, 2000.

[16] Graham, C., Cagiltay, K., Lim, B., Craner, J., and Duffy, T. M. "Seven principles of effective teaching: A practical lens for evaluating online courses." The Technology Source 30, no. 5, 2001, pp. 14-16.

[17] 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, 2010, pp. 73-77.

[18] Blackboard, http://www.blackboard.com, [retrieved: 20, Febuary, 2014].

[19] Moodle, http://www.moodle.org, [retrieved: 20, Febuary, 2014].

 ISBN: 978-1-61208-361-2

A Study of the Effects of Scaffolded Assessment on the Learning Effectiveness in Network Peer Assessment Activities

Chien-I Lee , Ya-Fei Yang, and Shin-Yi Mai
Department of Information and Learning Technology,
National University of Tainan,
Tainan, Taiwan
e-mail: {leeci@mail.nutn.edu.tw, cillia55@gmail.com, maitz3@hotmail.com}

***Abstract*—Peer assessment has been considered an important process for learning, However, students may not offer constructive feedback due to lack of expertise knowledge in network assessment activities. Scaffolding can be offered to students timely to assist in peer assessment. Therefore, to address this issue, this study proposed a scaffolded assessment approach accordingly. To evaluate the effectiveness of the proposed approach, the quasi-experimental design was employed to investigate the effects of scaffolded assessment for self-critiques and peer assessment on students' learning effectiveness in the network assessment activities. A total of ninety 7th graders participated in the experiment, and divided into three groups with or without the scaffolding critique. The results show that the validity in the network peer assessment has been well demonstrated, indicating scaffolded assessment could enhance the validity of the network peer assessment. Moreover, the participants' learning effectiveness is also enhanced. In addition, the participants showed a positive learning attitude toward the network assessment activities and agreed that the network assessment activities could enhance the participants' interactions between the peers and instructor. All in all, the proposed scaffolded assessment enables the participants to offer constructive feedback in the assessment activities.**

Keywords-Network Peer Assessment; Scaffolding; Validity; Learning Effectiveness; Learning Attitude

I. INTRODUCTION

Research on network peer assessment had been extensively studied and indicated that network peer assessment has positive effects on learning [11]. Due to rapid development of network technologies, disadvantages of paper-pencil based peer assessment have been gradually corrected and then replaced by network assessment. Individuals who share common interests, feelings, or ideas over the Internet form a virtual community in which learners evaluate peers' works, receive feedback from peers, modify their original works based on peer assessment, and then make their works become better [14][18].

Peer assessment enables learners to gain diverse ideas and inspiration, enhances their higher order thinking skills, and offers peers opportunities to learn from each other, which can promote learners' learning motivation and achievements [7][16][17][22]. Network peer assessment defined as students are given the needed knowledge to tasks review and feedback in peer assessment process through the Internet. Basically, network peer assessment is similar with peer assessment [12][13].

However, the lack of expertise may result in weak feedback and comments due to student incomprehension and misunderstanding of works during the peer assessment [5]. Thus, the appropriate support and scaffold given to student is needed, in which facilitate students to offer proper comments for their peers works [21]. Learners can observe peers' works, understand their learning progress, reflect on self-learning, and then provide feedback to peers for further improvement [1]. In addition, peers may gain more feedback from diverse backgrounds of peers understanding toward the works than that from teachers [2]. Previous research have proved the effectiveness of the scaffolded assessment. For instance, Cho, Schunn, and Wilson [3] stated that scaffolded assessment could facilitate evaluators to assess given tasks precisely.

Thus, this study proposes the scaffolded assessment for self-critiques and peer assessment by using self-critiques and peer feedback as an evaluation basis. To evaluate the effectiveness of the proposed approach, the quasi-experimental design was employed to investigate the effects of scaffolded assessment for self-critiques and peer assessment on students' learning effectiveness and learning attitude in the network assessment activities.

In this paper, a survey of related work will be studied in Section II. Section III is our research framework. The experimental process and activities will be given in Section IV. Section V describes the research tools. Section VI shows the experiment results. Finally, the conclusions and some future work are given in Section VII.

II. LITERATURE REVIEW

A. Peer Assessment

The concept of peer assessment is from Peer Assisted Learning (PAL) and multiple assessments. Types of PAL include peer tutoring, peer modeling, peer education, peer counseling, peer monitoring, and peer assessment [15]. Topping [14] stated that peer assessment is performed by students with similar degree of knowledge or background. Students not only have to learn knowledge and accomplish their assignments but also play a role as a tutor to observe and evaluate peers' works. Then, students receive peer

 ISBN: 978-1-61208-361-2

feedback and modify their original assignments. Isaacs [6] mentioned that peer assessment is an activity in which students demonstrate their own works to peers for assessment. Peers who participate in assessment activities must have the same educational background or be in the same class, and they play roles of being an author, evaluator, and evaluatee. Students have to observe, evaluate, and compare by playing different roles in multiple activities which can enhance their diverse thinking abilities and learning effectiveness.

B. Network Peer Assessment

Network peer assessment is also called as web-based peer assessment, in which students can facilitate contacts, assist in brainstorming and generate meaningful learning [10]. With the application of the Internet, students can observe and evaluate peers' assignments on a network teaching platform and receive feedback from peers, and then they can modify their original works based on the received feedback [8][14][18]. Knoy, Lin, Liu, and Yuan [9] stated that the three advantages of network peer assessment are: 1) to ensure anonymity and enhance peers' willingness to evaluate, 2) easy for teachers to monitor learning and evaluation process which reduce paper waste and copying cost, and 3) convenient for students to demonstrate their assignments for peer assessment.

C. Saffolding

The concept of scaffolding is based on Vygotsky's learning theory. Scaffolding is the process of teaching children, which is similar to the construction of a house, by realizing children's characteristics, offering appropriate assistance, and ensuring that children have full support. When children are able to independently solve specific problems, support or assistance may be gradually reduced to develop their own problem solving ability. Once they can independently solve various problems without assistance, then scaffolding can be removed. Vygotsky [19] indicated development is the transformation of socially shared activities into internalized processes. The concept of scaffolding is similar to the Zone of Proximal Development (ZPD), in which instructors or peers with better academic performance can provide scaffoldings to assist students to develop their learning ability and achieve the goal of transferring learning. As learners' ability is enhanced, they can learn independently and construct their own knowledge which is the time to gradually reduce scaffolding. Thus, with scaffolded assessment, students can construct new knowledge and enhance their ability through peer feedback.

III. RESEARCH FRAMEWORK

This study aims to investigate the effects of scaffolded assessment on the students' learning effectiveness and validity of the network peer assessment. Since Modular Object-Oriented Dynamic Learning Environment (Moodle) [23] is a free software e-learning platform with three functions: management of website, learning, and course, it can run on Windows and Mac and is easy to access and manage. Thus, this study uses Moodle to construct a network peer assessment system in which the instructor can upload and manage course content, and students can upload their assignments, conduct peer assessment, and view their scores. The students' assignments, scores, and behaviors are recorded in the database. After the network peer assessment activities are completed, this study uses the SPSS 17.0 statistical software [24] to analyze the data.

This study investigates the effects of different scaffolded assessment approaches on the learning effectiveness, learning attitudes, and validity of the network peer assessment. The variables include independent, dependent, and control ones, which are specified as follows:

1) The independent variable: Providing the scaffolded assessment.

a) The experimental group I is given the scaffolded assessment for self-critiques.

b) The experimental group II is given the scaffolded peer assessment.

c) The control group would not receive any scaffolded assessment when evaluating the other members' assignments in the network peer assessment activities.

2) The dependent variable:

a) Validity on the assessment: The validity refers to the consistency between the earned scores from the peer and instructor in this study.

b) Learning effectiveness: A student's final score on his or her assignment is the mean of the earned scores from the three network peer assessment activities.

c) Learning attitude: The leaning attitude indicates that the students' attitudes toward the network peer assessment activities.

3) The control variable:

a) Instructor: The instructor is a full-time computer instructor.

b) Course hour: The instructor would complete the teaching instruction in a 45-minute class in the first week of the network peer assessment activities.

c) Course content: The 5th and 6th chapters, which are "my creative photos" and "my creative photo frames", respectively, of the textbook.

d) Same softwares, PhotoCap 6 and Photo Magician, used to complete the course assignment in a computer lab.

The students of the two experimental groups and the control group would have the same course hours, learn the same content, and be taught by the same computer instructor in the same learning environment.

A. Participants

The participants of this study were from three classes of the 7th grade, with a total of 90 students who participated in the network peer assessment activities. They possess basic computer ability to implement the software, PhotoCap 6, and access computers, such as collection information, uploading and downloading data from the Internet. The participants are normally grouped by age in the junior high school. Three classes of the 7th graders are randomly selected and labeled as the experimental group I with 30 students conducted with self-critique for reference, the experimental group II with 29 students conducted with other peers' critique for reference,

 ISBN: 978-1-61208-361-2

and the control group with 31 students conducted without any reference.

B. *Network peer assessment*

Moodle is an open Course Management System (CMS) based on the theory of social construction. Moodle is supported by a large international community of educators because of the three main functions: management of website, learning, and course. It has an easy to learn user interface which enables instructors to manage their courses, and students can learn course content independently by opening a web browser and linking to the teaching platform. The proposed network peer assessment system is constructed by Moodle 2.4.1 and AppServ software. AppServ can configure Apache, PHP, and MySQL to form an integrated environment for the experimental activities. MySQL is used to construct the backend database.

The interfaces of the proposed network peer assessment system shows in Figure 1. The network peer assessment system is installed on the serve host system. The instructor and students have to access the Internet, link to the server, and enter the network peer assessment system. Student account and password required based on the course enrollment by teachers. Figure 2 shows the peer assessment activities announce and assessing records in the system. Students can observe and evaluate peers' assignments on a network teaching platform and receive feedback from peers.

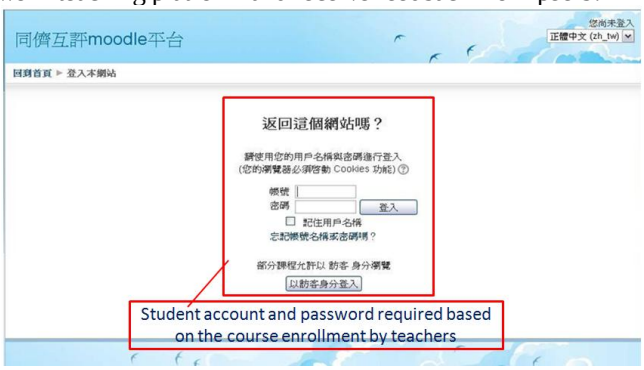


Figure 1. Student entry interface of the network peer assessment system

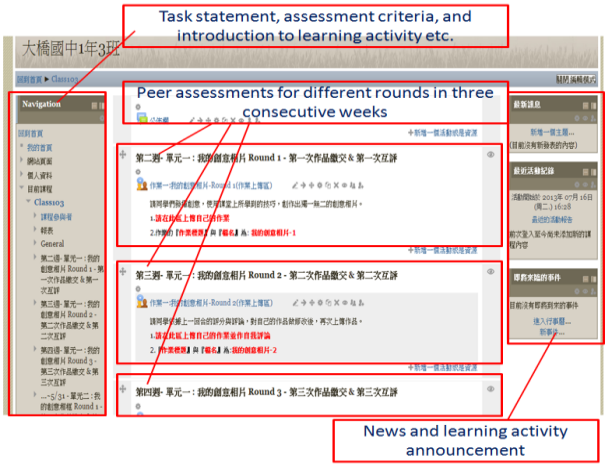


Figure 2. Peer assessment activity announced in the System

IV. THE DESIGN OF THE EXPERIMENTAL ACTIVITIES

A. *The experimental implementation phase*

After the preparation phase of the first four weeks, the experimental activities would be carried out for seven weeks. The details of the experimental activities are described in Table I.

TABLE I. THE DETAILS OF THE EXPERIMENTAL ACTIVITIES

1st week	Experimental groups *I* and *II*. The control group
1. Explanation of the course assignment and grading policy 2.The participants are required to upload their assignments to the system after the instructor delivered instruction	
2nd week	Experimental groups *I* and *II*. The control group
Completed the assignments, evaluated the other peers' assignments, offered feedback.	
3rd week	The experimental group *I*
Modified their original assignments based on the received feedback from their peers, provided self-critiques on their modified assignments, and then uploaded them to the proposed system in the first activity	
3rd week	Experimental group *II*. The control group
Modified their original assignments based on the received feedback from their peers and then uploaded them to the proposed system in the first activity	
4th week	Experimental group *I*
Read the evaluatees' self-critiques and evaluated the assignments in the second activity.	
4th week	Experimental group *II*
Read the other members' feedback, which was offered in the first network peer assessment activity, and then evaluated the assignments in the second activity.	
4th week	The control group
Evaluated the peers' assignments in the second activity.	
5th week	Experimental groups *I*
Modified their assignments again based on the received feedback from their peers, provided self-critiques on their modified assignments, and uploaded them to the system	
5th week	Experimental groups *II*. The control group
Read the other members' feedback, modified their assignments again, and uploaded them to the system	
6th week	Experimental group *I*
Read the evaluatees' self-critiques and evaluated the assignments in the third activity.	
6th week	Experimental group *II*
Read the other members' feedback on the assignments, which was offered in the first and second assessment activities, and then evaluated the third activity.	
6th week	The control group
Evaluated the peers' assignments.	
7th week	Experimental groups *I* and *II*. The control group
After completing all the three assessment activities, the participants are required to fill in the network peer assessment attitude scale in the system	

 ISBN: 978-1-61208-361-2

B. The result analysis phase

After the experimental activities are completed, the instructor and the teaching assistant would also evaluate the participants' assignments. The mean of the two given scores from the instructor and teaching assistant would be the earned score from the instructor on each participant's assignment. The scores from the instructor and peers would be incorporated for further analysis and investigation. The instructor has been a computer instructor for five year and fully familiar with the course content. The teaching assistant is also familiar with the necessary knowledge and skills required in this computer course.

V. RESEARCH TOOLS

This study uses the peer assessment score sheet as a tool for the participants to score the other members' assignments during the network peer assessment activities. The instructor has to explain the scoring criteria to the participants in the first week of the experimental activities. The scoring guidelines are developed based on the discussion between the teacher and participants. The peer assessment score sheet is used to evaluate the participants' assignments based on accuracy, expression, and completeness. The lowest score is 1, and the highest score is 100 on the peer assessment score sheet.

This study modifies the Wen and Tsai's [20] questionnaire, which originally had 34 items with five-point Likert scale (1: strongly disagree, 5: strongly agree), with an aim to investigate students' perceptions of network peer assessment. The instrument contains four subscales: positive attitude (PAS), online attitude (OAS), understanding-and-action (UAS), and negative attitude (NAS). The analysis results showed that the reliability of the four individual subscales took values more than 0.63 and 0.8 for the overall reliability. Thus, the questionnaire has high reliability which can be used to evaluate the students' attitudes toward network peer assessment.

The modified questionnaire contains 20 items. One of the three classes was randomly selected to evaluate the reliability and validity of the modified questionnaire. The results show that the Alpha coefficient takes a value of 0.6 for the four individual subscales and 0.86 for all the four subscales, indicating that the modified questionnaire is adequately reliable [4].

VI. RESULTS

The participants are required to complete the three assessment activities in this experiment. The valid sample size of the experimental group I is 25, making the completing rate of 83.3%. The valid sample size of the experimental group II is 26, making the completing rate of 89.7%. The valid sample size of the control group is 25, making the completing rate of 90.3%.

A. The reliability analysis

This study investigates the reliability between the scaffolded and non-scaffolded assessment activities. The experimental group I is given the scaffolded assessment for self-critiques. The experimental group II is given the scaffolding peer assessment. The control group would not receive any scaffolded assessment when evaluating the other members' assignments in the network peer assessment activities. This study uses the Pearson's correlation coefficient to measure the reliability between the scores from the instructor and teaching assistant. The results show that the reliability between the scores from the instructor and teaching assistant on the participants' assignments are positively correlated in the three assessment activities. The Pearson's correlation coefficient takes values more than 0.8, indicating that the scores from both the instructor and teaching assistant are adequately reliable and can be an external criterion for the validation of the network peer assessment.

B. The validity analysis of the scaffolded assessment for self-critiques in the experimental group I.

This study uses the scores from both the instructor and teaching assistant as the external criterion to examine whether the validity of the scaffolded assessment for self-critiques is enhanced in the network assessment activities. As can be seen in Table II, the results show that the scores from the instructor and peers are significantly positively correlated in the three network assessment activities, and the correlation coefficients, which take values between 0.5 to 0.7, gradually increase.

TABLE II. THE PEARSON CORRELATION COEFFICIENT BETWEEN THE SCORES FROM THE INSTRUCTOR AND PEERS IN THE EXPERIMENTAL GROUP I

Assessment activities	Evaluator	n	Mean	SD	r
The 1st	Instructor	25	73.44	8.905	.525**
	Peers		76.68	12.284	
The 2nd	Instructor	25	77.14	5.467	.624**
	Peers		79.04	7.738	
The 3rd	Instructor	25	84.25	5.592	.695**
	Peers		83.96	6.465	

** $p<0.01$

The consistency between the scores from the instructor and peers are enhanced after the second network assessment activity.

C. The validity analysis of the scaffolded peer assessment in the experimental group II

This study uses the scores from both the instructor and teaching assistant as the external criterion to examine whether the validity of the scaffolded peer assessment is enhanced in the network assessment activities. As it can be seen in Table III, the results show that the scores from the instructor and peers are significantly positively correlated in the first network assessment activity when $p < 0.05$.

The scores from the instructor and peers are significantly positively correlated in the second network assessment activity when $p < 0.01$, and their correlation coefficient takes values between 0.4 to 0.8. Thus, the consistency between the scores from the instructor and peers is enhanced.

 ISBN: 978-1-61208-361-2

TABLE III. THE PEARSON CORRELATION COEFFICIENT BETWEEN THE SCORES IN THE EXPERIMENTAL GROUP II

Assessment activities	Evaluator	n	Mean	SD	r
The 1st	Instructor	2	67.23	8.823	0.478*
	Peers	6	67.35	10.737	
The 2nd	Instructor	2	71.43	7.407	0.694**
	Peers	6	71.35	11.788	
The 3rd	Instructor	2	82.60	8.404	0.713**
	Peers	6	83.31	9.388	

*p<0.05; ** p<0.01

D. The validity analysis of the control group

This study uses the scores from both the teacher and teaching assistant as the external criterion to investigate the consistency between the earned scores from the instructor and peers on the same assignment in the network assessment activities. As it can be seen in Table IV, the results show that the scores from the instructors and peers are significantly positively correlated in the first and second network assessment activities.

TABLE IV. THE PEARSON CORRELATION COEFFICIENT BETWEEN THE SCORES IN THE CONTROL GROUP

Assessment activities	Evaluator	n	Mean	SD	r
The 1st	Instructor	28	70.82	8.147	0.684**
	Peers		79.36	11.656	
The 2nd	Instructor	28	76.11	6.632	0.417*
	Peers		80.86	10.302	
The 3rd	Instructor	28	87.19	4.865	0.117
	Peers		82.18	10.890	

*p<0.05 ; **p<0.01

However, the results show that the scores from the instructor and peers are not significantly correlated in the third network assessment activity. The process of repeatedly turning in, evaluating, and modifying assignments reduce the validity of peer assessment due to the lack of patience and motivation.

E. The comparison between the validity analyses of the two experimental groups and control group

As it can be seen in Table V, in the condition of no scaffolded assessment, the Pearson correlation coefficients between the scores from the instructor and peers are significantly positively correlated in the experimental group I and control group when $p < 0.01$ in the first assessment activity. The Pearson correlation coefficient between the scores from the instructor and peers is significantly positively correlated in the experimental group II when $p<0.05$. The results show that the validity of the network assessment activities is significantly enhanced in the two experimental groups, and the validity of the network assessment activities is significantly reduced in the control group. Moreover, the validity of the network assessment activities in the experimental group I is higher than that in the experimental group II, indicating that the consistency of the scores from the instructor and peers in the experimental group II is higher than that in the experimental group I.

F. The analysis of learning effectiveness of the three groups

This study investigates the effects of scaffolded assessment on the students' learning effectiveness by comparing the results of the three assessment activities from the three groups. Since the three groups are not provided with the scaffolded assessment in the first assessment activity, this study performs the one-way ANCOVA where the students' scores in the first assessment activity are treated as a covariate to investigate the effects of the scaffolded assessment on the students' learning effectiveness in the second and third assessment activities. With regard to the second network assessment activity, since the assumption of the homogeneity of regression coefficient within groups was satisfied, the ANCOVA test was then performed. As it can be seen in Table VI, there is no significant difference among the three groups in the second assessment activity ($F = 1.603$ and $p = 0.208 > 0.05$), indicating that the students' learning effectiveness does not have a significant difference among the three groups in the second assessment activity.

TABLE V. THE PEARSON CORRELATION COEFFICIENT BETWEEN THE SCORES IN THE THREE GROUPS

Assessment activities	Group	n	r
The 1st	Experimental groups *I*	25	0.525**
	Experimental groups *II*	26	0.478*
	Control group	28	0.684**
The 2nd	Experimental groups *I*	25	0.624**
	Experimental groups *II*	26	0.694**
	Control group	28	0.417*
The 3rd	Experimental groups *I*	25	0.695**
	Experimental groups *II*	26	0.713**
	Control group	28	0.117

*p<0.05; **p<0.01

The results also show that the effects of providing the scaffolded assessment on the students' learning effectiveness in the second assessment activity are not significant. As it can be seen in Table VII, there is no significant difference among the three groups in the third assessment activity ($F = 2.625$ and $p = 0.079 > 0.05$), indicating that the students' learning effectiveness does not have a significant difference among the three groups in the third assessment activity. The results also show that the effects of providing scaffolded assessment on the students' learning effectiveness in the third assessment activity are not significant. Therefore, the students' learning effectiveness is enhanced in the three assessment activities, but the effects of providing the scaffolded assessment on the students' learning effectiveness are not significant in the three assessment activities.

TABLE VI. SUMMARY OF ANCOVA FOR THE THREE GROUPS IN THE SECOND ASSESSMENT ACTIVITY

Source of variation	Sum of Squares	df	Mean Square	F	Sig.
Covariates	1770.434	1	1770.43	22.109	0.000
Within Groups	256.790	2	128.395	1.603	0.208
Error	6005.839	75	80.078		

 ISBN: 978-1-61208-361-2

TABLE VII. SUMMARY OF ANCOVA FOR THE THREE GROUPS IN THE THIRD ASSESSMENT ACTIVITY

Source of variation	Sum of Squares	df	Mean Square	F	Sig.
Covariates	1384.810	1	1384.810	20.674	0.000
Within Groups	351.730	2	175.865	2.625	0.079
Error	5023.795	75	66.984		

G. The analysis of learning attitude of the three groups

To investigate the degree to which an individual agree or disagree the network assessment activity in the study, ANOVA was employed to analyze four dimensions of the attitude questionnaire. The analyzed result was shown in Table VIII. Table VIII shows both dimensions "positive attitude" and "online attitude" take values of significant difference among three groups ($F = 4.95$, $p < .05$ and $F = 3.76$, $p < .05$, respectively). The post hoc of positive attitude shows the mean of experimental group II significantly higher than that of control group, indicating that students of experimental group II have positively higher level of degree to the network peer assessment than those of control group. In addition, The post hoc of online attitude shows the mean of experimental group II significantly higher than that of control group, indicating that students of experimental group II have higher level of acceptance to the network peer assessment than those of control group.

TABLE VIII. ANOVA OF LEARNING ATTITUDE TOWARD NETWORK PEER ASSESSMENT FOR THREE GROUPS

Dimension	Group	SD	Mean	*F*	Post hoc
Positive Attitude (PAS)	(1)	0.61	3.90	4.95*	(2)>(3)
	(2)	0.57	4.32		
	(3)	0.70	3.55		
Negative Attitude (NAS)	(1)	0.98	2.36	0.78	
	(2)	0.66	2.59		
	(3)	0.58	2.59		
Online Attitude (OAS)	(1)	0.75	3.94	3.76*	(2)>(3)
	(2)	0.74	4.29		
	(3)	0.66	3.57		
Understanding and Action (UAS)	(1)	0.67	4.06	0.98	
	(2)	0.76	3.92		
	(3)	0.57	3.80		

*$p<0.05$, (1): Experimental groups *I*, (2): Experimental groups *II*, (3): Control group

VII. CONCLUSION AND FUTURE WORK

Scaffolding provided by teachers or peers can assist students to develop their learning ability and achieve the goal of transferring learning. This study uses the scaffolded assessment which provides scoring guidelines to investigate whether the students are benefited from the peer assessment activities. The results show that the learning effectiveness of the participants in the two experimental and control groups is not significantly enhanced during the first and second network peer assessment activities. However, the learning effectiveness of the participants in the two experimental groups is significantly enhanced, but the learning effectiveness of the participants in the control group did not show a significant difference.

The use of the scaffolded assessment in the network peer assessment activities did not show a significant difference in the students' learning effectiveness. Even though the results show that learning effectiveness of the participants in the two experimental groups is significantly enhanced, there is no significant effect of providing the scaffolded assessment on the participants' learning effectiveness in the three groups. Thus, the effects of the use of the scaffolded assessment on students' learning effectiveness should be further investigated.

The participants all have a positive attitude toward the network peer assessment activities. The results of the questionnaires which address the participants' perceptions of network peer assessment show that the participants agree that network peer assessment is beneficial for learning, enhances the sense of participation and motivation, increases interactions among peers and improves learning effectiveness. The participants with other peers' evaluation highly agree the network assessment in compared to those without scaffolded assessment in the network peer assessment activities. In addition, the scaffolded assessment provides a solid basis for the participants when evaluating the peers' assignments. The participants could also gain more diverse ideas and insights in the network peer assessment activities. In the future, to enhance the validity of the research, the number of the subjects would be increased. Furthermore, learning retention can be considered to investigate the effectiveness of the proposed approach.

REFERENCES

[1] C.-C. Chang and K.-H Tseng, "Using a web-based portfolio assessment system to elevate project-based learning performances," Interactive Learning Environments, vol. 19, no. 3, 2011, pp. 211-230.

[2] C.-H. Chen, "The implementation and evaluation of a mobile self-and peer-assessment system," Computers & Education, vol. 55, no.1, 2010, pp. 229-236.

[3] K. Cho, C. D. Schunn, and R. W. Wilson, "Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives," Journal of Educational Psychology, vol. 98, no.4, 2006, pp. 891.

[4] J. Cohen, Statistical Power Analysis for the Behavioral Sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, 1988.

[5] M. Freeman, "Peer assessment by groups of group work," Assessment and Evaluation in Higher Education, vol. 20, 1995, pp. 289-300.

[6] G. Isaacs, "Brief briefing: Peer and self assessment," The Effective Assessment Conference, University of Queensland, 1998.

[7] A. Jaillet, "Can online peer assessment be trusted?," Journal of Educational Technology & Society, vol. 12, no. 4, 2009.

[8] R. L. Johnson, F. McDaniel and M. J. Willeke, "Using portfolios in program evaluation: An investigation of

interrater reliability," American Journal of Evaluation, vol. 21, no. 1, 2000, pp. 65-80.

[9] T. Knoy, S.-J. Lin, Z.-F. Liu, and S.-M. Yuan, "Networked peer assessment in writing: Copyediting skills instruction in an ESL technical writing course," 2001, Unpublished dissertation, National Chiao Tung University, Taiwan.

[10] E. Z. F. Lin, C. H. Chiu, S. S. J. Lin, and S. M. Yuan, "Student participation in computer science courses via the Networked Peer Assessment System (NetPeas)," Proceedings of the ICCE' 99, 1999, pp. 774-777.

[11] C. Rushton, P. Ramsey, and R. Rada, "Peer assessment in a collaborative hypermedia environment: A case-study," Journal of Computer-Based Instruction, vol. 20, 1993, pp. 75-80.

[12] P. M. Sadler and E. Good, "The impact of self-and peer-grading on student learning," Educational assessment, vol. 11, no. 1, 2006, pp. 1-31.

[13] Y.-T. Sung, K.-E. Chang, S.-K. Chiou, and H.-T. Hou, "The design and application of a web-based self-and peer-assessment system," Computers & Education, vol. 45, no. 2, 2005, pp. 187-202.

[14] K. Topping, "Peer assessment between students in colleges and universities," Review of Educational Research, vol. 68, no. 3, 1998, pp. 249-276.

[15] K. J. Topping and S. W. Ehly, "Peer assisted learning: A framework for consultation," Journal of Educational and Psychological Consultation, vol. 12, no. 2, 2001, pp. 113-132.

[16] C.-C. Tsai and J.-C. Liang, "The development of science activities via on-line peer assessment: The role of scientific epistemological views," Instructional Science, vol. 37, no. 3, 2009, pp. 293-310.

[17] C.-C. Tsai, S. S. Lin, and S.-M. Yuan, " Developing science activities through a networked peer assessment system," Computers & Education, vol. 38, no. 1, 2002, pp. 241-252.

[18] S.-C. Tseng and C.-C. Tsai, "On-line peer assessment and the role of the peer feedback: A study of high school computer course," Computers & Education, vol. 49, no. 4, 2007, pp. 1161-1174.

[19] L. S. Vygotsky, Thought and language: MIT press, 2012.

[20] M. L. Wen and C.-C. Tsai, "University students' perceptions of and attitudes toward (online) peer assessment," Higher Education, vol. 51, no. 1, 2006, pp. 27-44.

[21] D. Wood, J. S. Bruner, and G. Ross, "The role of tutoring in problem solving," Journal of child psychology and psychiatry, vol. 17, no. 2, 1976, pp. 89-100.

[22] Y.-F. Yang and C.-C. Tsai, "Conceptions of and approaches to learning through online peer assessment," Learning and Instruction, vol. 20, no. 1, 2010, pp. 72-83.

[23] Dougiamas, M., & Taylor, P. (2003). Moodle: Using learning communities to create an open source course management system. In World conference on educational multimedia, hypermedia and telecommunications (Vol. 2003, No. 1, pp. 171-178).

[24] George, D. & Mallery, P. (2010). SPSS for Windows step by step: A sample guide and reference 17.0 update. Boston, Mass. : Allyn & Bacon.

 ISBN: 978-1-61208-361-2

Customer Satisfaction through E-Learning Software Product Line

Amina Guendouz
CS Department
Saad Dahlab University
Blida, Algeria
Email: guendouz.amina@yahoo.fr

Djamal Bennouar
CS Department
Akli Mohand OulHadj University
Bouira, Algeria
Email: dbennouar@gmail.com

Abdelouahab Ramdani, Hamza Mazari
CS Department, Saad Dahlab University
Blida, Algeria
Email: wahab.inf@gmail.com,
hamza.ntj@hotmail.com

***Abstract*— As online education becomes a basic need for several organizations, a variety of Learning Management Systems is proposed on the market. However, available systems do not satisfy all the needs of different institutions, which push them to develop their own systems. Since developing and maintaining new software are cost, time and effort consuming, and with the increasing demand on e-Learning systems, it becomes necessary to find an efficient solution that allows the fast development of systems and overcomes the before-mentioned issues. We strongly believe that adopting a software product line approach in e-Learning domain can bring important benefits. In this paper, we present the development process of an e-Learning software product line. Throughout the development process, we demonstrate how this approach allows us to satisfy the variable needs of customers and benefit from the systematic large scale reuse at the same time.**

Keywords–E-Learning; Software Product Line; reusability; variability management.

I. INTRODUCTION

Nowadays, the Internet knows a spread use in several fields, including the education. Taking advantage from the benefit of using the Internet, organizations seek to provide an efficient and less expensive way of education in terms of time, cost and effort. Remote training, virtual learning, or electronic learning (e-Learning) means the use of Information and Communication Technologies (ICT) in education to improve the process of teaching-learning.

In order to satisfy the needs of the different institutions, various e-Learning applications have been proposed, the most known are Learning Management Systems (LMSs). A LMS, also known as Virtual Learning Environments (VLE), is the infrastructure that delivers and manages instructional content, identifies and assesses individual and organizational learning or training goals, tracks the progress towards meeting those goals, and collects and presents data for supervising the learning process of organization as a whole [12].

In spite of their important advantages, LMSs present several limitations. Dalsgaard [3] argues that LMSs are limited to cover only administrative issues, and suggests the necessity to go beyond LMSs in e-Learning to improve interactions between students and instructors. On the other hand, a survey on LMSs that has been carried out for 113 European institutions [9] revealed that a large number of the LMS systems used in Europe are commercial systems developed locally, or self-developed systems built by the institutions. Only a few commercial systems are used by several institutions, which means that institutions tend to create their own e-Learning systems to fulfill their specific requirements. García-Peñalvo et al. [19] announced that, despite the high levels of LMS adoption, these systems have not produced the expected learning outcomes yet. They mentioned that among the main shortcomings of LMSs the failure to take into account the user. Other recent studies [20][21][22] show that LMSs do not satisfy all the needs of teachers and students which push them to use social networks, cloud based services and mobile applications in order to complement the lack of LMSs, and suggest that students need learning environments which are better adapted to their needs.

In order to overcome these issues, we suggest the use of Software Product Line (SPL) approach for the development of e-Learning applications. E-Learning applications could be implemented in a variety of settings: for schools and universities to compliment or enhance classroom learning, for corporations to provide training and certification for their employees, and for organizations to provide e-learning courses to a larger learners population virtually anywhere in the world.

However, all of these applications share a set of common software elements and differ by some variable parts. So, the adoption of a SPL approach in the e-Learning domain seems to be a promising solution. On one side, to overcome the limitations of LMS systems, and on the other side, to provide institutions with e-Learning applications that fit their own requirements. Furthermore, SPL Engineering (SPLE) aims to share the development work of a set of product using common means of production, in order to reduce the costs and effort of development, maintenance and test, decrease time to market and improve quality.

In this paper, we show how to build a SPL for e-Learning applications. The remainder of this paper is organized as follows: Section 2 introduces SPL approach and presents the SPLE process that we will follow after that to develop our e-Learning SPL. Section 3 shows the different steps of the development of an e-Learning product line, mainly domain engineering. Section 4 comments on related work, while Section 5 summarizes the paper and outlines future work.

II. SOFTWARE PRODUCT LINE ENGINEERING

A SPL is "A set of software-intensive systems sharing a common, managed set of features that satisfy the specific needs of a particular market segment or mission and that are developed from a common set of core assets in a prescribed way" [6]. SPL approach aims to systematize the reuse throughout all the software development process: from requirements engineering to the final code and test plans. The purpose is to reduce the time and cost of production and to increase the software

 ISBN: 978-1-61208-361-2

quality by reusing elements (core assets) which have been already tested and secured. These objectives can be realized by putting in common development artefacts such as requirement documents, design diagrams, architectures, codes (reusable components), procedures of test and maintenance, etc.

SPL approach aims to improve reuse while maintaining diversity between products. This could be done by "Variability management". Variability management is a key activity that usually affects the degree to which a SPL is successful [2]. Variability refers to the ability of an artefact to be configured, customized, extended, or changed for use in a specific context [1]. This variability must be defined, represented, exploited, implemented, evolved, etc. – in one word managed – throughout SPL engineering [11].

SPL Engineering (SPL) relies on a fundamental distinction between two activities [5][11]: SPL development and software production. SPL development aims to develop and maintain the base of reusable elements while software production aims to produce final applications according to customer's needs. As mentioned in Section 1, the main shortcoming of LMSs is the production of generic applications that do not meet the specific requirements of customers. Adopting a SPL process for the development of e-Learning applications will permit customers to be involved in the development process of their applications which gives them the opportunity to customize applications according to their specific needs. Moreover, e-Learning application developers will benefit from the large scale reuse and, thus, the reduction of time, effort and cost of development.

Based on SPL approach, we propose, as shown in Figure 1, a development process for e-Learning SPL. It is composed of two sub-processes: Domain engineering and Application engineering. Domain engineering (correspond to SPL development activity) includes domain analysis, domain design and domain implementation activities. The purpose of domain engineering is to produce reusable core assets and to provide the effective means that help in using these core assets to build a new product within a product line. A core asset is a reusable artifact or resource that is used in the production of more than one product in a SPL. A core asset may be an architecture, a software component, a domain model, a requirements statement or specification, a document, a plan, a test case, a process

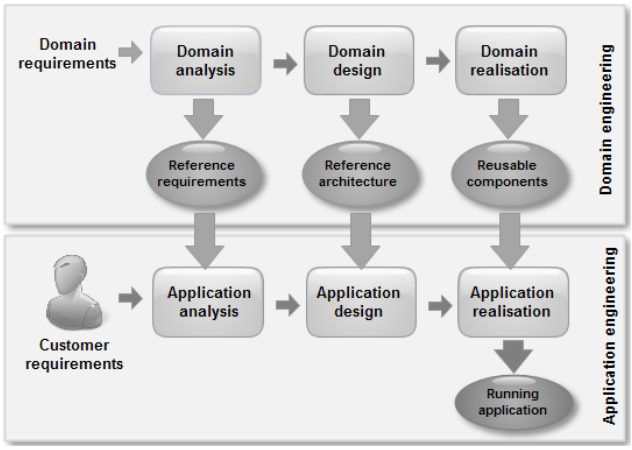


Figure 1. Software product line engineering process.

description, or any other useful element of a software production process [6]. The main outputs of this process are: reference requirements, reference architecture and reusable components.

Application engineering (corresponding to software production activity) consists in developing the final products, using the core assets and the specific requirements expressed by customers. This process is similar to traditional development process; however, each step is facilitated by the reuse of the outputs of the first process. The result of this process is an application ready to be used.

III. E-LEARNING SOFTWARE PRODUCT LINE ENGINEERING

In this section, we show the development process of our e-Learning product line focusing on the first sub-process: domain engineering.

A. Domain Engineering

As a preliminary activity of domain engineering, the scope of the SPL must be defined. In our case, e-Learning product line intends to cover the e-learning applications used by schools and universities providing online courses to their students, companies which provide online training to their employees and organizations that supply online courses to learners anywhere in the world. Domain engineering consists of three activities that are domain analysis, domain design and domain realization (Figure 1).

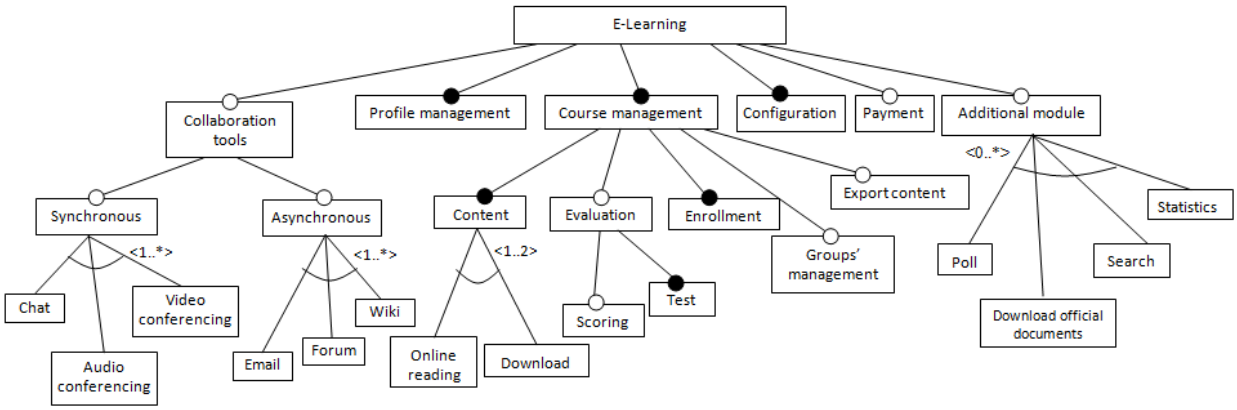


Figure 2. Capability feature diagram for e-Learning product line.

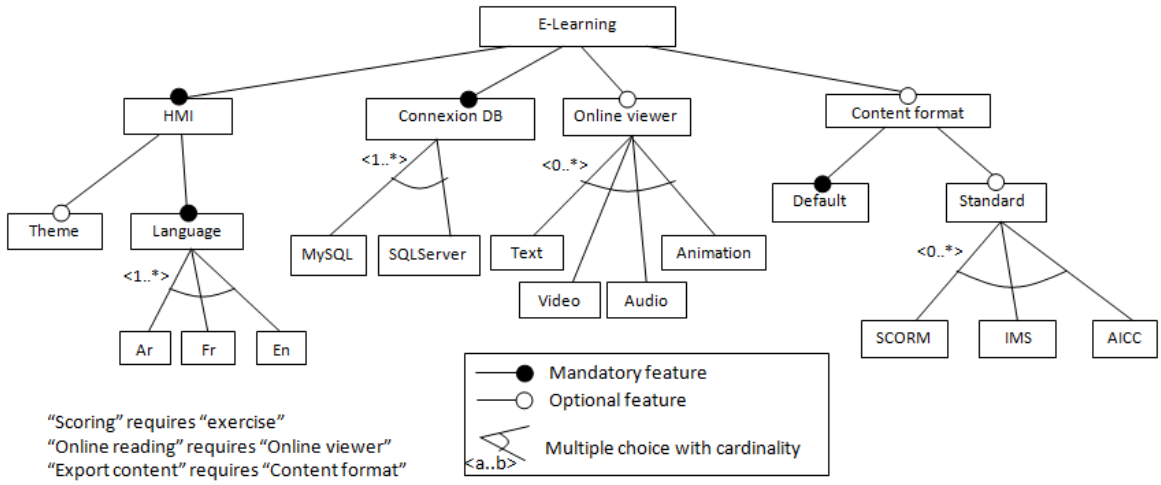


Figure 3. Implementation feature diagram for e-Learning product line.

1) Domain Analysis: The goal of domain analysis is to extract and document the similarities and variations between the SPL members. To document the common and variable features of our product line, we have used the Feature Model. The Feature Model is the first language dedicated to the modeling of variability; it was first introduced in the Feature-Oriented Domain Analysis (FODA) method [4]. It has known a broad use in the field of SPLE and several extensions [13][14][15], since it is a simple and easy to use language in comparison with other more complex modeling languages such as: Unified Modeling Language (UML) [23] and Business Process Modeling Notation (BPMN) [24]. The feature model is generally described by a hierarchy of the set of features of a system or what is called feature tree [5]. Figures 2 and 3 show a part of the feature model of our case. For the notation, we must note that in the case of multiple choices we have used only cardinality notation to avoid cluttering the diagram with different notations, for us cardinality is sufficient to represent all kinds of choices.

The feature model we constructed is divided into two diagrams according to the type of features it includes. Features in the first diagram called capability features, (Figure 2), represent the functionalities provided by the system. This diagram shows that the main features of an e-Learning application are "Profile management" and "Course management" and "Configuration". However, the application may include other functionalities such as:

- Online payment in the case of paid courses provided for example by private organizations.
- Interaction with learners through "Collaboration tools". Collaboration tools may be synchronous (chat, video conferencing, etc.) or asynchronous (e-mail, forum, wiki, etc.).
- Additional modules such as: statistics, search, download official documents (bulletin, attestation, etc.) and others. The cardinality 0..* for the feature "Additional module" means the possibility of adding new sub-features and so the possibility to extend the product line to cover new requirements.

A "Course management" must contain at least "Content" and "Enrollment" features, but it can include other optional features according to the usage context, such as: "groups' management", "Export content", and "Evaluation". The evaluation (if selected) may include several types of questions, for instance: in some cases text questions are sufficient, in other cases diagrams or audio recordings are needed. We do not show the whole feature model for the sake of brevity and space reasons.

The second diagram reported in Figure 3 represents the implementation features of the system; it means implementation details at lower and more technical levels. An e-Learning application must connect to a data base and supply a Human Machine Interface (HMI). If the application provide "Online reading" of the course's content, this require an "Online viewer" which differ according to the type of the content (text, video, sound or animation). The content of a course can be exported in several formats: default format provided by the system or other standard formats such as: Sharable Content Object Reference Model (SCORM) [26], IMS Global Learning Consortium (IMS GLC) [27] or Aviation Industry Computer-Based Training Committee (AICC) [28]. Constraints at the bottom of the diagram are used to express dependencies between features in the same diagram or between capability and implementation features.

2) Domain Design: The purpose of the domain design is to establish the generic software architecture of the product line. Variability identified during domain analysis must be explicitly specified in the product line architecture.

In our case, we have chosen Orthogonal Variability Model (OVM) [5] to represent variability in the design model. OVM consist of a set of Variation Points (VP) and Variants (V). OVM is based on a separation between variability model and other artefacts in order to decrease models complexity [16]. A variation point or variability subject shows an aspect of variability within the product line. Variants or variability objects are the different shapes of a variability subject. Using OVM allows us to represent variability in the architecture view without having to extend the design language. Variability is modeled separately from the component diagrams and related to this latter by means of traceability links.

 ISBN: 978-1-61208-361-2

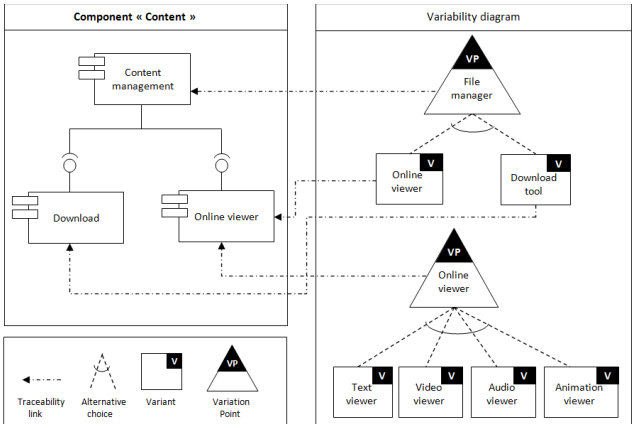


Figure 4. Variability model for "content" component.

To create our OVM, we relied on the Feature Model constructed in domain analysis phase. Figure 4 shows the variability modeling of the component "content" from the set of components of e-Learning product line. In this example, the component "Online viewer" will be implemented in one or more versions: text viewer, video viewer, audio viewer, or animation viewer.

3) Domain Realization: The main object of this step is to create a set of reusable software components. The components that have been identified in the previous step are detailed, planned and implemented to be reused in different contexts. To build our components we have used Java Enterprise Edition (J2EE) [25]. The result of this step is not a running application, but rather, a set of configurable and loosely coupled reusable components, that will be assembled during application derivation step.

B. Application Engineering

One can distinguish between two kinds of variability [18]: (i) product line variability which is specific to SPLE and describes the variation between the members of a SPL, and (ii) software variability that refers to the ability of a software element to be changed or customized for use in a particular context. Product line variability is resolved (bound) during Application Engineering through several binding times (design, implementation, compilation, assembly) [17], while software variability is bound after the delivery of an application (configuration and runtime). In order to ensure an efficient support of customer's needs, customers are involved not only at runtime but also at application derivation (Application Engineering) as well as application configuration steps (Figure 5).

During application engineering, e-Learning applications are derived using the core assets and customer's specific requirements. The feature model that we have defined represents the decision space for our SPL. For each new application, we select the relevant features from the feature diagram according to the specific customer requirements. The feature model of the particular application is then used to specify which variants must exist in the architecture model. As a result, we obtain an architecture model without variability, and which includes only the components of the derived application, in addition to components that implement the specific requirement if they exist. Finally, according to the application's architecture model, we select the components from the base of reusable components obtained in domain realization. In the case where the particular application needs components that have not been predicted in Domain engineering, these application-specific components must be implemented and then assembled with the selected reusable components to create the final running application.

During application configuration, customer can decide about variation points that have been delayed to this binding time, for instance, he can specify the language, the website address and name, the database driver, the administrator profile, etc. Other variations such as: courses organization, additional modules, theme selection, access rights of users, etc., might be delayed until runtime. Allowing customers to customize their applications through several steps (derivation, configuration and runtime) lead to more flexible applications and thus better user satisfaction.

As mentioned in section III, customers could be teachers in schools and universities and institutions providing free or paid online courses. When delivering the final applications, they will have neither extra-functionalities that they does not need, nor lacking functionalities that they must integrate themselves as in the case of LMSs. So, the customer's satisfaction is assessed according to the conformance of the provided functionalities to the required ones.

IV. Related Work

SPL was first used in e-Learning domain to develop and reuse digital educational content [7][8]. Pankratius et al. proposed the Product Lines for Digital Information Products (PLANT) approach to deal, in a general way, with the issues encountered in content reuse for e-Learning platforms. In this case, the reusable elements are a mixture of content and software, since online courses may contain, more than texts, programs and animations.

Another work using SPL approach to develop an auxiliary e-Learning application is presented by Sanchez et al. [10]. They use SPL engineering to develop e-Learning Web-miner product line, a family of data-mining applications aiming to assist educators involved in virtual education by extracting and providing useful information that these educators can use to improve the learning-teaching process.

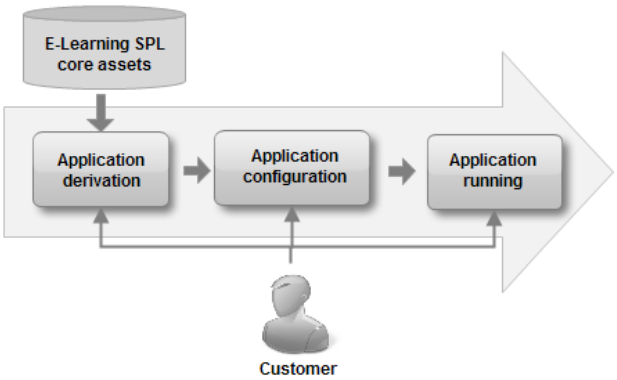


Figure 5. Application customization.

 ISBN: 978-1-61208-361-2

To the best of our knowledge, the use of SPL principles in the domain of e-Learning was limited to reuse online courses [7][8], or to develop data-mining applications related to e-Learning platforms [10]. But, there is no other SPL for e-Learning applications similar to the one we presented in this paper. By the present work, we show that e-Learning is a wide domain that includes several applications, characterized by an important set of common features and vary in some aspect, the adoption of SPL approach in this field can obviously bring important benefits not only to developers but also to users.

V. CONCLUSION AND FUTURE WORK

In this paper, we proposed the use of SPLE approach to develop e-Learning applications. The presented work aims to overcome the shortcomings of LMSs, mainly by satisfying the variable requirements of customers, providing more flexible applications, and to benefit from the advantages of SPL engineering. Using SPL approach in such a broad field allows developers to reduce costs and effort of both production and maintenance, to decrease significantly time of development and to improve quality.

The paper presented the different steps of the development process of an e-Learning product line, focusing on domain engineering. This latter result in a set of core assets: the domain requirements documented by the feature model, the reference architecture models including variability presented by OVM model and the software components. This base of core assets will be reused to simplify the development of each new member of e-Learning SPL during application engineering. Moreover, customers are prompted to express their needs throughout the application instantiation steps in order to reach better satisfaction.

As future work, we intend to improve our e-Learning product line by decomposing it into a set of sub-SPLs, each one intended for an e-Learning subfield (primary, secondary, university, paid courses, etc.). This will allow us to cover a broader scope while ensuring efficient variability management. It will be also important to define an automatic method of derivation to improve the application engineering process.

REFERENCES

[1] F. Bachmann and P. Clements, “Variability in software product lines,” technical report CMU/SEI, 2005.

[2] L. Chen, M. Babar, and A. Nour, “Variability management in software product lines: a systematic review,” in SPLC, San Francisco, California, 2009, pp. 81-90.

[3] C. Dalsgaardy, Social software: e-learning beyond learning management systems. European Journal of Open, Distance and E-Learning, 2006, no. 2, [Online]. Available from: http://www.eurodl.org/index.php?article=228 2014.6.6

[4] K. Kang, S. Cohen, J. Hess, W. Nowak, and S. Peterson, “Feature oriented domain analysis (FODA) feasibility study,” Technical Report CMU/SEI-90-TR-21, 1990.

[5] P. Klaus, G. Bockle, and F. van der Linden, Software product line engineering: foundations, principles, and techniques. Springer, 2005.

[6] L. Northrop and C. Clements, “A framework for software product line practice,” Version 5.0. [Online]. Available from: http://www.sei.cmu.edu/ 2014.6.6

[7] V. Pankratius, Product Lines for Digital Information Products. Information Systems, 2007.

[8] V. Pankratius and S. Wolried, “A strategy for content reusability with product lines derived from experience in online education,” in International Conference on Software Engineering (ICSE), USA, 2005, pp. 128-146.

[9] M. Paulsen, “Experiences with learning management systems in 113 european institutions,” Educational Technology and Society, vol. 6 (4), 2003, pp. 134-148.

[10] P. Sanchez, G. Diego, and Z. Marta, “Software product line engineering for e-learning applications: a case study,” in 2012 International Symposium on Computers in Education (SIIE 2012), Andorra, 2012, pp. 1-6.

[11] F. V. der Linden, and E. R. K. Schmid, Software product lines in action: the best industrial practice in product line engineering. Springer, 2007.

[12] R. Watson, “An argument for clarity: what are learning management systems, what are they not, and what should they become?” TechTrends, vol. 51, 2007, pp. 28–34.

[13] K. C. Kang, K. Sajoong, L. Jaejoon, K. Kijoo, J. Gerard, and S. Euiseob, “FORM: A feature-oriented reuse method with domain-specific reference architectures,” Annals of Software Engineering, vol. 5, 1998, pp. 143–168.

[14] M. L. Griss, F. John, and A. Massimo, “Integrating feature modeling with the RSEB,” in Proceedings of the Fifty International Conference on Software Reuse, Victoria, Canada, 1998, pp. 76-85.

[15] K. Czarnecki and C. H. P. Kim, “Cardinality-based feature modeling and constraints: a progress report,” in International Workshop on Software Factories at OOPSLA’05, San Diego, California, USA, 2005, pp. 16-20.

[16] K. Pohl and A. Metzger, “Variability management in software product line engineering,” in Proceedings of the 28th international conference on Software engineering, New York, NY, USA, 2006, pp. 1049-1050.

[17] R. Capilla and J. Bosch, “Binding time and evolution,” Systems and software variability management, 2013, pp. 57–73.

[18] A. Metzger, K. Pohl, P. Heymans, P. Schobbens, and G. Saval, “Disambiguating the documentation of variability in software product lines: a separation of concerns,” in 15th IEEE International In Requirements Engineering Conference, Delhi, 2007, pp. 243 - 253.

[19] F. J. García-Peñalvo, M. Á. Conde, M. Alier, and M. J. Casany, “Opening learning management systems to personal learning environments,” Journal of Universal Computer Science, 17(9), 2011, pp. 1222-1240.

[20] V. Stantchev, R. Colomo-Palacios, P. Soto-Acosta, and S. Misra, “Learning management systems and cloud file hosting services: A study on students’ acceptance,” Computers in Human Behavior, Vol. 31, February 2014, pp. 612–619.

[21] M. A. Conde, F. García, M. J. Rodríguez-Conde, M. Alier, and A. García-Holgado, "Perceived openness of Learning Management Systems by students and teachers in education and technology courses,"Computers in Human Behavior, Vol. 31, February 2014, pp. 517-526.

[22] Z. Du, X. Fu, C. Zhao, Q. Liu, and T. Liu, "Interactive and Collaborative E-Learning Platform with Integrated Social Software and Learning Management System,"Proceedings of the 2012 International Conference on Information Technology and Software Engineering, Lecture Notes in Electrical Engineering, Vol. 212, 2013, pp 11-18.

[23] T. Baar, A. Strohmeier, A. Moreira, and S. J. Mellor, UML 2004 - The Unified Modeling Language. Springer, 2004.

[24] J. Mendling, M. Weidlich,and M. Weske , Business Process Modeling Notation. Springer, 2011.

[25] E. Armstrong, J. Ball, S. Bodoff, D. B. Carson, I. Evans, D. Green, and E. Jendrock, The J2EE 1.4 tutorial. Sun Microsystems, 2004. Available from: http://docs.oracle.com/javaee/1.4/tutorial/doc/ 2014.7.4

[26] “Sharable Content Object Reference Model,” [Online]. Available from: http://scorm.com/scorm-explained/ 2014.7.4

[27] “IMS Global Learning Consortium,” [Online]. Available from: http://www.imsglobal.org/ 2014.7.4

[28] “Aviation Industry Computer-Based Training Committee,” [Online]. Available from: http://www.aicc.org/joomla/dev/ 2014.7.4

Collaboration and community building in an on-line Teacher Community of learning: A Social Network Analysis

Panagiotis Tsiotakis
Department of Social and Educational Policy
University of Peloponnese
Korinthos, Greece
e-mail: ptsiotakis@uop.gr

Athanassios Jimoyiannis
Department of Social and Educational Policy
University of Peloponnese
Korinthos, Greece
e-mail: ajimoyia@uop.gr

***Abstract*—This paper presents an investigation of teachers' engagement and collaboration toward building a Teacher Community in the context of a blended post-graduate course about e-learning and ICT in education. The design of both pedagogical and technological dimensions of the teacher community framework are presented. Research data were analysed using Social Network Analysis methods and revealed important information regarding critical indicators of interaction and collaboration among participants and the whole community performance. Conclusions are drawn for future development and research about on-line teacher communities.**

Keywords- On-line Teacher Communities, community learning; Social Network Analysis.

I. INTRODUCTION

Virtual learning communities are generally thought as social structures that provide enhanced opportunities to communicate and collaborate with peers sharing the same interests, and to continually support learning and professional development of the participants [1][2][3][4]. An efficient community requires a set of rules, habits, strong ties and interaction between members to be established in order to achieve common goals [5]. In the last decade, the idea of Teacher Communities (TC) has received a growing interest among academics, policy makers and educators as an alternative to both isolated manner of work and the traditional teacher professional development approaches [6][7][8].

In the context of situated learning, many scholars have come to emphasize learning in professional communities as a dynamic and social participation process, captured in collaborative activities, working artifacts, routines, stories or perceptions [1][9][10]. The main idea is to support teachers' learning and collaboration, and strengthen their professional work by a) sharing common interests, experiences, educational resources and material, b) developing meaning and constructing new knowledge in a participatory and collaborative way, and c) enhancing their ability to put innovative instructional approaches into practice [11].

The widespread interest about on-line TC is rooted in their potential to create unique conditions for informal learning and a sustainable environment for teacher interaction and collaborative learning. There is growing research evidence supporting the impact of communities on teachers' professional development as well as on student achievements [7][11][12][13]. TC are also claimed to contribute to the improvement of teaching and schooling practices [7][11][12] while they are considered as a way to embed teacher collaboration into the school culture [13]. In addition, on-line TC constitute a promising idea and a new model for teacher professional development [8][14][15]. The rapid growth and diffusion of Web 2.0 tools has led to increased interest in creating dynamic on-line TCs due to their affordances to support, without temporal or spatial restrictions, sustainable participatory environments for communication, interaction, content sharing, collaboration, self-directed and collaborative learning, peer- and self-assessment [7][8][15][16].

Many ideas have been proposed to describe decentralized, on-line learning communities where members work together and support each other, use a variety of tools and resources, and endeavour to achieve their learning goals through collaboration and problem solving activities [17][18]. Since each member offers his knowledge, experience, abilities and creations to the whole community, he/she can contribute decisively to the establishment of *collective thinking* and *sharing knowledge* among participants [10].

The design, implementation and evaluation of on-line TCs are still an open research problem in both teacher development and e-learning contexts [1][7][11][16][19]. This paper reports on the investigation of teachers' engagement and collaboration toward building a structured TC, designed and implemented in the context of a blended post-graduate course about e-learning and ICT in education. The contribution of this study is, therefore, two-fold: a) to extend previous research findings concerning TCs in formal learning contexts by using an authentic learning design framework [20] driven by the ideas of social learning [5][10], and b) to apply a combined analysis of teachers' learning-community activities using Social Network Analysis (SNA) methods, in order to shed light into the different ways of individual contributions, the dynamics of social interaction, and member ties appeared within the community.

The paper is structured as following. The design framework of the TC and its pedagogical and technological dimensions are outlined. Teacher learning activities and the community workflow are presented in detail. Preliminary SNA findings regarding teachers' participation, interaction and collaboration, as well as the community structure are presented. Finally, conclusions are drawn for future development and research in the area of on-line TCs.

 ISBN: 978-1-61208-361-2

II. TEACHER COMMUNITY DESIGN FRAMEWORK

According to Wenger [5], there are three core dimensions in a TC, which reflect the nature of the community, what it is about and how it functions:

- **Group identity:** Mutual engagement that bind teachers together in a social entity.
- **Shared domain:** A joint enterprise as understood and continually negotiated by community members.
- **Shared interactional repertoire:** Shared practices, communal resources and beliefs that teachers have developed over time.

Literature review suggests that asynchronous discussion forums and traditional Learning Management Systems (LMS) were widely used to support on-line TC [2][3[[21]. However, LMS are tutor-centred environments, designed to support e-learning in the context of formal education; they offer limited opportunities for learner-directed initiatives and actions. A TC platform should incorporate features and tools beyond conventional LMS.

Towards outlining a conceptual and pedagogical framework of community interactions and collaborative learning, we have defined four constitutional components in relation to the dimensions of a TC [22]:

Content sharing area: It includes various content sharing tools (file repository, blog, wiki, tagging, Web links, etc.).

Communication area: This component is structured around various communication tools (messaging, chat, discussion forum, videoconference, microblogging).

Community area: Community pages, group supporting, e-portfolio, task schedule and content management are the main tools of this component.

Personal and supportive tools: Personal repository, dashboard, timeline, profile and searching tools.

However, there are tools that could be assigned into more than one conceptual category. For example, wiki acts as both, content sharing and communication tool.

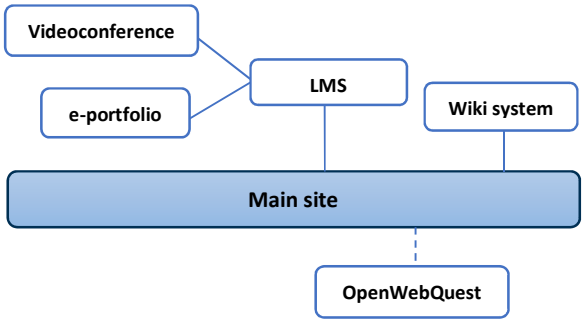


Figure 1. Architecture and components of TC platform.

Using the conceptual framework above, the TC platform was designed to provide in an integrated way a variety of constructive, collaborative and community tools [22]. Figure 1 presents the architecture of the TC platform and its components. The main site supported by Joomla offers a single-sing-on access to the community platform. The Learning Management System (Moodle) operates as both, the instructional/tutor area and the community dashboard, offering information about current news and events, forum topics, blog articles, group members, recent activities, etc. Linking to the other components of the platform, e.g., the e-portfolio system (Mahara), the wiki system (Mediawiki), videoconference (BigBlueButton), is also available through Moodle. OpenWebQuest (http://openwebquest.org) is an additive tool developed in this project with the aim to support teachers to create and publish their own WebQuest-scenarios, and share them with peers in the community.

III. THE STUDY

A. Research questions

In accordance with the research objectives and consistent with the related literature, the following research questions were addressed in this study:

- What type of teacher activities were effective and facilitated learning presence within the community? To what extent did teachers participate and contribute to the community learning activities?
- What were teachers' patterns of presence and social interaction within the TC? What were the main features of the structure and the dynamics of the TC?

B. Participants and context

The community ran during spring semester of 2013, in the context of a masters' degree course entitled "e-learning and ICT in education", at the Department of Social and Educational Policy, University of Peloponnese, in Greece. Twenty three students (20 of them were in-service primary and secondary education teachers) attended this course. The teachers in the sample, though familiar with ICTs and the Web, had no previous experience with on-line collaborative tools and e-learning platforms; they had never before participated in learning communities.

C. Community activities and workflow

Our educational intervention followed the philosophy of structured community of learning [11][14] and the CIMO-logic design model [23] by treating in an integrated manner the four community components: Context, Intervention, Mechanism and Outcomes. Using a blended format, it included five (5) face-to-face sessions in the classroom combined with on-line collaborative work, between January and June 2013. Following the pilot study [22], we shaped an ongoing cooperation framework with the aim to support a high level of dialogue, interaction, and collaboration among members. The instructor (T) was acting as the e-moderator by setting the context, the expectations and the processes of the community-based learning. Guidelines and technical assistance were also given by the course assistant.

Teachers were asked to work both individually and collaboratively. They were free and encouraged to contribute by reflecting on themes presented in both face-to-face and on-line sessions, starting new discussion topics, debating and interchanging ideas, sharing experiences, writing articles in the community blog, commentating peer contributions, uploading content material, suggesting information/content

 ISBN: 978-1-61208-361-2

resources, creating specific interest groups, undertaking roles and responsibilities within their group, designing collaboratively educational scenarios applicable in school practice, etc. However, each student-teacher was requested a) to write on the journal-blog area, at least, one article per month (5 in total) and b) to create and publish a WebQuest scenario using the OpenWebQuest platform. They were also asked to collaboratively create new educational artefacts (articles and documents, learning scenarios and educational material). To achieve this goal, they were encouraged to use e-portfolio (Mahara) and wiki (Mediawiki) subsystems for communication, ideas sharing and negotiation, co-authoring and co-creating.

During this period, the participants were exposed into detailed on-line discussions regarding various issues arising into the community. Teachers' individual and collaborative work was visible within the platform. In addition, participants were continually informed about any community event by receiving e-mail notifications through the platform.

D. *Sourse data and analysis*

Each member contribution was considered as the analysis unit. We used, therefore, three main sources of data:

- Postings to various topics in the discussion forums
- Publications on e-portfolio subsystem (articles/pages authoring, commentaries on peer articles, educational scenarios proposed etc.)
- Contributions to the collaborative authoring of educational scenarios and to the related discussion topics on wiki-pages (Mediawiki subsystem).

The analysis of students' engagement, learning presence, and individual position within both, teacher groups and the whole community, was implemented by using SNA methods through NetMiner 4.0 (http://www.netminer.com). SNA has been effectively applied to analyze network operation (e.g., interactions among members, communication, information exchange, knowledge sharing, etc.) and community structure in various e-learning situations [24][25][26]. It provides a set of algorithms to quantify and give insight into member relations and group dynamics in terms of network structure parameters, e.g., *cohesion*, *power centrality* and *betweeness centrality*. In addition, SNA provides multiple graphs which represent the relations-interconnections among members, individual contributions as well as the structure-operation of the whole community.

IV. RESULTS

A. *Descriptive analysis of teachers' contributions*

Table I depicts an overall picture of teachers' engagement and the main community activities/contributions recorded for each member: a) teacher working groups they participated in, b) individual articles published in the journal area, c) article commentaries received by community members, d) article views by the other community members, and e) posts they uploaded in the community forum.

It is quite clear that the majority of the participants were active community members. A total of 14 working groups were created during the community timeline. They were not addressed by the tutor who created just one group. Rather they appeared as the outcome of teachers' interests and their spontaneous initiatives to collaboratively work with peers in order a) to study a new educational topic and b) to design new educational scenarios. Teachers S15 and S22 created three groups, S23 initiated two groups, and members S6, S14, S17, S19, S20 and S22 created one group each. Within working groups, 14 pages were recorded and 6 complete learning scenarios were collaboratively created in the wiki system. In addition, 21 WebQuests were individually constructed and were available to the community members for commenting and peer reviewing. To organize and support their work and collaboration the teachers uploaded, in total, 206 forum postings.

A total of 135 original articles were published in the e-portfolio area which were dealing with both, theoretical and practical themes; they were related to various educational topics (e.g., contemporary pedagogy and ICT, learning design and scenarios, educational practices with ICT, e-learning, Web 2.0 tools in practice etc.). Comprehensive discussions were evolved around the topics above; 647 article commentaries were uploaded. The number of views per article is also an indicator of teachers' social interactions through sharing their ideas, beliefs and creations.

TABLE I. MEMBER ACTIVITIES WITHIN COMMUNITY

Member	*Member in groups*	*Articles*	*Article Comm.*	*Article Views*	*Forum Posts*
S1	4	7	23	641	6
S2	5	5	7	491	3
S3	3	7	46	729	17
S4	4	6	27	357	8
S5	4	5	15	318	2
S6	6	5	22	558	20
S7	3	7	58	594	3
S8	5	9	25	674	15
S9	1	6	15	239	2
S10	6	4	20	338	10
S11	3	0	1	0	0
S12	4	6	1	270	0
S13	7	6	66	425	4
S14	7	6	33	434	5
S15	10	6	41	555	8
S16	8	6	23	315	1
S17	8	6	19	420	1
S18	6	4	15	375	21
S19	3	7	19	417	2
S10	11	6	25	308	6
S21	5	7	44	446	2
S22	8	5	24	591	40
S23	8	7	48	588	24
T1	8	2	30	591	6
Total	*14*	*135*	*647*	*10674*	*206*

Figure 2 shows a screenshot of the structural (main) page of a typical teacher e-portfolio. It presents her activities and the artifacts she produced during the community workflow. It is organized in three columns projecting teacher articles, individual reviews and educational scenarios, wiki groups, suggested Web links and references. The majority of the teachers used a similar e-portfolio structure, organized in two or three columns.

 ISBN: 978-1-61208-361-2

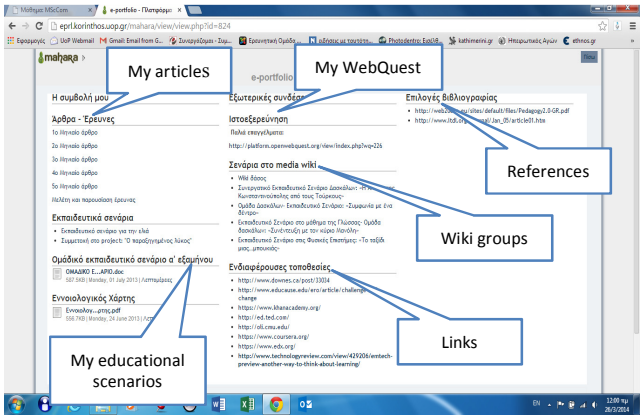


Figure 2. Structure of teachers' e-portfolio.

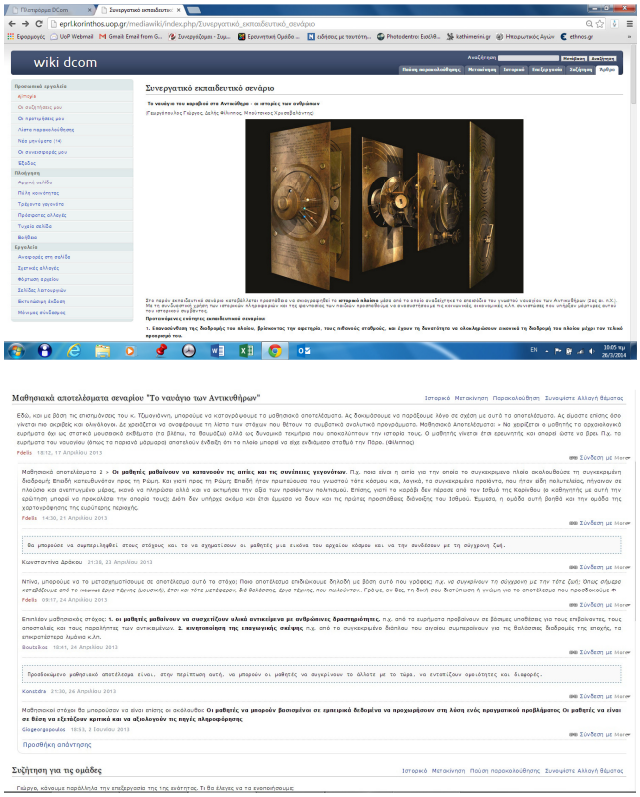


Figure 3. Main wiki page and the related discussion supporting the collaborative development of an educational scenario.

The upper screenshot in Figure 3 presents a wiki page hosting an educational scenario about the Antikythera mechanism (Astrolabe) which was collaboratively created. Teachers' discussion posts showing the negotiation of new ideas and suggestions regarding the expected learning outcomes of this scenario are presented at the bottom.

B. Social Network Analysis

Individual contribution, member relations, group dynamics and community operation were analyzed in terms of three network structure parameters, namely cohesion, power centrality and betweeness centrality. Cohesion analysis aims at revealing the architecture of the TC, e.g., the existence of cliques (subgroups) of community members who were connected internally more than externally. If the *cohesion index* in a sub-group is greater than 1 then the links within the clique are stronger on average than the links with the members outside [26]. A total of 58 cliques revealed in the community. The e-portfolio sub-system was adopted by the teachers as the main area around which their community activities were evolved; 49 cliques were recorded there. In addition, 4 cliques were in Mediawiki, and 5 in the Moodle forum topics.

It is important to be pointed here that the majority of the Mahara cliques (46 of them) included a great number of members, ranging from 7 to 12. This indicates that the community subgroups were very cohesive. In other words, TC members tended to develop strong interrelations and a wide scope of interaction offering enhanced opportunities for collaborative knowledge construction and knowledge sharing among teachers.

Power (or centrality) analysis is an effective SNA method to measure network activity, to reveal the operation of the community network and to assess the impact each member had with respect to spreading information and influencing others in the community [26][27]. *In-degree centrality* represents the number of interactions a teacher receives from other members in the community. Accordingly, *out-degree centrality* is the number of connections a teacher has to the other members. *Betweensess centrality* represents the capacity of a teacher to act as a connector between other members, e.g., it is an indicator of individual position within the community.

TABLE II. POWER ANALYSIS OF THE TC

Teacher	*In-degree Centrality (%)*	*Out-degree Centrality (%)*	*Node Betweeness Centrality*
S1	50.00	58.33	0.00718
S2	25.00	25.00	0.00108
S3	75.00	83.33	0.02999
S4	66.67	54.17	0.00653
S5	37.50	54.17	0.00367
S6	70.83	50.00	0.00861
S7	70.83	83.33	0.03263
S8	66.67	58.33	0.05189
S9	16.67	33.33	0.00129
S10	79.17	62.50	0.0264
S11	4.17	4.17	0
S12	33.33	4.17	0.00016
S13	62.50	83.33	0.02454
S14	58.33	75.00	0.02195
S15	79.17	79.17	0.10178
S16	45.83	54.17	0.01652
S17	70.83	41.67	0.0132
S18	45.83	50.00	0.01682
S19	66.67	50.00	0.00841
S10	62.50	54.17	0.00752
S21	54.17	66.67	0.01351
S22	75.00	75.00	0.06532
S23	70.83	75.00	0.02398
Average	53.50	53.50	0.02

 ISBN: 978-1-61208-361-2

Table II shows the results of the network activity measures and presents the power distribution among members in the community. The great majority of the teachers were active community members since they have interacted, at least, with 50% of their peers. Teachers S10 and S15 were the most influential members, since they received a great number of connections (posts) from their peers (79.17%). Teachers S3, S7 and S13 were the most effective and successful members in the community towards triggering other teachers (they were connected with 83.33% of the participants). On the other hand, S9 and S11 had a marginal community contribution, since they influenced only one member (4.17%).

Figure 4 shows the power centrality map of teachers' activities, which includes member connections through the main systems of the community platform (LMS, e-portfolio, wiki). It is a measure of the influence each participant had in the community. A large group of teachers were placed near the center of the map (e.g., S3, S10, S22, S15, S4, S6, S7, S8, S17, S19, and S23). They were the most active, influential and powerful members in the community and they had many ties and connections to other powerful participants. On the other hand, as moving to the periphery, teachers were less powerful and important community members, e.g., S9 and S11 had a marginal community contribution.

Figure 5 presents the betweeness (intermediation) centrality map. Teacher S15, who placed at the center, was the most effective member to connect others and, consequently, he had more control of the interaction and information interchange within the community. Teachers S22, S3, S10, S23, S14, S4, and S7 were also good connectors compared to their peers in the periphery. As an overall view, this was a very cohesive community; the majority of the participants had significant contribution while there is only one member (S11) with marginal engagement.

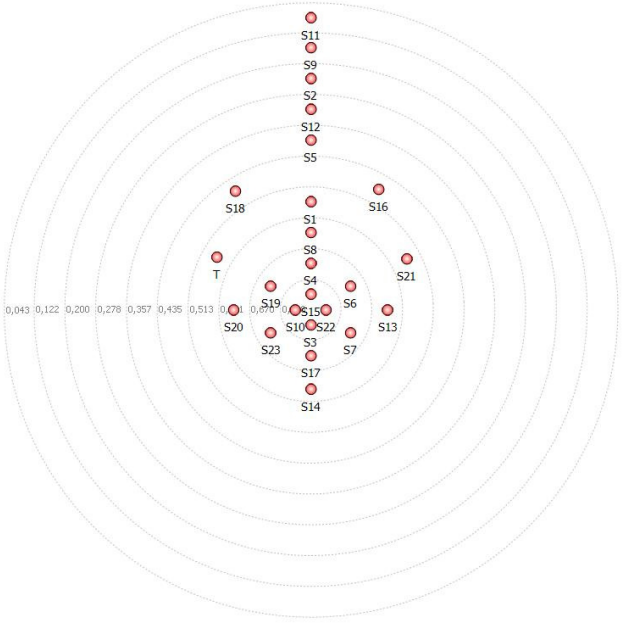


Figure 4. Power centrality map.

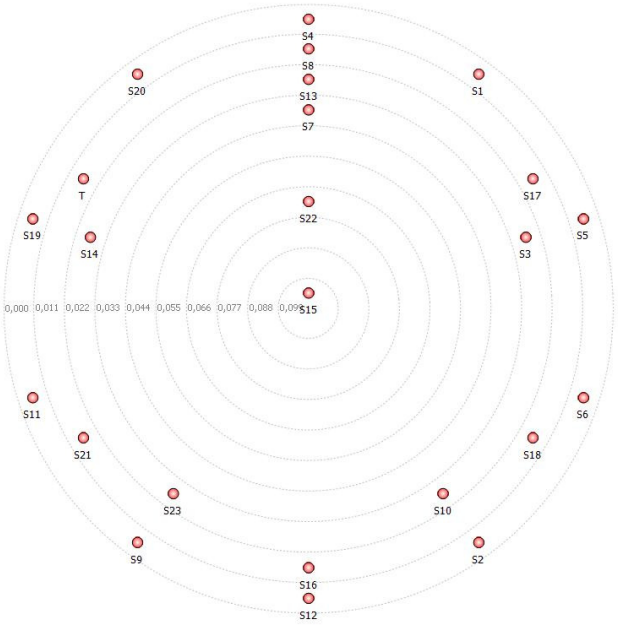


Figure 5. Betweeness centrality map.

V. CONCLUSIONS AND FUTURE WORK

This paper reported on an investigation concerning a TC, designed to support blended and collaborative learning, in the context of a post-graduate course. The research findings of SNA provided supportive evidence that the structured community framework presented here was effective and offered valuable source data for the investigation of teachers' engagement and learning actions occurring within a TC.

The results revealed important information about the structure and the cohesion of the community, the teacher groups developed therein, teacher connections and information flow, as well as the power and the influence each participant had within the community. The majority of the participants demonstrated enhanced interest and they were actively engaged into the community activities (uploading articles and postings, supporting dialogue and discussion topics, interchanging ideas, sharing content and resources, co-creating educational material, etc.).

The outcome of this non-formal e-learning program was a promising evidence of a decentralized teacher community. SNA revealed that the tutor was not the central member of the learning community network. On the other hand, teachers' engagement and interaction patterns were strong indicators of a) enhanced learner control and motivation to keep learning activities evolving, and b) supporting collaboration and building a community of learning among participants.

This investigation contributes to the existing knowledge and could guide both, future research in this area as well the design and the implementation of efficient on-line TC [6][14]. Our current efforts are directed towards combining SNA findings with qualitative data from content analysis of teachers' on-line discourse and teachers' interviews to further

 ISBN: 978-1-61208-361-2

analyze their engagement, learning presence and knowledge construction within the community. In addition, a comparative analysis of data extracted from an open, non-structured TC is under completion.

ACKNOWLEDGMENT

The authors are grateful to the students/teachers who participated in this teacher community and in the research.

REFERENCES

[1] B. Baran and K. Cagiltay, "The dynamics of online communities in the activity theory framework", Educational Technology & Society, 2010, vol. 13(4), pp. 155-166.

[2] A. P. Correia and N. Davis, "Intersecting communities of practice in distance education: the program team and the online course community", Distance Education, 2008, vol. 29(3), pp. 289-306.

[3] M. Delfino, G. Dettori, and D. Persico, "Self regulated learning in virtual communities", Technology, Pedagogy and Education, 2008, vol. 17(3), pp. 195-205.

[4] C. Gray and K. Smyth, "Collaboration creation: Lessons learned from establishing an online professional learning community", The Electronic Journal of e-Learning , 2012, vol. 10(1), pp. 60- 75.

[5] E. Wenger, "Communities of practice: Learning, meaning, and identity", New York: Cambridge University Press, 1998.

[6] P. Brouwer, M. Brekelmans, L. Nieuwenhuis, and R.-J. Simons, "Fostering teacher community development: A review of design principles and a case study of an innovative interdisciplinary team", Learning Environments Research, 2012, vol. 15, pp. 319-344.

[7] T. O. Jackson, "Towards collective work and responsibility: Sources of support within a freedom school teacher community", Teaching and Teacher Education, 2009, vol. 25, pp. 1141-1149.

[8] A. Jimoyiannis, M. Gravani, and Y. Karagiorgi, "Teacher professional development through Virtual Campuses: Conceptions of a 'new' model". In H. Yang & S. Yuen (eds.), Handbook of Research on Practices and Outcomes in Virtual Worlds and Environment, 2011, pp. 327-347, Hershey, PA: IGI Global.

[9] J. Lave and E. Wenger, "Situated learning: Legitimate peripheral participation", Cambridge: Cambridge University Press, 1991.

[10] G. Siemens, "Learning Ecology, Communities, and Networks extending the classroom", 2003, Retrieved 2011.12.10 from http://www.elearnspace.org/Articles/ learning_communities.htm.

[11] T. H. Levine and A. S. Marcus, "How the structure and focus of teachers' collaborative activities facilitate and constrain teacher learning?", Teaching and Teacher Education, 2010, vol. 26, pp. 389-398.

[12] A. Skerrett, "There's going to be community. There's going to be knowledge': Designs for learning in a standardised age", Teaching and Teacher Education, 2010, vol. 26, pp. 648-655.

[13] V. Vescio, D. Ross, and A. Adams, "A review of research on the impact of professional learning communities on teaching practice and student learning", Teaching and Teacher Education, 2008, vol. 24, pp. 80-91.

[14] P. Graham, "Improving teacher effectiveness through structured collaboration: A case study of a Professional Learning Community", Research in Middle Level Education, 2007, vol. 31(1), pp. 1-17.

[15] A. L. Luehmann and L. Tinelli, "Teacher professional identity development with social networking technologies: Learning reform through blogging", Educational Media International, 2008, vol. 45(4), pp. 323-333.

[16] H. T. Hou, K. E. Chang, and Y. Ting Sung, "What kinds of knowledge do teachers share on blogs? A quantitative content analysis of teachers' knowledge sharing on blogs", British Journal of Educational Technology, 2010, vol. 41(6), pp. 963–967.

[17] J. S. Brown and P. Duguid, "Organizational learning and communities-of-practice: Toward a unified view of working, learning, and innovation", Organization Science, 1991, vol. 2(1), pp. 40–57.

[18] B.G. Wilson, "Metaphors for instruction: why we talk about learning environments", Educational Technology, 1995, vol. 35(5), pp. 25-30.

[19] W.-M. Roth and Y.-J. Lee, "Contradictions in theorizing and implementing communities in education", Educational Research Review, 2006, vol. 1, pp. 27-40.

[20] J. Herrington and L. Kervin, "Authentic Learning Supported by Technology: Ten suggestions and cases of integration in classrooms", Educational Media International**,** 2007, vol. 44, pp. 219-236.

[21] J. M. Zydney, A. deNoyelles, and K. Kyeong-Ju Seo, "Creating a community of inquiry in online environments: An exploratory study on the effect of a protocol on interactions within asynchronous discussions", Computers & Education, 2012, vol. 58, pp. 77-87

[22] P. Tsiotakis and A. Jimoyiannis, "Developing a Computer Science Teacher Community in Greece: Design framework and implications from the pilot", Procs. of EDULEARN13 Conference, 2013, pp. 70-80, 1st-3rd July 2013, Barcelona, Spain.

[23] D. Denyer, D. Tranfield, and J. E. van Aken, "Developing design propositions through research synthesis", Organizational Studies, 2008, vol. 29, pp. 393-413.

[24] A. Martínez, Y. Dimitriadis, B. Rubia, E. Gómez, and P. de la Fuente, "Combining qualitative evaluation and social network analysis for the study of classroom social interactions", Computers & Education, 2003, vol. 41(4), pp. 353-368.

[25] P. Shea, S. Hayes, J. Vickers, M. Gozza-Cohen, S. Uzuner, R. Mehta, A. Valchova, and P. Rangan, "A re-examination of the community of inquiry framework: Social network and content analysis", The Internet and Higher Education, 2010, vol. 13(1), pp. 10-21.

[26] A. Jimoyiannis and S. Angelaina, "Towards an analysis framework for investigating students' engagement and learning in educational blogs", Journal of Computer Assisted Learning, 2012, vol. 28(3), pp. 222-234.

[27] A. Jimoyiannis, P. Tsiotakis, and D. Roussinos, "Social network analysis of students' participation and presence in a community of educational blogging", Interactive Technology and Smart Education, 2013, vol. 10(1), pp. 15-30.

 ISBN: 978-1-61208-361-2

Context-Aware Leisure Service:

A Case-Study based on a SOA 2.0 Infrastructure

Guadalupe Ortiz, Juan Boubeta-Puig
UCASE Software Engineering Group
Department of Computer Science and Engineering
University of Cádiz
Cádiz, Spain
{guadalupe.ortiz, juan.boubeta}@uca.es

Adrián Brenes Ureba
Higher School of Engineering
University of Cádiz
Cádiz, Spain
adrian.brenesureba@alum.uca.es

Abstract— **Service-Oriented Architectures (SOAs) have settled as an efficient solution for the implementation of systems in which modularity, loose-coupling and communication among third parties are key factors. However, although there are excellent tools and frameworks for service development, their adaptation to context has not been properly focused on to date. In this paper, we have made use of a SOA 2.0, where the core element is an enterprise service bus, in order to improve context-awareness for services. The proposal is illustrated trough a real case-study scenario implementation, where the results show the benefits of using such an architecture for web service context-awareness.**

Keywords- Web Service; Context-Awareness; Service-Oriented Architecture; Enterprise Service Bus.

I. INTRODUCTION

In recent years, Service-Oriented Architectures (SOAs) have settled as an efficient solution for the implementation of systems in which modularity, loose-coupling and communication among third parties are key factors. This fact has led to the increasing development of distributed applications composed of reusable and sharable components (services). These components have well-defined platform-independent interfaces, which allow SOA-based systems to quickly and easily adapt to changing business conditions.

However, although there are excellent tools and frameworks for service development, their adaptation to context has not been properly focused on to date. Even though this is a field in which many industry and scientific community are starting to provide their proposals [1]–[5], there are no clear solutions in the scope of web services. To illustrate the need for adaptation, let us provide an example: for instance, we may have services that would be suitable for their adaptation to the invoking client's specific context- such as his location or the weather conditions in his location. This would imply that service answers should be adapted depending on these contextual situations. In the past, we proposed a method for adapting services to the invoking device [6], as well as to adapt them to the client-specific context in general [7]. These approaches are good for the specific type of context dealt with – adapting to device and client-specific context – but are not prepared to deal with the external context.

In this regard, adapting services to context and current conditions might require the analysis of context information very often. Nevertheless, SOAs are not suitable for environments where it is necessary to continuously analyze the information flowing through the system, a key factor for an appropriate context-aware service implementation. This limitation may be solved by the joint use of Complex Event Processing (CEP) [8] together with SOA, the so-called event-driven service-oriented architecture or SOA 2.0 [9]: an extension of SOA to respond to events that occur as a result of business processes. However, most approaches implementing context-aware services do not take advantage of the use of CEP and SOA 2.0, therefore having to continuously access a context manager [1]–[3], [10]. See Section III analysis on related work for further details.

We already envisaged an architecture for this purpose in [11], in which the key element is an Enterprise Service Bus (ESB), which currently is the core of SOA 2.0. The main contribution of this paper is the definition of the exact architecture required and its implementation through a real case-study scenario. To this end, we have chosen a service which provides leisure activities. Particularly, the provided activities will be based on the location and weather conditions of the user; weather conditions will also be used to send special offers to subscribers.

The rest of this paper is organized as follows. Section II provides some background on the paper main areas of interest: context-awareness and event-driven SOAs. Afterwards, Section III describes and compares more relevant related work to the one presented in this paper. Then Section IV addresses the implemented architecture, first of all including the case-study description, secondly the architecture definition and finally the flows required in the ESB for materializing the good use of the SOA 2.0 architecture. Following, Section V provides the final application overview from the point of view of the different user roles, specially focusing on how context-awareness is dealt with. The article ends with Section VI, which discusses the proposal and conclusions.

II. BACKGROUND

In this section, we will introduce the main concepts of context-awareness and event-driven SOA.

 ISBN: 978-1-61208-361-2

A. Context-Awarenes

Dey et al.'s context definition in [12] is specially well-known: "Context is any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves".

Context-awareness supports the fact that the context information provided by the client, or taken from the environment, is properly used by the system so as to improve its quality; that is, using such contextual information to customise system outputs to improve final user satisfaction. Therefore, a system is context-aware if it uses the context to provide relevant information or services to the user, adapting the system behavior to the user particular needs.

A context-classification can be found in [13]; in this work, we will focus on dealing with the environmental context: sensors and/or specific services are currently used in order to provide such kind of information as location, temperature, precipitations, wind, etcetera. This type of context will imply filtering the information sent to the client; for instance, if location is taken into account when looking for leisure activities, the result would be restricted to those available within a limited distance; when taking into account weather conditions, the results might be narrowed and only activities suitable for these conditions might be provided.

B. Event-Driven SOA (SOA 2.0)

Event-driven architectures promote the detection of events and the subsequent intelligent reaction to them [14]. These architectures rely on complex event processing, a technology that provides a set of techniques to help discover complex events by analyzing and correlating other basic and complex events [8]. Therefore, CEP allows detecting complex and meaningful events in a particular context and inferring valuable knowledge for end user interests. Let us suppose again that we are looking for leisure activities for today; kite-surfing is fine when is windy; however it is not the same that it is windy when you are going to kite-surf, than that it is windy and raining more than 10 cm^3/h. This is an example on how complex events may help make decisions on the information to be provided to the user.

Currently, the integration of Event-Driven Architecture and SOA is known as Event-Driven SOA or SOA 2.0 [9]. SOA 2.0 will ensure that services do not only exchange messages between them, but also publish events and receive event notifications from others. For this purpose, an Enterprise Service Bus (ESB) will be vastly helpful to process, enrich and route messages between services of different applications. Further information on the integration of CEP with SOA in other scenarios can be found in [15].

III. RELATED WORK

In this section, we will focus on the main research for CEP and SOA integration and context-aware service implementations.

Several works on CEP and SOA integration in different domains can be found in the literature; for instance, Taher et al. [16] develop an architecture that integrates a CEP engine and input/output adapters for SOAP messages in order to adapt Web service messages between incompatible interfaces: input adapters receive messages sent by Web services, transform them into the appropriate representation to be manipulated by the CEP engine and send them to the latter. Accordingly, output adapters receive events from the engine, transform them into SOAP messages and send them to the corresponding to Web services.

There are some approaches which use CEP for monitoring such as the one from Xu et al. [17], where CEP is used to detect events in an Ambient Assisted Living (AAL) domain so that proper palliative actions can be taken in real time. The paper from Li et al. [18] is also worth a special mention. They provide an adaptive approach to context provisioning and automatic generation of actions. The latter definitely bears similarities with our proposal; however we focus on non-intrusive service result adaptation rather than action taking.

Most of the work found in the context adaptation area specially focuses only on websites [19] or in general on client side adaptation. We can mention, for instance, the paper from Laakko and Hiltunen [1] where content adaptation is done through a proxy; we can also mention the one by Mohomed et al. [2] where the system can learn about context through the interaction with the user. Both are interesting works, but they overhead the client computation; opposite to our proposal which deals with all the heavy tasks in the server side. Another example is the proposal from Keidl et al. [3], which consists of an approach for services to deal with client contextual information through a context framework. In their case, the context is always included in the client SOAP header, as well as in service messages. This implies that not only services, but also clients have to process the context included in the header; however they do not explore how the client can deal with the received context, and again they are overheading client communications.

Bucchiarone et al. [4] focus on the role of context in adaptation activities and describe a life-cycle for designing and developing adaptable service-based applications. They consider necessary to build contextual monitors and adaption mechanisms to detect context changes and trigger the subsequent actions. Furthermore, they propose rule engines as possible candidates for this purpose. However, implementations using rule engines are slower and less efficient in handling and receiving notifications, compared to those using CEP engines [20].

Sheng et al. [5] proposes *ContextUML*: a modeling language for context-aware model-driven web services. Several years later they improved their proposal supplying a platform for developing context-aware web services [21]. This platform, named *ContextServ*, is based on *ContextUML* and provides an integrated environment where developers can specify and deploy context-aware services, as well as generating Business Process Execution Language code. The main drawback of this proposal is the high learning curve required for their modelling methodology; in addition, it does not take any advantage of the use of the ESB and CEP,

 ISBN: 978-1-61208-361-2

which leverages the context-aware system scalability, usability and maintenance.

There are also several approaches which only consider location and personal preferences [22], [23], but no other environmental contexts; those are not relevant in this scope.

To summarize, our proposal mainly differs from others in benefiting from the advantages of the use of CEP and an ESB to adapt services to context information in a decoupled and scalable way, where the context can be automatically detected through real time events.

IV. CASE-STUDY DESCRIPTION

In this section, we are going to describe the case-study requirements and the proposed architecture to implement it.

A. *Description*

The goal is having a service-based application which offers leisure activities, as well as special offers depending on the weather conditions in a particular location.

The *application manager* (web master) will be the one in charge of defining a set of categories corresponding to weather conditions; for instance, wind speed above 40km/h means it is a windy day; otherwise it is not.

The *activity providers* will define on the one hand which activities are suitable under specific weather conditions (for instance a camera obscura is not a place to recommend when it is already dark whilst it is really nice in a sunny day). On the other hand, special offers which might be triggered under specific weather categories: the camera obscura provider might recommend it as long as there is day light, but since many people would not go there when suddenly raining or cloudy, he might be interested in sending a special price offer in such conditions (to attract some additional visitors).

The prospective *application users* are several:

First of all, a visitor might check what he can do now/today in the visited city. The result should take into account both weather-based recommended options and any available offer. This kind of user should not be required to get registered in the platform, since most of prospective users would not register for a short visit.

Secondly, a local user might be interested in receiving suggestions and offers in his city continuously; he should register for this purpose. He also has the chance to indicate under which weather conditions he is interested in doing leisure activities, to therefore receive customised alerts in his email account.

Finally, another kind of user would be future visitors who might check what is possible to visit in a location before they go there. For these visitors, the activities might be based on the weather forecast or on the weather historic data when no forecast information is available.

We would like to highlight that the system should be readily accessible both from a computer and a mobile device.

B. *Architecture*

If we start from the top-left of Figure 1, we can see that weather information is constantly arriving to the system (1); this information might be provided by sensors or any other event producer element; we have used web suppliers.

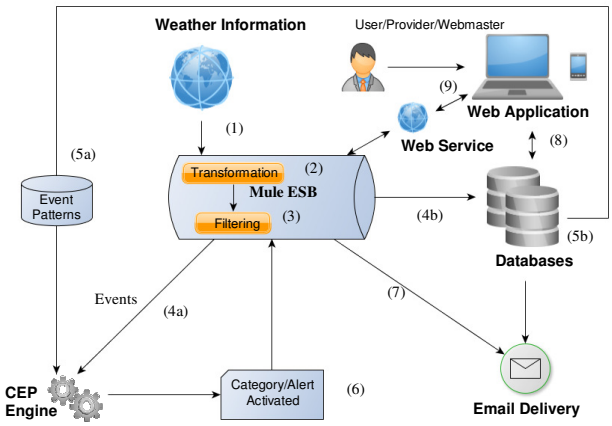


Figure 1. System Architecture

Therefore, in this system, the events reaching the system consist of the weather information. This information is transformed in the required objects of the system (2) and filtered to keep only the information/events of interest (3). Then these events are automatically redirected to the Esper CEP engine [24] (4a) in order to detect the predefined complex events pattern, as well as to a non-SQL database (4b) in order to keep a history table. Such patterns (5a) and their correspondence to the system weather categories have been designed by the system manager and stored permanently in the database (5b); besides they can be updated at any time.

If the patterns of interest are detected then two things happen: on the one hand categories are activated for their use in the web site (6); on the other the corresponding alert emails are sent to the subscribed users (7).

The provider has previously established the conditions for the alerts to be triggered or the activities to be offered at special prices (8). Activities and their associated conditions are stored in a SQL database.

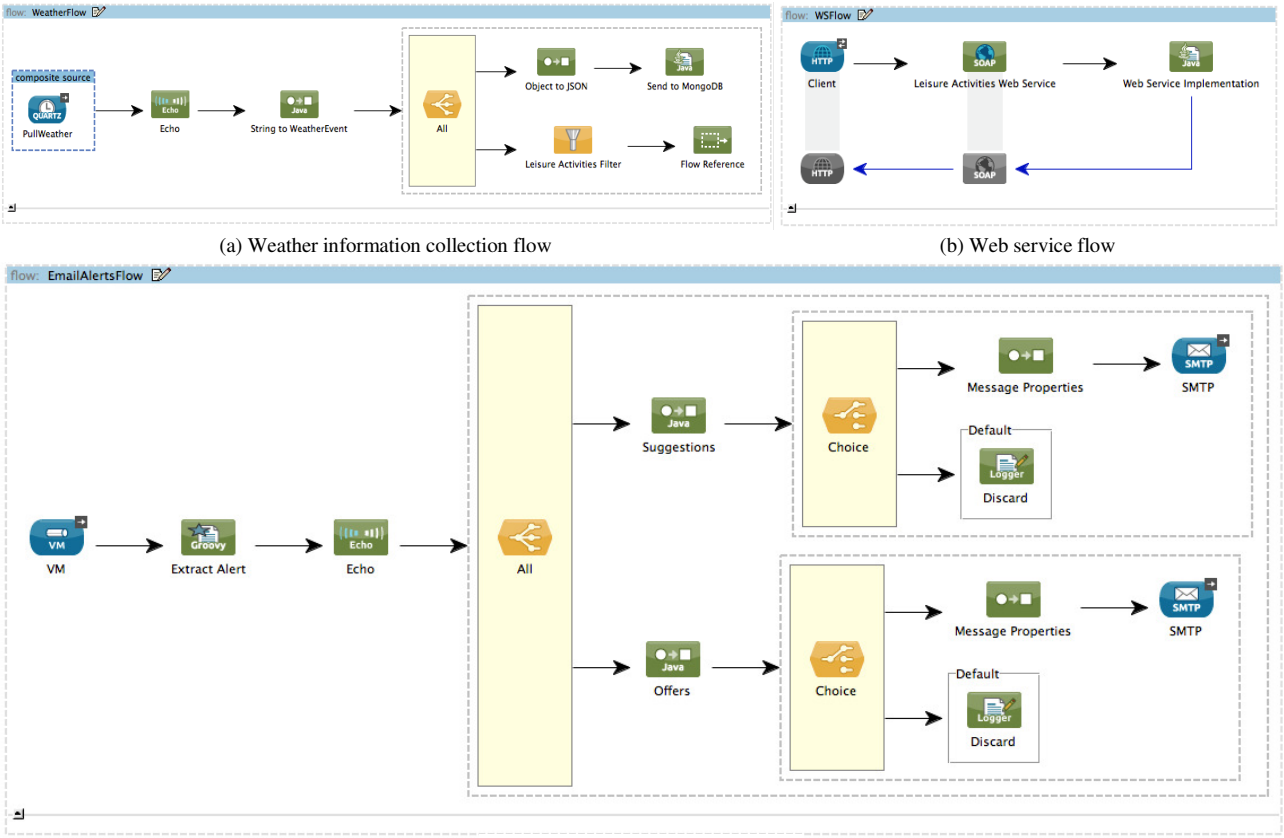
The web service is used as the intermediary between the user and the system. The user, when looking for leisure activities in the web site is transparently invoking the web service which provides this information (9), the latest is already adapted in real time – thanks to our architecture – to the current weather conditions.

C. *Flows in the Enterprise Service Bus*

Most of the business logic of the system has been implemented inside the enterprise service bus, specifically as Mule flows. This provides us with a twofold benefit: on the one hand, Mule facilitates the interoperation of multiple inputs/outputs formats; on the other, all the core functionality will be located together in the ESB. In Figure 2, we have included the most relevant flows in the bus.

Particularly, we have a specific flow for weather information collection/detection, which is shown in Figure 2(a). In this flow, information from several locations is obtained every minute; then it is immediately processed and transformed into the required format and immediately after sent to the non-SQL database, as well as to the complex

 ISBN: 978-1-61208-361-2



(a) Weather information collection flow

(b) Web service flow

(c) Notification flow

Figure 2. Flows in the enterprise service bus

event processing engine. We have filtered the information sent to the engine, so that only relevant information related to current leisure activities in the system reach the CEP engine. Please also bear in mind that additional sources could be added at any time should it be necessary.

Figure 2(b) shows the flow corresponding to the leisure web service. The client (in our case the web site) can invoke the service; the implementation of the service in Mule already takes into account the patterns detected in the engine, adapting the results to the weather conditions.

Finally, Figure 2(c) shows how suggestions and offers are sent based on the alerts triggered by the complex event processing patterns: first of all, we extract the alerts detected by the CEP engine according to the weather patterns defined in the system and to the weather events entering into the latest. Then, based on this information, and on the client interests stored in the system, the corresponding suggestions and offers are sent to their email accounts. The emails sent to the users are by default limited to one per day.

V. APPLICATION OVERVIEW

In this section, we describe the relevant functionality of the resulting application from the point of view of the system manager, the activity provider and the final user.

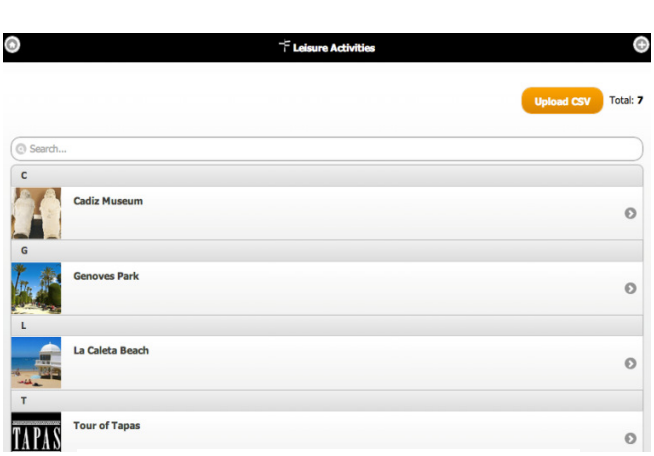
A. *System Manager Role*

The manager (web master) will be the one in charge of administrating categories. Even though the system already includes common categories related to weather situation; the manager will be the one in charge of including new categories – should it be necessary – and the patterns matching the named category. Those patterns have to be defined using EPL of the chosen CEP engine (Esper). The selection of this language was not only based on the efficiency of the CEP engine, but also on its close syntax to the well-known SQL, as well as its native support for multiple event format types. To give an example, if we want to include the category “windy”, an example of pattern for a windy day using Esper EPL would be the following:

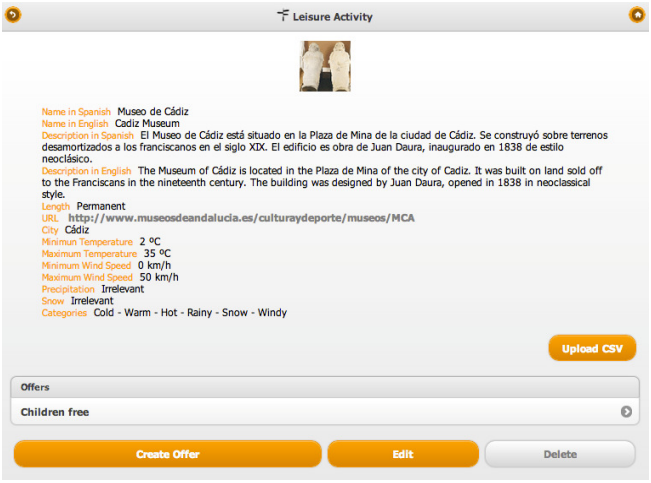
```
@Name("windy")
insert into windy
select ‘wind’ as alertName, a.windSpeed as
   windSpeed
from pattern [every a =
   WeatherEvent(windSpeed > 50)]
```

The remaining tasks of the system manager/web master would be usual web sites maintenance tasks.

 ISBN: 978-1-61208-361-2



(a) Activities uploaded by the provider



(b) Activity configuration by the provider

Figure 3. Application functionality from the point of view of the activity provider

B. *Provider Role*

The providers will include in the system the different activities. They can fill in all the information manually per each activity or can do it uploading a CSV file (see Figure 3 (a)). The relevant issue here is that they will indicate the weather conditions which will trigger the offers for each activity and the activities which will be provided under particular weather conditions (see Figure 3 (b)).

C. *User Role*

Non-registered user will enter the system and will be able to see the activities to be done now in his location, as well as those which have a special offer for today (see Figure 4).

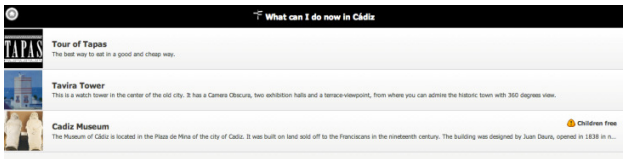


Figure 4. User activity search result

When the user registers he can predefine for which weather categories he wants to receive suggestions and offers (Figure 5).

VI. DISCUSSION

We have presented an application which satisfies both the provider and the consumer: the first one may trigger offers to profit from a larger number of clients when weather conditions are not suitable for his offered activities. The consumer not only receives information about the more suitable activities for current weather conditions but also might benefit from special offers.

It could be thought that limiting your plausible clients might not be good for your business. However, imagine you are happily visiting Granada in winter time and you are about to decide what to do today; the system offers you the option of visiting the Alhambra or going to ski. You go for the second, you rent all the equipment before you go up to Sierra Nevada and once there you discover that all the ski tracks are closed due to the blizzard. What would you do next time? Would it have not been better that the system had not suggested skiing for such a day?

Regarding the limitations of our proposal, we are aware of the difficulty for the system manager to create new event patterns in the system. This is why (1) we provided a large set of categories predefined in the system and (2) we pretend to integrate this architecture with other results of our research: the user-friendly editor for complex event pattern which will generate and deploy automatically the code in the CEP engine for the patterns to be detected [25].

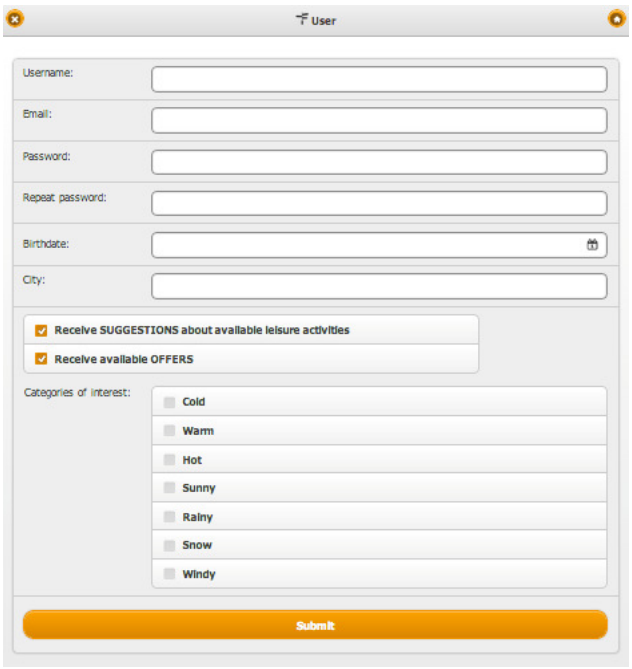


Figure 5. User preferences configuration

 ISBN: 978-1-61208-361-2

VII. CONCLUSION AND FUTURE WORK

In this paper, we have presented an event-driven service-oriented architecture and case-study implementation for providing a context-aware leisure service. Thanks to the use of an ESB to connect the different inputs and outputs of the system and to the use of a CEP engine, we can provide the activities adapted to the weather conditions in real time. Even more, activity providers benefit from a system which may send offers and suggestions to prospective clients based on weather condition real time changes, therefore improving their revenues, as well as the users satisfaction.

In our future work, we plan to extend the architecture with additional features which facilitate different contexts dealing and adaptation mechanism. As we mentioned before, we also pretend to integrate this architecture with a user-friendly editor for event pattern definition.

ACKNOWLEDGMENT

G. Ortiz and J. Boubeta-Puig acknowledge the support from the Spanish Ministry of Science and Innovation under the National Program for Research, Development and Innovation, project MoD-SOA (TIN2011-27242).

REFERENCES

[1] T. Laakko and T. Hiltunen, "Adapting Web Content to Mobile User Agents," *IEEE Internet Comput.*, vol. 9, no. 2, pp. 46–53, Mar. 2005.

[2] I. Mohomed, J. C. Cai, S. Chavoshi, and E. de Lara, "Context-aware interactive content adaptation," in *Proceedings of the 4th international conference on Mobile systems, applications and services - MobiSys 2006*, Uppsala, Sweden, 2006, p. 42.

[3] M. Keidl and A. Kemper, "Towards context-aware adaptable web services," presented at the 13th international World Wide Web conference on Alternate track papers & posters, New York, NY, USA, 2004, pp. 55–65.

[4] A. Bucchiarone, R. Kazhamiakin, C. Cappiello, E. di Nitto, and V. Mazza, "A context-driven adaptation process for service-based applications," presented at the 2nd International Workshop on Principles of Engineering Service-Oriented Systems, New York, NY, USA, 2010, pp. 50–56.

[5] Q. Z. Sheng and B. Benatallah, "ContextUML: a UML-based modeling language for model-driven development of context-aware web services," in *International Conference On Mobile Business*, 2005, pp. 206–212.

[6] G. Ortiz and A. García De Prado, "Improving device-aware Web services and their mobile clients through an aspect-oriented, model-driven approach," *Inf. Softw. Technol.*, vol. 52, no. 10, pp. 1080–1093, Oct. 2010.

[7] G. Ortiz and A. García de Prado, "Web Service Adaptation: A Unified Approach versus Multiple Methodologies for Different Scenarios," presented at the Fifth International Conference on Internet and Web Applications and Services (ICIW), 2010, pp. 569 –572.

[8] D. C. Luckham, *The power of events: an introduction to complex event processing in distributed enterprise systems*. Addison-Wesley, 2002.

[9] B. Sosinsky, *Cloud Computing Bible*. John Wiley & Sons, 2011.

[10] Q. Z. Sheng, S. Pohlenz, J. Yu, H. S. Wong, A. H. H. Ngu, and Z. Maamar, "ContextServ: A platform for rapid and flexible development of context-aware Web services," in *International Conference on Software Engineering*, 2009, pp. 619–622.

[11] G. Ortiz, J. Boubeta-Puig, A. García de Prado, and I. Medina-Bulo, "Towards Event-Driven Context-Aware Web Services," in *Adaptive Web Services for Modular and reusable Software Development: Tactics and Solutions*, IGI Global, 2012, pp. 148–159. DOI: 10.4018/978–1–4666–2089–6.ch005.

[12] G. D. Abowd, A. K. Dey, P. J. Brown, N. Davies, M. Smith, and P. Steggles, "Towards a Better Understanding of Context and Context-Awareness," London, UK, 1999, pp. 304–307.

[13] A. García de Prado and G. Ortiz, "Context-Aware Services: A Survey on Current Proposals," in *The Third International Conferences on Advanced Service Computing*, Rome, Italy, 2011, pp. 104–109.

[14] H. Taylor, Ed., *Event-driven architecture: how SOA enables the real-time enterprise*. Upper Saddle River, NJ: Addison-Wesley, 2009.

[15] J. Boubeta-Puig, G. Ortiz, and I. Medina-Bulo, "An Approach of Early Disease Detection using CEP and SOA," in *Service Computation 2011, The Third International Conferences on Advanced Service Computing*, 2011, pp. 143–148.

[16] Y. Taher, M.-C. Fauvet, M. Dumas, and D. Benslimane, "Using CEP technology to adapt messages exchanged by web services," New York, NY, USA, 2008, pp. 1231–1232.

[17] Y. Xu, P. Wolf, N. Stojanovic, and H.-J. Happel, "Semantic-based Complex Event Processing in the AAL Domain.," in *ISWC Posters&Demos*, 2010, vol. 658.

[18] F. Li, S. Sehic, and S. Dustdar, "COPAL: An adaptive approach to context provisioning," 2010, pp. 286 –293.

[19] D. Carlson and L. Ruge, "Towards Augmenting Legacy Websites with Context-awareness," presented at the 10th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services,, Tokyo, Japan, 2013.

[20] K. M. Chandy and W. R. Schulte, *Event Processing: Designing IT Systems for Agile Companies*. USA: McGraw-Hill, 2010.

[21] Q. Z. Sheng, J. Yu, A. Segev, and K. Liao, "Techniques on developing context-aware web services," *Int. J. Web Inf. Syst.*, vol. 6, no. 3, pp. 185–202, 2010.

[22] K. Cheverst, N. Davies, K. Mitchell, A. Friday, and C. Efstratiou, "Developing a context-aware electronic tourist guide: some issues and experiences," 2000, pp. 17–24.

[23] R. A. Abbaspour and F. Samadzadegam, "Building a context-aware mobile tourist guide system based on a service oriented architecture," *Int Arch. Photogramm. Remote Sens. Spat. Inf. Sci.*, vol. XXXVII, no. B4, pp. 871–874, 2008.

[24] E. Inc, *Esper - Reference Documentation*. 2014.

[25] J. Boubeta-Puig, G. Ortiz, and I. Medina-Bulo, "A model-driven approach for facilitating user-friendly design of complex event patterns," *Expert Syst. Appl.*, vol. 41, no. 2, pp. 445–456, Feb. 2014.

 ISBN: 978-1-61208-361-2

Asynchronous Learning Management System - The Case of Federal University of Technology (UTFPR)

Nádia P. Kozievitch, Eduardo Manika, Robinson Noronha, Leandro Almeida, Rosamelia Parizotto
Henrique da Silva, Laudelino Bastos, Thainã Monteiro
Federal University of Technology
Curitiba, PR, Brazil
Email: {nadiap,manika,vida,leandro,rosamelia,hosilva}@utfpr.edu.br,
bastos@dainf.ct.utfpr.edu.br, thaina128@gmail.com

***Abstract*—Mobile Learning has been emerging as a new research branch of learning, in which mobile devices are explored during the learning process. The development of these applications, however, is a non trivial task since it requires skills and knowledge of various domains (computational, pedagogical, etc.), along with technologies to support the process. This paper describes a case study at UTFPR, with the objective of integrate mobility, not only in terms of learning through mobile devices, but also in terms of mobile participants and mobile administration. The case study focuses on a generic and asynchronous computer architecture (PEDRO), based on a the Moodle-based adaptation. Our proposal allows the Learning Management System and mobile devices to work together in order to provide a complete set of instructional materials to students.**

Keywords–Moodle; Learning Management Systems; Mobile Devices.

I. INTRODUCTION

In spite of common beliefs, learning is not confined and the inclusion of remote experiments on mobile devices is a reality. In fact, students are always learning, no matter the place they are. However, there are learning environments and learning management systems [1] where this skill is maximized, being the most common: the schools. Traditional schools have some disadvantages [2], namely student's attention is not well focused and teachers and students must all meet in a common place, so that knowledge can be transmitted.

The popularization of Mobile Devices (MDs), such as "tablets" and "smartphones", for example, enabled an alternative learning method for students. MD has been used as tool access to information and execution of activities. Examples include teaching of science [3] and mathematics [4], among others.

Text editors, spreadsheets, production of videos and photos or playing electronic games can be accessed from any place and time, and can be easily integrated with a learning environment. Nevertheless, participants can also function as mobile learners in the sense that they may use the application any time and anywhere, in informal settings, in the course of their everyday activities.

The potential use of these devices has been promoting a new set of research possibilities and challenges for the community of Computing in Education [5], and their final use in a Virtual Learning Environment (VLE). The possibility of a student using a MD to access classes instigates alternatives of how to explore traditional environments. One of the great challenges for developing for MD applications [6] is their limited capacity when compared to personal computers. Another challenge appears when we consider that not necessarily all MDs have 24 hour access to Internet.

This paper addresses these issues in a case study at UTFPR, with the objective of integrate mobility. The case study focused on a generic and asynchronous computer architecture (PEDRO - portuguese acronym for "Programa de Ensino e Desenvolvimento Remoto Off-line" - Offline Learning and Remote Development Tool), based on a Moodle [7] adaptation. The novelty resides on a flexible architecture which explores (i) mobility not only in terms learning through MDs, but also in terms of mobile participants, and mobile administration; and (ii) the background infrastructure (software, hardware), considering network traffic and different locations.

A. A Motivating Example

For instance, consider the infrastructure for the Federal University of Technology - Paraná (Brazil), with 12 different campus (Utfpr Pato Branco 'A' , Utfpr Medianeira 'B' , Utfpr Campo Mourão 'C' , Utfpr Cornélio Procópio 'D' , Utfpr Curitiba 'E' , Utfpr Dois Vizinhos 'F' , Utfpr Francisco Beltrão 'G' , Utfpr Guarapuava 'H' , Utfpr Londrina 'I' , Utfpr Apucarana 'J' , Utfpr Ponta Grossa 'K' , Utfpr Toledo 'L'), as shown in Figure 1.

The university aims to integrate mobility, not only in terms learning through MDs, but also in terms of mobile participants and mobile administration. In this case, the participants are mobile learners in the sense the same course (in a VLE) can be available to several locations, and participants can remotely interact with MDs. In the same way, the system administration should be flexible in order to have structural and data updates from different locations, maintaining an architecture which can be easily adaptable in order to receive new requirements.

The background infrastructure (software, hardware) should also have an easy integration with the new proposed architecture. Since the institution has several locations (far apart from each other), with their respective users, the network traffic is another parameter which can result in a bottleneck (consider for example, that a specific course might be composed by several videos, which are transfered to one central campus to others).

 ISBN: 978-1-61208-361-2

Figure 1: The 12 UTFPR campus location provided by Google Search Engine.

Briefly, the interdisciplinary efforts may vary in several domains: (i) from the database perspective, the challenge is how to explore available learning systems within the institution with the minimal set of tables in order to have efficiency (and to enable any future integration with other systems); (ii) from the network perspective, the challenge is how to exchange a great amount of information within several campus and do not have a bottleneck; (iii) from the hardware perspective, the challenge is how to store a set of courses within the minimal amount of memory and space available within the MD; (iv) from the MD lifetime perspective, the challenge is how to decide which information should be deleted/compressed, due to limited space; (v) from the pedagogycal perspective, the challenge is how to explore all this structures and have a motivated student which is responsable for his main learning; (vi) from the design perspective, the challenge is how to adapt an interface which explores the MD interface; among others.

In this paper, we explore the term of mobility, along with the mentioned interdisciplinary efforts, through a case study PEDRO, a generic architecture proposed for mobile learning and administration. The novelty resides on the idea that not only students may interact with a remote and offline learning system, but also administrators and professors.

B. Organization

The remainder of this paper is organized as follows. Section 2 contains a description of related work. An overview of our solution is described in Section 3. Finally, we offer our conclusions and future work in Section 4.

II. Related Work

Mobile learning involves the use of mobile technology, either alone or in combination with other information and communication technology, to enable learning anytime and anywhere [8]. Learning can be explored in a variety of ways: people can use mobile devices to access educational resources, answer assignments, connect with others, or create content.

Multiple initiatives have explored ways to integrate or use MDs in learning situations [9]–[13]. If we consider room classes [14], for example, solutions with good results with MDs have already been presented. At this work, students are questioned by teacher during the class and answers should be provided by MDs. A central computer system summarizes these answers, processes them, and the teacher instantly presents a summary of choices of students, by means of graphs and tables.

Other initiatives focused on distance education and MDs (allowing students to interact in discussion forums [15]) or generic frameworks delimited by Pedagogical and Technological domain [5].

Some other research initiatives have been developed with the purpose of allowing customization of MD contents. For example, da Silva et al. [16] presents a recommendation system for learning objects based on Sharable Content Object Reference Model [17], using agents and ontologies.

In their system, agents located on the server monitor the activities of the student. The selection of Learning Objects (LO) is made by comparing the choices of students. The LOs which were more accepted by students of similar behavior were recommended to a new student added to an activity.

Another example of a system which allows the customization of learning objects sent to the MD is presented by Orlandi and Isotani [18]. In this system, a tool for initial diagnosis identifies which objects should be available to students. The architecture defined within this paper also provides lists of exercises with automatic correction. The main difference resides with the ”feedback” messages which can be sent to students.

All the systems presented above centralize a data server or applications. But, there are also other initiatives which do not require a centralized element (such in [19]). In this architecture, three heterogeneous agents (Social, Interface, and Collector) are scattered within different mobile devices to interact. Similar to [16], it explores the concept of Context Environment, storing geographic information to recommend suggestions for student interaction.

Among the more than 200 VLEs available on the market [20], Moodle [7], along with several extensions [21] has been highlighted by the number of users and servers installed around the world, mainly by its low cost, it is free software, and for different features depending on the configuration [22].

In the context of the research presented within this paper, the MD remains disconnected from the Internet most of the time. The internet connection is sporadic and should receive assignments and evaluations (accessed through an adapted version of Moodle), returning the tasks performed by the student. In this context, none of the investigated architectures could be easily adapted as a solution.

III. Overview of our Solution

In this section, we outline PEDRO architecture, along with the Moodle-based adaptation, and the respective usage.

A. PEDRO Architecture

From a theoretical perspective, PEDRO architecture is composed by a Embedded Learning Environment (ELE) and a VLE, as shown in Figure 2. Basically, the participant interacts

 ISBN: 978-1-61208-361-2

with the system through MB (the ELE at Figure 2). After the login, the interaction and embedded modules are activated. Here, the participant can explore the available courses and assignments, having all the activities logged. When the participant accesses the internet, the communication module gets the next available assignments and sends the assignments already answered, along with any additional information that should be exchanged within the main server.

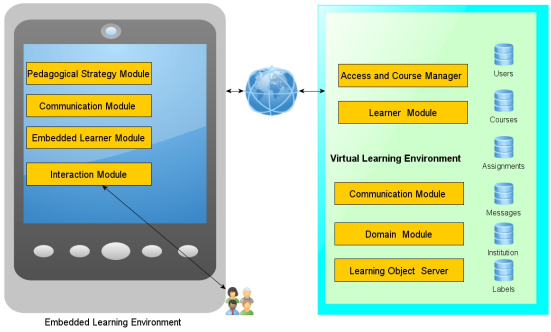


Figure 2: PEDRO architecture.

The ELE consists of four modules: the pedagogical strategy model, the communication module, the embedded learner module, and the interaction module.

The pedagogical strategy module is responsible for setting and controlling how instructional materials (videos, PDFs, assignments, etc.) will be accessed by the student. Three types of assignment strategies were adopted:

- Simple: assignments are accessed one by one, according to a specific order. This order was set by the teacher during the authoring process. Consider for example, that a specific course is divided into 12 topics, and all the available participants explore them, one by one, in an ordered sequence.
- Conditional: assignments are accessed only if aspecific condition happens. Consider for example, that a participant have to provice an answer to an assignment only if the average grade is less than a specific value.
- Random or free: the student has the freedom to access the assignment if he/she wants. It is not necessary to follow any order.

Note that the pedagogical strategy model is implemented through modules and sections of the available course.

The communication module is responsible for communication with the MD with the VLE. It mainly which information (such as assignment answers) is exchanged with the VLE. Note that the exchange information is stored within XML files.

The embedded learner module is a structure which stores information activities and instructional materials that still have to be accessed.

And finally, the interaction module is responsible for the student interaction with the MD, along with the log of this interaction.

The VLE provides five modules: the access and course manager module, the learner module, the communication module, the domain module, and the learning Objects server. The access and course manager module identifies which courses a student is enrolled (along with related information). This module comprises the information modified by the system administration.

The learner module comprises a structure which stores the tasks and instructional materials accessed by participants. The communication module is responsible for communicating with the MD. The domain module stores the course instructional materials. The learning objects server manages a set of learning objects, evaluation, and pedagogical strategy defined by the course author.

At the server side, the course and access manager receives the participant request access, and processes the following requests:

- Verify if the participant is registered within the system (through Moodle);
- If the participant is not registered within the system, the user will have access to the registration form available in the system;
- If the participant is registered within the system:
  1) Log the activities already performed by the participant;
  2) Check the course assignments for the respective participant;
  3) Compare the available activities with the performed activities by the participant;
  4) If the lists have different items, proceed with the next assignments that should be send to the participant;
  5) Check available space at MD, and compare with the required space for the next assignments that should be sent to the participant. If any issue arise, the manager should be informed;
  6) Pack all the required information and send to MD though the communication module; and
  7) Log all the required information.

The random access to the ELE with the VLE is the main requirement of the architecture presented within this paper. Since the user does not have full-time access to internet, there is an effort to identify which set of information should be presented to the student.

So, in this sporadic access, the environments should exchange the greater amount of relevant information to the educational process. This type of information exchanged between the mobile device and the VLE have already been called ”Context Sensitive Information” [23].

In closing this sporadic access, the ELE is able to partially track and monitor the student. This task consists of monitoring and recording the following information: i) answers to assignments, ii) tasks, iii) readings, iv) videos, v) applications used in MD, and vi) MD storage space.

Unlike some of the initiatives presented in Section 2, the selection of information to be sent to the MD does not consider

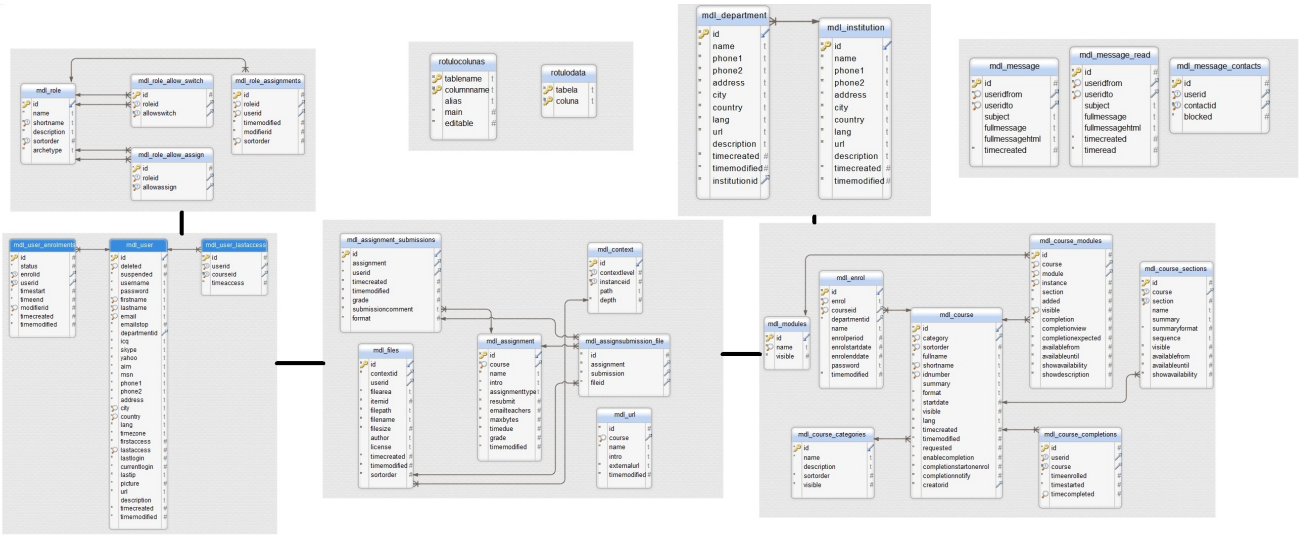


Figure 3: The complete set of database tables grouped by module.

the location of the student [16], [23], the effect of long-term reading texts on screen [24], among other criteria.

In summary, the architecture shares the following features between ELE and VLE: (i) log the student's interactions with the assignments; (ii) report student activities in MD; (iii) identify which instructional materials should be sent to the MD; (iv) allow the student to have access to free and guided instructional materials; (v) interpolate the various types of instructional materials; and (vi) explore mobility not only within an learning environment, but also through the management context.

B. Adapting Moodle

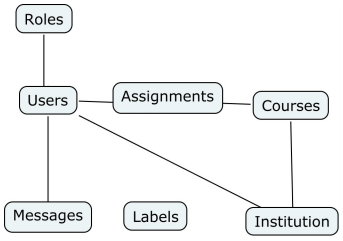


Figure 4: The main modules used within the project.

As a background to store all the information within the ELE and VLE environments, Moodle was adopted (the environment was already used as a basic software within the institution). The initial 2.4 version database (305 tables) was enhanced to a smaller set of tables (27 - Figure 3), resumed in 5 modules (Figure 4): roles, users, course, messages and assignments. The role module is responsible for defining roles, their names, permissions and how they are assigned to users. The users module is used to registering all the users within the system, along with their last access, and enrollments. The course module stores all the information within available courses. The message module is responsible for storing messages exchanged among users. Finally, the assignment module is used to assign an activity for a student enrolled within a course.

Besides the available modules, two new groups of information were created (as shown in Figure 4): institution and labels. The institution module is responsible for storing all available institutions within the project, with respective departments. Since Moodle is an english-based system, the label module has the objective of creating user-friendly descriptions for the available tables.

Note that *mdl_enrol* table, initially available at Moodle enrollments module, is represented here within the *course* module. The complete list of tables is listed within Figure 3, divided by module. Each table was also adapted to store only the relevant attributes for the project. In order to enhance optimization, foreign keys and tablespaces were added.

C. PEDRO Usage

PEDRO usage is explored under the VLE by a remote management web page, and under the ELE by the MD interface. The remote management web page has the objective of providing a flexible way of locally or remotely update the database structure, data, and labels to administrative users.

Within PEDRO architecture (Figure 2), the remote management web page explores the access and course manager module. The 27 Moodle-based tables (shown in Figure 4) can be explored through (i) their table structure (with respective columns, primary and foreign keys), (ii) data, and (iii) labels.

Figure 5 presents a printscreen from the database tables. If an administrator needs to update the data structure from table mdl_assignment for example, the administrator is redirected to another web page (Figure 6). Labels are used in order to present a more friendly portuguese description to users. So, despite using an english-based system, a manager could update a specific table (such as mdl_assignment shown in Figure 5) to have a friendly translation to participants.

 ISBN: 978-1-61208-361-2

Figure 5: The main development web page for to visualize and alter the structure, data and labels.

Figure 6: The main development web page for alter table mdl_assignment table.

Finally, Figure 7 presents the graphical interface available at the ELE. Note that at the left side three courses are presented (mathematics, portuguese, and history), along with the portuguese course structure at the left side of the image. The information exchange between VLE and ELE is provided by an XML file, as shown in Figure 8. Note that (i) the first block shows the studen information (such as name, number of courses, etc.), (ii) the second block presents the information for the first course (such as name, author, and amount of classes), (iii) the third block presents the information for the first class, followed by (iv) the list of the available activities comprising the first class.

As long as the participant finishes each of the available tasks, tag ”when” is completed with the respective date/time of the action (as shown in the last line from the fourth block from Figure 7). Note that the ELE interface (in java) is quite simple (in sense of design) in order to process the XML file along with the integrated files, under the available MD memory.

Our solution used Moodle version 2.14, Postgres 9.1.9, PHP, Java, and Android-Java. In particular, the same architecture could be easily translated to other languages, and explored under other domains, such as basic learning in industry (for home office employees). This strategy can provide a flexible way of foster knowledge dissemination, under a low cost, and reusing an available infrastructure.

The course “Portuguese Language: necessary revisions I”, is the pilot project, which will be realized in School Francisco Zardo (Curitiba, Paraná), with 18 students from 12th grade, using PDF and MP4 video files. The available Moodle server will be located at UTFPR Curitiba, which distances 11 KM from the school. Next experiments include all other campus locations provided by Figure 1.

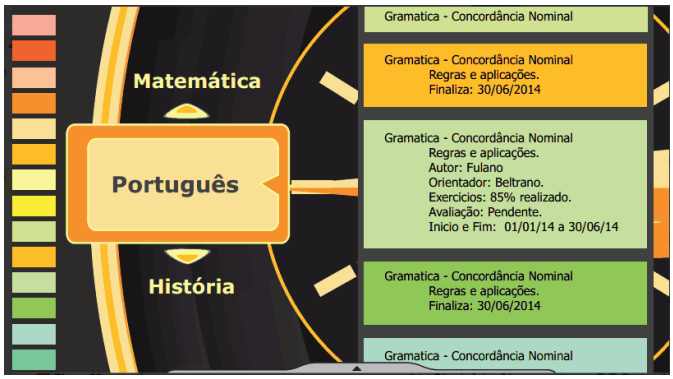


Figure 7: The graphical interface available at the ELE.

```
<student>
   <name>John</name>
   <password>1234</password>
   <numberOfCourse>3</numberOfCourse>
   <typeOfNavigation>1</typeOfNavigation>
   <course>
      <id> 1</id>
      <name>Gramática - Concordância Nominal</name>
      <author> Robinson Vida Noronha </author>
      <status> on  </status>
      <amountOfClass> 2 </amountOfClass>
      <class>
        <id>1</id>
        <title>Introdução as conceitos básicos da disciplina</title>
        <intro>Neste tópico, voce irá estudar alguns dos conceitos
que definiram essa disciplina</intro>
        <assessment>false </assessment>
        <amountOfTask>3</amountOfTask>
        <task>
          <id>9</id>
          <titleTask>Video to be watch</titleTask>
          <typeTask> VIDEO </typeTask>
          <file>File9.mp4</file>
          <completed>false</completed>
          <when>null</when>
        </task>
        <task>
          <id>1</id>
          <titleTask>File to be read</titleTask>
          <typeTask> PDF </typeTask>
          <file>File1.pdf</file>
          <completed>false</completed>
          <when>null</when>
        </task>
        <task>
          <id>2</id>
          <titleTask>File to be read</titleTask>
          <typeTask> PDF </typeTask>
          <file>File2.pdf</file>
          <completed>true</completed>
          <when>03/15/2014 ; 15:00:00</when>
        </task>
      <completed>33.0</completed>
       </class>
       ....
</course>
</student>
```

Figure 8: The XML file comprising the information exchanged between VLE and ELE.

 ISBN: 978-1-61208-361-2

IV. CONCLUSION

With the use of mobile technologies, each student will have a personal interaction in the sense that they may use the application any time and anywhere, in informal settings, in the course. This does not mean, however, that the use of mobile devices is a panacea [5].

In this paper, we addressed this issue with a case study (within Federal University of Technology) through the Moodle adaptation, a remote management web page, and PEDRO, a generic architecture concerning an Asynchronous Learning Management System and an Offline Development Tool. The proposed architecture describes the integration of i) the VLE and ii) the ELE for MD. The novelty resides on the idea that not only participants may interact with a remote and offline learning system, but also administrators and professors.

A straightforward future work consists in the inclusion of other Moodle modules (such as Scorm, Wiki, etc.), along with the architecture performance study, and a georeferenced location module.

ACKNOWLEDGMENT

We would like to thank CAPES, Fundação Araucária and CNPq.

REFERENCES

[1] P. McGee, C. Carmean, and A. Jafari, Course Management Systems for Learning: Beyond Accidental Pedagogy. Information Science Publishing, 2005.

[2] V. J. E. Miguel, S. M. M. Guerreiro, and R. P. C. do Nascimento, "Web tool to support online inquiries: adapting moodle to meet some of tutors and teachers needs," in Proceedings of the 2007 Euro American conference on Telematics and information systems, ser. EATIS '07. New York, NY, USA: ACM, 2007, pp. 53:1–53:4. [Online]. Available: http://doi.acm.org/10.1145/1352694.1352749

[3] C. Cortez, M. Nussbaum, R. Santelices, P. Rodríguez, G. Zurita, M. Correa, and R. Cautivo, "Teaching science with mobile computer supported collaborative learning (mcscl)," in Proceedings of the 2nd IEEE International Workshop on Wireless and Mobile Technologies in Education, ser. WMTE '04. Washington, DC, USA: IEEE Computer Society, 2004, pp. 67–74. [Online]. Available: http://dl.acm.org/citation.cfm?id=977408.978764

[4] L. de Lima, E. M. B. Filho, J. W. Ribeiro, R. M. de Castro Andrade, W. Viana, and A. J. M. Jnior, "Guidelines for the development and use of m-learning applications in mathematics," IEEE Multidisciplinary Engineering Education Magazine, vol. 6, no. 2, 2011, pp. 1–12. [Online]. Available: http://doi.acm.org/10.1145/1217299.1217304

[5] L. F. T. Meirelles and L. M. R. Tarouco, "Framework para aprendizagem com mobilidade," in Anais do XVI Simpósio Brasileiro de Informática na Educação, ser. SBIE '05. Sociedade Brasileira de Computação, 2005, pp. 623–633.

[6] B. Al-Hamadani and J. Lu, "An investigation in potential technology in compressing mobile learning xml documents," in Learning With Mobile Thecnologies, Handheld Devices, and Smart Phones: Innovative Methods. Edited by Zhongyu (Joan) Lu. IGI Global, 2012, pp. 37–55.

[7] Moodle, "Moodle, available at http://www.moodle.org, last accessed on May 30, 2014," 2009.

[8] UNESCO, Police Guidelines for Mobile Learning. United Nations Educational, Scientific and Cultural Organization, 2013.

[9] O. R. E. Pereira and J. J. P. C. Rodrigues, "Survey and analysis of current mobile learning applications and technologies," ACM Comput. Surv., vol. 46, no. 2, Dec. 2013, pp. 27:1–27:35. [Online]. Available: http://doi.acm.org/10.1145/2543581.2543594

[10] J. C. Sánchez Prieto, S. O. Migueláñez, and F. J. García-Peñalvo, "Mobile learning: Tendencies and lines of research," in Proceedings of the First International Conference on Technological Ecosystem for Enhancing Multiculturality, ser. TEEM '13. New York, NY, USA: ACM, 2013, pp. 473–480. [Online]. Available: http://doi.acm.org/10.1145/2536536.2536609

[11] M. H. Dabney, B. C. Dean, and T. Rogers, "No sensor left behind: Enriching computing education with mobile devices," in Proceeding of the 44th ACM Technical Symposium on Computer Science Education, ser. SIGCSE '13. New York, NY, USA: ACM, 2013, pp. 627–632. [Online]. Available: http://doi.acm.org/10.1145/2445196.2445378

[12] Y. Zhou, G. Percival, X. Wang, Y. Wang, and S. Zhao, "Mogclass: Evaluation of a collaborative system of mobile devices for classroom music education of young children," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ser. CHI '11. New York, NY, USA: ACM, 2011, pp. 523–532. [Online]. Available: http://doi.acm.org/10.1145/1978942.1979016

[13] T. Liu, J. Kiang, H. Wang, T. C. Wei, and LiHsing, "Embedding educlick in classroom to enhance interaction," in Proceedings of the International Conference on Computers in Education, ser. ICCE. IEEE Press, 2003, pp. 117–125.

[14] K. L. Nielsen, J. B. Stav, G. Hansen-Nygrd, and T. M. Thorseth, "Designing and developing a student response system for mobile internet devices," in Learning with Mobile Technologies, Handheld Devices, and Smart Phones: Innovative Methods, ed. Zhongyu (Joan) Lu. IGI Global, 2012, pp. 56–68.

[15] M. A. A. Sibaldo, E. Loureiro, I. I. Bittencourt, and E. de Barros Costa, "Infra-estrutura para acesso a comunidades virtuais na web através de dispositivos móveis," in Anais do XVII Simpósio Brasileiro de Informática na Educação, ser. SBIE '06. Sociedade Brasileira de Computação, 2006, pp. 58–60.

[16] L. C. N. da Silva, F. M. M. Neto, and L. J. Júnior, "Mobile: Um ambiente multiagente de aprendizagem móvel para apoiar a recomendação sensível ao contexto de objetos de aprendizagem," in Anais do XXII Simpósio Brasileiro de Informática na Educação, ser. SBIE '11. Sociedade Brasileira de Computação, 2011, pp. 254–263.

[17] SCORM, "Sharable Content Object Reference Model (SCORM), available at http://www.adlnet.gov/scorm/ , last accessed on May 30, 2014," 2000.

[18] B. H. Orlandi and S. Isotani, "Uma ferramenta para distribuição de conteúdo educacional interativo em dispositivos móveis," in Anais do XXIII Simpósio Brasileiro de Informática na Educação, ser. SBIE '12. Sociedade Brasileira de Computação, 2012.

[19] S. Rabello, J. L. V. Barbosa, J. Oliveira, A. Wagner, D. N. F. Barbosa, and P. B. S. Bassani, "Um modelo para colaboração em ambientes descentralizados de educação ubíqua," in Anais do XXIII Simpósio Brasileiro de Informática na Educação, ser. SBIE '12. Sociedade Brasileira de Computação, 2012.

[20] J. Itmazi, "Sistema flexible de gestión del elearning para soportar el aprendizaje en las universidades tradicionales y abiertas." Ph.D. dissertation, Universidad de Granada, December 2005.

[21] G. Röand A. Kothe, "Extending moodle to better support computing education," in Proceedings of the 14th annual ACM SIGCSE conference on Innovation and technology in computer science education, ser. ITiCSE '09. New York, NY, USA: ACM, 2009, pp. 146–150. [Online]. Available: http://doi.acm.org/10.1145/1562877.1562926

[22] E. Magalhães, V. Gomes, A. Rodrigues, L. Santos, and T. Conte, "Impacto da usabilidade na educação a distância: um estudo de caso no moodle ifam," in Proceedings of the IX Symposium on Human Factors in Computing Systems, ser. IHC '10. Porto Alegre, Brazil, Brazil: Brazilian Computer Society, 2010, pp. 231–236. [Online]. Available: http://dl.acm.org/citation.cfm?id=1999593.1999626

[23] P. Moore, B. Hu, M. Jackson, and J. Wan, "Intelligent context for personalised m-learning," in Complex, Intelligent and Software Intensive Systems, 2009. CISIS '09. International Conference on, 2009, pp. 247–254.

[24] L. L. Fabris and M. D. Finco, "Percepção de escolares no uso de laptops educacionais no contexto do projeto uca," in Anais do XXIII Simpósio Brasileiro de Informática na Educação, ser. SBIE '12. Sociedade Brasileira de Computação, 2012.

 ISBN: 978-1-61208-361-2

Rated Tags as a Service: A Cloud-based Social Commerce Service

Daniel Kailer

Munich University of Applied Sciences
Department of Computer Science and Mathematics
Munich, Germany
Email: dkailer@hm.edu

***Abstract*—The rise of social media has influenced many web applications, particularly in the area of e-commerce. Especially small and medium enterprises (SMEs) benefit from user-generated content, because these enterprises often do not have the dedicated resources to generate or categorize content. Moreover, SMEs often do not have the resources to create social media features by themselves. As a solution to these problems, a Cloud-based social commerce service named Rated Tags as a Service is presented in this paper. The intention of the Rated Tags system is to improve the decision making process of customers. This was already successfully evaluated in a user study. This paper discusses client- and server-side challenges for providing such a feature in a service-oriented way and proposes a corresponding Cloud-based architecture.**

Keywords–SaaS; E-Commerce; Architecture; SME.

I. INTRODUCTION

The rise of Web 2.0 and social media has influenced many web applications in recent years. Especially e-commerce companies make use of social media to transform their business into a more customer-centered environment [1]. The term social media stands for interactive, web-based applications that allow the creation and exchange of user-generated content [2]. The use of social media in e-commerce is often referred to as social commerce [3].

Social media, particularly user-generated content, is especially interesting for small and medium enterprises (SMEs), because it is an easy and cheap way to generate content and SMEs typically do not have dedicated resources to create or categorize content. SMEs usually also do not have the resources to implement and maintain their own infrastructure for these social interaction features. A possible solution to this is the service-oriented paradigm, i.e., the ability to use existing services instead of implementing and maintaining them by themselves.

This paper proposes a client- and server-side architecture for a novel social commerce service named Rated Tags as a Service. The Rated Tags system itself was designed to support the decision making of e-commerce customers through social tagging. The next step is the service-oriented provisioning of the system for online retailers, especially small and medium ones. This paper discusses some client- and server-side challenges and proposes an Cloud-based architecture for such a service.

The remainder of this paper is organized as follows. In Section II, the concepts behind the Rated Tags system and its related work will be briefly explained. After that, the proposed architecture and its implications are discussed in Section III. Finally, Section IV concludes this article and presents areas for future research.

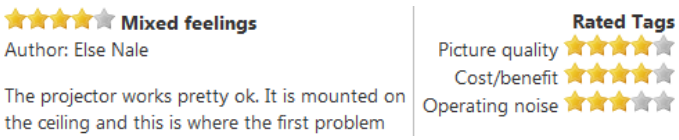


Figure 1: Exemplary excerpt of a customer review with attached Rated Tags

II. BACKGROUND AND RELATED WORK

Many online retailers make use of user-generated product ratings and reviews. Especially consumer reviews are a helpful resource for information seeking customers. However, it is a time-consuming task to analyze the vast amount of reviews, especially because of the unstructured nature of textual reviews. A social commerce feature named Rated Tags was developed by the author of this paper to reduce this decision making effort of customers.

A Rated Tag is the combination of a user-generated tag and a 5-star rating [4]. Similar to traditional social tagging, reviewers can create a Rated Tag and assign it to their review. For example, if a reviewer writes about the picture quality and operating noise of a projector, he or she can assign the Rated Tags “picture quality” and “operating noise” to their review and rate them on a 5-star scale. An exemplary excerpt of a customer review that contains Rated Tags is shown in Figure 1. Because of the user-generated nature of Rated Tags, other reviewers can reuse the created Rated Tags and assign them in their reviews. As a consequence, other information seeking customers are able to display only reviews that discuss a specific aspect of a product, for example the picture quality of a projector. Additionally, the ratings of the Rated Tags can be aggregated, which gives customers a good overview about the discussed aspects in the reviews, as shown in Figure 2. Thus, Rated Tags can be classified as an interactive decision aid (IDA).

To determine the helpfulness of Rated Tags as an IDA, the system was evaluated in a case-control study with 34 participants. The participants were provided with 5 products, whereas each product had 20 customer reviews. The participants of the

 ISBN: 978-1-61208-361-2

Figure 2: Exemplary summary of aggregated Rated Tags

Rated Tags group also had access to the corresponding Rated Tags (as shown in Figure 1). The instructions for the participants were to select the product with the lowest operating noise. The study results show, that Rated Tags users had a significant increase in decision quality. This was measured by comparing the selected product of the participants with the dominating product, i.e., the product with the lowest operating noise. Participants of the Rated Tags group significantly more often chose the dominating product than participants of the control group. The results also show a decrease in decision effort. This was measured by the time required to chose a product. The participants of the Rated Tags group were significantly faster in their decision making.

Similar approaches to Rated Tags were proposed by Vig et al. [5] and Lee et al. [6]. Their research also conducted the combination of user-generated tags and ratings. However, their research focused on different aspects, for example the user acceptance or user interface for the created tags. The research for Rated Tags primarily focuses on an improvement of the decision making process of e-commerce customers.

The combination of e-commerce and Cloud-based services is subject to several scientific articles [7][8] [9][10]. However, these works only discuss the integration of Cloud-based e-commerce services from a theoretical point of view. In contrast to that, this paper presents a conceptual model and architecture for a concrete e-commerce service. Therefore it is more likely to find solutions for real world problems, which can later be generalized to a common Cloud-based social commerce framework.

III. Rated Tags as a Service

This section constitutes the core of this article. It describes the proposed client- and server-side architecture to make the Rated Tags system available in a service-oriented way. As depicted in Figure 3, the three relevant actors for the architecture are service provider, service consumer and the e-commerce customer.

A. Service provider architecture

Because the proposed service needs to support a rising number of service consumers, an important aspect for the underlying system is scalability. Scalability means that the underlying system can cope with an increasing amount of load or traffic without modifying the system's architecture [11]. For this reason, a Cloud-based solution is proposed, because Cloud-resources allow a dynamic, on-demand scaling without the need to maintain an own data center [12].

The proposed Cloud-based architecture for the service provider is depicted in Figure 4. It is based on the service

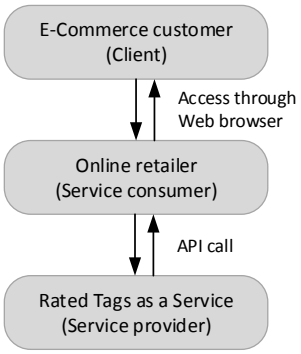


Figure 3: Conceptual model for Rated Tags as a Service

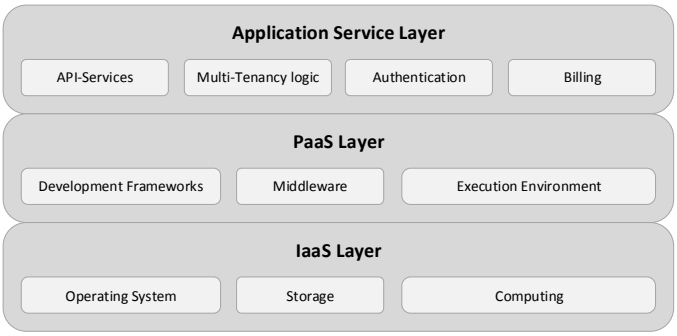


Figure 4: Proposed Cloud-based architecture for the service provider

orchestration architecture from the NIST cloud reference architecture [13, p. 3]. The topmost layer is the application service layer. This layer contains all the application specific logic for the service, while only a subset of all the components is displayed in Figure 4. The service layer makes use of the underlying PaaS and IaaS layers. The PaaS layer provides the execution environment for the service among other things. Finally, the bottom (IaaS) layer provides access to various resources, e.g., files or other storage structures.

An important aspect for the architecture is the underlying data model. To fully take advantage of the Software as a Service (SaaS) paradigm, a multi-tenant architecture is proposed. Multi-tenancy means, that a single software instance runs on a server, which serves multiple, independent service consumers (tenants) [14]. Therefore the data model needs to be designed for data partitioning, which is also a requirement to improve the scalability of Cloud-based applications [14].

The proposed partitioning scheme is horizontal data partitioning by tenants. This means that every tenant has its dedicated table. An example for this is shown in Table I. The name of the table is "RatedTag_ACME", whereas the name or id of the tenant (in this case "ACME") appears after the underscore. The shown columns do not contain any tenant-specific data. This is different from the shared table approach, where a table is shared among tenants and the identification is done via a column that identifies the tenant [15]. The advantage of the proposed dedicated table approach is that the data of tenants is physically separated from each other. It is also

TABLE I: An example of a tenant-specific table

RatedTag_ACME						
Id	ProductCategoryId	ProductId	ReviewId	UserId	Tag	Rating

assumed that this approach performs better, but this needs to be evaluated in a future simulation.

As shown in Table I, specific ids from the service consumer need to be stored to connect the reviews of the service consumers to the saved Rated Tags, for example the id of the review. Because the used format for these ids is tenant-specific and can be numeric or text-based, a common format is required. The proposed type for these columns is text, because the common types of ids (numeric ids and GUIDs) can be transformed into a textual format.

Another important aspect is the type of data storage. Traditionally, data for web applications is stored in relational databases like MySQL or Microsoft SQL Server, but with the rise of cloud computing and the increasing need for scalability the paradigm of NoSQL emerged. NoSQL databases do not require a fixed table scheme and mostly use an eventually consistent model in favor of availability and partition tolerance. Eventually consistent means that "the storage system guarantees that if no new updates are made to the object, eventually all accesses will return the last updated value" [16]. Because the Rated Tags system has no strict consistency requirements, an eventually consistent model is proposed.

B. Service consumer architecture

The main challenge for service consumers is the integration of the service in their own infrastructure. Three imaginable integration scenarios for service consumers are presented and discussed. These integration scenarios are depicted as sequence diagrams in Figures 5, 6 and 7.

One possible approach for service integration is shown in Figure 5. It shows a client (e-commerce customer), who requests a product page from the online store where Rated Tags are used. The online store (service consumer) then forwards an API call to the service provider to get all relevant data for the current context. After the completed API call, the retrieved data is included in the HTML and the rendered response is delivered back to the client. This can be considered as a synchronous approach, because the page rendering waits for the API call to complete.

Another possible approach is shown in Figure 6. This approach is similar to the previous one, but it applies an asynchronous communication. The requested page is rendered immediately and includes JavaScript code to reload additional content while loading the website. The JavaScript code calls the server of the service consumer, which forwards the API call to the service provider to fetch additional data and deliver the rendered results back to the client. This approach is more responsive than the previous one, because the Rated Tags-specific content is asynchronously loaded after the main web site is loaded.

Finally, a third approach is shown in Figure 7, where the client directly communicates with the service provider. This is

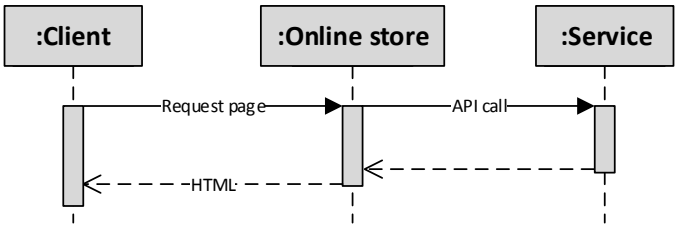


Figure 5: Integration scenario 1: Load content at once (no JavaScript required)

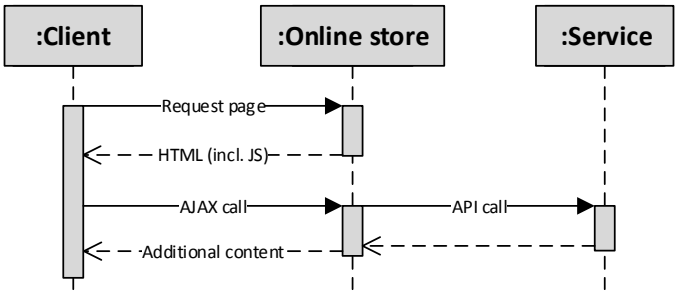


Figure 6: Integration scenario 2: Load additional content via AJAX

different to the two previous approaches, where the service consumer acted as a proxy, that forwarded an API call to the service provider. Such a direct communication approach is typically taken by mashups or widgets, that include mainly JavaScript and do not have any server side code. An exemplary service that uses this technique is Disqus [17], which is a discussion service, that can be integrated in any website to facilitate social interactions between users of this website. The approach has the advantage that it is easy to integrate, but it also has issues that needs to be considered. Some of these issues are discussed below.

A problem of the above approach is the required cross-domain request, because the JavaScript same-origin policy forbids browsers to send and receive content from a different domain. There are two solutions to get around this problem: JSON with padding (JSONP) and cross-origin resource sharing (CORS). JSONP is a technique to include and execute JavaScript from a different domain by specifying a callback function [18]. A restriction of JSONP is that it only supports HTTP-GET requests. CORS can be used to allow JavaScript requests from other domains. For this to work, the client (browser) and server must specify special CORS HTTP-headers [19]. Unfortunately, CORS is a newer mechanism, which means this technique is not supported by older browsers. The direct communication approach also lacks a server-side component, which is required to control the access to the service. Because of these issues, the integration scenario from Figure 7 is not suited for the architecture.

Eventually, the decision about the integration scenario depends upon the service consumer. If it is acceptable to use JavaScript, the approach from Figure 6 should be preferred, because it is more responsive. Alternatively, the approach from Figure 5 can be used when the online store of the

 ISBN: 978-1-61208-361-2

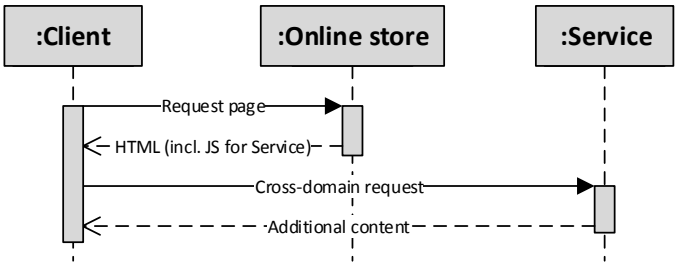


Figure 7: Integration scenario 3: Client-side call to the API

service consumer needs to be available for customers who have JavaScript disabled. The proposed architecture supports both approaches.

IV. Conclusion and Future Work

This paper presented the conceptual model and a Cloud-based architecture for a novel service called Rated Tags as a Service. The challenges for client- and server-side implementation were discussed. An important aspect for the client-side is the ease of integration, for which possible integration scenarios were presented and assessed. From the server-side perspective, the overall cloud-based architecture, a tentant-specific architecture and an eventually consistent model for data storage were proposed.

In a future work, a prototype of the proposed model will be implemented and evaluated. This includes not only the server-side components of the framework, but also the client-side. We will demonstrate by example, how third party online stores can integrate the social commerce service into their system. The evaluation of the prototype will concentrate on the non-functional requirements of the service, e.g., the scalability or the ease of integration.

References

[1] Z. Huang and M. Benyoucef, “From e-commerce to social commerce: A close look at design features,” Electronic Commerce Research and Applications, vol. 12, no. 4, 2013, pp. 246–259.

[2] J. H. Kietzmann, K. Hermkens, I. P. McCarthy, and B. S. Silvestre, “Social media? get serious! understanding the functional building blocks of social media,” Business Horizons, vol. 54, no. 3, 2011, pp. 241–251.

[3] C. Wang and P. Zhang, “The evolution of social commerce: An examination from the people, business, technology, and information perspective,” Communications of the AIS (CAIS), vol. 31, no. 1, 2012, pp. 105–127.

[4] D. Kailer, P. Mandl, and A. Schill, “Rated tags: Adding rating capability to collaborative tagging,” in 2013 IEEE Third International Conference on Social Computing and Its Applications 2013 (SCA 2013), Karlsruhe, Germany, 2013, pp. 249–255.

[5] J. Vig, M. Soukup, S. Sen, and J. Riedl, “Tag expression: Tagging with feeling,” in ACM Symposium on User Interface and Technology, 2010, pp. 323–332.

[6] S. E. Lee, D. K. Son, and S. Han, “Qtag: Tagging as a means of rating, opinion-expressing, sharing and visualizing,” in 25th Annual International Conference on Design of Communication, 2007, pp. 189–195.

[7] J. Yu and J. Ni, “Development strategies for sme e-commerce based on cloud computing,” in 2013 Seventh International Conference on Internet Computing for Engineering and Science (ICICSE), Sept 2013, pp. 1–8.

[8] X. Wang, “Research on e-commerce development model in the cloud computing environment,” in 2012 International Conference on System Science and Engineering (ICSSE), June 2012, pp. 539–542.

[9] H. L. Qing, “Research on the model and application of e-commerce based on cloud computing,” in 2012 International Conference on Computer Science Service System (CSSS), Aug 2012, pp. 142–146.

[10] G. Lackermair, “Hybrid cloud architectures for the online commerce,” Procedia Computer Science, vol. 3, no. 0, 2011, pp. 550–555, world Conference on Information Technology.

[11] A. B. Bondi, “Characteristics of scalability and their impact on performance,” in Proceedings of the 2Nd International Workshop on Software and Performance, ser. WOSP ’00. New York, NY, USA: ACM, 2000, pp. 195–203.

[12] M. Armbrust et al., “A view of cloud computing,” Communications of the ACM, vol. 53, no. 4, Apr. 2010, pp. 50–58.

[13] F. Liu et al., “NIST Cloud Computing Reference Architecture,” http://www.nist.gov/customcf/get_pdf.cfm?pub_id=909505 (retrieved: March, 2014), 2011.

[14] W.-T. Tsai, Q. Shao, Y. Huang, and X. Bai, “Towards a scalable and robust multi-tenancy saas,” in Proceedings of the Second Asia-Pacific Symposium on Internetware, ser. Internetware ’10. New York, NY, USA: ACM, 2010, pp. 8:1–8:15.

[15] Z. H. Wang et al., “A study and performance evaluation of the multi-tenant data tier design patterns for service oriented computing,” in IEEE International Conference on e-Business Engineering, 2008. ICEBE ’08, Oct 2008, pp. 94–101.

[16] W. Vogels, “Eventually consistent,” Communications of the ACM, vol. 52, no. 1, Jan. 2009, pp. 40–44.

[17] http://www.disqus.com (retrieved: March, 2014).

[18] B. Ippolito, “Remote json – jsonp,” http://bob.ippoli.to/archives/2005/12/05/remote-json-jsonp (retrieved: March, 2014), 2005.

[19] A. van Kesteren, “Cross-origin resource sharing: W3C Recommendation 16 January 2014,” http://www.w3.org/TR/2014/REC-cors-20140116 (retrieved: March, 2014), 2014.

 ISBN: 978-1-61208-361-2

Services to Support Use and Development of Speech Input for Multilingual Multimodal Applications for Mobile Scenarios

António Teixeira, Pedro Francisco, Nuno Almeida, Carlos Pereira, Samuel Silva
Department of Electronics, Telecommunications & Informatics / IEETA
University of Aveiro
Aveiro, Portugal
{ajst, goucha, nunoalmeida, cepereira, sss}@ua.pt

***Abstract*—Speech is our most natural form of interaction. Developing speech input modalities for several languages, combining speech recognition and understanding, presents various difficulties. While automatic translators ease the translation of normal text, the adaptation of grammars for several languages is currently performed based on an *ad hoc* approach. In this paper, we present a novel service that enables a multilingual speech input modality and helps developers in the creation of the grammars for different languages. The service itself uses two additional services for parsing and translation. The use of the service is exemplified in the context of AAL PaeLife project multilingual personal assistant.**

Keywords - service; speech; grammars; multilingual; multimodal interaction; translation.

I. INTRODUCTION

Advances in technology have brought mobile devices to our everyday life. With the growing number of features provided by devices such as smartphones or tablets, it is of paramount importance to devise natural ways of interacting with them that help to deal with their increasing complexity. Natural interaction is, therefore, an important goal, striving to integrate devices with our daily life by using gestures, context awareness or speech.

The importance of natural interaction is also boosted by the needs of various user groups, such as the elderly, that might present some kind of limitation at physical (e.g., limited dexterity) or cognitive (e.g., memory) level and lack the technological skills to deal with devices that can play an important role in improving their daily life [1].

The increased mobility and the multitude of devices that can be used impose important challenges to interaction design. Nevertheless, the "always connected" nature of most of these devices, in a multitude of environments (e.g., home, work and street), offers the possibility of using resources located remotely, including computational power, storage or on-the-fly updates to currently running applications to serve a new context.

Speech and natural language remain our most natural form of interaction [2][3] and a number of recent applications use speech as part of a multimodal system [4] in combination with other modalities. Nevertheless, despite its potential, the inclusion of input and output modalities based on speech poses problems at different levels. On a higher level, speech modalities involve several complex modules that need to work together and ensure speech recognition and speech synthesis. Tailoring these modules to different applications is a tiresome task and we have recently proposed a generic, service-based, modality component [5] that can work decoupled from the application, thus providing easier deployment of speech modalities. Another important issue concerning speech is its inclusion in applications targeting multiple languages. Therefore, our generic modality component also aims at being able to internally handle several languages.

Several well-known applications use speech. A representative example is mTalk [6] multimodal browser developed by AT&T, a tool to support the development of multimodal interfaces for mobile applications. The mTalk uses cloud-based services to process most of the multimodal data. Siri [7] and Google Voice Search [8] are other examples of speech enabled applications, that use cloud based services to process multimodal data.

Automatic Speech Recognition (ASR) takes as input the speech signal and produces a sequence of words. Speech recognition engines are typically based in Hidden Markov Models [9], which provide a statistical model to represent the acoustic model for the utterances. In addition to the acoustic model, a language model or a grammar is also needed to define the language. Language models, such as the ones defined by the ARPA format, are statistical n-gram [10] models that describe the probability of word appearance based on its history. Grammars can be defined as a set of rules and word patterns which provide the speech recognition engine with the sentences that are expected. The Java Speech Grammar Format (JSGF) [11] and GRXML [12] are examples of grammar formats.

Although grammars are more limited in the amount of sentences that will be recognized, they are capable of being more specific to each particular context of use, which often translates to a more accurate recognition.

These models and grammar are language dependent and, therefore, require language specific training. Usually, acoustic models and language models are trained generically to support a broad part of the language. They only need to be trained once for each language. Since grammars are created based on the context of one application, it is necessary to translate the grammars of each application to each language that the application aims to support.

The PaeLife project [13] is aimed at keeping the European elderly active and socially integrated. The project is developing AALFred, a multimodal personal life assistant (PLA), offering the elderly a wide set of services from

 ISBN: 978-1-61208-361-2

unified messaging (e.g., email and twitter) to relevant feeds (e.g., the latest news and weather information). The platform of the PLA comprises a personal computer connected to a TV-like big screen and a portable device, a tablet. One of the key modalities of the PLA is speech; speech input and output will be available in four European languages: French, Hungarian, Polish, and Portuguese. One of the demanding tasks on using the speech modality, due to the several languages involved, is to help developers and user interaction designers in the derivation of the grammars for each language.

Therefore, in this context of multi-language support, our main goals for the generic speech modality include:

- Streamlining of internationalization support;
- Reduce variance among grammars contributing for easier update and maintenance;
- Customization of any of the different grammars, if needed;
- Additionally to manual editing, allow automatic expansion of the recognized sentences and word corpora using existing services.

To approach these goals and in the context of a multimodal personal assistant, AALFred [14], part of the aforementioned project PaeLife, we present a first instantiation of a service which explores automatic translation to provide initial versions for the grammars in the different languages based on the definition of the semantic grammar (in English). The service receives a grammar, translates it and supports the needs of the speech modality.

The multimodal architecture integrating the multilingual support for speech input is directly related to the recent work of the W3C on a distributed architecture for multimodal interaction [15]. In fact, as described in [4][16] we have been working on the application of such architecture to mobile and AAL applications.

The use of services to support the functionalities in speech input has been adopted in several mobile architectures, such as the mentioned mTalk [6] and SIRI [7], but none, to the best of our knowledge, explored the use of automatic translation of grammars to support multilingual speech input.

The remainder of this document is organized as follows: Section II describes the main aspects of the proposed service regarding its architecture and main components; Section III discusses prototype implementation; Section IV provides some application examples; finally, Section V presents some conclusions and ideas for future work.

II. SYSTEM OVERVIEW

The system's main objective is to be able to automatically generate a derived grammar in other target languages. That is achieved by preserving as much of the main grammar structure as possible, generating coherent phrases in the target language and having in consideration the process of word reordering.

The system is dual in functionality. It supports both development and use in real interaction contexts.

In the development stage, developers use the system to make semantic grammars available, to produce the translated versions of such grammars. At this stage the service can also be used remotely to check and make corrections to the grammars. This can be done by native speakers or, if available, language specialists.

In interaction contexts, the system is in charge of the natural language understanding, making use of the grammars sent to the service at development stage. It receives the output of speech recognition and returns the semantic information extracted. The service also returns, on request, to the speech modality, the necessary information on words and sentences needed to configure the speech recognition engine.

A. Architectural Definitions

The architecture, in Fig. 1, is composed of four main components: the speech modality, the core service, the access APIs and the external resources (both parser and translator services). Further details about each component are provided in what follows.

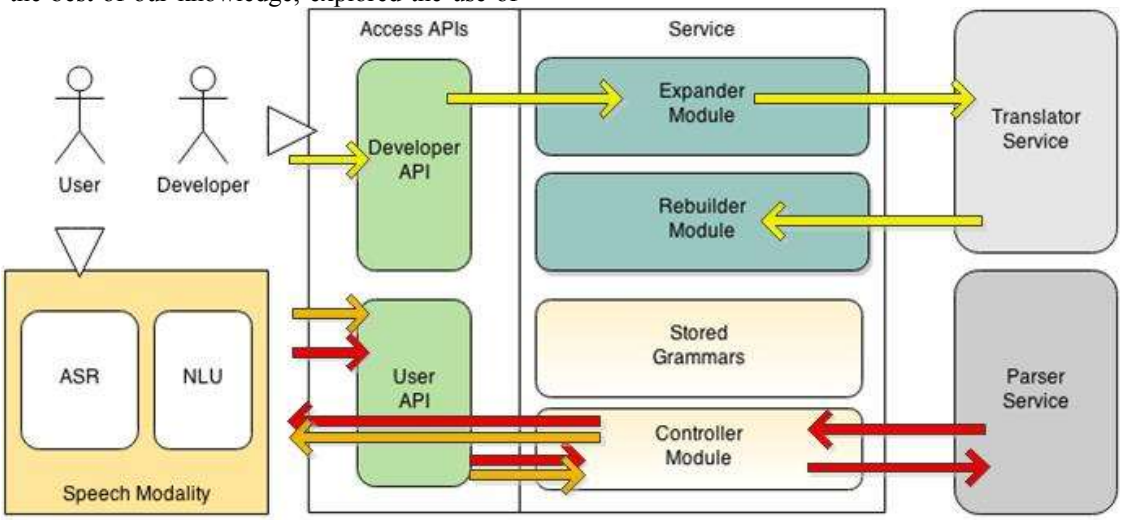


Figure 1 - Conceptual Architecture.

1) Speech Modality

The speech modality allows the recognition of speech in a specified language, previously selected. In practice, speech is processed by the Speech Recognizer (ASR) to produce a list of words that are sent to the Natural Language Understanding (NLU) interpreter to process. The NLU goal is to extract semantic information from the sequence of recognized words. The implementation is aligned with the modalities in a multimodal architecture, integrating in general a recognizer and an interpreter.

2) Main Service

The main service is responsible for the manipulation of the grammar. It allows to: a) upload files and input to be analyzed, and retrieval of the parsing result; b) get all statements generated by the specified grammars and on-demand translation of grammars; c) submit corrections to derived grammars and get a listing of all available grammars.

The service also requires the definition of a format for representation of the input grammar. From this representation, using the Expander Module we are able to generate all possible statements recognizable by the grammar, which are the statements submitted for translation, using the Translator Module.

The service has several ways of being used. The simplest, illustrated in Fig. 2, consists in the submission of a grammar and the selection of an intended language which results in the subsequent generation of valid phrases, to ease the configuration of an ASR by a third party.

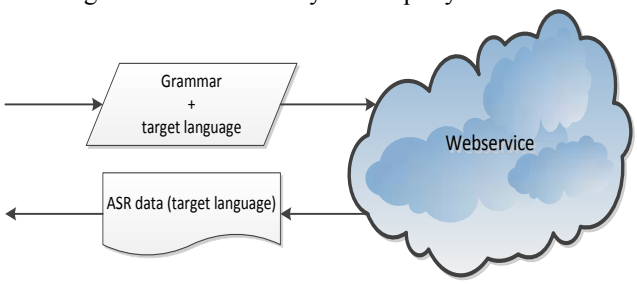


Figure 2 - Simple use of the service to get list of sentences for ASR in a target language.

Figure 3 shows a case where, assuming previous configurations and a working ASR, the service is used to extract semantic tags of a given text, and return them to the caller. This way of using the service implements the multilingual NLU processing.

Given the limitations of automatic translations, the service also supports manual revision and subsequent update of grammars (Fig. 4). This use is particularly suited when developing an application – such as AALFred – allowing the creation of an initial semantic grammar in English and using the service to provide translated grammars in other languages, enabling each involved partner in the project to revise and correct the automatically generated grammars. Each revised version becomes part of the service, after upload, and is used as described in the previous use cases.

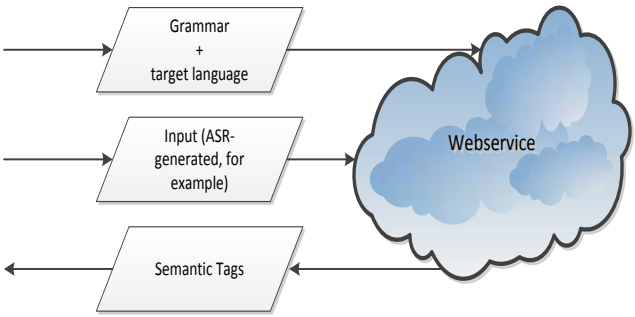


Figure 3 - Service used as multilingual NLU.

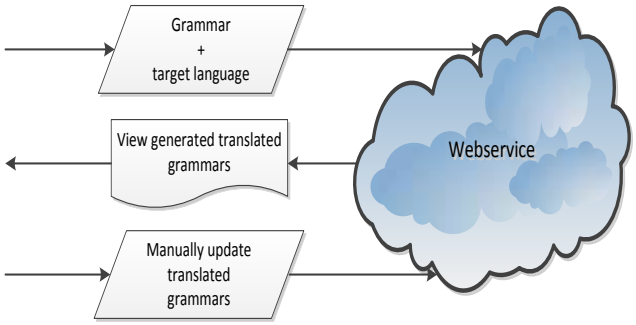


Figure 4 - Service used to manual revision and update of grammars.

After the translation is accomplished, a Rebuilder Module recreates new grammars according to the translated languages. Afterwards, these new grammars are stored within the Stored Grammars module for further usage.

3) Access APIs

All operations are made through the access APIs, ensuring a consistent and complete operation control.

To enable the introduction of new grammars, a specific interface is required for the developer. This interface can be seen as a frontend which allows the developer to submit a grammar and check the results of grammar translation, both in terms of generated grammar and of generated sentences. In our current implementation, it supports editing a grammar and its resubmission. This method enables faster feedback cycles of grammar enhancement.

For the speech modality, a user API is provided, allowing sequences of words from the speech recognizer to be processed in order to obtain semantic tags (i.e., to perform NLU in speech recognizer output).

4) Parser and translator services

The service is connected to two external services. The first one provides parsing; the second provides translation.

 ISBN: 978-1-61208-361-2

III. PROTOTYPE IMPLEMENTATION

To test our architecture and associated ideas, a prototype service has been created and used. Phoenix [17] was chosen as both the parser and grammar specification format. The advantages of this choice are explained by Phoenix's robustness to errors in recognition and parsing abilities. For translation, the choice fell on Bing due to its ability of providing reordering information. Later on, a more detailed explanation will be given on this.

The following sections provide information on the implementation and features of key components within the prototype.

A. Parser service

The objective of the parser is to extract the semantic tags, as defined in the semantic grammar, from the list of words received from the ASR, and return the text plus the semantic tags to be processed by the Interaction Manager and ultimately used by the application.

Internally, the analysis is done by Phoenix. Phoenix uses an automatic translated semantic grammar that allows tags existing on the original grammar to be preserved on the target language grammar.

In order to have an integrated support for the multiple languages of the project – or even other languages – the NLU parser is coupled with the management and process of automatic derivation of grammars by automatic translation.

B. Translation of Semantic Grammars

The goal is to translate to a target language all the terminal words while preserving the semantic tags. Translation must also produce a complete list of sentences defined by the grammar.

The process adopted and implemented is composed of three steps: 1) full expansion of the grammar; 2) translation; and 3) grammar rebuild.

1) Grammar Expansion

In order to be able to manipulate the Phoenix Grammar, one of two approaches had to be followed: either change the Phoenix Parser or have a separate parser to parse the Phoenix Grammar structures onto a separate data structure, on which we would then apply our modifications. We decided to implement a separate parser so as not to change the Phoenix code, allowing us to use C# for our work and rely on the Phoenix Parser only for its already defined and well-tested function: parsing the input text based on a defined Grammar.

In order to properly translate the grammar to take in consideration word reordering, we need to submit the full sentence for the translator to properly evaluate which translation to provide. While a word-by-word translation would yield a non-natural result, submitting the whole sentence allows us to retrieve a translated sentence that sounds natural and takes in consideration language specific connectors and variances which may not exist on the original language.

The algorithm developed makes use of two data structures: an "in progress" stack and a "done so far" queue. On the first, the algorithm stores the current rule while on the second it stores the translated words. Expanding all the rules is done by keeping the history of the rules visited along the expansion.

2) Translation

The translation process consists in submitting the result of the expansion (words plus their history/grammars rules) and receiving the resulting translated sentences (pairing of words in the translation with the correspondent words in the source).

In our prototype, we selected Bing Translator as the translator service. The usage of the Bing Translator is an advantage to us since it provides the realignment info [18] necessary to get word reordering support during the grammar rebuild process. That realignment info both eases the matching of translation with source words and is what allows us to support word reordering when reconstructing grammar rules. In addition, Bing Translator also allows us to obtain multiple translations per request, which enables the expansion of an existing grammar to support several similar sentences, with no need of additional input by the developer. We can thus increase the coverage of our grammar in an automatic and effortless way.

TABLE 1 – EXAMPLE OF BING TRANSLATION REORDERING INFO.

Source text: The answer lies in machine translation. Translated Text: La réponse se trouve dans la traduction automatique. Alignment info: 0:2-0:1 4:9-3:9 11:14-11:19 16:17-21:24 19:25-40:50 27:37-29:38 38:38-51:51 The -> La answer -> réponse lies -> se trouve in -> dans machine -> automatique translation -> traduction . -> .

3) Grammar Rebuild

When the grammar is parsed (in order to expand it afterwards), a different object is created for each instance of any rule. As such, for each Terminal Word present in the statement resulting from the expansion of the grammar, we can determine exactly which rule gave origin to the path that lead to it after the sentence is submitted for translation. Since we have reordering info available, we know which rules generated the text resulting from the translator.

The developed algorithm uses the saved Grammar Expansion history and the translated sentences of the Translation Process. It consists of analyzing the ancestors' historic information to remake the grammar. This is done by merging Non-Terminals of the same level throughout the

 ISBN: 978-1-61208-361-2

grammar in a top-bottom approach. Fig. 5 and 6 show an example.

[Main]	[Main]	[Main]	[Main]
[OPEN_NEWS]	[OPEN_NEWS]	[OPEN_NEWS]	[OPEN_NEWS]
A	[ITEM_NUM]	Elem	Megnyitása
	Második		

Figure 5 - Initial representation of the grammar.

[Main]			
[OPEN_NEWS]			
A	[ITEM_NUM]	Elem	Megnyitása
	Második		

Figure 6 - Resulting data after application of rebuild algorithm.

Duplicates are eliminated automatically, thus obtaining the grammar according to the translation given.

IV. First Results

Currently, the developed service module supports the translation of text in English to French, Hungarian, Polish and Portuguese. Furthermore, it supports translations from French, Hungarian, Polish and Portuguese to English. Two examples of service usage are presented in this section.

A. Example of grammar translation

After the submission of a new grammar, either via a direct API or via a website (in development), the submitted grammar will be parsed and stored in memory after which all phrases will be generated. As an example, the grammar in Fig. 7 will be converted to the Hungarian translation presented in Fig. 10.

```
[Main]
      ([VIEW_DAYS])
      ([OPEN_NEWS])
;
[DAY]
      (yesterday)
      (today)
;
[ITEM_NUM]
      (first)
      (second)
      (third)
;
[VIEW_DAYS]
      (news from [DAY])
      (open [DAY] news)
;
[OPEN_NEWS]
      (open the [ITEM_NUM] item)
;
```

Figure 7 – Example of original grammar (in English) sent to the service by an application developer.

```
news from yesterday
news from today
open yesterday news
open today news
open the first item
open the second item
open the third item
```

Figure 8 - Result from the expansion of the original grammar (in English).

```
a tegnapi hírek
a mai hírek
nyissa meg tegnap hírek
nyissa meg mai hírek
nyissa meg az elso elemet
a második elem megnyitása
a harmadik elem megnyitása
```

Figure 9 - Results from translation of the sentences in Fig.8 to Hungarian.

```
[Main]
      ([VIEW_DAYS])
      ([OPEN_NEWS])
;
[DAY]
      (tegnapi)
      (mai)
      (tegnap)
;
[ITEM_NUM]
      (elso)
      (második)
      (harmadik)
;
[VIEW_DAYS]
      (a [DAY] hírek)
      (nyissa meg [DAY] hírek)
;
[OPEN_NEWS]
      (nyissa meg az [ITEM_NUM] elemet)
      (a [ITEM_NUM] elem megnyitása)
;
```

Figure 10 - The resulting Hungarian grammar.

As can be seen following the steps, the grammar in English is used to generate all sentences (Fig. 8), which are then translated. The translation (Fig. 9) is then used, in conjunction with word generation history, to rebuild the grammar in Hungarian, with flexibility to deal with word reordering (in bold) and to synonyms/alternatives (underlined).

B. An example of grammar manual fine tuning

The system autonomously generates a grammar ready to be used on any language. However, it is possible to fine-tune the grammar to achieve a higher degree of correctness. This can be done by the developer or by a third party. The web based grammar editor allows previewing the sentences that the edited grammar describes and resubmission of the grammar (Fig. 11). To complement the first example in Hungarian, this example is in French.

 ISBN: 978-1-61208-361-2

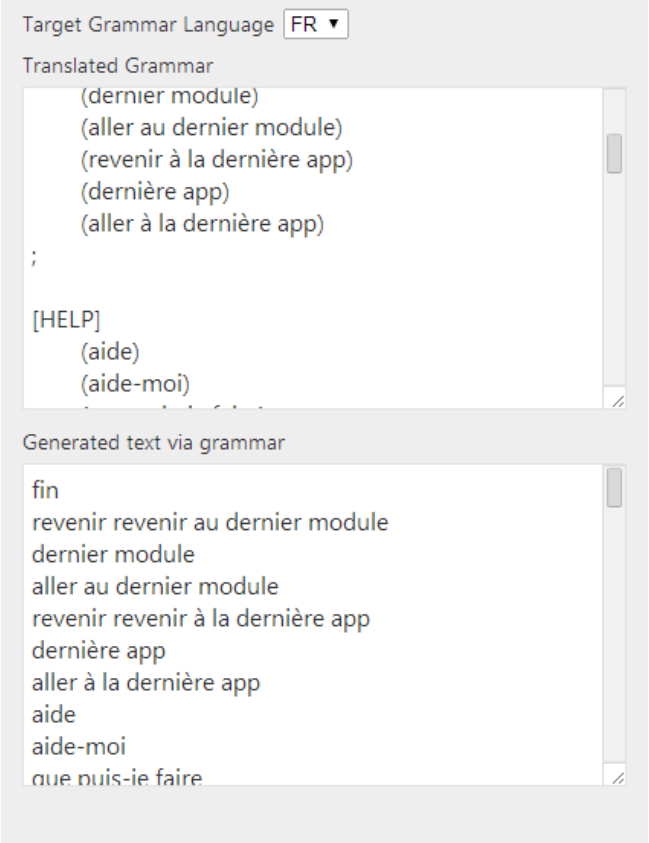


Figure 11 - Web base editor showing the translation to French in manual edition and the corresponding list of sentences defined by the edited grammar. Some generation problems are noticeable, such as repetition of word "revenir".

V. CONCLUSIONS AND FUTURE WORK

Multilanguage support in speech modalities is a complex task. In the context of a generic service-based speech modality, proposed by the authors, a service is presented which aims to provide support for easy deployment of applications supporting several languages. The main highlight of the proposed service is the possibility to generate grammars for different languages by automatic translation of an existing grammar (in English). A first prototype has been implemented and tested and several application examples are provided.

Future developments should explore the use of multiple translation services, increasing the probability of having, in the set of translated sentences, the correct ones. The evaluation in real use, both by users of the personal assistant integrating the speech modality and developers, will be performed in the next months, as part the development process in project PaeLife.

ACKNOWLEDGMENT

Authors acknowledge the funding from AAL JP, FEDER, COMPETE and FCT, in the context project of AAL/0015/2009, project AAL4ALL (www.aal4all.org), IEETA Research Unit funding FCOMP-01-0124-FEDER-022682 (FCT-PEstC/EEI/UI0127/2011) and project Cloud Thinking (funded by the QREN Mais Centro program, ref. CENTRO-07-ST24-FEDER-002031).

REFERENCES

[1] T. H. Bui, "Multimodal Dialogue Management - State of the art" , Technical Report no. TR-CTIT-06–01, 2006.

[2] N. O. Bernsen, "Towards a tool for predicting speech functionality," Speech Communication, vol. 23, no. 3, pp. 181–210, 1997.

[3] C. Munteanu et al., "We need to talk: HCI and the delicate topic of spoken language interaction," in CHI'13 Extended Abstracts on Human Factors in Computing Systems, April 2013, pp. 2459–2464.

[4] A. Teixeira et al., "Multimodality and Adaptation for an Enhanced Mobile Medication Assistant for the Elderly," in Proc. Third Mobile Accessibility Workshop (MOBACC), CHI 2013, April 2013.

[5] N. Almeida, S. Silva and A. Teixeira, "Design and Development of Speech Interaction: A Methodology", in Proc. HCI International, 2014, in press.

[6] M. Johnston, G. Di Fabbrizio and S. Urbanek, "mTalk - A Multimodal Browser for Mobile Services.," in Interspeech, 2011, pp. 3261–3264.

[7] "iOS - Siri.", http://www.apple.com/ios/siri/ [Accessed: 21-Mar-2014].

[8] "Voice Search", http://www.google.com/insidesearch/features/voicesearch/index-chrome.html [Accessed: 21-Mar-2014].

[9] B. Singh, N. Kapur and P. Kaur, "Speech Recognition with Hidden Markov Model: A Review," Int. J. Adv. Res. Comput. Sci. Softw. Eng., vol. 2, no. 3, pp. 400–403, 2012.

[10] P. Clarkson and R. Rosenfeld, "Statistical language modeling using the CMU-cambridge toolkit", Eurospeech, 1997, vol. 97, pp. 2707–2710.

[11] T. Brøndsted, "Evaluation of recent speech grammar standardization efforts.," in Interspeech, 2001, pp. 1089–1092.

[12] A. Hunt and S. McGlashan, "Speech Recognition Grammar Specification Version 1.0", http://www.w3.org/TR/speech-grammar/ [Accessed: 21-Mar-2014].

[13] "PaeLife: Personal Assistant to Enhance the Social Life of Seniors.", http://www.microsoft.com/portugal/mldc/paelife/ [Accessed: 21-Mar-2014].

[14] A. Teixeira et al., "Speech-Centric Multimodal Interaction for Easy-To-Access Online Services : A Personal Life Assistant for the Elderly," Procedia Computer Science, pp. 389–397, 2013.

[15] D. A. Dahl, "The W3C multimodal architecture and interfaces standard," J. Multimodal User Interfaces, vol. 7, no. 3, pp. 171–182, Apr. 2013.

[16] A. Teixeira, N. Almeida, C. Pereira and M. Oliveira e Silva, "W3C MMI Architecture as a Basis for Enhanced Interaction for Ambient Assisted Living," in Get Smart: Smart Homes, Cars, Devices and the Web, W3C Workshop on Rich Multimodal Application Development [online], July 2013.

[17] W. Ward, "Understanding spontaneous speech: the Phoenix system," in International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 1991, pp. 365–367.

[18] "Word Alignment Information from the API", http://msdn.microsoft.com/en-us/library/dn198370.aspx [Accessed: 21-Mar-2014].

 ISBN: 978-1-61208-361-2

Investigating Aspects of Visual Clustering in the Organization of Personal Document Collections

Hoda Badesh
Faculty of Information Technology
Misurata University
Misurata, Libya
hoda.badesh@gmail.com

Anwar Alhenshiri
Faculty of Information Technology
Misurata University
Misurata, Libya
alhenshiri@gmail.com

Jamie Blustein
Faculty of Computer Science
Dalhousie University
Halifax, NS, Canada
jamietab@cs.dal.ca

Evangelos Milios
Faculty of Computer Science
Dalhousie University
Halifax, NS, Canada
eem@cs.dal.ca

***Abstract*—Organizing personal collections of digital documents can be frustrating for two main reasons. First, the effort required to work with the folder system on personal computers and the possible misplacement and loss of documents. Second, the lack of effective organization and management tools for personal collections of digital documents. The research in this paper investigated specific visualization and clustering features intended for organizing collections of documents and built in a prototype interface that was compared to a baseline interface from previous research. The results showed that those features helped users with: 1) the initial classification of documents into clusters during the supervised stage; 2) the modification of clusters; 3) the cluster labelling process; 4) the presentation of the final set of organized documents; 5) the efficiency of the organization process, and 6) achieving better accuracy in the clusters created for organizing the documents.**

Keywords- information organization, management, retrieval, clustering, visualization, human factors.

I. INTRODUCTION

Personal documents grow in size and number rapidly. In the current state, desktop documents can be organized either manually in folder hierarchies or using special software such as: OpenText[15], IBM's Document Manager, and Google Desktop[16], which has been discontinued). Manual organization can be very demanding since desktop computers may involve large collections of documents. Every type of software has its advantages and disadvantages. For instance, Google desktop presented its search results in a list provided from searching the index of keywords Google built from the user documents. This type of presentation may require the user to go through large lists of result hits, formulate several queries, and eventually may or may not find the intended document.

When the pile of documents on a user's desktop grows extensively, organizing those documents into folders may become very time consuming. The use of software that presents lists of results may also be very ineffective. The use of clustering for organizing user desktop documents has had little consideration. User interfaces for assisting users with organizing their documents using aspects of clustering and visualization have not been thoroughly investigated.

Clustering is grouping together documents of the same type, genre, topic, etc. A categorization scheme has to be defined prior to applying clustering. Topical clustering and genre clustering have been investigated [5][15]. The use of clustering in document presentation has been investigated for desktop retrieval as well as web retrieval [1][14]. Clustering makes use of overviews of documents for conveying the different topics or genres covered in the document collection.

Visualization can help the presentation of multiple features of search results [1][2]. Document features such as its size, last update, and type can be visualized. Features of the collection as a whole can also be visualized by showing documents of the same type connected or by showing documents with similar content under one category. Such visual clustering combines the benefits of visualization and clustering. Adding clustering and visualization to the presentation of search results can help users organize large collections of documents and find results more effectively and efficiently.

There are several problems associated with managing and organizing personal documents on desktop computers. The following summarizes those problems:

1. The size of the collection of documents on computers of personal nature grows very rapidly as users keep using their machines.
2. Manual organization of documents on desktop computers necessitates the use of folder structure which may result in:
3. Excessive time consumption in the case of large collections.
4. Losing documents due to the complex structures and the difficulties associated with manually searching those structures.
5. Organization tools may drive the user away for one or more of five reasons: visibility, integration, co-adoption, scalability, and return to investment [13].
6. Search using desktop tools has problems associated with the presentation of the search results and the interaction with the user.

The research discussed in this paper attempts to answer the following questions:

1- What is the effectiveness of using three options of document views (abstract, text cloud, full content) on how users classify their documents for organization?
2- What is the effectiveness of presenting the initial clusters during the classification process as bubbles containing glyphs of documents inside each

 ISBN: 978-1-61208-361-2

corresponding cluster with different modification capabilities?

3- What is the effectiveness of having different views of clusters, as a list of cluster labels and as labelled bubbles?
4- What is the effectiveness of presenting the final set of documents clustered and organized in bubbles representing topics with their documents represented as glyphs?

The features indicated in the questions above were investigated in a prototype interface called the Bubbles Interface. The prototype was essentially intended to investigate these particular visualization features (also shown in Table 1) in improving the organization of collections of personal documents using clustering. To investigate the usefulness of those features, the interface was compared to the Pie Interface, another project developed in [9] for the same purpose. The results of the research showed that the new interface had a better layout and assisted its users with the initial classification of documents. Modifying and labelling clusters was also enhanced using the new interface. The interface improved the final organization of the documents by improving the accuracy of the clusters created.

The remainder of this paper is organized as follows. Section 2 discusses related work. Section 3 illustrates the study conducted in this research. Section 4 provides a detailed discussion of the results. The paper is concluded in Section 5.

II. RELATED WORK

Managing and organizing information has been explored in different directions. Knoll et al. [14] investigated how users view and manage desktop information in general. Jones et al. [13] investigated important reasons behind giving up on certain personal information management tools. The strategies users follow to manage web information in order to be able to relocate and reuse information previously found are discussed in [12]. Their work showed that users follow different keeping strategies to re-find and compare information later. The variety of managing and organizing strategies for personal information can be attributed to the fact that current tools lack important reminding, integration, and organization schemes [7].

Jones et al. [14] found that users abandon the use of an information management tool for one or more of five closely related reasons: visibility, integration, co-adoption, scalability, and return on investment. Jones [11] reviewed research in support of a more general preference for *way finding* methods that depend on a sense of digital location vs. direct search as the primary means for access to personal information [6]. Bergman and Nachmias [4] indicated that direct search becomes the user choice for retrieving personal information after attempts for search by navigation fail.

Jaballah [10] designed a desktop personal library manager to overcome the problems associated with the use of folder-based organization schemes. Users could browse and search their personal collections of documents by the document type, title, filename, date of modification, and so on. The interface was evaluated using a pilot study (two experts) followed by a learnability study and final diary study (6 participants). The results showed that even with the prototype's ability to harvest metadata about the files in the collection, the users preferred the standard folder system. They reported that some actions on the prototype were difficult and that users spent most of the time trying to familiarize themselves with the interface.

To further emphasize the value of visual access to information for managing and organizing personal collections, Bauer [3] built an interface intended to arrange piles of images or PDF documents in portraits. Each PDF file in the portrait is shown as one page containing images and parts of the text in the documents. Images are shown in their own piles. The closer the image to the user, the larger the size of the document is. The prototype allowed interactions with collections of documents to be logged over long periods. The prototype was not evaluated and it was expected to improve the user's experience with managing piles of personal documents and images.

Civan et al. [6] compared the user behavior for organizing information using folders and using labels (tags). For the purpose of the comparison, Gmail, which is Google's email service, and Hotmail, which is Microsoft's email service, were selected. Users organized their e-mail messages using different methods in the two systems. Gmail's users labeled or tagged their messages; Hotmail users put messages into folders. The two approaches were compared with respect to: "retrieval performance, evolution in mappings between articles and folders/labels over time, and limitations to fully express one's internal conceptualization" [6]. No clear winner was identified between tagging and placing. The study concluded that "better support for information organization may need to go well beyond folders and tags or their artful combination" [6].

Managing information is concerned with how people store, organize, and re-find information [8]. Information management systems are methods by which users find, categorize, and re-find information on daily basis. Research has considered personal information management. However, there is further need for investigating organizing and finding information in cases where the personal collection of documents grows extensively and when standard folder-based organization becomes overwhelmingly demanding.

III. RESEARCH STUDY

The study discussed in this paper compared two interfaces, the Pie Interface from the work in [9] and the Bubbles Interface designed for the purpose of this study. The Pie Interface was selected based on the results of a previous study that showed some drawbacks in the prototype during the evaluation. The Bubble Interface was designed and compared to the Pie Interface for evaluating the features embedded in the Bubble Interface to overcome difficulties users encountered with the Pie Interface in the previous study as discussed in [9]. The interfaces are briefly described as follows:

 ISBN: 978-1-61208-361-2

Figure 1. The Pie Interface.

A. *The Pie Interface*

The Pie Interface is divided into four sections: 1) the supervision panel, 2) the un-assigned document view, 3) the cluster view, and 4) the labeled document view. They are shown in Figure 1.

B. *The Bubble Interface*

This interface was designed to allow users to organize their personal collections of documents based on clustering and using aspects of visualization in both the classification stage, which is the supervised portion of the process, and the final presentation stage of the organization process. The Bubbles Interface (shown in Figure 2) was designed to overcome several disadvantages in the Pie Interface.

C. *Study Design and Population*

The study design was complete factorial and counterbalanced. It accounted for the possible effects of *order* using two conditions in a *within-subject* design. The possible main effect of the independent variable (the interfaces) was controlled by randomly selecting with which interface the participant started.

 ISBN: 978-1-61208-361-2

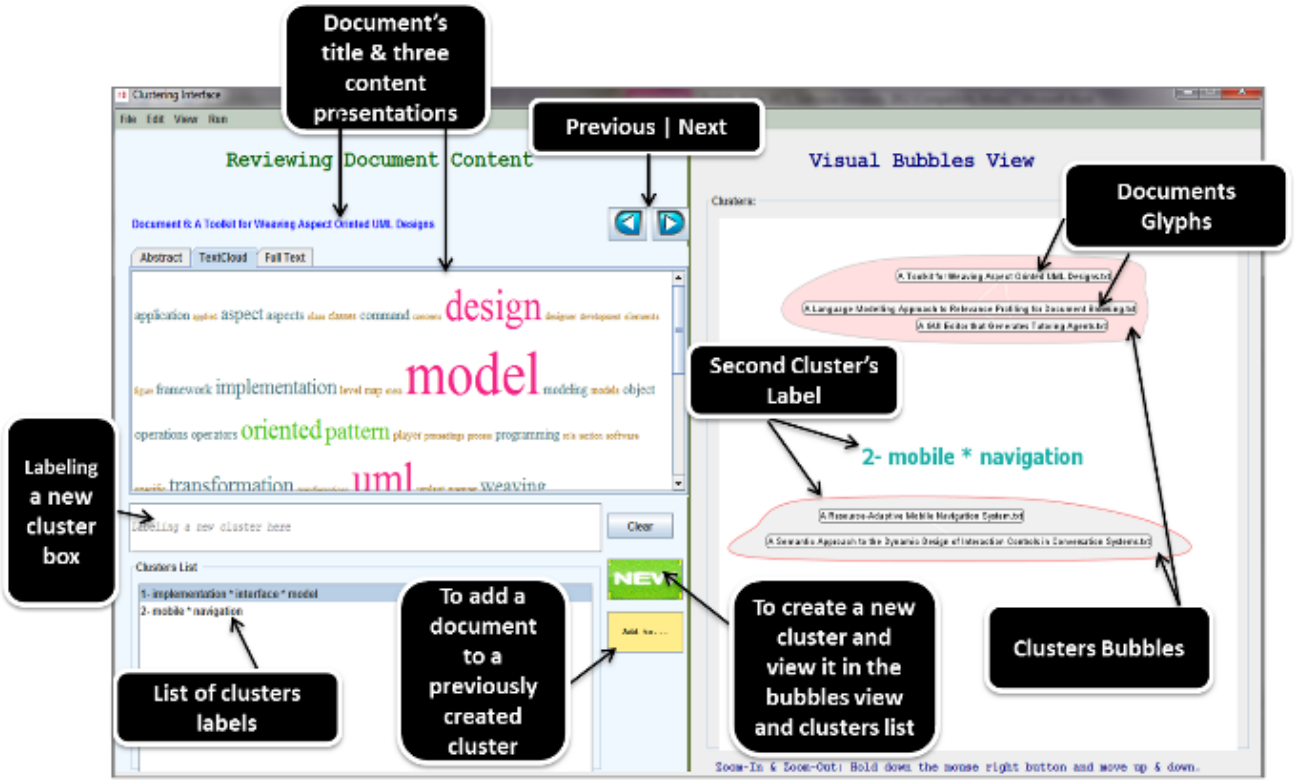


Figure 2. The Bubble Interface.

Ten participants—all computer science students—took part in this study. The small sample was meant to provide evidence of effectiveness of the prototype for further studies. The participants, even though are few, they represent early adopters. Of the participants, eight were males and only two were females. The ages of two participants were between 18 and 22. The ages of the remaining eight participants were between 23 and 30. All participants were graduate students.

D. Study Methodology

Every participant was given 30 documents (randomly selected from the collection used in the previous experiment) from which they could select 12 documents as seeds to clusters (1–12 clusters). They were given 15 minutes to classify the 12 documents into initial clusters. This was the supervision stage. The ten participants were split into two teams (Team 1 consisted of four participants while Team 2 consisted of six participants). The two teams met on two different days. On the first day of the study, each team had a meeting and the evaluation was completed as follows:

1. The team was divided into two groups (Group 1 and Group 2).
2. Each group was given a training session (approximately 5 minutes) on how to work on each interface.
3. Group 1 started working on the Bubbles Interface while Group 2 started working on the Pie Interface.
4. The participants were given the 30 documents used in the study two days ahead to familiarize themselves with the collection.
5. Every participant was asked to classify 12 documents from the collection of 30 documents into any number of clusters (1-12 clusters). After completing the classification process, the interface called the underlying clustering algorithm used in the work discussed in [9] and the remaining 18 documents were assigned by the algorithm to complete the clustering stage.
6. Every participant was asked to evaluate the clustering process by deciding whether or not each of the documents was assigned to the correct cluster from the participant's point of view.
7. Every participant was asked to complete a post-testing questionnaire about the interface they used.
8. The groups were then switched to follow the same steps 5 through 8 as described above.
9. A focus group discussion took place after completing the task on both interfaces.

 ISBN: 978-1-61208-361-2

TABLE I. A COMPARISON OF THE FEATURES ON THE PIE AND BUBBLE INTERFACES

No.	Features	Pie Interface	Bubble Interface
1.	Document representation	In a circle with a document ID	As a document title with a document index
2.	Mechanism of showing documents	Automatically after classifying the previous document	Using "*Previous*	*Next"* buttons
3.	Permanent document content view(s)	Plain text cloud + whole content	Abstract only + colorful text cloud + full text
4.	Other document content view	None	PDF format in a new window
5.	Creating clusters mechanism	Drag and drop a document into the "*New Cluster*" sector to create a new sector (cluster) containing a yellow stripe (document)	Click the "*New"* button. The new cluster label will be added to the Clusters List. A new bubble (cluster) with a glyph (document) will appear in the Visual Bubbles View
6.	The case of creating a cluster without a label	Although it is incorrect, the interface allows users to do so.	An error massage will pop up asking the user to create a label first.
7.	Visual view of clusters	a) Pie chart presentation b) Stripes (documents) within a sector (cluster) c) Not zoom-able	d) Visual bubbles presentation e) Glyphs (documents) within a bubble (cluster) f) Zooming in and out, and moving the bubbles around
8.	Viewing one cluster at a time	Not applicable	Allowed
9.	Skipping document(s)	Users are allowed to do one of two things to a document. 1) Assign it to a cluster. 2) Send it to the "*Trash"* sector (will not be considered in the clustering phase).	Allowed by hitting the "*Next"* button

A background questionnaire was used to gather demographic data about the participants. It was also used to collect information about the size of the participants' personal collections of documents and any tools they use to organize their documents.

The study was meant to evaluate the effectiveness, efficiency, and enjoyment of each interface and compare the two interfaces. The efficiency was measured using the time and the number of mouse clicks needed to complete the study. The perceived effectiveness and the engagement of the interfaces were measured through the data accumulated in the questionnaires and the accuracy of the clustering process.

The design of the experiment in [9] influenced the design of the current experiment in many ways. First, both studies used the same collection of documents. Second, the current study gave users the documents in advance since they were unhappy about the time they took to familiarize themselves with the documents in the study illustrated in [9]. The design of the Bubbles Interface attempted to change the visualization used in the Pie Interface and provide more interaction and display of content features, as seen in Table 1.

E. Study Results

1. Efficiency Result

The number of mouse clicks (left, right, and middle) during the study was logged. The number of mouse clicks on the Pie Interface was 301.3 on average (SD=66.6). In the case of using the Bubbles Interface, the average number of mouse clicks was 208.5 (SD=142.41). A two-sample-for-means z-test showed that no significant difference existed between the number of mouse clicks on the Bubbles Interface and the number of mouse clicks on the Pie Interface ($\boldsymbol{z}$ = -1.86, p = 0.06). However, by removing the outliers in the case of the Bubbles Interface, the difference became significant ($\boldsymbol{z}$ = -6.22, p<0.0001).

2. Effectiveness Result

To measure the effectiveness of the interfaces and compare the Bubbles Interface to the Pie Interface, every participant was asked to evaluate the accuracy of the final clustering of the 30 documents used. Every participant was asked to determine which documents were assigned to the correct clusters and which documents were assigned to the incorrect cluster based on the cluster topic built by the participant. The two-samples-for-means z-test results ($\boldsymbol{z}$ = -

 ISBN: 978-1-61208-361-2

2.93, p <0.003) indicated that there was a significant difference between the two interfaces with respect to the number of documents accurately clustered as perceived by the participants.

3. Enjoyment Result

The study used a post-task questionnaire for each interface after the user completed the task. Each questionnaire had 16 five-point Likert-scale questions that measured engagement factors considered in the study. The questions used involved the option of 'other' in most cases so that the user could provide different answers from the choices given. In all of the questions that used Likert-scales, the neutral case (i.e. the answer of 'not sure') was ignored from the analysis.

The first and second choices of the 5-point Likert-scale were merged and considered as one choice. The same procedure was followed with the fourth and fifth choices. The data was evaluated using the *z-test* (Downy et al., 2004) for comparing two proportions (equivalent to *Chi Square*). The following discussion goes through the results in each individual case measuring the engagement of the interfaces.

a. **How easy was the selection of documents for each cluster?** Nine participants chose 'easy' and 'very easy' for the Bubbles Interface, while only three participants found the Pie Interface to be 'easy' with regard to selecting documents for each cluster. The difference between the two proportions of participants (9/10 and 3/10) was significant ($\mathbf{z} = 2.739$, $p < 0.007$).

b. **How effective (helpful and useful) did you find creating labels for a new cluster?** On the Bubbles Interface, eight participants (8/10) indicated that creating cluster labels was 'effective'. The remaining two participants selected the neutral choice 'not sure' on the Likert-scale. On the Pie Interface, five participants chose 'effective' while three participants selected the 'not effective' choice. The difference between the two proportions of participants who considered the labelling feature on either interface as effective (8/10 and 5/10) was not significant ($\mathbf{z} = 1.41$, $p = 0.16$).

c. **How easy was modifying a cluster to add or remove documents?** On the Bubbles Interface, 70% of the participants (7/10) found it easy to modify clusters created during the supervision stage. Two participants indicated that it was difficult while the remaining one selected the neutral choice 'not sure'. On the Pie Interface, eight participants (8/10) found modifying clusters to be easy. One participant found it to be difficult while the remaining one was 'not sure'. The difference between the proportions of participants was not significant ($\mathbf{z} = -0.52$, $p = 0.60$).

d. **How clear did you find the view of your selected documents in the initial clusters?** On the Bubbles Interface and during the supervision stage, six participants (6/10) liked the clear presentation of their initial clusters. Two participants indicated that it was not clear while the rest selected the neutral choice 'not sure'. During the supervision stage on the Pie Interface, five participants liked the clear presentation of their initial clusters. Three participants found it unclear while two participants selected the 'not sure' choice. The difference between the proportion of participants who found the presentation of the initial clusters clear on either interface was not significant ($\mathbf{z} = 0.45$, $p = 0.56$).

e. **How helpful and effective did you find the final view of the clusters created by the system?** On the Bubbles Interface, four participants (4/10) found the final presentation of clusters to be helpful and effective. Three participants (3/10) indicated that it was neither helpful nor effective because of the overlapping of the documents' names while the three remaining participants (3/10) selected the neutral choice 'not sure'. On the Pie Interface, four participants (4/10) found the final presentation of the clusters to be helpful and effective. Four participants (4/10) considered it neither helpful nor effective while two participants (2/10) were 'not sure'. The difference between the proportions of participants who found the final presentation of the clusters helpful and effective on the Bubbles Interface and those who found it helpful and effective on the Pie Interface was not significant ($\mathbf{z} = 0$, $p = 0.99$).

f. **How do you rate the presentation of elements on the interface?** All participants (10/10) rated the presentation of elements on the Bubbles Interface as effective. Four participants (4/10) rated the presentation on the Pie Interface as effective while four participants rated it as not effective. There was a significant difference between the proportions of participants who found the presentation of elements on the Bubbles Interface to be effective and those who found the presentation of the elements on the Pie Interface to be effective ($\mathbf{z} = 2.93$, $p < 0.003$).

g. **How do you rate the positioning of the document view and cluster view on the screen?** The positioning of the document view and cluster view on the Bubbles Interface were considered effective by 70% of the participants (7/10). Two participants rated the views as not effective while only one participant selected the 'not sure' choice. On the Pie Interface, the positioning of the document view and cluster view were considered as effective by three participants (3/10). Four participants (4/10) rated the view as not effective and the remaining three participants (3/10) selected the 'not sure' choice. There was a significant difference between the proportions of participants who rated the positioning of the document view and cluster view on the Bubbles Interface as effective and those who rated the positioning of the document view and cluster view on the Pie Interface as effective ($\mathbf{z} = 2.25$, $p < 0.02$).

h. **How easy was it to undo actions on the interface?** On the Bubbles Interface, eight participants (8/10) rated the

ability to reverse actions as easy. One of the remaining two participants rated it as difficult and the other one selected the 'not sure' choice. On the Pie Interface, three participants (3/10) rated the ability to reverse actions as easy while three other participants (3/10) rated it as difficult. The remaining four selected the neutral choice of 'not sure'. The difference between the two proportions of participants who found reversing actions to be easy on either interface was significant ($z = 2.25$, $p < 0.02$).

i. **Was the feedback from the interface helpful to you?** The feedback from the Bubbles Interface was considered as clear and helpful by eight participants (8/10), not clear or helpful by one participant (1/10), and not applicable by one participant (1/10). The feedback from the Pie Interface was considered as clear and helpful by only three participants (3/10), not clear or helpful by two participants (2/10), and not applicable by one participant (5/10).There was a significant difference between the proportions of participants who found the feedback from the Bubbles Interface as clear and helpful and those who found the feedback from the Pie Interface as clear and helpful ($z = 2.25$, $p < 0.02$).

j. **How helpful and effective do you think the interface will be with organizing your collection of documents?** Seven users (7/10) predicted that the Bubbles Interface will be helpful and effective with organizing their own collections of documents. Two participants (2/10) anticipated that it will neither be helpful nor effective. Two participants predicted that the Pie Interface will be helpful and effective with organizing their own collections of documents. Four participants (4/10) anticipated that it will neither be helpful nor effective. There was a significant difference between the proportions of participants who expected the Bubbles Interface to be helpful and effective and those who expected the Pie Interface to be helpful and effective ($z = 2.24$, $p < 0.02$).

F. Study Limitations

The study had volunteers who were computer science students. The population of the study was very limited with regard to the number of participants involved due to limited resources. The number of documents used in the experiment was also limited because of the time required to manage more documents and investigate the effectiveness of the final clustering. The accuracy of clustering was manually examined which would have required more time and funding if more documents had been used.

IV. DISCUSSION

The study showed that users worked more efficiently on the Bubbles Interface than they did on the Pie Interface. The Bubbles Interface required significantly fewer mouse clicks by the user than the Pie Interface to complete the same task. However, there was no significant difference between the times needed to complete the task on the Bubbles Interface and the times needed by users to complete the same task on the Pie Interface.

Performing more clicks on the Pie Interface can be attributed to the user's need for very frequent scrolling in order to see the document content. This kind of scrolling was not needed as frequently on the Bubbles Interface. The reason for completing the tasks on both interfaces with no significant difference in the time needed can be attributed either to the nature of the task itself or to other factors that were not measured in the study.

Users achieved higher clustering accuracy with the Bubbles Interface than they did with the Pie Interface. One participant indicated that "*navigation among the document content views was much easier with the Bubbles Interface*". The Bubbles Interface may have helped users with assigning the appropriate documents together to represent a topic (cluster). It may have also helped users with identifying the documents in each cluster in the final results. The labelling process on the Bubbles Interface may have also helped with identifying the accurate topic of both the documents during the supervised classification stage and the clusters during the final presentation stage. One participant mentioned that "*I did very well in assigning documents into correct clusters."*

Several engagement factors have been addressed in the study. For example, the difference between the number of participants who found the process of selecting documents for clusters to be easy on the Bubbles Interface and those who found it easy on the Pie Interface was significant. This may indicate that the approach that was used to show the document content to the user was more effective on the Bubbles Interface. It may also indicate that users found it easier to perceive the cluster content and see where the new document belonged during the supervised initial classification.

Users also found the presentation of elements on the Bubbles Interface to be more effective than the presentation of elements on the Pie Interface. Users commented that the layout was intuitive and easy to understand and that no confusion or frustration was caused with the organization of the Bubbles Interface elements. During the group discussion, one participant stated "*I had some difficulties viewing the document content both with the text cloud and the whole content view. The area customized for displaying the content was not sufficient. It should be larger on the Pie Interface.*" Another participant indicated "*I got really lost with the Pie Interface because I always forget how to review the document and cluster content.*"

The positioning of the document view and cluster view on the display was considered effective and helpful by significantly more users of the Bubbles Interface than users of the Pie Interface. The participants reported that they "*found the interaction with the Bubbles Interface easier because of the nice layout that was easy to understand.*"

The feedback given by the interfaces was different. Significantly more users favoured the feedback given by the Bubbles Interface. For example, all the messages given by the Bubbles Interface were clear and were given to serve many purposes. However, the only feedback message given to the user while using the Pie Interface was the delete

 ISBN: 978-1-61208-361-2

conformation message when the user attempted to delete a cluster or a document. The users stated that the feedback of the Bubbles Interface was more helpful and reduced the need for asking the researcher questions to clarify the reactions of the interface. Two participants stated that they liked "*to have something on the interface indicating how many documents have already been classified and how many remain.*"

Users favoured the Bubbles Interface for future use with organizing their collections. Moreover, participants preferred the categorization of documents in the final results provided by the Bubbles Interface over the categorization provided by the Pie Interface. They indicated that in the case of the Pie Interface, there was little information about the documents in each cluster. The interaction with the interface to obtain more information about the clusters and the documents was hard.

Even though the Bubbles Interface has promising features in the organization of documents, it may have issues of clutter with very large collections. Different design parameters may need to be adjusted such as the glyph size for the document and the size of the bubble representing the cluster. The quality of clustering of a large collection of documents can be evaluated in the case of using the Bubbles Interface by evaluating the seeds selected for the clusters. It will be almost impossible (very time consuming) to ask users in a laboratory experiment to measure the accuracy of the final results in the case of very large collections of documents. However, the seeds chosen by the users to be given to the clustering algorithms can be evaluated by comparison to a ground truth.

There are several guidelines that can be drawn from the findings of the study. First, visualization through the use of intuitive bubble clusters would assist users in isolating clusters to locate documents in retrieval-based interfaces. Second, the use of labeled bubbles representing documents within clusters eases the process of identifying documents within clusters. Third, the interactive clustering in which changes are applied immediately to the visual view of clusters during the classification stage makes organization more effective. Fourth, providing different views of clusters may help in allowing users to continuously modify the groups of documents assigned to clusters according to changes in the observed topics. Finally, the use of intuitive clustering such that of the bubble interface would improve the user judgment of the documents assigned to each cluster.

V. CONCLUSION AND FUTURE WORK

The investigation compared specific features in a prototype interface against a baseline interface from previous research [9]. The results of the investigation showed that the new prototype interface had a better layout and helped users with: 1) the initial classification of documents into clusters during the supervised stage; 2) the modification of clusters; 3) the cluster labelling process; 4) the presentation of the final set of organized documents; 5) the efficiency of the organization process, and 6) the actual accuracy of the cluster for the organization process.

Further work will focus on the use of visualization and clustering parameters for more effective retrieval of documents in personal and even larger collections of documents. Studies will focus on improving the current prototype to provide efficient environment for organizing and managing in addition to retrieving documents.

REFERENCES

[1] A. Alhenshiri, and J. Blustein,"Exploring visualization in web information retrieval," International Journal for Internet Technology and Secured Transactions, vol. 3, issue 3, July 2011, pp. 320-330.

[2] H Badesh, and J. Blustein, "VDMs for Finding and Re-finding Web Search Results," iConference, Feb. 2012, pp. 419-420, Toronto, ON, Canada: ACM.

[3] D. F. Bauer, "Spatial Tools for Managing Personal Information Collections," HICSS2005, 2005, pp. 104-106, Hawaii, USA: IEEE Computer Society.

[4] O. R. Bergman, and R. Nachmias, "The project fragmentation problem in personal information," SIGCHI Conference on Human Factors in Computing Systems, April 2006, pp. 271-274, Montréal, QC, Canada.

[5] C. Carpineto, S. Osiński, G. Romano, and D. Weiss, "A Survey of Web Clustering Engines," vol. 41, issue 3, ACM Computing Surveys, July 2009.

[6] A. Civan, W. Jones, P. Klasnja, and H. Bruce, "Better to organize personal information by folders or by tags?: The devil is in the details," Journal of the American Society for Information Science and Technology, vol 45, issue 1, 2008, pp. 1-13.

[7] E. Cutrell, S. Dumais, and J. Teevan, "Searching to Eliminate Personal Information Management," In Communications of the ACM (Special Issue: Personal Information Management), Jan. 2006, pp. 58-64.

[8] D. Elsweiler, and I. Ruthavan, "Towards Task-based Personal Information Management Evaluations," the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2007, pp. 23-30, Amsterdam, The Netherlands: ACM.

[9] Y. Hu, E. Milios, J. Blustein, and S. Liu, "Personalized Document Clustering with Dual Supervision," the 12th ACM Symposium on Document Engineering, 2012, pp. 161-170, Paris, France: ACM.

[10] I. C. Jaballah, "Managing Personal Documents with a Digital Library," 9th European Conference, Research and Advanced Technology for Digital Libraries, 2005, pp. 18-23, Vienna, Austria.

[11] W. Jones, "Keeping Found Things Found: The Study and Practice of Personal Information Management," 2007, San Francisco, CA, USA: Morgan Kaufmann Publishers.

[12] W. Jones, H. Bruce, and S. Dumais, "How do People Get Back to Information on the Web? How Can They Do It Better? 9th IFIP TC13 International Conference on Human-Computer Interaction, 2003, Zurich, Switzerland.

[13] E. Jones, H. Bruce, P. Klasnja, and W. Jones, "I Give Up! Five Factors that Contribute to the Abandonment of Information Management Strategies. 68th Annual Meeting of the American Society for Information Science and Technology (ASIST 2008). 2008, Columbus, OH.

[14] S. H. Knoll, A. Hoff, D. Fisher, S. Dumais, and E. Cutrel, "Viewing Personal Data Over Time," CHI 2009 Workshop on Interacting with Temporal Data, 2009, pp. 1-4, Boston, MA, USA.

[15] M. Santini, and S. Sharoff, "Web Genre Benchmark under Construction," Language Technology and Computational Linguistics (JLCL), vol. 25, issue 1, 2009.

[16] OpenText, http://www.opentext.com

[17] Google Desktop, http://desktop.google.com

 ISBN: 978-1-61208-361-2

Towards a Mobile Application Performance Benchmark

Florian Rösler
Department of Cooperative Studies
Berlin School of Economics and Law
Berlin, Germany
florian.roesler@gmail.com

André Nitze
Department of Cooperative Studies
Berlin School of Economics and Law
Berlin, Germany
andre.nitze@hwr-berlin.de

Andreas Schmietendorf
Department of Cooperative Studies
Berlin School of Economics and Law
Berlin, Germany
andreas.schmietendorf@hwr-berlin.de

***Abstract*—In this work-in-progress paper, we present our current findings concerning performance efficiency in cross-platform mobile applications (apps) and how they can contribute to a general benchmarking approach. At first, several test cases for evaluating performance of mobile applications are described. Then, the performance efficiency of native and hybrid apps is compared on a mobile device using IBM Worklight. The results show that hybrid applications still suffer performance issues in comparison to native apps. The performance deviations and reasons for them are discussed and evaluated. It is concluded that the performance of mobile applications is crucial to user experience and satisfaction. Software quality should thus not be sacrificed, despite the economic attractiveness of hybrid development approaches. The results provide a starting point for a general approach to benchmark mobile application performance, which is discussed in the end.**

Keywords- Mobile applications; benchmark; software quality; performance efficiency.

I. INTRODUCTION

The market for mobile devices is currently contested by several Operating system (OS) providers. The two most popular OSs, Android and iOS, currently, make up about 90% of the market [5], but both bear big differences in their development processes. The remaining 10% are made of less popular OSs including BlackBerry, Windows Phone and Symbian. Therefore, when developing smartphone applications, a wide range of skills is required to cover all available platforms in their native environments. To supply this diverse market, software companies need a competent workforce that is capable of handling the development for the required platforms within multiple codebases. This leads to expensive development processes and costly maintenance.

To overcome the differences among the various OSs, several cross-platform development frameworks have been published to streamline the creation of apps for multiple platforms. As most of these frameworks are based on web technologies, web developers are able to build apps without first learning specific programming skills required by the individual platforms, eliminating the need for specialists for each targeted platform. This enables companies to employ a smaller and less specialized workforce, creating a more cost efficient way to create apps for multiple OSs. On the downside, web technologies bear limitations that confront development companies with a number of tradeoffs. Some of these limitations have been already made public by scientific research, whereas others still remain unclear.

There are several aspects of performance measurement in mobile app development. Delivering products in an efficient manner demands short development cycles with high quality (i.e., few errors), which can be seen as a form of process performance. The product performance (the app itself) is primarily reflected by user ratings in the app distribution platforms. Consumer apps with poor performance can lead to disgruntled users, who delete the app and subsequently cause negative publicity. In the future, apps meant for the business sector will continue to significantly affect business processes and revenues. In this case, the impact of performance will be much more of an issue since it can significantly constrain the operation of a company. For example, when there are contracts to be approved, sales representatives must be able to quickly receive customer data or conduct other time-critical processes dependent on mobile interaction with business data.

This paper shows performance related problems that come with cross-platform approaches comprising web technology. It aims to emphasize that mobile app development should not be conducted as economically as possible, but rather in a manner that is the most appropriate for the customer.

After considering related work in the field, we will describe the technical concept of hybrid applications. Then, we will describe the method used to gather data and present our results. Eventually, we will interpret our findings and outline an approach for further research.

II. RELATED WORK

Charland and LeRoux explain the key problems of cross-platform development, which include code execution time and User Interface (UI) issues [4]. They also point out that end users care about the quality of the app more so than they do about the efforts put into its development.

According to Ohrt and Turau, the use of cross-platform frameworks results in slower launch times and bigger application package sizes in comparison to their native counterparts [10]. The results for each individual framework vary widely, from being unremarkably less efficient to being slower and bigger by several orders of magnitude.

Corral, Sillitti and Succi test the performance of cross-platform apps in terms of accessing hardware features of an Android phone [6]. They conclude that most routines, except

 ISBN: 978-1-61208-361-2

one (launching a sound notification, 35% faster), are slower than native code. Whereas some routines are only slower by a factor of around 2, some are considerably slower, by a factor of 30 or even 500.

Toca compares several cross-platform development frameworks by measuring various functions, including start-up time and scroll performance [12]. He states that the usage of some frameworks may lead to a bad user experience; frame rates during scrolling drop to insufficient values and starting the apps sometimes takes longer than 10 seconds.

Heitkötter, Hanschke and Majchrzak identify criteria to rate cross-platform and native development [7]. Their work is based on interviews with domain experts and developing prototypes. As one of the reviewed frameworks, called PhoneGap appears as fast as its native counterparts, they conclude that cross-platform frameworks could also be an alternative when developing for a single platform.

III. HYBRID APPLICATIONS

Hybrid mobile apps are wrapped local web applications, which allow the execution of native code. This requires the native code pieces to be called out of the browser. Such a technique is known as the "PhoneGap Hack", which led to a library for calling several device APIs [2]. These are currently included and maintained in the PhoneGap framework, also known as Apache Cordova. Apache Cordova enables the creation of cross-platform apps using only Hypertext Markup Language (HTML), Cascading Style Sheets (CSS) and JavaScript. Moreover, developers are enabled to access a device's camera, Global Positioning System (GPS) sensor and many other device functionalities using JavaScript [1]. Cordova currently supports all the major Oss [2] and offers the possibility to implement plug-ins by the developer, which are own pieces of native code [3]. These plug-ins can then also be used with JavaScript calls.

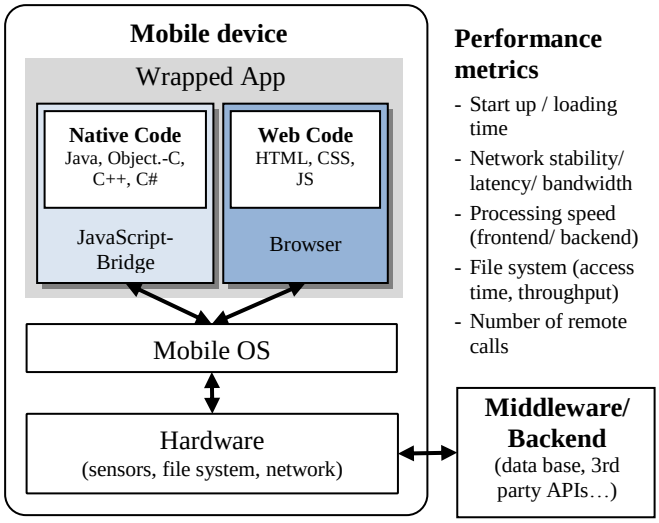


Figure 1. Architectural overview of hybrid mobile applications with performance metrics.

Among today's most prominent hybrid frameworks stand the already mentioned Apache Cordova and IBM Worklight. Unlike Cordova, Worklight is distributed under a proprietary license. While Worklight comprises Apache Cordova, it adds several features aimed at business applications. These include the operation of a backend server, which supports access to different data sources [8]. It also provides multiple authentication mechanisms and security concepts for accessing business data (ibid.).

Figure 1 shows the architecture of hybrid mobile applications and exemplary metrics to measure performance in mobile software systems, including device configuration, network characteristics, and backend and third party services.

As hybrid apps at their core are web applications, they utilize UI toolkits to display user interfaces. Because UI toolkits can only imitate certain behaviour of native controls, they sometimes lack the native look and feel that most users expect [11].

IV. METHOD

For a meaningful comparison two nearly identical Android apps, containing all in the following presented test cases are developed. Besides a native app, a hybrid IBM Worklight app is created, each utilizing jQuery Mobile as its UI toolkit. The defined test cases are meant to compare these apps by the subcharacteristics of performance efficiency as described in ISO/IEC 25010 [9], namely time behavior and resource utilization. Both versions only comprise basic UI elements and no rich media. In order to minimize the interference of background threads, the used smartphone is put into flight mode during testing. Every test case is executed ten times to obtain an arithmetic mean value.

A. System under Test

In order to retrieve comparable results, the test cases are each executed on the same device. The chosen device is a Samsung Galaxy Tab 2 10.1, which can be classified as a mid-class tablet and should therefore provide satisfactory performance.

When measuring time behavior, two timestamps are taken; one before a test case is executed and one right after execution has finished. In the case of resource utilization, instead of timestamps, a representational key figure for the memory consumption is recorded. As apps share resources among each other on Android, we use the Private Dirty Random Access Memory (RAM) as a representative key figure. The Private Dirty RAM indicates which amount of memory is only consumed by the specific app and is therefore freed upon closing the app.

B. Test cases

Although Ohrt and Turau already compared the start-up time of hybrid apps as well as their memory consumption after start-up [10], we recreate their experiments. This is because their tested apps were virtually empty and the hybrid app did not contain a UI toolkit. We expect a remarkable increase in time and memory consumption when the app's web resources are loaded. Those parts of an app cannot rely on an intelligent library sharing mechanism like Zygote, which shares Java libraries across apps.

 ISBN: 978-1-61208-361-2

It must be noted that Worklight apps are in general much more extensive than a basic Cordova app due to the included backend functionality. It is currently not possible to exclude these libraries from Worklight projects even when they are not used by an app.

In order to retrieve comparable values for the basic UI performance of an app, a certain amount of items are added to a list view. List views represent a common way of navigation and display of data, thus its performance is crucial to the overall impression of an app. In the case of the hybrid apps, the list items are added by utilizing standard DOM (Document Object Model)-methods. As Android utilizes data binding to connect an array to the list view, the creation of the array and its items are excluded from the time measurement. The described test case is additionally tracked in terms of memory consumption, thus indicating how efficient list items are handled by the specific system. Such a test case cannot act as a precise performance benchmark, but shall rather point out a general performance comparison as list view operations are a basic feature that should be executed close to real time. If an app struggles adding 100 items in a benchmark environment, where the number of background processes is minimized, it may have stronger execution issues when adding these items in a real life situation, where other processes take large amounts of processing power.

V. RESULTS

The outcome of the first test case reveals that, while the native application is nearly immediately loaded, the hybrid counterpart is significantly slower by a factor of around 20 (see Figure 2). With a startup time of more than two seconds,

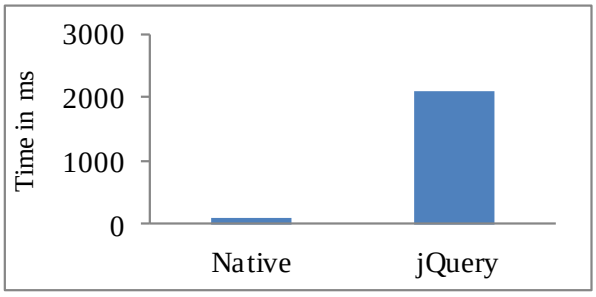


Figure 2. Start-up time comparison of native (Android) and hybrid (jQuery) apps.

the hybrid app shows a remarkable delay, which is noticeable by the user. In an app, which contains real content, this additional loading time may negatively influence a user's satisfaction.

When analyzing the start-up process further, it becomes clear that the native shell, which wraps hybrid apps, takes up a majority of the time span followed by the UI toolkit's loading time (see Figure 3). During this time, the internal Cordova server is started, which refers JavaScript calls to their native counterparts. Additionally, required JavaScript libraries as well as web resources are loaded into the browser view, which hosts the app. Thus, loading a native app is a more trivial process to the operating system and can be executed much faster.

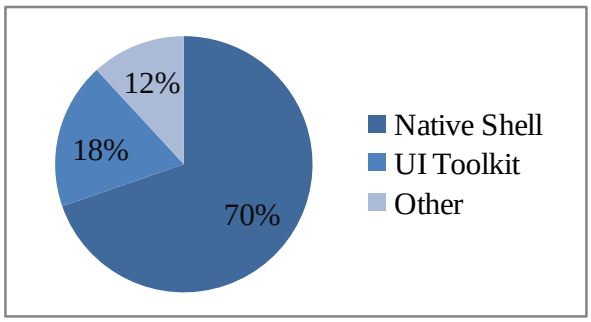


Figure 3. Start-up time details for hybrid (jQuery) apps.

The measurement of the memory consumption during start-up shows similar results. The differences in memory utilization after start-up are significant, with the hybrid application consuming more than four times the memory of the native implementation, which takes up less than 700KB (see Figure 4). This difference is explainable by the Worklight shell and the UI toolkit, which cannot be shared among hybrid apps. When running multiple hybrid apps at the same time, each utilizes its own copy of the aforementioned resources. Native Android apps on the other hand can share libraries that bring in basic functionalities like UI operations among each other, which decreases the overall memory footprint. Additionally, the DOM, which is required to display web pages within Cordova is also stored in the phone's RAM.

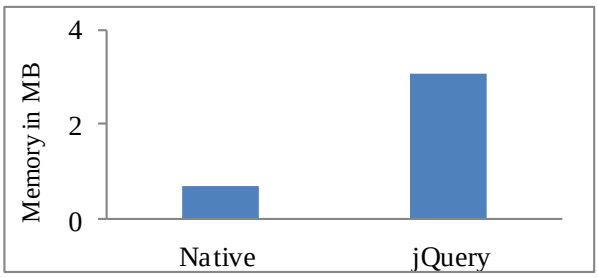


Figure 4. Memory consumption after start-up.

The test case for adding 100 list items to a list view shows that the native implementation performs close to real time (see Figure 5). In the case of the hybrid app, the process takes nearly half a second, therefore being slower by an

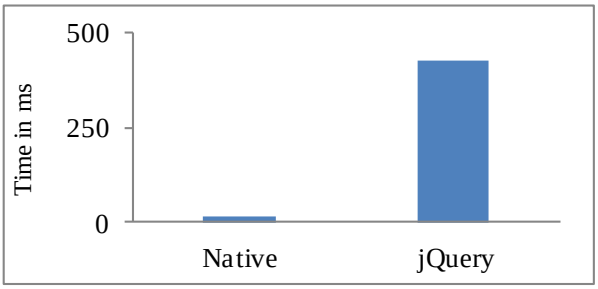


Figure 5. Time to add 100 list items.

order of magnitude. The reason for the difference when adding the items in the hybrid implementation could be the utilization of the DOM, which cannot compete with the efficiency of native UI mechanisms. Although the performance of the hybrid app is still acceptable, users might feel a delay when loading the screen with the list items, which again could affect the user's satisfaction. It also should be mentioned that low-end phones may show worse results. On low-end phones, adding list items can lead to long waiting times, which may be unacceptable for such a common operation.

The results for measuring the memory consumption when adding list items indicate that the native implementation is remarkably more efficient than the hybrid

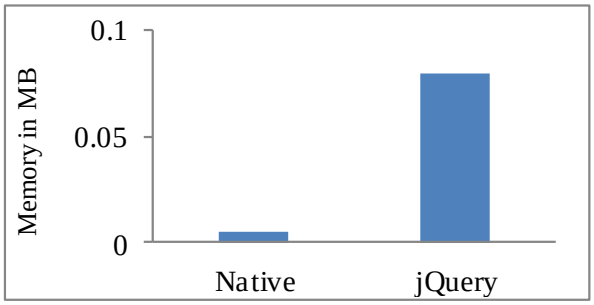


Figure 6. Increased memory consumption when adding 100 list items.

version (see Figure 6) as the jQuery Mobile app utilizes 16 times more memory than the native implementation. The reason for the higher memory increase of the hybrid implementation may again be the DOM, which stores the document in an expensive tree structure, which usually includes redundancies like recurring element names. On the contrary, the OS can handle native apps in a more efficient way and store the values in inexpensive data structures.

VI. CONCLUSION AND FUTURE WORK

Hybrid apps were analyzed in terms of performance efficiency, which is an important factor for the software quality of apps. In all conducted tests, native apps were superior to hybrid apps. Since performance is considered crucial for user experience, low performance is likely to influence a user's satisfaction and rating of the app. Users of low-end phones seem to be particularly disadvantaged by a market shift towards hybrid apps. A large share of the market for hybrid apps is currently advertised by many consulting companies due to the economically efficient development process. Despite this, companies should focus on their clients who expect a satisfying performance, which is more likely to be achieved with the native approach. Some cases cannot yet be covered sufficiently in terms of responsiveness using hybrid approaches. Although web technologies and hybrid frameworks are progressing steadily, native development prevails, at least for consumer-facing apps.

While many papers have already covered performance efficiency of hybrid mobile apps, there is still no clear statement of which approach to choose for a certain project. We therefore suggest the creation of a general benchmark method that can be implemented at the beginning of a software development project for evaluation purposes. It should cover the most important aspects of an app's performance, including the utilization of hardware features or UI performance. These tests should support lead developers and managers in deciding whether the disadvantages in performance are negligible for the certain use case.

As the environment of a hybrid app can differ in many factors like OS, hybrid shell, UI toolkit and smartphone hardware, it should be possible to implement the benchmark for a specific system in a cost efficient manner with low time expenses. However, a more general mobile application performance benchmark would need to include a set of configurations to cover the most widely used technological pathways. To achieve this, more factors apart from performance have to be incorporated as comparison criteria. Furthermore, a more typical set of UI elements should be derived from practical use cases. Also, more economical factors have to be included as their impact on the platform choice can be significant.

Model-driven development approaches like those discussed in [13][14] and [15] have not yet found wide adoption outside of academic projects and hence shall for now be excluded of performance evaluations.

Regarding future developments, it can be assumed, that the typical increase of computing speed and memory capacity of mobile devices will lead to improved performance. Nevertheless, decisions on the trade-off between performance and other factors will always have to be made.

REFERENCES

[1] Adobe PhoneGap 2013a. PhoneGap Documentation Overview. [online] Available at: <http://docs.phonegap.com/en/2.9.0/guide_overview_index.md.html#Overview> [Accessed 10 May 2014].

[2] Adobe PhoneGap 2013b. Adobe PhoneGap Build. [online] Available at: <https://build.phonegap.com/> [Accessed 10 May 2014].

[3] Adobe PhoneGap 2013c. Plugin Development Guide. [online] Available at: <http://docs.phonegap.com/en/2.8.0/guide_plugin-development_index.md.html> [Accessed 10 May 2014].

[4] Charland, A. and LeRoux, B. 2011. Mobile Application Development: Web vs. Native. In *Communications of the ACM*, vol. 54, 5 (May 2011), pp. 49-53. DOI= http://dx.doi.org/10.1145/1941487.1941504.

[5] comScore, 2013. US Smartphone Subscriber Market Share April 2013. [online] Available at: <http://www.comscore.com/Insights/Press_Releases/2013/6/comScore_Reports_April_2013_U.S._Smartphone_Subscriber_Market_Share> [Accessed 10 May 2014].

[6] Corral, L., Sillitti, A., and Succi, G. 2012. Mobile multiplatform development: An experiment for performance analysis. In *Procedia Computer Science*, vol. 10, pp. 736-743.

 ISBN: 978-1-61208-361-2

[7] Heitkötter, H. Hanschke, S., and Majchrzak, T. 2012. Comparing Cross-Platform Development Approaches For Mobile Applications. *Lecture Notes in Business Information Processing*, vol. 140, pp. 120-138.

[8] IBM 2012. IBM Worklight V5 Technology Overview. [pdf] Available at: <ftp://ftp.software.ibm.com/software/pdf/mobile-solutions/worklight/WSW14181USEN.pdf> [Accessed 10 May 2014].

[9] ISO 2011. ISO/IEC 25010:2011. Geneva: ISO.

[10] Ohrt, J. and Turau, V. 2012. Cross Platform Development Tools for Smartphone Applications. *IEEE Computer*, vol. 45, pp. 72-79.

[11] Quilligan, A. 2013. HTML5 Vs. Native Mobile Apps: Myths and Misconceptions. [online] Available at: <http://www.forbes.com/sites/ciocentral/2013/01/23/html5-vs-native-mobile-apps-myths-and-misconceptions/> [Accessed 10 May 2014].

[12] Toca, F. 2011. Cross-Platform-Entwickung unter iOS und Android: Technologieüberblick und Prototyp-basierte Bewertung. (Cross-Platform Development on iOS and Android) [Diploma Thesis] University of Magdeburg. Available at: <http://wwwiti.cs.uni-magdeburg.de/iti_db/publikationen/ps/12/thesisAlcalatoca.pdf > [Accessed 10 May 2014].

[13] Balagtas-Fernandez, F.T. 2008. Model-Driven Development of Mobile Applications. In: *23rd IEEE/ACM International Conference on Automated Software Engineering*, pp. 509-512.

[14] Dunkel, J. and Bruns, R. 2007. Model-Driven Architecture for Mobile Applications, In: *Proceedings of the 10th International Conference on Business Information Systems*, Springer, *v*ol. 4439, pp. 464-477.

[15] Kramer, D., Clark, T., and Oussena, S.: MobDSL: A Domain Specific Language for multiple mobile platform deployment. In: 2010 IEEE International Conference on Networked Embedded Systems for Enterprise Applications (NESEA 2010). Suzhou, S., pp. 1–7.

 ISBN: 978-1-61208-361-2

Redundancy-Driven Vertical Domain Explorer

Celine Badr
Dipartimento di Ingegneria
Università Roma Tre
Rome - Italy
badr@dia.uniroma3.it

***Abstract*—Entities, generally, represent real-world concepts, such as a person (writer, singer, etc.), a product (book, camera, etc.), a business, etc. In large data-intensive websites, sections related to an entity in a given vertical domain consist of a thousands of data-rich pages, each displaying attribute values for one instance of the given entity. Ideally, to build a rich repository of entity instances that serves the unlimited search needs of Web users, data aggregators aim to collect all the possible instances available for that given entity and apply data extraction for its attributes. A manual approach would be costly in time and effort. In this work, we propose a system that automatically discovers new large websites publishing pages about a conceptual entity, by exploiting the large amount of overlap on the Web among sources in the same vertical domain. Starting with information from one training site, specific queries are generated and results returned by search engines are analyzed and filtered. The sources retained from these search results undergo then a semantic, syntactic, and structural evaluation to detect data-intensive pages for the domain entity. Semi-structured attributes location is also identified on the discovered entity pages. Our approach can thus be exploited by vertical search engines in pre-processing to enhance web page crawling, as well as in data extraction.**

Keywords–entity discovery; vertical domain; search; keywords.

I. INTRODUCTION

Large data-intensive websites are composed of categories, each listing thousands of pages related to real-world conceptual entities. We refer to this set of pages, generally sharing a common structure or template, as *entity pages*. In Web data extraction, inferred wrappers extract selected attribute values on a subset/all of the site's entity pages. However, the extracted attribute values remain limited to the data available from the site's repository. Ideally, to build a rich vertical warehouse of instances of a given entity, data aggregators aim to collect all the possible instances available for that entity and extract its attributes on a large web scale. A manual approach would be costly in time and effort. Figure 1 shows, for example, a page representing an instance of the `Book` entity offered on a book selling website. Common attributes for the `Book` entity, like title, price, ISBN, publisher, etc., are listed on each such page. The actual values displayed on the page pertain to one `Book` instance and originate from one record in the underlying database. We use *instance page* to refer to one individual example of the entity pages.

In order to complement, enrich, or validate data gathered from a training site, it is useful to find instance data available on other sites that offer similar information on the same type of entities. This requires performing two main operations:

1) Find other large data-intensive websites offering pages on the given entity,
2) Download entity pages to extract their instance data.

Given one large website in a vertical domain, we presented in [2] a synergic method to locate its entity pages and extract instance data on them, with our CoDEC system that implemented it. The approach was facilitated by exploiting redundancies in the site's HTML structure, navigation paths, and content tokens. To extend it to other websites, in this work we address the problem of automatically locating large data-intensive sources in a given vertical domain. For that, we propose an approach that can be exploited by vertical search engines to enhance web page crawling and filtering by using domain knowledge in a pre-processing phase. Our solution uses as input the information collected during inference on a training site to find other large websites containing instance pages about the entity of interest. This is based on the fact that there is a large amount of overlap on the Web among sources in the same vertical domain [1], so information we have from one site, or possibly more, can lead to overlapping information in sources not yet discovered. For example, if our training website in the `Book` domain contains distinct instances of `Jane Eyre` book, `Oliver Twist`, and `Madame Bovary`, and we can find another website that also lists instance pages for these 3 books, it is very likely that this new website is a data-intensive source in the `Book` domain.

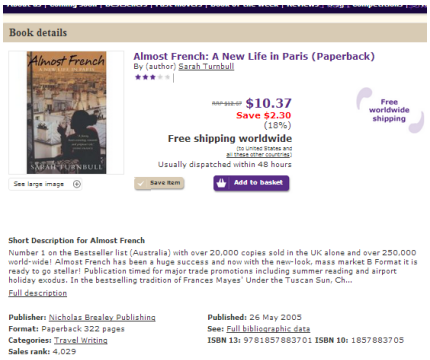


Figure 1: Data-Rich Instance Page of the Book Entity

The search is made possible by extracting from the collected information repository, keywords that represent the domain, the entity of interest, and a few selected instances, in order to *run queries* through a search engine. Many of the returned results are not data-intensive entity pages (e.g., blogs, news, reviews), which are of no relevance to our targeted approach. Thus we need to apply a two-level *filtering mechanism* to confirm or discard a result page based on its relevance: First, at the level of the returned URLs to determine the subsets of search results to be further examined; second, the pages pointed to by these URLs need be checked for

 ISBN: 978-1-61208-361-2

semi-structured data formatting to be considered candidates for eventual data extraction tasks.

We aim to perform these tasks in a fully automated way, independently of the training website. Our system, then, outputs a set of similar instance pages found on newly discovered websites and points out content nodes on them.

II. Proposed System Model

Entity instance pages are spread out on the Web, populating many large data-intensive websites. Thus finding and collecting these pages in a repository for a pre-determined domain consists mainly of a targeted web search activity, followed by an appropriate result filtering process. To conduct the above activities, we propose a system that exploits the data and domain knowledge collected on a training website. The gathered information is then used to facilitate discovering other potential large websites in the same vertical domain and eventually extract instance data from their entity pages.

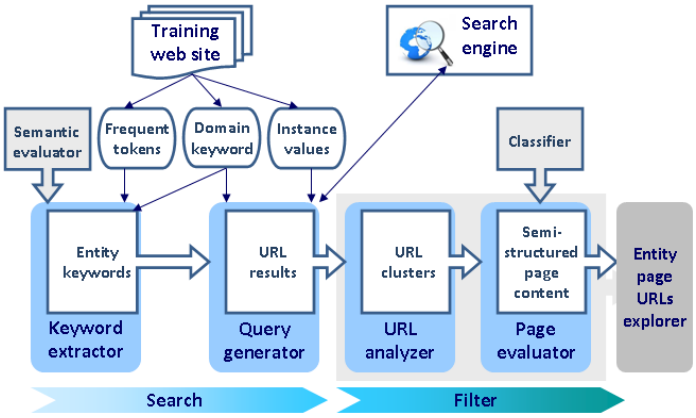


Figure 2: System components

The system is conceived of 4 main components illustrated in Figure 2. The components' goal is to direct and enhance the entity page search and filtering process as follows:

- The input to the first component is the information obtained from the training website, in particular, the template tokens from the sample pages. This component then analyzes these tokens in combination with the domain identifying terms and with the help of a semantic evaluator to output keywords for the entity of interest (Section III).
- Keywords are then passed to the query generator component that builds a series of boolean queries and sends them to a web search engine (Section IV).
- The URL results from the respective queries are passed as input to the following component, the URL analyzer. The latter applies a clustering algorithm on the set of URL results and keeps only relevant clusters for further analysis (Section V).
- These are then passed to the page evaluator component that downloads the pages from the Web and identifies entity-related semi-structured content (Section VI).

When various instance pages are identified on a website, a customized crawler can find the page that links to them on that site and download all the other instances that it contains. In addition, the system keeps record of the location (containers) of instance data on the pages, facilitating further data extraction efforts. The following sections describe in more detail each system component and the functionality that it implements.

III. Finding Entity Keywords

In general, any effort to find content on the Web passes through a search engine by providing keywords. Search engines today use implicit techniques (contextualization, approximation, search history, PageRank, etc.) to offer useful results in a ranking order estimated relevant to the information need. However, keywords remain essential to state the initial intent of the search, and choosing "good" search terms minimizes the distance between what's sought and what's found in the search engine's indexed content repository. This section presents the keyword extractor component for selecting query terms likely to generate relevant results. Results are relevant if they lead the system to discover new data-intensive large websites for the given domain entity.

In our case, we would like to use search engines to find semi-structured instance pages for a given entity. For an efficient web search, we need to have well formulated queries and a subsequent mechanism to confirm or discard a returned result page based on its relevance. For this operation, a user's manual intervention would not scale. Consequently, in our automated approach, it is the system's responsibility to formulate the queries by choosing a combination of "useful" keywords likely to lead to new web pages for the entity of interest. In later sections, we also explain how the system automatically filters returned search results. Our result acceptance criteria combines both the page content and the format in which it is presented.

To automatically select *entity keywords* for our search queries, that is, keywords tightly associated with the entity of interest, we rely on the information derived from the training site. We aim to find terms that are both quantitatively and qualitatively related to that entity. Common tf-idf measures are counter-intuitive here as we want template words occurring frequently on all sample pages, and not the opposite. When estimating fixed template tokens during the inference process [2], the system kept text nodes appearing frequently at the same location in the set of sample pages as potential template tokens, while tokens that were particular to one or few pages were considered to be variable content. Many may be labels typical of the domain, while others may be more general. Some may be stop words or extra information present on the page but not related to the domain entity. The recurring text fields collected constitute then a good starting point to find candidate domain keywords related to the entities. We propose to narrow down on entity vocabulary with the help of a semantic evaluator. The semantic evaluator takes two words or sets of words and returns a value that represents their semantic distance. The closer they are semantically, the higher the score returned by the evaluator. Since the domain identifier token is already given to the system as input, the system is able to rate, for each of the words collected, its relationship to the domain identifier expression. The words that score highest are kept. After also cross-evaluating these top-scoring words, the final keyword candidates set consists of terms that

- Appear frequently on instance pages of the vertical domain,
- Are closely related to the domain identifier string,

 ISBN: 978-1-61208-361-2

- And, additionally, are closely related among each other from a semantical aspect.

This automatically selected collection of terms replaces the need to manually select entity attribute labels for each domain without knowing the relevance or frequency of these labels when applying the search to the remaining pages on the Web.

Book	NBA Player	Restaurant	University
paperback	player	hotel	college
publish	soccer	shop	campus
author	game	pub	faculty
bible	tennis	cuisine	school
write	sports	store	graduate
chapter	golf	food	student
stock	team	establishment	education
item	stadium	nightlife	teaching
note	season	drink	institutional
	jersey	club	institute
	complete	city	sciences
			undergraduate

Figure 3: Entity keywords

Examples of resulting word sets for 4 domains are shown in Figure 3. Two entity keywords (highlighted) are derived from the related words pool, such that, with the domain keyword, they have pairwise a high semantic correlation.

IV. Constructing Queries

Since we are interested in finding large websites containing entity instances of one vertical domain, we build on the observation in [1] that there is a significant amount of connectivity and redundancy in content among data sources within the same domain on the Web. The existence of overlapping entities among different sources permits the discovery of new websites starting with entities already discovered, operating with a set-expansion approach. Based on this fact, we propose to start with the entities gathered from our inference website and conduct a search for overlapping entities on the Web.

The domain knowledge acquired on the training website consists of the domain identifier terms, the entity keywords described in Section III, and all the values of the entity attributes extracted on the instance pages by the inferred wrappers. With this knowledge, we propose to build search queries for different instances that are likely to find other pages on the Web describing these respective instances. When some discovered instances are determined to belong to a new data-intensive website, they can constitute the seeds for a crawler tailored to find the rest of its instances.

Each query is composed of the domain identifier terms, the entity keywords extracted for that domain, and one record from the instance attributes stored in the information repository. Various queries, each for a specific instance in the repository, run in parallel. For each instance query, pages returned by a search engine are considered relevant if they are semi-structured pages about the same or a similar entity instance. To restrain the search scope, we require that attribute values used in the query be matched on the result pages as they are part of the instance-related content. The identifying attribute is required, while other attributes can be optionally added to the query terms. The domain and entity identifier terms are rather descriptive and not necessarily expected to be in the content of an instance page. Therefore, in each query composition, attribute instances are combined with logic AND, while the domain and entity terms are joined with the OR conjunction.

For example, in the book domain, the domain identifier term is `book`. The entity keywords derived from our inference website are `publish` and `write`. A random instance extracted on the inference site has the attribute values `Great Expectations` for title and `Charles Dickens` for author. The title attribute is the instance identifier. The query composition for this example is then:

```
``Great Expectations'' AND ``Charles Dickens'' ``book'' ``publish'' ``write''
```

The `OR` operator is implied by default. Queries are constructed for a number of different instances and each is sent to the search engine. The search results URLs are collected for each query, but results pages are not yet downloaded.

We note that some attribute values are likely to yield less matching search results than others, due to differences in dates or measures format, for example, or the presence of abbreviations and spelling variations in the values of search phrases. This can be mitigated by diversifying instance values and attributes selection for a wider search coverage.

V. Filtering URL Results

In this section, we explain how the system goes about the numerous query results returned by the search engine, and how the first step of the automatic filtering is performed. The main idea is to select only a useful subset of the URLs collected from different instance queries for further processing, instead of downloading all the web pages listed in the results.

When a query is sent out to a search engine, the latter returns a very large number of results in response. Because of the entity redundancy principle discussed in section IV, the more frequently a website appears in the result sets, the higher the probability of it being a large website with content redundancy responding to our search needs. Equally, the more overlap there is, the more frequently that website will show up in the result sets. We aim to exploit similarities among URLs to group result pages into clusters, such that pages in a cluster respond to distinct instance queries, but belong to a single website and have little dissimilarity in their URL patterns. Each qualifying cluster is then analyzed to see if it contains candidate pages belonging to a large website.

http://www.rakuten.com/prod/the-cold-war-a-new-history/31198770.html
http://www.abebooks.com/book-search/kw/fact%F3tum-charles-bukowski/page-1/
http://www.alibris.com/Power-Multi-Level-Marketing-Mark-Yarnell/book/5267499
http://amblingbooks.com/books/view/emotional_alchemy_2
http://www.shelfari.com/books/8515328/The-Landscape-of-History
http://www.rakuten.com/prod/your-first-year-in-network-marketing/30327356.html
http://www.alibris.com/African-Cry-Jean-Marc-Ela/book/158642
http://productsearch.barnesandnoble.com/search/results.aspx?ATH=Mark+Yarnell
http://www.shelfari.com/books/373097/Stone-Rain
http://www.abebooks.com/book-search/title/sharpes-prey/page-1/

Figure 4: Sample query result URLs

There are some existing works that propose algorithms to cluster large websites [3]. However, these approaches assume that the website is already given and they try to discover its structure. Inspired by Blanco et al. [4] that combine URL analysis with some simple content features, we use URLs clustering as a starting point for a further website exploration.

 ISBN: 978-1-61208-361-2

By looking at a reduced sample set of URLs gathered from a results pool for queries in the book domain (Figure 4), we can already spot some recurrences. To operate on a large scale, the system has to automatically determine which URLs have similar patterns and group them together. Thus, we opt for a hierarchical agglomerative clustering (HAC) algorithm to process the result URLs collected from all the instance queries. HAC is a simple "bottom-up" technique that fits our data set, where the percentage of results to be merged into clusters is small with respect to the entire set of URLs returned by the search engine for the instance queries. We need to specify for our problem a metric defining a distance measure for two URLs. A URL can be broken into different parts (protocol, host, port number, path, query string, etc.), some of which are optional. We define the distance, or dissimilarity, between two URLs based on the parts they have in common and those that are different, as shown in Figure 5. For successive iterations,

Algorithm 1: Measuring Dissimilarity of Two URLs

Input : URLs $\{u_1, u_2\}$
Output: The distance D between u_1 and u_2

```
if u1.authority == u2.authority then
    Let D be the number of different parts between
    u1.path and u2.path;
    if u1.protocol != u2.protocol then
        D++;
    if u1.query != u2.query then
        D++;
    if u1.ref != u2.ref then
        D++;
else
    D = INFINITY;
return D;
```

Figure 5: Measuring Dissimilarity of Two URLs

a linkage criterion needs to be defined for the algorithm to compute the distance between two sets of elements as a function of the distance of the elements these sets contain. A possible function is the minimum distance between elements of each cluster, referred to as single-linkage clustering. For clusters $C1$ and $C2$, their distance is:

$$\forall x \in C1, \forall y \in C2 : d(C1, C2) = min(d(x, y)) \quad (1)$$

where x and y are URLs. Hence, at a given iteration, the two clusters separated by the shortest distance are merged. This implies that for a subsequent iteration, the minimum distance between clusters is larger than that at the previous step. A stopping condition for the algorithm can be either a threshold that says clusters have become too distant to be merged, or, when applicable, a predefined number of clusters to reach. Based on the URL distance metric, we set our stopping condition as a small integer, between 1 and 3.

From the clusters generated by HAC, we can consider as valid candidates for instance pages those URLs that:

- Occur in multiple distinct instance search results,
- Originate from the same website,
- Share a pattern with low dissimilarity value.

URLs satisfying these properties make it through the first filtering step of the system and are then further examined to determine if they constitute semi-structured instance pages from a data-intensive website. URLs occurring only occasionally or not matching any cluster (other than their own singleton) are not considered of interest and their pages are not downloaded.

VI. Page Evaluation

In this section, we describe how pages at the resulting URLs undergo the second stage of filtering through our system. We adopt as a reasonable pre-condition the fact that pages belonging to large data-intensive websites are generated by regular templates with some level of structure. The responsibility of the page evaluator is then to determine whether the pages at the given URL addresses contain semi-structured data sections that can be of interest for the data extraction task. Pages not contaning any data sections with structure can be discarded for our purposes. For each URL in the clusters returned by the previous system component, the corresponding web page is downloaded to be analyzed. To separate interesting from non-interesting pages, we proceed in three steps:

- Locate relevant fragment(s) on the page,
- Extract features from page fragment(s),
- Classify the page based on extracted features.

Given a downloaded page, the first task in this component's process consists in identifying in its content the fragments to evaluate with regards to the originating search purposes. In large data-intensive websites, instance pages have at least a section that displays the entity attributes in semi-structured format. Since our data extraction goal is to retrieve semi-structured attributes values where available, we need first to locate the HTML part where they are displayed on the page. Some works propose vision-based analysis and DOM tree alignment, but we opt for a less complex approach. Instead, we take the query that generated this result page and we search for the least common ancestor HTML container of the attribute values in that query on the given result page.

Depending on the occurrence of the attribute values, the following are the possible scenarios that can be encountered:

- Attribute values are located in one page fragment: the corresponding HTML container is returned for analysis.
- Attribute values are found in several page fragments: a list of HTML containers is returned and each will be analyzed separately.
- Attribute values are not found on the page: the given result page is discarded.

Given an HTML fragment where these attribute values appear, the automatic page filtering sequence proceeds to analyze it with respect to some structural and content features that are commonly observed on data-intensive web pages. Occurrences in text sections are not of any interest for our data extraction purposes. Recurrent characteristics have been identified and used to train a classifier in order to automatically distinguish structured from non-structured content. Namely, we look into the text length, recurrence of characters such as the column, usage of lists or table cells, and other HTML formatting aspects. A result page is boosted as potential instance page if at least one of its extracted fragments is classified as semi-structured. Otherwise, the page is not considered a valid candidate and is discarded.

 ISBN: 978-1-61208-361-2

The output from the page evaluator component consists then of search result pages already retained in a URL cluster, where the respective query attribute terms occur in a semi-structured layout. These are the final candidate instance pages of the system. Clusters with more classified candidate pages will have higher priority to be processed by the site explorer for crawling more instances from the discovered website to populate the vertical domain repository.

VII. EXPERIMENTS

We describe here our Vertical Domain Explorer system *(VerDE)* implementation, the experiments we conducted, and the results obtained, with some analysis and comments.

VerDE system is implemented in Java as a set of packages. For the semantic evaluator, we use the semantic similarity service provided by the University of Maryland, Baltimore County, which combines Latent Semantic Analysis (LSA) and knowledge extracted from WordNet to evaluate word similiarity. For the clustering module in the URL analyzer, we opt for the HAC algorithm with single linkage, and base our implementation on the Java library provided by the *Sape* research group at the University of Lugano. Support scores are computed for each URL results cluster to reflect their coverage of the different instance queries. Clusters with a score below a set threshold are not processed any further. Finally, the page evaluation component requires a classifier to assess the structure of the page content where the instance attributes are located. Because of the binary nature of the expected output (semi-structured content or not), the classifier is implemented as a logistic regression. Eight features related to text length, punctuation, and HTML formatting are used. We report the accuracy scores listed in Table I for the performance of our classifier model on positive and negative HTML content collected on various pages on the Web. All the implemented VerDE components are integrated to smoothly deliver the functionalities of the automated search and filter approach that the entire system builds on.

TABLE I: CLASSIFIER ACCURACY

Training set accuracy	81.82%
Cross-validation set accuracy	93.75%
Test set accuracy	100.00%

We evaluate the results obtained from VerDE by running experiments in 4 vertical domains: Restaurant, University, Book, and NBA Player. Once retrieved from the inference site tokens, entity keywords are stored for future use to enhance performance. For attribute values, random records are selected from the database to formulate queries during the system execution. A match for these values is then sought in the search results. For evaluation, we include in query constructions an entity identifier attribute and a variable second attribute and we run experiments with 5 and 50 random instances for each domain. The instance number and attribute selection of each run are specified in the experiment configuration.

In total, 10104 URLs were collected during the experiments, while only about 15% of them became part of clusters with URL redundancies and 11.73% were finally classified as semi-structured. This translates into a considerable effort saved from downloading unuseful pages. From the pages where the attribute values were matched in the first phase of automatic filtering, 79.6% were classified positive in the second filtering phase, highlighting the benefit of the URL pre-processing step.

Table II lists the percentage of examined pages with respect to the total URLs collected from search engine results, precision of the semi-structured classification, and number of distinct new discovered sources. The results reflect a clear

TABLE II: EXPERIMENT RESULTS

Domain	% Examined	Precision	# Discovered
RESTAURANT	3.32	0.83	54
UNIVERSITY	25.89	0.91	342
BOOK	8.03	0.83	245
NBA PLAYER	26.32	0.84	469

optimization in the automated effort of exploring websites for crawling and data extraction, allowing a site explorer to focus the processing on likely candidates of large data-intensive websites. Due to the huge amount of results returned by the search engine, recall is hard to evaluate, but an estimate of 73.4% was calculated by manual verification on a sample URL subset. We note that we observed poor accuracy on numerical attributes, e.g., height, phone numbers, etc., mostly due to wide variations in formatting on the Web and our heuristics being based on exact matching. However, diversifying the attribute selection in queries for a given entity can compensate any loss in results coverage. For example, queries with books title and publisher would find matching sources that queries with title and ISBN numeric values did not find. Another option would be to relax the boolean search with approximation or regular expressions when matching values on pages.

VIII. RELATED WORK

The VerDE system we presented touches on many subjects in information retrieval. The research work in [5][6][7][8] presents topical crawlers that filter the portions of the Web to be crawled. However, topical crawlers do not necessarily find large data-intensive websites. Challenges are also highlighted in selecting non-biased crawl seeds and the ability of the crawler to distinguish relevant from non-relevant documents. Our approach relies on filtering the URLs to be downloaded by identifying redundancies. It automatically selects candidate URL seeds likely to yield entity pages matching the crawling objectives. It also allows to limit the crawl to sections of the website that satisfy constraints both on content and format.

In recent years, the need for vertical search engines dedicated to topical services like news, finance, shopping, etc., has motivated efforts for highly accurate information retrieval techniques. Nie et al. [9] build object search engines in the academic and product search vertical domains. They statistically estimate that 12.6% of randomly crawled pages are product pages. This echoes the estimation range we also report in our result findings. Similarly, Nguyen et al. [10] consider the issues of data extraction, schema reconciliation, and data fusion, in building a system that synthesizes products for shopping verticals or product search engines. Hao et al. [11] start with one labeled example site from a given vertical and trains a system to extract data from unseen sites in that same vertical, independently of the domain. Also, Song et al. [12] exploit entity redundancy in different websites to learn entity attributes from inner- and cross-site features. In all these works, input pages are provided either by topical crawlers or by manually specifying the web source to analyze. No

 ISBN: 978-1-61208-361-2

optimization is performed at the search and download stages to only target data-intensive and semi-structured entity pages, which would reduce considerably the amount of later page processing. On the other hand, our approach discovers new data-intensive sites automatically, thus reducing human effort and also avoiding unnecessary page downloads. Moreover, VerDE individuates the HTML containers where the semi-structured data are displayed. This can resolve some extraction obstacles the other systems face where an attribute like book title appears several times on the page with different values, as various recommended instances are listed on the same page with the entity instance data.

One approach similar to ours is by Blanco et al. [13]. While both our work and theirs address the domain-independent page gathering task, the two approaches differ in several aspects:

- They analyze several similar websites to perform quantitative keyword extraction and bootstrap the system. Our approach uses data from one training site with a quantitative and qualitative semantic evaluation to find meaningful keywords for the vertical domain.
- Their system determines instance pages based on entity keyword appearance and hyperlinks location. This may lead to false positives in websites that do not offer semi-structured data-intensive pages, while our page filtering is based on a trained classifier that evaluates sections of interest considering content, structure, and formatting, which also reduces noise.
- For each URL result returned by the search engine, they run a full website evaluation, which is not a trivial task, especially when the URL belongs to a large website composed of thousands of pages. The website exploration is even repeated if a derived template evaluates as a potential instance page. In contrast, we minimize the number of candidate URLs before they are processed any further. The second step of page classification avoids exploring a website if no semi-structured content is detected.

Another related work is by Weikum and Theobald [14]. They present a knowledge harvesting technique to construct a comprehensive knowledge base of facts by extracting semantic classes, mutual relations, and temporal contexts of named entities. In this context, the semi-structured nature of data-intensive pages cannot be exploited in a pattern-based facts extraction without substantial postprocessing of the output.

Some recent data extraction approaches address wrapper generation using visual content features from the web sources [15][16][17]. Such approaches examine the resemblance of data records to build a block tree, then proceed to data record extraction and data item extraction, assuming the relevant information block is centered in one main region on the page. In contrast, our page evaluator component detects and classifies the block containing relevant attribute values.

IX. Conclusion and Future Work

In this work, we presented an approach to automatically and efficiently locate large data-intensive web sources in a given vertical domain, by starting with knowledge gathered from a training site and exploiting redundancies in entity occurrences on the Web. The system prototype implemented is composed of 4 logical components: a keyword extractor and an automatic query generator for the search tasks, then a URL analyzer and a page evaluator for the filter step. Our experiment results show a great advantage in using the automatic 2-stage filtering before exploring websites returned by search engines, and a high level of precision of our classifier. Future work can address the implementation of the website crawler that exploits output page clusters, in addition to application of data extraction techniques on the semi-structured containers identified on the pages.

References

[1] N. Dalvi, A. Machanavajjhala, and B. Pang, "An analysis of structured data on the web," Proc. of the VLDB Endowment, vol. 5, no. 7, 2012, pp. 680–691.

[2] C. Badr, P. Merialdo, and V. Crescenzi, "Synergic data extraction and crawling for large web sites," in ICIW 2013, The 8th International Conference on Internet and Web Applications and Services, 2013, pp. 200–205.

[3] I. Hernandez, C. R. Rivero, D. Ruiz, and R. Corchuelo, "A tool for link-based web page classification," in Advances in Artificial Intelligence. Springer, 2011, pp. 443–452.

[4] L. Blanco, N. Dalvi, and A. Machanavajjhala, "Highly efficient algorithms for structural clustering of large websites," in Proc. of the 20th international conference on World wide web. ACM, 2011, pp. 437–446.

[5] 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, 1999, pp. 1623–1640.

[6] M. Chau, "Spidering and filtering web pages for vertical search engines," in Proc. of the Americas Conference on Information Systems, AMCIS, 2002.

[7] S. Sizov et al., "The bingo! system for information portal generation and expert web search." in CIDR, 2003.

[8] A. Patel and N. Schmidt, "Application of structured document parsing to focused web crawling," Computer Standards & Interfaces, vol. 33, no. 3, 2011, pp. 325–331.

[9] Z. Nie, J.-R. Wen, and W.-Y. Ma, "Object-level vertical search." in CIDR, 2007, pp. 235–246.

[10] H. Nguyen, A. Fuxman, S. Paparizos, J. Freire, and R. Agrawal, "Synthesizing products for online catalogs," Proc. of the VLDB Endowment, vol. 4, no. 7, 2011, pp. 409–418.

[11] Q. Hao, R. Cai, Y. Pang, and L. Zhang, "From one tree to a forest: a unified solution for structured web data extraction," in Proc. of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. ACM, 2011, pp. 775–784.

[12] D. Song, Y. Wu, L. Liao, L. Li, and F. Sun, "A dynamic learning framework to thoroughly extract structured data from web pages without human efforts," in Proc. of the ACM SIGKDD Workshop on Mining Data Semantics. ACM, 2012, p. 9.

[13] L. Blanco, V. Crescenzi, P. Merialdo, and P. Papotti, "Supporting the automatic construction of entity aware search engines," in Proc. of the 10th ACM workshop on Web information and data management. ACM, 2008, pp. 149–156.

[14] G. Weikum and M. Theobald, "From information to knowledge: harvesting entities and relationships from web sources," in Proc. of the 29th ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. ACM, 2010, pp. 65–76.

[15] P. L. Goh, J. L. Hong, E. X. Tan, and W. W. Goh, "Region based data extraction," in Fuzzy Systems and Knowledge Discovery (FSKD), 2012 9th International Conference on. IEEE, 2012, pp. 1196–1200.

[16] K. Simon and G. Lausen, "Viper: augmenting automatic information extraction with visual perceptions," in Proc. of the 14th ACM international conference on Information and knowledge management. ACM, 2005, pp. 381–388.

[17] L. Li, Y. Liu, and A. Obregon, "Visual segmentation-based data record extraction from web documents," in Information Reuse and Integration, 2007. IRI 2007. IEEE International Conference on. IEEE, 2007, pp. 502–507.

 ISBN: 978-1-61208-361-2

Comparing the Twitter Usage of Online Retailers in Germany and in the UK

Georg Lackermair, Daniel Kailer
Munich University of Applied Sciences
Department of Computer Science and Mathematics
Munich, Germany
Email: {georg.lackermair, dkailer}@hm.edu

***Abstract*—The usage and acceptance of Twitter microblogging differs from region to region as various works have shown. As this platform is gaining importance as a channel to reach customers in the E-commerce, the question arises whether this differences get apparent in the Twitter usage of online retailers as well. This paper investigates this question by comparing German and UK based online retailers empirically. A data set composed of the top selling companies from both countries is analyzed quantitatively. For this purpose, a conceptual model is presented to classify different interaction strategies for microblogging in the E-commerce domain. There are four different strategies used to distinguish between a more bidirectional, interactive communication and a more unidirectional, promotional communication. The model used distinguishes between direct dialogs inside the borders of Twitters and the redirection of users to other social networks.**

Keywords–E-commerce; Twitter; Social Web; Germany; UK.

I. INTRODUCTION

The transformation of the Web from an unidirectional media of linked content towards an interactive communication platform affects the E-commerce massively. Due to its distributed nature, the Web 2.0 extends the range of channels for distributing information to the (potential) customers respectively. Another aspect is the bidirectional flow of communication. This requires companies to take care of information published by users and to react somehow to this.

A major actor in the social web is the microblogging platform Twitter. Every day, about 500 million status messages are published on this platform. Reasons for its success can be found in its simplicity, scalability, ubiquity, and interactivity. Due to its publish/subscribe capabilities, traditional newsfeeds based on Rich Site Summary (RSS) or Atom are shifting gradually to Twitter. This development applies to the E-commerce as well, as within this domain microblogging plays the role of a update notification capability for tools for personalization and direct customer interaction, e.g., discussion boards, weblogs or newsfeeds.

In academia, a considerable amount of research was already conducted to understand the usage of Twitter, in particular the communication conventions, user intentions and the network structure. Several works suggest that the usage of this platform is related to regional characteristics (1)(2). A blog post by The Economist states that Twitter is in Germany less popular than in other countries (3). The data shows that the ratio of Twitter accounts related to the size of population is a multiple times higher in Great Britain than in Germany.

But to the best knowledge, there is no study that explicitly investigated the Twitter usage of online retailers and compares two different samples against each other. This paper examines the following research questions:

1) How many retailers in Germany are using Twitter compared to the UK?
2) Are there differences in how German retailers are using Twitter compared to the UK?

To answer the above questions, an empirical study was conducted. A sample was collected to compare the Twitter usage in the E-commerce in Germany and UK. Question 1 is answered by hand of account related data retrieved from the Twitter profiles. In order to answer question 2, a model is presented to classify a communication strategy based on directed messages and embedded URLs.

This paper is organized as follows: In Section II, the theoretical background is presented. Then, the research design of the empirical study is explained in Section III. After that, the results of the collection and analysis are presented in Section IV and discussed in Section V. Finally, this paper concludes with an outlook for future research.

II. THEORETICAL BACKGROUND

The main research areas addressed in this paper comprise the social web and E-commerce. The combination of social media and E-commerce are often denoted as *Social Commerce* (4)(5). Most of the studies in Social Commerce investigate the customers' perspective to the platform Twitter. This study is focused on the retailers' perspective instead.

A. Conventions

The publish/subscribe capability is Twitter's fundamental pattern. The different possibilities for routing a status message to other users are summarized in Fig 1. Users subscribe either to other users or add another user to a list. Every message issued gets by default submitted to the following users and to all lists. To broaden the audience, one can embed *Hashtags (HTs)*. The use of HTs is a communication convention that enables authors to post a message either to a community or to add a content information (6). Besides that, messages can be directed to a single user by annotating the user's name with an @-sign, which is called *User Mention*. Retweeting means the redistribution of a Tweet to a user's own audience (7). From the Twitter API's view, Hashtags, User Mentions and URLs are treated as special entities. The use of these entities are examined by (2) in a large scale study.

 ISBN: 978-1-61208-361-2

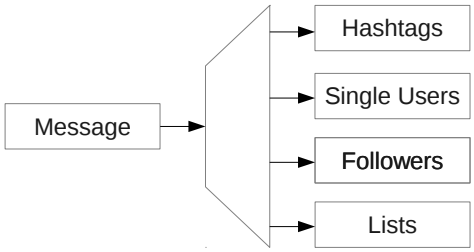


Figure 1: Routing of status messages on Twitter.

B. Purposes

Previous works have investigated that Twitter is used for various communication purposes (8)(6). Twitter entries can be classified in five different genres (9)(10): personal updates, directed dialog, real-time sharing of news and links, business broadcasting, and information seeking. For the study of E-commerce related communication, the broadcasting behaviour and dialogs are interesting. The success of broadcasting activities can be measured in the size of audience, redistribution rate, and the ability to engage users in a dialog (11)(12)(13). Dialogs can be identified by the use of the @-sign, which is often used to direct a message to the user mentioned. Though, this is no exact measure, the vast majority of @-signs are used for UMs in Twitter (14) and, thus, can be used for indicating the interactivity of an online relationship (15, p. 21 - 25).

Related to the E-commerce, by means of an explorative observation of 200 Tweets from 10 online-retailers domain-specific communication practices could be discovered:

- *News:* Published news that are related to the business domain of the retailer, without any direct relation to a product or service offered by the company.
- *Review:* Messages that contain links to a review about a procuct or service offered by the company. The status can be issued by the retailer or by a customer an republished (retweeted) by the company.
- *Promotion:* Tweets that promote products or services directly. Those kind of messages often contain an URL linking to the product page and the respective price of the offer.
- *Dialog:* Directed messages flow betweet the retailer and a customer. The purpose of such dialogs vary: there are examples, where customers request further information about a product or date of delivery via Twitter.
- *Agent:* Retailers mark their messages with an attribute representing the issuing agent personally. This is a communication convention, which is not explicitely supported by Twitter, but used to facilitate directing a message to a certain agent acting on behalf of a company's account. Mainly, the circumflex notation (^Name) is used to annotate an agent.

C. Linking

Another interesting aspect about the communication on Twitter are URLs that are embedded in Tweets. URLs are by default shortened by the platform's own shortening service http://t.co (16). The targeted URLs can be categorized as self-links, social media links and other external links. For the study presented, the former two contain interesting information. A *self-link* points to the own website of an online retailer and indicates the promotion of a product. *Social media* URLs direct users to discussions on other social networks, e.g., Facebook. This indicates a more community-centric activity than links to product pages.

III. Study Design

In this section, the design of the empirical study will be presented.

A. Data Collection

To acquire a sample of E-commerce related communications, two lists of the 115 best-selling online-retailers in Germany (17) and the 100 best-selling shops in UK (18) were used. The selection of retailers for further analysis was performed in four consecutive steps:

1) *Find account:* Twitter accounts were matched to the shop sites by querying search engines. In case that the query did not return a valid result, the search process was continued by examining the shopping site manually.
2) *Targeting specific country:* With this step, it was checked, if a given Twitter account is really targeting the respective country (Germany of UK). For this purpose, a manual examination of each profile's description and timeline was performed.
3) *Retrieval of account data:* The profile information was collected, including the accounts' lifetime, the number of connections to other accounts (followers, friends, listed) and the number of statuses issued since creation.
4) *Activity check:* In order to filter out inactive accouts, the last status, issued by the regarding account was retrieved. Then, the time elapsed between the publication of the last Tweet and the retrieval was calculated. If the last status was issued more than 30 days ago, an account was considered as not being active anymore.

After the selection and retrieval of profile information, the timeline consisting of the last 100 status messages were collected for each account passing the preceding process. Twitter's REST API was queried to retrieve the data set. The data collection was performed on 14th of February, 2014 for the German subset and on 25th of February for UK.

B. Account data

First, the profile information was analyzed. The lifetime of an account in days is defined as L. In order to calculate the Tweet rate R_T per lifetime as an indicator for broadcasting activity, the total number of Tweets since the creation of the account (T) was related to L (see (1)).

 ISBN: 978-1-61208-361-2

$$R_T = \frac{T}{L} \quad (1)$$

In order to analyze the links to other users, the number of followers f_{in}, the number of friends f_{out}, and the listed count l_{in} can be used. As stated in Section II, those values are considered as in- and outdegree measures. To reflect the lifetime of an account, those values were related to L and, thus, they define the indregree rate R_{in} and R_{out} in (2) and (3).

$$R_{in} = \frac{f_{in} + l_{in}}{L} \quad (2)$$

$$R_{out} = \frac{f_{out}}{L} \quad (3)$$

C. Interaction strategies

First, different strategical categories were defined for the study. As described in Section II-B, there are different purposes of Twitter communication. One purpose is the dialog with users, which is called *interactive strategies* subsequently. Two major forms of such strategies can be distinguished: S_1 for communicating inside Twitter and S_2 for distributing links pointing to social networks for dialogs outside Twitter. While S_1 and S_2 were used, when the corresponding attribute dominates, a third form S_3 is introduced for cases, where both attributes dominate only when combined. Thus, S_3 describes an interactive focus, combining both Twitter dialogs and linking to other Web 2.0 sites. A promotional strategy S_4 is indicated by the use of URLs pointing to the shop, owned by the issueing account. S_x is assigned, when none of those attributes dominates in such a manner that one of the other strategies could be assigned.

To determine the Twitter interaction strategies of the conducted accounts, the following definitions are introduced:

- A: All Twitter accounts whereas each element a represents an online retailer from the sample.
- P_M: The fraction of Tweets that address other Twitter users relative to.
- P_S: The fraction of Tweets that contain at least one URL linking to another social network.
- P_P: The fraction of Tweets that contain at least one URL pointing to the online store of the issuer.

A conceptual model was derived, which is based on the Tweets that contain User Mentions (P_M), URLs to social networks (P_S) and URLs to the online store of the account owner (P_P). As shown in Figure 2, we identified four different strategies, which will be explained below.

The first strategy S_1 is characterized by a frequent communication with other Twitter users. Accounts that apply this strategy make use of User Mentions in at least two-thirds of their Tweets.

Strategy S_2 is applied by accounts that intend to direct Twitter users either to the weblog of the company or to a website of another social network (e.g., Facebook) to continue communication there. Accounts were assigned with this strategy when at least two-thirds of their Tweets contain URLs to other social networks or to a company weblog.

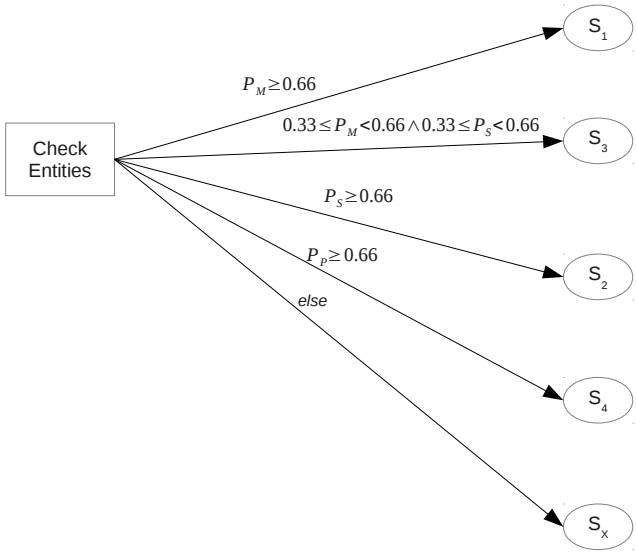


Figure 2: Decision tree for assigning strategical category.

Strategy S_3 is categorized by accounts that use strategy S_1 and S_2 moderately, i.e., accounts that make moderate use of User Mentions and moderate use of URLs to other social networks or weblogs. „Moderate use“ means that User Mentions and social network URLs are present in at least one third, but no more than two-thirds of the Tweets.

The last strategy S_4 is based on the URLs in Tweets that refer to the website of the online retailer. An account is using this strategy, when at least two-thirds of the Tweets contain URLs to the retailers online store, i.e., URLs for promotional purposes.

Finally, all accounts that did not fit into the four above strategies were classified as having no clear strategy (S_X).

IV. Results

In this section the results of the study are presented. First, an analysis of the profile information is given, and, second, accounts are assigned with interaction strategies.

A. Analysis of accounts

For each step of the selection and retrieval process described in III-A, the resulting population size of retailers passing the respective step is examined. The passing of this process is summarized in Table I and divided into the two samples Germany and UK. The value $\frac{x_i}{n}$ is the percentage of the number of accounts passing the step i relatively to the sample's n. The fraction $\frac{x_i}{x_{i-1}}$ represents the number of accounts passing the step i relatively to the previous step $i-1$.

It can be noted that both subsets differ from each other strongly in steps 1 and 4. While the corresponding Twitter

TABLE I: RETRIEVAL OF TWITTER ACCOUNTS.

	DE (n = 115)			UK (n = 100)		
Step i	x_i	$\frac{x_i}{n}$	$\frac{x_i}{x_{i-1}}$	x_i	$\frac{x_i}{n}$	$\frac{x_i}{x_{i-1}}$
1. Find account	88	77%	77%	95	95%	95%
2. Targeting	77	70%	88%	77	77%	81%
3. Retrieving	75	65%	97%	73	73%	95%
4. Active Accounts	58	50%	77%	70	70%	96%

account for 95% of the retailers could be identified for the UK subset, only 77% of the account could be matched for the German subset. While passing 96% of the incoming accounts the activity check for the UK set, only 77% are considered as being active in the German sample.

The variables lifetime, status per day, indegree and outdegree derived from the profile information are summarized in Table II both for Germany and UK. For those values, the .25, .50 and .75 quantiles and Geary's skewness indicator were calculated. For a easier comparison of the location parameters of these values, boxplots visualizing the differences of both samples are depiced in Figure 3. For the sake of readability, the outliers on the right side were cut out, mainly affecting the UK plots.

The German sample is characterized as follows: The mean account lifetime is slightly above 4 years, while the values are slightly left-skewed. Since its creation, an account issued on average about seven Tweets per day, whereas the data is strongly right-skewed. The majority of accounts issued less than two Tweets per day. The mean increase of indegree is about 1.6 users per day, while the data is right-skewed. For 75% of the retailers, this value is at about 1.5 or less. The outdegree value for 75% of the accounts is at about 0.6 or less.

The UK sample is characterized as follows: The mean account lifetime is $4^1/_2$ years, and the values are slightly left-skewed. Since creation, an account issued on average about 17 statuses per day, whereas the values are right-skewed. The mean increase of indegree is at about 33 users per day, while the values are right-skewed. The average increase of outdegree is slightly above two user per day, which is almost identical to the third quantile.

B. Twitter interaction strategies

In order to analyze interaction strategies, each URL was resolved and classified by the hostname in one of the categories self-link, social media and other for both data sets. The German sample consisted of 5792 Tweets (99.86 per account) with 4661 URLs (79%). 4479 of those links could be resolved (96%) and those pointed to 411 unique hosts. The UK sample consisted of 6997 Tweets (99.96 per account) that contained 2726 URLs (39%), therefrom 2546 (93%) could be resolved. Those URLs were pointing to 431 unique hosts. Both sets differ notably in the overall URL usage in microblogging (UK: 39%, Germany: 80%) which indicates that in the German sample, Twitter is dominantly used to direct users to other content located on the web, whereas among the retailers in the UK, Tweets are more self-contained.

TABLE III: COMPARISON OF TWITTER INTERACTION STRATEGIES.

	Strategy	Occurrences DE	UK
S_1	Interactive (Twitter)	19%	80%
S_2	Interactive (other platforms)	24%	0%
S_3	Interactive (mixed)	5%	0%
S_4	Promotional	21%	6%
S_x	No clear strategy	31%	14%

Table III shows the comparison of strategies for Germany and UK. Strategies as described in Section III-C were assigned to each account. 80% of the retailers in UK use one of the interactive strategies (S_1, S_2, S_3), while only 48% of the German retailers are assigned such a strategy. While in the German subset, the values distribute somehow across those three, in the UK sample, the whole category is concentrated on S_1. The promotional strategy S_4 is much more prevalent in the German subset, as well as S_X.

V. DISCUSSION

In Section IV-A, a comparison of the overall activity by Twitter accounts managed by online retail companies is given. The results of the retrieval process distinguish the two samples from each other in two major aspects: The ratio of assigned Twitter accounts is substantially higher in UK compared to Germany, as well as the ratio of active accounts among all accounts. The account lifetime value is used as an indicator for the adoption behaviour among the retailers. The UK sample of accounts is characterized by a slightly longer account lifetime, which indicates an earlier adoption of this technology. Another interesting measure for the adoption is the frequency of usage. According to the number of status messages issued per day, UK retailers are publishing Tweets more frequently. Besides that, a comparison of common influence measures was performed. The data showed that UK retailers are more successful in generating user followers. Another interesting observation is that the UK retailers are also a more likely to follow other accounts.

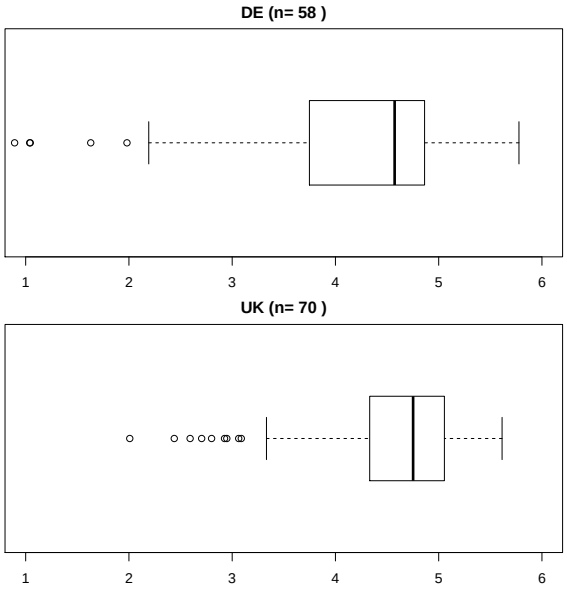
For the identification of different communication strategies, the use of URLs and UMs in Section IV-B was compared. The collected data shows that among the UK sample, interactive strategies – particularly dialogs inside Twitter – are much more prevalent in the German sample. In return, the ratio of promotional strategies is much higher among German retailers. This is also true for the share of retailers, that could not be assigned a strategical category. Although the model defined in Section III allows the occurrence of multiple strategies per account, there were no accounts that actually applied more than one strategy. This shows that the defined strategies are disjunctive and clearly separated from each other.

But the choosen approach has also several limitations: First, the use of UMs and URLs for the indication of the communication purpose. As a previous work showed, the special communication conventions supported by the Twitter platform are not always used correctly (14). Thus, e.g., a locational "@" can be mistaken as indicator for a directed message by the classification approach. Besides that, only the target of an URL points to was examined and not the content

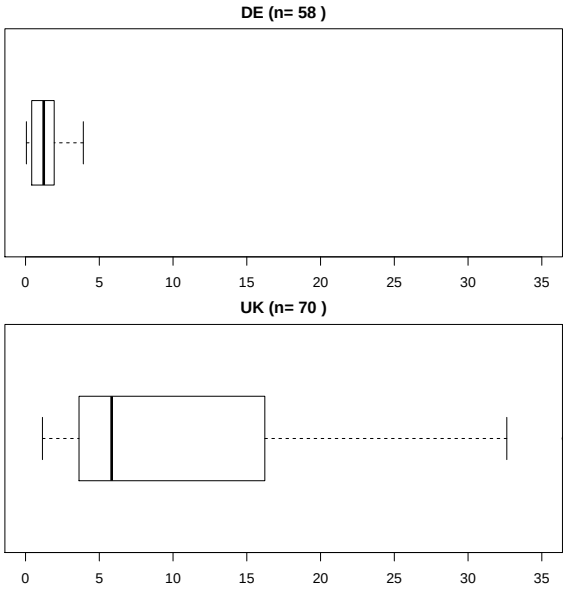
 ISBN: 978-1-61208-361-2

TABLE II: VARIABLES CHARACTERIZING THE RETAILERS' TWITTER PROFILES.

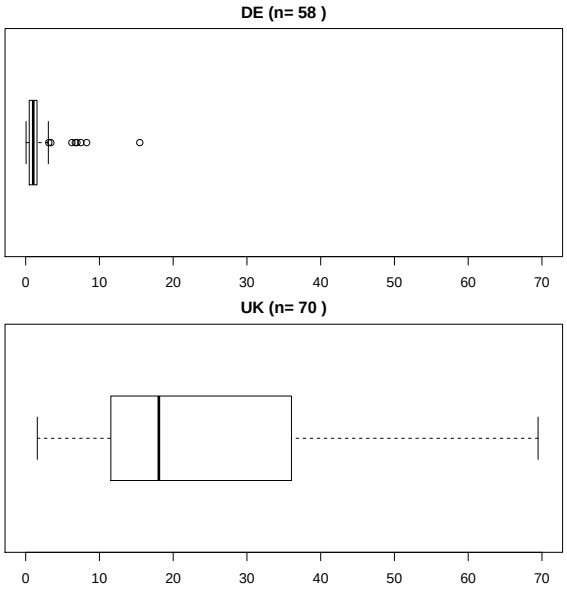
	min		$Q_{.25}$		$Q_{.50}$		$mean$		$Q_{.75}$		max		$skewness$	
	DE	UK	DE	UK	DE	UK	DE	UK	DE	UK	DE	UK	DE	UK
L	0.893	2.008	3.748	4.334	4.563	4.752	4.139	4.508	4.822	5.055	5.777	6.400	-1.350	-1.0412
R_T	0.057	1.145	0.418	3.632	1.208	5.839	7.711	17.840	1.894	15.510	367.300	458.200	7.345	7.338
R_{in}	0.055	1.576	0.482	11.570	0.987	18.060	1.575	32.980	1.471	35.68	8.258	174.800	2.209	2.094
R_{out}	0.004	0.009	0.068	0.333	0.198	0.804	0.620	2.199	0.629	2.094	8.008	32.940	4.314	4.872



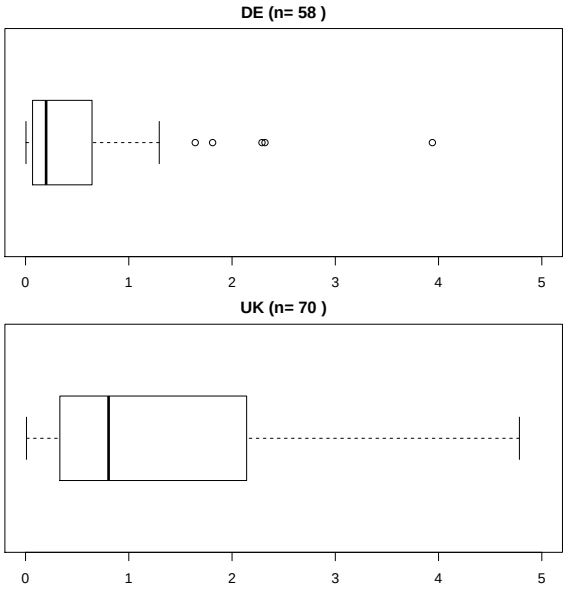
(a) Account lifetime



(b) Status messages per day



(c) Indegree per day



(d) Outdegree per day

Figure 3: Comparison of account life time and number of status per day.

 ISBN: 978-1-61208-361-2

residing at that location. A link to a post on Facebook linking back to the product site of the shop for example would be falsely classified as a "social link".

VI. Conclusion and Future Work

This paper contains contributions to the subjects E-commerce and the social web. The comparison of the usage of communication strategies in the Web 2.0 on hand of the example Twitter shows a different adoption between German online retailers and the UK. For studying the usage of Twitter in general this work provides a model for assigning of four basic communication strategies to an examined account. Besides that, an interesting finding is that there are enormous differences in the communication between retailers and customers. The ratio of dialogs on the Twitter platform is much higher among the UK sample. This could reflect the earlier adoption and higher acceptance of Twitter in the UK as other works suggest.

Since there are some limitations, the refinement and improvement of the approach used in this paper will be the next steps. Besides the sole quantitative analysis of communication patterns presented, it is planned to analyze a subset of the data qualitatively to evaluate the precision of the classification process described in Section III-C. Another limitation of this work lies within the composition of the sample's population, consisting of various different business domains. An examination based on a sample narrowed down to single branch will be carried out.

References

[1] A. Java, T. Finin, X. Song, and B. Tseng, "Why we twitter: Understanding microblogging usage and communities," 2007, uRL: http://aisl.umbc.edu/get/softcopy/id/1073/1073.pdf [retrieved: 24th April, 2014].

[2] L. Hong, G. Convertino, and E. H. Chi, "Language matters in twitter: A large scale study." in ICWSM, 2011, pp. 518–521.

[3] "Why do germans shun twitter?" 2013, uRL: http://www.economist.com/blogs/babbage/2013/12/social-media [retrieved: 6th March, 2014].

[4] M. Bächle, "Economical perspectives on web 2.0 – open innovation, social commerce and enterprise 2.0 (Ökonomische perspektiven des web 2.0 – open innovation, social commerce und enterprise 2.0)," WIRTSCHAFTSINFORMATIK, vol. 50, no. 2, 2008, pp. 129–132. [Online]. Available: http://dx.doi.org/10.1365/s11576-008-0024-2

[5] A. Richter, M. Koch, and J. Krisch, "Social commerce: An analysis of the change in the e-commere (social commerce: eine analyse des wandels im e-commerce)," 2007.

[6] L. Yang, T. Sun, M. Zhang, and Q. Mei, "We know what @you #tag: Does the dual role affect hashtag adoption?" in Proceedings of the 21st International Conference on World Wide Web, ser. WWW '12. New York, NY, USA: ACM, 2012, pp. 261–270, uRL: http://doi.acm.org/10.1145/2187836.2187872 [retrived: 24th April, 2014].

[7] A. Bifet, G. Holmes, B. Pfahringer, and R. Gavalda, "Detecting sentiment change in twitter streaming data," in Workshop on Applications of Pattern Analysis (WAPA) 2011 Proceedings, 2011, pp. 1–15.

[8] H. Kwak, C. Lee, H. Park, and S. Moon, "What is twitter, a social network or a news media?" in Proceedings of the 19th International Conference on World Wide Web, ser. WWW '10. New York, NY, USA: ACM, 2010, pp. 591–600, uRL: http://doi.acm.org/10.1145/1772690.1772751 [retrieved: 24th April, 2014].

[9] S. Westman and L. Freund, "Information interaction in 140 characters or less: Genres on twitter," in Proceedings of the Third Symposium on Information Interaction in Context, ser. IIiX '10. New York, NY, USA: ACM, 2010, pp. 323–328, uRL: http://doi.acm.org/10.1145/1840784.1840833 [retrieved: 24th April 2014].

[10] S. A. Paul, L. Hong, and E. H. Chi, "Is twitter a good place for asking questions? a characterization study." in ICWSM, 2011, pp. 578–581.

[11] B. Krishnamurthy, P. Gill, and M. Arlitt, "A few chirps about twitter," 2008, uRL: http://www2.research.att.com/~bala/papers/twit.pdf [retrieved: 24th April, 2014].

[12] B. A. Huberman, D. M. Romero, and F. Wu, "Social networks that matter: Twitter under the microscope," 2008, uRL: http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2317/2063 [retrieved: 24th April, 2014].

[13] M. Cha, H. Haddadi, F. Bevevenuto, and K. P. Gummadi, "Measuring user influence in twitter: The million follower fallacy," 2010, uRL: http://snap.stanford.edu/class/cs224w-readings/cha10influence.pdf [retrieved: 24th April, 2014].

[14] C. Honeycutt and S. C. Herring, "Beyond microblogging: Conversation and collaboration via twitter," 2009, uRL: http://ella.slis.indiana.edu/~herring/honeycutt.herring.2009.pdf [retrieved: 24th April, 2014].

[15] H. Edman, "Twittering to the top: A content analysis of corporate tweets to measure organization-public relationships," 2010, uRL: http://etd.lsu.edu/docs/available/etd-04292010-162453/unrestricted/edmanthesis.pdf [retrieved: 24th April, 2014].

[16] About Twitter's link service (http://t.co), uRL: https://support.twitter.com/entries/109623# [retrieved: 24th April, 2014].

[17] D. Kailer, P. Mandl, and A. Schill, "An empirical study on the usage of social media in german b2c-online stores," International journal of advanced Information technology, vol. 3, no. 5, 2013, pp. 1–14.

[18] Top 100 online retailers in the UK 2013, uRL: http://www.digitalstrategyconsulting.com/intelligence/2013/06/top_100_online_retailers_in_the_uk_2013.php [retrieved: 24th April, 2014].

 ISBN: 978-1-61208-361-2

A Run-time Life-cycle for Interactive Public Display Applications

Alice Perpétua[1,2]
[1]Faculty of Engineering
University of Porto
Porto, Portugal

ei08060@fe.up.pt

Jorge C. S. Cardoso
[2]CITAR/School of Arts
Portuguese Catholic University
Porto, Portugal

jorgecardoso@ieee.org

Carlos C. Oliveira
Faculty of Engineering
University of Porto
Porto, Portugal

colive@fe.up.pt

***Abstract*—Public display systems are becoming increasingly complex. They are moving from passive closed systems to open interactive systems that are able to accommodate applications from several independent sources. This shift needs to be accompanied by a more flexible and powerful application management. In this paper, we propose a run-time life-cycle model for interactive public display applications that addresses several shortcomings of current display systems. Our model allows applications to load their resources before they are displayed, enables the system to quickly pause and resume applications, provides strategies for applications to terminate gracefully by requesting additional time to finish the presentation of content, allows applications to save their state before being destroyed and gives applications the opportunity to request and relinquish display time.**

Keywords-interactive public displays; run-time life-cycle.

I. INTRODUCTION

In this paper, we propose a run-time life-cycle model for interactive public display applications. This model allows both the display application and the display system to better manage their resources.

The most common and simple approach for content scheduling in public displays is to follow a timetable where each content item is given a pre-determined amount of display time. In this approach, display systems usually have only one active application at a time, using all the display's resources. Applications are simply instantiated and killed by the display system. This approach works well with time-based content where the content's duration is known, such as in videos, or with non-time-based content where the display owner can easily decide how much display time the content should have, as in still images or text.

However, the movement towards open display systems [1] creates a more complex environment where the traditional scheduling approach may compromise the user's experience. In an open network, display owners can easily interconnect their displays and take advantage of various kinds of existing content, including rich interactive applications. Application developers can create applications and distribute them globally, to be used in any display. Users can not only watch the content played on the display, but also appropriate it in various ways such as interacting with it, expressing their preferences, submitting and downloading content from the display.

In this environment, while display owners may still have control over what is displayed, display systems must be prepared to manage an increasing number of applications in a more flexible and unanticipated way. For example, imagine an interactive video application for public displays where users can somehow select videos to play next. Before displaying another application, the display system should make sure the video is allowed to finish, in order not to disturb the viewing experience. Other applications, such as "background" applications, may require display time in response to asynchronous events such as user interactions or other external events. For example, an application may wish to briefly display a calendar notification only when a specific user or group of users, who subscribed to those calendar notifications, are present. In these situations, the currently displayed application that is about to be interrupted should be able to quickly resume operation after the notification. A more detailed analysis of the challenges of content scheduling in open display networks can be found in [2].

This type of environment requires display systems to function more as operating systems, and it also requires a specific application framework that defines a more fine-grained run-time life-cycle. This will allow a better display resource management just like we have in other platforms. For example, the Android platform defines a rich run-time application life-cycle that breaks down all the possible states and transitions between states of an application from the time it is loaded into memory and started, to the time it is shut down and removed from memory. This break down of possible states allows application programmers and system to negotiate the resources that an application needs in each state, guaranteeing an efficient usage of those resources on the one hand, and rapid application switching and loading, on the other hand. For example, an application may be paused if another application comes to the foreground (e.g., because the user requested another application), stopping animations and other CPU consuming operations and save its state to persistent storage (because paused applications may be destroyed by the system if it needs memory). When the application is resumed, it can start the animations again. It is easy to imagine that display systems will need this kind of resource management when the number of applications that each display handles grows.

In this paper, we present our initial effort in this direction. We have looked at existing computing platforms (mobile and desktop) and their typical application run-time life-cycles and synthesized and adapted those models

 ISBN: 978-1-61208-361-2

according to the specific requirements of a public display system. We have also a first implementation of the proposed model as a Google Chrome extension for web-based public display applications.

The rest of this paper is organized as follow. Section II is dedicated to present relevant related work. Section III addresses the observed shortcomings in existing public displays systems and associated design goals for the run-time life-cycle presented in Section V. Section IV summarizes all information gathered about run-time life-cycles of existing computing platforms. Section V describes our proposed life-cycle model, and Section VI concludes.

II. RELATED WORK

Many public display content players / content schedulers have been implemented by researchers and industry.

For example, Linden et al. [3] proposes a web-based framework for managing the screen real estate of the UBI-hotspot system - a public display system that supports concurrent applications on a single display. The framework was implemented using Mozilla Firefox browser and custom JavaScript code that manages the temporal and spatial allocation of the screen to various applications. These hotspots support two modes: a passive broadcast mode, and an interactive mode. These two modes represent different ways for deciding when and which application/content should be loaded by the display system. The framework does not support any type of fine-grained control over the execution of an application. For example, if an application takes a long time to load, the user will be aware of this (at best the application may use a splash screen). Similarly, when unloading, the system simply unloads the content, giving no possibility for the application to run clean-up operations. Even if an application is often used, it will always have to be completely loaded and unloaded every time it is used; the system does not put applications in a suspended state for rapid resuming.

Yarely [4] is a public display player for open pervasive display networks that was developed to replace the existing software infrastructure of the Lancaster e-Campus system [5]. Yarely uses a subscription management system where each display node receives a content descriptor set that lists the content that the player should play and how it should be scheduled. It also supports caching of content items so that displays still function under network failures and disconnections. Even though Yarely is a very powerful software player, even capable of running native content, it is still geared towards passive content that is scheduled consecutively and where the content length can be known *a priori*. Yarely supports dynamic schedule changes that allow it to display unforeseen content such as emergency broadcasts, but it does not provide any specific support for interrupted content to be resumed.

III. EXISTING PROBLEMS AND DESIGN GOALS

Work on interactive public display applications [6][7] has identified a number of shortcomings in existing public display systems. In this section, we present the observed problems and the associated design goal for the run-time life-cycle we propose in this paper.

A. Application loading

Many interactive applications have noticeable loading times that designers usually address by showing a splash screen or loading indicator. Loading times may be, in some cases, avoidable or reduced by leveraging on caching techniques, but they are not generally solvable. Many applications, particularly web-based applications, have to set up communication channels with their own servers and with external services. These initialization processes may be hard to circumvent to give users the impression of instant loading. On public displays these loading times represent wasted resources and reduce the user experience: the time an application takes to load could have been used to display the previous content for a bit more time.

Our goal is to create a display system that efficiently manages the screen in these situations by assigning display time only when the application is ready to display useful content.

B. Graceful termination

Interactive applications have no intrinsic duration that display owners can use when setting up their display's schedule. The result is that applications may be assigned an arbitrary time slot for running. For some applications, this results in a suboptimal user experience because they are sometimes interrupted in the middle of an important operation. The interactive video player application is a paradigmatic example: an application that lets users search/select videos to play next. The public display player may terminate this application before the video finishes, representing an obvious failure for users.

Our goal is to allow applications to, within system-defined bounds, request additional display time to finish an import operation or process. Obviously, these requests may not be honored by the system if another content with higher priority needs display time.

C. Forced unloading, pausing, and resuming

Another issue we noticed in interactive applications was the difficulty of running proper finishing processes before the application is terminated. Usually, applications are simply unloaded from the browser component without warning. This results in added difficulty for the application to save state and terminate connections in a proper manner. Although standard web events could be used in this case, they would still be very dependent on the concrete implementation of the player (some players assign browser tabs to applications, others reuse a single tab). Additionally, in some situations it is more efficient to pause and resume an application instead of unloading and reloading it again in the future. For example, if an alert must be displayed, the interrupted application probably does not need to be unloaded, but simply taken to a paused state where it stops most activity, until the alert is removed from the display.

Our goal is to support application termination, pausing, and resuming. The system should allow applications to

 ISBN: 978-1-61208-361-2

terminate properly if the application is to be killed. Additionally, applications should be able to quickly resume operation if they are interrupted by the system, without having to be completely loaded again.

D. Application-requested loading and unloading

Another problem faced by interactive applications for public displays is that they usually have no way to request display time by themselves, or to relinquish the display if they have no possibility to continue. Although some public display players do allow unanticipated content to be displayed, this usually requires manual intervention. Ideally, applications should be able to request display time in order to display short-term notifications, for example. Conversely, applications that find themselves in a situation where they can no longer continue to execute (e.g., because a fundamental resource could not be loaded) should be able to inform the display system and relinquish the display. Obviously, this requires additional management policies on the display system to guarantee that applications do not misbehave and take over the display.

Our goal is to support this kind of operation, allowing display applications to request display time for short periods, and to give up the display time if they are unable to continue operating.

IV. ANALYSIS OF EXISTING PLATFORMS

The main objective of this paper is to describe our initial model for a run-time life-cycle for public display applications. To arrive at this model, we have looked at existing computing platforms in order to learn about the existing run-time life-cycles. We then synthesized these models and adapted the result to take into account our design goals.

We have analyzed the Android platform, iOS, Windows Phone, Windows 8, and Applets platforms. The main event callbacks associated with each platform are presented in Table 1. Each platform has different ways to manage applications and give applications different levels of granularity for managing their resources. However, we can identify commons categories of application states/event callbacks:

Initializing refers to callback methods that are invoked only once by the system, while the application is in memory. All initial routines related to the user interface or data should be done here.

Starting/Resuming refers to callback methods that are called before the application is put into the foreground, either for the first time, or because the user is resuming the application. Different platforms handle this process differently, but in general these callbacks allow applications to start graphical animations, sounds, and other quick initializations. These callbacks may be invoked several times during the lifetime of the application in memory.

Pausing refers to callbacks that signal the application that it is being interrupted and is being taken out of the display, at least partially. In these cases, applications should stop animations, sound, and other CPU intensive operations.

Stopping/Destroying refers to callbacks that signal the application to stop executing, unload all unnecessary resources, and perform state saving routines. Stopped applications may not be immediately removed from memory, but are good candidates to be destroyed and removed from memory if the system needs the resources.

V. RUN-TIME LIFE-CYCLE

The model for a run-time life-cycle for public display applications is presented graphically in Fig. 1, and described next.

onCreate() – This represents the application's entry point method and is called only once while the application is in memory. Depending on the implementation, it is possible that application code may execute before this method is called. In our Javascript implementation for example, we cannot prevent applications from executing before the onCreate() method is invoked. However, only after onCreate() can an application interact with the display system and it should not be assumed that the display system is ready before the onCreate() is called.

onLoad() – The onLoad() method is called when the display system decides to give display time to the application. Before the display time is actually assigned to the application, the system calls onLoad() and expects applications to reply with a loaded() method call. At the onLoad() stage, applications should perform all necessary loading routines to ensure the application is ready to be displayed.

onResume() – this callback is called immediately before the application is put visible on the display. At this phase, applications should make sure they are ready to show content. This callback can be used to perform very fast initialization routines such as starting animations. When this

TABLE I. SUMMARY OF ANALYSED PLATFORMS

Callbacks categories	**Platforms**					
	Android	***Android services***	***iOS***	***Windows Phone***	***Windows 8***	***Applets***
Initializing	onCreate()	onCreate()	WillFinishLaunchingWithOptions() DidFinishLaunchingWithOptions()	Launching()	onLaunched()	Init()
Starting/Resuming	onStart() onResume() onRestart()	onStartCommand() onBind()	DidBecomeActive()	Activated()	Activating() Resuming()	Start()
Pausing	onPause()		WillResignActive() WillEnterForeground()	Deactivated()	VisibilityChanged() Suspending()	
Stopping/Destroying	onStop() onDestroy()	onUnbind() onDestroy()	DidEnterBackground() WillTerminate()	Close()		Stop() Destroy()

 ISBN: 978-1-61208-361-2

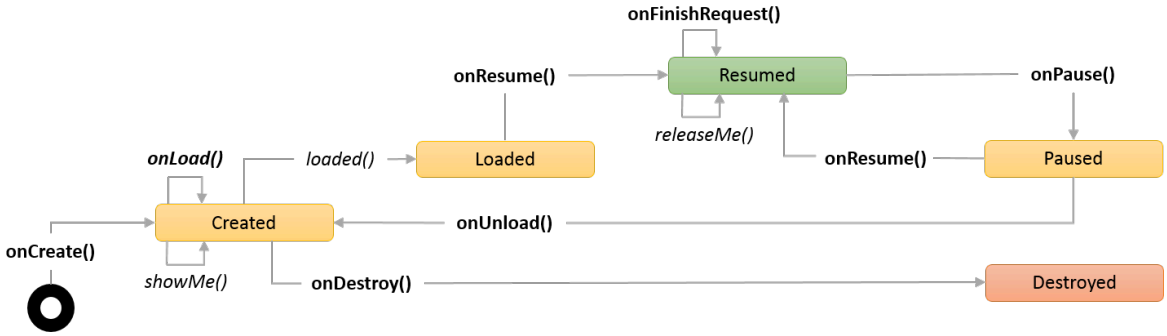


Figure 1. Application lifecyle for public displays.

method is called there should be no noticeable delay before content is displayed by the application.

onFinishRequest() – this callback signals the application that it should finish. In this stage applications should notify the system about how much more time they need to finish gracefully. The system will honor the application's time request, within pre-defined limits, and call onPause() when the time required by the application expires. This callback may not be invoked if the system has another urgent content to display, in which case the onPause() callback will be used immediately.

onPause() – called to signal that the application should pause animations, sounds and other unnecessary operations. In this stage the application is either not visible or only partially visible. Paused applications may be resumed quickly by the system by invoking the onResume() callback.

onUnload() – when an application is closed, it should release all processing resources and clean navigation data as well as state information;

onDestroy() – signals the application that it is being removed from memory. Applications should perform any finalization routines here, perhaps saving state to persistent storage either locally or remotely.

showMe() – Applications can signal the system that they want display time by calling the showMe() method. The system will then apply its internal policy to determine if and when the application should be given display time.

releaseMe() – Conversely, applications can signal the system that they cannot display any more content (perhaps due to a server error or other condition). The system will then take the necessary steps to bring another application to the display.

VI. CONCLUSIONS

We have presented a run-time life-cycle model for public display applications that allows a better resource management for display systems that have to handle a high number of independent applications. The model allows applications to load their resources before they are displayed, system to, allows applications to terminate gracefully, allows rapid pausing and resuming, and allows applications to request and relinquish display time.

We have started to implement this model as a Google Chrome Extension where each application is assigned a browser tab. Our implementation manages the life-cycle of each application determining which tab should be displayed at any time. We support two types of applications: foreground and background applications. The display owner schedules foreground applications, to be shown for pre-defined periods of time. Background applications are loaded at startup by the system, but are only assigned display time when they request it. Our system will apply a priorities scheme to determine which applications can interrupt which applications. It will also manage the system memory resource by dynamically destroying and creating applications based on their memory footprints and usage pattern.

ACKNOWLEDGEMENTS

This paper was financially supported by the Foundation for Science and Technology — FCT — in the scope of project PEst-OE/EAT/UI0622/2014.

REFERENCES

[1] N. Davies, M. Langheinrich, R. Jose, and A. Schmidt, "Open Display Networks: A Communications Medium for the 21st Century," Computer (Long. Beach. Calif)., vol. 45, no. 5, pp. 58–64, May 2012.

[2] I. Elhart, M. Langheinrich, N. Davies, and R. José, "Key Challenges in Application and Content Scheduling for Open Pervasive Display Networks," in Work in Progress Session PerCom 13, 2013, pp. 393-396.

[3] T. Linden, T. Heikkinen, T. Ojala, H. Kukka, and M. Jurmu, "Web-based framework for spatiotemporal screen real estate management of interactive public displays," in Proceedings of the 19th international conference on World wide web - WWW '10, 2010, p. 1277-1280.

[4] S. Clinch, N. Davies, A. Friday, and G. Clinch, "Yarely: a software player for open pervasive display networks," pp. 25–30, Jun. 2013.

[5] O. Storz, A. Friday, and N. Davies, "Supporting content scheduling on situated public displays," Comput. Graph., vol. 30, no. 5, pp. 681–691, 2006.

[6] J. C. S. Cardoso and R. José, "Evaluation of a programming toolkit for interactive public display applications," in Proceedings of the 12th International Conference on Mobile and Ubiquitous Multimedia - MUM '13, 2013, pp. 1–10.

[7] J. C. S. Cardoso and R. José, "PuReWidgets: a programming toolkit for interactive public display applications," in Proceedings of the 4th ACM SIGCHI symposium on Engineering interactive computing systems - EICS '12, 2012, p. 51-60.

 ISBN: 978-1-61208-361-2

Driving the Learning of a Web Aapplication Framework by Using Separation of Concerns

Daniel Correa Botero, Fernando Arango Isaza, Carlos Mario Zapata Jaramillo
Universidad Nacional de Colombia
Medellín, Colombia
Emails: {dcorreab, farango, cmzapata}@unal.edu.co

***Abstract*— Web Applications Frameworks (WAFs) have become very popular tools for developing software applications. These tools lead to the implementation of a big amount of classes, components, and libraries which support developers for saving costs, time, and effort. Due to the big number of WAF elements, a developer needs to invest considerable effort and time in order to understand the WAF usage. Some authors had proposed different framework learning techniques, but these techniques focus on how to document or show the framework information. Then, how to drive the framework learning is a developer concern. Commonly, developers follow a guide containing too much information, but in some cases developers only need to learn an incomplete WAF usage. After analyzing some software projects, we define in this paper a list of web application concerns. This list is connected to a list of WAF components, indicating for each concern the specific elements a developer should know for understanding and covering the concern. Such a list helps the developer to drive the WAF learning. We also develop a web application for driving the WAF learning and an example with a real case of driving WAF learning.**

Keywords-Framework learning; WAF; concerns; framework comprehension; WAF components.

I. INTRODUCTION

Developing web systems is a complex, time-consuming, and expensive task that often requires the coordination of efforts across organizational and technical boundaries [1]. Web Applications Frameworks (WAFs) provide different elements and components to develop effective web systems. They are powerful techniques for large-scale reuse promoting developers to improve quality and save costs and time [2][3]. Since WAFs are considered crucial for rapid web development [4], several frameworks are available, and this topic is being included in research and development [14][23][24].

Developers usually face the need for developing an application by using a specific WAF (perhaps an unknown one); consequently, they need to learn how to use the WAF for developing the application. Several authors have proposed different framework documentation techniques such as: patterns [5], example-based learning [6], cookbooks [7], and visualizations [8]. Some of these techniques are adapted to WAF development. However, they only focus on how to document or show the framework information (architecture, components, classes, relationships, libraries, etc). Currently, when a developer has to use a specific WAF, he/she has to invest considerable effort on understanding it [9]. This problem is due to the big amount of WAF components and the increasing number of documents. Sometimes, developers need to face the reading of hundreds of documentation pages with information they never going to use.

However, in most cases developer learning is primarily influenced by the specific requirements of the application he/she wants to develop. So, developers only need to be concerned on the WAF elements needed to fulfill those requirements. Then, how to drive the WAF learning to be focused on those concerns is an important issue.

In the software development context, a concern is a particular goal, concept, or area of interest [10]. Based on this perspective, we have faced the driving WAF learning by using a separation of concerns. In this approach, each concern represents an application feature supporting a kind of application requirements. For example: authorization, data storage, internationalization and client-side validation are different types of concerns supporting different kinds of application requirements.

Separation of Concerns (SoC) has been used in multiples software areas during the last years, e.g., requirements specifications [11], framework architectures [1], and aspect-oriented programming [12]. SoC is a basic principle of software engineering. Derived from common sense, SoC essentially means that dealing successfully with complex problems is only possible by dividing the complexity into sub-problems which can be handled and solved separately from each other [13].

We use these separation of concerns connected to WAF components [14], giving a specific structure of the elements that a developer should learn for supporting the application requirements. By following such ideas, we develop a new WAF learning technique in which a developer only needs to select the concerns related to his/her development or project, and it will show to him/her the specific components and documentation related. This technique helps developers to save time and to focus on what really they need to learn.

In this paper, we propose a list of 29 basic concerns, and a connection between the list of the concerns and a list of WAF components. Next, we develop a representation of the driving the WAF learning. After that, we develop a simple web application for driving WAF learning, and finally we

 ISBN: 978-1-61208-361-2

develop an example with a real case of driving *Codeigniter* learning.

II. FRAMEWORK UNDERSTANDING

Over the past decade, several documentation techniques have been proposed to support the framework understanding such as: patterns, example-based learning, and cookbooks, among others. However, such techniques are still immature and unused for developing software [2].

Shull et al. [6] show an evaluation of the role examples play in framework reuse. As the main hypotheses, they propose example-based techniques as appropriate to be used by beginning learners instead of hierarchy-based techniques because the latter have a larger learning curve. However, the case study they use is based on a specific example with no patterns, preventing reuse for other frameworks.

Flores and Aguiar [9] present some pattern-based proven solutions to recurrent problems for framework understanding. However, such solutions are top-level basic suggestions.

Jackson et al. [8] support the programmers in understanding the framework code by providing animated visualizations of example programs interacting with the framework. However, a comparison with other methods is not provided.

Cookbooks are commonly used as a documentation technique for web-based framework development. Cookbooks are designed to be carefully read by programmers as reference manuals. Cookbooks also describe the entire framework composition. However, they provide too much information, so they slow the framework learning process.

Most of the aforementioned studies are only focused on documenting the framework information—architecture, components, classes, relationships, libraries, etc.—instead of addressing the framework learning for developers.

Flores [25] presents an approach to guide the framework learning process. His study presents DRIVER, a platform to teach how to use a framework in a collaborative environment. In such platform, learners can search and rate available knowledge and get recommendations for the best course of action. In this approach, learners should decide by themselves—with no guidance based on their needs—on the way they want to follow the documents. Besides, DRIVER is still under development and improvement.

In conclusion, several framework documentation techniques have been proposed, but how to address the framework learning is still a developer task. Besides, these techniques are applied to general frameworks, so WAFs are still underspecified.

III. CONCERN LIST

Developers use WAFs for different reasons: developing a software project, acquiring more knowledge, applying for a job position, accessing the training about tools in organizations, etc. No matter the reason, the final goal for learning a WAF usage is to develop specific web applications.

These specific web applications could be very different from one to another. For example:

- Developer A could be requested to develop a complex Customer relationship management (CRM) system.
- Developer B could be requested to develop a simple static website.
- Developer C has to develop a simple under-construction home page.

In the first case, CRM system involves a lot of requirements, more than the other applications. It means developer A should learn and read more information than the other developers. We could also recognize application B maybe involves less data persistence and less database effort, and maybe application C only involves displaying information on screen (i.e., developer C is focused on a very specific concern). In other words, different developers are driven by different interests or concerns.

In the software development context, a concern is a particular goal, concept, or area of interest. For example, the core requirements of a library borrow card processing system is related to processing book transactions; while its system level concerns would be handle logging, transaction integrity, authentication, security, performance, etc. [10].

Some authors have defined different concern lists or methods to define concerns [1][11][15][16], but in most cases the definition of these concerns is delegated to an analyst. In other cases, the concern list is just a list of non-functional requirements or a list of ambiguous elements like: immunity, integrity, precision, robustness, among others.

However, these concern lists are very general and are difficult to adapt to the specific WAF components and elements that a developer should learn. So, based on the idea of driving WAF learning through a concern list, we developed a new web application concern list. In order to develop this list, we analyzed more than 20 web projects that were develop by computer science students in a course during the last 2 years.

These projects are based on real industry needs. We found similarities among each project requirements and we grouped them in a concern list. In this analysis we registered how many projects required a specific concern. Also, this analysis shows that no matter how different seems each application from one another, they use similar concerns. After this process, we define in Table I, 29 concerns and we categorize them in different groups.

This concerns list should be used by a developer. At the beginning a developer has to recognize the specific requirements for the project he/she is working on. After that, he/she has to carefully read each concern and its specific description. Finally, he has to select the concerns which are involved in his/her project requirements.

Later on, each concern will be connected to the specific components or elements of a WAF. This generates a personalized learning guide.

 ISBN: 978-1-61208-361-2

TABLE I. LIST OF WEB APPLICATION CONCERNS

#	**Concern (Times of appearance on projects)**	***Category***	***We suggest to select this concern if:***
1	Display information on screen (20)	User Interface	You have to display information on a screen.
2	Stylized screens (20)	User Interface	Your screens have to be edited and stylized usually through a CSS file. Sometimes WAFs are based on prefabricated styles.
3	Tools and accessories for creating views (20)	User Interface	You have to create forms, tables, or other view elements. (Some WAF support to create faster view elements usually using front-end languages like html).
4	Routes and navegability (20)	User Interface	You need to display a screen. Each application section or link has a specific route. These routes and their connections are very different from WAF to WAF.
5	Capture and assign data (20)	User Interface	Your application involves creating forms, to capture data, or to send data from a controller to a view.
6	Client-side data validation (20)	User Interface	You need to do validation in client side like guarantee not empty forms or specific type of data or validations using AJAX. Besides, don't forget to revalidate in server-side.
7	Upload files (13)	Architecture and data flow control	You need to upload files like images, and documents, among others.
8	Error handling (20)	Architecture and data flow control	Your application generates client errors, or database errors, or any kind of errors. It is important to know how to treat them, how to capture them and show them.
9	Internationalization (3)	Architecture and data flow control	Your application requires multiple languages or to have the screens texts centralized (which improves maintainability).
10	Localization (2)	Architecture and data flow control	The information displayed on your application screens depends on user location (e.g., show a specific app to a user on US and another to a user in UK).
11	Caching (3)	Architecture and data flow control	Performance is a very important requirement. Some WAF use caching systems to have pre-storage of the information.
12	Testing (7)	Architecture and data flow control	You need to know how to debug the application information or to apply some test.
13	Portability (7)	Architecture and data flow control	You need to develop a version of your application for desktops and another for mobiles.
14	Data Selection (20)	Data modeling and persistence	You need to extract data from a class model (usually connected to a table of your database).
15	Data Selection with pagination (19)	Data modeling and persistence	You need to extract data by pages from a class model (usually connected to a table of your database).
16	Data selection using filters (20)	Data modeling and persistence	You need to select filtered data (usually using specific searches).
17	Multiple data selection (20)	Data modeling and persistence	You need to extract data from multiple class model (usually connected to various table of your database).
18	Data storage (20)	Data modeling and persistence	You need to save data from a class model (usually save data on your database).
19	Data editing (19)	Data modeling and persistence	You need to edit data from a class model (usually update data your database).
20	Deleting Data (14)	Data modeling and persistence	You need to delete data a class model (usually delete data your database).
21	Creating model functions (20)	Data modeling and persistence	You need to create specific functions for your classes.
22	Model-side data validation (20)	Data modeling and persistence	You need to apply model-side validations.
23	Authentication (20)	Security	You need a login in your application.
24	Authorization (20)	Security	You need to grant access to different areas in your application.
25	Control data in session (20)	Security	You need a login, a shopping cart or other functionality that require control data in session.
26	Server-side data validation (20)	Security	Your application require validate data (usually additional data that data from models).
27	Coupling modules (14)	Modules and extensions	You need to couple a specific module in your application (some WAFs have websites plenty of specific modules like calendars, pdf generation, transformation to csv and much more). You have to search if the module you need is available or you have to develop it.
28	Creating modules (14)	Modules and extensions	You need to create a new module in your application.
29	Auto-generated code (14)	Modules and extensions	Your WAF offers the possibility to auto-generate a CRUD (create-read-update-delete) of a class model.

 ISBN: 978-1-61208-361-2

IV. CONCERN LIST VS COMPONENT LIST

Several WAF comparison studies show many similarities between them [17][18]. In a previous work, we defined WAF components based on these similarities. We divided the learning process for each component in a set of fundamental tasks, each task details very specifically how components are composed. Furthermore these tasks guide developers in what they should learn in order to learn and use each component [14].

In a real application, concerns are traduced and codified into different components. Commonly, concerns are related to aspect-oriented programming and they are codified into aspects [19]. In object-oriented programming concerns could be traduced into classes and/or components. They also could be traduced in several ways: functions and libraries, among others. It depends on the developer techniques, the tools, or the programming architecture.

TABLE II. WAFS CONCERNS LIST VS WAFS COMPONENTS LIST

Component	*Task*	*# of related Concerns*
Superclass model	Identify what functions are available	14, 15, 16, 17, 18, 19, 20, 21
	Identify how to create model classes and what functions should be override	14, 15, 16, 17, 18, 19, 20, 21
	Identify how to create new class functions	21
	Identify how to call attributes and functions classes	14, 15, 16, 17, 18, 19, 20, 21
Superclass Controller	Identify what functions are available	1
	Identify how to create controller classes and what functions should be override	1
	Identify how to call model classes	14, 15, 16, 17, 18, 19, 20, 21
	Identify how to call libraries or plugins	27, 28
	Identify how to call views	1
	Identify how to do redirects	8, 23, 24
	Identify how the variables get, post, session, and files are treated	5, 23, 25
	Identify how to receive and send data to views	5, 7
	Identify how to show results by pages	15
	Identify how to manage different packages of languages	9
	Identify how to show information depending on user's location	10
	Identify how to manage login and logout	23
	Identify how to upload files	7
	Identify how to design an application for desktop and mobile	13
Route Manager	Identify how URLs are and what means each part of the URLs	4
	Identify how to send and receive data from URLs	4
Error Handler	Identify what the sections to catch errors are	8
	Identify what the types of errors are	8
	Identify how to capture and show these errors	8
Database Class	Identify how to connect to a specific database	14, 15, 16, 17, 18, 19, 20
	Identify how to add data to the database	18
	Identify how to delete data from the database	20
	Identify how to edit data from the database	19
	Identify how to filter data	16
	Identify how to select data from the database (even information from various tables)	14, 15, 16, 17
	Identify additional functions or functionalities	--

Component	*Task*	*# of related Concerns*
Template Manager	Identify if a different syntax is used in the view layer and how it works	1
	Identify how the communication between controller and view layers is achieved	1, 5, 7
	Identify what functions are available	1
	Identify how the variables get, post, session, and files are treated	5
	Identify how to create styles (css files) and where are located	2
Role Manager	Identify how to validate permissions in the application	24
	Identify how to grant access to specific areas.	24
	Identify how to add types of roles	24
Data Validation	Identify how validations in control layer are treated	26
	Identify how validations in view layer are treated	6
	Identify how validations in model layer are treated	22
	Identify what kinds of validations are predefined	6
	Identify how to create new validation types	--
Cache	Identify how to call cache	11
	Identify where cache is used	11
Helper	Identify what kinds of helpers exist	3, 27
	Identify what facilities give each helper and how to use them	3, 27
	Identify how to create and connect a new helper or library	28
Tester	Identify how to create unit tests	12
	Identify how to debug information	12
ORM	Identify how the transformation among relational databases and class objects is achieved	14, 15, 16, 17, 18, 19, 20
	Identify how various objects are gathered from different classes	17
	Identify how one-one and many-many relations, among others, are treated	17
Automatic code generator	Identify how to call specific SQL statements	--
	Identify how to call and use auto-code generators.	29
	Identify what information is created and how to edit it	29
	Identify how to delete that information	29

 ISBN: 978-1-61208-361-2

Due to the WAF features and taking advantage of WAF components separations, we connected application concerns to WAF components and their tasks. This connection gives the possibility to know for each concern what are the specific components and tasks related to start the personalized learning process.

Table II exhibits the common connection between the lists. The connection is not an ultimate one; a senior WAF developer could make adjustments as he/she considers. The main idea is each task as a solution for a specific WAF (it could be a link to website, forum or blog; could be a video or a specific explanation text). Later, a real example is developed.

We need to emphasize that one concern could be related to a specific task or multiple tasks, of one or multiple components.

These lists also give a perspective of the components all developers should take advantage of. If a WAFs first-time user read the concern list, he/she could find component for crucial elements unknown to him/her (e.g., internationalization, caching, and portability, among others). This means that if he/she implements these elements at the beginning of the development; the final application would have more quality.

The final step given the learning tasks is to associate the specific learning material for each task in a specific WAF. As these associations are very different for each WAF, and are out of our scope, we suggest this process should be done by a senior WAF developer. In our work we developed an application capable to register these associations.

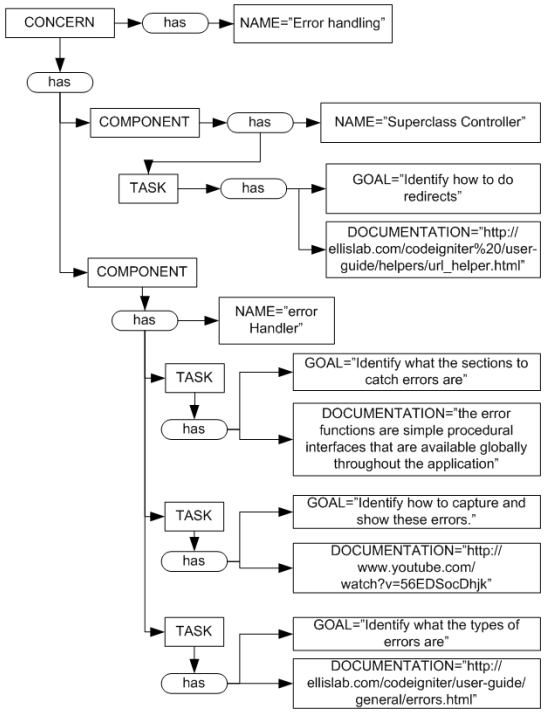


Figure 1. Example of "Error handling" concern, connected to its respectives components and tasks in Codeigniter.

Figure 1 is developed by using an executable pre-conceptual schema [20]. In this figure, we show an example about how concerns, components and learning tasks are connected. If a developer is only interested on capturing and fixing errors, he/she has to analyze and learn tasks documentation. If a developer is interested on the error handling concern, he/she could be also interested on others concerns like: "display information on screen" or maybe "Client-side data validation", which increase the number of components and tasks he/she has to analyze and learn.

In Figure 2, we summarize the driving of the WAF learning process. Developer first step is to choose the specific WAF in which he/she wants to develop the application. The second step is analyzing the application to develop and extract the requirements. Third, he/she has to choose the concerns related to the application that support the previously requirements. Finally, he/she has to work with the specific elements and documentation tasks (previously filled by a senior WAF developer) in order to build the application.

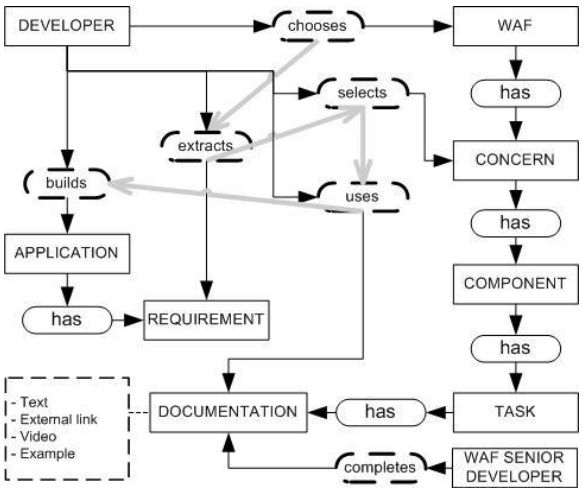


Figure 2. Representing the drive of WAF learning process.

V. CONCERNS SELECTION EXAMPLE

A developer is requested to build an application module by using a new framework. After the requirements elicitation process, a requirements list is presented:

- The application has to extract the real estate information from the main database.
- Only admin users—already created in the database—can access the real estate information. Then, a login system is required.
- Admin can filter real estate information ordered by name, location or type.

We suppose the requested developer should select the concerns listed in Figure 3. Similar to Figure 1, each concern of Figure 3 will be connected to its related components and tasks. Concerns of the Figure 3 support developers as personalized learning guides, i.e., before starting the learning process, developers can discard some documentation unrelated to his/her needs.

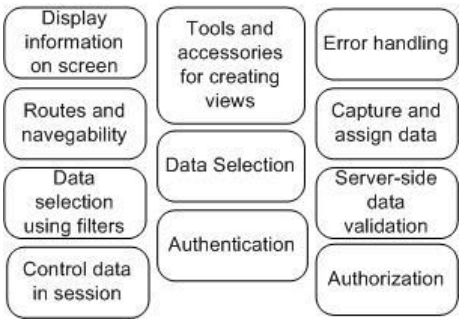


Figure 3. Example of concerns selection.

VI. DL APPLICATION

Tasks documentation is a WAF senior developer job. We developed a driving learning (DL) application [26] with the aim of having this documentation available online and only documenting tasks once in a specific WAF. This is a simple web application which authorizes developers to build a personalized learning guide (see Figure 4).

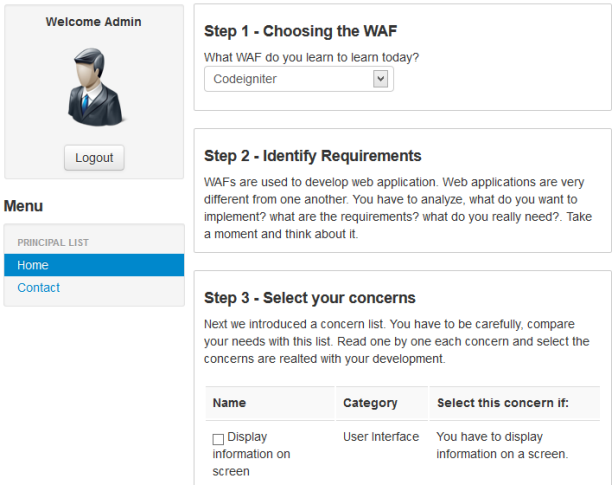


Figure 4. Home of DL application.

Figure 5 shows a real case on *Codeigniter*. The developer complete the previously form and only chose "Error handling" concern. The applications show him/her the solution to this concern with the specific components and tasks he/she has to develop. This guide also allows WAF senior developer to create some notes for each task in order to better complete the information.

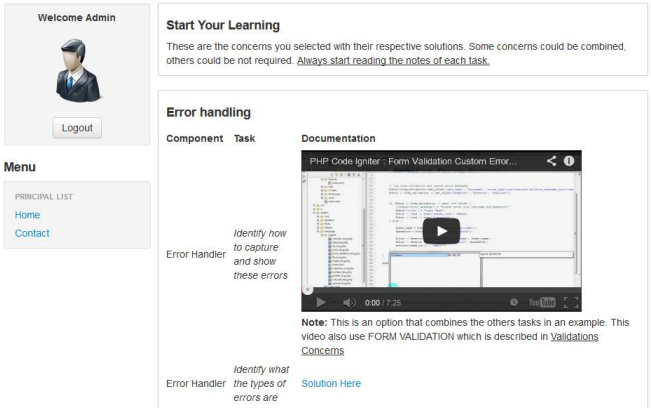


Figure 5. An example of a personalized learning guide on DL application.

Some advantages of this approach are:

- Developers will find a way to guide their learning focusing only in what is concerned to them.
- By learning the basic concerns—first concerns—developers has to understand the framework fundamentals as architecture, folder layout, and basic syntax. E.g., 'display information on screen' concern will give developer the framework fundamental elements.
- Material should be developed by WAF senior developers which guarantee no time wasting on deprecated or wrong internet solutions.
- Future work will connect concerns with a specific example gluing together the components and tasks. As a bonus, exercises provide a source of code reuse —e.g., 'display information on screen' concern connected with 'hello world' example portraits the framework architecture and a base code for all apps—.

VII. CONCLUSIONS

WAF learning is an important issue. Nowadays, WAF learners have to face hundreds of documentation pages and web documents, but they really need to read and follow some parts of the documentation. The main objective of these developers is to build web applications which have different requirements from one to another. By related the documentation to the developer needs, we reduce the amount of documents they have to face, and focus them on what they need. Web application concerns are connected to WAF components giving the possibility to know for each concern what are the specific components and tasks related. This connection is completed by a WAF senior developer; he/she develops all documentation in a specific WAF. Our DL application supports this documentation. In the final step, a WAF learner starts his/her learning process by selecting the concerns related to his/her requirements over DL application and accessing to their personalized learning guide.

 ISBN: 978-1-61208-361-2

VIII. FUTURE WORK

Programmers frequently use a copy-and-paste process to develop their applications [21]. We will improve this learning guide with examples for each concern. A developer has the basic example of each concern and he/she could use it in his/her applications. We will use micro-learning [22] to separate the different steps of WAF learning and finally we will develop a real experimental design to obtain stats and better results.

DL application could also be improved allowing forum discussions and star rating documentation, also increasing the amount of material.

Comparison between different learning techniques like example-based learning, cookbooks, micro-learning and other techniques could be developed.

IX. REFERENCES

[1] X. Kong, L. Liu, and D. Lowe, "Separation of concerns: a web application architecture framework," Journal of digital information, vol. 6, no. 2, 2005, pp. 1-8.

[2] N. Flores and A. Aguiar, "Understanding Frameworks Collaboratively: Tool Requirements," International Journal On Advances in Software, vol. 3(1 and 2), 2010, pp. 114-135.

[3] D. Hou, "Investigating the effects of framework design knowledge in example-based framework learning," Proceedings of 24th IEEE International Conference on Software Maintenance, Beijing, China, 2008, pp. 37-46.

[4] J. An, A. Chaudhuri, and J. S. Foster. "Static typing for Ruby on Rails," Proceedings of 24th IEEE/ACM International Conference on Automated Software Engineering, Auckland, New Zealand, 2009, pp. 590-594.

[5] R. E. Johnson, "Documenting frameworks using patterns," ACM Sigplan Notices, vol. 27, no. 10, pp. 63-76, 1992.

[6] F. Shull, F. Lanubile, and V.R. Basili, "Investigating reading techniques for object-oriented framework learning," IEEE Transactions on Software Engineering, vol. 26(11), pp. 1101-1118, 2000.

[7] G. E. Krasner and S. T. Pope, "A cookbook for using the model-view-controller user interface paradigm in Smalltalk-80," Journal of Object-Oriented Programming, vol. 1(3), 1998, pp. 26–49.

[8] K. Jackson, R. Biddle, and E. Temper, "Understanding frameworks through visualisation," Proceedings of 37th International Conference on Technology of Object-Oriented Languages and Systems, Sydney, Australia, 2000, pp. 304-315.

[9] N. Flores and A. Aguiar, "Patterns for understanding frameworks," Proceedings of 15th Conference on Pattern Languages of Programs (PLoP), Nashville, TN, USA, 2008, pp. 8.

[10] G. Kamble, "Aop-Introduced Crosscutting Concerns," Proceedings of International Symposium on Computing, Communication, and Control (ISCCC), October. 2009, pp. 140-144.

[11] L. Rosenhainer, "Identifying crosscutting concerns in requirements specifications," Proceedings of OOPSLA Early Aspects 2004: Aspect-Oriented Requirements Engineering and Architecture Design Workshop, Vancouver, Canada, October. 2004.

[12] T. Elrad, M. Aksit, G. Kiczales, and K. J. Lieberherr, "Discussing aspects of AOP," Communications of the ACM, vol. 44, no. 10, pp. 33-38, 2001.

[13] D. L. Parnas, "On the Criteria To Be Used in Decomposing Systems into Modules," Communications of the ACM, vol 15, no. 12, pp. 1053–1058, 1972

[14] D. Correa, C. M. Zapata, and F. Arango, "Learning of web application frameworks components," IADIS AC, October. 2013, pp. 155-162.

[15] G. Sousa, S. Soares, P. Borba, and J. Castro, "Separation of crosscutting concerns from requirements to design: Adapting the use case driven approach," EA, pp. 93-102, 2004.

[16] I. S. Brito, F. Vieira, A. Moreira, and R. A. Ribeiro, "Handling conflicts in aspectual requirements compositions," Transactions on aspect-oriented software development III, Springer Berlin Heidelberg, pp. 144-166, 2007.

[17] P. Wang, "Comparison of Four Popular Java Web Framework Implementations: Struts1. X, WebWork2. 2X, Tapestry4, JSF1. 2," Doctoral dissertation, Master's Thesis, University of Tampere, 2008.

[18] M. Canales, "A Comparative Study of Rapid Development Frameworks for the Creation of a Language Placement Exam Template," Doctoral dissertation, Texas A&M University, 2010.

[19] M. Marin, A. van Deursen, L. Moonen, and R. van der Rijst, "An integrated crosscutting concern migration strategy and its semi-automated application to JHotDraw," Automated Software Engineering, vol. 16, no. 2, pp. 323-356, 2009.

[20] C. M. Zapata, G. L. Giraldo, and S. Londoño, "Esquemas preconceptuales ejecutables," Avances en Sistemas e Informática, vol. 8, no. 1, p. 2, 2011.

[21] M. Kim, V. Sazawal, D. Notkin, and G. MurphyKim, "An empirical study of code clone genealogies," ACM SIGSOFT Software Engineering Notes, vol. 30, no. 5, pp. 187-196, 2005.

[22] T. Hug, "Didactics of microlearning: concepts, discourses and examples," Waxmann Verlag GmbH, Germany, 2007.

[23] X. Shi, K. Liu, and Y. Li, "Integrated Architecture for Web Application Development Based on Spring Framework and Activiti Engine," The International Conference on E-Technologies and Business on the Web (EBW2013), The Society of Digital Information and Wireless Communication, May. 2013, pp. 52-56.

[24] J. Weinberger, P. Saxena, D. Akhawe, and M. Finifter, "A systematic analysis of xss sanitization in web application frameworks," Computer Security–ESORICS 2011, Springer Berlin Heidelberg, pp. 150-171, 2011.

[25] N. Flores, "Patterns and Tools for Improving Framework Understanding: a Collaborative Approach," Doctoral dissertation, University of Porto, December 2012.

[26] Driving Learning Application. [Online]. Available from: http://www.frameworkg.com/dl/.

 ISBN: 978-1-61208-361-2

A Method to Achieve Automation in the Development of Web-Based Software Projects

María Consuelo Franky
Department of Systems Engineering
Pontificia Universidad Javeriana
Bogotá, Colombia
lfranky@javeriana.edu.co

Jaime A. Pavlich-Mariscal
Department of Systems Engineering
Pontificia Universidad Javeriana
Bogotá, Colombia
jpavlich@javeriana.edu.co

***Abstract*— This paper proposes a method to achieve a high degree of automation in the development of Web software projects. This method is based on the experience of two consecutive university-industry projects that have received funding from the Colombian government. These projects aim to improve the software development tools of a large-scale software company, applying techniques based on Model Driven Engineering (MDE) and software building tools to achieve a high level of automation in generating new Java Platform, Enterprise Edition (Java EE) projects and in integrating existing components developed by the company. The tools developed in the first project significantly improved the development speed in the company. In the final state of the second (ongoing) project, we expect that MDE transformers will improve flexibility in generating Java EE projects with different architectures and different types of user interfaces, such as JavaServer Faces (JSF) or Java FX2. We believe that the steps performed during those two projects can serve as a guide for other software organizations to effectively automate their development for large scale projects.**

Keywords- Web technologies; Frameworks; Web applications development; Software Reuse; Automatic Software Generation; Model-driven development of Web applications.

I. INTRODUCTION

Competition and market requirements lead to companies developing large software projects to find higher competitiveness through shorter development cycles and lower costs. One way to achieve these goals is through better automation in the development of software projects and also through higher reuse of software components that are useful for multiple projects [1].

This paper proposes a method comprising a series of stages with associated techniques for achieving a high degree of automation and reuse in the development of Web software projects. The steps and techniques described in this paper have been applied to the specific case of Heinsohn Business Technology (HBT) [2], a large-scale Colombian software development company that develops Java EE [3] applications for governmental and financial organizations.

This method was developed as part of two joint projects between the Pontificia Universidad Javeriana [26] (the university of the authors) and HBT, which were funded by the Colombian government. In these projects, we have applied techniques based on MDE [8] and software tools [13] to achieve a high level of automation in generating new Java EE projects and effectively integrating and reusing components developed by the company. As a result, the company has been able to reduce significantly the initial stages of development of Java EE projects.

We are currently working on a second project to further improve these tools and processes. This new project will develop MDE transformers that will increase flexibility in generating Java EE projects with different architectures and different types of user interfaces (JSF [5] or Java FX2[29]).

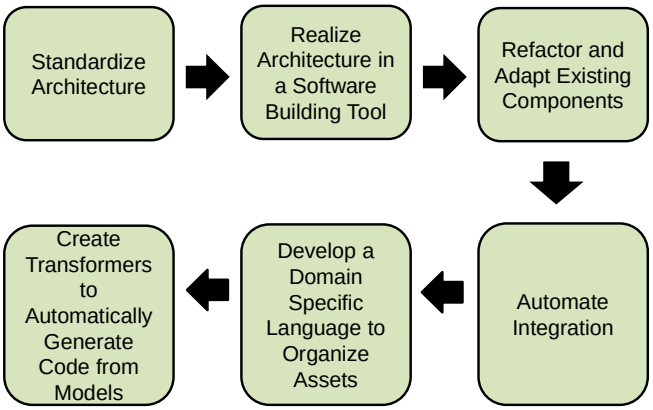


Figure 1. Proposed Method.

Figure 1 is an overview of the proposed method. First, it is necessary to standardize a multilayer architecture for the organization of new projects. Software building tools are used to materialize that architecture when creating the codebase of these projects. It is also necessary to refactor and adapt existing components in the organization, so they can be efficiently integrated into new projects. The next step is to automate the integration of such components into new software projects. A domain specific language (DSL) [27] is developed to create models that effectively reference and organize all of the above assets with the structure and behavior of a web application. Code generators automatically transform those models into working software applications.

The remainder of this paper details the proposed method for achieving high automation in the development of Web software projects in a company. Section II describes the initial stage of defining and standardizing a multilayer

 ISBN: 978-1-61208-361-2

architecture for a company. Section III describes the materialization of such architecture through a software building tool (Maven [4]). Section IV describes the refactor of the company reusable components in order to be compatible with the architecture. Section V describes the automated integration of components in Java EE projects. Section VI describes how to incorporate DSL to allow the modeling of web applications (including reusable components) independently of technology. Section VII describes the construction of MDE transformers in order to automatically generate Java EE projects that integrate the reusable components and with different types of user interfaces (JSF or Java FX2). Section VIII analyzes related work, and Section IX presents the conclusions and future work.

II. STANDARDIZING ARCHITECTURE

The company that participated in our two research projects (HBT) develops large-scale software, with a focus on Java EE. During the development of several software applications, HBT determined the necessity of adopting a standard reference architecture to organize the application code.

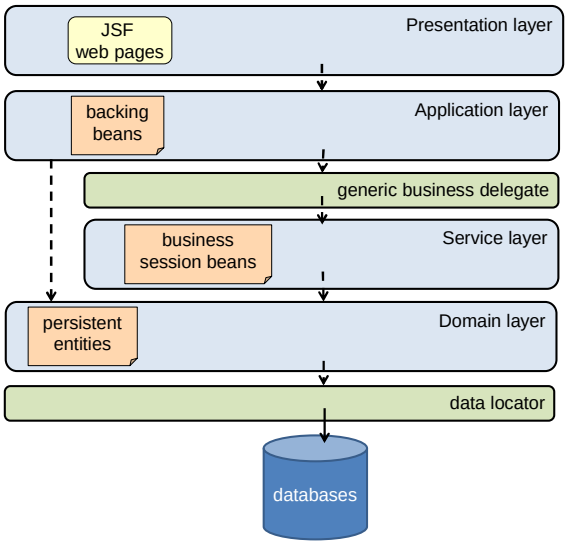


Figure 2. Multilayer Architecture adopted at HBT.

Based on this experience, the first stage of our proposed method is the standardization of the software architecture utilized by the organization in its projects. An important premise for this stage is that the software projects developed by the organization must be of a similar nature and should be effectively addressed by a standard architecture.

Figure 2 depicts the standard architecture adopted by HBT. The architecture effectively separates the application into several decoupled layers. For instance, web pages are supported by backing beans that manage the displayed information. These beans are decoupled from business session beans and entity beans that are in the lower layers.

III. MATERIALIZING THE ARCHITECTURE THROUGH A SOFTWARE BUILDING TOOL

To properly adopt the architecture, it is very important to materialize it in the software building tools that are used to create, integrate, and build software projects. In the case of HBT, since its focus is on Java EE projects, the chosen tool was Maven [4]. Maven is a tool to automate the building lifecycle of a software application, dependency management, and software variants. Maven defines a Project Object Model (POM) file, an Extensible Markup Language (XML) file that stores all of the above information about a software project. A detailed discussion of the reasons for choosing this tool in HBT can be found in [28].

In the context of HBT, Maven was used to materialize the adopted architecture. Each Java EE project is described as a Maven project with the following sub-modules:

- Presentation Layer. JSF [5] web pages realizing CRUD ("Create, Read, Update, and Delete") and business operations.
- Application Layer. Descriptors and backing beans [6] to support the JSF pages.
- Service Layer. Session Beans that realize all of the functionality of use cases.
- Domain Layer. Descriptors and persistent entities [6].
- Persistence Layer. Structured Query Language (SQL) scripts to populate tables with initialization data.

Each of the above sub-modules has a POM file that describes the library dependencies of each sub-module (including dependencies to other sub-modules), the type of artifact that yields after building and packaging (e.g., a Java Archive – JAR – or Web Application Archive – WAR - file [3]), and the identification of each sub-module in a Maven component repository of the organization.

To provide an adequate flexibility in the creation of the codebase of new projects, a useful tool is Maven Archetypes [4], templates based in Maven to instantiate the adopted architecture into new projects that are parameterized by specific design decisions.

Our joint project with HBT created an archetype that includes several profiles that parameterize new projects according to different database engines (Oracle [15], Postgresql [16], MySQL [17], and SQLServer [18]), and different application servers (JBoss [19], Glassfish [20], WebLogic [21], and Websphere [22]).

IV. REFACTORING AND ADAPTATION OF EXISTING COMPONENTS

The next step in the process is to refactor existing software components in the organization, to adapt them to the adopted architecture. Section A explains the existing components at HBT and Section B describes the process to adapt them into the chosen architecture.

 ISBN: 978-1-61208-361-2

A. *Description of company reusable components*

Software development organizations usually create several software components in order to reduce costs and capitalize the knowledge of previous solutions. In the context of HBT, these components address requirements, such as security (originally based on [34]), audit, notifications, batch data processing, text files processing, etc., or improve functionality of existing COTS (commercial off-the-shelf) components [28].

Component reuse may be complex, since a component may span several layers of the architecture, to provide a complete solution to programmers. Components also include several different types of files, each one associated with specific layers in the architecture. For instance, JSF pages in the presentation, session beans for the business logic, persistent entities, SQL scripts to populate databases, etc.

To properly reuse a component in a new project, many of the above elements need to be adapted to the specific project requirements. This task could be made easier by using automation techniques.

B. *Component Refactoring to a Maven Multi-Module Structure*

To properly incorporate existing components into the adopted architecture, it is necessary to convert them to the format of the software build tool. For our project with HBT, the existing component had to be converted to Maven multi-module components [13], to facilitate reuse and integration with the Maven-based architecture. Particularly, the conversion to Maven facilitated the identification of the parts of each component that are associated with each layer of the architecture.

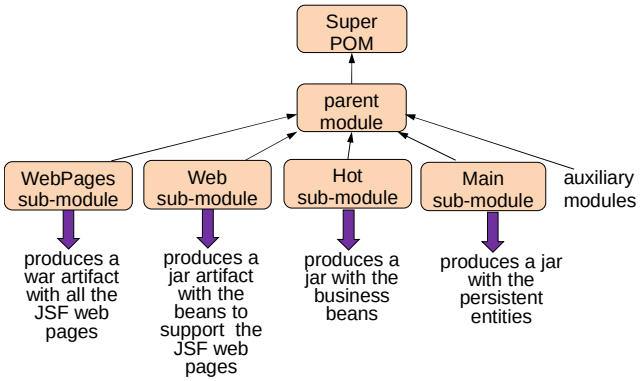


Figure 3. Maven-based Structure of Refactored Components.

Figure 3 depicts the general structure of each refactored component. A refactored component has a root module that contains a sub-module for each layer in the architecture; Figure 3 shows the "Super POM" that is the root module that sets the standards for any Maven hierarchy.

Figure 4 depicts the folder structure of one of the refactored components (in this case, the security component). Each folder corresponds to a Maven sub-module, each one with its own POM file. The pom.xml descriptor at the root folder of the component denotes dependencies on third-party libraries and on other components. These dependencies are inherited by the modules inside the component.

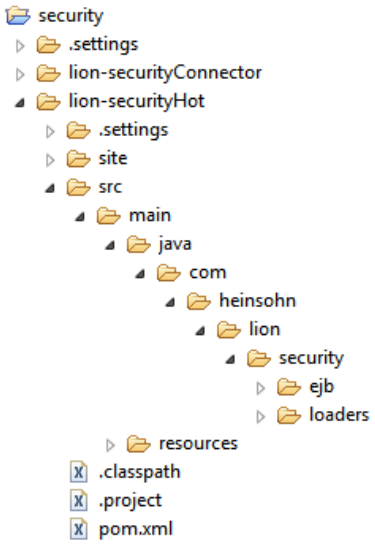


Figure 4. Maven Multi-Module Structure of a Refactored Component.

Each sub-module yields a small artifact that can be indexed and stored in a Maven repository of the company. This facilitates the integration of new multi-module components in the future.

V. AUTOMATING THE INTEGRATION

Although refactoring components to adapt them to the chosen architecture may reduce development times, there are other tasks that can be performed to further improve this process. Particularly, an adequate automation of the component integration process may speed up the creation of a project codebase.

In the context of our projects with HBT, component integration was automated through a tool called LionWizard. This tool automatically generates a new Java EE codebase utilizing a Maven archetype. The tool further transforms that codebase to automatically integrate all of the components selected by the user, effectively eliminating most of the manual tasks.

Figure 5 is the main window of the wizard. This tool receives as input a series of properties given by the programmer to parameterize the new Java EE project: project folder, Maven project name, group, and version to identify the artifacts of the project, the database engine, the application server to deploy to, etc.

With the above information, LionWizard generates a Maven multi-module Java EE project, based on an archetype.

In addition, LionWizard lets the programmer select specific components to be integrated into the generated codebase. The wizard automatically integrates those components by transforming the POM files of the required sub-modules and root. That transformation incorporates new dependencies to other Maven artifacts and to the sub-

 ISBN: 978-1-61208-361-2

modules of the selected components. The wizard also modifies the required configuration files to effectively integrate the components into the codebase.

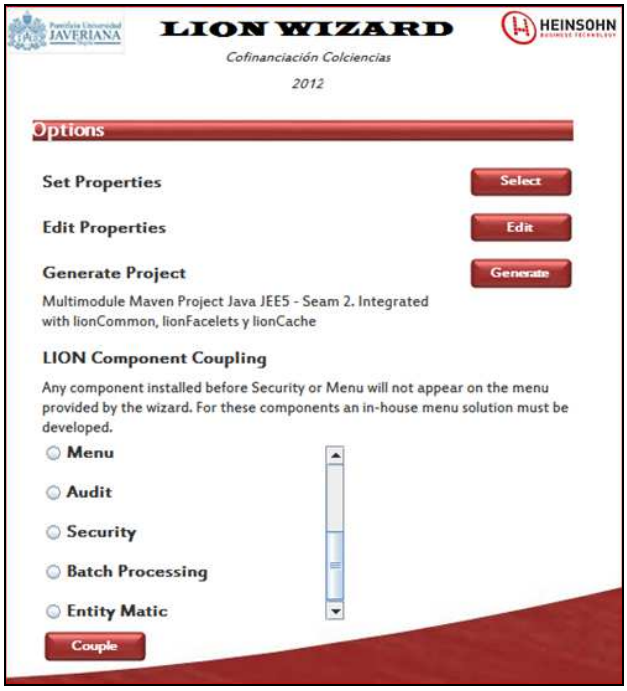


Figure 5. LionWizard's Main Window.

The wizard was designed with sufficient flexibility to seamlessly integrate components that could be developed in the future. To achieve that flexibility, each component has a special configuration file that describes the way to integrate that component into the codebase, originally based on, all of the tasks required to transform the component and the codebase (file modifications, file creations, and properties insertions) to effectively compose them together. Another paper submitted by the authors [28] illustrates the XML configuration file that is used to integrate some components into the codebase of new generated projects.

The automation provided by LionWizard is almost complete. A few tasks are still manual, such as a few SQL script executions and security realms configuration through the console of an application server.

Before the creation of LionWizard, HBT had developed several components. However, their reutilization was hindered by the difficulty to manually integrate them into new codebases. Typically, such integration could take up to three weeks at the beginning of the project. After the creation of the wizard, integration times were reduced to just a few hours, which comprise the time to execute the automatic integration, plus the time to execute a few manual tasks. Another paper submitted by the authors [28] quantifies the benefits of the proposed technology automation through an evaluation of the framework.

VI. DEVELOPING A DOMAIN SPECIFIC LANGUAGE

After executing the previous stages, the integration process of a new codebase can be significantly accelerated. However, the benefits obtained by this automation can be hindered when the codebase evolves during the project, since changes performed to the code may make it harder to automatically integrate new components. Furthermore, as new projects demand the utilization of more recent technologies (e.g., from Java EE 5 to Java EE 6 or 7), make it necessary to modify the wizard to reflect these changes in technology.

Our second joint project with HBT is performing a further step in automation by relying on Model Driven Engineering (MDE) [8] to utilize models as the main development artifact in a software Project.

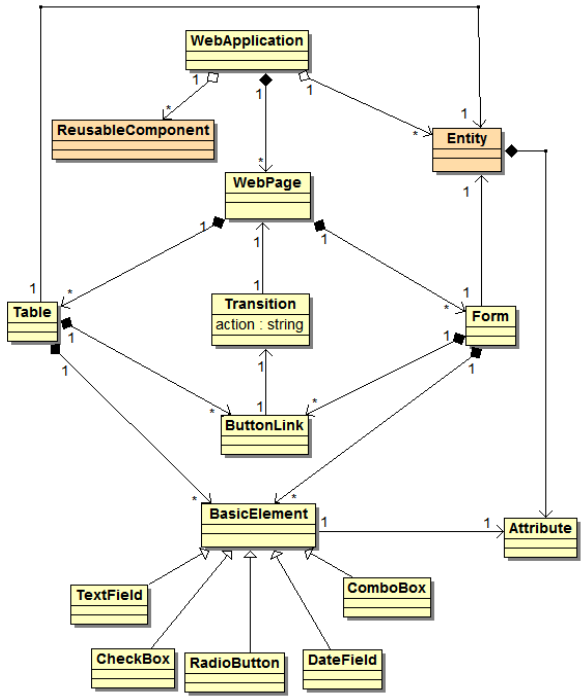


Figure 6. Fragment of the DSL's Meta-Model.

MDE aims to utilize models for every stage in the software engineering process (requirements, design, etc.). In particular, MDE's intent is to utilize transformation tools to automatically generate the source code from models. As a result, a set of models of an application can be used to generate software for different platforms and frameworks. An adequate utilization of MDE may facilitate the evolution and seamless integration of components across the entire development lifecycle and also provide a degree of independence over the technology used to implement the software applications.

In the context of our project with HBT, ongoing work is creating a DSL [27] with an aim to model concepts [32] associated to the structure and behavior of large-scale web applications, while adhering to the adopted multilayer architecture (see Section II). In addition, this language aims to provide clear abstractions of the available components and

 ISBN: 978-1-61208-361-2

their integration mechanisms. The implementation of that language is based on Eclipse EMF [9] and Xtext [33].

Figure 6 depicts a Unified Modeling Language (UML) diagram with the main elements of the language's meta-model, such as the following:

- A web application is composed of multiple web pages and works on multiple persistent entities.
- A web page can contain multiple forms and tables.
- Each form or each table may contain different widgets (basic elements) to depict attributes of the entities.
- Each form or table may also contain links and buttons to connect to other pages after executing some business action.
- Abstractions to model reusable components and their mechanisms of integration.

VII. CREATE MDE TRANSFORMERS

The last stage towards automation is the creation of transformers that can automatically generate application code and integrate components from the information provided by the models specified by the DSL.

In our experience, to create effective transformers it is necessary to capitalize good practices of previous development efforts and also in the previous stages of the proposed method, so that the generated code can include the best strategies for implementation.

Our ongoing work with HBT is creating a transformer that generates Java EE code with the following elements:

- A Maven multi-module web application codebase with the same structure as described in the previous sections of this paper.
- Automatic integration of the existing components. To achieve this, the generator utilizes the information provided by the models (based in the DSL) and the automatic integration code of LionWizard.
- CRUD operations and user interfaces for each entity of the application
- A set of JSF web pages, their backing beans, and a configuration file that describes the page flow (faces-conFigurexml) based in the transitions expressed in the model.

The transformer is based on EMF [9] and Acceleo [7]. Future work is to create transformers for other user interfaces different from JSF, such as Java FX2 [29]. In all of those cases, the models will stay independent of the specific implementation technology.

VIII. RELATED WORK

Maven [13] is widely used to manage the building process in software projects. The evidence of this is the high amount of projects stored in public Maven repositories [12]. The common way to utilize Maven is to automate the compilation and packaging process. Our contribution is the utilization of Maven and its enhancement with additional programs to automate the component integration process, based on a standardized architecture.

There are some tools based in Maven that create codebases from archetypes and generate CRUD operations for entities, such as JBoss Forge [23] and AppFuse [24]. However, these tools do not address the automatic integration of complex components.

There are several strategies for automatic reuse and integration in software projects. Some of the strategies are code generation and program transformation, software product lines, and web services (see a description of these strategies in [28]). Our approach can be classified as a simpler, scope-bounded code transformation approach. Because of its simplicity, it is more maintainable while being adequate for the integration of components into the codebase of software projects.

There are several tools to apply MDE principles for the development of web-based applications, such as Web Ratio [31], Magic [10], and Integranova M.E.S. However, these environments have a high licensing cost, which makes adoption difficult. Moreover, they lack the flexibility to seamlessly adapt their language and transformers to the specific architectures required by a software development organization. The MDE environment that we are building is based on open source software tools that allow HBT to further enrich the DSL modeling language and add new transformers for other technologies.

IX. CONCLUSION AND FUTURE WORK

This paper proposed a method to achieve automation in the development of large web-based applications and the integration of existing components in an organization. This proposal is based on the experience of executing two joint projects with HBT in order to automate their development tools.

The increment in productivity obtained by the first project was promising, since it significantly reduced the times to create new codebases. The ongoing work is expected to yield similar improvements.

Overall, these joint projects have improved the synergy between the University and a software development organization. Future work is to enhance the MDE tools to incorporate additional transformers to other technologies, such as .NET and mobile applications, platforms for which HBT also have components that could be reused. In addition, all of the above will be integrated into the existing product line framework that is being developed in a parallel project [14].

X. ACKNOWLEDGMENTS

Contract/grant sponsor: This article is part of the projects Lion and Lion2, executed by the SIDRe research group of the Pontificia Universidad Javeriana and Heinsohn Business Technology, co-financed by Colciencias (Colombian Administrative Department of Science, Technology and Innovation).

We thank Colciencias [25] for funding the projects described in this paper. We also thank all of the other participants in both joint projects with HBT:

- From HBT: Alvaro Javier Infante, María Catalina Acero, Leonardo Giral, Angee Zambrano, Cristián Fernández, Rubén Darío Betancur, Carlos Díaz, Jorge Camargo.
- From the Pontificia Universidad Javeriana: Andrea Barraza-Urbina, Luisa Barrera, John Carlos Olarte, Francisco Mora.

REFERENCES

[1] C. Larman, "Agile and Iterative Development: A Manager's Guide", Addison-Wesley Professional, 2003.

[2] Heinsohn Business Technology, URL http://www.heinsohn.com.co 01.03.2014.

[3] Oracle, "Java EE at a Glance", URL http://www.oracle.com/technetwork/java/javaee/overview/index.html 01.03.2014.

[4] Foundation A. Maven, "Introduction to Archetypes", URL http://maven.apache.org/guides/introduction/introduction-to-archetypes.html 01.03.2014.

[5] D. Geary and CS. Horstmann, "Core JavaServer Faces", Prentice Hall 3 edn., 2010.

[6] M. Keith and M. Schincariol, "Pro EJB 3: Java Persistence API", Apress, 2006.

[7] Obeo, "Acceleo", URL http://www.eclipse.org/acceleo/ 01.03.2014.

[8] S. Kent, "Model driven engineering", Springer Berlin / Heidelberg, 2002, p. 286–298, URL http://www.springerlink.com/content/9vuqb4hp8fyg2adv/abstract/ 01.03.2014.

[9] The Eclipse Foundation, "Eclipse modeling framework (EMF)", URL http://www.eclipse.org/modeling/emf/ 01.03.2014.

[10] No Magic, "MagicDraw", URL http://www.nomagic.com/products/magicdraw.html 01.03.2014.

[11] Integranova, "Integranova M.E.S.", URL http://www.integranova.com/integranova-m-e-s/ 01.03.2014.

[12] Foundation A. Maven, "The Central Repository", URL http://search.maven.org/#browse%7C47 01.03.2014.

[13] Sonatype Company, "Maven: The Definitive Guide", O'Reilly Media, 2008.

[14] C. Parra, L. Giral, A. Infante, and C. Cortés, "Extractive SPL adoption using multi-level variability modeling", Proceedings of the 16th International Software Product Line Conference - Volume 2, SPLC '12, ACM:New York, NY, USA, 2012, p. 99–106, URL http://doi.acm.org/10.1145/2364412.2364429 01.03.2014.

[15] Oracle, "Oracle Database: Introducing Oracle Database 12c: Plug into the Cloud", URL http://www.oracle.com/us/products/database/overview/index.html 01.03.2014.

[16] Group PGD, "PostgreSQL", URL http://www.postgresql.org/ 01.03.2014.

[17] Oracle, "MySQL", URL http://www.mysql.com/ 01.03.2014.

[18] Microsoft, "Microsoft SQL server", URL http://www.microsoft.com/en-us/sqlserver/default.aspx 01.03.2014.

[19] JBoss Community, "JBoss application server 7" , URL http://www.jboss.org/jbossas 01.03.2014.

[20] Oracle, "GlassFish", open source application server - project kenai, URL https://glassfish.java.net/ 01.03.2014.

[21] Oracle, "Oracle WebLogic Server", URL http://www.oracle.com/technetwork/middleware/weblogic/overview/index.html 01.03.2014.

[22] IBM software, "WebSphere software", URL http://www-01.ibm.com/software/websphere/ 01.03.2014.

[23] JBoss, "JBoss Forge", URL http://forge.jboss.org/ 01.03.2014.

[24] Atlassian, "AppFuse", URL http://appfuse.org/display/APF/Home 01.03.2014.

[25] Colciencias: Departamento administrativo de Ciencia, Tecnología e Innovación,. URL http://www.colciencias.gov.co/ 01.03.2014.

[26] Pontificia Universidad Javeriana, Department of Systems Engineering, URL http://puj-portal.javeriana.edu.co/portal/page/portal/Facultad%20de%20Ingenieria/dpto_sist_presentacion 01.03.2014.

[27] M. Fowler, "Domain Specific Languages", Addison-Wesley Professional, Firts Edit., 2011, p. 413.

[28] M. C. Franky et al., "Achieving Software Reuse and Integration in a Large-scale Software Development Company: Practical Experience of the Lion Project", submitted to IET Software in July- 2013.

[29] J. Weaver, G. Weiqi, S. Chin, D. Iverson , and J. Vos, "Pro JavaFX 2: A Definitive Guide to Rich Clients with Java Technology", Apress, 2012.

[30] G. Booch, J. Rumbaugh, and I, Jacobson, "The Unified Modeling Language User Guide", Addison-Wesley, 2006.

[31] WebRatio, "The New Business-IT Equation", URL http://www.webratio.com/ 01.03.2014.

[32] S. Kelly and J.-P. Tolvanen, Domain-Specific Modeling: Enabling Full Code Generation, 1st ed. Wiley-IEEE Computer Society Pr, 2008.

[33] M. Eysholdt and H. Behrens, "Xtext: implement your language faster than the quick and dirty way," in Proceedings of the ACM international conference companion on Object oriented programming systems languages and applications companion, 2010, pp. 307–309.

[34] M. C. Franky and V. M. Toro, "CincoSecurity: Automating the Security of Java EE Applications with Fine-Grained Roles and Security Profiles", International Journal On Advances in Security (IARIA Journal), vol. no 3&4, 2011, URL http://www.thinkmind.org/index.php?view=article&articleid=sec_v4_n34_2011_10 01.03.2014.

 ISBN: 978-1-61208-361-2

Search Query Share for Enhancing Communication in a Small Community

Tatsuya Ogawa, Nobuchika Sakata, and Shogo Nishida
Engineering Science
Osaka University
Osaka, Japan
e-mail: {ogawa, sakata, nishida}@nishilab.sys.es.osaka-u.ac.jp

***Abstract*—Queries entered into search engines, such as Google, are used for a variety of purposes, such as satisfying user interests and finding some solutions. Some researchers focus on the search queries in order to extend the user interaction and information support. This is because the search queries include user intentions and circumstances. We assume that sharing the search queries can be applied not only to the online virtual world, but also to the real world; in particular, we focus on applying these in a small community. The opportunity of conversation is increased by sharing search queries, which show the user's interests and intentions in a closed community.**

Keywords-search query; communication; small community; search query share.

I. INTRODUCTION

Queries entered into search engines, like Google, are used for a variety of purposes, such as satisfying user interests and finding some solutions. Some researchers focus on the search queries in order to extend the user interaction and information support. This is because the search queries include user intentions and circumstances. Matsui et al. [1] support information exchange among people who have the same interests. It provides online chat conversation to the people who use the same query at the same time on the Internet. Previous studies [2][3][4][5] show that the efficiency of web search activity is improved when referring to the browsed history of other people who used the same search queries. These existing researches support online communication, collaboration retrieval, and other online collaboration. In this research, we assume that sharing the search queries can be applied not only to the online virtual world, but also to the real world; in particular, we focus on applying these in a small community.

Shared spaces, such as lounges, corridors or kitchens, are important areas for informal communication and sharing information. Some researches provide opportunities for communication in order to utilize specific informal communication [6][7]. Although those researches support incidental encounters and provide opportunities for a starting point of a conversation, they do not provide the conversation topic. HuNeAS and Cyber-IRORI intend to provide the conversation topics [8][9]. However, these studies focus their attention on the method of how to express information among community participants. Also, there is no mention on how to start a conversation topic.

We assume that starting a conversation from incidental encountering happens quite often in small communities, such as a university laboratory or a small office. Also, we assume that events in daily lives, interests and thinking among participants of small communities tend to be shared as conversation topic. Therefore, the opportunity of conversation is increased by sharing search queries which show user's interests and intentions in a closed community.

The remainder of this paper is structured as follows. Section II describes related researches. Section III describes search queries and conversation topic. Section IV shows an architectural overview of our system and the user study conducted by our system. In Section V, we discuss our findings. Finally, Section VI concludes this paper.

II. RELATED WORK

In this section, we describe related researches of enhancing communication among the community.

A. Provide an Opportunity for Conversation

Meeting Pot matches the timing of people to meet each other [6]. This system detects the presence of people in a break room by monitoring the state of a coffee maker. When it detects people gathering, the system informs the colleagues in personal offices by using a coffee aroma generator as implicitly communication.

Also, communications tend to be neglected in office work spaces by partition panels that make up a personal cubic space. A psychological barrier not to visit other personal space tends to be caused by the physical partition panels without any special reason. TravelingCafe compensates the psychological barrier in such an office [7]. This system monitors the remaining amount of coffee in their cup with pressure sensors. By displaying the remaining amount of all member’s coffee cups beside the coffee maker, this system give a reason to visit some colleagues to "come to pour some coffee", and that might start a conversation. The psychological barrier of hesitating to visit other personal space is reduced by that system. Furthermore, they supposed that the subjects became tolerant to "visit" and "being visited." Also, they reported that the system can create new communication because the communication between two persons is extended to communication among multiple people at the same rate as the one used in a break space.

This research shows that synchronizing break time among people can cause a chance of occurring

 ISBN: 978-1-61208-361-2

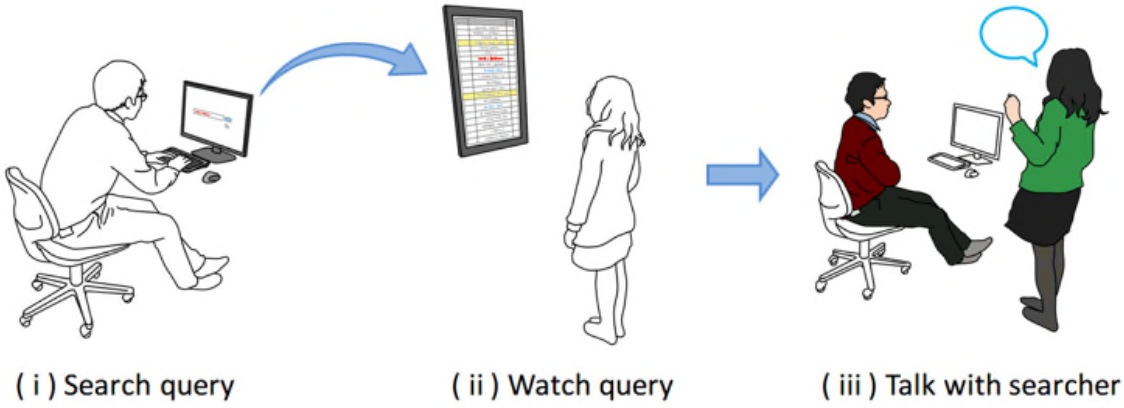


Figure 1. Concept of proposed system

communication. However, a common topic of conversation is not provided in the system. We assume that providing a common topic among colleagues is important as well as the synchronization of break time.

B. *Conversation Support in a Shared Space*

HuNeAS studied encounters of people in a shared space [8]. HuNeAS promotes information sharing by indicating the required information on a large display placed in a common space such as a corridor or a lounge. At the same time, to enhance the information sharing in that community, HuNeAS gathers information from all users and places it on the large display.

Cyber-IRORI showed that coziness is important for supporting informal communication [9]. It provides a reason and the justification to visit/stay at the shared space. Cyber-IRORI consists of a touchable table-top display, a vertical display, and imitated hearth. This system displays news and web pages read by the community members on the touchable table-top display as articles. Also, when a member choses a certain article, the article is displayed on another larger display which can be seen from anywhere in the community room. They described that Cyber-IRORI encouraged the users to stay/visit the shared informal space and was able to enhance conversations.

III. SUPPORT CONVERSATION WITH SEARCH QUERY

As mentioned before, queries input to search engines like Google are used for a variety of purposes, such as satisfying user interests and finding some solutions. Some researchers focus on the search queries in order to extend the user interaction and information support. This is because the search queries include user intentions and circumstances. Matsui supports communication among people who search the same query [1]. Yamaguchi, Tan, Tanaka, and Takeda made online-searching more efficient by using searched queries used before as reference [2][3][4][5]. According to existing studies, we assume that the advantage of sharing query for people in a small community is increases the opportunities of conversation.

A. *Search Query and Conversation Topic*

Our research focuses on a certain community where people are acquaintances and talk to each other quite often, such as a university laboratory and a small office. We assume that conversation topics tend to be about events in daily lives, interests and thinking among participants of a small community. When the conversational partner has no interest in a topic, the communication tends to become a one-sided conversation or the conversation is over in a minute. As we know, it is very important how much the partners of a conversation are interested in the topic in terms of having comfortable conversations. As we have described previously, search queries reveal user interests. There is a possibility that the search query is able to be the topic. Therefore, it is expected that opportunities of conversation are increased by sharing search queries as conversation topics.

In this research, to enhance communication among members of a small community, we propose a new query sharing system which indicates that queries searched by community members are available on large display in a shared space. Hence, we designed a system to increase the amount of conversation by sharing queries which someone are interest in when they visit a shared space. Fig. 1 shows the basic concept of the system. Fig. 2 shows the actual display placed in shared space in this experiment. Note that, at this time, we place the display in a shared space because we suppose to observe face to face communication in real space, and it is easy to sample data of people behavior in a shared space. Also, we did not want to interfere with someone's work by sharing queries in personal space. Basically, we can realize the query sharing system on each PC in a personal space, such as an office cubicle space. We believe that our proposed system can be applied to full online style and can produce same or more effect when queries that are shared only on one display.

IV. USER EXPERTIMENT

In this paper, we defined "search query shared among small community" as "closed query". We want to measure the effect of a closed query. This section describes the details

 ISBN: 978-1-61208-361-2

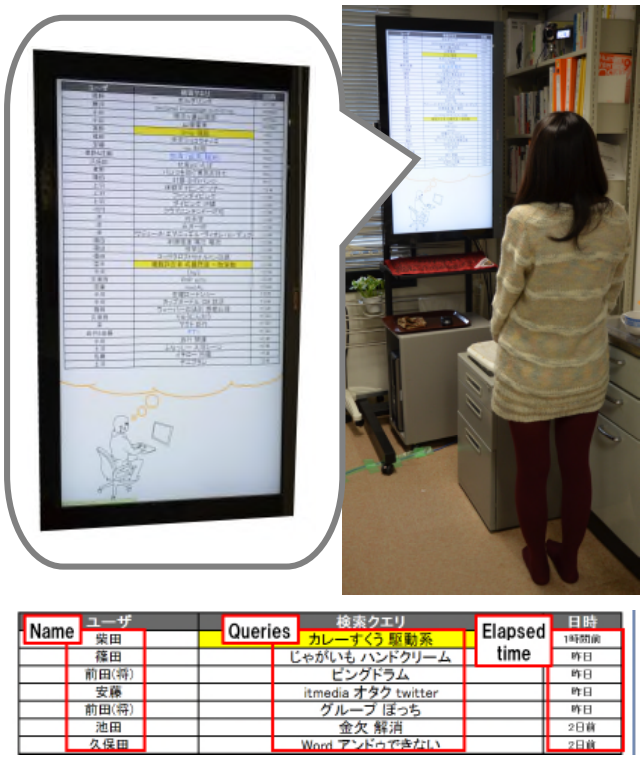


Figure 2. Large display placed in shared space and displayed closed query and part of list of queries.

of the experiment and evaluation gathered from a questionnaire.

A. *"Global" and "Local" Query*

We conducted a user study to compare a closed query with a "hot word", which is a word used in lot of people searches. We thought that a hot word cloud be regarded as a global query which is symmetrical about closed query. This way, we investigated whether a closed query is more effective to increase opportunities of conversation than any queries. This section describes the design and the result of the comparative experiments.

The system indicates list of "closed query" on a large LCD (SONY KDL-40EX500:40-inch LCD TV) in 7 days, as shown in Fig. 2. The list is composed of 35 “closed query”s and the name of the person who searches “closed query”, as shown in bottom of Fig. 2. After that, the system indicates the list of "hot words” for 7 days as well. In this situation, we believe that the order effect was negligible. We conducted a questionnaire investigation after each period. In this experiment, hot words were displayed one week after the queries were closed. Closed queries were obtained from search form implemented Google Chrome extension (Fig. 3). Therefore, participants can adopt inputting the search query to the implemented form or usual form whether or not they want to share the query. Also, we deploy a video camera beside the display to observe whether subjects are interested in displayed queries when they visit the shared space. This experiment was conducted with 17 subjects (gender: 16 male and 1 female; age: 22 to 25). Also, trend words of Yahoo! Japan Search data are displayed as hot word [10]. The trend words are updated every hour and 3 words are added on that screen.

Figure 3. Search form implemented Google Chrome extension

TABLE I. QUESTIONNAIRE ABOUT SHARING CLOSED QUERY

P1	Did you check the search query frequently when you were in the shared space?
P2	Did you have interest in search queries?
P3	Did you incline to talk about the search query?
P4	Did you try to talk about query with the searcher?
P5	Did you talked about search query?
P6	Did you feel uncomfortable sharing own queries?
P7	Did you have interest in other's queries before experience?
P8	Did you have interest in other's queries after experience?
P9	Did you feel conversation was increased in small community?
P10	Did you feel conversation is efficient for small community?
P11	Do you want to use this system in future?

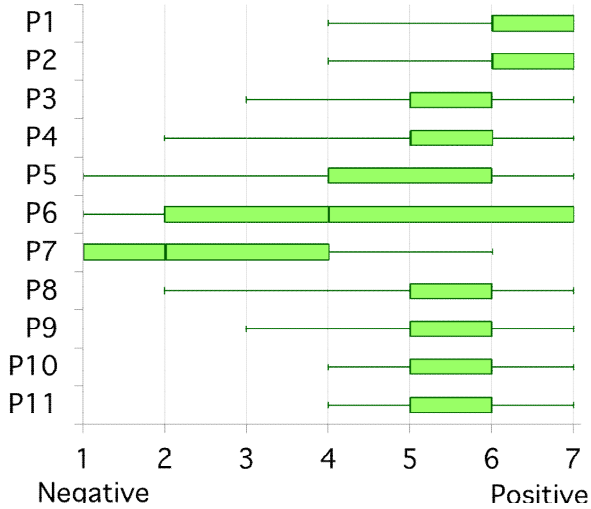


Figure 4. Evaluation of closed query

B. *Influence on Communication in Closed Query*

We carried out a questionnaire survey on an absolute scale from 1 to 7 (1: Negative 7: Positive) to 17 subjects after experiments. Table I shows the questionnaire items and Fig. 4 shows the results after the closed queries were shared.

Almost all subjects who visited the shared space checked the closed queries (P1), and they were interested in the closed queries (P2). They said that "It was interesting that I was able to see other's new side, such as interests, thinking, and progress of work." According to the user comments, almost all subjects had a good impression about the closed queries of others. Also, they shared their closed queries (P3, P4, P5). Even subjects who did not like to share their search query were interested in the closed queries and had conversations about the closed query. Also, there were variable comments about sharing closed queries. Some comments were on the negative side, such as "I didn't want to be known that I searched something very basic. I do not want others to think I am not very knowledgeable.", "Sharing my hobby makes me feel ashamed". In addition, as the opposite opinion, "I share the queries about my hobby mainly because I thought that it was boring even if a serious conversation with queries about research." All subjects tended to be aware that queries cloud be shared as topic. Subjects who were not interested in other's queries, became interested in closed queries. We suppose that their decision to share search queries increased because they could obtain a variety of information from closed queries.

Also, we observed that some people searched the same query at the same time. This means that they had deep interest and conversations with those close queries.

C. *Comparison between "Closed Query" and "Hot Word"*

We deployed a video camera beside the display to observe whether subjects are interested in the displayed search queries when they visit the shared space. Therefore, we counted the frequencies of visiting the shared space and checking queries at that time. Table II shows the number of times that subjects have visited by subjects shared space. Also, we have compared search queries during the most crowded four hours between 15:00 to 19:00 on each day of displaying closed query and hot word. In addition, the date of the observation was chosen when it seemed that enough time has passed since the start of the experiment. This was done in order to compensate for the initial curiosity of deploying the new system. However, this data seems to be just for reference because we analyze only one day.

When subjects visited the shared space, they checked the display at a frequency of 34.8% in case a closed query was displayed. They checked display at a frequency of 18.3% in case a hot word was displayed. Also, few subjects checked the hot word more than once a day. However, almost all subjects check the hot word just once a day. We can say that subject who is interested in the closed query does not have an interest in the hot word. Therefore, it was indicated that there is a difference in the frequency with which the search query was seen.

Next we describe comparison experiment of closed query and hot word. Fig. 5 shows questionnaire item and Fig. 6 shows results after the experiment.

Q1	Is your amount of checking closed query / hot word frequently?
Q2	Did you have interest in closed query / hot word?
Q3	Did you incline talking about closed query / hot word?
Q4	Did you talked about closed query / hot word?
Q5	Did you feel conversation was increased in small community?
Q6	Did you feel conversation is efficient for small community?
Q7	Do you want to use closed query / hot word sharing system in future?

Figure 5. Questionnaire for sharing closed queries

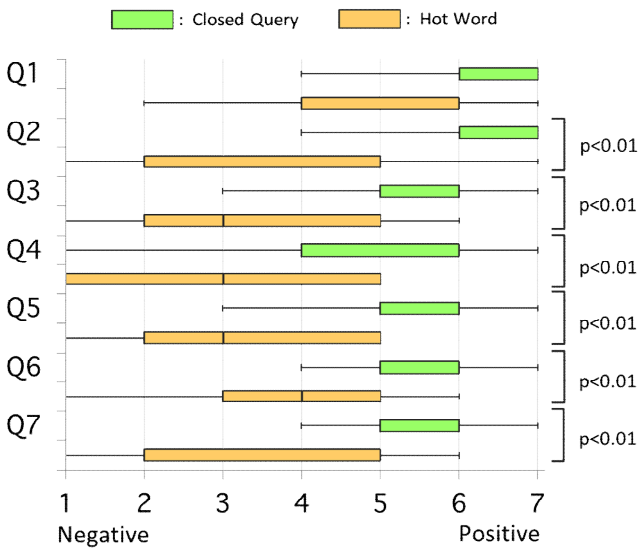


Figure 6. Result of comparing closed query and hot word .

TABLE II. FREQUENCY OF WATCHING CLOSED QUERY AND HOT WORD

	Number of times	
	Visit shared space	***Check display***
Closed query	69	24
Hot word	71	13

Using Wilcoxon matched-pairs signed-rank test, a significant difference was found in all items except for Q1 ($p=0.08$). It means that most subjects had often seen both the closed queries and the hot word (Q1). However, conversations caused by the hot word did not occur so much, and it was less than in case of sharing closed query (Q2, Q3, Q4). Closed queries made the subjects determine who was interested in what topic easily. On the other hand, sharing the

 ISBN: 978-1-61208-361-2

hot word cannot initiate conversation as well as the closed query. Therefore, conversations caused by the hot word as topic occurred almost always when some subjects were "in a shared space" and "at the same time". Also, there were many comments such as "I was good to know the interest of the other party before a conversation."

D. Difference of Queries by Different Communities

Closed queries are classified into two types, such as "amusing and attractive word" and "word concerning work". In the previous experiment, a certain group tends to share "amusing and attractive word". On the other hand, other groups tend to share "word concerning work". Also, those two groups are researching difference topics and obeying a different group leader. Therefore, we divided the subjects into two groups and conducted the experiment in which closed queries were shared in each group independently. We observed the effect of the classification of these groups on the communication.

As a result, there were some positive comments, such as “It was easier to share a search query than in the previous experiment.” On the other hand, many subjects, especially within the group that tends to share queries of type "word concerning work", had a bad impression compared to the case when they shared a query with the entire laboratory. They said that “It was limited in what I said about the closed query.” and “It is difficult to talk about the query related study.” This reason is estimated that it is hard to talk about query because queries about "amusing and attractive" are decreased. This result might show that amusing and attractive words tend to be nice topics to start a conversation.

V. DISCUSSION

As a result of our system, conversation is promoted by sharing search queries in a small community. Indicating the closed queries and the user names who searched the closed query at the time, the closed query have become interesting topics for all conversation participants.

In this experiment, words relating to hobby were many in closed queries. It can be said that subjects who searched about hobby tended to use the proposed query sharing system. On the other hand, subjects who did not share the query much said "It was difficult to synchronize a timing to talk." or "I abstained from sharing queries about a hobby." These differences in type and amount of closed query were observed by groups. We consider that there is a difference of viewpoint in members of each group how to regard closed queries as. In future work, by conducting experiments in another community, there is a possibility that usage of the system and the type of closed query becomes different. We believe that we can obtain a significantly different result from this experiment when applied to a more intimate community such as friends, family members, and lovers. In addition, these experiments were conducted in collegial relationships. However, search queries were shared in reporting relationships like superiors, subordinates and supervisors, it is supposed that the proposed system become a useful tool in order to help with some projects.

Table III shows certain moments of closed queries in the experiments. We describe the searcher's conditions which can be estimated from those words. In this example, we classify those words by hand, not automatically. When the query is about a hobby, the news, or a topic word such as "*football match Serbia schedule*", "*Entrance examination disestablish*", or "*LEGAL HIGH* (TV drama in Japan)", it is supposed that subject is resting. Some subjects presumed to talk about fast-reading of English because they searched "*English fast-reading*" at the same time. When searching this type of words, the condition of the user can be categorized as "Resting". While in this condition, it is easy to start the conversation with the person who searches the same kind of words without feeling that you disturb them. Also, when the subjects search solutions, such as "*c++ log file input/output*", "*opencv plane detection*", especially when the subject searches the same or similar words repeatedly, such as "*php js function*", "*php function*", this may seem like a problem occured while working. These conditions are categorized 'In need'. Besides, in case queries are titles of paper and English words, the searcher is working. These searchers conditions are categorized 'Busy'. We assume that indicating closed query and user condition at the same time make starting communication more smoothly. For example, when someone is interested in closed query noticed with the 'Resting' condition, they might start communication without any hesitation. Also, by noticing 'in need' condition to people during a break, there is a possibility that a problem is resolved soon in case another person in the community knows the solution. We supposed that sharing search queries with searchers' conditions is efficient for enhancing opportunities of conversation start. Also, we suppose to consider and implement a classifier for deciding user’s condition.

TABLE III. CERTAIN MOMENT OF THE SEARCHER’S CONDITION ESTIMATED FROM THE SEARCH QUERY

Search Query	condition
football match Serbia schedule Entrance examination disestablish English fast-reading Grave of the Fireflies Kobe PE line leader LEGAL HIGH Osaka brass band regular concert	Resting
c++ log file input/output opencv plane detection php js function php function boost::serialization	In need
PocketTouch Skinput Large-scale learning of word ~	Busy

VI. CONCLUSION AND FUTURE WORK

In this paper, we proposed query sharing system to encourage conversations in a small community. Also, we studied whether there is a possibility that shared queries become trigger of conversation. By sharing closed queries and hot words in a small community, we observed that

members of small community are more interested in their closed query. Finally, we discuss and suggest that indicating a closed query and search condition at the same time make starting a communication smoother.

In future work, by conducting the experiment with another community, there is a possibility that usage of the system and type of closed query becomes different. We believe that we can obtain a significantly different result from this experiment when applied to a more intimate community such as friends, family members, and lovers. For example, search queries were shared in reporting relationships like superiors, subordinates and supervisors. Also, the authors believe that the proposed system becomes a useful tool in order to help with some projects.

REFERENCES

[1] Y. Matsui, Y. Kawai, and J. Zhang, "Page as a Meeting Place: Web Search Augmented with Social Communication", Intelligent Control and Innovative Computing, Lecture Notes in Electrical Engineering Volume 110, pp. 303-317, 2012.

[2] T. Yamaguchi, A. Niimi, and O. Konishi, "A Method of Organizing the Group to Improve Search Efficiency with Sharing History of Web Browsing", The 71th National Convention of IPSJ, pp. 1-681-1-682, March, 2009 (in Japanese).

[3] H. Tan, I. Ohmukai, and H. Takeda, "QueReSeek: Community-Based Web Navigation by Reverse Lookup of Search History", in Proceedings of the 2008 International Workshop on Information-Explosion and Next Generation Search, pp. 31-38, 2008.

[4] Y. Tanaka, Y. Suhara, N. Hiroshima, H. Toda, and S. Susaki, "Snippet generation by Identifying Attribute Associated Information". the 9th Asia Information Retrieval Societies Conference, pp. 50-61, December, 2013.

[5] T. Takeda and T. Igarashi, "Improving the efficiency of web search by reusing search history of group members", IPSJ SIG technical report HCI 2008(11), pp. 93-98, 2008 (in Japanese).

[6] I. Siio and N. Mima, "Meeting Pot : Coffee Aroma Transmitter", Poster/demonstration in ACM UbiComp 2001: International Conference on Ubiquitous Computing, Sep.30-Oct.2, 2001.

[7] T. Nakano, et al., "The Traveling Cafe: A Communication Encouraging System for Partitioned Offices", ACM CHI2006 Conference on human factors in computing systems, pp. 1139-1144, 2006.

[8] K. Nishimoto and K. Matsuda, "Informal Communication Support Media for Encouraging Knowledge sharing Creation in a Community", International journal of information technology and decision making (IJITDM), Vol.6, No.3, pp. 411-426, September, 2007.

[9] M. Usuki, K. Nishimoto, K. Sugiyama, and T. Matsubara, "Evaluation of the IRORI: A Cyber-Space that Catalyzes Face-to-Face Informal Communication", In knowledge-Based Intelligent information and Engineering Systems, Proceedings of 8th International Conference, KES 2004 volume 3213/2004 of Lecture Notes in Computer Science, pp. 314-321, Springer, October, 2004.

[10] Trend Words of Yahoo! Japan Search data , http://searchranking.yahoo.co.jp/realtime_buzz/ [retrieved: 28th May, 2014].

 ISBN: 978-1-61208-361-2

Psychology for Predicting Internet Behavior Patterns

Nadejda Abramova[1], Olga Shurygina[1], Alexander Kondratiev[1], Ivan Yamshchikov[1`2]
[1]Yandex, Moscow, Russia
[2] University of Applied Sciences Zittau/Goerlitz, Germany
{abramovanadia, o-shurygina, fefa4ka}@yandex-team.ru , i.yamshchikov@hszg.de

***Abstract*—Yandex[10] is an international search engine with more than 19 million users daily only in Russia. In order to improve their daily search-experience and make interfaces more user-friendly, Yandex carries out a great number of research activities. This research was aimed at finding a reasonably fragmented audience segmentation that should be based on psychological principles and be automatically processed. We assume that every person has his search behavior characteristics that could be explained by some stable psychological type. By search behavior we mean all the actions the users undertake to find the answers on the Internet: request formulation, quantity, timing and duration of clicks and views, returns to the pages visited before, the habitual usage of tabs and windows, navigation within the service and so on. Using cognitive psychology, we have managed to create the segmentation that has certain predictive power and, therefore, could be used in industrial applications. Based on the qualitative, as well as on the quantitative, analysis of the user-behavior, we have developed two binary scales that split the audience in four groups. The first scale "Analytic – Synthetic" could be roughly characterized as a scale describing the style of information processing, that is more natural for a user and classifies the attention markers. The second one "Logical – Ethical" deals with the informational priorities. We describe these two scales and show how they can be found from the users' behavior. We give a brief explanation of how interactions with the service could be improved based on the knowledge of user's psychological characteristics. Though all results obtained are based on a specific service and can vary from country to country due to the cultural differences, we strongly believe that these two scales, addressing the fundamental priorities of human personality, could be applied for other products and services. Moreover, understanding of general principles that guide the behavior of each user group could help generate user behavior hypotheses to other behavior researchers.**

Keywords - user segmentation; cognitive psychology; Internet behavior.

I. INTRODUCTION

Internet is an information environment that is a part of almost every person's life nowadays. Though every person is unique, in order to improve daily user-experience, a company needs to classify users according to their similarities. When behavior is classified qualitatively, somebody needs to find a quantitative metric that would allow targeting all the users that share the same characteristics. The knowledge of such metrics and the understanding of how we can improve the product or the service for these specific people could be applied to the product and increase user-satisfaction, because it would make uses' lives a bit better and simpler. This approach is called user segmentation and our research is carried out exactly in this framework for Russian search engine Yandex. Our method is based on concepts of cognitive psychology and is to give a possibility to track down people of each psychological type automatically, without any questionnaires, which is rather new and interesting. Firstly, psychological insights can give a broad and detailed description of many different aspects of human behavior; thus giving us great predictive power. Secondly, psychological theories claim to be universal and therefore, applicable to all users. A variety of theories and approaches gave us hope, that we could find an insightful idea of a segmentation that could be general, stable, measurable and weakly dependent on the timely factors such as the user's physical conditions or current mood.

To find out that, we divided our work into three parts: theory and hypothesis setting, qualitative experiments deployment and quantitative checking to verify our hypothesis. So, this paper is organized as follows: firstly, we describe findings of related works and theory behind our classification, and then, we will present the hypothesis, qualitative and quantitative experiments and findings. The conclusion and acknowledgement close the article.

II. RELATED WORKS

The whole concept of using psychology for Internet behavior studies is not really new. A number of researchers have applied psychological ideas and principles in their research of the Internet behavior. Therefore, we mention the most important areas that have a direct connection with this paper.

One of the most successful syncretic areas of psychology and IT (Information Technologies) is the area of adaptive interfaces. In 2011, Susan Weischenk published a book "100 Things Every Designer Needs to Know About People" [1], where she gives a cookbook of practical advices for the interface-designers that is entirely based on psychological theories and concepts. The whole industry of usability is largely based on the idea of applying empirical psychology to the interface-design. The concept of improving the whole user experience with the knowledge of the user's motivation, social-demographics and personal qualities is profoundly discussed by Anderson in [2].

 ISBN: 978-1-61208-361-2

Another area is behavior analytics, where one can find two major research concepts. The first one could be called a unified approach, where one tries to find a psychological feature applicable to all human beings. Stecher and Counts in [3] studied attention, memory and thinking process and came to the conclusion that some characteristics are shared among all users. For example, they proved that trait information is remembered preferentially to the content. But as noted, such patterns are inherent to all people, without going into details why some people are more successful at remembering context or other type of information. The second approach could be called a segment-based approach. The concept behind is that there are distinguishing markers - personality parameters - that could split people into certain groups with the same behavioral patterns. Such approach seems to be really useful in the area of targeting, e.g., [4] [5]; therefore, we decided to follow this path. While it is typical to use segmentation for a specific use-case, we wanted to create a psychological segmentation that could be applied to a vast majority of cases and, consequently, would be of a great industrial interest.

Since the Internet is an informational environment, we decided to focus on the psychological theories covering issues of dealing with information. To be more specific, we were interested in information perception and processing. As a starting point, we used a classical concept of a cognitive style that was first introduced by Bieri [6] in 1955. This concept is well known and widely used in academic psychology, as well as in industrial applications. It was profoundly studied and applied to the needs of education by Riding in [7], [8] and [9]. Assuming that interaction with the Internet is an elementary exchange of information between users and environment, we could apply the concept of cognitive styles to every user action. Moreover, cognitive styles are stable, they are hardly altered by mood or physical conditions and they characterize people behavior over the long haul. All these advantages make cognitive styles extremely promising and interesting for our problem.

III. PSYCHOLOGICAL SEGMENTATION

Our research is simultaneously carried out in several directions. We started with a pure theoretical hypothesis based on the classical psychological concepts and our demands on the segmentation, such as stability and measurability that we have already briefly mentioned. Then, we verified this hypothesis in a qualitative as well as in a quantitative way and now are trying to apply obtained expertise to the products.

A. Theory and Hypothesis

Relying upon the ideas and the theoretical background mentioned in the Introduction we were looking for a segmentation that could correspond to the following conditions:

- Applicable to a variety of use cases, problem contexts and users
- Relevant for our target audience
- Measurable with some standard face-to-face means
- Time-stable
- Detectable, leaving digital imprint
- Able to find predictable behavior patterns applicable to the whole segment

Finally, we have found two cognitive scales that seem especially interesting and promising for us, see Figure 1. The first scale is Analytic - Synthetic. It describes how people deal with different information. Analytic tends to specify, go into details instead of scanning and building up the whole picture. Analytic tends to be focused on and act stepwise solving one problem at a time. Synthetic, on the contrary, is a multitasking person, who tends to generalize and does not like to pay any attention to the details. He switches between issues quickly and could be easily distracted.

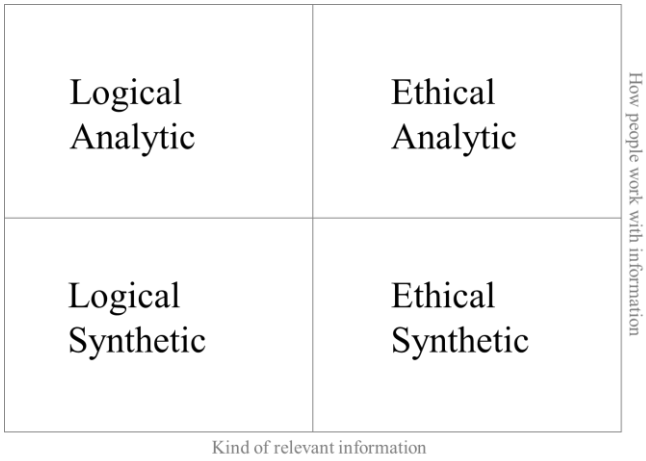


Figure 1: Four psychological types based on the cognitive concept.

Logical - Ethical is the second scale that describes which kind of information a person would consider as most valid, interesting and reliable. Logical people tend to focus on facts, measurements, depersonalized and objective arguments, while Ethical people value subjective, person-focused feedback and experience. They also tend to look for personal emotional insights, rather than to look at bare facts.

These two scales with the binary outcome for each split the users in four different groups. In our qualitative and quantitative research, we tried to find and describe these four groups in a greater detail.

B. Qualitaive experiments

The aim of this part of research was to develop a portrait of each psychological type through the deep interviews of 40 people. We were going to find out their Internet preferences and habitats. Therefore, we have carried out a series of interviews with the following procedure:

- At first, a respondent was asked to fill in the forms based on the standard questionnaires used for these two cognitive styles in the classical psychology to understand a person's type.
- Then, the user was asked to carry out a series of simple tasks that involved the usage of the interface.

 ISBN: 978-1-61208-361-2

- We tracked his attention, focus and motor activity to compare our expert-based impression with the results of the test and with the collected data.
- Afterwards, we tried different user cases with the same user varying his level of competence in the area and given time for the task.

Based on these interviews and collected data we have managed to see representatives of all four groups and build a qualitative description of the typical representative of each group. Here is the summary of our results, describing four psychological types behavioral patterns:

- Logical Analytic type
  - searches step-by-step
  - plans his actions before doing them
  - concentrates on objective facts
  - specifies every search parameter
  - tends to fact-check and compare parameters
  - eager to solve the problem in the most effective way
- Ethical Analytic type
  - searches step-by-step
  - does not perform multi-tasks
  - pays attention to any related type of user generated content eagerly
  - tends to find the most useful solution of the problem
- Logical Synthetic type
  - searches multidimensionally
  - performs spontaneous changes in the task
  - tends to prefer facts and numbers over a personal experiences, but does not bother with fact-checking
- Ethical Synthetic type
  - searches multidimensionally
  - sometimes does three or more tasks in parallel
  - switches easily from one task to another one
  - can be easily distracted
  - focuses on personalized experience
  - looks for the most impressive solution of the problem

We have also found some interesting features of the behavior that could be attributed just to the Analytic - Synthetic scale. For instance, Analytics tend to do one thing at a time. Even if you ask them a question, while they are surfing the Internet, they tend to continue browsing only after the question is answered. This observation gave us a clue that content of the site and advertisement should be strictly connected to the query. Synthetics can have more objects in their attention at once but they pay for it with some negligence and carelessness. They even make more typos. That means that content of the web page can contain both context information and additional information as well. So, the design for Synthetics should be easy enough for them to follow several topics or objects they are interested in.

Moreover, as we tracked eye-paths, we have discovered that Analytic people, especially Logical Analytical type people are mostly consistent in their actions. They view every page slowly, from the top to the bottom and do it very attentively. Synthetic type people, on the contrary, are very chaotic. They scan the page in a random way. To get detailed attention patterns of each psychological type is an aim for our future research. However, these findings gave us a clue that interfaces design as well as content should differ for each type in order to increase convenience for the user.

Such substantial differences in the behavior of each type encouraged us to think of further qualitative experiments. Furthermore, all gathered qualitative research results motivated us to move on to quantitative experiments and gave lots of insights for the calculations and data-analyses that we are still doing at the moment.

C. Qualitative research

The first and foremost, we decided to focus on the scale Analytic - Synthetic as on more promising and more insightful one in our quantitative research. We have constructed a number of metrics to verify the fact that these two groups exist, such as reaction rate, number of entry points to start search, query length, regression in navigation and change of topics. But, in this paper, we would like to focus on only two of them, which are the most illustrative.

1) Regressions in navigation: Based on the evidence we have got from the qualitative experiments that Analytics should have more linear navigation patterns in comparison with Synthetics, which we believe have more Internet navigation recursions and loops, we have constructed the following metric.

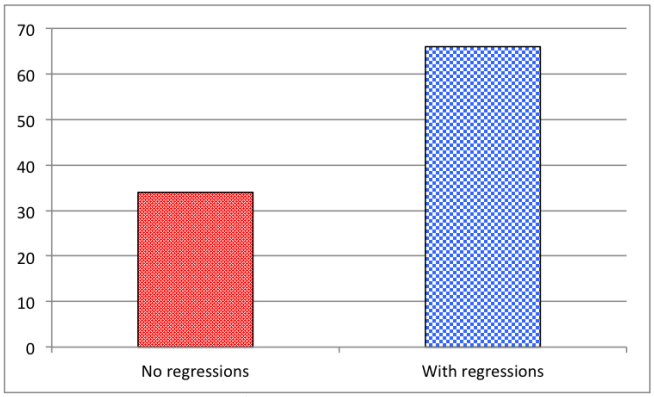


Figure 2: Percentage of the users without and with regressions in the monthly audience of Yandex Russia.

Let us assume that the user has opened a certain page A then moved from it to the page B. Now, if this user does not show any activity on the page B and comes back to the page A within a short interval let us call such pattern a regression. The time interval was estimated as an average time between the click opening the page and the first click on it calculated for this user. We have used one month dataset of approximately 35 million unique users. As it turned out, almost a third of the users did not have any regressions within a month; see Figure 2.

 ISBN: 978-1-61208-361-2

2) Change of the topic: When users search something on the Internet, they naturally input search queries. One can classify if two queries going one by one within one session have the same general topic or are not even slightly connected. As we have seen from the qualitative experiments, Analytics should have no unconnected queries and Synthetics, on the contrary, should change query topics from time to time during one session. It turned out, see Figure 3, that there is another strong dichotomy with approximately 30 percent of the audience not having any single change of the topic within a month.

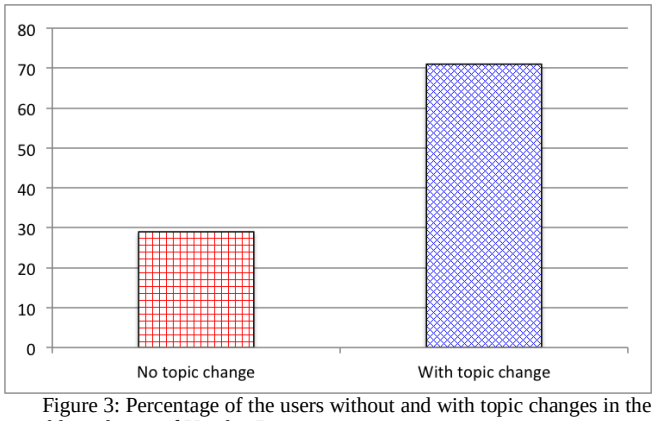


Figure 3: Percentage of the users without and with topic changes in the monthly audience of Yandex Russia.

3) Further quantitative research: We compared these two metrics and got 70% correlation. This fact is extremely interesting because metrics are built on different counters and describe different activities. Synthetics tend to have regressions and changes of the topic and Analytics tend to do everything in a more organized step-by-step way.

We are currently running a number of experiments in order to prove the direct connection between the cognitive styles and the other metrics found. This proof is extremely important for the possible results generalization.

IV. CONCLUSION AND FUTURE WORK

We believe that we have found a very fruitful psychological hypothesis that can be a basement of very general user segmentation. Segments existence was proved by a series of qualitative and quantitative experiments that we briefly mentioned, but could not describe in a greater detail. Our cross-function approach is new in terms of the scalability and generality. It characterizes human personality fundamental qualities and can be applied for a great range of issues from interface improvements, content customization to user targeting. Moreover, we have another series of experiments that should produce additional targeting metrics and prove behavioral changes dependence on cognitive styles and not on some other hidden factor. We also plan to apply the results obtained to a specific industrial product in the nearest future.

ACKNOWLEDGMENT

The authors would like to thank Valentina Nochka and Denis Popovtsev for their help and support. We also are extremely grateful to Andrey Sebrant for his bright ideas and guidance.

REFERENCES

[1] S. Weinschenk “100 Things Every Designer Needs to Know About People”, Pearson Education, 2011.

[2] P. Anderson, “Seductive Interaction Design: Creating Playful, Fun, and Effective User Experiences”, New Riders Pub, 2011.

[3] K. Stecher and S. Counts, “Spontaneous inference of personality traits and effects on memory for online profiles”, in ICWSM 2008 Conference Proceedings, ICWSM, 2008

[4] S. T. D. Ryen, W. White, and J. Teevan, “Characterizing the influence of domain expertise on web search behavior,” in WSDM 2009 Conference Proceedings. WSDM, 2009.

[5] D. S. M. Kosinskia and T. Graepel “Private traits and attributes are predictable for digital records of human behavior”, 2012.

[6] J. Bieri, “Cognitive complexity-simplicity and predictive behavior,” Journal of Abnormal and Social Psychology, vol. 51, 1955, pp. 263–268

[7] R. Riding and I. Cheema, “Cognitive styles - an overview and integration,” Educational Psychology, vol. 11, no. 3/4, 1991, pp. 193–215

[8] R. Riding and E. Sadler-Smith, “Type of instructional material, cognitive style and learning performance,” Educational Studies, vol. 18, no. 3, 1992, pp. 323–340

[9] R. Riding and E. Sadler-Smith, “Cognitive Styles and Learning Strategies: Understanding Style Differences in Learning and Behavior”, London, David Fulton Publishers, vol. 5, no. 4, March 2002, p. 217

[10] company.yandex.com [accessed June 2014]

 ISBN: 978-1-61208-361-2

A Structured Approach to Architecting Fault Tolerant Services

Elena Troubitsyna, Kashif Javed
Åbo Akademi University, Finland
Elena.Troubitsyna@abo.fi, Kashif.Javed@abo.fi

***Abstract*— Service-oriented computing offers an attractive paradigm to designing complex composite services by assembling readily-available services. The approach enables rapid service development and significantly increases productivity of the development. However, it also poses a significant challenge in ensuring quality of created services and in particular their fault tolerance. In this paper, we propose a systematic approach to architecting complex fault tolerant services. We demonstrate how to graphically model the architecture of composite services and augment it with various fault tolerance mechanisms. We propose an approach facilitating a systematic analysis of possible failures of the services, recovery actions and alternative solutions for achieving fault tolerance. Our approach supports structured guided reasoning about fault tolerance at different levels of abstraction. It allows the designers evaluate various architectural solutions at the design stage that helps to derive clean architectures and improve fault tolerance of developed complex services.**

Keywords - services; fault tolerance, architecture, service composition, service orchestration; failure modes and effect analysis

I. INTRODUCTION

Web-services [13] constitute one of the fastest growing areas of software engineering. With a strong support for compositionality, the process of developing an application essentially becomes a process of composing available services. Services – the basic building blocks of complex applications are platform and network independent components implementing computations that can be invoked by clients or other services.

To enable a rapid service composition, services define their properties in a standard and machine readable format. It enables service discovery, selection and binding. Service composition introduces the orchestration of the basic services to build applications. However, usually research on service orchestration focuses on defining the language for service composition that does not support reasoning about such essential features as fault tolerance. Such reasoning can be supported by dependability analysis and architectural modelling [5].

In this paper, we propose a systematic approach to architecting fault tolerant services. We demonstrate how to graphically model the architecture of composite services and augment it with various fault tolerance mechanisms. We propose static and dynamic solutions for introducing fault tolerance into the service composition. The structural solutions rely on availability of redundant service providers that can be requested to provide services in case of failures of the main service providers. This mechanism allows the designers to mask failures of the individual service providers. The dynamic solutions rely on re-execution of failed services to recover from the transient faults of services. This solution requires modifications of the service execution flow.

To facilitate design of complex fault tolerant services, in this paper, we introduce a systematic approach to analysing possible failure modes of services and defining fault tolerance measures. Our approach is inductive – it progressively analyses one component after another in the service execution flow, explores possible fault tolerance alternatives and systematically introduces them into the service architecture.

We believe that our approach supports structured guided reasoning about fault tolerance and enables efficient exploration of the design space. It allows the designers to evaluate various architectural solutions at the design stage that helps to derive clean architectures and improve fault tolerance of developed complex services.

The paper is structured as follows: in Section II, we demonstrate how to model a fault tolerant service from a service user's perspective. In Section III, we demonstrate how to unfold service architecture, i.e., explicitly represent the service composition and the service execution flow. We also propose different fault tolerance mechanisms that can be introduced to enhance fault tolerance. In Section IV, we introduce a structured approach to designing a fault tolerant architecture. Finally, in Section V, we overview the related work and discuss the presented work.

II. ABSTRACT MODELING OF FAULT-TOLERANT SERVICES

The main goal of introducing fault tolerance in the service architecture is to prevent a propagation of faults to the service interface level, i.e., to avoid a service failure [7] [9]. A fault manifests itself as *error* – an incorrect service

 ISBN: 978-1-61208-361-2

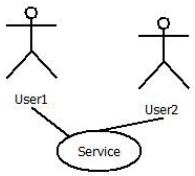


Figure1. Use case representation of a service.

state [9]. Once an error is detected, an error recovery should be initiated. Error recovery is an attempt to restore a fault-free state or at least to preclude system failure.

Error recovery aims at masking error occurrence or ensuring deterministic failure behaviour if the error cannot be masked. In the former case, upon detection of error, software executes certain actions to restore a fault-free system states and then guarantee normal service provisioning. In the latter case, the service provisioning is aborted and failure response is returned.

In this paper, we focus on the architectural graphical modelling [12] of fault tolerant services [13]. We demonstrate how to explicitly introduce handling of faulty behaviour into the service architecture. We follow the model-driven development paradigm and start our modelling from a high level of abstraction [8]. The consecutive model transformations introduce the detailed representation of the service architecture.

The high-level model of a fault tolerant service is given in Fig.1. The service is defined via its interactions with different service users. Each association connecting an external user and a service corresponds to a logical interface, as shown in Fig.2. The logical interfaces are attached to the class with ports. At the abstract modelling level, we treat a service as a black box with the defined logical interfaces.

The UML2 interfaces *I_ToService* and *I_FromService* define the request and request parameters of the service user. We formally describe the communication between a service and its user(s) in the *I_Communication* state machine as illustrated in Fig.3. The request *ser_req* received from the user is always replied: with the *ser_cnf* in case of success, with the *ser_fail_cnf* in case of unrecoverable failure and with the *ser_tfail_cnf* in case of a recoverable failure. Let us point out, that already at the abstract level of modelling, we explicitly introduce representation of faulty behaviour and reaction on it.

To exemplify an abstract modelling of a fault tolerant service, let us consider a *positioning service*. It provides the services for calculating the physical location of the service user.

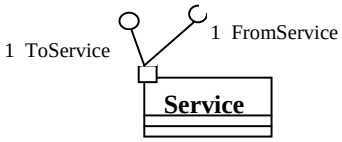


Figure 2. Abstract architectural diagram.

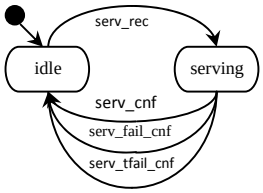


Figure 3. State diagram of communication.

As shown in Fig.4, the abstract model represents an interaction of the service with a user. An abstract architectural diagram defines an interface for communicating with the user. The state diagram formally defines the communication between the user and the service.

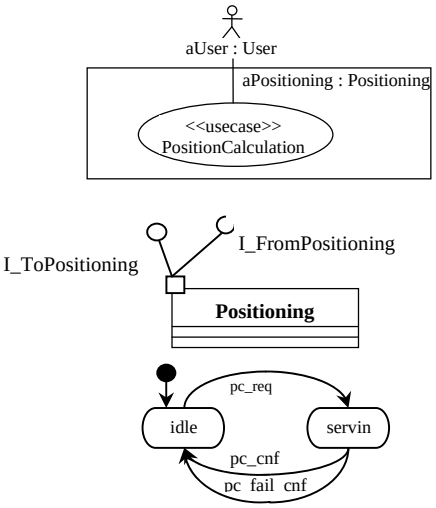


Fig.4. Modelling positioning service

The request to calculate the position is modelled by the event *pc_req*. In case of a normal execution, the positioning service returns the reply *pc_cnf*. Let us observe, that in our modelling we explicitly define the possibility of a service failure following the pattern proposed above. Indeed, in case of the unrecoverable failure, the positioning service returns *pc_fail_cnf*. In case of a recoverable failure, the service returns *pc_tfail_cnf*. Such a fault-tolerance explicit approach to modelling ensures that the service execution always terminates, i.e., the service never becomes unresponsive.

III. ARCHITECTURAL DECOMPOSITION

Our abstract modelling has defined the service from the service user's point of view. The model transformation presented next focuses on defining the composition that constitutes the overall service.

An execution of a composite service consists of executing several subservices. Coordination of a service execution is performed by a *service manager* (sometimes

 ISBN: 978-1-61208-361-2

called *service composer*). It is a dedicated software component that on the one hand, communicates with a service user and on the other hand, orchestrates the service execution flow.

To coordinate service execution, the service manager keeps the information about subservices and their execution order. It requests the corresponding service components to provide the required subservices and monitors the results of their execution.

Let us note, that any subservice might also be composed of several subservices, i.e., in its turn, the subservice execution might be orchestrated by its (sub)service manager. Hence, in general, a composite service might have several layers of hierarchy [5].

To model a composite service, we introduce the providers of the subservices into the abstract architectural service model. The model includes the external service

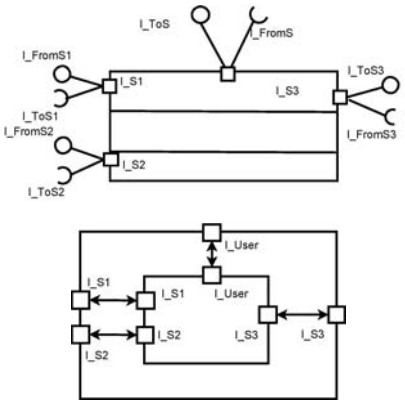


Figure 5. Architecture of a positioning service.

providers communicating with the aggregated service via their service director. For each association between the main service and the corresponding subservice, we define a logical interface. The logical interfaces are attached to the corresponding classes via the corresponding ports. This enables a structured representation of the modular structure of the composite service. The functional architecture is defined in terms of the service components, which encapsulate the functionality related to a single execution stage of another logical piece of functionality.

The architectural diagram of the position calculation [5] [14] – the composite service example described above is presented in Fig. 5. The service manager role is two-fold: it orchestrates service execution flow and handles communication with the service user. The dynamics of the execution flow is refined by introducing the corresponding sub-states in the service state as shown in Fig. 6.

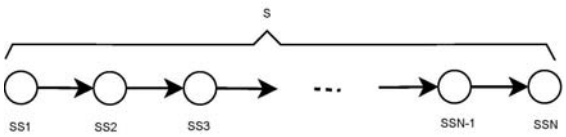


Figure 6. Unfolded dynamic behaviour.

Now, let us discuss the fault tolerant aspect of the composite services. Execution of any subservice can fail. To ensure fault tolerance of composite services, we propose a two-fold approach. On the one hand, we define a set of patterns [11] that allow us to introduce structural means for fault tolerance using various forms of redundancy. On the other hand, we propose to extend the responsibilities of a service manager, to implement dynamic error recovery. Next, we propose the architectural patterns for introducing structural fault tolerance and define the corresponding modeling artifacts.

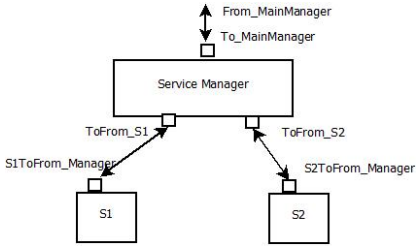


Figure 7. Duplication scheme.

Duplication pattern. The duplication is a simplest arrangement for structural fault tolerance. It can be introduced if there are two service components available that provide the same functionality. In this case, the services can be executed in parallel. A successful execution of a service by any out of two service components suffices for the successful service provisioning.

An architectural diagram of the duplication arrangement is given in Fig. 7. We introduce a dedicated service manager to take care of the execution of the duplicated service. The dynamical behavior of the duplication pattern is shown in Fig. 8. An alternative architectural approach would be to allow the main service manager to orchestrate this arrangement.

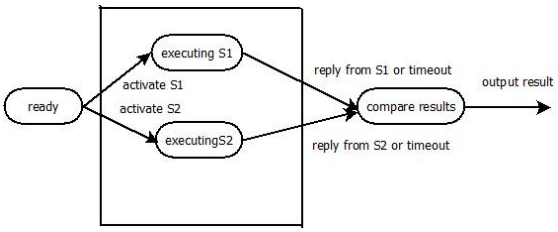


Figure 8. Dynamic behavior of duplication pattern.

Stand-by spare. This arrangement relies on availability of a spare service component implementing the desirable service. The spare is used only if the execution of the service by the main component fails. If the main service component succeeds in executing a service, the spare service component remains inactive. However, if the main service component fails to execute a service then the spare service component is requested to provide the service.

The stand-by spare arrangement can be implemented with and without an introduction of the dedicated service

 ISBN: 978-1-61208-361-2

director. The design decision depends on the complexity of the composite service, i.e., whether the design of the main service manager would become too complex with the introduction of this additional responsibility.

The architecture of the stand-by-spare implemented with the dedicated service manager coincides with the duplication pattern. However, the dynamic behavior is different as shown in Fig.9.

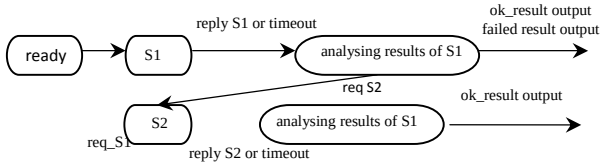


Figure 9. Dynamic behavior of stand-by spare.

Triple modular redundancy pattern. A more complicated scheme for structural redundancy – triple modular redundancy is shown in Fig.10. The precondition for implementing it is that we have three service components available that provide identical services with the same functionality. All three service components receive the same service request and work in parallel. The results of the service execution are sent to a voting element.

The voting element is a dedicated software component that performs comparison of the results and produces the final result. The voting element takes a majority view over the produced results of the successfully executed services and outputs it as the final result of the service execution.

In the context of the service-oriented computing, the voting component might be implemented in two different ways: it might output the results after receiving the first two replies or it might start to act only after the certain deadline when all non-failed services have replied.

Let us discuss a difference between triple modular redundancy scheme adopted in hardware and services. In hardware context, the scheme can mask failure of a single component by adopting the majority view. In the service-oriented context, it gives more fault tolerance options. Indeed, if two out of three services failed to reply within the timeout, the voter component can be design to simply output the result of the non-failed service. Obviously, in case of a failure of a single service, it gives better fault tolerance guarantees, because it can compare the results of two non-failed services and take the one, which is more accurate as the output.

Since the triple modular redundancy scheme has a rather complex architecture by itself, we propose to introduce a dedicated service manager to integrate the arrangement in the architecture of a composite service. The proposal is depicted in Fig. 10.

The dynamic behavior of the triple modular arrangement is depicted in Fig.11. Here, the dedicated service manager performs voting before outputting the service result.

The static redundancy schemes require availability of redundant service components and hence, sometimes,

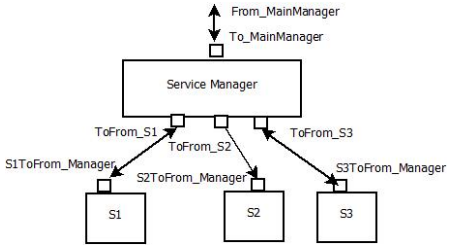


Figure 10. Architecture of triple modular redundancy.

might be non-implementable. However, they provide an efficient means to cope with permanent service failure. In contrast, dynamic fault tolerance relies on service re-execu-

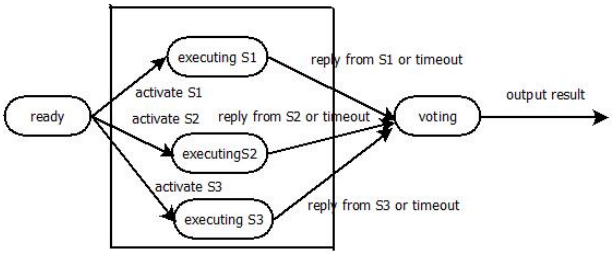


Figure 11. Dynamic behaviour of triple modular redundancy.

tion to increase the chances of the successful service execution and does not require an availability of the redundant service components. Obviously, the dynamic fault tolerance solutions can cope with transient failures.

To leverage fault tolerance of a composite service, the service manager might alter the normal flow of service execution to dynamically cope with failures. For instance, it might repeat service execution, roll-back or abort service execution.

If service execution failed, but the returned exception indicates that the error is transient then by re-executing the failed subservice, the service manager might recover from the error. The service execution flow is shown in Fig.12.

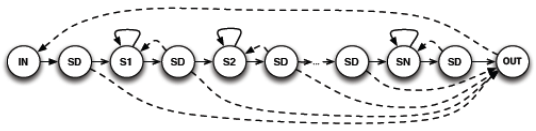


Figure 12. Service execution flow.

If service execution failed but the returned exception indicates that the error is unrecoverable and there are no alternative services available, then the service manager can abort the entire service execution and return failure response.

Obviously, designing fault tolerant composite services is a non-trivial task that requires a systematic support. In the next section, we propose an approach to systematic development of fault tolerant architecture by a structured

 ISBN: 978-1-61208-361-2

analysis of failure modes of the services and fault tolerance schemes.

IV. DEVELOPMENT OF A FAULT TOLERANT SERVICE ARCHITECTURE

The main motivation behind our approach is to facilitate a structured disciplined derivation of fault tolerant service architecture. Essentially, we define the guidelines for analyzing faulty behavior of the services and deciding on the mechanisms for fault tolerance.

Our approach is inspired by the Failure Modes and Effect Analysis (FMEA) technique. FMEA [16] is an inductive analysis method, which allows designers to systematically study the causes of components faults, their effects and means to cope with these faults. FMEA is used to assess the effects of each failure mode of a component on the various functions of the system as well as to identify the failure modes significantly affecting dependability of the system.

FMEA step-by-step selects the individual components of the system, identifies possible causes of each failure mode, assesses consequences and suggests remedial actions. The results of FMEA are usually represented in the tabular form that contains the following fields: component name, failure mode, possible cause, local effect, system effect, detection, and remedial action.

Let us exemplify the proposed approach. Assume that a service *S1* is a part of the composite service *S*. The services *S11* and *S12* have identical functionality. Assume that the service *S1* might experience transient silent failures, i.e., become temporally irresponsive. Such failures can be detected by timeout. Then we can arrange services into a triple modular redundancy scheme. The structured analysis of the fault tolerance arrangement around the service *S1* according to the proposed approach is shown in Table I.

TABLE I. TRANSIENT FAILURE ANALYSIS

Service	S1
Failure mode	Transient silent failure
Detection	Timeout
Available redundancy	S11, S12
Structural redundancy	Triple modular redundancy arrangement. Result is produced upon timeout
Recovery	Masking failure by use of triple modular redundancy arrangement. In case of simultaneous failure of S1, S11 and S12 repeat execution

Let us now assume that a service *S2* is also part of the composite service *S*. Assume that the service *S2* might experience transient failures that are identified by receiving the response *S2_tfail_cnf* from it. Since no redundant service components are available for this case and the service failure is detectable with the corresponding notification, we can rely on dynamic redundancy to cope with failures of S2. The structured analysis of the fault tolerance arrangement around the service *S2* according to the proposed approach presented in Table II.

TABLE II. FAILURE MODE ANALYSIS

Service	S2
Failure mode	Transient detectable failure
Detection	*S2_tfail_cnf* response
Available redundancy	No
Structural redundancy	No
Recovery	Re-execute service. Maximal allowed number of retries is 3.

It easy to observe that reliance on the proposed approach facilitates structured derivation of fault tolerance architecture for both structured and dynamic fault tolerance schemes.

As a result of introducing various means for fault tolerance, we also should modify the design of the service manager. Fig 13 depicts the modified flow with a retry.

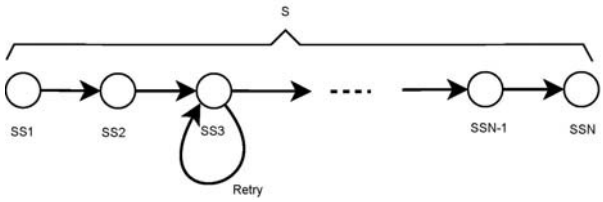


Figure 13. Execution flow with retry.

The process of introducing fault tolerance mechanisms can be iteratively applied to unfold all the architectural layers. As a result of this process, we obtain a hierarchical structure of service managers augmented with fault tolerance properties.

V. RELATED WORK AND CONCLUSIONS

While the topic of service orchestration and composition has received significant research attention, the fault tolerance aspect is not so well addressed. Liang [10] proposes a fault-tolerant web service on SOAP (called FT-SOAP) using the service approach. It extends the standard WSDL by proposing a new element to describe the replicated web services. The client side SOAP engine searches for the next available backup from the group WSDL and redirects the request to the replica if the primary server failed. It is a rather complex mechanism that hinders interoperability.

Artix [2] is IONA's Web services integration product. It provides a WSDL-based naming service by Artix Locator. Multiple instances of the same service can be registered under the same name with an Artix Locator.

 ISBN: 978-1-61208-361-2

When service consumers request a service, the Artix Locator selects the service instance based on a load-balancing algorithm from the pool of service instances. It provides useable services for the service consumers. An active UDDI mechanism [4] enables an extension of UDDI's invocation API to enable fault-tolerant and dynamic service invocation. Its function is similar to the Artix Locator. A dependable Web services framework is proposed in [1]. Once a failure for one specific service occurs, the proxy raises a "WebServiceNotFound" exception and downloads its handler from DeW. The exception handling chooses another location that hosts the same service and re-invoks the method automatically. The main goal of DeW is to realize physical-location-independence. Providing fault-tolerance capability for composite Web service has also been discussed in [3].

A formal approach [15] [17] to introducing fault tolerance to the service architecture has been proposed in [5] [6]. This work extends the set of architectural patterns that can be introduced to achieve fault tolerance as well as propose a systematic support for deriving fault tolerance solutions.

In this paper, we have proposed a systematic approach to architecting fault tolerant services. We demonstrated how to graphically model the architecture of composite services and augment it with various fault tolerance mechanisms. We defined a set of static and dynamic solutions for introducing fault tolerance into the service composition. The proposed mechanisms can cope with different types of failures to increase reliability of complex composite services.

To facilitate design of fault tolerance mechanisms, we proposed an approach to a structured analysis of possible failure modes of services and introducing fault tolerance measures. The proposed approach is inductive – it progressively analyses services in the execution flow, explores possible fault tolerance alternatives and systematically introduces them into the service architecture.

We believe that our approach supports structured guided reasoning about fault tolerance and enables efficient exploration of the design space while architecting complex composite services.

ACKNOWLEDGMENT

Troubitsyna thanks the Need for Speed program. http://www.digile.fi/N4S for a financial support.

REFERENCES

[1] E. Alwagait, S. Ghandeharizadeh, "A Dependable Web Services Framework" 14th International Workshop on Research Issues on Data Engineering 2004. [Online]. Available from http://fac.ksu.edu.sa/alwagait/publication/31143 2014.30.05.

[2] Artix Technical Brief. [Online]. Available from http://www.iona.com/artix 2014.30.05.

[3] V. Dialani, S. Miles, L.Moreau, D. Roure, M. Dialani, "Transparent fault tolerance for web services based architectures". 8th Europar Conference (EULRO-PAR02), Springer 2002, pp. 889-898. ISBN:3-540-44049-6.

[4] M. Jeckle, B. Zengler, "Active UDDI-An Extension to UDDI for Dynamic and Fault Tolerant Service Invocation" 2nd International Workshop on Web and Databases, Springer 2002, pp. 91-99. ISBN:3-540-00745-8.

[5] L. Laibinis, E. Troubitsyna and S. Leppänen, "Service-Oriented Development of Fault Tolerant Communicating Systems: Refinement Approach" International Journal on Embedded and Real-Time Communication Systems, vol. 1, pp. 61-85, Oct. 2010, DOI: 10.4018/jertcs.2010040104.

[6] L. Laibinis, E. Troubitsyna, A. Iliasov, A. Romanovsky, "Rigorous development of fault-tolerant agent systems", In in M. Butler, C. Jones, A. Romanovsky and E. Troubitsyna (Eds.), Rigorous Development of Complex Fault-Tolerant Systems, LNCS 4157, pp. 241-260, Springer 2006, ISBN 978-3-642-00867-2.

[7] L. Laibinis, E. Troubitsyna, S. Leppänen, J. Lilius and Q. Malik, "Formal Service-Oriented Development of Fault Tolerant Communicating Systems", in M. Butler, C. Jones, A. Romanovsky and E. Troubitsyna (Eds.), Rigorous Development of Complex Fault-Tolerant Systems, LNCS 4157, pp. 261-287, Springer 2006, ISBN 978-3-642-00867-2.

[8] L. Laibinis, E. Troubitsyna "Fault Tolerance in use-case modelling", In Workshop on Requirements for High Assurance Systems (RHAS 05), [Online]. Available from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.84.4950 2014.01.05.

[9] J. C. Laprie. Dependability: Basic Concepts and Terminology. Springer-Verlag, 1991.

[10] D. Liang, C. L. Fang, C. Chen, F. X, Lin. "Fault-tolerant web service". Tenth Asia-Pacific Software Engineering Conference, IEEE Press, Dec. 2003, pp.56-61, ISBN 973-4-642-01867-1.

[11] I. Lopatkin, A. Iliasov, A. Romanovsky, Y. Prokhorova, E. Troubitsyna, "Patterns for representing FMEA in formal specification of control systems" High-Assurance Systems Engineering Conference (HASE), IEEE Nov 2011, pp. 146 – 151, ISBN 978-1-4673-0107-7.

[12] J. Rumbaugh, I. Jakobson, and G .Booch, The Unified Modelling Language Reference Manual. Addison-Wesley, 1998.

[13] Web Services Architecture Requirements. [Online] Available from http://www.w3.org/TR/wsareqs. 2014.01.05.

[14] 3GPP. Technical specification 25.305: Stage 2 functional specification of UE positioning in UTRAN. Available at http://www.3gpp.org/ftp/Specs/html-info/25305.htm. Accessed 01.05.2014.

[15] K. Sere, E. Troubitsyna, "Safety analysis in formal specification" In Formal Methods (FM'1999), Springer Sep. 1999, pp 1564-1583, ISBN:3-540-66588-9.

[16] N. Storey. Safety-critical computer systems. Addison-Wesley, 1996.

[17] E. Troubitsyna. "Elicitation and specification of safety requirements". In Third International Conference on Systems (ICONS 08), IEEE Apr. 2008, pp. 202-207, ISBN978-0-7695-3105-2.

 ISBN: 978-1-61208-361-2

Scalable Web Content Understanding Framework

Yang Sun, Hyungsik Shin, Sayandev Mukherjee, Ronald Sujithan, Hongfeng Yin, Yoshikazu Akinaga and Pero Subasic

DOCOMO Innovations, Inc.
Emails: {ysun, hshin, smukherjee, rsujithan, hyin, akinaga, psubasic}@docomoinnovations.com

***Abstract*—The contextualization of an unknown web page is a fundamental need in many online applications. We propose a new framework known as the Content Understanding Engine (CUE) that allows computational stages to be composed with different technologies to contextualize an unknown URL. We describe how this computation pipeline interfaces with our Big Data infrastructure and how this approach simplifies deployment to private or public cloud environments. The implementation details of this framework are provided along with a use case to demonstrate the value of the CUE. We provide the results from our evaluation of this pipelined architecture with a wide range of URL from different topics.**

Keywords–contextual tagging; advertising; content understanding engine.

I. INTRODUCTION

Many online applications heavily rely on automated systems to analyze the context of web pages. These contextual systems are typically required to take URLs, fetch web pages, parse content, extract keywords, classify text and find the most relevant tags. The resulting contextual profiles will not only present the opportunities for advertisement matching, but also unveil personal preferences, interests and trending concepts [1][2]. For example, to optimize advertising campaigns, advertisers often develop models based on analyzing historical contextual consumption of users and then match product offerings to specific contextual profiles. Contextual advertising systems also rely on contextual information to match advertisement with web pages.

In practice, a number of different techniques drawn from diverse fields, such as Text Mining, Natural Language Processing, Machine Learning and Information Retrieval need to be combined into a pipelined architecture to meet these requirements. However, current frameworks do not handle the whole process from URL crawling to the semantic analysis. Our proposed framework integrates existing technologies and handles the end to end complexity of the contextual profiling process.

Many components are required to build a scalable and useful contextual profiling system, including (1) a scalable web content fetching module for billions of URLs, (2) a fast web page parsing and ad removing module, (3) a scalable document storage and processing module, (4) a content understanding module to extract and tag web pages with brief and representative text, and (5) a tag generation module that finds the most relevant tags from the representative text. Each module has unique requirements in terms of response time, scalability and accuracy.

Our proposed system framework has the following contributions to the community:

- A Software Architecture for a scalable computation pipeline that can be deployed to any private or public cloud infrastructure;
- A staged architecture, allowing the use of different technologies for each stage, and an interface with big data infrastructure for truly large scale processing;
- A Wikipedia-based contextual tagging solution that reflects the latest events and trending concepts; and
- A uniform way of communicating the contextual tags to all players involved in the marketing process.

The rest of this paper is organized as follows. In Section II, we describe the related work done in the areas of distributed workflow engines and prior research on Contextual Profiling Systems. In Section III, we describe the stages of our Content Understanding Framework in detail. In Section IV, we briefly describe the cloud deployment of our Content Understanding Engine (CUE). In Section V, we describe a compelling use case for the CUE and discuss evaluation results. Finally, we state our conclusions and describe further work.

II. RELATED WORK

The contextualization of the contents of any web page, including content tagging using a controlled vocabulary, is a fundamental task in many online applications. Existing approaches focus on three aspects of the context understanding problem: (1) the free-text approach – using free text labels to tag articles [2][3]; (2) the classification approach – classifying articles to a well-organized hierarchical taxonomy of topics [1][4]; and (3) the semantic approach – using semantic analysis to determine advertising needs [5]. Unfortunately, none of these by itself is suited to our task, as discussed below.

The first approach summarizes articles with free text that is rich enough to represent the meaning as well as abstract enough to fit to specific applications. The feature space of free text has dimension in the millions. Therefore, it is typically difficult for advertising systems to use. The second approach maps complex concepts to well-structured categories. These categories typically have very general terms, so that specifics of the articles are not represented in the approach. The third approach, semantic analysis, is an evolving field that has potential, but does not at present provide a mature solution to the content understanding problem.

 ISBN: 978-1-61208-361-2

A. Principal Challenges

There are several challenges in the domain of content understanding systems: (1) such systems need to function well in the presence of "noise", such as spelling and grammatical errors, different languages, markup errors, handling boilerplate content, etc.; (2) they need to create features from the extracted text so that the features represent the original content in a satisfactory form; (3) they need to map the representative content extracted from the page to a set of known vocabulary terms; and (4) they need to be able to scale and deal with petabyte-scale volumes on model cloud infrastructures such as the now-ubiquitous Amazon Web Services.

In order to address these challenges, many different technologies are used in practice. These technologies arise from many different related fields, such as Text Mining, Natural Language Processing, Machine Learning and Information Retrieval. A number of pipelined architectures have been developed to solve the problem of applying these different technologies to solve the end-to-end problem.

Our work is closely related to work on web usage mining, automatic discovery and analysis of patterns in click streams, user transactions, and other associated data collected or generated as a result of user interactions with Web resources [6][7][8]. In particular, our work is related to web mining systems that use other sources of knowledge: either semantic domain knowledge from ontologies (such as product catalogs, concepts and categories) or a more generic knowledge base such as the freely-available Wikipedia concepts and categories [9]. However, the existing approaches focus primarily on challenges 1, 2 and 3, but do not address challenge 4 as an integral part of the solution. While building on this early work on web mining, we needed a workflow solution that can scale to process petabyte-scale volumes that can be deployed and operated as a cloud-based service.

B. Existing Technologies and Frameworks

Hadoop [10], the open source implementation of the MapReduce framework, has become the dominant environment for building an architecture for solving the scalability problems described in the previous section. In particular, systems such as Oozie [11] and Azkaban [12] provide a workflow abstraction on top of Hadoop. Both support defining a workflow as a Directed Acyclic Graph (DAG) [13] made up of a composition of individual steps. In Azkaban the job type, any parameters, and any dependencies are specified. However, Azkaban does not have any notion of a self-contained workflow, so a job can depend on any other job in the system. In Oozie, a workflow is defined in an XML file, which specifies a start action. However, Oozie is tightly coupled with Hadoop and HDFS, thereby making it harder to deploy computational stages that use other technologies. Cascading [14] is a popular workflow engine for building flexible enterprise data processing solutions without having to worry about how to distribute the workload.

Recently, real-time workflow engines that overcome the batch oriented nature of Hadoop, such as Storm [15] and Spark [16], have become very popular. Storm is a distributed realtime computation system that provides a set of general primitives, Spouts and Bolts, for composing computation stages. However, Storm flow is based on individual items flowing through the system as a stream. Spark is a MapReduce-like cluster computing framework designed to support low-latency iterative computations. Spark aims to overcome the batch-oriented nature of Hadoop by distributing the data in slices and storing it in memory, thereby gaining a significant performance boost. However, Spark maintains a tight coupling with Hadoop and HDFS, making the integration of non-Hadoop distributed systems non-trivial.

C. Why we design our own framework

After evaluating the frameworks mentioned above, we decided to develop our own content understanding engine from first principles for a number of reasons. Firstly, our requirements are such that we need a flexible architecture to combine very different technologies into a uniform pipeline. Secondly, we need the ability to produce detailed output at each stage and carefully study the results. Thirdly, we need the ability to substitute a completely different technology for a given stage without impacting other stages or the final results. Finally, we wanted to leverage best-of-breed Open Source technologies at each stage so that we can focus on the broader solution we need to build.

III. FRAMEWORK

A. Requirements

The keys to the web content fetching module are flexibility and scalability. The module is typically required to fetch content for millions or even billions of URLs. It has to be flexible enough to handle many exception cases such as invalid URL, redirects, connection timeout, lost connection and so on. It also has to be scalable at the same time in order to fetch from billions of URLs in a reasonable time period.

The key requirements for the web page parsing module and document storage module are scalability, processing speed and accuracy. With billions of web pages, the parsing module and document storage module have to be able to process web pages fast and be scalable at the same time. To process 1 billion web pages with 100 machines where each machine can process a page in 100 ms, the system needs more than 11 days to finish. The parsing module also has to be very accurate in extracting core content from the web page markup to eliminate irrelevant tags, scripts, and ads.

Algorithms and methods play a key role in representing complex articles via short text labels. Web pages are typically written or edited with long natural language text for readability. Contextual profiling and advertising systems cannot directly utilize the web page content because it contains many HTML and scripting tags and commonly used words that do not contribute to the core meaning of the page. Content extraction and summarization systems typically scrape the web pages and convert the natural language text into much shorter words and phrases.

In contextual advertising, marketers typically prefer well-defined categories over free text for reasons of communication, consistency and measurement. Well-defined contextual elements, such as tags and categories, are easy to communicate from marketers to advertising operators. It is also easy to measure the campaign complexity and potential gain and cost with well-organized categories and tags. However, predefined category hierarchies cannot be changed dynamically as new concepts become available. It is also hard for category structures to capture the meaning of articles. On the other hand, free

 ISBN: 978-1-61208-361-2

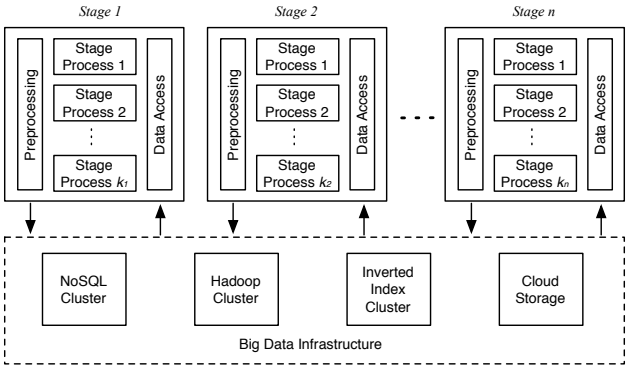


Figure 1. The Framework: separation of Computational Workflow from the Big Data Infrastructure allowing each module to scale separately in a cloud environment.

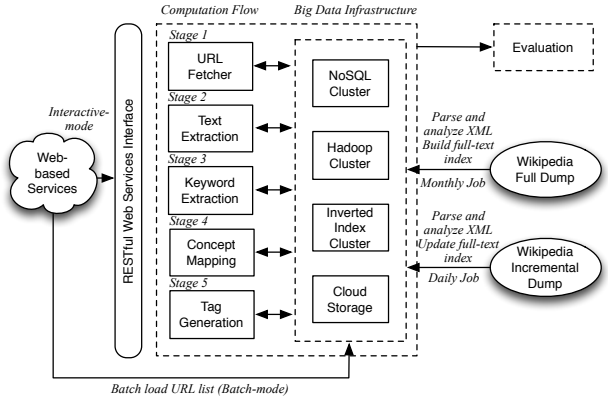


Figure 2. The Content Understanding Engine: internal modules and interfaces.

text summarization provides richer contextual information than categories. However, the huge dimension size makes it harder for advertisers to prepare marketing strategies.

B. Architecture

To balance the granularity of information representation and the ease of management, we propose a Wikipedia based concept system to extract and match contextual tags from news articles to the most relevant Wikipedia categories and concepts. Instead of free text summaries, the CUE generates a set of tags that map the contextual meanings of news articles to user-defined and user-maintained concepts and categories.

We now describe our architecture in detail. Our architecture separates computation-tier from the data-tier. As shown in Fig. 1, the computation-tier consists of a sequence of stages where each stage takes a predefined set of inputs and produces some predefined output. Each stage can be configured to interface with a data-tier cluster to access data-at-rest as well as to distribute large-scale data crunching. Our Heterogeneous Big Data Infrastructure allows us to use different technologies, such as NoSQL Database (e.g., HBase [17], Hadoop, Inverted Index (Solr [18] or ElasticSearch [19]) and Cloud Storage.

The same pipeline can be run in batch-mode (a list of URLs), interactive-mode (one URL in near realtime) or service-mode. In service-mode, a RESTful Web Service interface is provided so that a run can be initiated posting a URL. When initializing the pipeline, a configuration can be provided to specify the sequence of stages that need to be run and the additional parameters needed by each stage. This framework allows us to perform comparative evaluation of different technologies and to understand their relative merits.

The CUE is architected from first principles to address our specific needs. Our approach allows us to combine disparate technologies into a unified pipeline as a sequence of stages with clearly defined input and output for each stage. The first stage in our pipeline is a custom Fetcher that can take a list of URLs and fetch the HTML content of the web page. In the next stage, we extract the main textual content from the web page using a Support Vector Machine (SVM) model trained from many different kinds of web pages (e.g., blogs, news articles, product reviews, etc.). Once the plain text is extracted, we perform sentence and phrase detection followed by keyword extraction in the next stage.

As part of this project, we have built an inverted index of the Wikipedia content, including the titles (concepts) and categories. Thus, in the next stage, we search the inverted index using the extracted keywords (and phrases) from the previous stage. This provides us with the list of Wikipedia concepts and categories related to the original content of the web page. In the final stage, we propagate the relevance scores attached to the concepts through the category hierarchy and find the most relevant categories representative of the original article. Next, we describe the stages of the pipeline in detail.

C. Web Service Interface

As shown in Fig. 2, the CUE offers a RESTful Web Service interface so that it can be integrated into other services. This service can be invoked either in interactive mode or bulk mode. Several applications as mentioned before can be integrated with this service to gain a contextual understanding of a web page. In interactive mode, the service provides a browser interface to view the detailed output from each stage of the pipeline. In bulk mode, a batch of URLs (e.g., from a weblog) can be posted to the service to be processed in the background. For processing very large URL batches, a file location can be provided as input. Files can be local, network-mounted or cloud-based (e.g., Amazon S3). With the latter approach, a large URL batch file can be uploaded to a cloud storage first and CUE can be requested to process this file and send the final output to another local, network-mounted or cloud-based file system.

D. URL Fetcher

This stage (Stage 1 in Fig. 2) takes a list of URL as input and stores the contents of that page in original form including the HTML or other markup present in the document. The fetcher can be configured to deal with the complexities involved in crawling a web page (such as setting the user agent, dealing with scripts, images, etc.). We use the well known techniques for fetching the contents given a set of URLs [6][20]. The harvested raw web content is stored in the NoSQL cluster with the URL plus timestamp as a key and the raw content as the value.

E. Text Extraction

This is an important stage (Stage 2 in Fig. 2) in the pipeline where we wanted to evaluate and use different

technologies for extracting text from the downloaded page content from the Fetcher. After evaluating several solutions in this space [6][21][22], we settled on the open source project Boilerpipe [23] since this library yielded the best results for our purposes. The Boilerpipe library provides us a way to strip out all the boilerplate content, such as irrelevant tags, scripts, and ads from the web page and extract only the main content (e.g., the main article of a news page) with high accuracy [24]. This library, under the covers, uses an SVM classifier trained from thousands of web pages to extract the main content of a web page.

F. Keyword Extraction

This stage (Stage 3 in Fig. 2) performs sentence detection, tokenization, phrase detection and keywords extraction. First, we segment the extracted text from the previous stage into sentences by applying several rules to detect the sentence boundary. Then we tokenize each sentence and detect the most frequent one to three word phrases. We experimented with several ways of finding the significant phrases from the extracted text including TF-IDF scores. From this sorted list of phrases, we select the top N (with $N = 20$, say) as the extracted keywords from the original text document. This stage can be configured with the rules for sentence segmentation, stemming or stopword removal [25].

G. Concept Mapping

This stage (Stage 4 in Fig. 2) constructs a query based on the extracted keywords and performs a full-text search. We utilize the Open Source search engine ElasticSearch that uses the Lucene [26] toolkit to build a full-text index of a document collection. Using the Wikipedia Plugin provided by ElasticSearch we have built a full-text index of the entire Wikipedia content. This provides us the ability to search a set of keywords and find the matching Wikipedia Articles and retrieve the Concepts (Titles) and Categories with the highest scores. The underlying Lucene Engine provides scores for the matching concepts that we use as a way of ranking the results. We apply a boosting factor to keyword matches in article titles in order to improve the accuracy of the search results.

H. Tag Generation

This stage (Stage 5 in Fig. 2) finds the related higher-level Wikipedia Categories from the highest ranked Wikipedia concepts from the previous stage. The Wikipedia categories are first cleaned to remove categories that do not represent a meaningful concept (e.g., lists of famous dead people) and to remove cycles. We also perform case normalization since categories for the same concept are present with different capitalization (e.g., *Computer science* and *Computer Science*). After these preprocessing steps, we get a clean graph representation of the Wikipedia category hierarchy.

With the concepts as the leaf-nodes of this graph structure, we propagate the Lucene scores associated with each concept to find the most significant categories that are present at the intersection of several concepts [9]. These categories will be taken as representative of the contents of the page, providing the semantic meaning of the original page.

I. Evaluation Module

Evaluation is the core component of the framework that bridges the automated system and the business applications. A flexible framework has to be able to support variety of evaluation methods. The architecture design of our framework sets evaluation module as a pluggable component which communicates with modules via RESTful services.

IV. Deployment

We deployed our Web Content Understanding Framework to different cloud environments and compared our experiences. For this part of our evaluation, we deployed our system on the following cloud services and conducted experiments with various workloads: (1) Internal Private Cloud, (2) Amazon Web Services, and (3) HP Cloud Services.

Our development environment is an internal private cloud and the computational stages are distributed as shown in Fig. 2. Each stage can be scaled by adding more nodes as necessary. For instance the ElasticSearch server used in the Concept Mapping stage is shared across four nodes to index the entire Wikipedia dump and search matching articles given a set of keywords. With this configuration we were able to reduce the search time to be < 100 ms. Moreover, we are able to deploy this architecture to an AWS cloud using medium-powered EC2 instances as well as to deploy to an HPCS cloud using a similar server configuration. We are able to deploy and be operational within a day and are able to scale both horizontally and vertically depending on the workload.

V. Use Case: Profiling Weblogs

In this section, we will describe a use case for the CUE - web usage mining to analyze the behavioral patterns and profiles of users interacting with web sites. With the continual growth and proliferation of mobile devices and the Internet of things, the volume of interaction logs generated by web-based service providers has reached several petabytes in size.

Analyzing such data can help organizations determine the life-time value of customers, design cross-marketing strategies, evaluate effectiveness of campaigns and personalize content to visitors. This type of analysis involves the automatic discovery of meaningful patterns and relationships from very large collection of weblogs collected from various operational data sources. Such weblogs typically contain a list of URLs accessed by each user along with information such as (1) the user's IP address, (2) the user's authentication name, (3) the date and time stamp of the access, (4) the HTTP request, (5) the response status, (6) the size of the requested resource, and optionally, (7) the referrer URL and (8) the user's browser identification.

Mobile service providers are collecting these logs from multiple sources. We are using CUE in several projects to enhance user experience with the goal of increasing customer retention and revenue (ARPU) (see Fig. 3). URL collections are processed through CUE to contextualize each URL with concepts and categories. This information can then be associated with user profiles to ascertain the interests of the users.

A. Evaluation

Evaluating the effectiveness of a contextual service is a difficult task. Unlike traditional evaluation of text classification

 ISBN: 978-1-61208-361-2

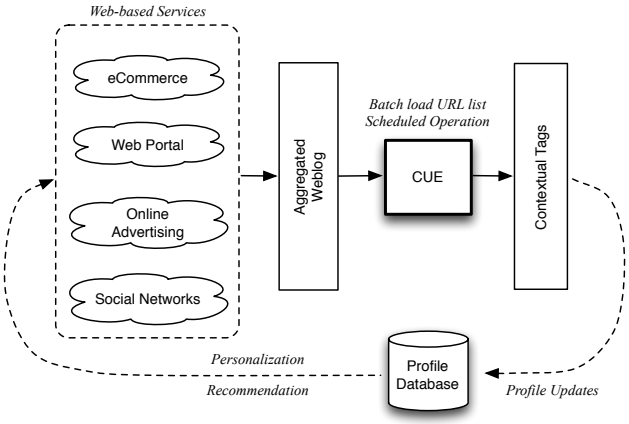


Figure 3. The Content Understanding Engine (CUE) in Context. The figure shows how CUE is deployed along with existing web based services.

systems where the metric is the accuracy of articles being classified to a predefined category, the effectiveness of a contextual service can be measured from different perspectives when they are used for different goals. Accuracy does not always translate to marketing strategies and goals directly. In digital advertising, marketing strategies and goals lead to different measurements of successfulness.

The effectiveness of the contextual profiling from web usage logs can be examined from different perspectives when the results are used in different applications. The contextual profiling of web users is typically used for online advertising. Advertisers can design marketing strategies based on users' content consumption and trends. For the advertising application use case, we implemented three evaluation modules, consistency, accuracy, and usefulness, to measure the effectiveness of the system.
We use three metrics to evaluate our CUE system:

1) *Consistency*, the ability of a tagging system to produce similar results for similar contents, is one of the most important measures in advertising applications. The wording of news articles can be significantly different for different news sources. The free text based approach may generate very different summaries for the same news content from different sources, while the classification based approach typically fails at capturing the details of the content and the model can rapidly become outdated.
We want the output of CUE conceptual tags of, for example, the "Crimea Crisis" news story from New York Times and from CNN to be similar, so that advertisers are assured that users on different websites are consuming the same content.
For a given topic, the consistency evaluation module collects news articles from several news sources. Label these articles $1, \ldots, K$, say. For the kth article, let A_k denote the set (actually, ranked list) of conceptual tags output by CUE from this article. The CUE outputs the same number n ($= 10$, say) of conceptual tags from each article, so $|A_k| = n$ for all k. For each topic, we compute two consistency metrics:
   a) The *overlap consistency* for each pair (i, j): the Sørensen-Dice similarity of A_i and A_j,
$$C_{i,j}^{\text{overlap}} = D(A_i, A_j) = \frac{2|A_i \cap A_j|}{|A_i| + |A_j|}. \quad (1)$$
The overlap consistency shows the degree of overlap of tags for a given topic.
   b) The *average rank correlation*: the average of the Kendall tau rank correlation coefficient $\tau(A_i, A_j)$ over all distinct tags for the topic,
$$\overline{C}^{\text{rank}} = \frac{\sum_{i=1}^{K-1} \sum_{j=i+1}^{K} \tau(A_i, A_j)}{|\cup_{k=1}^{K} A_k|}. \quad (2)$$
The average rank correlation shows the consistency of the ranking of tags for a topic.
2) *Accuracy* is defined as the degree of agreement between automatically generated tags and professional editors. This metric reflects how accurate the CUE tagging system is when processing news articles. We collect K Wikinews articles with editor-labeled conceptual tags and measure the accuracy as
$$\text{Accuracy} = \frac{1}{K} \sum_{k=1}^{K} D(A_k, B_k), \quad (3)$$
where for the kth article, A_k is the set of CUE tags, while B_k is the set of editor-assigned tags.
3) *Meaningfulness* is defined as how a general news reader agrees with the automatically generated tags. To measure the meaningfulness of the CUE system, a traditional expert rating evaluation module is attached to the system. The module takes the output of CUE system, randomly selects and ranks URLs and corresponding conceptual tags and then presents the results to experts. Experts rate each conceptual tag for the level of relevance to the article pointed by the corresponding URL. The rating results are summarized by the module and the relevance score will be presented automatically.

B. Experiments

To measure consistency, we use Google news as our data source. In Google news, articles from different news sources are grouped together. We collected 425 news topics with 2,571 articles. CUE system generated contextual tags for each articles. The consistency score distribution comparing to random scoring is shown in Fig. 4 and Fig. 5. The relatively low consistency scores are due to the strict matching rule implemented in the module where only the exact concept matchings are counted as success. For example, "variance" and "standard deviation" are considered not a match although they are conceptually close. More relaxed matching algorithms should get a higher score.

We also collected 278 articles from Wikinews with categories assigned by editors. Our CUE system has an average accuracy of 35.25% (see Eq. 3). Finally, 5 experts are hired to evaluate the meaningfulness of tags generated by CUE from 411 news articles. An average of 41% of tags are considered to be meaningful to the articles by the experts. The results show that the CUE system is fairly useful to identify key concepts despite the limitations of our evaluation methods. The accuracy and meaningfulness are greatly influenced by the strict

 ISBN: 978-1-61208-361-2

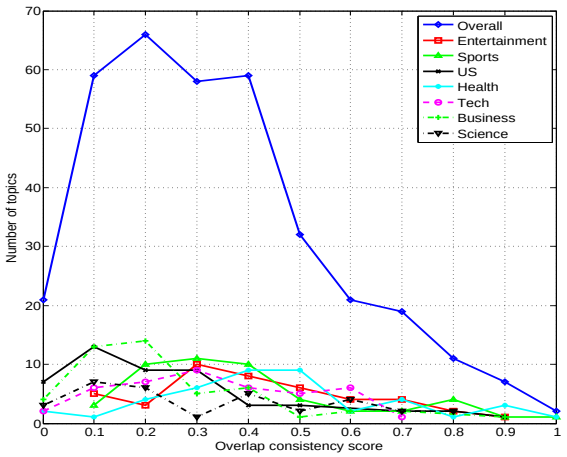


Figure 4. Overlap consistency score distributions for the top categories of news articles.

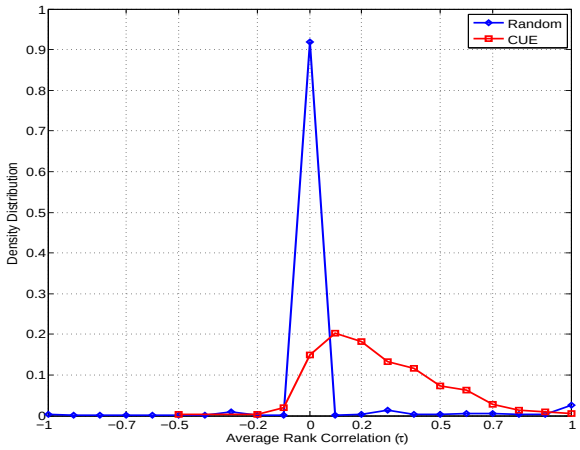


Figure 5. Rank correlation consistency distribution for CUE compared to randomly-assigned tags from top categories.

matching rule and binary judgments, where closely connected or similar tags are not considered a match. We are working on more fair and sophisticated matching algorithms that are appropriate for complex hierarchical matching problems.

VI. Conclusions and Future Work

This paper presented a scalable and modular architecture of a web Content Understanding Engine (CUE) for finding contextual tags for potentially large collection of URLs. There are two key aspects to the software framework presented in this paper: (1) computational flow that allows different technologies to be composed into a unified pipeline and (2) Big Data infrastructure that allows the use of different technologies such as NoSQL Database, Hadoop, Full-text Index or Cloud Storage. We are using CUE in several projects to enhance the user experience of web-based services with the goal of increased customer retention and incremental revenue. One specific example use of CUE is to generate trending concepts on a daily basis from URLs harvested from popular news aggregators and to leverage this information to produce better recommendations. Our further work is to explore alternative technologies for each stage of the computation flow to optimize the end-to-end scalability and performance of the CUE.

References

[1] A. Addis, G. Armano, and E. Vargiu, "Profiling users to perform contextual advertising," in Proc. 10th Workshop dagli Oggetti agli Agenti (WOA), Jul. 9–10, 2009, Parma, Italy, 2009, URL: http://cmt.math.unipr.it/woa09/papers/Addis2.pdf [accessed: 2014-05-01].

[2] G. Armano, A. Giuliani, and E. Vargiu, "Experimenting text summarization techniques for contextual advertising," in Proc. 2nd Italian Information Retrieval Workshop (IIR) Jan. 27–28, 2011, Milan, Italy, ser. CEUR Workshop Proceedings, vol. 704. CEUR-WS.org, 2011, Melucci, M., Mizzaro, S., and Pasi, G., Eds., ISSN: 1613-0073, URL: http://ceur-ws.org/Vol-704/12.pdf [accessed: 2014-05-01].

[3] G. Armano, A. Giuliani, A. Messina, M. Montagnuolo, and E. Vargiu, "Content-based keywords extraction and automatic advertisement associations to multimodal news aggregations," Studies in Computational Intelligence, vol. 439, Jul. 2011, pp. 33–52, ISSN:1860-949X.

[4] J.-H. Lee, J. Ha, J.-Y. Jung, and S. Lee, "Semantic contextual advertising based on the open directory project," ACM Trans. Web, vol. 7, no. 4, Oct. 2013, pp. 24:1–24:22, ISSN:1559-1131.

[5] B. Zamanzadeh, N. Ashish, C. Ramakrishnan, and J. Zimmerman, "Semantic advertising," CoRR, vol. abs/1309.5018, 2013, URL: http://arxiv.org/abs/1309.5018 [accessed: 2014-05-01].

[6] S. Chakrabarti, Mining the Web: Discovering Knowledge from Hypertext Data. Morgan-Kaufmann Publishers, San Francisco, 2003, ISBN: 978-1558607545.

[7] T. W. Yan, M. Jacobsen, H. Garcia-Molina, and U. Dayal, "From user access patterns to dynamic hypertext linking," Comput. Netw. ISDN Syst., vol. 28, no. 7-11, May 1996, pp. 1007–1014, ISSN: 0169-7552.

[8] R. Cooley, B. Mobasher, and J. Srivastava, "Web mining: information and pattern discovery on the world wide web," in Proc. Ninth IEEE Intl. Conf. Tools with Artificial Intelligence, Nov. 3–8, 1997, Newport Beach, CA, USA. IEEE, Nov. 1997, pp. 558–567, ISSN: 1082-3409.

[9] M. Strube and S. P. Ponzetto, "Wikirelate! computing semantic relatedness using wikipedia," in Proc. 21st Natl. Conf. Artificial Intelligence, Jul. 16–20, 2006, Boston, ser. AAAI'06, vol. 2. AAAI Press, Jul. 2006, pp. 1419–1424, ISBN: 978-1-57735-281-5.

[10] Hadoop. [Online]. Available: http://hadoop.apache.org/

[11] Oozie. [Online]. Available: http://oozie.apache.org/

[12] Azkaban. [Online]. Available: http://data.linkedin.com/opensource/azkaban/

[13] N. Christofides, Graph theory: An algorithmic approach. Academic press New York, 1975, vol. 8.

[14] Cascading. [Online]. Available: http://www.cascading.org/

[15] Storm. [Online]. Available: http://storm.incubator.apache.org/

[16] Spark. [Online]. Available: http://spark.apache.org/

[17] Hbase. [Online]. Available: http://hbase.apache.org/

[18] Solr. [Online]. Available: http://lucene.apache.org/solr/

[19] Elastic search. [Online]. Available: http://www.elasticsearch.org/

[20] C. C. Aggarwal, F. Al-Garawi, and P. S. Yu, "Intelligent crawling on the world wide web with arbitrary predicates," in Proc. 10th Intl. Conf. World Wide Web, May 1–5, 2001, Hong Kong, ser. WWW '01. ACM, May 2001, pp. 96–105, ISBN: 1-58113-348-0.

[21] S.-H. Lin and J.-M. Ho, "Discovering informative content blocks from web documents," in Proc. Eighth ACM SIGKDD Intl. Conf. Knowledge Discovery and Data Mining, Jul. 23–26, 2002, Edmonton, Canada, ser. KDD '02. ACM, Jul. 2002, pp. 588–593, ISBN: 1-58113-567-X.

[22] S. Debnath, P. Mitra, and C. L. Giles, "Automatic extraction of informative blocks from webpages," in Proc. 2005 ACM Symp. Applied Computing, Mar. 13–17, 2005, Santa Fe, USA, ser. SAC '05. ACM, Mar. 2005, pp. 1722–1726, ISBN: 1-58113-964-0.

[23] Boilerpipe. [Online]. Available: http://code.google.com/p/boilerpipe/

[24] C. Kohlschütter, P. Fankhauser, and W. Nejdl, "Boilerplate detection using shallow text features," in Proc. Third ACM Intl. Conf. Web Search and Data Mining, Feb. 3–6, 2010, New York, ser. WSDM '10. ACM, Feb. 2010, pp. 441–450, ISBN: 978-1-60558-889-6.

[25] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval. Cambridge University Press, New York, 2008, ISBN: 978-0521865715.

[26] Apache lucene. [Online]. Available: http://lucene.apache.org/

 ISBN: 978-1-61208-361-2

A Tool to Assist the Social Search on Facebook

Cleyton Souza, Jonathas Magalhães, Evandro Costa, Joseana Fechine, Ruan Reis
Laboratory of Artificial Intelligence - LIA
Federal University of Campina Grande
Campina Grande, Brazil
cleyton.caetano.souza@gmail.com, jonathas@copin.ufcg.edu.br, evandro@ic.ufal.br, joseana@dsc.ufcg.edu.br,
ruan.victor.amorim@ccc.ufcg.edu.br

***Abstract*— Sharing questions is a new way of getting answers on social networks. However, the usual strategy of broadcasting questions could be optimized. In this work, we propose a Social Query mobile app to assist users sharing their questions on Facebook. Before publishing the question, the app will guide the user through some steps to enhance the probability of getting an answer by someone. It is a tool to help the users phrasing their problems, restrict the social search to a certain demographic group and find people to help them. As far as we know, this is the first work to merge these three aspects of the social search. To validate our proposal, we run a questionnaire so that people could value what we are offering, and we received great feedback.**

Keywords-social query; social search; query routing; expertise finding; Facebook; system.

I. INTRODUCTION

Currently, the Social Networks (SNs) is the most popular service on the Web, surpassing even E-mail [1]. In this scenario, Facebook stands out as the most popular social network throughout the world: it has more than one billion users [2]. If Facebook were a country, it would be the third largest country in the world, bigger than the U.S. and Indonesia; and if it keeps growing in population, in three years, it would be larger than China [3]. Nowadays, one sixth of the world has a Facebook account [2].

These SN sites were first designed to allow remote interaction between geographically dispersed people [4]. One of the goals of interacting with other people is the knowledge exchange [5]. Thus, naturally, a version of knowledge exchange emerged inside these virtual spaces: users using the available features to exchange knowledge and find information through online SN.

One of the ways of knowledge exchange is the *social query*: people, trying to take advantage of the *crowd's knowledge*, share problems with their contacts, usually in the form of a question, aiming to find a solution [6]. It is an attempt to transform social relationships in practical knowledge [4]. This strategy is particularly useful when the solution requires a degree of personalization, maybe impossible to reach through other channels, because it is assumed that the friends of the questioner hold privileged information about his/her preferences [6][7].

In this work, we aim to improve the social query process. Broadcasting questions to all contacts has become a popular strategy to share problems on SNs. However, this is not the best way to benefit from the social capital [8], specially, in context of the most popular SN. Facebook feed works differently from Twitter timeline; the feed showed to each user is based on a personalized algorithm [9]; therefore, when users broadcast messages, there are no guarantees that these messages will be seen by all their contacts. Some studies defend that directing questions to experts is more efficient than broadcasting, but knowing to who the question should be directed is not always easy [10]. In addition, the way the question should be formulated could be decisive to receive an answer or not. Teevan et al. [11] found that characteristics of the question itself predicted the quality, quantity, and speed of responses. Thus, it is noticed that turning social relationships into practical knowledge is not a simple task and several factors could and should be considered in order to guarantee a solution to the problem.

In order to help users, we propose a mobile app called *Social Query*. It will guide the users through some steps before the disclosure of the questions on Facebook, enhancing their chances of getting answers. We propose a system to help people to find other people to help them. It is not only an Expertise Finding System (EFS), as most people can think, but also a tool to assist users to formulate their problems and restrict the social search to a certain social group. As far as we know, this is the first work to merge these three aspects of the social search. Through this app, the users inform their questions and receive suggestions to increase the chances of receiving an answer. The suggestions range from tips to rephrase the question to indications about whom probably might know the answer (a person or group), so the user could direct the question to specific contacts, ensuring that it will be visualized by them.

We used a questionnaire to get feedback about our proposal. Through the questionnaire people could express their opinions about the functions available in the Social Query app. The results were excellent; people considered useful most of the available functions, but we highlight the acceptance of the Expertise Finding mechanism and the Filtering mechanism.

The remainder of this paper is organized as follows. First, we will present a brief review of literature about the practice of sharing questions on SNs; next, we will detail how our proposal works; then, in Section IV, the questionnaire results are presented; later, we close with our conclusions and future work proposals.

 ISBN: 978-1-61208-361-2

II. RELATED WORK

The habit of sharing questions on the web was born on Community Questions and Answering (CQA) sites and was extended to SNs [12]. Asking a question on Facebook, for instance, is an explicit action performed by users in order to convert the social relationships maintained on the site into actionable information and other social capital outcomes [4].

The broadcasting feature, a key component of most SN sites, allows users to distribute content to their network, including requests for informational or emotional support; this ability is particularly helpful when the information is held by weak ties only available through the SN [4]. Both technical and social features made SNs the ideal place to share questions: (1) the possibility of contacting a large and diverse audience through only one post is quite useful when we are looking for information [13], (2) the fact that the majority of this audience is composed by people who know us [12], and (3) the possibility to reach weak ties [4].

Morris et al. [6] presented statistics confirming social query as a viable method to obtain answers online. In their case study, 93.5% of users had their questions answered after sharing them (in online SNs or using status update in Internet Messengers) and these responses, in 90.1% of the cases, were provided within one day. The main motivations pointed by the users who practice the social query were (1) their trust on their contacts and (2) the hope of a personalized answer. These motivations highlight the advantages of posing questions on SNs compared to more generic CQA sites; in addition, some patterns identified by the research on information seeking suggest that certain information needs, such as those revolving around quotidian occurrences, are more commonly solved by individuals that one already knows [14].

Nichols and Kang [8] confirmed that directing questions significantly increases the response rate, while the number of the answer depends to who the question will be directed to. In this sense, EFSs play an important role: if we identify an expert on the topic of the question and direct it to that expert, the answer would come faster and with higher quality [15].

The process of directing questions to appropriate helpers is known in literature as Query Routing (QR) and there is a vast research about this topic, especially when we looking the CQA's context [5]. However, most work about QR concentrates on the Expertise Finding (EF) aspect of the problem. In addition, it is usually considered a *global context* of candidates. However, our goal is the detection of specialists into the set of the questioner's friends (*local context*). Thus, the "level" of expertise that characterizes someone as "an expert" will constantly change. Moreover, EF often considers mainly the expertise about some topic; what we propose is taking into account several factors to improve the probability of finding relevant information through the help of friends.

In [10], we had proposed a QR system that routes questions to followers on Twitter based on three criteria: knowledge, trust and activity. However, in this work, we are not only proposing a system that recommends someone, but that also assists the users in the process of sharing their problems. (a) Our app will (a) analyze the question and suggest modifications, (b) suggest to restrict the search for help to a certain group of friends, and (c) will suggest people based in their bounds, availability and expertise. As far as we know, this is the first work to merge these three aspects of the social search. Next, we will detail how our proposal fits into the Q&A process.

III. SOCIAL QUERY ON FACEBOOK: MOBILE APP

The Social Query app was developed in Android. It helps users to use the potential of their social capital to transform social connections in practical knowledge. In the next sections, we will detail how our app works and the ideas behind its views.

A. First View

The First View of the app is the Login page, presented in Fig. 1 (a), where users must inform their Facebook credentials (b).

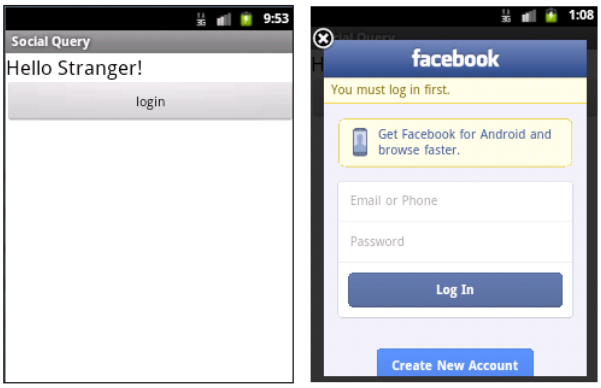


Figure 1. (a) First View and (b) Facebook's Mobile Login Dialog

After logging in, users must give us permission to access the information of their Facebook accounts and to publish content in their feeds, as presented in Fig. 2 (a). After that, they are directed to the Main Page, as presented in Fig. 2 (b).

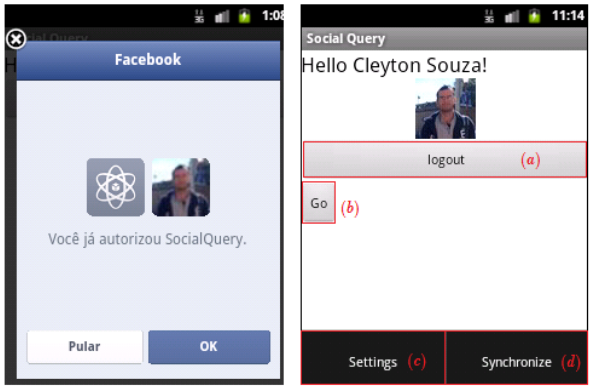


Figure 2. (a) Permission's Dialog and (b) Main View

The options in the Main page are Logout of the app (a); go to Settings (c); Synchronize again with the Facebook

 ISBN: 978-1-61208-361-2

account (d); and Go make a question (b). The Logout option directs users to the Login page again. The Settings option allows users to choose what EF model to use (currently, there are three available) and to define Filters to the EF search. The Synchronize option is an opportunity for users update the app information about them (catch more recent information about them, their contacts and new connections); it will start the same thread initiated after the Login. The Go button guides the users to the main functionality of the app. The next section will detail how it works.

B. Q&A Process

After clicking the Go button, users are directed to the New Question View, where they can inform their question. Fig. 3 illustrates this use.

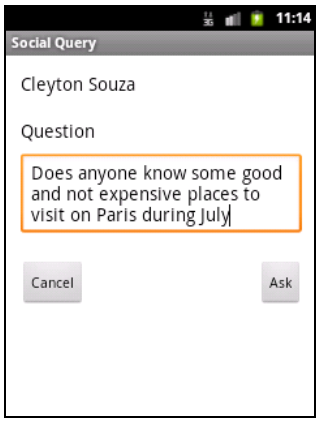


Figure 3. New Question View

In Fig. 3, the user 'Cleyton Souza' has a question about places to visit in Paris. After typing the question, the user will click the Ask button, being directed to the Tips View, as presented in Fig. 4.

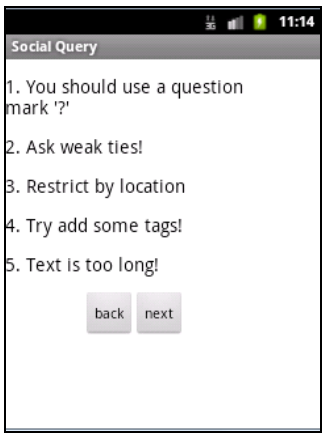


Figure 4. Tips View

Fig. 4 illustrates some of the tips that could be given to the user. Basically, we analyze the text of the question searching for specific information (e.g., terms or mentions to place or people); next, we select some pre-established tips to show users, who has the option to follow them or not.

The decision about which tips will be presented is based on the characteristics of the question, determined by our *Question Analyzer*. Teevan et al. [11] found that a concise style of question-asking, a defined scope (or audience), and the inclusion of a question mark were associated with more and higher quality responses within shorter periods of time. The Question Analyzer processes the questions and extracts their characteristics. Then, it associates these characteristics with pre-established tips, which were decided based on literature review and interviews conducted by us.

The chosen tips are displayed to users. If they decide to follow any tip, they must to click the Back button (to edit the question text) or Settings menu (to turn on some filters). After that, they can click the Next button to be directed to the Recommendation List View, where they choose who they want direct the question to. This view is presented in Fig. 5.

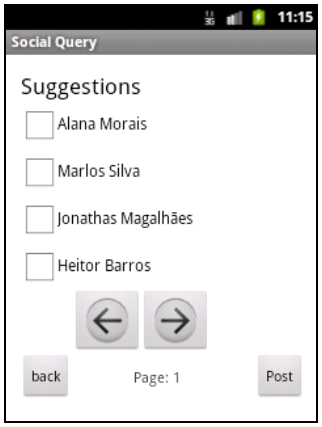


Figure 5. Recommendation List View

Friends of the questioners are ordered according their score of utility calculated by the EF model chosen on Settings. The users check the people and click the Post button. Then, the Social Query app posts on the Facebook users' feed the question, but tagging the friends that they checked.

C. Settings

In Fig. 6, the Settings View is presented, where users can edit the features of the social search.

 ISBN: 978-1-61208-361-2

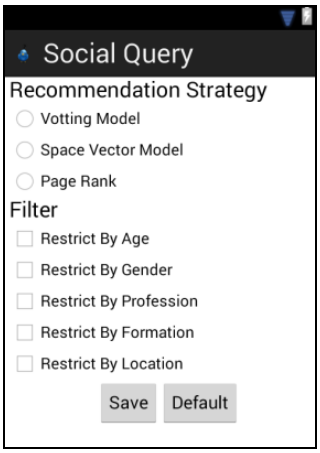


Figure 6. Settings View

Currently, the Settings limit to choose the EF model and establish filters to the Expert Search. The EF model is the technique that will be used to represent the contact's expertise. The filters to Expert Search restrict the recommendation to a certain group of contacts.

In the current version of the app, there are three EF models implemented, next they will be detailed and after the five Filters available will be explained too.

1) Expertise Finding Models

a) Voting Model: Proposed by Macdonald and Ounis [17], it considers the task of ranking experts as a voting problem. The profile of each expert candidate is associated with a set of documents that represents their expertise. The request for expert is assumed as a query into a search engine that retrieves some of theses documents. Each retrieved document is associated with one or many users and counts as an implicit vote for them. The ranking of experts is based on the total of votes of each candidate. Several strategies could be used to retrieve documents, associate the documment with the users or weighting the votes.

b) Vector Space Model: A classical approach from Information Retrieval (IR), was originally proposed by Salton et al. [18]. *The idea behind the model is to represent content in multidimensional vectors. In our context, the vector represents the content associated with each user, the coordinates represent the words, and the coordinate values are calculated using TD-IDF. The expertise score is the similarity between the expertise profile and the question vector using cosine similarity.*

c) PageRank: It is a classical algorithm that measures the importance of a node counting the number and quality of nodes pointing to it [19]. If we consider that the scenario where "a user X, author of question Q, receives an answer A, from user Y" represents a graph like X→Q→A→Y, that could be simplified to X→Y. One of the goals of PageRank is to estimate our probability of randomly getting in a node; the higher this probability, the greater the chances the node of being a good recommendation.

2) Filters

The Filters are used to restrict the social search to a certain social group. Currently, there are five Filters implemented. Restrict by: age, gender, profession, formation and location. In the current version of the app, each filter restricts the expert search to people with the same characteristic of the user. For instance, if the user is a man and he checks the "Filter by gender", it will be only recommended men; if he lives in Paris and he also checks the "Filter by location", it will be only recommended men who live or lived in Paris too.

However, for next releases of the app, we are planning the improvement of the filter mechanism. One of the improvements will be allowing the users to choose the filter value (e.g., the hometown city that they want use to restrict the search). Another improvement, it is the prediction the ideal filter value (e.g., find what would be the most indicated hometown city). In literature, there is already some research in this direction [16]. In addition, we are constantly thinking about new Filters.

IV. EVALUATION

To validate our tool, we shared a questionnaire in Facebook groups. The questionnaire was answered by 250 volunteers. To know our volunteers, the first part of questionnaire asked them about their experience with SMQA; the second part requested them to value the main function of the Social Query app.

Among the volunteers, 159 confirmed that had already shared questions through an SN. For this reason, only their answers were considered for the questions about their SMQA experience. Regarding their habits before sharing the question, most volunteers search for the answer by themselves before turning to friends for help, only 5% of them admitted that they go straight to SNs. In addition, most people (84%) often think carefully about how to phrase the problem. It is already known that a short period and a well-defined audience are associated with better answers [11]. However, only 1/3 thinks about people they know who probably can help. Moreover, 1/3 of volunteers also make the "mistake" of being thorough. Regarding their opinion about how easy is finding help through SMQA, 130 (81%) consider it easy while 29 (19%) consider it hard. Moreover, 94% said that they usually do not need repost their problem to receive an answer.

Then, volunteers evaluate the aspects of the application described in previous sections. Initially, we asked them to directly value the functions of the app. There was a template question like "How useful would be this [function]?" followed by one of the functions of the Social Query app. The options were "Don't know", "Somewhat Useful", "Useful" and "Very Useful". The results are summarized in Fig. 7.

 ISBN: 978-1-61208-361-2

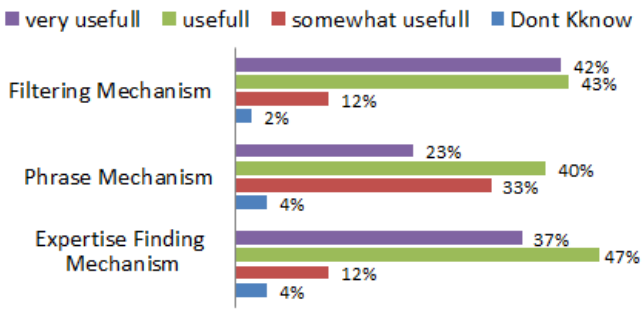


Figure 7. Results of Questionnaire (A)

As can be seen in Fig. 7, most of functions were considered, at least, useful. We compare these values using a one-tailed-right binomial test and we found statistical significance in "Useful" percentage for all functions (α=0.01). This means that the "Useful" appearance is statistically greater than "Somewhat Useful" and this should continue regardless of the amount of feedback that we received and regardless the function. In addition, the most useful functions, according the answers, were the Expertise Finding mechanism (84% of aggregate usefulness) and the Filtering Mechanism (85% aggregate of usefulness).

Regarding the Expertise Finding, we asked volunteers about what they are looking for in answers from their Facebook friends. "Truth" (27%) was the most desired characteristic followed by "Detail" (21%). "Personalization" was the less desirable characteristic (2%). This, particularly, was an unexpected result, because, many appreciated that their private SN was familiar with their additional context, such as knowledge of their location, family situation, or other preferences [6]. The popularity of these characteristics that reflect a mastery over a subject (Truth and Detail) results in a need to prioritize expertise rather than other more subjective criteria (availability, trust, etc.) when estimating the utility of each candidate.

Finally, we asked if they believed that certain questions were implicitly restricted to certain kind of people. We used a template question like "Do you agree that some questions can only be answered by a certain [characteristic]?" followed by each Filter option. This question aimed to evaluate the volunteer's perception about the utility of the Filtering mechanism and their results are summarized in Fig. 9.

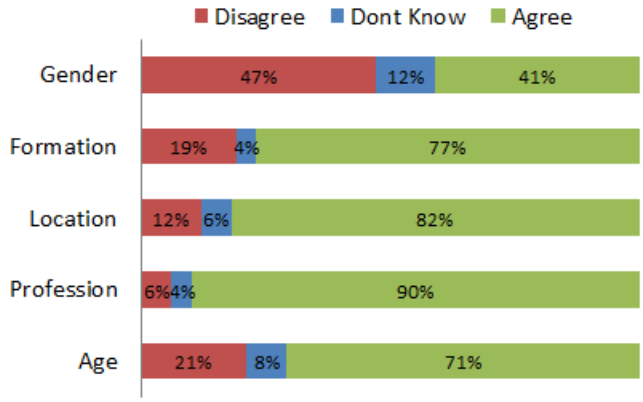


Figure 8. Results of Questionnaire (B)

In general, all the Filters were considered useful by most of volunteers, except by the gender filter, which was a polemic subject. We observe the highest divergence between male and female opinion. We believe that this rejection was due to the sexist aspect of our question. This may be absurd, but men and women may have understood that they were not able to answer questions made by the other gender and rejected the filter by this reason. But, this is just a guess; we could not confirm this without individually interviewing each respondent. The fact is that the Gender filter was not well received by our audience

V. CONCLUSION AND FUTURE WORK

In this work, we presented the Social Query app to assist users to search for information on SNs. While most part of previous work focused on the Expertise Finding mechanism, we propose a tool to help the users through several steps of the social search process. First, our solution helps the users to rephrase the questions enhancing its chances of be answered. Second, the app offers three different approaches to finding experts. Last, there is an option to filter the expert finding search to a certain group with the same demographic characteristic of the questioners (age or gender, for instance).

To evaluate our proposal we run a questionnaire, which was answered by 250 Facebook users. Through the questionnaire, these users could give their impressions about the functional aspect of the Social Query app. The results were excellent. The main functions (Expertise Finding mechanism, Filtering mechanism and Rephrase mechanism) of the app in average were considered at least useful by more than 40% of users. In addition, we obtained great feedback that allows us to think about improvements to our proposal.

As future work, we are planning the following improvements: (1) use of other Expertise Finding models, some of them considering semantics; (2) enhance the Question Analyzer, besides suggesting changes in problem specification, automatically applying some or all of these changes; (3) improve the Filtering use to specify the input; (4) allow the user to maintain a list of contacts; (5) allow users to maintain lists of friends; (6) considering additionally the reputation of the users, based on previous; and (7) make friends of friends available as expert candidates.

 ISBN: 978-1-61208-361-2

ACKNOWLEDGMENT

We want to thank people who answered our questionnaire.

REFERENCES

[1] T. Wayne, “Social networks eclipse e-mail”, May. 2009. [Online]. Available from: http://www.nytimes.com/2009/05/18/technology/internet/18drill.html [retrieved February, 2014]

[2] M. Ross, “Facebook turns 10: the world's largest social network in numbers,” Feb. 2014. [Online]. Available from: http://www.abc.net.au/news/2014-02-04/facebook-turns-10-the-social-network-in-numbers/5237128 [retrieved February, 2014]

[3] G. McMillan, “If Facebook was a country, it’d be larger than China in three years,” Feb. 2013. [Online]. Available from: http://www.digitaltrends.com/social-media/facebook-could-be-larger-than-china-in-three-years-time/ [retrieved February, 2014]

[4] R. Gray, N. Ellison, J. Vitak, and C. Lampe, “Who wants to know? question-asking and answering practices among Facebook users,” Proc. 16th Conference on Computer Supported Cooperative Work (CSCW), ACM Press, 2013, pp. 1213–1224.

[5] B. Furlan, B. Nikolic, and V. Milutinovic, “A survey of intelligent question routing systems,” Proc. 6th International Conference Intelligent Systems (IS), IEEE Press, 2012, pp. 14–20.

[6] 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.

[7] G. Manoranjitham and S. Veeraselvi, “Mobile question and answer system based on social network,” International Journal of Advanced Research in Computer and Communication Engineering, vol. 2, September 2013, pp. 3620–3624.

[8] J. Nichols and J. Kang. “Asking questions of targeted strangers on social networks,” Proc. 15th Conference on Computer Supported Cooperative Work (CSCW), ACM Press, 2012, pp. 999–1002.

[9] L. Magid, “Facebook tweaks news feed algorithm again,” Jan. 2014. [Online]. Available from: http://www.forbes.com/sites/larrymagid/2014/01/21/facebook-tweaks-news-feed-algorithm-again/ [retrieved February, 2014]

[10] C. Souza, J. Magalhães, E. Costa, and J. Fechine, “Social query: a query routing system for Twitter,” Proc. 8th International Conference on Internet and Web Applications and Services (ICIW), IARIA Press, 2013, pp. 147–153.

[11] J. Teevan, M. Morris, and K. Panovich, “Factors affecting response quantity, quality, and speed for questions asked via social network status messages,” Proc. 5th International Conference on Weblogs and Social Media (ICWSM), AAAI Press, 2011.

[12] C. Souza, J. Magalhães, E. Costa, and J. Fechine, “Routing questions in Twitter: an effective way to qualify peer helpers”. Proc. International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), IEEE Press, 2013, pp. 109–114.

[13] M. Burke, R. Kraut, and C. Marlow, "Social capital on Facebook: differentiating uses and users,” Proc. Conference on Human Factors in Computing Systems (CHI), 2011, pp. 571–580.

[14] R. Savolainen, “Everyday life information seeking: approaching information seeking in the context of way of life,” Library & Information Science Research, vol. 17, 1995, pp. 259-294.

[15] 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), Spring Press, 2012, pp. 714–729.

[16] 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.

[17] C. Macdonald and I. Ounis, “Searching for expertise: experiments with the voting model,” The Computer Journal, vol. 52, 2009, pp. 729–748.

[18] G. Salton, A. Wong, and C. Yang, “A vector space model for automatic indexing,” Communications of the ACM, vol. 18, 1975, pp. 613–620.

[19] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," Journal Computer Networks and ISDN Systems, vol. 30, April 1998, pp. 107–117.

[20] G. Comarela, M. Crovella, and V. Almeida, “Understanding factors that affect response rates in Twitter,” Proc. 23th International Conference on Hypertext and Social Media (HT), ACM Press, 2012, pp. 123–132.

 ISBN: 978-1-61208-361-2

Webpage Resource Protection via Obfuscation and Auto Expiry

Zhuhan Jiang and Jiansheng Huang
School of Computing, Engineering and Mathematics
University of Western Sydney, Sydney, Australia
{z.jiang, j.huang}@uws.edu.au

***Abstract*—Content delivery via web pages or sites are becoming increasingly popular due to the effectiveness and versatility of the readily available delivery mechanism, especially in the e-education and training. While the copyright laws are there to protect the ownership and commercial rights of the intelligent properties, the openness of the web architecture often makes it impossible to prevent the content source being misappropriated or incorporated illegitimately elsewhere after some modifications on the downloaded source. We propose here an obfuscation mechanism for the HTML5 to convert a site of raw content into a site of obfuscated pages and images. With the advent of canvas on HTML5 and the AJAX to stop certain unauthorized access, the whole site of documents can be rendered meaningless or useless on both the server and the client side if just a small key part is modified or hidden. Several masquerading algorithms have been proposed for this purpose. The obfuscation will become permanent if a webpage is merely downloaded or even DOM-saved without having all necessary intermediate data or keys tracked by a specialist, before the auto-expiry of such a process, at a cost tantamount or exceeding the reconstruction of the original documents from scratches hence defeating the purpose of piracy. We applied the scheme to the delivery of a university subject by automating the whole process.**

Keywords-Content obfuscation; document ownership protection; e-training; HTML5 and AJAX; document auto-expiry.

I. INTRODUCTION

Web has long since surpassed its original purpose of publishing material on the Internet. It has evolved into a very effective and interactive platform to operate business, create social media, and run educational or training services, to name a few [1]. The largely sharing-by-all paradigm of the web during its inception has been gradually diverted into controlled access and restricted content or media deliveries, especially in e-business and e-education. The question of how much one can safeguard the ownership of the delivered content and to what extent naturally arises and becomes increasingly pertinent. It is generally understood that, if a piece of material is delivered via web to its authorized recipients, the document is practically fully surrendered in that almost all text and images there are at the disposal of the recipients in their original digital format, as long as the recipients have the minimal expertise on the web technologies. This means that these recipients may easily modify the content to reproduce and redistribute the original material. This can be highly undesirable for the protection of intelligent properties, especially when there are powerful web crawlers or site copiers [2]. Due to the common inherited belief that not much can be done in this regard, not much research efforts [3] have been made to seek as much as possible the protection of the delivered material, apart from setting up a few superficial obstacles such as [4] disabling the copy/paste, disallowing printing certain parts of the pages, using data URI scheme, and replacing a portion of a web page by a Javascript (JS) which converts the coded counterpart of the portion, e.g., in base 64, back to the Hyper Text Markup Language (HTML) format. However, these superficial tricks are only effective to the people of no or shallow technical skills.

Before we undertake to investigate how to protect our web source, we have to first establish the level of protection that we seek. If a piece of web content is to be delivered to a client's browser screen, there is no way one can stop the client from taking a picture or a video of the delivered material. Hence, a separate hardcopy or recording of the essential content does not belong to our protection scope. For convenience we refer this as the *knowledge scope*, and web content is thus unprotectable there. Next, we consider the *reproduction scope* in that the web content can be saved outside the original server and utilized to achieve essentially the same browsing experience. Currently, almost all regular web pages can be fully saved and are thus unprotectable in this scope apart from those live stream multimedia objects. Since most web contents are not ideal or practical to be streamed live, we will exclude the consideration of such objects in this work. The last scope we will also look into is the *source scope* in which we will examine whether the source can be saved in a clear and manageable format. A piece of source is considered to be in a readily manageable format if it is close to the original format that is suitable for modification or editing. Our aim in this work is to achieve a reasonable protection in the reproduction scope as well as in the source scope through obfuscation. Because the web delivered material can't be realistically protected in the knowledge scope, then the extent of protection is only limited to the understanding that the efforts required to reproduce the same browsing effect or manageable source are *not less* than the efforts required to start from scratch on the mere basis of "hardcopy" saved in the knowledge scope. Just like the industrial encryption algorithms are often theoretically breakable if there were an infinitely fast computer, but will be nonetheless treated as secure, as long

 ISBN: 978-1-61208-361-2

as the current technologies are still way behind the power required to break the encryption, the protection of the web resource is often in comparison to what is required to reconstruct the resource from scratches. In this sense, if an unbearably significant amount of time or effort would be required to break or circumvent a protection, it still serves as a good protection.

It is worth noting at this point that while pure cryptography, with or without the Public Key Infrastructure, will in principle secure the transmission of almost anything digital, it often requires [5] additional support platforms, the measures to secure the keys, as well as the willingness to sacrifice certain performances or flexibilities. This also explains the needs for the obfuscation systems like the one we are proposing here.

This paper is organized as follows. First, Section II describes the basic design principle for the obfuscation of both the images and the text, utilizing HTML5 features, Asynchronous Javascript and XML (AJAX), and time-dependent keys. The automation process of the obfuscation is then outlined in Section III, with an implementation done for the demonstration in Section IV. Section V finally gives a conclusion.

II. THE DESIGN PRINCIPLE AND METHODOLOGIES

A. The Basic Design Principle

When a browser renders a page, it first loads a preliminary source of the page and then gradually loads other sources specified in the current or updated page content on a need to basis or recursively. Because additional page or resource loading may be activated by JS after interaction with a client user, the required resources in principle cannot be totally identified before a browsing session is fully completed. This means when JS are properly engineered, a

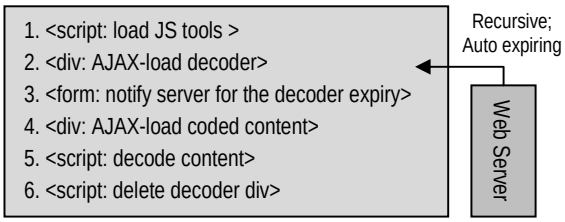


Figure 1: A simplistic scheme

browser will have no memory of its intermediate interactive page modifications or loading by the time of saving a page, leaving the fully saved version still missing the intermediate components to function independently on its own. Some browsers, such as Chrome, as opposed to IE, will save maximally the page content by using the *resulting* Document Object Model (DOM) at the time of making a copy, and thus store elements additionally loaded through such as AJAX. However, the unavoidable loss of the intermediate execution memory still leaves room to craft the protection or obfuscation of the web source. A simplistic scheme is depicted in Fig. 1 in which a "decoder" in JS is dynamically loaded, directly or recursively, via AJAX, and the decoder expires immediately or automatically after a certain period of time. We note that a direct page saving will miss the decoder, and a manual tracing may hit the expiry time quantum particularly when recursive AJAX loading is utilized. When the same page is reloaded, it may load a different decoder or key valid within a different time quantum. Hence, if a part of the content to is to be dynamically reconstructed at the expired viewing time it will produce an illegible result. This is what we will call *lapse protection* in that the pages are unrecoverable after the site is disabled even though they are downloaded and "saved". That is, completely recoverable pages must be constructed before the site life expectancy quantum has lapsed.

The coded content does not necessarily imply an encryption, and in many cases it is not encrypted at all. Some browsers can still save partly the images and text in the content. However this can be avoided too to the similar extent of lapse protection. The basic idea is to use JS to generate and rearrange the general text and use canvas in HTML 5 to superpose images and wipe out the loading traces of the component images. Because the JS portions need to be preserved so as to maintain the original interactivity, mixture of JS and plain HTML portion cannot be "flattened" out into just HTML without losing the interactivity. This explains that all browsers save the mixture of JS and text as they are, but the coded form of such a mixture wouldn't be much more useful compared to a hard camera copy.

B. Conceal the Images

Since images constitute an important part of a typical web page, how to conceal the source of the original images deserves a separate consideration. Whenever an image is rendered in a web page, it appears in the form of an image element and will be typically stored as a part of DOM with its hosting page, even though some browsers may not do so when images appear in an AJAX-loaded DIV section. Our strategy comes with the advent of the CANVAS object

1. Replace all image tags of a given page with the canvas tags;
2. AJAX-load JS which is used to load all images into variables of image objects;
3. AJAX-load JS that maps images to canvases;
4. Draw and decode images on the corresponding canvases
5. Remove the element containing the JS loaded in steps 1 and 2

Figure 2: Draw images in canvases

which is meant to support interactive image modifications at the pixel level. Just like JS interactivity in general can't be flattened into a piece of JS-free text, the dynamic nature of canvas implies that it is not stored state-wise when a page containing a canvas is saved. Hence the simplest strategy to protect images will be to display each image within a canvas, as described in Fig. 2.

However, images will stay in the cache for some time when they are loaded even if loading paths may be removed by JS within a web page. To avoid such cached images to be directly retrieved and made use of, one may load their distorted counterparts instead and use JS and a rectification

 ISBN: 978-1-61208-361-2

key to dynamically to convert the images back to the original on the canvases. As the images are meant to be delivered to the recipient for viewing and can be camera-copied anyway, there is not much point to hide all picture details, and a reasonable image distortion will render the source not directly useable by a third party. There are infinitely many ways to distort an image, and we will consider several here in detail.

We start with a simple block-wise permutation of the images, with each block optionally subject to an additional "rescaling". For a given picture image P of resolution $M \times N$ pixels, we carve it up into blocks $\{ B_{i,j}: i=1,..,m, j=1,..,n \}$, or simply $\{ B_k: k=0, ..., K-1 \}$ with $K=m \cdot n$, of $\lfloor M/m \rfloor \times \lfloor N/n \rfloor$ pixels starting from the top left corner of the image, see Fig.

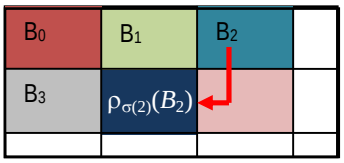


Figure 3: Image blocks

3. Let σ be any permutation of 0,1,.., K-1, and ρ_k an invertible mapping that can be used to transform any images. If each image block B_k in P is replaced by $\rho_{k'}(B_{k'})$ where $k'=\sigma(k)$ is the index permuted from k by σ, then the resulting image P' is our distorted image which can be reverted back if one knows σ and all the ρ_k. For simplicity, we have kept intact the potential strips left over on the right and bottum due to the incomplete block partition. For any key ω in the form of a sequence of random characters, we can derive a corresponding permutation σ in the following way: (i) Repeatedly concatenate ω with itself if $|\omega|<K$; (ii) Let $c=\omega(0)$ be the 1st character of string ω, and $d \equiv c \bmod (K)$, and we assign $\sigma(0)=d$, i.e., 0 is permuted to d; (iii) Let $c=\omega[1]$ be the next character in ω, and $d \equiv c \bmod (K-1)$, set the ordered indices I as $\{0, 1, .., K-1\} \backslash \{\sigma(0)\}$ in increasing order, then the index $I(d)$ is the index that 1 will be permuted to, i.e.,

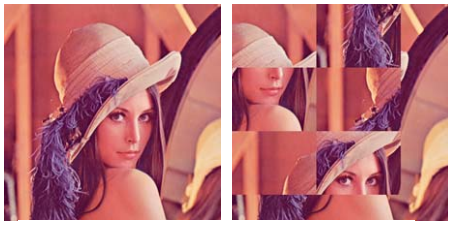


Figure 4: Image block permutation

$\sigma(1)= I(d)$; (iv) Repeat essentially step (iii) to have the permutation σ completely constructed. We have thus shown how a random string can uniquely determine a permutation of K indices. Fig. 4 shows an image block permutation with block size 200×150 pixels and ω=abc.

Next, let us explore what image transformers ρ_k we can design. The simplest transformers will be an identity mapping or an inverse mapping which produces a "negative" image. Although in theory one could decompose any image into a form of combination of two images one of which is pseudo randomly generated pixel-wise, such decomposition would result in massively reduced image compression [6][7] and would thus be impractical. Hence we propose to define an image transformer to be of the form ρ_k: $x_{i,j} \rightarrow x'_{i,j}=x_{i,j}+ \psi_{i,j}$ where image $\Psi=(\psi_{i,j})$ is a smooth image so that any smoothly varying portion of the image $X=(x_{i,j})$ will be mapped to a similarly smooth portion of image X' so that compression will work at a similar scale to the original one. As a simple example, one can use the previously mentioned random key ω to create seeds for a pseudo random number generator, and then generate the pseudo integer constants τ_k and set $\psi_{i,j} =\tau_k$ for all i and j. There are many algorithms to generate pseudo random number, including for instance the Mersenne Twister [8], or the generator by Marsaglia [9]. We note that permutation of image blocks can of course be done in a variety of ways, including the one implemented in [5] which tries to smooth the block borders in addition. In contrast to their work which focuses more on the transmission of images alone, we are more concerned with the speed and efficiency, and will thus have to avoid any algorithms that would lead to the massive increase on the size of the transmitted images.

C. *Obfuscate the Text*

Since the text of an HTML page is delivered in plain to its recipient, the only way to obfuscate the content or make it illegible or unusable is to resort to the client side scripts that would dynamically convert the obfuscated text back to the intended version within a browser, or vice versa. Since the role of JS is to provide client interactivity rather than solely manipulate the text, it is not possible to holistically emulate the execution of all the JS by just the "resulting" text. This thus lends us means to encrypt, hide, obfuscate or camouflage the text. In principle, one may additionally make use of the current visibility of the DOM elements within a browser, using for instance [10] getBoundingClientRect, so that only those elements within the current viewport of the browser will be guaranteed to be dynamically converted back to the intended text or format and some of the other DOM elements will still contain the coded or "incorrect" content.

Assume that a decoders array of algorithms are loaded into the current page as in Fig. 1, then a piece of coded text T may be converted back to the intended format by executing *decode*(T, k), where k is a key, and then having its result written back to the page via the JS function w defined as document.write if possible, and can be achieved via the DOM element editing anyway if necessary. In fact, one can place T into an invisible element with a designated element ID so that a JS can easily further decide whether to convert the coded text. A decoder could be a decrypter, with a key extracted from ω if needed, or could be as simple as a word or letter shuffler. If the coded text T is made to retain the demarcation of the sentences, then one can also use JS to parse T sentence wise and have each sentence decoded by a different decoder sequentially. If the word structure is also

 ISBN: 978-1-61208-361-2

preserved or identifiable, then the decoding can be moved towards the word level too.

Due to the nature of our goal, cryptographic encryption is generally an over kill. Instead, an effective design of text scramblers will serve the purpose better in general. Our text scrambler will permute a selected group [11] G of g characters containing at least a-Z, A-Z and 0-9, that is, all the characters from base 64 apart from "+" and "/". Hence, for any string $s \in G^*$, any character $s[k]$ indexed by x in G can be permuted to the character indexed by $x'=\sigma(x+k)$ in G. For notational convenience, we may simply identify the character in G with its index. This is simple to implement and hides away also the original character statistics. We note that additional cryptographic features can also be added at this stage.

D. Document Corruption and Presentation Erosion

Although a typical protection application will have most of the publically transmitted web resource in a "ciphered" or "corrupted" form and the correct content and presentation converted back real-time, the methodology is essentially the same as corrupting or eroding a proper page into something "illegible". To corrupt or obfuscate a web page, we can erode the textual accuracy, image content or correspondence, Cascading Style Sheets (CSS) styling, as well as JS interactivity, to name a few. One may even choose to erode a web page to the extent that is proportional to the time elapsed from the expiry date of the decoding key. There can be countless ways to corrupt or destroy the page content so as to protect the content ownership; we will thus examine a few prominent and effective strategies that can be employed to achieve this goal.

First, global variables will be ideal to represent the expiry status which can be easily retrieved elsewhere using a non-telling name such as $w[x]$ where w contains the window object and x contains the string name of a global variable. Additional JS tools, if needed, can be loaded dynamically into the HEAD element of the DOM with callback enforced. Next, one can traverse the DOM tree to scramble the page content of text nodes, and to remove randomly or selectively some of the nodes for embedded or external CSS. If one wants to select only *some* CSS definitions to disable for an external CSS file, the document.styleSheets array allows one to do so. As for JS, one can enumerate all the properties and functions for any given object, e.g., for(var p in obj) { if(typeof obj[p] == 'function') { .. } } much like playing the role of a name space, to determine which ones are to be modified or deleted for instance. If obj is the window object, then all global variables and functions can be located or modified. We note that any functions or properties defined via such as obj.f=function(){..} or obj.p='foo'; can be located this way.

As for the images on a web page, they can be temporarily drawn on an invisible canvas, modified and then saved to replace the original for the display. One can also use JS to distort the ordering of the images within the same page. We have thus shown that there should be no technical obstacles to have a JS code implemented to distort, encrypt or camouflage a web page.

E. Issuing Time-dependent Ticket of a Key

How to create a ticket for a timestamp and a given key of characters so that the ticket gives a random look and can be used to recover the original key if another (new) timestamp not exceeding a specified time difference is passed to it? In other words, if a mapping T=genT(t,k) generates a ticket T from a timestamp t and key k, and a mapping k'=genK(t',T) generates a key k' from a timestamp t' and a ticket T, then $k=k'$ only if t' has not exceeded t by a specified amount. For this purpose, we allocate M bits of which N bits are to cater for the timestamp scope, leaving thus $K=M-N$ bits to represent part of the key. The procedure to generate a ticket is as follows: (i) convert into binary the timestamp t and the key k which is padded bits 0 to make it a multiple of K bits; (ii) insert a block of N random bits after every K bits of the key k to form an expanded key k^*; (iii) collect M bits on the right of timestamp t, padding bits 1 on the left if necessary, to make an expanded timestamp t^*; (iv) XOR t^* and k^* bitwise into T^*; (v) pad bits 0 to T^* on the right to make it a multiple of 6 bits and then convert it into the final ticket T by mapping successively each block of 6 bits into a base 64 character or an equivalent; see Fig.5 for an illustration. The reverse function genK can be similarly constructed. To avoid the use of "+" and "/" characters which have a special meaning in a web page, we used "." and "_" instead. As an example, for M=30, N=15, k="Random", t=1388494800 (1 Jan 2014), we have T=gwUAj_dGtMungJk6WQi3 as one of the valid tickets for the same key and timestamp.

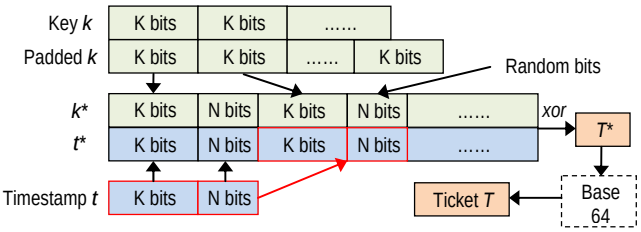


Figure 5: Bitwise mixing key with timestamp

Our experiments show that the generated ticket does appear quite random because of the distributed random bits in k^*, particularly when K is relatively small. For a much larger K, one can also apply an additional pattern permutation to make the appearance more random. This can be regarded as a form of *variation divergence*, and can also be used to masquerade a pattern like a timestamp, such as the bitwise mixing in Fig. 5 if desired, after T^* is obtained there.

III. Automate the Protection Process

In terms of the eventual applications of our proposed webpage source protection scheme, a typical scenario would be that a web designer first creates a web site mostly in a normal manner, embedding additional automation directives within some web pages if necessary. Then, a web server will dynamically create a serving version of the pages that are injected with the content obfuscation mechanism. In this section, we first outline a general framework that can be employed to serve a regular web page *a.html* in a protected

 ISBN: 978-1-61208-361-2

mode, and then examine a number of implemental techniques. Even though the original pages and images may be stored in a database as in a typical Content Management System (CMS), we will for simplicity leave them as files behind a firewall, hence not directly accessible to the outside.

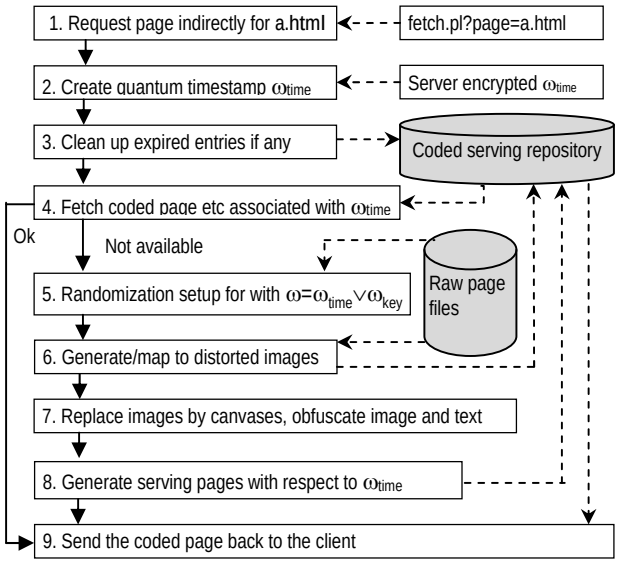


Figure 6: Work flow of protected document delivery

The workflow of delivering the content of a "raw" page *a.html* that resides behind a firewall to a remote client in a protected mode can be illustrated in Fig. 6. When a request for a page, say, *a.html*, is made in step 1 through a server side script, say, *fetch.pl*, a timestamp ω_{time} is created in step 2 and such timestamps will differ for every quantum of time, say 1 hour or a single day or a month. For each new quantum timestamp we can create a new serving version of the document which will be associated with a different set of algorithms and keys to convert the serving page back to the intended format via JS. The version associated with an outdated timestamp will thus expire and be removed in step 3 at the next appropriate time after which full reconstruction of the pages saved for the outdated timestamp will become impossible. If the coded serving version is available for the current timestamp, then the page will be immediately returned via steps 4 and 9 to the client. Hence repeated access of the same page within the expiration time quantum will be straightforward and may be cached on the client.

If the serving version for the current timestamp is not available, then steps 5-8 will immediately generate one. The *race condition* [12] can be easily avoided if such a creation has to first acquire a lock that would expire in a given period of time such as several seconds. The unprotected raw page files will be archived behind the firewall or protected by the access permissions so that they are not directly accessible by any remote clients but are accessible to certain server scripts. Step 5 will typically accomplish the following tasks:

i) Read in the raw page *a.html*, extract all image element IDs and other element IDs, and assign distinct new IDs to the image elements if they don't have an existing one. For sections of text, wrap them with tags along with distinct new IDs. To facilitate the processing and conversion, processing directives may be inserted in the raw pages to indicate such as code skipping via <!--wdp--skip--> up to <!--wdp--/skip--> and text for obfuscation via <!--wdp--paragraph--> up to <!--wdp--/paragraph-->.

ii) Construct an array TxtId for all the IDs of the text elements that are to be encoded, and associative arrays such as ImgViaId and ImgAlgorithm for proper identification of the images, recovery algorithm, etc.

iii) All the generated supporting files such as the algorithms in JS and distorted images may be kept in subdirectory ω, or each file is prefixed by the "ω_". The ω_{key} can be regarded as an optional seed for the randomization in this Step 5.

Step 6 will generate distorted images from those original ones dynamically. Such a distortion can also be constructed offline by separate image processing software with the mapping saved in the same directory as the raw page. On top of the images being distorted, the randomization of the image names could further obfuscate the correspondence between the images and their rightful positions in the document. As for the image distortion, it should be dynamically achievable through the use of such as *ImageMagick* package and the *PerlMagic* module.

To replace the images by canvases in step 7, we load all the distorted images into image objects created by JS in a single go as explained in Fig. 2, and then paint them on the respective canvases before deleting those loaded image objects. As for the text coding, we use the *seed* ω to generate a pseudo random list of scrambling algorithms, and sequentially code the *i*-th section of text by the *i*-th algorithm. Combining all these together we can complete the task of step 8. Once the page is loaded through step 9, the decoders will unscramble the text and rectify all the distorted images.

IV. IMPLEMENTATION AND EXPERIMENTS

Although all strategies and techniques studied earlier may be implemented to protect the original webpage source by obfuscation and timed expiry, a selected subset often suffices the main needs while not inflicting too much implementation cost. Our first application is on the web delivery of the practicals for a university database subject. The main procedure is as follows. First, we created the normal web pages for the practicals. We then developed a PERL script extractImages.pl which reads in a regular webpage, say, a.html, and generate another page, say, b.html. When this new page is loaded into a browser, it will generate all the corresponding obfuscated images, appending "-x" to the corresponding image name while leaving the file extension intact. The image conversion is done through the JS we implemented, utilizing the support for the *canvas* element and the existing JS tools such as FileSaver.js and canvas-toBlob.js. Then using another PERL

 ISBN: 978-1-61208-361-2

script convertPage.pl we developed for this implementation to convert the raw webpage to a new version to be eventually made available on the web server. In the raw page a.html we inserted <!--wdp--paragraph--> and <!--wdp--/paragraph--> to delimit the body of the page: all the tag-free text portions will be separately and randomly obfuscated while tags and JS and CSS are preserved. In fact, the tag-free text will be first parsed into "sentences" and the obfuscation is independently done at the sentence level.

As an example, a part of the converted page will appear as in Fig. 7 when the page is being served in the normal

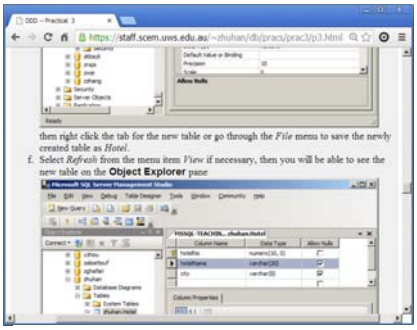


Figure 7: Normal view of the page

manner. Fig. 8 shows how it appears when the page is expired. We observe that both the text and images are now distorted or obfuscated. If the page is saved as source code,

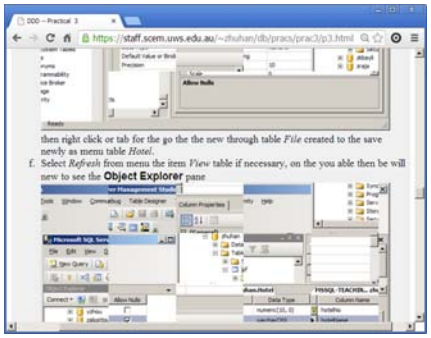


Figure 8: Expired view of the page

typically of the saving with IE, it will actually save not much. If the page is saved as a complete page, practically dumping the DOM elements as with Chrome, the part displayed on the saved page will appear obfuscated as in Fig. 9. Besides, all interactivity via the JS on the page will cease to function. The code for the 2nd image in Fig. 7 reads

```
<a href='viewimage.php?u=lab3h-x.png'
id='image_lab3h_png'>
<canvas id='canvas_lab3h_png' width=0 height=0
        class=iconwidth _width=450></canvas>
<script>imageOnCanvas('lab3h-x.png',
        'canvas_lab3h_png',128,96);</script></a>
```

that corresponds to the original text <img lab3h.png, and part of the (auto converted) source code for the above takes the following typical out of order form

```
... <i><span id='sid_82_0'>View</span><script>a('sid_82_0');
</script> </i><span id='sid_83_0'>able new table on you will
the be to then the necessary, if
see</span><script>a('sid_83_0'); </script> <b><span
id='sid_84_0'>Object Explorer</span> ...
```

which means the source being served is completely obfuscated in both text and images. When the source code gets into the client's browser, it becomes rectified under proper display styles but otherwise undecipherable as in Fig.

Figure 9: Viewing the saved page

9. Fig. 10 gives a comparison of the same page when being displayed in normal, expired and saved manners respectively. Although Fig. 10 already well shows the broad

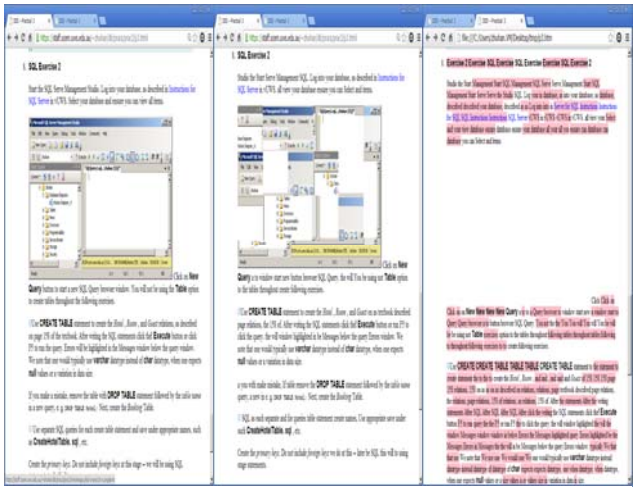


Figure 10: Comparison on normal/expired/saved page view.

differences, we further magnify parts of Fig. 10 in Fig. 11 to see more clearly the details. In Fig. 11, the main background comes from the 2nd picture in Fig. 10, the top left with a light green background comes from the 3rd picture, and the one at the bottom with a purple background extracts the 1st paragraph in the 1st picture of Fig. 10. We note that the extent of the text obfuscation can be controlled and will be somewhat proportional to the expired time. We also note that the image is missing from the saved version because the content on a canvas is not saved when a page is "completely" saved by a browser, and we also deliberately highlighted in pink the random text automatically inserted for the obfuscation.

Since the "decoding" part is to be completed by the client browser, there will be an overhead on the speed of page rendering. Hence, how complex an algorithm is allowed to be must be tied to the allowed performance degradation. Fortunately, the rectification of the deliberately

 ISBN: 978-1-61208-361-2

distorted images does not have a visible effect in general, and the complexity of the textual obfuscation can also be controlled by the operating parameters. For the system we

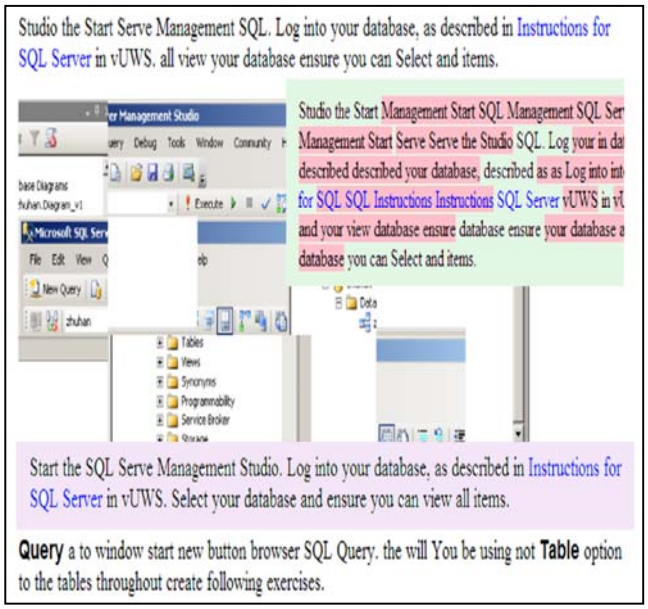


Figure 11: Magnified parts of Fig. 10

implemented here, the overhead is observable but easily tolerable, and is in general in the scale of half a second. We however plan to explore more quantitatively the overhead impact in our future work.

V. CONCLUSION

The protection of web page content has been explored in terms of the reproduction scope and the source scope. We proposed an obfuscation mechanism that protects the page resource on both the server and the client side. On the server side, by removing or altering a simple key on the server, the otherwise fully functional website can be made instantly useless as all the resources there are obfuscated and unrecoverable without the key. On the client side, the web pages will be initially loaded in the distorted format for both the images and the text from the server, and dynamically rectified if the clients' authorization has not expired. If a client saves the "complete page" using a browse saving tool, the saved page remain fully obfuscated at the source level, and will not be able to get dynamically converted into its intended proper format if the deciphering key is not saved, as is the case for all auto-saving, or expired. A client user, having saved the complete pages, will find making use of the saved resource to imitate the original server not any easier than building everything from scratches, according merely to the screenshots of the whole site. This, therefore, defeats the purpose of saving the page resources.

REFERENCES

[1] F. Greyling, M. Kara, A. Makka, and S. Van Niekert, "IT worked for us: online strategies to facilitate learning in large (undergraduate) classes", Electronic Journal of e-Learning, vol. 6, 2008, pp. 179-188.

[2] HTTrack Web Site Copier, http://www.httrack.com, last accessed on 27 May 2014.

[3] Y. Gao, Y. Zhang, B. Bai, and X. Wang, "Survey of webpage protection system", Computer Engineering, or 计算机工程 as its original name, vol. 30, no. 10, 2004, pp. 113-115.

[4] Web Protection, http://www.wiscocomputing.com/ articles/ protect_web_sites.htm, also http://thenetweb.co.uk/obfuscate-hide-and-obscure-e-mail-addresses-telephone-numbers-and-text, last accessed on 27 May 2014.

[5] A. Poller, M. Steinebach, and H. Liu, "Robust image obfuscation for privacy protection in Web 2.0 applications", Proc. SPIE 8303, Media Watermarking, Security, and Forensics, 2012; doi: 10.1117/12.908587.

[6] K.S. Thyagarajan, Still Image and Video Compression with Matlab, Wiley, 2010.

[7] Z. Jiang, O. de Vel, and B. Litow, "Unification and extension of weighted finite automata applicable to image compression", Theoretical Computer Science A, vol. 302/1-3, 2003, pp. 275-294.

[8] M. Matsumoto, and T. Nishimura, "Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator", ACM Transactions on Modeling and Computer Simulation, vol. 8, no. 1, 1998, pp. 3-30.

[9] G. Marsaglia, "Seeds for random number generators", Commun. ACM vol. 46, no. 5, 2003, pp. 90-93; also https://groups.google.com /forum/#!msg /comp.lang.c/ qZFQgKRCQGg/rmPkaRHqxOMJ, last accessed on 27 May 2014.

[10] http://stackoverflow.com/questions/123999/how-to-tell-if-a-dom-element-is-visible-in-the-current-viewport, last accessed on 27 May 2014.

[11] J. A. Beachy and W. D. Blair, Abstract Algebra, 3rd edition, Waveland Pr Inc, 2006.

[12] Y. Yu, T. Rodeheffer, and W. Chen, "Race track: efficient detection of data race conditions via adaptive tracking", ACM SIGOPS Operating Systems Review – SOSP'05, vol. 39, issue 5, 2005, pp. 221-234.

 ISBN: 978-1-61208-361-2

Message Spreading Model over Online Social Network with Multiple Channels and Multiple Groups

Sungmin Hwang and Kyungbaek Kim
Dept. Electronics and Computer Engineering
Chonnam National University
Gwangju, Republic of Korea
e-mails: {sungmin1511@gmail.com, kyungbaekkim@jnu.ac.kr}

***Abstract*—Understanding the characteristics of message spreading over online social network is important for estimating the influence of message initiated from arbitrary users. Past researches present some models, such as independent cascade model and linear threshold model, to explain the message spreading. Recent studies show many variations of previous models focused on different issues. In this paper, we focus on multiple channels that are used for communicating with each other, and multiple groups that react differently to the message coming from each channel, in order to observe a more detailed aspect of message spreading, such as spreading speed or chances to accept the message. Considering these properties, we propose a new message spreading model that has multiple member groups and multiple channels. We examine the impact of channel and group preference in message spreading by conducting extensive simulations of our suggested model. Through the simulations, we observed that considering multiple channels and multiple groups explains the speed and the coverage of message spreading in more detail.**

Keywords-Message Spreading; Online Social networks; Multiple Channels; Multiple Groups.

I. INTRODUCTION

Observing information diffusion has always been an interesting research area, and, thanks to the Internet, there are more researches coming out, which are related to maximizing the effect of viral marketing [1][2][5], or social influence in Social Network Services [3]. To understand the message spreading in social networks, many models have been suggested and some of them are frequently discussed [1]. Based on each model, many related researches have been done which focus on different aspects of social network such as finding the source of information [6], or finding effective way to spread information [1][2][4]. Also, there are newly suggested models for different attributes and goals [7][8]. We focus on the means which people use in the social network to communicate with others.

With the help of technologies developed recently, the number of ways which people use to communicate with each other is increasing. We call these ways to communicate as channels. Since members of social networks have many choices to send messages, we need to consider the properties of channels to study about message spreading speed. For the cases that require urgent message delivery, like accident or disaster aware services, not only coverage, but also speed of message spreading is an important property. Considering speed, every channel has its own unique properties, such as time it takes to send message or time it takes until receiver checks the message. Texting and phone calls can be examples. Since texting requires typing and cannot be sure whether the receiver checks the message instantly, it has longer time expectation compared to phone call which makes the receiver react instantly. For the message that has time constraints, or to observe the speed of message spreading, considering these properties is needed to properly estimate the diffusion.

While considering channels, we find one more thing to think about, namely, the preference of channel. Channel preference can be different from one user to another. According to this, considering each individual user for applying channel preference is encouraged, but it is practically impossible. As an alternative, we considered a user group that shares similar properties, such as age, income and professions, and the user group explains the characteristics and the preference of channels.

After grouping members in network, their common behavior can be considered. Focused on message spreading, we considered what channels they prefer when sending messages, and from what channel they accept the message and resend to others. As an example, teenagers will prefer using Short Message Service (SMS) [12] or instant messenger to send messages and may have higher chance to accept messages through these channels compared to messages from other channels. When communicating with same groups, this is not an issue. But, when communicating with members in other groups, this can bring a different aspect of diffusion due to their different preference of channels.

In this paper, we propose a new message spreading model which considers the properties discussed above such as multiple channels and multiple groups. In this model, a member who receives the message reacts differently by which channel the sender used and which group the member belongs.

The rest of this paper is structured as follows. In Section 2, related works of message spreading are explored, such as linear threshold model and independent cascade model. Section 3 describes the structure of our proposed model and how this model works related to real situations. In Section 4,

 ISBN: 978-1-61208-361-2

we evaluate our model with the real-world social network data, and we conclude in Section 5.

II. Related Works

There are many models that explain how the information diffusion happens in social networks and two famous models are independent cascade model and linear threshold model. Each model shows a different aspect of influence.

A. Linear Threshold Model

Linear Threshold (LT) Model [1] describes the activation of a node as following the neighbors' major opinion or behavior. In this model, a social network is modeled as directed graph $G = (V, E)$, where the vertices of V represents individuals in the network and edges in E represents relationships and direction of an edge shows who is influenced by whom. Every node *v* in G is in state of either active or inactive and can only be activated once. Also, node *v* chooses a random threshold θ_v that has range of [0, 1].

A node *v* is influenced by each neighbor w according to the edge weight $b_{v,w}$ such that

$$\sum_{w \text{ neighbor of } v} b_{v,w} \leq 1 \quad (1)$$

If the total weight of activated neighbors reaches the threshold θv, the node gets activated and affects other inactive nodes. Being affected by neighbor nodes explains the tendency of adoption of message or product when other neighbors already adopted it. A man will feel like buying something if many of his coworkers already have one. Some researches modified this model to explain product adoption in social network [5].

B. Independent Cascade Model

LT model expresses the diffusion process well, but does not fit into our purpose since its main idea lies in tendency of majority. So, we used the other model that is based on probability, the Independent Cascade (IC) Model [1]. Fig. 1 illustrates the activation process of the IC model. From the initial stage of the diffusion process, every node starts with an inactive state except the nodes in initial node set *A0*. A node can only be activated once, and when it is activated, it has one chance to activate neighbor nodes. In Fig. 1, when node *v* becomes active by the node *x*, it tries to activate each neighbor node *w* and *y* which are in inactive state. By the chance $P_{v,w}$, which is 0.5, *v* becomes active. If node *v* fails to activate node *w*, node *v* cannot try to activate node *w* again. Once node *w* is activated, it can no longer be activated again and stays in activated state. The process continues until there is no further chance to activate neighbor nodes from each node. Unlike the LT model, this model describes the diffusion process as cascading of independent decisions made by each node. Since a node making own decision better fits our idea of separating channels and groups, we modify this model to explain the message spreading over social network, which considers multiple channels and multiple groups.

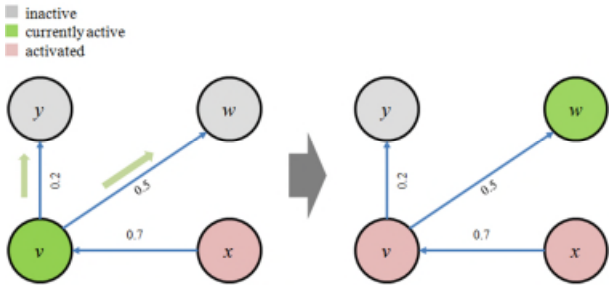


Figure 1. Activation process of Independent Cascade Model

C. Considering Groups in Social Network

Classifying groups in social networks has been researched in order to improve the performance of services in social networks, such as language learning [9][10].

III. Message Spreading Model with Multi-Cahnnel Multi-Group

A. Modeling

To consider channel attributes and preferences, we modify the IC model discussed in Section 2. In our new model, a node *v* belongs to one of groups in the network based on grouping rules, which can be profession, age, or whatever that member of each group can share same channel preference. Each user group g_i, has distinguishing channel preference S_{ij} for each channel c_j to send messages,

$$S_{m\times n} = \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1n} \\ S_{21} & & & \vdots \\ \vdots & & \ddots & \\ S_{m1} & & \cdots & S_{mn} \end{bmatrix} \quad (2)$$

where m is the number of groups and n is the number of channels. Since a node *v* must select at least one channel, $\sum_{j=1}^{n} S_{ij} = 1$. It also has acceptance A_{ij}, the chance to accept and resend the message that comes through channel c_j,

$$A_{m\times n} = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & & & \vdots \\ \vdots & & \ddots & \\ A_{m1} & & \cdots & A_{mn} \end{bmatrix} \quad (3)$$

Here, $0 \leq A_{ij} \leq 1$, since acceptance of each channel works independently. A message sent through c_j has time delay d_j, which is determined by the channel used for activation.

Fig. 2 illustrates what happens in the model when a node *v* tries to activate its neighbor node *w*. When the process starts, node *v* decides which channel to use from c_1 to c_n based on S_{ij}, the channel preference of the user group it belongs. After the node *v* chooses channel c_j, *v* tries to activate a neighbor node *w* through that channel. After the time delay d_j, the time it takes a message to reach its target,

 ISBN: 978-1-61208-361-2

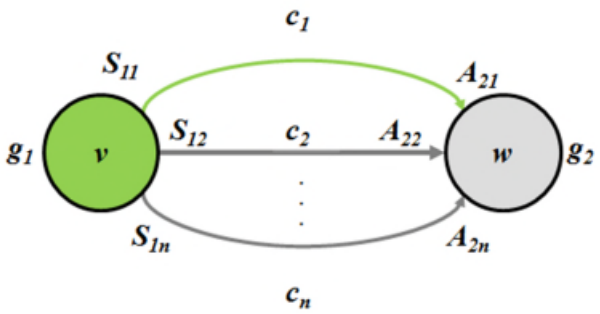


Figure 2. Channel selection and acceptance

the neighbor node *w* determines activation by the acceptance of the channel c_j, A_{ij}. With this individual activation process, Algorithm 1 in Fig. 3 shows the entire node activation process.

B. Correspondence between groups

Intra-group communication, which takes place between members of the same group may differ from that of inter-group communication. People's tendency of relying on the most preferred channel is one of the reasons that change how people react to information from certain channels. Considering correspondence between members of the same group, each member will have higher chance to accept the information from the channel they prefer most. This leads to the assumption that people may show higher acceptance to the information coming from the channel they prefer.

Algorithm 1 Node activation algorithm

```
1:  set time=0
2:  for each node v∈AO do
3:      add v to activation queue Q
        (Q is timestamp priority queue)
4:      set T_v = time
5:  end for
6:  while (Q is not empty)
7:      if time = T_f, timestamp of first node in Q do
8:          set state of v activated
9:          for each neighbor w
10:             generate random number X,
                0 < X ≤ 100
11:             set i = group of v
12:             for each channel c_j
                    if X > Σ_k^{j-1} S_ik and X ≤ Σ_k^j S_ik then
13:                     set edge E_vw = c_j
14:             end for
15:             generate random number Y,
                0 < Y ≤ 100
16:             set i = group of w
17:             if Y ≤ A_ij then
18:                 set timestamp of w = time + d_j
19:                 add w to Q
20:         end for
21:     else time++
22: end while
```

Figure 3. Node activation algorithm

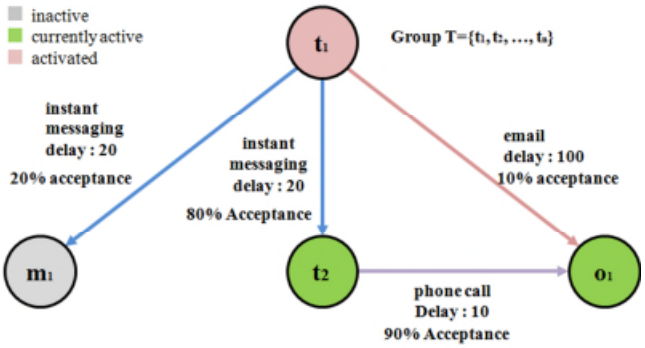


Figure 4. An example of MCMG model

TABLE I. CHANNEL PREFERENCE OF GROUPS

Channels	Channel selection			Acceptance		
	Instant messaging	*Phone call*	*Email*	*Instant messaging*	*Phone call*	*Email*
Group T	0.5	0.2	0.3	0.8	0.15	0.2
Group M	0.3	0.5	0.2	0.3	0.6	0.3
Group O	0.1	0.8	0.1	0.1	0.9	0.1

Fig. 4 shows an example of the activation process of the Multi-Channel Multi-Group (MCMG) model that has three groups and three channels. Based on the assumption that intra-group acceptance is strong and people rely on what they use most, each group shows highest acceptance on the channel which has highest selection chance. Following the group channel selection rate in Table 1, member t_1 in group T shows highest usage of instant messaging channel. After member t_1's attempt to activate inactive member m_1, t_2, and o_1, each member accepts or ignores the message based on acceptance chance in Table 1. t_2 shows 80% acceptance for the message through instance messaging channel while m_1 shows lower acceptance of 20% for the message from same channel. Group O as a group of members over age of 70, shows lower acceptance on email since it is possible that they cannot check the email at all but they show high acceptance in classical way of communication such as phone call. The reaction difference to each channel shows why we need to consider groups more seriously when we consider channels.

C. Simplified model of MCMG model

In some of previous researches, simple Independent Cascade Model that has single channel and single user group

 ISBN: 978-1-61208-361-2

has been used to get the result of experiment due to calculation time and complex algorithm when considering complex model. A simplified version of our MCMG model is needed for those reasons, and for comparing MCMG model's aspect of message spreading with simple Independent Cascade Model. To simplify the model, we consider simplifying channels into one single channel that presents all other channels.

$$C = \{c_1, c_2, \cdots, c_n\} \qquad C \rightarrow c \qquad (4)$$

$$\text{delay } D = \{d_1, d_2, \cdots, d_n\} \qquad D \rightarrow d_{avr} \qquad (5)$$

The average delay of channel becomes delay of the simple channel. Also, due to the simplifying channels, preference of channels changes into

$$S_{m\times n} \rightarrow S_{m\times 1} \qquad A_{m\times n} \rightarrow A_{m\times 1} \qquad (6)$$

where preference $S_{m\times 1}$ and $A_{m\times 1}$ has value of average preference for every channel.

Simplifying groups into one whole group should consider preferences conversion that fits to one group.

$$G = \{g_1, g_2, \cdots, g_m\} \qquad G \rightarrow g \qquad (7)$$

$$S_{m\times n} \rightarrow S_{1\times n} \qquad A_{m\times n} \rightarrow A_{1\times n} \qquad (8)$$

where simplified preferences

$$S_n = \text{avr}\{S_{1n}, S_{2n}, \cdots, S_{mn}\}, \text{ and} \qquad (9)$$

$$A_n = \text{avr}\{A_{1n}, A_{2n}, \cdots, A_{mn}\} \qquad (10)$$

show general channel selection and acceptance when group ratio in the networks are some.

If group ratio is considered,

$$R_{1\times m} = [R_1 \quad R_2 \quad \cdots \quad R_m] \qquad (11)$$

$$RS_{mcmg} = S_{simplified} \qquad (12)$$

where R is the group ratio.

This model simplifies all nodes into one group and all channels into single channel that has general tendency. The model is used for comparison in evaluation.

IV. EVALUATION

A. Setup

1) Dataset

We use real world graph data set from Slashdot [11] to simulate our MCMG model. Slashdot is a technology-related news website known for user-submitted technology oriented news and it has Slashdot Zoo feature, which allows users to tag each other as friends or foes. We use this friend/foe links between users of Slashdot as our dataset which is obtained in November 2008.

Every node in the graph has its unique ids and randomly assigned to one of groups based on ratio of the groups. The first message starts from the fixed initial node.

TABLE II. NETWORK GRAPH STATISTICS

Dataset statistics	
Nodes	77360
Edges	905468
Diameter(longest shortest path)	10
90-percent effective diameter	4.7

TABLE III. CHANNEL SELECTION PROBABILITIES OF GROUPS

Channel Selection	**Channel1**	**Channel2**	**Channel3**	**Channel4**
g_1	0.70	0.20	0.5	0.5
g2	0.15	0.60	0.15	0.10
g3	0.10	0.40	0.40	0.10
g4	0.10	0.15	0.60	0.15
g5	0.10	0.10	0.10	0.70

TABLE IV. ACCEPTANCE PROBABILITIES OF GROUPS

Acceptance	**Channel1**	**Channel2**	**Channel3**	**Channel4**
g1	0.4	0.10	0.5	0.5
g2	0.10	0.40	0.10	0.5
g3	0.5	0.25	0.25	0.5
g4	0.5	0.5	0.40	0.5
g5	0.5	0.5	0.10	0.40

We consider 5 groups,

$$G = \{g_1, g_2, g_3, g_4, g_5\}$$

and delays of channel, D, where

$$D = \{80, 150, 300, 450\}$$

to see the variations of channels and groups. Delay differences between channels show the time differences caused by sending and receiving. As an example, email usually has long delay due to their time taken from sending and checking the message while calling on a phone has relatively short delay. We consider 4 channels in this case.

Table 3 and Table 4 show the probabilities of channel selection and acceptance, respectively. Each group has been set to have higher chances of acceptance to the channels they show more selection chances, due to the assumption we have made in modeling.

 ISBN: 978-1-61208-361-2

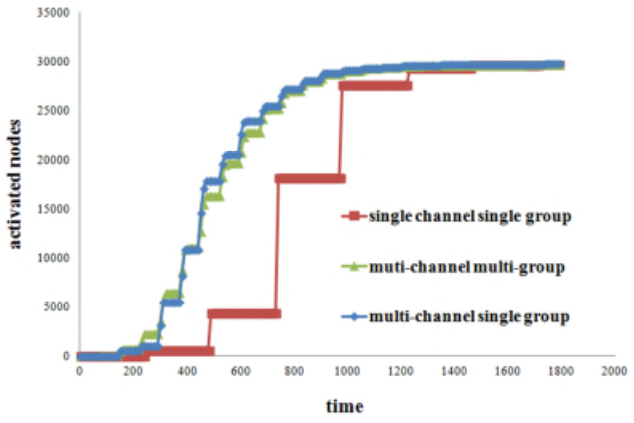


Figure 5. Activation progress of MCMG model and simplified model

2) Simulation

We implemented a simulator that examines activation of nodes per time based on the proposed model. It starts from a set of initial nodes and tries to activate the neighbor nodes. When the node is successfully activated, the status of node changes to activated and the node is stored in priority queue with time stamp value of current time plus delay of channel used. When the time reaches the value of time stamp, the node tries activation of its neighbor node and deleted from the queue after the process. Time starts from 0 and keeps increasing until there is nothing in the queue. The result contains time and total activated nodes at that time.

B. Result

1) Multiple channels

Fig. 5 shows the coverage of message spreading with 3 different models, as a function of time. "multi-channel multi-group" represents the proposed model. "multi-channel single group" represents a simplified model which merges all groups into a single user group. "single channel single group" means the simplified model described in section 3.C. In this evaluation, the percentage of each group is set to equal, 20%. As shown in Fig. 5, cases using multiple channels show similar result, the fluent curves. Multiple choices a node can make brought diversity of time delays that causes different progress in diffusion. But the case that has only single channel shows step shaped line due to lag of variance in delay. Though it has same coverage (the total number of activated nodes), it shows different speed in the middle of spreading message.

This can be critical under the circumstances that have time constraint, or that require intermediate values since those cases require spreading speed at the certain point of time. Short period marketing, such as limited sale policy, can use the proposed model to get proper analysis of effect. Also, analyzing warning message spreading such disaster-aware service, which is time sensitive, requires activation rate at certain point too.

The simplified model (single channel single group) is not appropriate for observing detailed progress of diffusion. It might save the calculation time to conduct message spreading process and to find the eventual coverage, but it does not consider the diversity that it can easily miss few important characteristics of channels. Also as the IC model, the MCMG model surely follows the concept 'independent' since every node decides what channels to use on its own.

2) Multiple groups

To observe the impact of groups to message spreading, we tried different user ratio for each groups. Fig. 6 shows the result of considering group ratio in network. In Fig. 6(a), group1 is the major group which prefers fastest channel such as internet phone and SNS instant messaging. In Fig.

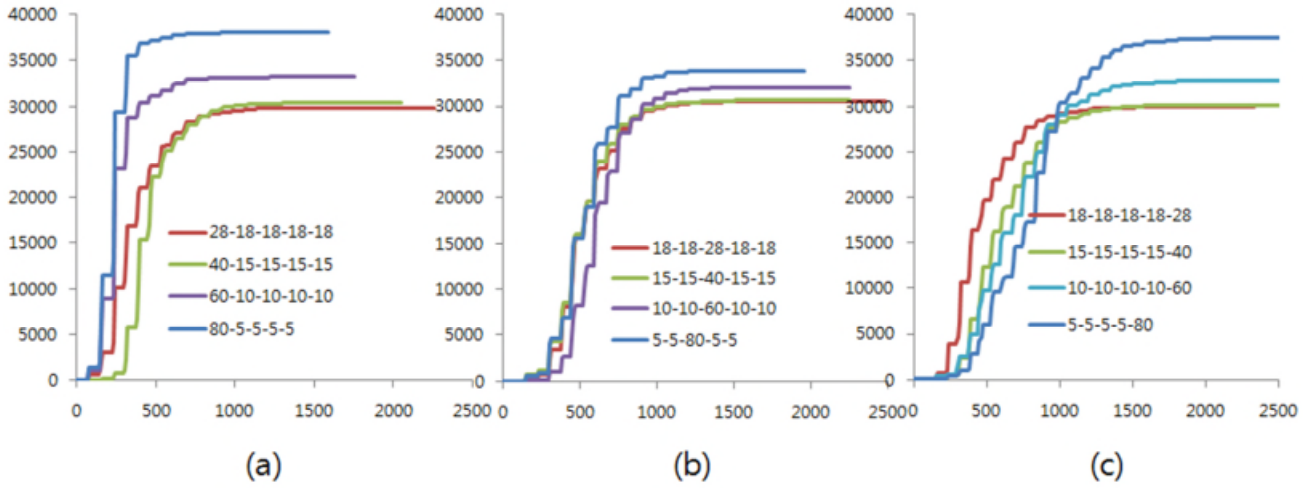


Figure 6. Message spreading under various group ratios. Number in each legend represent percentage of groups by sequence

6(c), major group is group5 which prefers slowest channel such as SNS posting (not instant messaging). In Fig. 6(b), major group is group 3 whose preference is in the middle between group 1 and group 5. In Fig. 6(a), as the percentage of group 1 increases, the speed of message spreading increases. On the other hand, in Fig. 6(c), as the percentage of group 5 increases, the speed of message spreading decreases. That is, the speed of message spreading depends on which channel the major group prefers.

Besides the speed of message spreading, there is one more issue to discuss, the coverage. In Fig. 6(a) and Fig. 6(c), it is observed that the coverage of message spreading increases as the percentage of major group increases. It is because that the intra-group acceptance set higher than other in the current evaluation setting. With the same reason, when the major group is group 3 which prefers two channels, the coverage is less than other two cases like Fig. 6(b). These evaluation results show that the coverage of message spreading may be highly affected by the ratio of user group and preferences of groups.

These results lead us believe that considering groups is an important issue when those groups have their distinguishable channel preference. Since online social network in these days consist of many groups with unique attributes of their own, examining channels they prefer can help us to follow complex message spreading in the network. Also, predicting message spreading without considering groups can cause an overestimation or underestimation compared to which considers groups.

V. CONCLUSION AND FUTURE WORK

Nowadays, the number of channels people use for communication is increasing with the fast development of messenger programs and technology. Google Talk, Skype, or Face Time can be the examples. These channels have unique characteristics that should be considered when observing message spreading over social networks. In addition, every social networks and groups in the network have preferred channel of their own and this can make the relation between channels and groups. The key point is that each group reacts differently to message sent through each channels. This affects not only the speed of message spreading, but also coverage of the message. For that reason, both channel and group in social network should be considered to examine message spreading.

In this paper, we proposed a new message spreading model over online social network which considering multiple channels and multiple groups. With multiple channels, we considered delay and preference of channels, and we found meaningful results from dividing channels. We expect considering the other properties of channels also will lead to more detailed/accurate aspect of message spreading. Likewise, more detailed grouping policy will be helpful to get various results.

As a natural extension of this work, our future work is finding relations between channels and groups and applying them into information diffusion model, which are important to find more effective ways to spread certain messages over online social network.

ACKNOWLEDGEMENTS

This research was supported by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the ITRC(Information Technology Research Center) support program (NIPA-2014-H0301-14-1014) supervised by the NIPA(National IT Industry Promotion Agency).

REFERENCES

[1] D. Kempe, J. Kleinberg, and E. Tardos, “Maximizing the spread of influence through a social network,” Proc. ACM SIGKDD 2003, pp. 137-146.

[2] W. Chen, Y. Wang, and S. Yang, “Efficient influence maximization in social networks,” Proc. ACM SIGKDD June. 2009, pp. 199-208.

[3] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts, “Everyone’s an influencer: quantifying influence on twitter,” Proc. ACM international conference on Web search and data mining, 2011, pp. 65-74.

[4] A. Najar, L. Denoyer, and Patrick Gallinari, “Predicting information diffusion on social networks with partial knowledge,” Proc. 21st international conference companion on World Wide Web, April. 2012, pp. 1197-1204.

[5] S. Bhagat, A. Goyal, and L. V.S. Lakshmanan, “Maximizing product adoption in social networks,” in Proc. the fifth ACM international conference on Web search and data mining, February. 2012, pp. 603-612.

[6] T. Lappas, E. Terzi, D. Gunopulos, and H. Mannila, “Finding effectors in social networks,” Proc. the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, July. 2010, pp. 1059-1068.

[7] J. Yang, and J. Leskovec, “Modeling information diffusion in implicit networks,” Proc. the IEEE International Conference on Data Mining, 2010, pp. 599-608.

[8] X. Song, Y. Chi, K. Hino, and B. L. Tseng, “Information flow modeling based on diffusion rate for prediction and ranking,” Proc. the 16th international conference on World Wide Web, May. 2007, pp. 191-200.

[9] C. Troussas, M. Virvou, J. Caro, and K. J. Espinosa, “Language learning assisted by group profiling in social networks,” iJET, 8(3). 2013, pp. 35-38.

[10] C. Troussas, M. Virvou, J. Caro, and K. J. Espinosa, “Multivariate clustering for group language learning in Facebook,” IIMSS, 2013, pp. 80-88.

[11] Slashdot. [Online]. Available from http://slashdot.org [accessed May 2014]

[12] A. Stoica, Z. Smoreda, and C. Prieur, “A Local Structure-Based Method for Nodes Clustering: Application to a Large Mobile Phone Social Network”, The Influence of Technology on Social Network Analysis and Mining, Lecture Notes in Social Networks 6, pp. 157-184, 2013.

 ISBN: 978-1-61208-361-2

A New Semantic Role-based Access Control Model for Cloud Computing

Masoud Barati
Department of Computer Engineering, Islamic Azad University- Kangavar branch, Kangavar, Iran
emsbarati@yahoo.com

Mohammad Sajjad Khksar Fasaei
Department of Computer Engineering, Islamic Azad University- Songhor branch, Songhor, Iran
sajjadkhaksar@gmail.com

Soheil Lotfi
Department of Computer Engineering, Kermanshah Science and Research branch, Islamic Azad University, Kermanshah, Iran
soheillotfi1983@gmail.com

Azizallah Rahmati
Department of Computer Engineering, Islamic Azad University- Kangavar branch, Kangavar, Iran
m_aziz_rahmati@yahoo.com

***Abstract*- One of the main topics in Cloud computing is access control. Among the approaches of access control in this environment, semantic role-based access control is an interesting issue. In current methods of role-based access control used in Cloud, when a user has no permission for a specific function, its request may be aborted. In this paper, we want to propose a new semantic role-based access control model being compatible with cloud. In our model, a number of functions will be semantically suggested for a user with a certain role. These offered functions can be perfectly used by the user without rejection of its request. In fact, in our approach, by using of two agents called request agent and permission agent, the permissions will be issued based on the semantic similarity between the function asked by a user having a certain role and the predefined functions being in Cloud environment.**

Keywords-Cloud computing; Role-based access control; Ontology; Semantic similarity function; SPARQL query.

I. INTRODUCTION

Cloud computing provides computing services through the Internet. Cloud services let businesses and individuals to tap software and hardware, which are handled by third parties in remote places. For example, these services include file storages, webmail, social networks, and online business applications. With Cloud computing model, users are allowed to access the information and computer resources from everywhere in a network [1].

The service models of Cloud computing are Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS). In the SaaS model, an application, along with any needed software, operating system, and network are provided. In PaaS, an operating system, hardware, and network are provided, and the customer installs or develops its own software and applications. The IaaS model offers only the hardware and network; the client installs or promotes its own operating systems, and software applications [2].

One of the main issues in Cloud is access control. Access control is divided into: Discretionary Access Control Model (DAC), Mandatory Access Control Model (MAC) and role-based access control model (RBAC) [3]-[6]. Among the existing methods of access control, RBAC is the way of permissions combination relying on permissions defined in the functional role. In addition, RBAC models are more flexible than their mandatory counterparts because users can be assigned several roles and a role can be associated with several users.

It is clear that in the Cloud system, autonomous domains [6] have a separate set of security policies. Hence, the access control Mechanism has to be flexible to support various kinds of policies and rules. With the progress of distributed systems, role based access control has become quite significant [7]-[10].

Among some new approaches in the case of access control in Cloud environment, Sun et al. [11] analyzed existing access control methods and presented a semantic-based access control model which considers semantic relations among different entities in Cloud. Besides, Jung and Chung [12] proposed an adaptive security model for Cloud computing environment. The model is based on the improved RBAC model and adapts the role switching model [6].

In this paper, by using semantic descriptions and ontology, we propose a new semantic model for RBAC used in a Cloud environment. In our model, each user requirements and roles are predefined semantically, and agents as brokers are able to give a permission to a user based on semantic similarity function. In fact, allocating a permission is flexible in our approach, since with determining a threshold, a role with the most similar functions set could be dedicated to user instead of simply exact ones. In fact, a user with a certain role may have no permission to an exact requested function, whereas in Cloud may be a number of functions having the most similarities with the function which user asked. Besides, it is possible that these similar functions could meet the user needs. So, in our method, these similar functions can be found and suggested to user by using semantic similarity function.

 ISBN: 978-1-61208-361-2

The structure of the rest of the paper is as follows: in Section II, we present the definitions and primary concepts. Then, in Section III, our semantic model for access control in a Cloud environment is proposed. Finally, in Section IV, the conclusion is highlighted.

II. DEFINITIONS AND PRIMARY CONCEPTS

A. RBAC

RBAC is a method to limit the system access for authorized users [5]. In this respect, access is the ability of a user to perform a specific task, such as delete, create, or update a record. Roles are defined according to job authority, and responsibility in an organization. In this organization, roles are defined for a variety of job functions. The permissions of performing specific functions are devoted to certain roles. The users of system are assigned particular roles, and through such roles assignments get the permissions of computer to perform a group of specific system functions. As users are not directly assigned permission, but only get them through their roles, management of individual user rights becomes a subject of simply assigning appropriate roles to the user's account. This straight forwarded ordinary operations, such as changing a user, or adding a user's institute [6].

RBAC has three primary rules, namely, *Role assignment, Role authorization, and Permission authorization.* In the first rule, if an individual has been assigned a role, he can use a permission. In the second rule, a person's active role must be authorized for that person. Finally, in permission authorization rule, a man can get a permission only if the permission is authorized for the man's active role. This rule ensures that users could get only permission for which they are authorized [3].

B. Ontology and semantic similarity function

Ontology is a formal structure including information about semantic description of data and a group of concepts and the relations between them. It will be used to retrieve information about user requests. A formal definition of ontology [13] in a certain domain, as follow:

$$O=\{C, \leq c, R, \leq r, A\},$$

where C is a set of concepts, R as set of relations, $\leq$ c is an order on C, and $\leq$ r is a partial order on R. In this definition, *A* is considered as a set of axioms [13]-[15].

Semantic similarity function is used for computing similarity between two concepts. The similarity between two concepts illustrates the degree of likeness between them [16]. Similarity function is defined as: sim(x,y):c×c→[0,1]. The result of this function is a real number in the interval [0,1] that shows the rate of similarity between two concepts x, y. In this case, zero means no similarity and one indicates complete similarity between the two concepts [16]-[19]. We compute semantic similarity based on the method from [18]:

$$sim\,(x,y) = \rho \frac{|\alpha(x) \cap \alpha(y)|}{|\alpha(x)|} + (1-\rho)\frac{|\alpha(x) \cap \alpha(y)|}{|\alpha(y)|} \quad (1)$$

Here, ρ is a real number in the interval [0,1] and it is used to determine the degree of influence of generalizations depending on the hierarchical graph of ontology. Here, we can assume that $is\ \frac{1}{2}$, as there is no difference between sim (x,y) and sim(y,x) in our ontology graph. $\alpha(x)$ is the set of nodes which are upwardly reachable from node *x* in the ontology graph. Also, $\alpha(x) \cap \alpha(y)$ is the reachable nodes which are shared by node *x* and node *y* [20].

For instance, an example of ontology with hierarchical graph is depicted in Fig. 1. It has 7 concepts with *'is a'* relationships.

As indicated in Fig. 1, we define Thing as a root node, and which has sub-nodes including Account, Centralize and Decentralize. Account also includes sub-nodes Short Account, Current Account and Long-term Account.

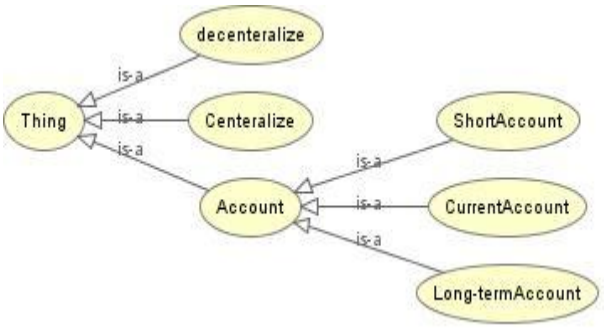


Figure 1. A simple ontology graph of bank account

In case of Eq. 1, the concepts Account and Centralize have 2 reachable upward nodes from themselves. Hence, α(Account)=2 and α(Centralize)=2.
Besides, the similarity of α(Account)∩ α(Long-term Account)=2 is more than α(Account)∩ α(Decentralize)=1.

C. SPARQL

SPARQL [21] is a query language that enables us to retrieve and manage the data saved in Resource Description Framework (RDF) format [22]. The forms of SPARQL queries include a set of triple patterns named a basic graph pattern. In the triple patterns of SPARQL, each of the subject, predicate and object may be a variable. Moreover, SPARQL provides aggregation, sub-queries, negation, and creating values by expressions, and constraining queries with source graph of RDF. The outputs of SPARQL queries could be outcome sets or RDF graphs.

 ISBN: 978-1-61208-361-2

In general, SPARQL graph patterns containing paths are converted to subject-object joins in the SQL [21], and those involving multiple attributes about the similar entity contain subject-subject joins in the SQL.

An example of SPARQL query which models the question of *"What are all the country capitals in America?"* is shown in Fig. 2.

```
PREFIX abc: <http://example.com/exampleOntology#>
SELECT ?capital ?country
WHERE {
  ?x abc:cityname ?capital ;
     abc:isCapitalOf ?y .
  ?y abc:countryname ?country ;
     abc:isInContinent abc:America .
}
```

Figure 2. An example of SPARQL query

A variable is indicated by a "?" prefix, bindings for ?capital and, the ?country will be returned.

III. THE PROPOSED MODEL

A three layers model is presented for our semantic access control model in Cloud computing. As illustrated in Fig. 3, the layers are known as *User Layer*, *Broker Layer*, and *Knowledge Layer*.

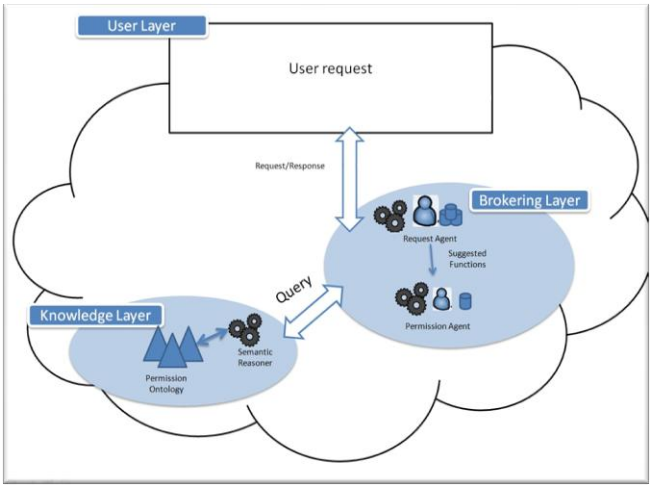


Figure 3. Our proposed semantic model

User Layer

Each user having his role may ask a permission from network to accomplish a function. In this layer, the user request is received and then translated into a format of (*Role, Function*). Following that, the request will be delivered to *request agent*, in the next layer.

Broker Layer

This layer is responsible of getting user requests in the right formats, and issuing permissions for them. In fact, there are two agents called *request agent* and *permission agent* in this layer. The duty of *request agent* is getting the binary set of (Role, Function) from User Layer, and suggesting a sort of functions which are selected based on the semantic similarity function.

To do so, by regarding the ontology graph formed for functions in Knowledge layer, and also by using the semantic similarity function, a matrix of similarities among functions is made by *request agent*. a semantic similarity matrix SIM($n \times n$) can be constructed, as follows:

$$\text{SIM (n} \times \text{n)} = \begin{pmatrix} sim\,(f1, f1) & \cdots & sim\,(f1, fn) \\ \vdots & \ddots & \vdots \\ sim\,(fn, f1) & \cdots & sim\,(fn, fn) \end{pmatrix}$$

Then, this agent based on a predefined threshold (i.e., 0.9), may offer and find more functions having the most similarities with the asked function of user. Following that, it may deliver a number of binary sets of (Role, Function) to permission agent in this layer.

Once Permission agent gets the suggested binary sets of (Role, Function) from request agent, it runs a SPARQL query in the predefined ontology graph in Knowledge layer to search the relationship between the user role and suggested functions. So, the permission of offered functions can be issued, if there are direct relationships between role and functions. Finally, this agent gives the right permissions to the user layer.

Knowledge Layer

In this layer, there is an ontology graph with three primary concepts of permission, roles, and functions. The direct relationships between a role and a function in this graph indicates a permission between that role and function. A general schema of this graph is illustrated in Fig. 4.

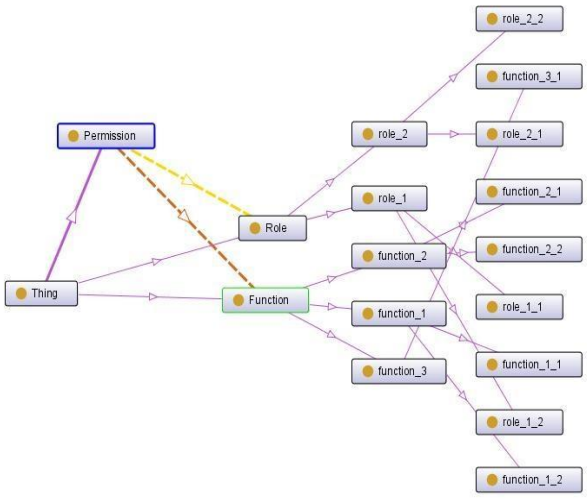


Figure 4. A scheme of ontology graph in knowledge base layer

For example, we assume that a manager in an office wants to calculate his small budgetary computations with an account application A. So, he could join cloud environment with his role, and with his requested function (application A). Then, the user request is semantically translated with the agents in broker layer. In this case, regarding the knowledge layer, should permission agent find the exact application A, then it will issue the permission for the user, otherwise, it tries to find the most similar function for user (e.g., account application B). What is more, this suggested application can properly do the user function.

IV. CONCLUSION AND FUTURE WORK

In this paper, we introduced a new semantic access control model for Cloud computing based on RBAC. In our presented model, the permissions can be assigned to users based on semantic similarity function. In fact, to give a permission, the most similar functions of a role is selected by the agents in broker layer instead of exact ones. So, in our approach, may be found and suggested more than one function for a certain role. Moreover, our model is scalable and it is able to use into different large scale environment.

In future work, we would focus on how we can offer a semantic discovery algorithm to find suggested functions, and we will compare the algorithm with some existing algorithms related our work.

REFERENCES

[1] L. Wang, J. Tao, and M. Kunze, "Scientific Cloud Computing: Early Definition and Experience," in: Proceedings of the 2008 International Conference on High Performance Computing and Communications (HPCC 2008), 2008, pp. 825–830.

[2] L.M. Vaquero, L. Rodero-Merino, J. Caceres, and M. Lindner, "A break in the clouds: towards a cloud definition," In: ACM SIGCOMM. Computer communication review 2009. New York: ACM Press,2009, pp. 50–55.

[3] D.F. Ferraiolo and D.R. Kuhun, "Role Based Access Control," Proceeding of 15th National Computer Security Conference, Baltimore MD, 1992, pp. 554-563.

[4] B. cha, J. Seo, and J. Kim, "Design of Attribute Based Access Control in cloud computing," Proceeding of International conference on IT convergence and Security, Springer. 2011, pp. 41-50.

[5] R. Sandhu, "Role-based access control," In M. Zerkowitz, editor, Advances in Computers, vol. 48. Academic Press, 1998.

[6] R. Sandhu, E. J. Coyne, H. L. Feinstein, and C. E. Youman, "Role-based access control models," 1996. IEEE Computer, 29(2), 1996, pp. 38–47.

[7] L. Obrst, D. McCandless, and D. Ferrell, "Fast Semantic Attribute-Role-Based Access Control (ARBAC) in a Collaborative Environment," The 7th IEEE International Workshop on Trusted Collaboration (TrustCol 2012), October 14–17, 2012, Pittsburgh, PA, 2012.

[8] S. Ullah, Z. Xuefeng, and Z. Feng, "TCloud: A Dynamic Framework and Policies for Access Control across Multiple Domains in Cloud Computing," CoRR abs/1305.2865, vol. 62, no. 2, January 2013.

[9] M. Amirreza and J. Joshi, "OSNAC: An Ontology-Based Access Control Model for Social Networking Systems, Social Computing (SocialCom)," 2010 IEEE Second International Conference on Social Computing, 20-22 Aug. 2010, Minneapolis, MN, 2010, pp. 751 – 759.

[10] C. Ngo, P. Membrey, Y. Demchenko, and C. Laat, "Policy and Context Management in Dynamically Provisioned Access Control Service for Virtualized Cloud Infrastructures," Seventh International Conference on Availability, Reliability and Security (ARES), 2012.

[11] L. Sun, J. Yong, and G. Wu, "Semantic access control for cloud computing based on e-Healthcare," Proceedings of the 2012 IEEE 16th International Conference on Computer Supported Cooperative Work in Design, China, 2012, pp. 512-518.

[12] Y. Jung and M. Chung, "Adaptive Security Management Model in the Cloud Computing Environment," In: 2010 the 12th International Conference on Advanced Communication Technology (ICACT), vol. 2, 2010, pp. 1664–1669.

[13] D. Fensel, "Ontologies: A silver bullet for knowledge management and electronic commerce," Springer-Verlag New York, Inc. Secaucus, NJ, USA, 2003.

[14] H. Stuckenschmidt, "Ontology-based information sharing in weakly structured environments," Ph.D. thesis, AI Department, Vrije University, Amsterdam, 2002.

[15] T.R. Gruber, "Toward principles for the design of ontologies used for knowledge sharing," KSL-93-04, Knowledge Systems Laboratory, Stanford University, 1993.

[16] T. Andreasen, H. Bulskov, and R. Knappe, "From ontology over similarity to query evaluation," in: R. Bernardi and M. Moortgat (Eds.): 2nd CoLogNET-ElsNET Symposium - Questions and Answers: Theoretical and Applied Perspectives, Amsterdam, Holland , 2003, pp. 39–50.

[17] O. Resnik, "Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity and natural language," Journal of Artificial Intelligence Research, vol.11, 1999, pp. 95–130.

[18] R. Richardson, A. Smeaton, and J. Murphy, "Using WordNet as a knowledge base for measuring semantic similarity between words," Tech. Report Working paper CA-1294, School of Computer Applications, Dublin City University, Dublin, Ireland, 1994.

[19] M.A. Rodriguez and M.J. Egenhofer, "Determining semantic similarity among entity classes from different ontologies," IEEE Transactions on Knowledge and Data Engineering, vol. 15, 2003, pp. 442–456.

[20] N. Seco, T. Veale, and J. Hayes, "An intrinsic information content metric for semantic similarity in WordNet," Tech. Report, University College Dublin, Ireland, 2004.

[21] SPARQL Query Language for RDF. W3C Working Draft 4 October 2006. http://www.w3.org/TR/rdf-sparql-query/, 2006.

[22] P. Muster, "Quantitative and Qualitative Evaluation of a SPARQL Front-End for MonetDB," in Department of Informatics, University of Zurich: Zurich, 2007.

 ISBN: 978-1-61208-361-2

Enhancing the Energy Efficiency in Enterprise Clouds Using Compute and Network Power Management Functions

Kai Spindler, Sven Reissmann, Sebastian Rieger
Department of Applied Computer Science
University of Applied Sciences Fulda
Fulda, Germany
{kai.spindler, sven.reissmann, sebastian.rieger}@cs.hs-fulda.de

***Abstract*—Enterprise cloud infrastructures and virtualization technologies constitute a growing proportion in today's data centers. For these data centers the ongoing operational costs are not negligible, especially for electricity which is also increased by cooling. Solutions that raise the energy efficiency allow to reduce these operational costs and to optimize the utilization of the data center infrastructure. The following paper presents a solution to optimize the energy efficiency by observing the current utilization parameters of compute resources and network devices and by taking appropriate actions based on this data. This optimization will be carried out by an automated instance with a comprehensive view on the data center assets, which is relocating virtual machines and optimizing the network structure. The paper presents a lightweight prototype that can be integrated in enterprise cloud environments using standard OpenStack components and application programming interfaces. By monitoring the energy consumption of resources in the environment and combining state of the art in energy-efficient cloud computing with upcoming power management techniques for compute, storage and especially network resources, new possibilities to increase the energy efficiency in enterprise clouds are introduced.**

Keywords–Enterprise Clouds; OpenStack; Energy Efficiency; Computer Networks; Power Management.

I. INTRODUCTION

Enterprise or private cloud solutions are currently gaining more and more momentum, mainly driven by the success of cloud-based services [1] and virtualization, but also by the ongoing eavesdropping scandals that hinder the usage of public cloud providers for sensitive information. One of the major benefits of cloud-based services is formed by their scalability. This scalability is supported by the "elasticity" [2] of the underlying infrastructure that allows providers to support large-scale applications and services for a vast number of mobile devices (e.g., smart phones, tablets) and users from all over the world. However, the improvement in scalability is achieved at the cost of larger data centers and a growing energy consumption. Energy is not only needed to supply the IT infrastructure itself with electricity, but also for appropriate cooling. Hence, energy costs are one of the major challenges for current data centers. Since cloud services are based on distributed systems, besides compute and storage, another essential resource is the network, enabling fast and decentralized access to the services over the Internet and especially the Web. This is also described as "broad network access" in [2]. To provide cloud and web-based services, efficient IT virtualization techniques and computer networks are necessary. These technologies in turn have an impact on energy consumption and cost. Hence, adaptive power management based on the current requirements, i.e., the load on the applications and services, helps to increase the energy efficiency by turning components on and off or reducing their performance (e.g., throttling, energy saving functions). Such adaptive power management functions can also balance or consolidate the power consumption in enterprise cloud environments. As cloud services are provided on an "on-demand" basis according to [2], an adaptive management based on the current load of the resources is supported by this major cloud paradigm.

This paper presents a solution to enhance the energy efficiency in OpenStack-based enterprise cloud environments. A special focus is put on the efficient placement of virtual machines (VM) and the reduction of power required by network connections and components. Adaptive placement of VMs also permits a reduction of compute and storage power consumption by consolidating them on specific hosts, addressing the "resource pooling" requirement for cloud computing environments given in [2]. The paper presents a prototype that was implemented to monitor the energy efficiency (e.g., compute, storage and network utilization as well as temperature and thermal efficiency of the cooling) in cloud environments and throttling, enabling or disabling resources based on the current demand and given constraints (e.g., required fault tolerance, redundancy, quality of service parameters and network connectivity). The prototype uses standard cloud APIs (application programming interfaces) (i.e., OpenStack, Open Cloud Computing Interface – OCCI). Therefore, it can easily be integrated in existing cloud infrastructures using standard OpenStack components.

The paper is laid out as follows. Section II gives an overview on enterprise clouds based on OpenStack and describes the requirements for energy efficiency in such private cloud environments. Also, examples for existing techniques to enhance the energy efficiency in computer networks and references to related research projects are given. Requirements for our prototype, to enhance the energy efficiency by combining the state of the art techniques and extending them, are defined in Section III. The implementation of our prototype and mechanisms to optimize the energy efficiency in enterprise clouds are presented in Section IV. Finally, Section V draws a conclusion, evaluates our research findings and outlines future work that will be pursued in the research project.

II. STATE OF THE ART

The following sections give an overview on the deployment of private clouds using OpenStack and examine the requirements for the energy efficiency of such environments. Additionally, the state of related research projects is discussed.

 ISBN: 978-1-61208-361-2

A. OpenStack-based Enterprise Clouds

The term cloud is an ambiguous concept and has been interpreted in many ways by vendors and customers of cloud services. One of the most sophisticated definitions is documented in NIST SP 800-145, expressing cloud computing as "a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction" [2]. NIST identifies five essential characteristics, three service models, and four deployment models. Our work focuses on private cloud deployments with OpenStack, which is a software project that provides an open source implementation of technologies for building and operating public and private cloud environments using the "Infrastructure as a Service" (IaaS) service model. In OpenStack, this infrastructure is built by offering networking resources (named Neutron), compute resources (Nova), and storage resources, i.e., object storage (Swift) and block storage (Cinder). Additionally, OpenStack is offering many more services for management and orchestration, such as Horizon and Heat, its identity service Keystone, and a telemetry service called Ceilometer.

The IaaS service model in OpenStack is implemented by providing VMs, which can run as Nova instances on the compute nodes of an OpenStack environment. The placement of VMs, being one of the main objectives of our work, can be on a specific Nova node or may depend on various parameters of the environment. Also, the migration of a running VM from one compute node to another or to start or stop VMs depending on the current load is possible during the lifecycle of a service. This flexibility is providing some interesting aspects in terms of resilience (i.e., by seamlessly moving VMs from one data center to another) but also in terms of energy efficiency as we will demonstrate in detail later in this paper.

B. Energy Efficiency in Enterprise Clouds

In todays rapidly growing IT infrastructures, energy efficiency is no longer a secondary requirement, but rather has become one of the main objectives when planning and operating new data centers. A reason for this development is the common sensitization for an ecologically sustainable use of global resources. Furthermore, large-scale data centers are consuming enormous amounts of electrical power not only for running the IT systems, but also for cooling them. A measure for the ratio between the energy used by the computing equipment and the overall energy consumption of a data center is the power usage effectiveness (PUE), which takes into account, i.e., the energy needed for cooling and losses by (uninterruptible) power supply [3]. At the same time, PUE has an impact on the operational overhead cost of a data center, hence its minimization is of great interest for todays data center operators, which have to be economical while facing increasing energy costs [4]. It can be said that cloud computing by definition leads to energy efficiency through its operation concepts, which include a better utilization of physical resources, dynamic scaling based on the current load, and location independent and efficient resource management. However, to take advantage of these concepts, the whole cloud infrastructure needs to be carefully adapted to the operators' individual needs. For instance, resource pooling allows a cloud operator to consolidate multiple VMs providing various services on only a few physical hosts, hence increasing the efficiency of these hosts. At the same time, rapid elasticity and on-demand self-service concepts require the immediate and automatic availability of compute power if needed [2], therefore instant availability of additional resources is required.

The energy consumption of a VM running in OpenStack mainly depends on the energy requirements of its physical IaaS components, including compute (i.e., CPU (central processing unit), RAM (random-access memory)), storage (i.e., SAN (storage area network), NAS (network-attached storage), HDD (hard disk drive)), and networking components (i.e., router, switch), but also on the distance of the involved components (e.g., the distance of the storage from the compute node). Consequently, the real power consumption ratio of a cloud service depends on the number of active compute, storage, and networking components needed to provide it. As VMs can be migrated from one physical host to another, it is possible to take advantage of fluctuating electricity prices or to adapt the load factor of a data center to climatic changes. This could be done not only by consolidating VMs in one data center, but also by sending the VMs to another geographical location, where operation costs are lower. In OpenStack, the placement of VMs on a specific cloud computing fabric controller (Nova) is mainly determined by nova-scheduler [5]. While offering several techniques for optimal VM placement, by default the so called Filter Scheduler is used. It supports the placement of a VM based on a physical location, available compute resources (e.g., CPU, RAM), or by its requirements to secondary resources, such as the availability of a specific storage or network capabilities. Moreover, the Filter Scheduler addresses the operational requirements for resilience or consolidation of VMs by explicitly allowing its placement on different hosts or by grouping them on a single host. However, it does not take into account any energy efficiency parameters, neither for initial placement nor the live-migration of VMs. Also, automatic migration of a VM in favor of load balancing or energy efficiency enhancements is not supported by nova-scheduler. Nevertheless, with its components for service orchestration (Heat) and telemetry (Ceilometer) OpenStack is providing interfaces to manage VM migration that can be extended to evaluate energy consumption or cooling requirements.

C. Energy-efficient Computer Networks

Another aspect to take into account when measuring the energy consumption of a VM running in OpenStack is the networking equipment. According to [6], computer networks typically account for 15–25% of the total energy consumption in data centers. The increasing number of users and the complexity of cloud services require high bandwidth, which leads to increasing link speeds and therefore rises the power consumption of each switch port. Additionally, redundant links are required to assure resilience of the network, again increasing the power consumption. Concepts like Equal Cost Multipathing or Multipath TCP are available to utilize the equipment up to its capacity. However, variable bandwidth requirements (e.g., decreased usage during nighttime) makes it economically reasonable to scale down the network as well. For wired local area networks (LAN), which we primarily focus on, there are already some power management techniques

 ISBN: 978-1-61208-361-2

being offered by the vendors of networking components. First and foremost, the LAN standard 802.3 was extended to include 802.3az, also called energy-efficient ethernet (EEE) [7]. Since this extension is part of the regular 802.3-2012 standard, it is likely that in the near future all equipment will support EEE.

While EEE manufacturers claim that 802.3az allows a reduction of the energy consumed by a single port by up to 81% [8], this benefit comes with the price of increased latency during the low power idle (LPI) phase [9]. Regarding the fact that currently data center network infrastructures are moving to 10 Gbit/s ethernet and beyond, where power consumptions per port are usually over 5 Watts [8], the power savings for the entire data center infrastructure are even higher. Furthermore, there are other vendor-specific power management functions of networking components (i.e., Cisco EnergyWise [4]) that are not covered by EEE. Compared to power management functions of compute and storage resources (e.g., APM, ACPI), that have constantly evolved over the last decades, power management functions for network components are relatively new and will supposedly be improved due to energy efficiency requirements in the near future.

All of the existing solutions are able to reduce the local power consumption on individual network components and ports, but they are unaware of the current global requirements in the entire network. Therefore, their scope is rather limited and the energy efficiency optimization is rather isolated. Some research projects, notably Stanford's ElasticTree have identified this problem, but did not integrate it with an appropriate placement of VMs and especially did not discuss the requirements of enterprise clouds [10]. By using a network controller that is aware of the entire topology, such links could be disabled or throttled during off-peak times while still maintaining fault tolerance requirements. Moreover, such a controller could also activate and deactivate entire networking components based on the current requirements to enhance the energy efficiency. These assumptions and possible solutions are presented in the forthcoming sections of this paper.

D. Related Work

Energy-efficient placement of virtual machines in OpenStack private cloud environments is also discussed in [11][12][13]. However, these approaches do not consider an optimal placement of the VMs with respect to temperature, cooling and network connectivity requirements. Additionally, the extensions presented in these papers cannot be used with the current Havana Release of OpenStack. Furthermore, an integration of additional custom criteria for scheduling decisions regarding the optimal placement of VMs is not supported. A more generalized and detailed evaluation of an energy-efficient placement of VMs in cloud environments and relevant parameters is given in [14][15]. However, these contributions do not offer testbeds for OpenStack environments. Common factors and algorithms to estimate the energy demand of VMs and their migration are discussed in [16].

Concerning energy-efficient computer networks, especially the ElasticTree project [10] presented interesting starting points and related work for power management and throttling of network components using OpenFlow. The ideas of ElasticTree were extended, e.g., in the ECODANE project [17] to include traffic engineering. Also, theoretical energy-aware optimizations of data center networks were presented in [18][6]. Requirements and constraints for energy-efficient placement of VMs regarding the network connectivity were explored in [19][20][21]. However, these solutions do not include existing power management techniques like we described for networking resources (e.g., [6][8][9]) in the previous sections. Furthermore, these approaches do not include power management functions like the Advanced Configuration and Power Interface (ACPI) and related solutions. In our work, we combine the existing power management mechanisms and the solutions that were discussed in the related work given in this section and present a lightweight extension to leverage power management techniques in existing OpenStack enterprise clouds.

III. Energy-efficient Placement and Network Connectivity of Virtual Machines

In the following sections we describe various capabilities of OpenStack regarding the placement of VMs and identify requirements for adding energy efficiency criteria to this process. A special focus is laid on the energy efficiency of the network connection between VMs in distributed enterprise clouds.

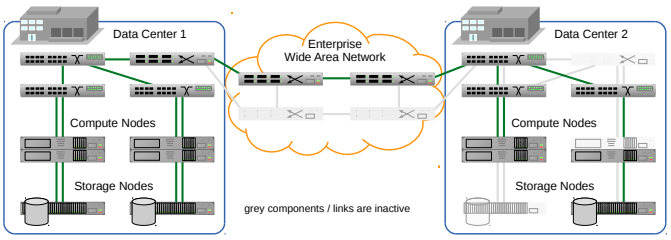


Figure 1. Power management for energy-efficient compute, storage and networking resources in enterprise clouds.

Figure 1 shows an example of an enterprise cloud IT infrastructure that is distributed over two data centers at different sites. Each data center provides compute, storage and network resources as described in Section II-A. Regarding the power management, each of these components consumes energy based on its utilization. Furthermore, as the components are connected to each other over the network, by deactivating or throttling individual components or links, the energy consumption of the enterprise cloud can be reduced, e.g., during off-peak times. Also, redundant components or links can be deactivated completely in favor of increased energy efficiency when active fault tolerance is not needed, e.g., due to low utilization. The deactivation or throttling is symbolized by the grayed out links and components shown in Figure 1.

A. Energy-efficient Placement of Virtual Machines in OpenStack Environments

As introduced in Section II-A, OpenStack is not by itself able to manage resources with respect to the energy efficiency. Therefore, we present concepts supporting the decision about when and how resources like VMs can be relocated to increase the energy efficiency while respecting required dependencies (i.e., storage, network). To decide whether or not to move a VM from one host to another, it is necessary to know various metrics about the system that runs the hypervisor. Basically,

 ISBN: 978-1-61208-361-2

two kinds of metrics are needed to support these decisions. The first are general resource informations, like free RAM, disk space or system load. Using this data it is possible to determine whether the system still has enough free resources, so additional VMs can be moved to this host. A second metric of importance is defined by the temperature and energy consumption of the system, which is closely related to the PUE. Since the current load and the temperature of a system are closely related, it is possible to correlate these metrics, and to draw conclusions about the energy consumption of the system. Another global metric we identified to be interesting to evaluate whether it makes sense to move VMs from one data center to another would be the current local electricity price at a specific site.

Having all these data, it is necessary to select the desired strategy regarding the optimization of the energy efficiency. First, it is a good idea to shutdown a server completely if other servers can provide enough free resources to take over its load. More important, however, it is possible to shutdown the servers switchport to reduce the energy consumed by the network as mentioned in Section II-C. Basically, there are two options to turn servers on and off. The first option is, to control the server using Wake-on-LAN (WOL) if the system was put into ACPI status S3 (Suspend to RAM), S4 (Suspend to disk) or S5 (soft off). Another option is to use IP-based switchable power distribution units (PDU) to switch sockets and attached devices on and off respectively. Using this technique, the BIOS should be configured to automatically boot the system after AC power is restored. Also, entire racks with multiple compute, storage and networking equipment could be powered on and off in a controlled way, if an appropriate mechanism exists to optimize the energy consumptions based on the strategy discussed in this section.

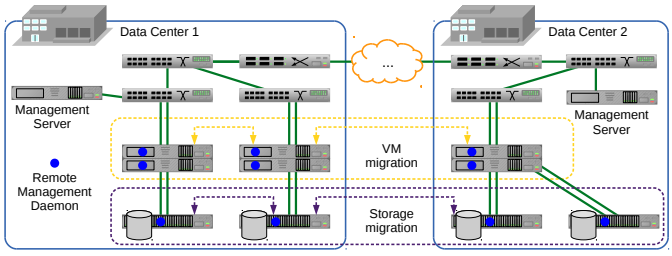


Figure 2. Integration of power management components to enable energy-efficient compute, storage and networking.

As shown in Figure 2, we introduce dedicated management servers in each data center, which have a global view over all servers in the data center. Additionally, management data from other data centers can be synchronized to have the same global knowledge. Each of the management servers is collecting data from the compute, storage and networking nodes in the data center using remote management daemons. Based on this data, they decide when to move the VMs by instructing the involved hypervisors to start a live migration process.

B. Energy-efficient Network Connectivity in OpenStack Environments

The complexity of computer networks with respect to energy consumption can be reduced to consist of nodes and links (Figure 1). Regarding the energy efficiency of a network, two factors driving the energy consumption can be identified. First and foremost, the energy requirements are defined by the number of nodes and links. This especially includes power dissipation at each component. Second, the utilization of each node and each link influences its individual energy consumption. The higher the utilization, the more energy is needed for each component. Nonetheless, a sufficient utilization of all links and components leads to increased efficiency. From a theoretical point of view, the network builds a graph, with each edge representing a link. To include the metrics of each link in the network a weighted graph can be defined, where the weights of the edges represent the load or utilization of the link, its performance (latency, bandwidth, jitter, failure rate) or in our specific example the energy consumption.

By using a graph database, it is possible to model the topology of a network and apply the metrics described above. The network connection of a VM is given by one or multiple paths in the graph. Querying the database, the energy requirements of the network can be evaluated. Also, constraints like fault tolerant links can be defined in the database, as already described in Section II-C. Furthermore, this way the management servers are able to identify redundant links that can be turned on or off depending on the current utilization of the active links or the requirements to resilience. Hence, graph databases can be used to support the decision for energy-efficient placement and network connectivity of VMs. Given the dependencies and metrics represented by weights in the graph, components and links can be deactivated or throttled, e.g., in off-peak times, or reactivated based on network utilization.

IV. Enhancing the Energy Efficiency of Virtual Machines in Enterprise Clouds using AEQUO

Based on the latin word for equal, we named our prototype AEQUO, as it implements a management component to balance the power requirements in OpenStack environments. The prototype is part of a research project at the University of Applied Sciences Fulda with the purpose of creating a proof of concept to enhance the energy efficiency of cloud environments. In this section, we describe the implementation of our prototype based on the requirements that we defined earlier in Section III.

A. AEQUO Testbed Based on OpenStack

For our proof of concept we used Rackspace Private Cloud [22], which provides a fast way to deploy an OpenStack environment with all its components. The installation and configuration of the components is done by Chef, which uses so called cookbooks to deploy the OpenStack services. Our proof of concept uses three virtual machines. Two of them are used as dedicated compute nodes while the other one is used as a hypervisor hosting the rest of the infrastructure. The hypervisor has two VMs running, one serves as the Nova Controller and includes block storage (Cinder), networking (Neutron), dashboard (Horizon), image service (Glance) and orchestration (Heat) components. The other machine serves as the Chef server, which is used for deploying all services during the installation process and later on for adding new components, which makes the system easily expandable. All virtual machines are set up using Ubuntu 12.04 LTS operating

 ISBN: 978-1-61208-361-2

system. To be able to move the VMs from one compute node to another during operation, it is necessary to install a shared storage on the nova controller and all nova compute nodes. The shared storage, which is realized using the distributed scale-out file system Gluster [23], is also used by AEQUO to exchange management information regarding the individual node.

B. Implementation of AEQUO

AEQUO is implemented in Python, which integrates well into the testbed, as most of OpenStack's components are written in the same language and offer a Python API. The current implementation consists of three components. First is the Collector Daemon that runs on each of the compute nodes and is responsible for accumulating performance data, like temperature or energy consumption. In our current implementation we are primarily evaluating the temperature because it is easy to collect for this early approach. The two other components are running on the Nova controller, whereby the Aggregator Daemon is collecting the data received from the Collector Daemon. Additionally, the Aggregator Daemon is writing its data into an SQLite database. Finally, our third component is the Balancing Daemon, which is querying the data from the SQLite database to evaluate it. This historical data is included in the process of making decisions whether or not to move a VM. Figure 3 illustrates AEQUO's components.

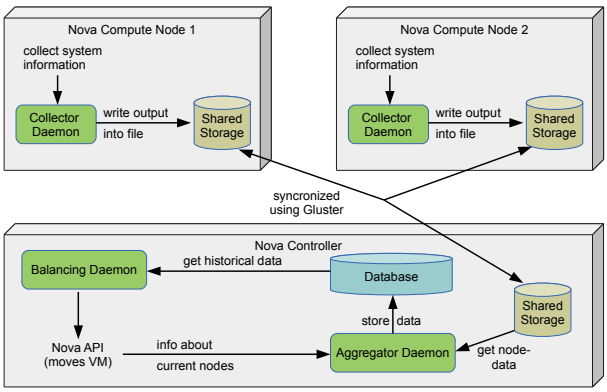


Figure 3. Components and architecture of AEQUO.

In the current version of our prototype, the database consists of three tables, which we illustrate in Table I. The measure table contains historical data from the monitored compute nodes. The table nodedata contains meta information about the compute nodes. Finally, the table vmmove keeps a log providing information about VM movements.

The Collector Daemon running on the compute nodes does not need any configuration, due to the fact that it is just collecting data and writes it to the shared storage. Aggregator Daemon and Balancing Daemon are implemented in a single Python script, as they run simultaneously on the Nova Controller. Using different arguments, the Python script can either be started to run the daemons, test whether the system is running properly or to show the latest data collected. Additionally, an option is offered to insert meta information regarding newly added compute nodes. When started, the script first checks whether the database already exists or needs to be created from scratch. Considering that a significant

TABLE I
DATABASE STRUCTURE USED BY AEQUO.

Table: nodedata

Field	Type	Description
hostname	TEXT	hostname of the compute node
ip	TEXT	IP address of the compute node
status	TEXT	status of the node

Table: measure

Field	Type	Description
hostname	TEXT	hostname of the compute node
time	INTEGER	timestamp of the data
temp1	REAL	temperature of the compute node

Table: vmmove

Field	Type	Description
hostname	TEXT	hostname of the compute node
time	INTEGER	timestamp
moved_from	TEXT	source
moved_to	TEXT	destination

part of electrical energy consumed by computing resources is transformed into heat [11], our current implementation uses simple thermal thresholds over a certain time to decide whether VMs should be moved. However, the prototype can easily be extended to include sophisticated algorithms, e.g., as presented in Section II-D.

C. Optimizing the Energy Efficiency of Virtual Machine Placement and Network Connectivity in OpenStack Environments

As we already mentioned in Sections II-C and III-B, there are also opportunities to reduce the energy consumption of the network components. Using AEQUO with its capability to monitor and control compute nodes, we currently prepare the infrastructure and graph database to extend our prototype to manage network devices. One possible scenario would be to completely power off a 19-inch rack, including all contained networking equipment like the ToR-Switch (top of rack) as well as the cooling for the rack. Therefore, it is necessary to make AEQUO aware of the components in each rack, and the energy consumption of these parts. This is necessary to support decisions, in which the entire load is moved from a rack and it is subsequently shut down. At this point, we are evaluating to include asset or facility management or monitoring tools serving as an additional data source for AEQUO.

Another possibility to save energy is to shutdown redundant paths and network devices or links that are only needed at peak times. The devices could be powered off completely by using power distribution units (PDU) like mentioned in chapter III-A. Alternatively, some network devices (e.g., Cisco IOS routers or CatOS switches) have CLI support to power modules or ports up or down. To use these functions, AEQUO needs to be aware of the network structure, to decide what parts of the network can be powered off. As mentioned above, we are currently implementing a graph database as defined in Section III-B. Instead of shutting down the links completely, network components that support energy-efficient Ethernet (EEE), as described in Section II-C or techniques that control the power used by individual ports of the switch, could also be integrated, e.g., to throttle the link speed or enter EEE's

 ISBN: 978-1-61208-361-2

low power idle mode. As described in Section II-C, the power reduction in this case comes with the drawback of increased latency, which has a negative impact especially on real-time applications. Hence, AEQUO can be used to temporarily turn on EEE and related mechanisms in the networking components when no real-time applications are used (e.g., less usage of VoIP applications or video conferencing traffic during the night). Furthermore, the activation and deactivation of power management mechanisms can also be configured on redundant network paths, as illustrated in Figure 1.

V. Conclusion and Future Work

In this paper, we presented a lightweight prototype to enhance the energy efficiency in enterprise cloud environments that uses standard OpenStack APIs and components. As a starting point, the prototype monitors the temperature and cooling of the components in our cloud testbed, allowing thermal-aware scheduling and migration of virtual machines. Compared to the related work described in Section II-D, our prototype can easily be integrated in OpenStack environments based on the current Havana release. Hence, it serves as a testbed in our research project to evaluate different strategies and state of the art research findings dealing with the energy efficiency in private cloud environments like [14][15]. These techniques can easily be integrated in our prototype thanks to its modularity as described in the implementation section of this paper. On the one hand, we are optimizing the placement and usage of compute and storage resources in OpenStack environments. On the other hand, we focused on network paths including links and devices connecting the virtual machines to the network. While energy-efficient networks were also discussed in [10][6][20], we built our prototype to leverage existing and upcoming local power management techniques of compute and networking components (e.g., [6][8]). This way, for example redundant links in the network can be throttled or even entire devices disabled when network and storage dependencies are integrated into the optimization. We will implement a correspondent scenario in our testbed using our prototype. As a next step of our research, we will measure and evaluate the power savings using the compute, storage and networking equipment in our OpenStack testbed, including models to calculate the cost for virtual machine live-migration [16]. Furthermore, our future work includes the evaluation of benefiting from different energy prices and lower temperature at multiple sites, e.g., to reduce energy costs for cooling. Additionally, we will evaluate the integration of the mechanisms we developed in OpenStack's orchestration framework Heat and the monitoring of energy efficiency metrics in OpenStack's Ceilometer.

Acknowledgment

The authors would like to thank the Hessen State Ministry of Higher Research Education, Research and the Arts for partially funding the research presented in this paper within the "Putting Research into Practice" program.

References

[1] C. Pape, S. Reissmann, and S. Rieger, "RESTful Correlation and Consolidation of Distributed Logging Data in Cloud Environments," in ICIW 2013, The Eighth International Conference on Internet and Web Applications and Services, 2013, pp. 194–199.

[2] P. Mell and T. Grance, "The NIST definition of cloud computing," NIST special publication, vol. 800, no. 145, 2011, p. 7.

[3] A. Greenberg, J. Hamilton, D. A. Maltz, and P. Patel, "The cost of a cloud: research problems in data center networks," ACM SIGCOMM Computer Communication Review, vol. 39, no. 1, 2008, pp. 68–73.

[4] S. S. Sandhu, A. Rawal, P. Kaur, and N. Gupta, "Major components associated with green networking in information communication technology systems," in International Conference on Computing, Communication and Applications (ICCCA). IEEE, 2012, pp. 1–6.

[5] OpenStack, "OpenStack Configuration Reference - Scheduling," 2014, URL: http://docs.openstack.org/trunk/config-reference/content/section_compute-scheduler.html, 2014.05.26.

[6] T. Cheocherngngarn, J. H. Andrian, D. Pan, and K. Kengskool, "Power efficiency in energy-aware data center network," in Proceedings of the Mid-South Annual Engineering and Sciences Conference, May 2012.

[7] D. Valencic, V. Lebinac, and A. Skendzic, "Developments and current trends in ethernet technology," in 36th International Convention on Information & Communication Technology Electronics & Microelectronics (MIPRO). IEEE, 2013, pp. 431–436.

[8] K. Christensen et al., "IEEE 802.3az: the road to energy efficient ethernet," Communications Magazine, IEEE, vol. 48, no. 11, 2010, pp. 50–56.

[9] Intel, "Energy efficient ethernet: Technology, application," 2011, URL: https://communities.intel.com/community/wired/blog/2011/05/05/energy-efficient-ethernet-technology-application-and-why-you-should-care, 2014.05.26.

[10] B. Heller et al., "ElasticTree: Saving Energy in Data Center Networks." in NSDI, vol. 3, 2010, pp. 19–21.

[11] A. Beloglazov, "Energy-efficient management of virtual machines in data centers for cloud computing," Dissertation, Feb. 2013, URL: http://repository.unimelb.edu.au/10187/17701, 2014.05.26.

[12] A. Beloglazov and R. Buyya, "Energy efficient resource management in virtualized cloud data centers," in Proceedings of the 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing. IEEE Computer Society, 2010, pp. 826–831.

[13] ——, "Openstack neat: A framework for dynamic consolidation of virtual machines in openstack clouds - a blueprint," Technical Report CLOUDS-TR-2012-4, Cloud Computing and Distributed Systems Laboratory, The University of Melbourne, Tech. Rep., 2012.

[14] A. Song, W. Fan, W. Wang, J. Luo, and Y. Mo, "Multi-objective virtual machine selection for migrating in virtualized data centers," in Pervasive Computing and the Networked World. Springer, 2013, pp. 426–438.

[15] N. A. Singh and M. Hemalatha, "Reduce energy consumption through virtual machine placement in cloud data centre," in Mining Intelligence and Knowledge Exploration. Springer, 2013, pp. 466–474.

[16] D. Versick and D. Tavangarian, "CAESARA - Combined Architecture for Energy Saving by Auto-Adaptive Resource Allocation," in 6. DFN-Forum Kommunikationstechnologien, 2013, p. 31.

[17] T. Huong et al., "ECODANEreducing energy consumption in data center networks based on traffic engineering," in 11th Würzburg Workshop on IP (EuroView2011), 2011.

[18] X. Wang, Y. Yao, X. Wang, K. Lu, and Q. Cao, "CARPO: Correlation-aware power optimization in data center networks," in INFOCOM, 2012 Proceedings IEEE. IEEE, 2012, pp. 1125–1133.

[19] V. Mann, A. Kumar, P. Dutta, and S. Kalyanaraman, "VMFlow: leveraging VM mobility to reduce network power costs in data centers," in NETWORKING 2011. Springer, 2011, pp. 198–211.

[20] W. Fang, X. Liang, S. Li, L. Chiaraviglio, and N. Xiong, "VMPlanner: Optimizing virtual machine placement and traffic flow routing to reduce network power costs in cloud data centers," Computer Networks, vol. 57, no. 1, 2013, pp. 179–196.

[21] M. A. Adnan and R. Gupta, "Path consolidation for dynamic right-sizing of data center networks," in Sixth International Conference on Cloud Computing (CLOUD). IEEE, 2013, pp. 581–588.

[22] Rackspace, "OpenStack Private Cloud Software," 2014, URL: http://www.rackspace.com/cloud/private/openstack_software/, 2014.05.26.

[23] RedHat, "GlusterFS," 2014, URL: http://gluster.org/community/documentation/index.php/OSConnect, 2014.05.26.

 ISBN: 978-1-61208-361-2

Return the Data to the Owner: A Browser-Based Peer-to-Peer Network

Dennis Boldt and Stefan Fischer
Institute of Telematics
University of Lübeck
Lübeck, Germany
Email: {boldt,fischer}@itm.uni-luebeck.de

***Abstract*—The paper covers the concept of a browser-based peer-to-peer network, which supports a decentralized, redundant and encrypted data storage. The core is a JavaScript-based socket API, which facilitates creating and accepting arbitrary TCP/IP connections from within a browser. This API builds upon a WebSocket SOCKS5 Proxy. This is essential, because the sandbox of a browser does not allow plain socket connections. We used this Socket API to implement, to the best of our knowledge, the first Browser-based peer-to-peer network based on the Chord protocol. Additionally, we implemented the first JavaScript-based forward error correction based on Reed-Solomon coding to handle the recovery of lost data. Our network circumvents user-generated content stored on powerful central servers operated by huge companies which allows the creation of user profiles, the placement of customized advertisements and a possible interface for intelligence agencies to access the central stored data. Our results show, that our approach works with reasonable performance for files up to 100 KB.**

Keywords-Browser-based peer-to-peer network; Berkeley Sockets API; SOCKS5; Chord; Reed-Solomon.

I. INTRODUCTION

For quite a while now, user-generated content web pages such as Wikipedia, Youtube, Twitter or Flickr are ubiquitous. Moreover, plenty of well-known desktop applications such as several widely-used office suites have been made usable through the browser interface. This works in a way that the owner of the web applications provide the platforms and the users are generating the corresponding content by accessing the applications through their web browsers (or specialized apps) and store it on the servers of the web applications' owners. Many users see this as the core problem: the server providers control the data in such a way, that they can create user profiles and provide customized advertisements. Moreover, they provide embeddable code snippets for like-buttons (e.g., Facebook and Google Plus), videos (e.g., Youtube) or messages (e.g., Twitter) which are embedded in millions of independent web pages [1]. Based on this, it is even more easy to generate perfect customized user profiles, especially if a single company provides dozens of heterogeneous applications which appear independent.

In our approach for a solution, we focus on the browser as an independent platform, because it is one of the most often used applications spread over all operating systems in the world. Furthermore, it is not per se necessary to install an additional runtime environment (e.g., Java or .NET Framework), because browsers support JavaScript out of the box. Finally the browsers are getting faster from release to release and new technologies such as HTML5 are better supported.

This paper provides a way to get rid of the dependency on web applications provided by huge companies. The user-generated data will be stored in an encrypted format, decentralized and redundant in the browsers of the users. The architecture we developed is a browser-based peer-to-peer (P2P) network based on well-known technologies and protocols such as Berkeley Sockets, SOCKS5, Chord, Reed-Solomon coding and HTML5 features like WebSockets.

The rest of this paper is organized as follows. Related work is introduced in Section II, the core architecture of the browser-based P2P network is proposed in Section III. The experimental results are provided in Section IV. Section V gives an outlook on future work. Section VI concludes the paper.

II. RELATED WORK

Classical applications in computer networks are based on client-server architectures, where a central server provides services which are used by many clients. In contrast there are peer-to-peer (P2P) networks, in which every participant is both client and server (peer) at the same time. In the last 15 years, plenty of P2P systems were developed. Well known examples are the first generation P2P networks Napster (1999) with a centralized lookup, or Gnutella (2000) with a decentralized flooding-based lookup.

Both approaches do not scale well which led to the development of a second generation of P2P networks like Tapestry [2] or Chord [3]. These approaches are more structured and, as a result, much more scalable. Also, P2P-based distributed storage systems were invented, where systems like OceanStore [4], Cooperative File System (CFS) or Wuala [5] were created. The latter is based on Network Coding [6].

A. Chord

A common structured P2P network protocol is Chord [3]. $P = \{p_1, \ldots, p_N\}$ is the set of N peers within the network. A hashing function h assigns to every peer $p \in P$ an unique ID $h(p) \in \{0 \ldots 2^m - 1\}$, where m should be sufficiently

 ISBN: 978-1-61208-361-2

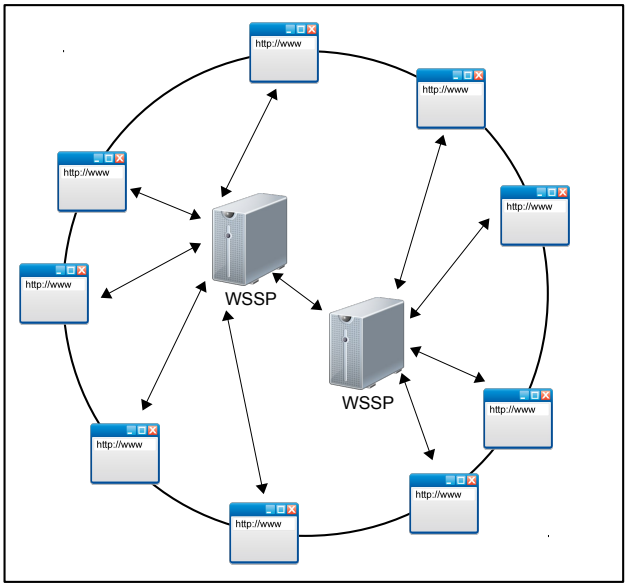


Figure 1: High level architecture: the browser-based peer-to-peer network based on a Chord-ring

huge. The ID for a peer is calculated by hashing the peer's IP address and port number, for example with SHA-1, where m is 160 bit. Based on these IDs, all peers are placed on a ring with a corresponding key space $\{0 \dots 2^m - 1\}$, the so-called Chord-ring. Based on this ring topology, peer p is in charge for the key space between the predecessor's ID and the own ID: $(h(pred(p)), h(p)]$.

The core operation of a P2P network is to find the node n which is in charge for a given m-bit key key; in Chord this can be expressed as $n = succ(key)$. This operation can easily be implemented by walking along the ring of peers in $O(N)$ steps. Chord improves this inefficient lookup to $log(N)$ steps with an additional routing table, the so-called *finger table*. Each finger table has m entries (fingers), where each $finger(i)$ bridges a distance of 2^{i-1} on the ring. Thus, the first finger is the successor and the last finger points to a peer which is at least half the ring away. The Chord protocol also handles joining and leaving of peers. To join to the network, a peer must know an existing peer from the network (bootstrapping). When a peer joins or leaves, the corresponding key spaces are changing.

All applications must be implemented on top of this routing protocol. Chord's core application is a *Distributed Hash Table* (DHT) for storing application data in the network. A DHT supports two operations, where *key* is a SHA-1 hash and *value* is an array of binary data:

1) *put(key, value)*
2) *value = get(key)*

Because the DHT and Chord are using the SHA-1 hashing function, both are using the same key space. Therefore, putting and getting data yields to the function *succ(key)* provided by Chord.

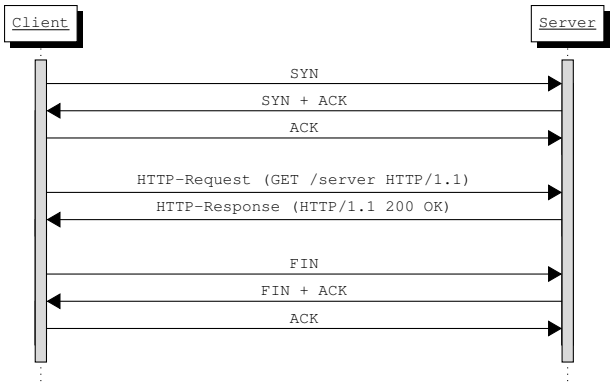


Figure 2: Schematic sequence diagram showing an XMLHttpRequest exchange.

B. WebRTC

WebRTC is a JavaScript API which enables web browsers to have real-time communications (RTC) [7]. WebRTC supports P2P audio and video communications. It also supports P2P data channels to send binary data between two peers. To create a P2P connection, WebRTC does a connection handshake between two peers over a signaling channel based on XMLHttpRequests or WebSockets using the Session Description Protocol (SDP) defined in RFC 4566 [8]. The WebRTC specification itself is still work in progress. To enable P2P connections between peers, some simple WebRTC-based P2P APIs like PeerJS [9] were developed. Current research topics on WebRTC are media streaming [10][11][12] and browser-based Content Delivery Networks (CDN) like Maygh [13], Tailgate [14] or PeerCDN [15].

III. System Architecture and Implementation

The architecture of our browser-based P2P network is shown in Figure 1. Because we implemented the Chord protocol, all browser-peers are placed on a ring. To set up the network, it has to be possible to create a socket connection to another browser, and to bind a socket to listen for incoming connections. This leads to the core problem: every browser-based application runs in a sandbox which has no capability to handle raw TCP/IP sockets. There are three ways for a browser to communicate to the world: *WebRTC*, *XMLHttpRequests* and *WebSockets*. The first was mentioned before, therefore we are focusing on the last two.

XMLHttpRequest Level 2 [16] (XHR) allows a web application to send HTTP requests to a server asynchronously. It is located on top of HTTP in the TCP/IP protocol stack. A connection starts with a kind of a HTTP-based handshake. Because HTTP is on top of TCP, this leads to a TCP handshake followed by an HTTP request. The HTTP response (typically containing XML, JSON, HTML, or binary data) from a server can be embedded in the application without reloading the web page. Because HTTP is a request-response protocol, a web application needs to request a resource continuously to get real time updates (polling). The message sequence for an XHR exchange is shown in Figure 2.

The WebSocket Protocol is defined in RFC 6455 [17] and allows real two-way communications. A connection starts

 ISBN: 978-1-61208-361-2

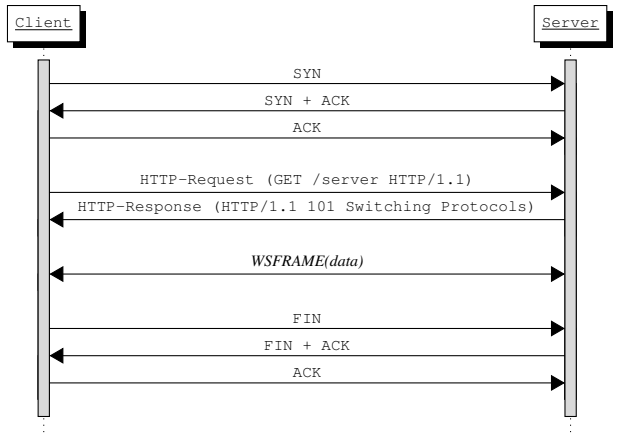


Figure 3: Schematic sequence diagram showing a connection with WebSockets

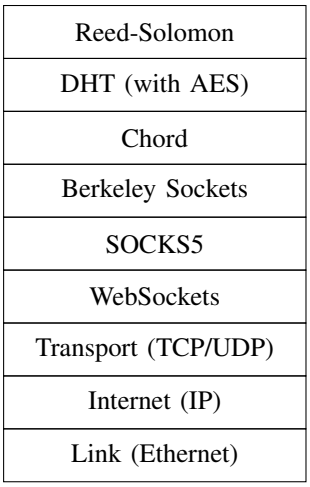


Figure 4: Low level architecture: Technology stack of the browser-based peer-to-peer network.

similar to a XHR connection. The difference is, that the HTTP response from the server signals that client and server are switching the protocol to WebSocket Frames (`HTTP/1.1 101 Switching Protocols`). These frames are similar to TCP segments. As soon the connection is initiated, client and server can send binary data, i.e., HTML5 Typed Arrays [18], to each other immediately. A typical WebSocket connection is shown in Figure 3.

The *SOCKS Protocol Version 5* (SOCKS5) is a protocol defined in RFC 1928 [19] to realize a proxy server. A proxy forwards connections of clients to servers and vise versa. SOCKS5 supports TCP/UDP, IPv4/IPv6 and authentication. It starts with an handshake to exchange the authentication methods with the *identifier/method selection message*. Afterwards a *socks request*, e.g., CONNECT/BIND request, with a given IP address and port number is send to the proxy which creates a connection or binds a socket. In the TCP/IP stack, SOCKS5 operates between the transport layer and the application layer.

A. The WebSocket SOCKS5 Proxy (WSSP)

The WSSP is the core component in our architecture which is implemented in Java and uses the Netty framework [20]. Netty supports zero-copy and uses New I/O (NIO), which is a non-blocking I/O based on Buffers and Channels. Netty perfectly fits in or architecture, because it already provides handlers for WebSockets. On top of the WebSockets, we implemented the SOCKS5 protocol to allow arbitrary TCP/IP connections.

Connect to a socket: To connect to a socket, a WebSocket connection to the WSSP must be established by a client. Afterwards, the WebSocket connection is used to send the method selection message followed by the CONNECT request. The WSSP connects to the given IP/port and sends a SOCKS5 reply over the WebSocket connection to the client. Finally, the connection is established.

Bind to a socket: To bind to a socket, a WebSocket connection to the WSSP must be established by a server. Afterwards, the WebSocket connection is used to send the method selection message followed by the BIND request. The WSSP binds to the given IP/port and sends a SOCKS5 reply over the existing WebSocket connection to the server and the binding is established. The Chord-ID is calculated by hashing this IP and port with SHA-1. On an incoming connection, the WSSP sends a new SOCKS5 reply over the existing WebSocket connection to the server. The server creates a new WebSocket connection to WSSP. The WSSP binds this new WebSocket connection to the incoming connection and a separate channel for each connection is established.

B. JavaScript APIs

In the following, we are going through the layers of our technology stack shown in Figure 4. We assume the link, internet and transport layer to be well-known and we focus on the application layers on top of the transport layer.

1) The WebSocket API: The WebSocket API [21] implements the WebSocket Protocol, which is located on top of HTTP; after switching the protocol on top of TCP/IP. The WebSocket API is supported by all common browsers.

2) JavaScript SOCKS5 API: Together with the WSSP, we implemented an appropriate JavaScript SOCKS5 API, which builds upon WebSocket API. With this API it is possible to create socket connections and to listen for incoming connections within sandboxed web applications.

3) JavaScript Berkeley Sockets API: The *Berkeley Sockets API* (BSD API) [22] provides well-known functions like `socket()`, `connect()`, `bind()` or `close()` to enable inter-process communication (IPC) for any application via TCP/IP. Figure 5 shows how this functions can be used by a server to bind to a socket and for a client to connect to a socket. The original BSD API is written in C and is part of every UNIX system. The BSD API is situated in the TCP/IP stack between the transport layer and the application layer as well. We adopted the BSD API on top of the JavaScript SOCKS5 API. With our JavaScript Berkeley Sockets API it is easy to create IPC within the browser.

 ISBN: 978-1-61208-361-2

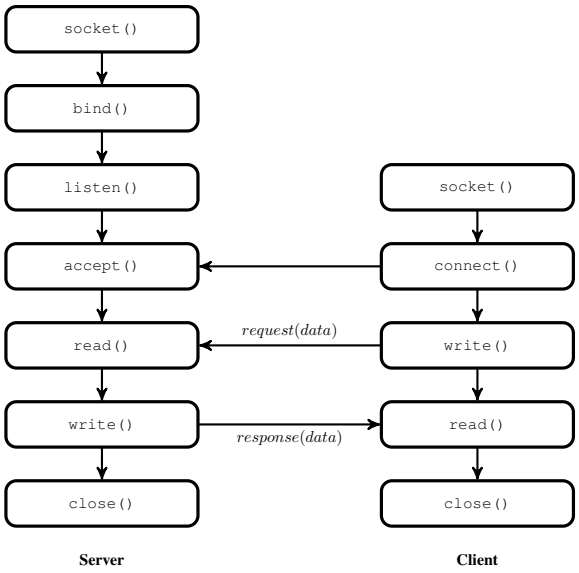


Figure 5: Client and Server with the Berkeley Socket API (based on [23])

4) JavaScript Chord API: With the *JavaScript Chord API* we implemented, to the best of our knowledge, the world-wide first JavaScript-based P2P network on top of the JavaScript Berkeley Sockets API. To join the network, we implemented a simple lookup server where all nodes which have joint are registered.

5) JavaScript Distributed Hash Table API: Our JavaScript Distributed Hash Table API uses the JavaScript Chord API to provide the two DHT operation *put* and *get*. It supports symmetric encryption and decryption of data using AES with a given *key_symm*, e.g., a password. This symmetric key can differ from data to data. The encryption and decryption are calculated on the client side. Therefore, the password never leaves the computer of the client. The key, which represents the storage position of the data in the Chord-ring, can be determined by calculating the SHA-1 hash of the data. The packet format can be seen in Listing 1. For transferring, this packet is converted into binary format using the library binarize.js [24]. The data are stored in-memory by using a JavaScript object.

```
{
  key : sha1(data),
  value : aes_enc(data, key_symm)
}
```

Listing 1: The DHT packet in JSON format. It contains the key and the encrypted data.

6) JavaScript Reed-Solomon API: We showed, how we store the application data in a decentralized and encrypted way in the users' browsers. In huge networks, nodes typically join and leave rather frequently. It is also possible that a node fails. In this case, all data stored at this node is lost. To ensure the availability of all data stored in the P2P network, identical copies located on different nodes, e.g., replicates located at different hashes, can be used. It is quite unrealistic to upload the same file multiple times to the network. Even the probability of losing all replicates is quite high. Therefore, our JavaScript Reed-Solomon API uses the Reed-Solomon Coding defined as polynoms on finite fields [25][26][27]. It is a so-called erasure code, which does forward error corrections on binary data. A file is split in n data parts. Based on these n data parts, m linear independent recovery parts are calculated. The coding uses Galois Fields – $GF(2^8)$ – for binary data, on which the mathematical operations $+$, $-$, $*$, $/$ are defined. To recover the whole file, just a subset of n arbitrary parts from all data and recovery parts is needed. This coding was first used for CDs.

IV. Experimental Results

Imagine the following use case: an user or a web application wants to store/load a file, e.g., a picture, within/from the DHT (see Figure 6). A password can be selected.

Putting data into the DHT: First of all, the application calculates a corresponding SHA-1 hash. Afterwards the Reed-Solomon encoding is called, which splits the file in n data parts and calculates m recovery parts. If a password is selected, these parts are encrypted using AES. Otherwise this step is skipped. Finally all (encrypted) parts are uploaded to the P2P network using the DHT.

Getting data from the DHT: Obtaining a file works the opposite way: an application downloads n arbitrary parts from the P2P network using the DHT. If a password is given, the AES decryption is called, which returns the decrypted parts. Then the whole file is decoded using Reed-Solomon. Finally the SHA-1 hashsum can be calculate to check the integrity.

We ran our experiments in-memory with the browsers Firefox 27, Chrome 33 and the server-side environment Node.js 0.10.21. Firefox is based on the JavaScript engine *SpiderMonkey* [28], while the last two are based on the *V8 JavaScript Engine* [29]. Node.js allows to run JavaScript as a standalone application without the need of a browser. Because of this, it also provides raw socket access to the application.

For our experiments we used a Dell Latitude E6420 with an Intel Core i5-2520M CPU with 2.50 GHz and 8 GB RAM. The file size in the experiments varies from 1 Byte up to 1 GB.

A. Network Independent Evaluation

The calculation of SHA-1 hashing values, AES encryptions or Reed-Solomon encodings for big files is quite CPU-intensive. This leads to the problem that the browsers are freezing while an API is working. To remedy this issue, the *Web Worker API* [30] was introduced. This API provides independent threads to JavaScript applications for long-running calculations. We use Web Worker within the browsers for the aforementioned APIs, which are computing intensive algorithms. We also use already existing implementations for SHA-1 [31] and AES [32], the latter is implemented in counter mode. We implemented the Reed-Solomon coding.

From Figures 7 to 9 we can see, that the duration for almost all network independent experiments is slowest in Firefox. Node.js is the fastest for small files, because it does not use Web Worker. When the file size increases, Chrome becomes

 ISBN: 978-1-61208-361-2

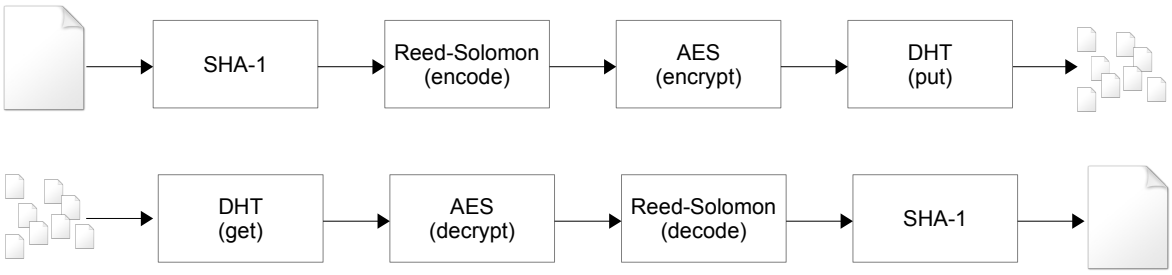


Figure 6: Workflow to put/get a file into/from the DHT.

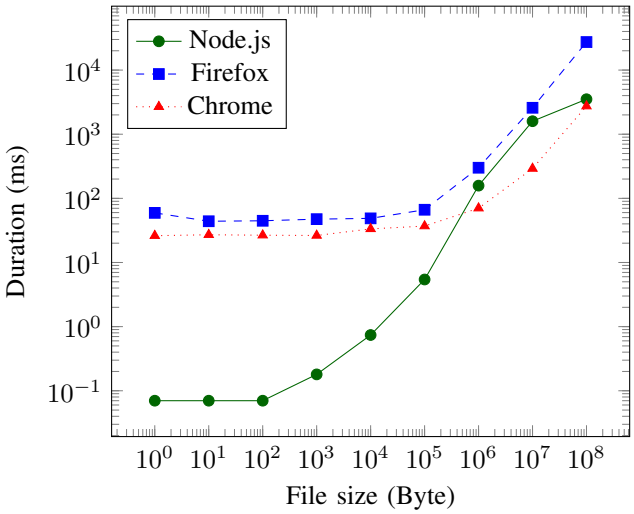


Figure 7: SHA-1 hashsums calculation for 100 files.

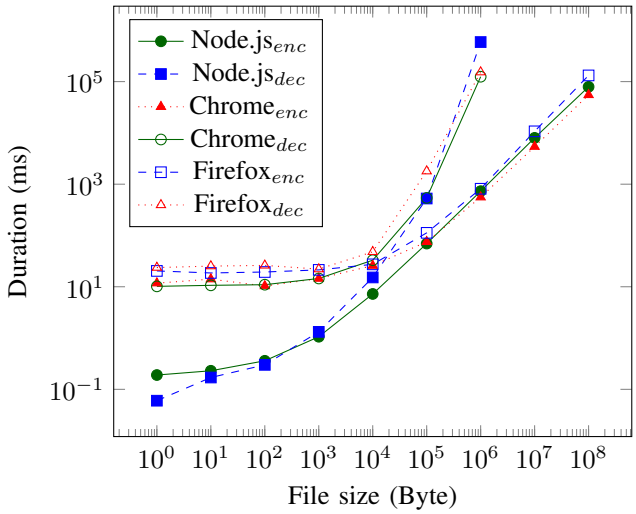


Figure 8: Encrypt/decrypt AES for 100 files with a key size of 256 bits.

faster than Node.js, even though both are using the same JavaScript engine. Figure 7 shows that Chrome is faster than Node.js for SHA-1 hashsum calculations for files bigger than 10 MB. The result for AES with key size of 256 bits can be seen in Figure 8. Up to a file size of 100 KB, encryption and decryption take almost the same time on all platforms (up to 1,7 seconds). For a file of size 1 MB the decryption varies from 2 to 6 minutes, which is not acceptable. Therefore, it is much faster to split a file in smaller parts of maximum 100 KB before the encryption step, than encoding the whole file. This fits in our architecture, because the splitting is done by the Reed-Solomon coding already. Figure 9 shows the result of the Reed-Solomon coding. With file sizes smaller than 1 MB, the coding does not need more than one second.

B. Network Evaluation

First of all we measured the round-trip time (RTT) of a packet with ICMP and afterwards we created a WebSocket connection to the WSSP. We observed that the duration of a WebSocket connection just depends on the RTT of the network. For example: the duration for a WebSocket connection is 80ms. Then the RTT is 40ms, because the WebSocket protocol needs two RTTs to set up a connection: one for the TCP handshake and one for the HTTP handshake (see Figure 3). We ran the following experiments just in Node.js, because the duration does not differ between Firefox, Chrome and Node.js.

We evaluated our WSSP-based approach, which is used in usual browsers, against a version with direct socket access provided by Node js. For this, we implemented a simple ECHO-server, which copies and returns the received message. We created a Chord-ring with two nodes, connected to the same WSSP with a RTT of 40ms.

WSSP-approach: The binding of a socket by a server takes 180ms. This results in four RTTs (160ms): two for the WebSocket connection to the WSSP and two for the SOCKS5 binding. The remaining 20ms are processing time by Node.js and the WSSP. A connection by the client to the bound socket takes 360ms. This yields to 8 RTTs (320ms): two to create the WebSocket connection and two for the SOCKS5 connect request by the client; two for the new WebSocket connection and two for the SOCKS5 request by the server to bind the incoming connection (see Section III-A). The remaining 40ms are processing time by Node.js and the WSSP. The ECHO of the data takes 85ms, which corresponds to 2 RTTs: one between the client and the WSSP and one between the WSSP and the server.

Raw sockets: With raw sockets we get rid of the WebSocket connections and the SOCKS5 messages. The binding happens immediately and the connection needs just one RTT (40ms).

The putting/getting of data with DHT just depends on the

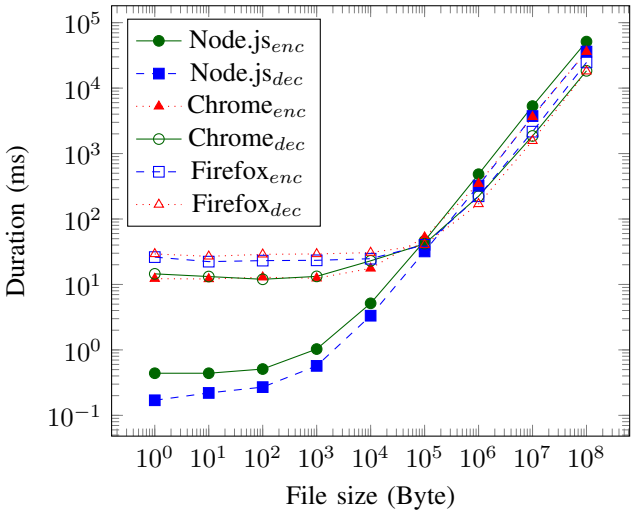


Figure 9: Encode/decode Reed-Solomon for 100 files with $n = 10$ and $m = 10$.

connection duration from our previous experiment and the network bandwidth. Therefore, we did not run a DHT-based experiment, because it will just show the maximum network bandwidth.

V. Future Work

Future work aims to improve the performance, the bootstrapping, the maintenance and the security.

Performance: To increase the performance, we are working on a WebRTC-based P2P network, even if it does not allow arbitrary TCP/IP connections like our approach. Therefore, we expect better experimental results, e.g., less RTTs, in the future.

Bootstrapping of the network: One big challenge is the bootstrapping of the network, because in the Chord protocol every node needs an entry node to join the network. The current implementation uses a simple lookup server where all active nodes are registered. This obviously does not scale in a network with millions of nodes.

Security: Currently we are using symmetric encryption of the data with AES. Thus, sender and receiver must know the *key_symm*, e.g., the password. An improvement will be the usage of asymmetric keys, e.g., RSA [33]. With this, we can update our packet format, where *key_async* can be the own public key (storage of data for the own purpose) or the private key (storage of data for public purpose). Like this, the *key_symm* is stored in the DHT packet itself (see Listing 2). This allow using random symmetric keys for every data part.

Maintenance of the data: In the current implementation, data stored at a node (DHT) is moved, if a node joins or leaves the network. This is a core feature of Chord. This does not scale, if a user must wait until all data, e.g., some hundred MB, are moved. Usually this leads to the failure of the node and the loss of the in-memory stored data (this also include a change of the IP address). Therefore, a challenge is the maintenance of the data stored in the P2P network. The network needs

```
{
  key : sha1(data),
  value : [
      aes_enc(data, key_symm),
      rsa_enc(key_symm, key_async)
    ]
  }
}
```

Listing 2: The improved DHT packet in JSON format. It contains the key, the encrypted data and the encrypted symmetric key.

to make sure that all data are always available, especially in the future, after millions of node joins, leaves and fails. To handle this, we already use the Reed-Solomon coding. Also the Reed-Solomon parts are lost over the time. If less than n parts are available, a Reed-Solomon encoded file cannot be decoded. Therefore, we need to maintain the Reed-Solomon parts, if the availability of a file is vulnerable, e.g., recovering the parts.

VI. Conclusion

In this paper, we introduced the approach of a browser-based P2P network, which is a possible platform to return the data to the owner. The experiments showed that our approach works with reasonable performance for files up to 100 KB, which fits the usual web traffic. Bigger files are split in smaller parts to keep the performance. The performance of the JavaScript engines of all browsers is going to be improved, while the limitation of the computing intensive algorithms SHA-1, AES and Reed-Solomon is mainly the CPU.

Acknowledgements

We would like to thank Philipp Abraham, Tobias Braun, Florian Burmann, Marvin Frick, Bennet Gerlach, Syavoosh Khabbazzadeh, Florian Lau and Dennis Pfisterer for helpful comments, debugging and testing our implementation.

References

[1] J. Schmidt, “Das Like-Problem (The like problem),” Heise Security, 04 2011. [Online]. Available: http://heise.de/-1230906 [Retrieved: May, 2014]

[2] B. Zhao, J. Kubiatowicz, and A. D. Joseph, “Tapestry: An infrastructure for fault-tolerant wide-area location and routing,” University of California Berkeley, Computer Science Department, Tech. Rep. UCB Technical Report UCB/CSD-01-1141, 04 2001.

[3] I. Stoica, R. Morris, D. Karger, M. F. Kaashoek, and H. Balakrishnan, “Chord: A scalable peer-to-peer lookup service for internet applications,” in ACM SIGCOMM Computer Communication Review, vol. 31, no. 4. ACM, 2001, pp. 149–160.

[4] J. Kubiatowicz, D. Bindel, Y. Chen, S. Czerwinski, P. Eaton et al., “Oceanstore: An architecture for global-scale persistent storage,” in Proceedings of the Ninth International Conference on Architectural Support for Programming Languages and Operating Systems, ser. ASPLOS IX. New York, NY, USA: ACM, 2000, pp. 190–201.

[5] “Wuala - Secure Cloud Storage.” [Online]. Available: http://www.wuala.com [Retrieved: May, 2014]

[6] M. Martalo, M. Picone, R. Bussandri, and M. Amoretti, “A practical network coding approach for peer-to-peer distributed storage,” in IEEE International Symposium on Network Coding (NetCod). IEEE, 2010, pp. 1–6.

 ISBN: 978-1-61208-361-2

[7] A. Bergkvist, D. C. Burnett, C. Jennings, and A. Narayanan, "WebRTC 1.0: Real-time communication between browsers," W3C Working Draft, WD-webrtc-20130910, Sep. 2013. [Online]. Available: http://www.w3.org/TR/webrtc/ [Retrieved: May, 2014]

[8] M. Handley, V. Jacobson, and C. Perkins, "SDP: Session Description Protocol," RFC 4566 (Proposed Standard), Internet Engineering Task Force, Jul. 2006. [Online]. Available: http://www.ietf.org/rfc/rfc4566.txt [Retrieved: May, 2014]

[9] M. Bu and E. Zhang, "PeerJS - Simple peer-to-peer with WebRTC." [Online]. Available: http://peerjs.com [Retrieved: May, 2014]

[10] F. Rhinow, P. P. Veloso, C. Puyelo, S. Barrett, and E. O. Nuallain, "P2p live video streaming in webrtc."

[11] J. K. Nurminen, A. J. Meyn, E. Jalonen, Y. Raivio, and R. G. Marrero, "P2P media streaming with HTML5 and WebRTC," in IEEE International Conference on Computer Communications. IEEE, 2013.

[12] A. J. Meyn, "Browser to Browser Media Streaming with HTML5," Master's thesis, Technical University of Denmark, Lyngby, Denmark, 2012.

[13] L. Zhang, F. Zhou, A. Mislove, and R. Sundaram, "Maygh: Building a cdn from client web browsers," in Proceedings of the 8th ACM European Conference on Computer Systems. ACM, 2013, pp. 281–294.

[14] S. Traverso, K. Huguenin, I. Trestian, V. Erramilli, N. Laoutaris et al., "Tailgate: handling long-tail content with a little help from friends," in Proceedings of the 21st international conference on World Wide Web. ACM, 2012, pp. 151–160.

[15] J. Wu, Z. Lu, B. Liu, and S. Zhang, "PeerCDN: A novel p2p network assisted streaming content delivery network scheme," in Computer and Information Technology, 2008. CIT 2008. 8th IEEE International Conference on. IEEE, 2008, pp. 601–606.

[16] A. van Kesteren, "XMLHttpRequest Level 2," W3C Working Draft, WD-XMLHttpRequest-20120117, Mar. 2012, retrieved: May, 2014. [Online]. Available: http://www.w3.org/TR/XMLHttpRequest2/

[17] I. Fette and A. Melnikov, "The WebSocket Protocol," RFC 6455 (Proposed Standard), Internet Engineering Task Force, Dec. 2011. [Online]. Available: http://www.ietf.org/rfc/rfc6455.txt [Retrieved: May, 2014]

[18] D. Herman and K. Russell, "Typed array specification," Khronos.org, 07 2011. [Online]. Available: https://www.khronos.org/registry/typedarray/specs/latest/ [Retrieved: May, 2014]

[19] M. Leech, M. Ganis, Y. Lee, R. Kuris, D. Koblas, and L. Jones, "SOCKS Protocol Version 5," RFC 1928 (Proposed Standard), Internet Engineering Task Force, Mar. 1996. [Online]. Available: http://www.ietf.org/rfc/rfc1928.txt [Retrieved: May, 2014]

[20] "Netty project." [Online]. Available: http://netty.io/ [Retrieved: May, 2014]

[21] I. Hickson, "The Websocket API," W3C Candidate Recommendation, CR-websockets-20120920, Sep. 2012. [Online]. Available: http://www.w3.org/TR/websockets/ [Retrieved: May, 2014]

[22] W. R. Stevens, UNIX network programming. Addison-Wesley Professional, 2004, vol. 1.

[23] S. Markey, "Manage mobile cloud socket connections," 01 2013. [Online]. Available: http://www.ibm.com/developerworks/cloud/library/cl-mobilesockconnect [Retrieved: May, 2014]

[24] E. Kitamura, "binarize.js – binarize arbitrary js object into ArrayBuffer." [Online]. Available: https://github.com/agektmr/binarize.js [Retrieved: May, 2014]

[25] I. S. Reed and G. Solomon, "Polynomial codes over certain finite fields," Journal of the Society for Industrial & Applied Mathematics, vol. 8, no. 2, 1960, pp. 300–304.

[26] J. S. Plank, "A tutorial on reed-solomon coding for fault-tolerance in raid-like systems," Softw., Pract. Exper., vol. 27, no. 9, 1997, pp. 995–1012.

[27] J. S. Plank and Y. Ding, "Note: Correction to the 1997 tutorial on reed–solomon coding," Software: Practice and Experience, vol. 35, no. 2, 2005, pp. 189–194.

[28] "SpiderMonkey." [Online]. Available: https://developer.mozilla.org/en-US/docs/Mozilla/Projects/SpiderMonkey [Retrieved: May, 2014]

[29] "V8 JavaScript Engine." [Online]. Available: http://code.google.com/p/v8/ [Retrieved: May, 2014]

[30] I. Hickson, "Web Workers," W3C Candidate Recommendation, CR-workers-20120501, May 2012. [Online]. Available: http://www.w3.org/TR/workers/ [Retrieved: May, 2014]

[31] "JavaScript sha1 function." [Online]. Available: http://phpjs.org/functions/sha1/ [Retrieved: May, 2014]

[32] C. Veness, "JavaScript Implementation of AES Advanced Encryption Standard in Counter Mode." [Online]. Available: http://www.movable-type.co.uk/scripts/aes.html [Retrieved: May, 2014]

[33] R. L. Rivest, A. Shamir, and L. Adleman, "A method for obtaining digital signatures and public-key cryptosystems," Communications of the ACM, vol. 21, no. 2, 1978, pp. 120–126.

 ISBN: 978-1-61208-361-2

A Comparative Study of Replication Schemes for Structured P2P Networks

Moufida Rahmani, Mahfoud Benchaïba
University of Science and Technology Houari Boumediene
LSI, Computer-science Department
Algiers, Algeria
Emails: {morahmani, mbenchaiba@}usthb.dz

***Abstract*—Structured Peer to Peer (P2P) networks provide efficient mechanisms for resource placement and lookup. However, these systems deal with irregular and frequent arrival/departure of nodes. Thus, these systems do not offer any guarantees about data availability. A well-known technique for improving resources availability and providing load balancing is replication. Many replication methods are proposed for structured P2P networks, with a specific main goal to achieve. This paper reviews and compares various existing replication techniques for structured P2P networks, which we classify based on their main objectives.**

Keywords–Structured Peer-to-Peer; Replication; DHT; Availability; Load balancing; Performance; Churn.

I. INTRODUCTION

Since 1999, P2P networks have been in continuous development. P2P networks are overlay distributed networks composed of a large number of autonomous nodes, i.e., virtual networks which may be totally unrelated to the physical network that connects the different nodes. These nodes, called peers, share a part of their own resources such as storage capacity, files and processing power. They play the role of both client and server. Communications between peers are direct, without passing intermediary entities.

Napster [1], a popular music exchange system, was the first to emerge as a P2P file sharing application. After that, several file-sharing software have succeeded, we quote: Gnutella [2], KaZaA [3], BitTorrent [4], Oceanstore [5], and PAST [6].

P2P networks can be classified as unstructured and structured, depending on the overlay structures. Unstructured P2P systems do not impose any structure on the overlay network and usually use flooding for searching objects. This method is expensive in terms of bandwidth consumption and is not efficient for locating unpopular files. Structured P2P systems, in the other hand, impose particular structures on the overlay networks (which are commonly referred to as Distributed Hash Tables (DHTs)). Any file can be located in a small number of overlay hops, which significantly reduces the search cost as compared to unstructured systems. Unfortunately, if connection/disconnection frequency is too high, data may be lost. To deal with these problems, the replication can be used as the efficient technique to improve resources availability and to provide load balancing enhance.

Many replication methods are proposed for structured P2P networks, with a specific main goal to achieve. Ktari et al. [7] have presented a comparative analysis of some replication algorithms for DHT architectures. Our paper's prime objectives are to:

- Highlight factors involved in replication, type of replication, and parameters affecting replication.
- Present the some existing replication techniques for structured P2P networks with a new classification. Each replication technique can be implemented for several objectives such as: improving availability, enhancing system performance, achieving load balancing. In this paper, we try to classify replication strategies existing in the literature based on their main objectives. Our classification lets to well study and compare them.

This paper is organized as follows: In Section II, structured P2P networks and examples of networks are presented. In Section III, we highlight some replication basic notions as well as factors and parameters involved in it. Section IV reviews, classifies and compares existing replication strategies. Finally, we conclude in Section V.

II. BACKGROUNDS: STRUCTURED P2P OVERLAY NETWORKS

In structured P2P overlay networks, the topology is tightly controlled and the data is placed at specific location which makes queries more efficient [8]. Structured P2P systems use DHT as a substrate, in which the location information of object (value) is placed deterministically, at the peers with identifiers corresponding to the data objects unique key. In DHT, a peer's identifier *ID* is chosen by hashing its IP address and objects's *key* is chosen by hashing its name for example. Both peers and objects are identified in the same namespace. Each node is responsible for some of the *keys* in the system and each data object is stored on this node if the identifier of the object belongs to the range which node is responsible. The main operations used in DHT are: *put(key, value)* and *lookup(key)*.

- *put(key, value)*: This operation is used when the peer wants to publish an object in the system. Peer computes the *key* of the object and then sends a message put(key, value) to the peer responsible for this *key*.
- *lookup(key)*: This operation returns the value associated with the *key*, if any.

Several systems employing DHTs have been developed; among the most well-known Pastry [9], Tapestry [10], CAN [11], Kademlia [12] and Chord [13]. We present Chord as an example of this category. Additionally, a drawing is added to clarify the functioning of chord.

 ISBN: 978-1-61208-361-2

- ***Chord:*** Chord is based on a ring topology, a Chord peer has knowledge of its predecessor and its successor. A hash function (SHA-1) generates a regular identifier, an m-bit for each peer from its IP address. Then, each peer is placed in the ring so as to arrange the identifiers in ascending order. The successor (respectively predecessor) of a peer n is the peer whose identifier is immediately higher (respectively lower) to identifier of peer n. Thus, each peer n with identifier ID is responsible for the interval of *keys*] predecessor (n), n]. For a given peer, mere knowledge of its predecessor and its successor is not sufficient to ensure good performance of the ring, particularly in terms of number of hops per request. To overcome this problem, for a *key* space in the range [0, 2m [, each peer ID connects to other neighboring nodes, called fingers, with ID successor(ID+ 2^i-1) with $1 \leq i \leq m$. These fingers constitute its routing table. Thus, the number of fingers per node is *O*(log N). Thus, the maximum number of peers traveled to forward a query is expressed in terms *O*(log (N)), where N is the number of peers in the system.

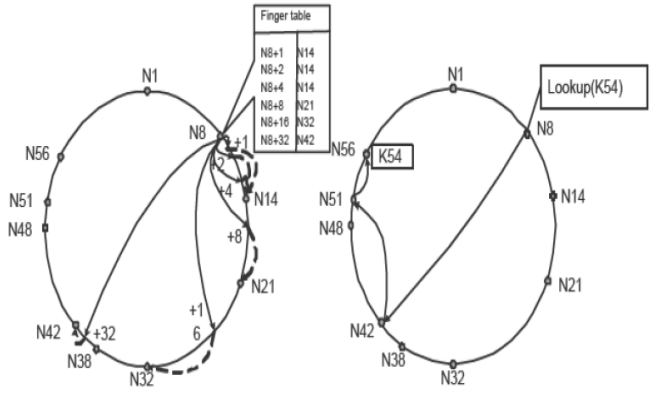


Figure 1. Chord ring with identifier circle consisting of ten peers and five data keys. It shows the path followed by a query originated from peer 8 for the lookup of key 54 [13].

When a node wants to find data object (value) for a key *lookup(key)*, it uses the following algorithm: it seeks among its fingers the peer whose identifier is the greatest and is lower than the key, and sends it the message. The node that receives the message, then in turn executes this same algorithm (see Figure 1) until the target node is found. Nodes are allowed to join and leave the system causing churn. A Chord regularly runs maintenance algorithms that detects failures and repairs routing tables, allowing requests for a *key* to be routed correctly to their owner despite node churn.

III. REPLICATION

The main idea of replication data is to maintain several copies or replicas of the same data at various different sites. Data replication is recognized as an effective way to increase the availability and performance of distributed systems such as P2P. Replication is a solution that offers several advantages:

- **Increases availability:** Replication removes single points of failure (data is accessible from multiple nodes), thus increasing availability and fault tolerance.
- **Improves performance:** Replication improves performance of system in terms of response time. Data can be located closer to their access points. Therefore, the success rate will increase, the response time and the overhead will decrease. The response time can be constant for all the users if all the data are replicated uniformly over the network.
- **Achieves load balancing:** Replication can provide load balancing between the nodes, such as multiple nodes can serve the same object simultaneously. Therefore, it reduces load on the nodes that own the original data.

As replication has advantages, it also has significant costs such as the storage cost and consumption of bandwidth. To control the cost of replication, there are important factors that replication must take into consideration. These factors have a direct impact on the performance of the system. In order to avoid wastage of network and peer resources, the excessive use of replication is not recommended.

A. Factors of replication

When designing a replication strategy, the following important aspects should be taken into consideration:

1) ***What should be replicated****:* Most strategies of replication choose the files to replicate based on their popularity values. A general way of measuring the file popularity on a peer is by counting the number of requests. Other strategies prefer, for their part, to replicate all shared files or only replicate the rare objects.

After choosing the file to replicate, there are important point to decide: is it necessary to replicate the file or its index (since the index size is smaller that the file's)? Or is it preferable to adopt erasure code to replicate file? This point is related to the type of replication, which will be detailed in Section "III.B".

2) ***Where should replica be placed****:* The second important factor is selection of the best node to host a particular replica. Replicated copies should be placed in proximity to peers who are likely to request the resource in order to reduce delays of search and downloading. Moreover, some peers' characteristics, such as available storage space and availability, should be taken into consideration. Each replication strategy which aims to improve file availability must consider the peer availability: If a new replica is hosted by a peer that has low availability and may leave soon, we may need another replica in order to maintain the required availability for the file.

3) ***When should be replicate****:* Replication can occur periodically or at event. For example, suppose that the peer considers file's popularity. When it receives a request for a file, the peer increases file's popularity value and checks whether it exceeds some threshold. If so, it might decide to replicate the file. Some strategies ignore this factor.

 ISBN: 978-1-61208-361-2

4) ***How much is the number of replicas per file****:* Each replication strategy determines the number of replicas per file according to its objectives and the system parameters which it takes into account. For example, if the aim of the strategy is to maintain a threshold level of availability, it needs to consider the system's parameters that affect availability and performance such as online availability of the peers in the system.

5) ***Replica replacement strategy****:* It is essential to deploy replica replacement strategy because storage space is limited. A replacement strategy consists on removing some less efficient replicas to create space for new replicas. Least Recently Used (LRU) is the most widely used in P2P networks.

B. Type of replication

In the context of P2P networks, the replication is redundancy by creating copies of data, called replicas, which can be stored in peers other than the source peer (which holds original data). Replica can be a complete file, the index of the file or bloc of the file.

1) ***Traditional replication****:* is called also data replication. In this type, replica is a complete (entire) file.

2) ***Erasure code replication****:* An erasure code provides redundancy as replication for achieving high availability and reliability in storage and communication systems without imposing high bandwidth and storage overhead [14]. In erasure code, a file is divided into b (equal size) blocks and recoded into c blocks, where $c > b$ (of same size as before). The erasure-coded blocks are dependent each other. The key property of erasure code replication is that any b out of c blocks is enough to reassemble the original file. We call c/b the storage overhead S which is sometimes stated as the stretch factor.

Lin et al. [15] provided a comparison between traditional replication and erasure code replication. They came to an important result: in erasure code replication with a storage overhead of S, if a file is divided into b blocks, then each file block is replicated S times. Therefore we have $S*b$ number of blocks in the system. They also pointed out that when $b=1$, erasure code replication is equivalent to traditional replication.

3) ***Index replication****:* In some cases, it is best to replicate the index of a file that is the pointer to the peer that holds the file. This solution is recommended when the file size is very large to avoid the problem of the data file coherence when the original file is updated. The index replication consumes little storage space and bandwidth.

4) ***Message replication****:* The idea of the message replication, introduced by Hassan and Ramaswamy [16], is to replicate a message several times within the network to enable a large query coverage in the network.

When we design a replication strategy for unstructured P2P, we can use one of the four replication types. We can also use two types in the same strategy for example traditional and index replication as in [17]. In case of structured P2P network, we can find all types of replication except message replication. Additional, in few cases, routing tables or information about neighbors are also replicated.

C. The parameters that affect replication

The parameters that affect the efficiency of the replication, and must be taken into consideration are mainly: the file popularity, the peer availability, rare objects, storage space and data consistency.

1) ***The file popularity****:* Jacky et al. [18] present a study of popularity measurements of P2P file systems in Gnutella and Napster. They came to an important result: caching or replicating the most popular files on the system is strongly suggested in order to greatly improve system performance. Some strategies presented in this article are based on popularity in order to determine the candidate files for replication. The file popularity can be calculated by keeping track of the number of requests. This value is local and can change rapidly and therefore increase the replication cost. To avoid this, the system should calculate and predict the overall popularity values (for all the network). Manel and Mahfoud [19] define a way to calculate a global file popularity based on local estimation of the peer and estimations done by the other peers participating in the network. The simulation results show that their measurement is closer to the real one.

2) ***The peer availability****:* In P2P networks, peers are volatile: they join and leave the network unexpectedly. The main consequence is that the files (original copy or replica) that a peer stores might become unavailable. Accordingly, the peer availability affects the file availability. Thus, the replication strategy which aims to improve availability of resources must take into account this parameter to calculate the number of the file replicas. In [20], a study was made to understand the peer availability.

3) ***Rare objects****:* Studies have shown for Gnutella that 18% of all queries return no responses even when results are available [21]. That is due to search algorithms used in unstructured P2P, the most typical query method is flooding. This method is effective for locating highly replicated data and is not suited for locating rare data (those with few replicas). Two solutions are proposed in the literature to solve this problem: hybrid search and replication. Hybrid search combines two search methods, it uses flooding techniques for locating popular items and structured (DHT) search techniques for publishing and locating rare items. Replication is a well-known technique for improving resource availability. To the best of our knowledge, there are few works that propose a replication technique in order to improve search for rare objects ([22][23][24]). A key challenge for the hybrid search and replication is how to identify rare items.

In our opinion, this last point is not satisfactorily addressed in the case of replication, because some strategy as presented by Ma et al. [24] are based only on the sampling technologies and number of copies to determine if object is rare or not. Therefore, before proposing a replication technique aiming to improve search for rare objects, one must first define what is a rare object, how one can identify it and at the end one discuss the factors involved in replication. This parameter is only for unstructured P2P network because in the case of structured P2P network, any file can be located even for rare data if any.

4) ***Storage space****:* Each peer in a P2P network shares a part of its storage capacity which will be used to store its shared files and the replicas of files of others peers. Unfortunately, this

 ISBN: 978-1-61208-361-2

storage space is limited and can store only a limited number of replicas. Therefore, replication strategy must consider this parameter when it chooses the best node to host a replica and it must deploy replica replacement strategy. However, most replication strategies do not consider storage capacity constraints thus the network and peer resources are abused.

5) ***Data consistency****:* Generally, we can not use data replication without evoking data consistency issue. It turns that in P2P systems (such as PAST, Gnutella) [25], the replicated resources often they are not subject to modification (static or read-only), this explains in part why replication techniques do not consider the consistency aspect. However, in case of resources update, it is recommended to trait the two aspects simultaneously. This can decrease the overhead of consistency maintenance and ensure that a file requester receive up-to-date files.

IV. REPLICATION TECHNIQUES FOR STRUCTURED P2P NETWORKS

A. Replication in DHTs

The replication strategies, for the structured P2P networks that employ DHTs, use two algorithms: Data replication algorithm (or called Replica placement algorithm) and maintenance algorithm. The choice of these algorithms can have significant impact upon performance and reliability.

1) Data replication algorithm: In this kind of algorithm the peers decide what should be replicated, how many replicas should be created and where to replicate them in order to realize a well-determined objective such load balancing or improve availability of resources. There are three main basic replica placement strategies, which are the basis of the most replication strategies present in the section.

- Neighbor replication: Called also the simple replication method, each peer maintains a list of neighbors such as successor-lists and predecessor-lists in Chord [13] or leaf-sets in pastry [9]. In neighbor replication, the data objects are stored not only in root peer but also on its successor, or on its predecessor, or on its leaf-sets and or on the nodes belonging to the same group as it. The root is node that stores the object location information and it can be different to the owner which is the node that stores the master copy of the object. Chord employs successors-lists replication. Pastry and Kademlia DHTs employ leaf-sets replication.

- Path replication: Path replication replicates a data object along the search path that is traversed by the lookup message, from the requester to the provider node (root peer). Tapestry [10] employs path replication scheme.

- Multi Publication Key Replication : A key is mapped into *r* points in the coordinate space and accordingly replicated at *r* distinct nodes in the system. CAN implements this solution by using Multiple hash functions.

2) Maintenance protocols: For each data replication algorithm, there is a special maintenance protocol. The idea is that the maintenance protocols must maintain k copies of each data objects without violating the initial placement strategy. It means that the k copies of each data object have to be stored on the root-peer neighbors in the case of the neighbors replication scheme, on the root peers in the Multi Publication Key Replication scheme and on all the peers that exist in the search path in the case of path replication.

B. Classification of replication techniques for structured P2P networks

A replication technique can be performed for several objectives such as: improving availability, enhancing system's performance, achieving load balancing. To fulfil these objectives, a number of replication strategies have been proposed for structured P2P networks. Each strategy is proposed for a specific DHT, and employs different algorithms for placement and maintenance. In this section, various existing replication schemes are presented and classified into four categories according to replication objectives as described above. Each category defines a main replication objective to achieve (Figure 2). However, some techniques, which belong to one category, may possess secondary properties of another category.

Category1: *"Achieve load balancing"* Godfrey et al. [26] say that the load unbalancing problem in structured P2P network may result due to non-uniform distribution of objects in the identifiers space. Therefore, some nodes having $O(logN)$ times as many objects as the average node, where *N* is the number of peers in system. Additional, if a single node stores a popular file, then all requests for this file are directed to this node and this will make it and the path leading to it overloaded. Category 1 includes proposed strategies that solve load balancing issue by efficient data replication.

Category2: *"Increase availability of resources"* This category shows the replication strategies that seek to increase the availability of resources caused by irregular departure of peers.

Category3: *"Enhance churn tolerance"* When new peer joins the system, if its identifier is closer to an object's key than the identifier of its current root, the data object needs to be migrated on the new peer and the new peer will become the root for this object. The migration process can be also appeared in the case when the peer quits the network or when neighbors list is changed. If a high churn rate arrived, then maintenance algorithms must often be applied in order to adapt to the new structure by migrating data objects, which generates more traffic and consumes much bandwidth. In order to avoid this issue, the replication strategy must be more tolerate under higher churn rates.

Category4: *"Improve search performance"* It includes replication strategies that decrease the response time and the overhead, thus the search performance is improved.

1) Achieve load balancing:

a) Lightweight Adaptive system-neutral Replication protocol LAR [27]: LAR can efficiently deliver a good load balance and low query latencies even when demand is heavily skewed. LAR functions as follows:

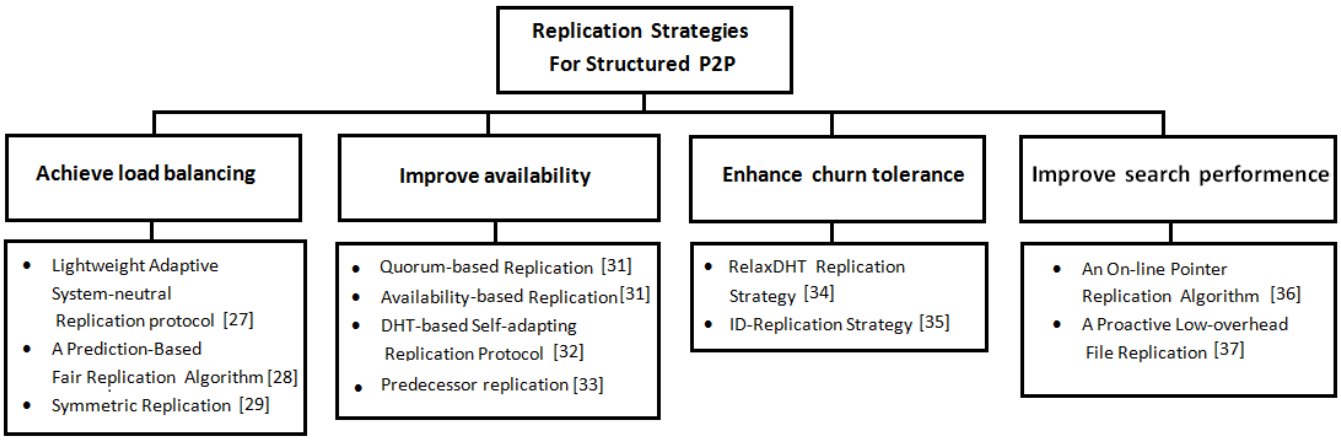


Figure 2. Classification of replication techniques for structured P2P networks.

Each time the peer *Pi* receives a request for data item from the peer *Pj*. It checks if its current load *Li* exceeds a certain threshold. Load is redistributed according to a per-node capacity *Lmax*, high-load *Lhi* and low-load *Llo* thresholds. The node capacity is the number of queries that it can route or handle per second, there are two different cases:

- if ($Li > Lhi$), it indicates that load redistribution is necessary. In this case, *Pi* attempts to create new replicas on Pj, if *Li* is greater than *Lj* by some fixed value K. Pi then asks *Pj* to create replicas of the *n* most highly loaded items in *Pj*, such that the sum of the local loads due to these *n* items is greater than or equal to *K*.
- if ($Llo < Li < Lhi$) then load is redistributed as above on *Pj* only if $(Li - Lj \geq Llo)$.

If *Pj* has no space storage for hosting the new replica, it then uses LRU algorithm, to choose a victim replica to leave the space for the new replica.

The locations of the new replicas is then spread, by piggy-backing them, on subsequent messages that contain requests for the same items. The pointers to the replicas are cached in the peers along the path traversed by requests for these replicas. In the same way as for replicas, LRU policy is used: the least recently used pointer is deleted.

The presence of pointers of the new replicas can then respond more quickly to the subsequent requests and potentially reduce the number of hops needed for routing. Indeed, if a peer reaches a node with a pointer to a replica rather than the node with the resource originally, it follows the pointer.

When a node that has a replica leaves the network, or simply when LRU algorithm is applied to replace cache or replicas, LAR does not apply maintenance protocol. Therefore, the initial placement strategy is violated and it can find pointers to inexistent content.

When adapting LAR to Chord, the finger list is the default item of replication. LAR replicates the data item only if the load on the server due to the data item is more than that due to the finger list.

b) A Prediction-Based Fair Replication Algorithm [28]: It aims to maintain an excellent system performance when the query is highly skewed. Through the use of a simple prediction method, it can foresee traffic surge and replicate beforehand. The basic idea of the PFR algorithm is to create replicas for the node whose predicted load or current load reaches certain predefined threshold and adaptively adjusts the Replication Speed (RS) for each replication process. RS can be measured by the ratio of the number of nodes chosen to hold replicas to the number of all nodes that have encountered along the query path. The light-loaded nodes are always chosen to hold replicas. For example, if RS equals to 3N/4, this means that load should be redistributed to 3/4 amount of the total number of nodes along the query path, where *N* is number of node in the query path. PFR function as follows:

When query packet is routed through a node, it computes and piggybacks its predicted load on the query packet. At the same time, it checks the value of the predicted load *Preload* and current load to determine whether the load rebalancing is necessary or not. With respect to nodes' predicted load fraction, PFR defines 5 different replication levels, which specify the RS for each query. There are two types of nodes along a query path: query's destination node and the nodes just forwarding queries. For the first type of nodes, node replicates according to these 5 levels. However, for the second type of nodes, replication is necessary only when its *Preload* has reached the level one of threshold. The first level threshold indicates that a node is approaching its capacity and it is in an emergency state to shed load in order to prevent itself from overloading and consequent dropping queries. The node capacity is the number of queries that it can route or handle per second. If the replication is necessary based on Preload, the node creates replicas of the *n* heaviest-loaded items on each selected node, such that the sum of the local loads caused by these n items will be greater than or equal to the difference in loads between the two nodes. Else, the node checks whether the replication is necessary or not based on the node's current load fraction. If yes, the node replicates in the same way, but the corresponding replication level should be decreased by 1. It means that value of RS when the replication is based on current load is lower compared to the value of RS when the replication is based on Preload. Whenever stored replicas reach

 ISBN: 978-1-61208-361-2

the nodes maximum storage size, the new replicas will replace the old replicas using the LRU algorithm.

PFR algorithm also uses the replica location dissemination method such as LAR. It helps replicas to be efficiently utilized in shedding load. But, it does not apply maintenance protocol. Therefore, we can find pointers to inexistent content.

When PFR is applied to Chord, the finger list is the default item of replication. PFR replicates the data item only if the load on the node caused by the actual data item is more than that caused by the finger list.

c) Symmetric Replication [29]: Symmetric replication can be used for load-balancing between the peers, end-to-end fault tolerance and to increase the security. The advantage of symmetric replication is that it can be applied to all structured P2P systems. It is closer to the methods that use several hashing functions for replication. The main idea behind symmetric replication is that each identifer in the system is associated with a set of f distinct identifers such that the following always holds: if the identifer i is associated with the set of identifers *r1, ..., rf*, then the identifer *rx*, for $1 \leq x \leq f$, is associated with the identifers *r1, ..., rf* as well. The identifier space is partitioned into *N/f* equivalence classes such that identifiers in an equivalence class are all associated with each other, where N is the size of the identifier space and f is replication degree. The replication degree is number of replicas made. The identifier i is associated to the f different identifiers given by the function H; $H(i,x)=i+(x - 1) N/f$.

In symmetric replication there is no root node, a data item with identifier i is stored on the f peers given by $Sp(H(x, i))$, for all x ($1 \leq x \leq f$). Sp is pseudo-metric space as hash function (SHA-1) in Chord. Thus, the responsible peer of identifer i stores every data item with an identifer associated with i. This implies that to find a data item with identifier i, a request can be made for any of the identifiers associated with i. Each node storing a data can easily calculate the keys of the different replicas of this particular data item, it is sufficient that each node knows about the replication scheme. Therefore, it can achieve load-balancing between replicas by sending requests to a random replica.

Symmetric replication has put a set of replication algorithms, these algorithms are used for joins and leaves, inserting and looking up items and failures. It can enhance performance by sending multiple concurrent requests and picking the first response that arrives. Unlike the other replication methods that use multiple hash functions, successor-lists, or leaf-sets to select replica nodes, symmetric replication needs only $O(1)$ messages for every join and leave operation to maintain any replication degree. The replica node is a node that stores a replica.

Every replica node cooperates to execute maintenance algorithm. The node storing the replica i checks if the replica $i+1$ is stored in the corresponding node. If not, it inserts a new replica in that node. The node storing the replica $i+1$ do the same with the replica $i+2$ an so on.

Discussion: The replication strategies presented in this category use different replica placement algorithm to achieve load balancing. The first technique, LAR, uses any type of replica placement algorithms described in Section IV.A.1. It resembles at owner replication, because only the requester node keeps the copy, the others nodes on the query path contain only the cache entries, pointers towards replicas. In the owner replication [30], if a search for an object is successful, this object is replicated on the requester node. The replicas can be efficiently utilized due to the use of the replica location dissemination method.

PFR is an improvement of LAR, because it uses the same principle. The both use the load value to determine whether load redistribution is necessary. After the replication, both use the dissemination method to spread the information about new replicas locations. In the case of LAR, RS value is always equal 1. Moreover, there are some difference between the two: First, in PFR not only the requester node keeps the copy, but also some node in the query path according to the RS value. Thus, PFR is a variety of replication path. Second, all the nodes in the query path can check the load value to determine whether load rebalancing is necessary or not. Unlike LAR, only the peer that receives a request checks. Third, LAR uses only the current load to check, but PFR uses two load values predict and current. Therefore, PFR adaptively adjusts the number of replicas created for each query process and can be scattered before flash crowd happens. Symmetric replication based on multi publication key replication is completely different compared to LAR and PFR, in different points:

First, symmetric replication replicates only data item, but in PFR and LRA, data item is not the default item of replication. Second, in symmetric replication, the replication degree is the same for all the replicas, but in PFR and LRA the replication degree is determined according to the load value. Third, only the symmetric replication applies maintenance algorithm. The maintenance protocol is applied in distributed manner and it needs cooperation of all replica nodes. Therefore, it is complex compared to the other techniques where the maintenance is provided by the root node. Fourth, LAR and PFR employ the replica location dissemination method in order to the replicas can be used. In symmetric replication, it is sufficient that each node knows about the replication scheme. Therefore, it can easily calculate the keys of the different replicas.

2) Increase availability:

a) Quorum-based replication and Availability-based replication [31]: Kim and Park have proposed efficient replication methods that can reduce network traffic enormously and achieve high data availability in DHT based on P2P storage system. In these methods, the replicas are loosely coupled to the consistent set such as the leaf-set and the successor-lists. Additionally, they are interleaved on the consistent set to reduce the compulsory copies which occur under churn, unlike the simple replication (neighbor replication) method that directly uses the consistent set. This set is tightly coupled to the current state of nodes and the traffic needed to support this replication can be high and bursty under churns.

Tow types of replication methods are proposed: *Quorum-based replication* and *Availability-based replication*.

The *Quorum-based replication* modifies the simple replication to prevent the compulsory copies which occur under churn. When a new node joins, it gets not only routing information such as DHT and consistent set, but also the replication set. The replication set indicates which node replicates the

 ISBN: 978-1-61208-361-2

object among the consistent set. The size of the consistent set is bigger than the number of replicas, the replication does not occur frequently and the replication set can interleave the replicas on the consistent set. This behavior increases the chance to reduce the compulsory copies.

The replication only occurs when the number of replicas is fewer than the target quorum. The quorum is the fixed minimum number replicas necessary to archive an objective. In this case, the number of replicas must be more than the target quorum to achieve target data availability. When a node leaves and it is not a member of the replication set of peers, there is no need to replicate the data. Otherwise, if it is a member, the selection of node among non-members of the replication set of peers as a new replica (target node) is necessary. The target node became among the replication set for this peer. The *Quorum-based replication* considers that each node has the same availability. However, if a new replica is assigned by the node which has low availability, this node may leave soon and we need another new replica. Therefore, *Availability-based replication* takes into consideration the node availability and guarantees the high data availability by selecting the more available nodes as replicas. To do this, each node manages its availability and advertises it to all members of the consistent set by piggybacking it to the periodic ping message which has been already used to detect node failures on the consistent set. In this approach, the replication only occurs when the data availability is below the target availability. When a node needs a new replica, the most available node among non-members of the replication set is selected as a new replica.

If all members of a consistent set have averagely low availability, *Availability-based replication* needs more replicas than the quorum based replication. Sometimes this behavior takes more bandwidth, but when nodes leave, this subtle replication can reduce much more bandwidth than the *Quorum-based replication*.

When a node fails and its neighbor gets the lookup request, this neighbor may not have the replicas for the requested object. In this case, the neighbors must forward the request to the replicas of the failed node for the routing correctness. To do this, each node should have the replication sets of all members of its consistent set by piggybacking this information to the periodic ping message for its consistent set.

The both *Quorum-based replication* and *Availability-based replication* used the same maintenance algorithm : When a new node joins and a target node gets this join request, the object ranges of the target node is divided into two object range and the new node is responsible for one of them. In this case, the new node simply copies the replication set and adds the target node as a new replica because it already has the object for this range. When a node leaves or fails, its neighbor node is responsible for its object range. In this case, both replication sets of the failed node and its neighbor node are merged.

b) DHT-based Self-adapting Replication Protocol [32]: Knezevic et al. have presented a fully decentralized replication protocol suited for any DHT networks, that adjusts autonomously the number of replicas to deliver a configured data availability guarantee.

The main idea of this replication technique is to associate for a given object many keys. The node that publishes the object simply calculates different keys and then inserts the object in the corresponding nodes. To calculate these keys, it uses correlated hashing. The first replica key is generated using a random number generator. All other replica keys are correlated with the first one, i.e., they are derived from it by using the following rule: *replicaKey(1)=c*, *replicaKey(ron)= H(replicaKey(1) + ron)* if $ron \geq 2$, where *ron* is a replica ordinary number, *c* is a random byte array, *H* is a hash function with a low collision probability. *replicaKey* is observed as byte arrays, and + is an array concatenation function. To calculate the key of the ron^{th} replica, the peer requires access to the first replica key and the replica ordinary number (*ron*). All this information is wrapped in an instance of Entry class. Additional, all the nodes use same H.

Every peer calculates the number of replicas *R* from measured average peer online probability and the requested data availability. During joining phase, a peer can get an initial value for *R* from others peers, or can get assume an initial value of *R* for it.

To meet the requested data availability in a DHT, the number of replicas of stored data at each peer has to be adjusted. Therefore, every peer measures the current average peer online probability, knowing the requested data availability and it calculates the new value for the number of replicas *R*. By knowing the previous value R1, a peer removes the replicas with ordinary number *ron* greater than *R* from its local storage replicas. If $(R < R1)$ higher number of replicas are needed, a peer creates new replicas of the data in its local storage under the keys *replicaKey(j)*, *j = R1 + 1, ..., R*.

c) Predecessor replication [33]: Predecessor replication is a simple and an efficient data replication approach. It ensures high data availability it can decrease the number of hops needed to locate the requested data. The node replicated each key whose it is responsible on its predecessor nodes in the same number of copies. According to the query routing mechanism used Chord, the lookup query can be routed to replica node (predecessor node) before reaching the root node. Therefore, the search path is minimized.

Two update strategies are used for the maintenance under churn: the basic update and the periodic update. In the basic update, when the node leaves the network, the data whose it is responsible will be migrate to its predecessor. The periodic update is periodically used to maintain the replication degree. Each node contacts all its replica nodes to ensure that they correctly maintain the appropriate replicas. Each replica node contacts the root nodes of all replicated keys which stores in order to keep the replicas up-to-date.

Discussion: Although the methods presented above are proposed to achieve high data availability in a DHT based P2P systems, there's a difference between them, we quote: The first and the third methods are based on neighbor replication and the second method is based on multi publication key replication.

In order that nodes can access to the replicas in case of the root node is failure, the information of the replication sets are spread in the first methods, by piggybacking this information to the periodic ping message. In case of the second method, it is not needed to know the information about the replica locations. When a peer wants to get a value, it is sufficient to return any available replica. The basic DHT lookup operation

 ISBN: 978-1-61208-361-2

is applied until the data is retrieved. Like the second method, the third method does not need to know the information about the replica locations, the nodes can access to the replica before reaching the root node.

In *Quorum-based replication* and predecessor replication, the replication degree is fixed constant and the is same for each data. They do not take into account some characteristics such as the requested data availability and the node availability. Therefore, the storage space can be abused by the data which is not popular. Additionally, the bandwidth consumption is increased by the migrating data and replication process to maintain the replication degree, because the replicas can be stored by the nodes which have low availability and can leave the network soon.

Availability-based replication and DHT-based Self-adapting Replication Protocol adjusts the number of replicas according to the node availability and the requested data availability. Thus, if the number of replicas is not sufficient to ensure the requested data availability, new replicas will be created. Additional in the second method, if there are more replicas than needed, peers will remove some of them. Therefore, the second method generates less storage costs.

3) Enhance churn tolerance:

a) RelaxDHT replication strategy [34]: Legtchenko et al. have proposed RelaxDHT replication strategy that enhances churn tolerance by building an efficient Replica placement and maintenance mechanisms. The RelaxDHT strategy is based on neighbor replication, exactly as the leaf-sets replication applied in DHT pastry.

When the root peer receives *put message* for new data block. In the case of the leaf-sets replication with the replication degree equals to R, the root peer stores a copy of the data block for which it is the root. After, it sends the R-1 copies of this data block for its replica-sets. The replicas sets of a root peer are a subset of its leaf-sets.

In case of RelaxDHT replication strategy, the root peer does not necessarily store a copy of the data blocks for which it is the root. It maintains metadata describing the localization of replica sets, the goal of using localization metadata allows to be anywhere in the leaf-sets. Therefore, when a new peer joins a leaf-sets, the application maintenance protocol is not necessary. The replica sets are selected randomly among the R peers around the center of the leaf-sets. This choice will reduce the probability that a chosen peer quickly quits the leaf-sets due to the arrival of new peers.

The root peer sends *Store message* for its replica sets peers that contains in addition to the data block itself such as the Leaf-sets replication, the identity of the peers in the replica set and the identity of the root. A peer may be root for several data blocks and a part of the replica set of other data blocks. In the first case, the peer must store a list of data block identifiers with their associated replica-set-peer list for blocks for which it is the root. In the second case, peer stores a list of data blocks for which it is a part of the replica set, additional it stores the identifier of this data block, the associated replica set peer-list and the identity of the root peer.

Periodically, each peer executes two maintenance protocols. In the first protocol, it checks for each data block that it stores if the root peer for this data block has changed. If so, the peer sends massage for the future root peer. When the new root of this data blocks receives the message, it adds the data block identifier and the corresponding replica set in the list. In the second, it checks for each block for which it is the root, if all the replicas are placed in its leaf-sets around the center. For a data block, if one of its replicas has changed, then the root peer chooses randomly a new peer in the center of the leaf-sets and changes the replica set. After, it sends a *Store massage* with the replicas set for each peer in its replica set. When the peer receives this message and already stores a copy of the corresponding data block, it updates the corresponding replica set if necessary. In other case, if the peer does not store the associated data block because it is a new peer in the replica set, it fetches everything necessary to store (data block) from one of the peers mentioned in the received replicas set.

b) ID-Replication strategy [35]: ID-Replication can be used in any structured overlay network, however Shafaat et al. have adapted it to Chord for the sake of simplicity. ID-Replication strategy is less sensitive to churn compared to successor-list replication. In order to achieve this goal, ID-Replication uses sets of nodes, called groups, instead of individual nodes. Thus, each group like a node in Chord has unique identifier and is responsible for some of the keys.

Within each group, the nodes that compose it possess two identifiers. A global identifier that is the same identifier as the group and a local identifier that is unique for each node. Each group has successor list, predecessor and fingers as node in Chord.

In successor-list replication, with the replication degree equal to *R*, the root peer stores a copy of the data block for which it is the responsible, the *R-1* copies of this data objects are replicated in its successor-list. In ID-Replication, there are not the root peer, and all the nodes within group store a copy of data blocks which the group is the responsible for. Therefore, a request can be routed to a random node in the group thereby load balancing between the replica nodes. Moreover, in successor-list replication, a request is first routed to the root peer. ID-Replication can send out multiple concurrent requests and picking the first response that arrives.

The replication degree is not fixed, it is between two parameters *Rmin* and *Rmax*. Thus, the number of the nodes within group is specified by these parameters. To allow more copies for popular data objects than other data objects, *Rmin* and *Rmax* must have higher values.

Periodically, each node *p* checks if the size of its group is smaller then *Rmin* because a node is failed, then *p* searches for a standby node by gossiping or contraction a directory and tries to include it in this *p*'s group. If a standby node cannot be found, *p* triggers a merge *p*'s group members with others group such that the size of merged group must be less than *Rmax*. If the size of a group is more than *Rmin*, then standby nodes are (size of group - *Rmin*) nodes. If size of a group is larger than *Rmax*, p initiates the split operation by dividing the group into two groups.

Discussion: The two strategies mentioned above are proposed to be less sensitive under the churn compared to the leaf-sets replication and successor-list replication respectively. There are some differences between them, we quote: In the

first, the root peer chooses randomly the replicas sets peers in the center of the leaf-sets then, it is necessary to maintain metadata describing the localization of its replicas set. Additional, each peer maintains information about replica set peer-list which it is part.

Unlike the first, in the second there is no root node and replicas set. Soon as a peer joins a group, it stores a copy of data blocks that the group is responsible for. Like the first, each peer can have information about the others replicas node. In order to realize this, the nodes in the group use gossiping between them.

The ID-Replication gives different replication degree, thus allowing popular data to have more copies, but in RelaxDHT replication strategy it is not mentioned.

The maintenance protocol in RelaxDHT strategy is more complicated then ID-Replication strategy, but the maintenance cost in the both is still moderate compared to the leaf-sets replication and successor-list replication. The maintenance cost is in term of generated overhead and bandwidth consummated by maintenance protocol.

4) Improve search performance:

a) Proactive Low-Overhead File Replication Scheme Plover [36]: Proactive Low-Overhead File Replication Scheme achieves high efficiency in file replication and supports low-cost and consistency maintenance, because it replicates files among physically close nodes based on node available capacities. Plover also includes an efficient file query redirection algorithm for load balancing between replica nodes.

In order to achieve efficient file replication, Plover uses clustering, the physically close nodes are grouped in clusters. Each cluster has a supernode, which is node with high capacity and fast connections. The others nodes are called regular nodes, which are nodes with lower capacity and slower connections.

Periodically, each lightly loaded node reports its information of available capacity. A node's capacity is presented by the number of bits it can transfer in responding file queries per second, and each heavily loaded node reports the information of its popular files to its super node. The lightly loaded node is the node whose actual load is no larger than its capacity, otherwise a it is heavily loaded node. The load caused by file access is determined by the file size and visit rate (popularity), which visit rate or popularity is measured by the number of visits during time unit (second). The super node collects all this information and arranges the file replication between them (among the clusters). The supernode notifies overloaded nodes to replicate popular files to lightly loaded nodes.

Plover addresses the problem of file consistency maintenance and tries to facilitate efficient file consistency maintenance with low-cost. When the node updates a file, it sends update message to its supernode.The super node forwards the message to the replica nodes following a predefined method instead to broadcast it.

When an overloaded node receives a file request, it should forward the request to one of the file's replica nodes. In order to load balancing between replica nodes, the overloaded node chooses the replica node according to Lottery scheduling method adopted by Plover.

b) An On-line Pointer Replication (OPR) algorithm [37]: It can efficiently reduce the query search latency. In order to realize this, OPR replicates the pointers of an object in multiple peers in the network. Therefore, the query for an object is forwarded to the nearest root among the others to fetch the location pointer, that reduces the query search latency as well as improves data availability. Latency incurred by a query is denoted by the number of hops it takes to route to the root.

OPR addresses two problems: Placement of Replicas and Extent of Replication. In Placement of Replicas, it decides how to place the replication pointers in the network to achieve the best performance. To find the best replica placement, OPR uses an heuristic approach called Greedy approach to select the roots. It bases on topologies hypercube to construct overlay topology of the network, because Greedy approach needs complete knowledge about the network layout. In a static network, hypercube can be easily embedded by connecting any two nodes that are one Hamming distance apart. In this article, Hamming distance of two nodes is the number of bits that are different in IDs of two nodes.

Using the greedy approach, the node with the largest distance to the nearest existing roots is selected as the next root and so until all roots will be selected. As the hamming distance represents a metric distance, OPR can easily identify the farthest node in the system.

In extent of Replication, it decides how to determine the replication degree for each object to achieve the best performance. OPR concludes that the optimal replication degree (number of pointers) is directly proportional to the query arrival rate and inversely proportional to the system churn rate.

Each root peer applies the maintenance protocol as follow: When a node wants to leave the network, it sends a message to its neighbors to inform them of its intention. Each neighbor receives this message should update its routing entries. The pointers kept on node leaving the network are pushed to the neighbor which has the ID numerically nearest to it.

Discussion: Plover strategy is different from all the strategies presented for structured P2P in different points:

Plover uses geographical clustering, such as making file replicas among physically close nodes based on nodes available capacities. By considering node available capacity and locality, plover achieves not only high efficiency in file replication but also facilitates efficient file consistency maintenance.

The consistency maintenance is an important issue. To the best of our knowledge, Plover is the only strategy which addressed this issue together with file replication relative to others strategies presented in this paper. It uses efficient file consistency maintenance with low-cost.

Additional, plover adopts lottery scheduling method to efficient achieve file query load balance between replica nodes, unlike symmetric and ID-Replication which use random method. In random load balancing, file query is routed randomly to a replica node. The random method is not efficient

 ISBN: 978-1-61208-361-2

TABLE I. COMPARISON BETWEEN DIFFERENT REPLICATION STRATEGIES IN STRUCTURED P2P NETWORKS.

Replication Technique and goal		Type of replica placement algorithm	Replica maintenance	Replication degree	Proximity	Replica nodes location	Feature
Achieve load balancing	Lightweight Adaptive system-neutral replication protocol [27]	owner replication	not mentioned	is one copy whenever the replication algorithm is applied	likely to find a replica preceding the root destination	piggybacking replica nodes location on messages	uses LRU as the replica replacement strategy
	A Prediction-Based Fair replication algorithm [28]	path replication	not mentioned	is adjusted according to the load value for each replication process	likely to find a replica preceding the root destination	piggybacking replicas nodes location on messages	uses LRU as the replica replacement strategy
	Symmetric replication [29]	multi publication key replication	more complex maintenance	theory is the same for each item	choose the numerically closet replica ID	each node can calculate the key of the different replicas	can achieve load balancing between replica by sending requests to a random replica. can be applied to all DHTs
Increase availability of resources	Quorum-based replication [31]	neighbor replication	provided by the root node	number of replicas can not exceed the neighbor list	request can be routed to the root node or to the replica node	piggybacking replicas node (replication sets) location on ping message	reduces the maintenance traffic under churn comparing the simple replication
	Availability-based replication [31]	neighbor replication	provided by the root node	is adjusted according to the node availability and the request data availability	request can be routed to the root node or to the replica node	piggybacking replicas node (replication sets) location on ping message	reduces the maintenance traffic under churn comparing the simple replication
	DHT-based self-adapting replication protocol [32]	multi publication key replication	not mentioned	is adjusted according to average peer online probability and the request data availability	choose the numerically closest replica ID	each node can calculate the key of the different replicas	generates less storage costs comparing to Availability based replication
	Predecessor replication [33]	neighbor replication	provided by the root node and by replica nodes	can not exceed the predecessor-lists size and is the same for each item	can find a replica node preceding the root node	root node maintains the localization of replica nodes	can reduce the number of hops needed to locate the requested data compared to neighbor, symmetric,
Enhance churn tolerance	RelaxDHT replication [34]	neighbor replication	provided by the root node and by replica nodes	number of replicas can not exceed the neighbor list size	can find a replica node preceding the root node	each node maintains metadata describing the localization of replicas sets	the root peer does not necessarily store a copy
	ID-Replication strategy [35]	neighbor replication	provided by replica nodes	is between two parameters Rmin and Rmax	request can be routed to random replica node in the group	it not necessary to maintain the information about the location of replica nodes	can allows more copies for popular data objects.
Improve search performance	Proactive Low-Overhead File replication scheme [36]	neighbor replication	not mentioned	is based on node available capacities	request is routed to the root node first	super node maintains metadata describing the localization of replica nodes for each replication file	adopts lottery scheduling method to achieve file query load balance uses efficient file consistency maintenance with low-cost
	An On-line Pointer replication algorithm [37]	multi publication key replication	provided by replica node	is directly proportional to the query arrival rate and inversely proportional to the system churn rate	choose the numerically closest replica ID	each node can calculate the keys of the different replicas	uses an heuristic called Greedy approach to select the roots

and it can choose the same replica node. Thus, the replica node can become overloaded.

Plover has not clearly stated how the peer applies the maintenance protocol. Unlike the others strategies which based on Multi Publication Key, OPR uses an heuristic approach called Greedy approach to select the roots. This method is simple to use and allows OPR to reduce the query search latency.

5) Summary: In this article, we explored all the techniques that have a significant scientific contribution. Each technique has a main replication objective to achieve and should take into consideration important factors and parameters. The latter have a direct impact on the system performance. Therefore, it is necessary to use the heuristics performing compromise between these factors and parameters proves to be necessary, in order to reduce the cost of replication without compromising its efficiency.

After this study, we try to draw some useful recommendations to take into consideration in replication strategies:

- Replication degree must not be the same for each data, because some of them are popular and others are not. Unpopular data must have less replicas than popular data. Thus, it decreases overhead of the maintenance protocol for unpopular data. Further, replication degree must be minimized as much as possible without compromising its efficiency. Therefore, it decreases the overhead of maintenance protocol and consistency maintenance.
- The search request must not be routed to the root node

first in order to not overcome it. This case can appear in the neighbor replication. Additional, it is necessary to increase the utilization of the replica nodes and to apply a mechanism of load balancing between replica nodes when it is possible.

- The consistency maintenance is an important issue, and it must be addressed together with the replication technique in order to reduce its overhead and to facilitate its execution. The replication technique must be designed to ease consistency maintenance like for example in neighbor replication, the root peer maintains the localization of replica nodes. Therefore, the root peer can easily forward the update message to the replica nodes if there is update. In path replication, it is very difficult to maintain the localization of all replica nodes. Then, not all data can be up-to-date.
- If P2P application requires mutual consistency, in our opinion the replication strategy which is based on multi publication key replication can facilitate mutual consistency. Each peer can calculate the key of different replica nodes. In this case, a replica peer can forward update message to other replica nodes. When a peer rejoins the network later, it contacts online replica nodes to recuperate updated data.
- Storage capacity constraint must not be ignored while the choice of suitable replica node and the replica placement strategy must applied when is necessary.

Table I summarizes the comparison of replication strategies for structured P2P networks presented above in function of some criteria: replica placement algorithm, replica maintenance, replication degree, proximity [7] (selecting a 'nearby' replica (in the ID space)), replica nodes location (existence of meta-information of the localization of replica nodes) and feature (other characteristics).

V. Conclusion and future work

Replication techniques are widely employed to improve the availability of data, enhancing performance of query latency and load balancing, in content distribution systems such as P2P. In this paper, a state of the art of the various replication techniques for structured P2P networks is presented. Thereafter, a new classification for these techniques is introduced, a detailed comparison is done. In our future work, we try to develop simultaneously data replication and data consistency maintenance methods and take into consideration recommendations that we presented in order to achieve high efficiency at a significantly lower cost.

References

[1] "Napster," [retreived: May, 2014]. [Online]. Available: http://www.napster.co.uk

[2] "Gnutella," [retreived: May, 2014]. [Online]. Available: http://www.gnutella.com

[3] "Kazaa," [retreived: May, 2014]. [Online]. Available: http://www.kazaa.com

[4] "Bittorrent," [retreived: May, 2014]. [Online]. Available: http://www.bittorrent.com

[5] J. Kubiatowicz et al., "Oceanstore: an architecture for global-scale persistent storage," SIGPLAN Not., vol. 35, no. 11, 2000, pp. 190–201.

[6] P. Druschel and A. I. T. Rowstron, "Past: A large-scale, persistent peer-to-peer storage utility," in HotOS, 2001, pp. 75–80.

[7] S. Ktari, M. Zoubert, A. Hecker, and H. Labiod, "Performance evaluation of replication strategies in dhts under churn," in Proceedings of the 6th international conference on Mobile and ubiquitous multimedia. ACM, 2007, pp. 90–97.

[8] E. K. Lua, J. Crowcroft, M. Pias, R. Sharma, and S. Lim, "A survey and comparison of peer-to-peer overlay network schemes," IEEE Communications Surveys and Tutorials, vol. 7, 2005, pp. 72–93.

[9] A. Rowstron and P. Druschel, "Pastry: Scalable, decentralized object location, and routing for large-scale peer-to-peer systems," in Middleware '01: Proc. IFIP/ACM international conference on Distributed Systems Platforms. Springer Berlin / Heidelberg, 2001, pp. 329–350.

[10] B. Y. Zhao et al., "Tapestry: A resilient global-scale overlay for service deployment," IEEE Journal on Selected Areas in Communications, vol. 22, 2004, pp. 41–53.

[11] S. Ratnasamy, P. Francis, M. Handley, R. M. Karp, and S. Shenker, "A scalable content-addressable network," in SIGCOMM, 2001, pp. 161–172.

[12] P. Maymounkov and D. Mazieres, "Kademlia: A peer-to-peer information system based on the xor metric," Peer-to-Peer Systems, 2002, pp. 53–65.

[13] I. Stoica, R. Morris, D. Karger, M. F. Kaashoek, and H. Balakrishnan, "Chord: a scalable peer-to-peer lookup service for internet applications," in SIGCOMM '01: Proc. 2001 conference on Applications, Technologies, Architectures, and Protocols for Computer Communications. ACM, 2001, pp. 149–160.

[14] H. Weatherspoon and J. Kubiatowicz, "Erasure coding vs. replication: A quantitative comparison," in IPTPS, 2002, pp. 328–338.

[15] W. K. Lin, D. M. Chiu, and Y. B. Lee, "Erasure code replication revisited," in In PTP04: 4th International Conference on Peer-to-Peer Computing. IEEE, 2004, pp. 90–97.

[16] O. A.-H. Hassan and L. Ramaswamy, "Message replication in unstructured peer-to-peer network." in CollaborateCom. IEEE, 2007, pp. 337–344.

[17] S. Mohammadi, H. Pedram, S. Abdi, and A. Farrokhian, "An enhanced data replication method in p2p systems," Journal of computing, vol. 2, 2010, pp. 1–5.

[18] C. Jacky, L. Kevin, and N. L. Brian, "Availability and popularity measurements of peer-to-peer file systems," 2004, [retreived: May, 2014]. [Online]. Available: http://forensics.umass.edu/pubs/chu.labonte.p2pjournal.pdf

[19] S. Manel and B. Mahfoud, "Toward a global file popularity estimation in unstructured p2p networks," in ICSNC 2013, The Eighth International Conference on Systems and Networks Communications, 2013, pp. 77–81.

[20] R. Bhagwan, S. Savage, and G. M. Voelker, "Understanding availability." in IPTPS, ser. Lecture Notes in Computer Science, M. F. Kaashoek and I. Stoica, Eds., vol. 2735. Springer, 2003, pp. 256–267.

[21] B. T. Loo, R. Huebsch, I. Stoica, and J. M. Hellerstein, "The case for a hybrid p2p search infrastructure," in IPTPS, 2004, pp. 141–150.

[22] G. Gao, R. Li, K. Wen, and X. Gu, "Proactive replication for rare objects in unstructured peer-to-peer networks," Network and Computer Applications, 2012, pp. 85–96.

[23] K. Puttaswamy, A. Sala, and B. Y. Zhao, "Searching for rare objects using index replication," in INFOCOM 2008. 27th IEEE International Conference on Computer Communications, Joint Conference of the IEEE Computer and Communications Societies, 13-18 April 2008, Phoenix, AZ, USA. IEEE, 2008, pp. 1723–1731.

[24] W. Ma, Y. Zhang, and X. Meng, "Distribution aware collaborative spread replication for rare objects in unstructured peer-to-peer networks," Journal of Networks, vol. 8, no. 5, 2013, pp. 991–998.

[25] L. A. Sung, N. Ahmed, R. Blanco, H. Li, M. A. Soliman, and D. Hadaller, "A survey of data management in peer-to-peer systems," School of Computer Science, University of Waterloo, 2005.

[26] B. Godfrey, K. Lakshminarayanan, S. Surana, R. Karp, and I. Stoica, "Load balancing in dynamic structured p2p systems," in INFOCOM 2004. Twenty-third AnnualJoint Conference of the IEEE Computer and Communications Societies, vol. 4, 2004, pp. 2253–2262.

 ISBN: 978-1-61208-361-2

[27] V. Gopalakrishnan, B. Silaghi, B. Bhattacharjee, and P. Keleher, "Adaptive replication in peer-to-peer systems," 24th International Conference on Distributed Computing Systems 2004 Proceedings, 2004, pp. 360–369.

[28] X. Zhu, D. Zhang, W. Li, and K. Huang, "A prediction-based fair replication algorithm in structured p2p systems," in ATC, 2007, pp. 499–508.

[29] A. Ghodsi, L. O. Alima, and S. Haridi, "Symmetric replication for structured peer-to-peer systems," in DBISP2P, 2005, pp. 74–85.

[30] Q. Lv, P. Cao, E. Cohen, K. Li, and S. Shenker, "Search and replication in unstructured peer-to-peer networks." in SIGMETRICS. ACM, 2002, pp. 258–259.

[31] K. Kim and D. Park, "Reducing replication overhead for data durability in dht based p2p system," IEICE Transactions, vol. 90, no. 9, 2007, pp. 1452–1455.

[32] P. Knezevic, A. Wombacher, and T. Risse, "Dht-based self-adapting replication protocol for achieving high data availability," in Advanced Internet Based Systems and Applications. Springer, 2009, pp. 201–210.

[33] F. Ben Guirat and I. Filali, "An efficient data replication approach for structured peer-to-peer systems," in Telecommunications (ICT), 2013 20th International Conference on. IEEE, 2013, pp. 1–5.

[34] S. Legtchenko, S. Monnet, P. Sens, and G. Muller, "Relaxdht: A churn-resilient replication strategy for peer-to-peer distributed hash-tables," TAAS, vol. 7, no. 2, 2012, p. 28.

[35] T. M. Shafaat, B. Ahmad, and S. Haridi, "Id-replication for structured peer-to-peer systems." Euro-Par'12 Proceedings of the 18th international conference on Parallel Processing, 2012, pp. 364–376.

[36] H. Shen and Y. Zhu, "Plover: A proactive low-overhead file replication scheme for structured p2p systems," in Proceedings of IEEE International Conference on Communications, ICC 2008, Beijing, China, 19-23 May 2008. IEEE, 2008, pp. 5619–5623.

[37] J. Zhou, L. N. Bhuyan, and A. Banerjee, "An effective pointer replication algorithm in p2p networks," in 22nd IEEE International Symposium on Parallel and Distributed Processing, IPDPS 2008, Miami, Florida USA, April 14-18, 2008. IEEE, 2008, pp. 1–11.

 ISBN: 978-1-61208-361-2