ICIW 2015

The Tenth International Conference on Internet and Web Applications and Services

ISBN: 978-1-61208-412-1

June 21 - 26, 2015

Brussels, Belgium

ICIW 2015 Editors

Steffen Fries, Siemens AG, Corporate Technology - Munich, Germany
Mario Freire, University of Beira Interior, Portugal

ICIW 2015

Foreword

The Tenth International Conference on Internet and Web Applications and Services (ICIW 2015), held between June 21-26, 2015, in Brussels, Belgium, 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 2015 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 2015. 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 2015 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that ICIW 2015 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 Brussels, Belgium, provided a pleasant environment during the conference and everyone saved some time to enjoy the charm of the city.

ICIW 2015 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 2015

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 2015 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
Marcos Baez, University of Trento, Italy
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
Jocelyn Aubert, Luxembourg Institute of Science and Technology (LIST), Luxembourg
Nahed A. Azab, The American University in Cairo, Egypt
Masoud Barati, Islamic Azad University - Kangavar branch, Iran
Bradley Barker, University of Nebraska-Lincoln, USA
Ana Sasa Bastinos, University of Ljubljana, Slovenija
Luis Bernardo, Universidade Nova de Lisboa, Portugal
Siegfried Benkner , University of Vienna, Austria
Emmanuel Bertin, Orange Labs, France
Giancarlo Bo, Technology and Innovation Consultant- Genova, Italy
Christos Bouras, University of Patras / Research Academic Computer Technology Institute, Greece
Laure Bourgois, INRETS, France
Mahmoud Brahimi, University of Msila, Algeria
Tharrenos Bratitsis, University of Western Macedonia, Greece
Maricela Bravo, Autonomous Metropolitan University, Mexico
Ruth Breu, University of Innsbruck, Austria
Claudia Canali, University of Modena and Reggio Emilia, Italy
Dung Cao, Tan Tao University - Long An, Vietnam
Miriam A. M. Capretz, The University of Western Ontario - London, Canada
Jorge C. S. Cardoso, Portuguese Catholic University, Portugal
Juan Carlos Cano, Universidad Politécnica de Valencia, Spain
Ana Regina Cavalcanti Rocha, Federal University of Rio de Janeiro, Brazil
Ajay Chakravarthy, University of Southampton, UK
Shiping Chen, CSIRO Digital Productivity Flagship (DPF), Australia
Xi Chen, Nanjing University, China
Zhixiong Chen, School of Liberal Arts - Mercy College, USA
Costin Chiru, University Politehnica of Bucharest, Romania
Dickson K.W. Chiu, University of Hong Kong, Hong Kong
Soon Ae Chun, City University of New York, USA
Paolo Ciancarini, University of Bologna, Italy
Gianpiero Costantino, Institute of Informatics and Telematics - National Research Council (IIT-CNR) of Pisa, Italy
María Consuelo Franky, Pontificia Universidad Javeriana - Bogotá, Columbia
Javier Cubo, University of Malaga, Spain
Roberta Cuel, University of Trento, Italy
Richard Cyganiak, Digital Enterprise Research Institute / NUI Galway, Ireland
Paulo da Fonseca Pinto, Universidade Nova de Lisboa, Portugal
Maria Del Pilar illamil Giraldo, Universidad de los Andes, Columbia
Maria del Rocío Abascal Mena, Universidad Autónoma Metropolitana - Cuajimalpa, Mexico
Gregorio Diaz Descalzo, University of Castilla - La Mancha, Spain
Ioanna Dionysiou, University of Nicosia, Cyprus
Matei Dobrescu, Insurance Supervisory Commission, Romania
Eugeni Dodonov, Intel Corporation- Brazil, Brazil
Ioan Dzitac, Aurel Vlaicu University of Arad, Romania
Matthias Ehmann, University of Bayreuth, Germany
Javier Fabra, University of Zaragoza, Spain
Evanthia Faliagka, University of Patras, Greece
Ana Fermoso García, Pontifical University of Salamanca, Spain

Adrián Fernández Martínez, Universitat Politecnica de Valencia, Spain
Gianluigi Ferrari, University of Parma, Italy
Stefan Fischer, University of Lübeck, Germany
Panayotis Fouliras, University of Macedonia, Greece
Chiara Francalanci, Politecnico di Milano, Italy
Steffen Fries, Siemens AG, Corporate Technology - Munich, Germany
Ingo Friese, Deutsche Telekom AG - Berlin, Germany
Xiang Fu, Hofstra University, USA
Roberto Furnari, Università di Torino, Italy
Ivan Ganchev, University of Limerick, Ireland
G.R. Gangadharan, IDRBT, India
David Garcia Rosado, University of Castilla - La Mancha, Spain
Rung-Hung Gau, National Chiao Tung University, Taiwan
Mouzhi Ge, Bundeswehr University Munich, Germany
Christos K. Georgiadis, University of Macedonia, Greece
Jean-Pierre Gerval, ISEN Brest, France
Mohamed Gharzouli, Mentouri University of Constantine, Algeria
Caballero Gil, University of la Laguna, Spain
Lee Gillam, University of Surrey, UK
Katja Gilly, Universidad Miguel Hernández, Elche, Alicante, Spain
Gustavo González-Sanchez, Mediapro Research, Spain
Feliz Gouveia, Universidade Fernando Pessoa - Porto, Portugal
Andrina Granić, University of Split, Croatia
Sven Graupner, Hewlett-Packard Laboratories - Palo Alto, USA
Carmine Gravino, University of Salerno, Italy
Patrizia Grifoni, CNR-IRPPS, Italy
Stefanos Gritzalis, University of the Aegean, Greece
Tor-Morten Grønli, Westerdals - Oslo School of Arts, Communication and Technology, Norway
Bidyut Gupta, Southern Illinois University - Carbondale, USA
Till Halbach, Norwegian Computing Center / Norsk Regnesentral (NR), Norway
Ileana Hamburg, Institut Arbeit und Technik, Germany
Sung-Kook Han, Won Kwang University, Korea
Konstanty Haniewicz, Poznan University of Economics, Poland
Takahiro Hara, Osaka University, Japan
Ourania Hatzi, Harokopio University of Athens, Greece
Martin Henkel, Stockholm University, Sweden
José Luis Herrero Agustin, University of Extremadura, Spain
Martin Hochmeister, Vienna University of Technology, Austria
Shigeru Hosono, NEC Corporation, Japan
Waldemar Hummer, Vienna University of Technology, Austria
Chi Chi Hung, Tsinghua University - Beijing, China
Hamidah Binti Ibrahim, Universiti Putra Malaysia, Malaysia
Muhammad Ali Imran, University of Surrey Guildford, UK
Raj Jain, Washington University in St. Louis, USA
Marc Jansen, Ruhr West University of Applied Sciences, Germany
Ivan Jelinek, Czech Technical University, Czech Republic
Jehn-Ruey Jiang, National Central University, Taiwan
Athanassios Jimoyiannis, University of Peloponnese, Greece

Nicolas Jullien, Institut TELECOM Bretagne & UEB, France
Monika Kaczmarek, Poznan University of Economics, Poland
Hermann Kaindl, Vienna University of Technology, Austria
Georgia M. Kapitsaki, University of Cyprus, Cyprus
Vassilis Kapsalis, Technological Educational Institute of Patras, Greece
Jalal Karam, Alfaisal University-Riyadh, Kingdom of Saudi Arabia
Vlasios Kasapakis, University of the Aegean, Greece
Brigitte Kerherve, UQAM, Canada
Suhyun Kim, Korea Institute of Science and Technology (KIST), Korea
Alexander Knapp, Ludwig- Maximilians-Universität München, Germany
Samad Kolahi, Unitec Institute of Technology, New Zealand
Kenji Kono, Keio University, Japan
Tomas Koubek, Mendel University in Brno, Czech Republic
George Koutromanos, National and Kapodistrian University of Athens, Greece
Leandro Krug Wives, Instituto de Informática - UFRGS, Brazil
Gurhan Kucuk, Yeditepe University, Turkey
Shuichi Kurabayashi, Keio University, Japan
Jaromir Landa, Mendel University in Brno, Czech Republic
José Laurindo Campos dos Santos, National Institute for Amazonian Research, Brazil
Friedrich Laux, Reutlingen University, Germany
Philipp Leitner, University of Zurich, Switzerland
Longzhuang Li, Texas A&M University-Corpus Christi, USA
Shiguo Lian, Orange Labs Beijing, China
Erick Lopez Ornelas, Universidad Autónoma Metropolitana, Mexico
Malamati Louta, University of Western Macedonia - Kozani, Greece
Zaigham Mahmood, University of Derby, UK / North West University, South Africa
Zoubir Mammeri, IRIT - Toulouse, France
Chengying Mao, Jiangxi University of Finance and Economics, China
Kathia Marcal de Oliveira, University of Valenciennes and Hainaut-Cambresis, France
Jose Miguel Martínez Valle, Universidad de Córdoba, Spain
Inmaculada Medina-Bulo, Universidad de Cádiz, Spain
Fuensanta Medina-Dominguez, Carlos III University Madrid, Spain
Andre Miede, University of Applied Sciences Saarbrücken, Germany
Fernando Miguel Carvalho, Lisbon Superior Engineering Institute, Portugal
Serge Miranda, University of Nice, France
Sanjay Misra, Federal University of Technology - Minna, Nigeria
Mohamed Mohamed, IBM US Almaden, USA
Nader Mohamed, UAE University, United Arab Emirates
Shahab Mokarizadeh, Royal Institute of Technology (KTH), Sweden
Arturo Mora-Soto, Mathematics Research Center (CIMAT), Mexico
Jean-Henry Morin, University of Geneva, Switzerland
Prashant R. Nair, Amrita University, India
T.R. Gopalakrishnan Nair, Prince Mohammad Bin Fahd University, KSA
Alex Ng, The University of Ballarat, Australia
Theodoros Ntouskas, Univeristy of Piraeus, Greece
Jason R.C. Nurse, Cyber Security Centre | University of Oxford, UK
Asem Omari, University of Hail, Kingdom of Saudi Arabia
Carol Ou, Tilburg University, The Netherlands

Christian Zirpins, Karlsruhe Institute of Technology, Germany
Jan Zizka, Mendel University in Brno, Czech Republic

Copyright Information

Table of Contents

Animating Information Dissemination Through a Gamified Online Forum

Shueh-Cheng Hu
Department of Computer Science & Comm. Engineering
Providence University
Taichung 43301, Taiwan, ROC
e-mail: shuehcheng@gmail.com

I-Ching Chen
Department of Information Management
Chung Chou University of Science and Technology
Chang Hua 51003, Taiwan, ROC
e-mail: jine@dragon.ccut.edu.tw

***Abstract*— Various online forums aim to facilitate information dissemination within particular knowledge domains. A successful forum needs active participation and contributions from individuals. Just like common information systems, most online forum systems were designed without consideration of motivating users. Gamification works of an online forum is presented in this article. The gamification work provides several game mechanisms motivating users to participate and contribute more actively, which eventually leads to a more effective platform for information sharing and opinions exchanging. The completed gamification work verified the feasibility of gamifying an open-sourced online forum system. In addition, a preliminary qualitative evaluation shows that diverse mechanisms brought different effectiveness and experiences to users.**

Keywords - Forum; information dissemination; motive; gamification; servlet.

I. Introduction

In the age of Internet, online forums have being deployed widely for people in similar knowledge domains or with close interests to share information and exchange opinions. Both professionals and amateurs expect to swiftly acquire knowledge and experiences via browsing contents in various online forums, or they will spend much more time to accumulate the same stuff by themselves. Obviously, to serve as an effective platform for information dissemination, an online forum needs active participation and contributions from its users. Many online forums gradually withered due to decreasing participation and contributions from users. Unfortunately, the presence of these gradually withered online forums will be further deteriorated by many search engines favoring Web sites with up-to-date and intensively-referred materials [1]. Even worse, the lack of participation/contribution and deteriorated search engine visibility will form a vicious circle. Accordingly, motivating users is a critical task for successfully operating an effective online forum.

There are many options for people who are looking for an online forum system, and many of them even come with free licenses. However, these online forum systems were designed by people who applied conventional software system development procedure and principles, which results in a system with complete and correct functions, but their users likely to have no strong motive to actively use these functions.

In light of the significance of an animating online forum system and the lack of relevant works, this work tried to renovate an online forum system: JForum [2], through gamifying it. The anticipated benefit is motivating its users to participate and contribute more actively.

The remaining parts of this article are organized as follows: Section II briefs prior works regarding the gamification and its applications in information systems; Section III describes the analysis and design works of gamifying the JForum; Section IV describes the corresponding implementation details; and the concluding remarks and future directions were provided in Section V.

II. Review of Prior Works

Just like the mentioned online forum systems, most information systems were designed without consideration of motivating users, because the traditional design philosophy only takes functionality and accountability into account, but ignore the role that user's motive play in an individual's overall performance. As a result, the so-called well-designed information systems offer complete functions that enable users to accomplish their assigned tasks correctly, but did not equip any mechanisms to motivate users to perform tasks more actively or even enthusiastically.

The routine tasks bore people; prior study indicated that employee's working quality will degrade if they experience boredom [3]. In highly informationized working environments, people heavily rely on various information systems to complete their routine tasks. Consequently, designers need to take user's motive into account while they are developing an information system. To create a more animating working environment in the age of informationization, gamification of information systems properly emerges as a popular approach.

Usually, gamification refers to embedding game mechanisms into a non-game environment, such as information system [4, 5]. The original idea is planting users' engagement and addiction that could be found while they are playing various games into their working environments. Its purpose is to strengthen users' motivation, i.e., it makes users perform tasks with more fun, stronger motive, and deeper engagement. Once each individual is well motivated, the overall performance of an organization could be improved consequently.

With widely recognition of its effects, gamification has been applied by enterprises to animate their employees, i.e., users of their information systems. According to a report from Gartner, over 70% of global 2000 enterprises have used gamification to renovate their information system before 2014 [6]. Among many successful cases, Starbucks gamified one of their supply chain management systems by ranking suppliers in a leaderboard according to their on-time delivery records. The operations of the gamified system motivated suppliers in the supply chain to fulfill orders on time, thus earned better ranks meaning better efficiency and administration. The consequence brought to Starbucks were higher percentage of on-time delivery of supplies, lower logistics cost, and more profits [7]. Delta airlines successfully used a mobile application (APP) with a number of game mechanisms to enhance its public recognition, customer loyalty, and revenue [8]. Not only in traditional manufacturing and service industries, gamification also was adopted by software developers, who integrated gamification components into software development process, and the preliminary results indicated that improved quality of software and corresponding documents [9].

Including the mentioned cases, the success of many prior experiences [10]-[13] collectively point out an important fact: successful gamification of an information system does not necessarily rely on complex game mechanisms. By contrast, the key factors are identifying the parts that do need motivating users, and then embed proper game mechanisms to increase users' motivation and engagement, which usually lead to better performance and outcomes [14].

III. Process of Gamifying An Online Forum

The JForum, an online forum system with Berkeley Software Distribution (BSD) license terms, was selected to gamify due to its openness and popularity. The gamification implies renovating an existing information system rather than creating a new one from scratch, thus openness is critical. Generally speaking, a gamification process comprises the following key activities before embedding game mechanisms into the existing system: understanding the users, setting missions, identifying motivations, and selecting effective game mechanisms accordingly [15]. This section delineates the particular process of gamifying the JForum in this work.

A. Users and Missions Analysis

Rationally, target users of an online forum share a common profile: they are willing to acquire knowledge during the course of interacting with others, so they likely to be socializers, explorers, and achievers according to the Bartle's player type categorization [16]. Socializers enjoy the interactions with peers in the forum, explorers are happy to find new information that they never know, and achievers can feel satisfaction by observing peers responded or recognized the helpful information that they provided.

Obviously, reasonable target business outcomes of a gamified online forum include more active participation and more productive contributions from users. Consequently, the missions of a gamification work should be to encourage people to join the forum, participate the discussions, and contribute (raise issues or provide information) more valuable contents.

B. Motivational Drivers Identification

With clear missions, then we need to identify factors that are able to drive people perform what we intend them to do. A number of motivational drivers [14] that suits with target users are described as follows:

1) Collecting: For a long time, people enjoy collecting of either physical or abstract things such as coins or number of friends on social networks, which are meaningful to the collectors, in terms of value, security, or social status. Furthermore, once a collection starts, people tend to complete it, so if a collection could be infinite, the collecting activities will keep going.

2) Connecting: Connecting with people, especially those who share common interests or characteristics with us, makes our life enjoyable. This explains the foundation of various associations, clubs, fellowships, etc. Although being a member of a forum itself is making connections with other people, users still have motivation to expand the connections with people outside of the forum.

3) Achievement: People get great satisfication from achievement, which usually means successfully dealing with challenges. The positive psychological feedback makes us be willing to rise to the same challenge repeatedly, even we know that we probably fail sometimes.

4) Feedback: The feedback means acknowledgement, recognition, or just response to initiators' actions or messages. Feedback enhances the sense of being noticed, so not receiving feedback is extremely demotivating to anyone. Providing feedback is very important to encourage continuous participation and contributions in a forum.

5) Autonomy: Just like average people, members of a forum does not want the contents quality or atmosphere of the forum they join shift to a situation that they dislike to see but unable to restrain it. Therefore, if there is a trigger available, they tend to take necessary actions when they encounter any latency leading to these situations, such as offensive or inappropriate contents.

6) Fear of punishment: Members in a forum, just like most members in a society, tend to avoid speech and behaviors leading to punishment, this tendency gradually develops social norms or the corresponding regulations in written. Unlike the other motivational drivers that stimulate people to do something, this factor prevents members from doing things that are improper in a particular circumstance.

After setting missions and identifying motivation drivers, the next step is to embed proper game mechanisms that are able to motivate the users.

C. Selecting A Set of Suitable Game Mechanims

A game mechanism refers to a component with which users interact during the process of playing games. Besides its visible part displaying on the user interface, a mechanism also includes a set of rules that govern how this mechanism works. To realize the mentioned missions for encouraging users' participation and meaningful contributions, the following 5 game mechanisms were selected to embed into the JForum. Each mechanism's characteristics and the motivational drivers it offers were described as follows:

1) Points: The most popular mechanism in various games and have been widely applied in commercial contexts to reward customers' loyalty. Points motivate users due to humans' intrinsic desire to collect things such as money, stamps, antiques, etc. Besides, awarding points to users for her/his participation and contributions is a rational approach to recognizing their activities in the forum. In this work, points will be awarded to encourage certain types of actions including login, raise new topic, post new message, reply a message; as well as will be deducted to discourage other types of actions such as posting inappropriate contents, abusive reporting, etc.

2) Leaderboard: This mechanism ranks users according to their achievements within a specific context, which usually be representable in the form of points. It forms a competitive atmosphere, which encourages those who dislike following behind peers to engage the forum more actively.

3) Badges: To award memebrs accumelating a certain amount of points, specific types of badges will be granted. Usually, there are multiple types of badges honoring achievements in different levels of difficulty, or with different types of works. In the latter case, collecting badges will motivate some members.

4) Facebook likes: To use the plugins Application Programming Interface (API) of the Facebook, members' messages could be exposed on the largest social network. This allows memebrs to make connections with other friends who are not in the same forum, as well as receive feedbacks of relevance from friends on Facebook.

5) Report of inappropriate contents: This mechanism echos members' motivations in two facets; one is autonomy and another is fear of punishment. The former one drives members to control the quality of contents or the atmosphere of the forum through supressing inappropriate contents. The latter one holds back members from posting contents that are evidently not sutiable in a particular forum. To avoid abusive reporting, the reported contents will be sent to administrators for judging whether they are really inappropriate or not. After judgement, the reported contents will be sustained or removed, and the point of the member who posted inappropriate contents or the member who abusively reported will be deducted accordingly.

IV. Implementation Details

The JForum is a Java servlet application being able to run on Apache Tomcat. Its design generally complies with the Model-View-Controller (MVC) architectural pattern [17], it has the 3-tier structure. Accordingly, planting game mechanisms needs to deal with components in different tiers. The gamification work needs to deal with a number of key components in each tier for accommodating the selected game mechanisms.

To minimize the interrelationship (coupling) between the original parts in JForum and the newly parts due to the gamification, the works were conducted with a loosely-coupled style, as Figure 1 shows. User identifiers were taken to serve as foreign keys of the data tables persisting mechanisms' status and records when it is rational. Besides, all gaming rules were collectively defined in a new package: "game". Doing so, it will make it easy to expand the gamification work, as well as to maintain either the original or the new gamification works.

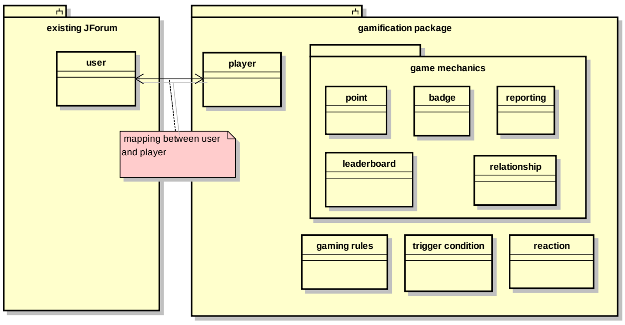


Figure 1. Loosely-coupled Architecture

Figure 2 illustrates the major components that need to be added and updated for gamifying the JForum, and each component is described as follows.

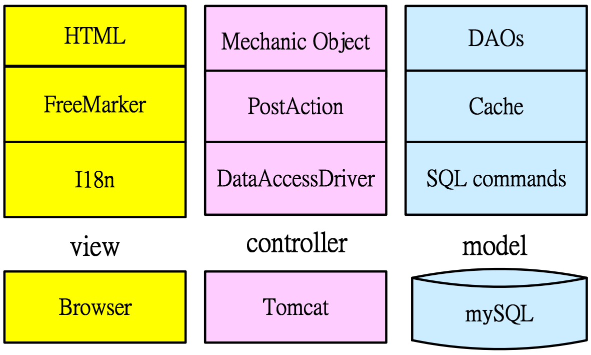


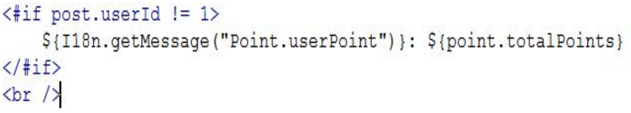
Figure 2. Key Components of the Gamification Work

A. User Interface (View)

The JForum uses the FreeMarker package [18] as a tool to separate Web pages design (view) and business logic programming (controller) works. The FreeMarker engine generates textual contents based on a template that contains HTML and FreeMarker Template (FTL) tags for dynamic Web contents, as well as a data model specifying data sources. Consequently, to embed a new game mechanism that will bring some new messages and graphics dynamically, it is necessary to update both the template and the data model of correspondence; the template decides the visual effect and format of the new components, and the data model tells where the displayed data come from dynamically when a particular page being accessed.

In addition, the JForum uses the I18 internationalization package for global users. For that reason, all new textual messages for embedding game mechanisms need to be added into the property file listing messages for a particular language (locale).

Figure 3 shows a FreeMarker fragment in a HTML file, which will display a message consisting of the word “point” in a default locale and the numeric value of the points earned by a non-administrator user.



```
<#if post.userId != 1>
    ${I18n.getMessage("Point.userPoint")}: ${point.totalPoints}
</#if>
<br />
```

Figure 3. A FreeMarker fragment

B. Game Objects and Corresponding Rules (Controller)

When thinking in the object-oriented way, each game mechanism obviously needs an object for holding its attributes and defining how it works. For example, to realize the “point” mechanism, a corresponding “Point” class will be defined; the attribute “userID” will bind it with a particular user, and the attribute “totalPoints” keeps track of the points being accumulated by the particular user. A method “changePoint” will be defined in the class to adjust the amount of points according to what that user did.

The PostAction object handles all actions for managing topics and posts, such as creating new topics or posting new messages, replying, deleting messages, etc. In other words, it deals with most major actions that the gamification work should focus on. So that, it needs linkage with some new game mechanisms. For example, when a user post a new message, the action of awarding points will be initiated by this object.

The DataAccessDriver abstract class defines the interface for linking the game mechanism and the persisting of the added game objects. Thus, all game mechanisms except the Facebook connection need to rely on one of its realization. Generally, the attributes of game objects will be fetched or stored via one of the corresponding concrete classes.

Since the servlet is the container for the JForum application, all business logics of the online forum systems and the new gaming rules were administrated and executed here.

C. Persistence (Model)

Taking flexibility of persistent storage into account, Data Access Object (DAO) design pattern was applied to separate the objects and its underlying storage mechanism. Thus, the object codes do not need to be changed due to switching to different data sources or APIs. DAOs provide abstraction and encapsulate all details for accessing the data source, which might be relational database, cloud storage, Lightweight Directory Access Protocol (LDAP), mainframe file systems, etc. The DAOs manage making connections with the data sources, as well as to fetch and store data. In the JForum system, each object was further divided into two layers of DAO for more flexibility, one is entity specific DAO, another is the generic DAO.

Cache mechanism in the JForum speeds up the object access operations through storing copies of object contents in Java virtual machine. So, many access operations do not need to access the database unless it is necessary to do so. To make contents of the persistent game mechanisms cacheable, addition of all persistent mechanisms needs the corresponding update in the cache mechanism.

All game mechanisms needing persistence require the corresponding tables in the database. Besides the properties of game mechanism objects, many gaming rules could also be encoded and then persisted in the database for more flexibility and maintainability. For example, the amount of points for awarding users should depend on the type of action they performed. However, the point amount for awarding a particular action could fluctuate rather than being constant for the sake of flexibility. Thus, these dynamics had better be encoded and persisted in the database, instead of being hardcoded in methods of mechanism class.

In summary, to illustrate the hierarchical operation of a game mechanism, the UML (Unified Modeling Language) [19] sequence diagram in Figure 4 shows the process of how components collaboratively worked to display the point of a user.

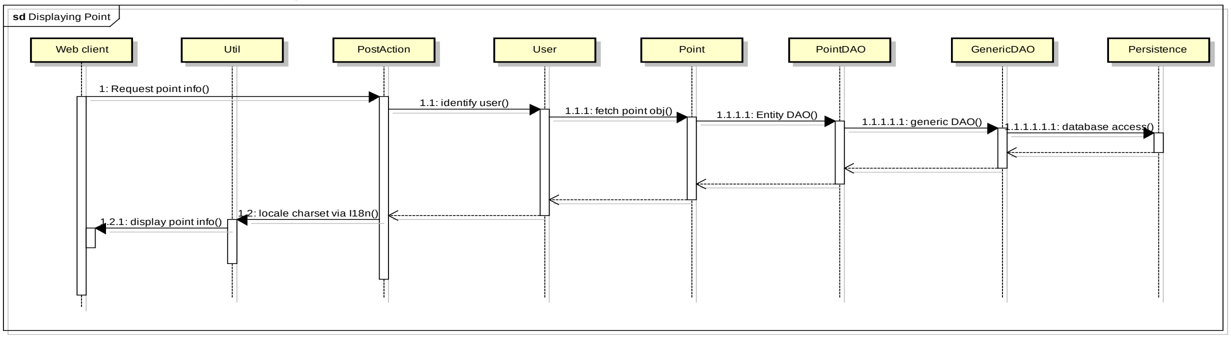


Figure 4. Working Process of the Point Mechanism

V. Discussion and Conclusion

The success of games relies on the proper application of motivational drivers that make players like and then be addicted to the game. Aiming to borrow the attractive features of various games, gamification techniques were used to improve users' experiences and engagement in non-game contexts for working, learning, or commercial purposes. The embedding of game mechanisms into an online forum system makes the system more engaging to users, who will feel animating while they are interacting and sharing knowledge with peers.

This article described the work of gamifying a JForum system realizing an online forum, including the work's motivation, process, and outcome illustrations. Besides, the present work shows the feasibility of gamifying an open-sourced software system by embedding 5 popular game mechanisms into the original code base with a loosely-coupled approach. A preliminary qualitative evaluation was conducted via interviewing with 8 users. They are computer science majored college students and usually spend a lot of time on visiting online forums to acquire information for completing their homework and projects. The responses from these target users of online forums indicated that the most significant impact brought by the present gamification on users is that they can observe the aspiring and competitive atmosphere, which was shaped by the badges and leaderboard and to some extent encouraged them to play more active roles in the information exchanging platform. The Facebook connections enable users to spread forum contents to their own social networks where there might be some constructive feedbacks regarding the topics in the forum. The mechanic of reporting inappropriate contents enable users to collectively maintain the quality of forum, which is important to the sustainable operation of a forum. By contrast, rewarding points made little difference due to the lack of redeeming mechanism enabling users to consume what they earned.

Besides studying how to embed other types of game mechanisms into an online forum, to better understand the effectiveness of particular game mechanisms in the context of online information sharing, the works being worthy of further investigations include the quantitative analysis of users' perceptions or satisfaction toward specific game mechanisms, and the quantitative evaluation of performance and productivity impact after particular gamification mechanisms being deployed.

Acknowledgment

This research work has being funded by the grant from the Ministry of Science and Technology, Taiwan, ROC, under Grant No. MOST 103-2221-E-126-014. We deeply appreciate their financial support and encouragement.

References

[1] J. B. Killoran, "How to Use Search Engine Optimization Techniques to Increase Website Visibility," Professional Communication, IEEE Transactions on, vol. 56, 2013, pp. 50-66.

[2] T. JForum. (Retrieved: April, 2015). Jforum Is a Powerful and Robust Discussion Board System Implemented in Java. Available: http://jforum.net/

[3] J. Watt and M. Hargis, "Boredom Proneness: Its Relationship with Subjective Underemployment, Perceived Organizational Support, and Job Performance," Journal of Business & Psychology, vol. 25, 2010, pp. 163-174.

[4] A. F. Aparicio, F. L. G. Vela, J. L. G. Sánchez, and J. L. I. Montes, "Analysis and Application of Gamification," in Proceedings of the 13th International Conference on Interacción Persona-Ordenador, 2012, p. 17.

[5] S. Deterding, D. Dixon, R. Khaled, and L. Nacke, "From Game Design Elements to Gamefulness: Defining "Gamification"," presented at the Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments, Tampere, Finland, 2011, pp. 9-15.

[6] C. Pettey and R. v. d. Meulen, "Gartner Predicts over 70 Percent of Global 2000 Organisations Will Have at Least One Gamified Application by 2014," Gartner, Barcelona, 2011.

[7] E. Sheely. (Retrieved: April, 2015). Case Study: How Starbucks Improved Supply Chain Efficiency with Gamification. Available: http://www.gamification.co/2013/11/18/fostercooperationgamfication/

[8] J. Hendricks. (Retrieved: April, 2015). Case Study: Delta's Nonstop Nyc Game That Got 190k Interactions in 6 Weeks. Available: http://www.gamification.co/2013/11/20/delta-nonstop-nyc/

[9] D. J. Dubois and G. Tamburrelli, "Understanding Gamification Mechanisms for Software Development," presented at the Proceedings of the 2013 9th Joint Meeting on Foundations of Software Engineering, Saint Petersburg, Russia, 2013, pp. 659-662.

[10] L. F. Rodrigues, C. J. Costa, and A. Oliveira, "The Adoption of Gamification in E-Banking," presented at the Proceedings of the 2013 International Conference on Information Systems and Design of Communication, Lisboa, Portugal, 2013, pp. 47-55.

[11] D. Krishna, "Application of Online Gamification to New Hire Onboarding," in The Third International Conference on Services in Emerging Markets, Mysore, India, 2012, pp. 153-156.

[12] O. Korn, "Industrial Playgrounds: How Gamification Helps to Enrich Work for Elderly or Impaired Persons in Production," presented at the Proceedings of the 4th ACM SIGCHI symposium on Engineering interactive computing systems, Copenhagen, Denmark, 2012, pp. 313-316.

[13] B. Gnauk, L. Dannecker, and M. Hahmann, "Leveraging Gamification in Demand Dispatch Systems," presented at the Proceedings of the 2012 Joint EDBT/ICDT Workshops, Berlin, Germany, 2012, pp. 103-110.

[14] J. M. Kumar and M. Herger, Gamification at Work: Designing Engaging Business Software. Aarhus C., Denmark The Interaction Design Foundation, 2013.

[15] G. Zichermann and C. Cunningham, Gamification by Design: Implementing Game Mechanics in Web and Mobile Apps: O'Reilly Media, 2011.

[16] R. Bartle, "Hearts, Clubs, Diamonds, Spades: Players Who Suit Muds," Journal of MUD research, vol. 1, 1996, p. 19.

[17] A. Leff and J. T. Rayfield, "Web-Application Development Using the Model/View/Controller Design Pattern," in Enterprise Distributed Object Computing Conference, 2001. EDOC'01. Proceedings. Fifth IEEE International, 2001, pp. 118-127.

[18] B. Geer, M. Bayer, and J. Revusky, "The Freemarker Template Engine," ed, 2004.

[19] Rumbaugh, James, Ivar Jacobson, and Grady Booch. Unified Modeling Language Reference Manual, The. Pearson Higher Education, 2004.

Web-based Graphic Design Framework to Support Users by Intuitively Reusing and Sharing Abstract Appearance Graphs

Nodoka Yamamoto
Faculty of Environment and Information Studies
Keio University
Fujisawa, Kanagawa, Japan
e-mail: mild.summer.y@gmail.com

Shuichi Kurabayashi
Faculty of Environment and Information Studies
Keio University
Fujisawa, Kanagawa, Japan
e-mail: kurabaya@sfc.keio.ac.jp

Abstract— **We propose a novel Web-based graphic design system that utilizes data-driven techniques to share and reuse the practitioners' knowledge of graphic design. A unique feature of this system is a user interaction model that utilizes an image being edited as a query for the knowledge base in order to recommend colors suited to it. The system stores images using a highly abstract graph structure, which we refer to as an Abstract Appearance Graph (AAG). The content being drawn at any given time is also modeled as an AAG. We have thus implemented a vector graphics design environment as an interactive graph database retrieval process to reuse existing designs. Our current prototype supports the extraction and reuse of the appearance properties of AAGs: the gradient of colors, their size, and positional relations. We implemented a cloud-based prototype system to create and share AAGs by using modern HTML5 technology in order to show the feasibility of our approach. We also performed experimental studies with impressional (*kansei*) graphic design knowledge. The experimental result shows that our system effectively supports graphic design by bringing to bear existing knowledge on new designs and rendering graphic design more flexible.**

Keywords- Graphic design; Color scheme; Image processing; Recommendation system;

I. INTRODUCTION

There is no end to the evolution of art and design. With the remarkable development of information and software technologies in recent times, it has become easier than ever before for anyone to create and share art. At the same time, however, this implies that anyone has come to be assessed and required artistic/aesthetic quality. It is evident that art and beauty are significant for a society [1]. Most universally, people recognize arts and aesthetic works as valuable things. Therefore, in today's society, it is necessary to constantly investigate ways to create high-quality works of art and design for anyone with advanced technology.

In the fields of art and design, a good sense in an artist or designer is accomplished only after having acquired considerable knowledge. In many cases, works of art and design consist of imitations of famous works. Artists and designers need to be versed in numerous technical details and styles used in existing works in order to create novel works of art and design by referring to them through a process of trial-and-error. Conventional works of art and design are sources of new works, so we can say that sharing more works contributes to more growth of art and design.

In light of the above, our aim in this study is to support graphic design by helping users effectively share and reuse existing works, which constitute design knowledge. In order to reuse conventional works for inspiration, artists and designers are required to abstractly comprehend the original author's intention through his/her works, and this accordingly requires technique. Nowadays, we have access to massive amounts of graphic design resources that are shared around the world. However, little research has been devoted to utilizing the large quantity of graphics-related resources available to us for the benefit of graphic design. While these innumerable resources have contributed to the worldwide growth of graphic design, the resources available to people at large are quite limited at present. Hence, a systematic framework that directly extracts the essence of graphic designs is necessary, not only for the growth of graphic design but also that of art.

Our concept is enhancement of graphic designer's creativity by systematizing the memory retrieval and the reuse of existing graphic design works. Our system supports graphic designers by suggesting design techniques and knowledge during his/her work. Therefore, in order to suggest technique of the graphic design for user's work interactively, we propose database construction methods, which analyze existing works for extracting technique and knowledge, and database retrieval methods.

In this paper, we propose a systematic way for graphic designers to accomplish the following: inputting and storing existing graphic designs, understanding and extracting typical features from stored graphics, and reusing the extracted features to create new graphic designs. To realize these objectives, we develop a new vector-based drawing tool on the Web specifically intended to facilitate the sharing and reusing of design knowledge possessed by designers. This knowledge consists of abstract and reusable features of existing graphic designs stored in an embedded database system. The system generates a query by capturing the status of a drawing in progress, which is submitted to the knowledge base to detect the most appropriate information related to the current drawing.

The most important feature of our method is a new data structure used to represent the flow of colors. It represents both the spatial distances in an image and the color distances

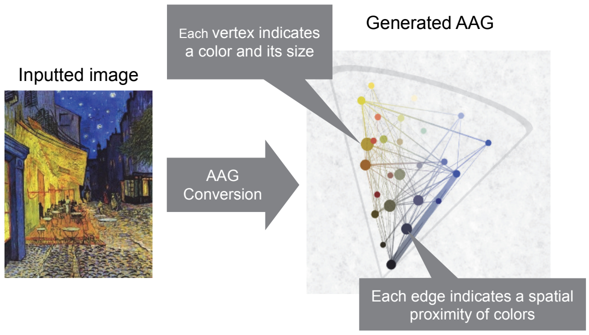

Inputted image
Each vertex indicates a color and its size
Generated AAG
AAG Conversion
Each edge indicates a spatial proximity of colors

Figure 1. An example of an Abstract Appearance Graph

measured according to a color metric in existing images. We refer to this data structure as an Abstract Appearance Graph (AAG). Representing the structure of colors, an AAG can be visualized in an intuitive manner, where it models color construction on a bitmap image by analyzing the relationship among all pixels using color distance and spatial distance. In general, the impression of colors is heavily affected not only by their discrete features, but also by their inter-relationships. This is because the visual stimulation of human senses is caused by relative mechanisms. Therefore, we assume that utilizing such graph structures is appropriate to comprehend the color design of graphics.

Figure 1 shows an AAG model, where a vertex represents a color, the size of the vertex denotes the color ratio in the original image, and an edge represents the spatial proximity between any two colors in the original image. Thus, the AAG represents the color status of image as a set of color vertices and edges connecting them.

In our proposed method, when a user creates a new graphic, the system recommends a relevant AAG that reflects the distribution of colors in the user's current work through a real-time analysis of its color structure. This smart graphic design environment also helps a user choose appropriate colors by reviewing his/her previous work, motifs, as well as the stored graphic design knowledge of all users. Our Web-based AAG sharing/reusing method is applicable to many fields, such as paint tools, Web design tools, and fashion outfitters because our method offers a smart method to determine sets of colors and distances between colors.

The remainder of this paper is structured as follows: Section 2 contains a summary of related research, and Section 3 is devoted to a structured overview of our system. We define the related fundamental data structure and algorithms in Section 4, and describe the implementation of our prototype system in Section 5. Section 6 contains an evaluation of the usability of our system, and we offer our conclusions in Section 7.

II. Related Works

With the recent advances in data processing, it has become popular to study design support systems, such as color scheme evaluation systems [2], color scheme support systems [3] [4], and automatic design support systems [5]. These systems exploit the relationships between colors and their evocative features by using the Color Image Scale [6] or fuzzy rules including it. The Color Image Scale [6] seeks to distill the emotional effects of color combinations, and consists of over 1,000 three-color combinations of 130 basic colors matched with key image words designed to reflect any mood or lifestyle. The Color Image Scale forms the basis of numerous studies and projects nowadays. Researchers studying color scheme support systems [3] [4] proposed a method to recommend harmonious color schemes from an input color and a keyword by utilizing fuzzy color schemes. These approaches are highly likely to correspond with direct user demand because they are based on textbook theories of design. However, because there is no "right" answer in the field of graphic design, supporting graphic design through specific rules cannot satisfy the designer's need for expansive and innovative work. From a broader perspective, a more open and flexible design support system is necessary to improve the design process.

Adobe Color [7] is a well-known service for creating and sharing color themes comprising five colors. Several color theme extraction methods have been proposed [8] [9] [10] by investigating sample models created by users. In order to reuse design knowledge, past research [11] [12] [13] has proposed methods that reflect the color scheme or color tone of an existing image in a target image. Furthermore, many community services have been developed to share art/design works, for example Dribbble [14], Behance [15], and Pixiv [16]. Indeed, there is considerable demand for means of sharing intricate knowledge in graphic design.

Based on the foregoing, we propose in this paper a systematic and direct method of supporting graphic design through a highly extensible and flexible social framework. An AAG, the data structure of graphic design in our system, is created from existing graphic design works selected by users as knowledge of color schemes. Therefore, the system permits all users to extend and benefit from the rule base as a knowledge sharing tool, as in [7] [14] [15] [16]. Prevalent research on design support, such as [3] [4] [5], does not have this feature. Furthermore, because an AAG consists of relationship among colors, the system can interactively support users' design work by recommending "flows of colors:" feasible colors for the work at hand. Thus, our system can construct a graphic design framework exhibiting extensibility and immediacy, and offers new possibilities for the development of graphic design.

III. System Architecture

The objective of our system is to recommend an appropriate color scheme for a user in real time by analyzing the color structure of the drawing in question in an on-the-fly manner. As shown in Figure 2, our design supporting mechanism consists of two phases: a storing design knowledge phase, and a real-time design support phase.

In the storing graphic design knowledge phase, our system constructs a database that stores design knowledge by converting existing graphic design objects, such as bitmap images, into simpler graph structures of color schemes. This color graph structure is the Abstract Appearance Graph. The AAG comprises the appearance properties (for example,

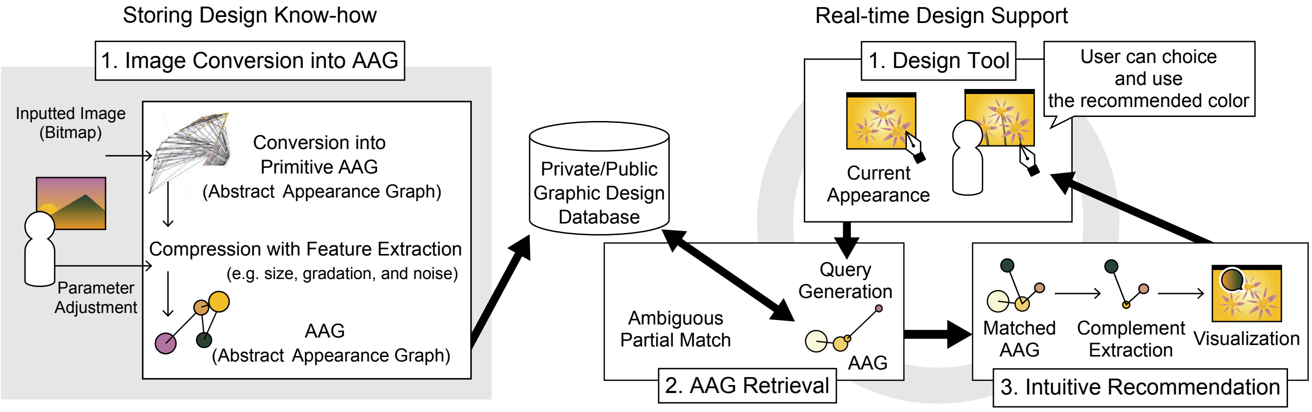

Storing Design Know-how
1. Image Conversion into AAG
Inputted Image (Bitmap)
Conversion into Primitive AAG (Abstract Appearance Graph)
Compression with Feature Extraction
(e.g. size, gradation, and noise)
Parameter Adjustment
AAG (Abstract Appearance Graph)
Private/Public Graphic Design Database
Real-time Design Support
1. Design Tool
User can choice and use the recommended color
Current Appearance
Query Generation
Ambiguous Partial Match
AAG
2. AAG Retrieval
Matched AAG
Complement Extraction
Visualization
3. Intuitive Recommendation

Figure 2. System architecture of the graphic design framework using AAG.

color and texture) and the positional relationships of colors. The system provides the user with an image conversion function, because of which the graphic design database of the system can be freely extended. When the system generates an AAG from an existing graphic design image, it analyzes the proximity and the contiguity between colors to detect the design elements of the image.

The AAG describes the proximity and contiguity between colors using vertices and edges. We can detect the spatial and semantic distance between colors by counting the number of hops between two vertices. The details of the AAG generation process are described in Section IV. It is important to render data structures simpler for accurate retrieval calculation because a pure bitmap image includes manifold colors in many cases. The graphic design storage module then stores the generated AAGs.

In the real-time design support phase shown in the right side of Figure 2, the system implements an interactive collaboration between a design tool and an AAG retrieval engine. The design tool provides functions for editing and creating graphic designs containing color schemes. It is important to mention that this design tool contains a function to convert current drawing content into an AAG data structure. The retrieval engine calculates the correlation between the working appearance of the design tool and the AAG objects stored in the graphic design knowledge storage module in order to obtain design elements relevant to the appearance of the drawing at hand. Following this, the retrieval engine performs an ambiguous partial match to extract reusable elements from the AAG obtained. The visualization interface then intuitively displays the obtained graphic design elements that complement the current drawing.

IV. Data Structure and Algorithms

As described in the previous section, AAG is a core data structure of our system. An AAG is represented as a graph $G = (V, E)$, where V is a set of vertices that denote simplex color elements in the image and E are edges that denote spatial proximity between colors (vertices) in the image. Each vertex has RGB color components and a size (a ratio to the image), and each edge has an angle of spatial proximity. In this section, we describe the algorithm to create an AAG object from an image and the retrieval function that scours the AAG database for design elements complementary to a drawing being edited using the design tool.

A. Conversion from Image to AAG

The AAG is constructed through the following two processes: gradient abstraction process and integration of similar vertices and edges. We detail the following conversion method on the assumption that the image to be converted is a bitmap image. A vector image data can then be easily converted into an AAG because its structure is already simple.

Process-1) Gradient Abstraction Algorithm: We designed a bottom-up abstraction method — in other words, a redundancy and noise reduction method — to convert a bitmap image into an AAG. The system aims to generate highly abstract data that includes texture information, such as gradient and noise, by analyzing relationships between adjacent colors. Our method uses the following algorithm to discover and abstract the gradient parts of an AAG. This is a recursive algorithm that repeats the analysis of three vertices.

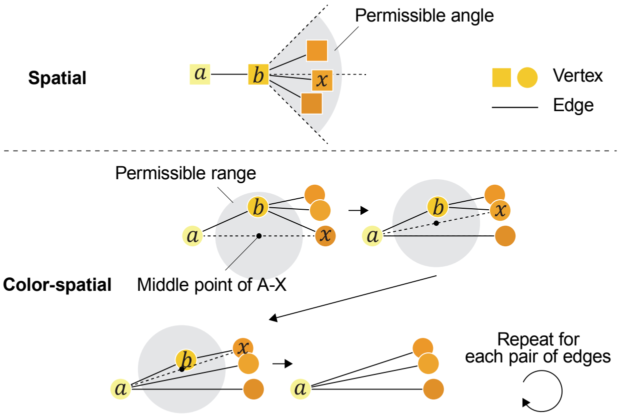


Figure 3. Schematic view of the gradient abstraction algorithm of the AAG.

Step 1) Convert an image into a color graph. Each vertex has a color and a size, which represents the pixel count. Each edge has an angle, a hop count (the initial value is 1), and gradient length (the initial value is 0). The initial angle of an edge is in units of 45 degrees because of contiguity on pixels, and edges consisting of the same colors at different angles are regarded as different edges.

Step 2) This step is the core phase of this algorithm. The system removes a redundant edge and introduces a directly connected edge instead. As shown in Figure 3, when we have three vertices (a, b, x) and two edges $\{a, b\}$ and $\{b, x\}$, the system deletes $\{b, x\}$ (and the opposite edge) and creates $\{a, x\}$ (and the opposite edge) if they meet the following conditions:

- We define the threshold for the difference between the magnitude of the angles formed by $\{a, b\}$ and $\{b, x\}$ as 45 degrees. This threshold is used to detect pixels in the same gradient direction in the bitmap image.
- In Euclidean color space, the distance between b and the midpoint of ax does not exceed the (constant rate of the) length of ax. This calculation uses RGB color components because the graphics software usually generates linear gradients in RGB space.

The newly created edge $\{a, x\}$ has the following properties: the angle is an average of the previous edges $\{a, b\}$ and $\{b, x\}$, and the number of hops is a summation of $\{a, b\}$ and $\{b, x\}$ because the new edge $\{a, x\}$ maintains the neighborhood relationships of pixels corresponding to the vertex b. The gradient length increases by the summation of the gradient lengths of the two edges because the previous edges are grouped into the gradient.

Following this, the system removes vertex b and the corresponding edges $\{a, b\}$ and $\{b, a\}$ if and only if vertex b is isolated from all vertices, except vertex a, by the above process.

Step 3) Repeat Step 2 until no edge meets the conditions.

Process-2) Integration of Similar Vertices and Edges: Using the above gradient abstraction method, the system reduces the level of detail in the AAG, by integrating smaller vertices into a larger vertex. The system calculates the distance between all vertices in the AAG and, the calculated distance is smaller than the predefined threshold if and only if the system integrates a smaller vertex into a larger one. The detailed process is as follows:

Step 4) The system measures the distance between every pair of vertices in the AAG. If and only if the distance is smaller than a threshold, the system integrates the smaller vertex into the larger vertex, which is a neighbor of the target vertex. This new vertex has a larger size than either of the original vertices because the size of the smaller vertex is added to that of the larger one. Due to this size control principle, a user can know the size of the color represented by the vertex in the original image. The vertices of our AAGs have two properties besides a color. The first is a size property calculated by summing up the pixel areas of same (or similar) colors. The second property is a range property regarding the area to which a specific vertex relates. The system calculates this range property by choosing the larger value from a range property of the larger vertex, and the distance between the smaller vertex and the larger one. The edges connecting the original smaller vertex are then shifted to the new, larger vertex.

Step 5) Every pair of edges connecting each vertex, whose angles are closer than a threshold, are integrated. When two edges are integrated, their properties are determined according to Step 2 of the gradient abstraction method.

B. AAG Retrieval

The AAG data structure plays two roles in this system. The first role is referred to as the “working AAG,” which appears on the image drawn by the user. The second role is referred to as the “model AAG,” which is obtained from existing images and is used as a source of design knowledge. Design support is achieved by recommending highly relevant sub-graphs from the model AAG, where the colored vertices in the working AAG are used as the starting point. To obtain sub-graph recommendations, the system calculates the relevance between vertices in the following manner:

Step 1) Selecting vertices in working AAG: The system selects a set of vertices from the working AAG, denoted as $V \in W$, where V are the selected vertices specified using a color and W is the working AAG. The system calculates the distance between each vertex in the model AAG, denoted as M, and each vertex in V. Each vertex, which is denoted as v, contains RGB color information. Moreover, each vertex in M has a range property. We implement the distance of vertices equation as in (1):

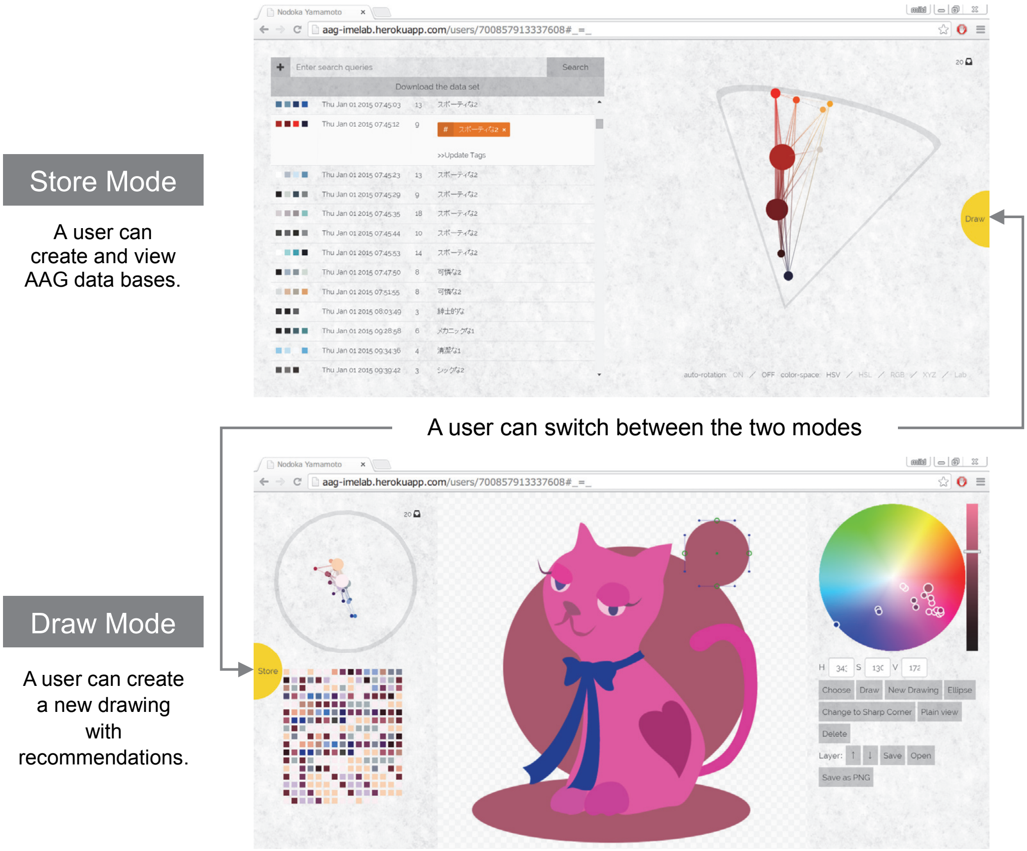
$$dist(v_w, v_m) := x\begin{cases} 0, & x < 0 \\ x, & x \geq 0 \end{cases} \tag{1}$$
$$x := \sqrt{(r_w - r_m)^2 + (g_w - g_m)^2 + (b_w - b_m)^2} - range_m$$

where v_w denotes the vertex in V and v_m denotes the vertex in M, r_w, g_w, and b_w are the components of v_w, and r_m, g_m, b_m, and $range_m$ are the components of v_m. This is a prototype implementation, and we plan to apply the CIEDE2000 delta equation [17] to improve the precision of distance calculations.

Step 2) Selecting the most relevant model AAG: The system selects the most appropriate AAG by calculating the relevance scores between the model AAGs and the current working AAG using (2):

$$score := \sum_{i=1}^{m} (t - d_i) \,\Big|\, d_i \leq t, \qquad d_i \leftarrow MIN\big(dist(i, \forall x \in M)\big) \tag{2}$$

where t denotes the threshold, i denotes the i-th vertex in the current working AAG W, and $\forall x \in M$ denotes each vertex in the model AAG M.



Store Mode

A user can create and view AAG data bases.

A user can switch between the two modes

Draw Mode

A user can create a new drawing with recommendations.

Figure 4. Prototype implementation system of the graphic design environment with HTML5 technology.

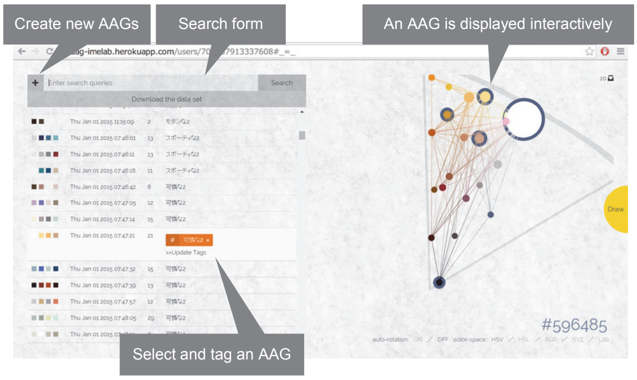
Step 3) Selecting a connected vertex in the selected model AAG: As the final step in design support, the system finds the connected vertex and places the partial graph of the selected model AAG into the current working AAG. The system retrieves the target vertices from the working AAG according to the relevance score. If the distance between a vertex in the model AAG and the working AAG is below the threshold t, the system selects it as a candidate. Thus, the system obtains the candidate vertices, which are denoted by $C \in M$. The system then obtains the vertices that are neighbors of each vertex in C. Finally, the system displays these neighbor vertices on the current drawing of the design tool as recommendations.

V. IMPLEMENTATION

In this section, we describe a prototype implementation system of our study.

A. *Prototype System*

We developed a prototype system using the modern Web technologies HTML5 Canvas, WebGL with Three.js, and d3.js, as shown in Figure 4. The interface of the prototype system comprises two modes: a store mode for accumulating the AAG database, which contains an AAG conversion module, and a draw mode for creating a new graphic by utilizing the AAG database, which contains a graphic design tool module, an AAG retrieval module, and an intuitive proposal module. A user can switch between the both modes at will.



Create new AAGs

Search form

An AAG is displayed interactively

Select and tag an AAG

Figure 5. Store mode of the prototype system: a list and 3D visualization of AAG.

1) *Store Mode*

In the store mode, a user can load existing images to construct his/her own AAG database. A list of AAG objects created by the user is displayed on the left side of Figure 5,

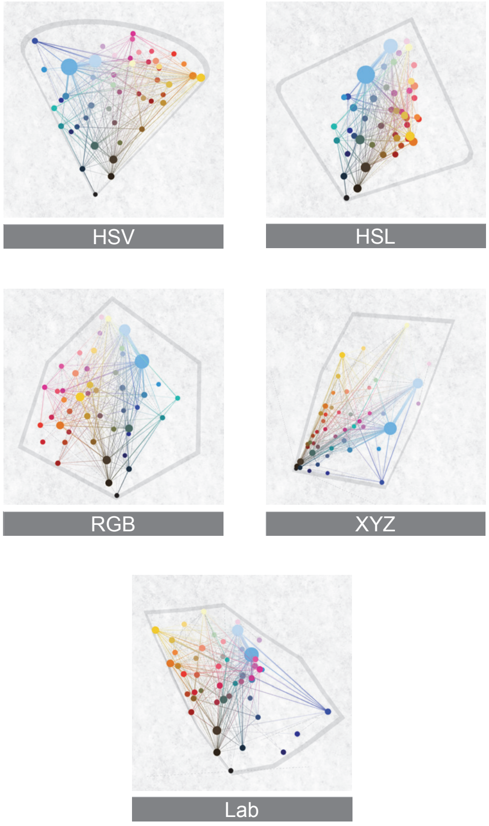
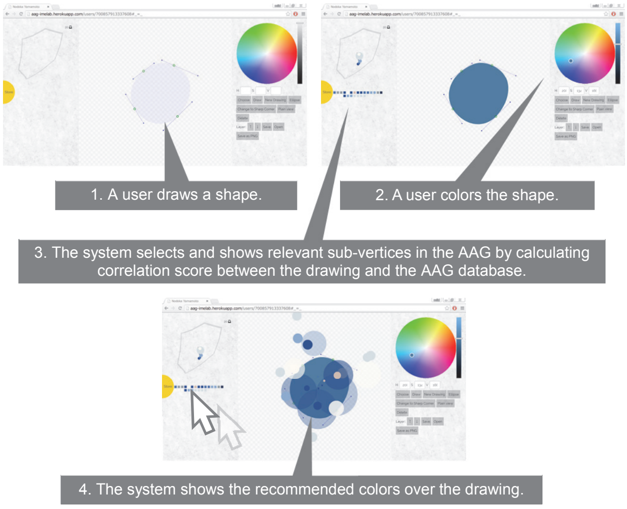


Figure 6. Five kinds of visualization color space in the prototype system.

and a 3D color space visualizing the relationships among colors in an AAG is displayed on its right side. The user can select bitmap images as new graphic design knowledge by clicking the "+" button. The system converts the input images into AAG objects, following which the AAG is represented in 3D color space developed using WebGL.

In this 3D visualization, the coordinates are equivalent to a color (color components). Each vertex of the AAG is plotted at its color coordinates regardless of its absolute position on the original image, and the relationships between colors are represented as edges connecting vertices. The dashed lines represent normal edges, and graded relations between two colors are indicated as gradient lines with a thickness corresponding to the gradient property. A user can interactively control the viewpoint. When a user clicks a color vertex, the connected vertices of that color are highlighted. Further, the user can choose the kind of the visualization color space from HSV, HSL, RGB, XYZ, and Lab, as shown in Figure 6. Hence, the user can easily grasp the color structure of the AAG.

The converted AAG can be stored in the database. In the store mode, a user can tag his/her own AAGs with appropriate words, and can search AAGs using tags from all users' databases. When the "Download the data set" option is selected, the listed AAG objects are downloaded and stacked as dataset of the retrieval module in the draw mode. In order to take advantage of its extensibility, one of the important features of our system, the dataset is not treated as a static entity. The system extracts the AAG dataset through queries.



1. A user draws a shape.

2. A user colors the shape.

3. The system selects and shows relevant sub-vertices in the AAG by calculating correlation score between the drawing and the AAG database.

4. The system shows the recommended colors over the drawing.

Figure 7. UI flow in the draw mode of the Web-based prototype system.

2) *Draw Mode*

In the draw mode, as shown in Figure 7, a user can draw an image consisting of Bezier curves using a graphic design tool that we developed on HTML5 Canvas. When a user provides or changes the color of a shape, the system searches for complementary AAG objects in the dataset. The interface then displays the color swatches of neighboring vertices that correspond to the colors on the current drawing. When the user positions the mouse pointer over it, the color is intuitively overlaid and dynamically reflects the size, distance, and angle in the model AAG. A user can thus re-color a shape by dragging and dropping the color swatches to the current drawing space.

Moreover, it is an important merit of our Web-based prototype implementation that the system can record all users' operation histories, such as the process according to which a user created a work, and the recommendation adopted by him/her for a work.

B. *Use Case*

We now present a use case and a use flow of the prototype system. The use case involves a user, Sharaku, who wants to create a drawing like Ukiyo-e, a genre of Japanese antique woodblock prints and paintings.

Sharaku first needs to construct the graphic design knowledge base of Ukiyo-e in store mode. Figure 8 shows that Sharaku has 20 favorite pictures from the genre Ukiyo-e; he enters them with the tag "ukiyo-e," and the system converts the input images into AAG objects without pausing, and lists them as shown in Figure 9. From this, the graphic design knowledge base associated with "ukiyo-e" is created and is open to the public. The AAG dataset for the recommendation system is then downloaded and stacked when the user clicks the "Download the data set" button.

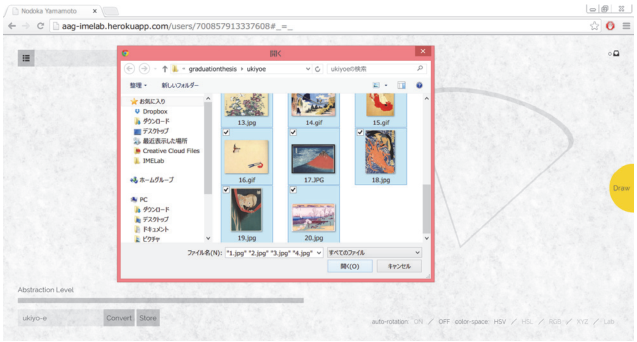
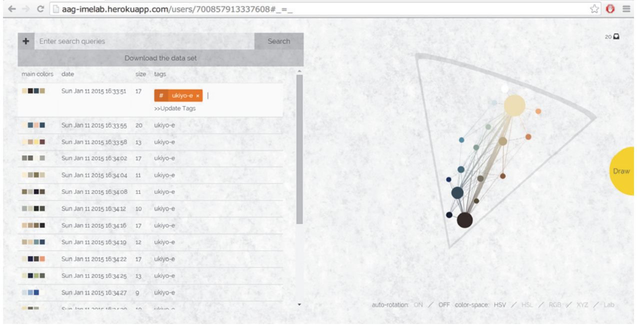

Nodoka Yamamoto
aag-imelab.herokuapp.com/users/700857913337608#_=_
開く
graduationthesis ▸ ukiyoe
13.jpg
14.gif
15.gif
16.gif
17.JPG
18.jpg
19.jpg
20.jpg
ファイル名(N): "1.jpg" "2.jpg" "3.jpg" "4.jpg"
すべてのファイル
開く(O)
キャンセル
Draw
Abstraction Level
ukiyo-e
Convert
Store
auto-rotation ON / OFF color-space HSV / HSL / RGB / XYZ / Lab

Figure 8. Use Case: Selecting Knowledge Images of *Ukiyo-e* in the Store Mode


aag-imelab.herokuapp.com/users/700857913337608#_=_

Enter search queries | Search

Download the data set

main colors	date	size	tags
■■■■	Sun Jan 11 2015 16:33:51	17	# ukiyo-e × >>Update Tags
■■■■	Sun Jan 11 2015 16:33:55	20	ukiyo-e
■■■■	Sun Jan 11 2015 16:33:58	13	ukiyo-e
■■■■	Sun Jan 11 2015 16:34:02	17	ukiyo-e
■■■■	Sun Jan 11 2015 16:34:04	11	ukiyo-e
■■■■	Sun Jan 11 2015 16:34:08	11	ukiyo-e
■■■■	Sun Jan 11 2015 16:34:12	10	ukiyo-e
■■■■	Sun Jan 11 2015 16:34:16	17	ukiyo-e
■■■■	Sun Jan 11 2015 16:34:19	12	ukiyo-e
■■■■	Sun Jan 11 2015 16:34:22	17	ukiyo-e
■■■■	Sun Jan 11 2015 16:34:25	13	ukiyo-e
■■■■	Sun Jan 11 2015 16:34:27	9	ukiyo-e

Draw
auto-rotation ON / OFF color-space HSV / HSL / RGB / XYZ / Lab

Figure 9. Use case: The list of created AAGs in store mode.

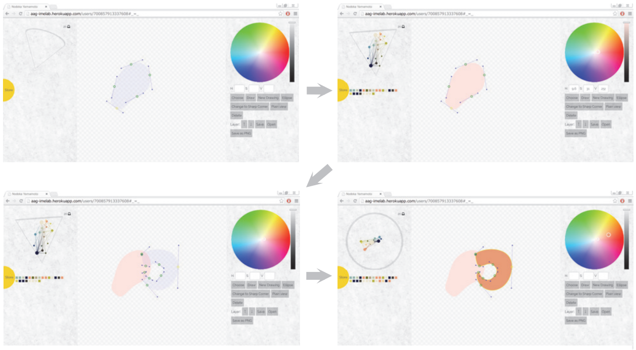


Figure 10. Use case: The flow of the drawing with recommendations in draw mode.

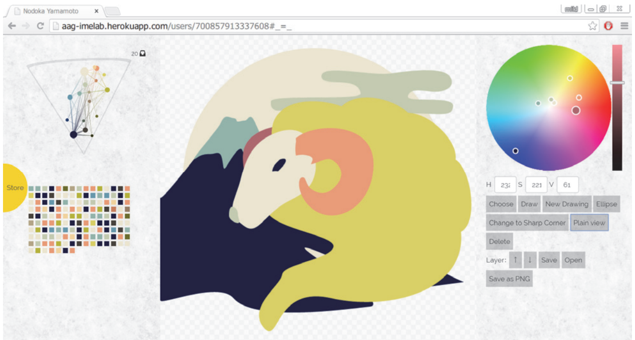


Figure 11. Use Case: The drawing of a sheep in the style of Ukiyo-e.

Sharaku can now create a drawing in the Ukiyo-e style using the color scheme support of the system in draw mode. He shifts to draw mode with the ukiyo-e dataset and starts working.

Sharaku is thinking of drawing the face of a sheep in the Ukiyo-e style. When he draws a shape resembling the face of a sheep and colors it beige, the system retrieves an AAG with a color matching the face color and automatically shows color swatches, as shown in Figure 10. All Sharaku has to do is pick up the ones he likes. Of course, if there is no appropriate AAG for the work at hand, the system does not recommend colors.

When Sharaku moves the mouse cursor over the swatches, the colors are dynamically displayed around the shape. In this case, when the mouse pointer hovers over the navy swatch, a big navy circle appears on the given shape.

Using the recommendation in this case, Sharaku adopts dull orange as the color of the sheep's horns. He drags and drops the swatch on the horn. The system once again retrieves an AAG following this color assignment.

In this manner, Sharaku creates a drawing using real-time recommendation from the system, as shown in Figure 11.

VI. Evaluation

In this section, we report an experiment to examine the effectiveness of our proposed system using our prototype implementation. The purpose of the experiment was to determine whether the AAG-based system adequately conveys graphic design knowledge from user to user and is sufficiently flexible for graphic design.

The experiment consisted of three phases: collecting images for design knowledge base, creating graphic works with the design knowledge base, and assessing the created works. We will explain these phases in order followed by the result of the experiment.

We assig(without/with recommendations).

ned seven subjects the task of collecting four groups of graphic design images from Internet severally. Each subject was provided with four impressions in order to create impressional (*kansei*) datasets for color recommendations. For example, a certain subject collected "clean," "modern", "energetic" and "antique". The four impression words were selected in order to not be similar to each other. We also told the subjects that the selected images should provide the relevant impression using colors because an AAG does not contain any information, such as shapes and structures, except colors. For example, images of one's favorite dishes are not an adequate dataset to represent the word "tasty" in this experiment, whereas images that appear "tasty" to many people according to their color distribution are considered fit. We thus created 28 AAG datasets from the collected graphic design images.

Following this, we assigned seven subjects the task of creating four pairs of "cat" or "dog" drawings conveying either impression, using the draw mode of our prototype system in two ways: with and without recommendations from the impressions dataset. That is, each subject had to create eight drawings. We distributed the tasks such that a subject who had collected images according to an impression

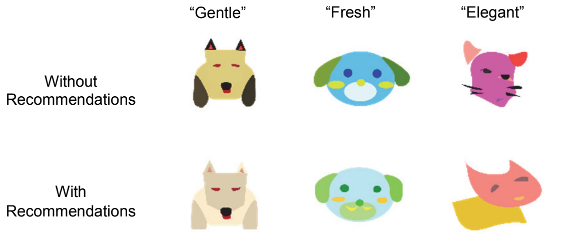


Figure 12. Samples of drawings created by subjects in the evaluation experiment.

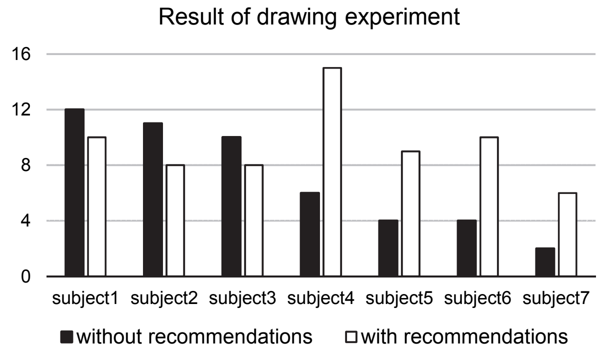


Figure 13. Seven subjects' scores in the drawing experiment.

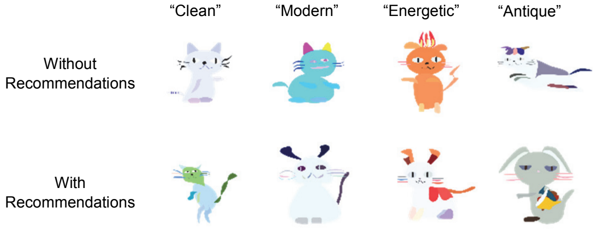


Figure 14. Drawings created by Subject 1.

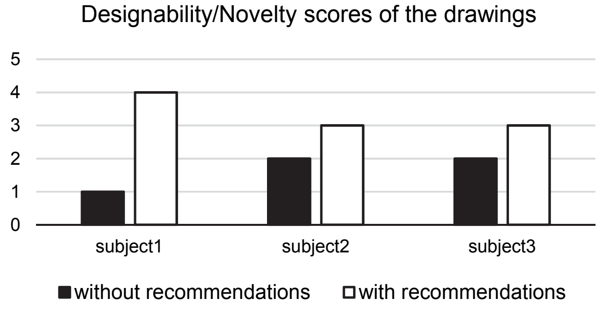


Figure 15. The result of novelty assessment in the drawing experiment.

in the previous phase was tasked to create images corresponding to a differ impression in this phase. Figure 12 shows sample images drawn by the subjects.

In the third phase of the experiment, we asked seven subjects to assess whether the created drawings conveyed the relevant impressions. Each assessor for an impression was the subject who had collected images according to the impression in the first phase of the experiment. For example, the subject who collected images corresponding to the impressions "lively," "fresh," "elegant," and "natural" in the first phase assessed drawings corresponding to these impressions created by another subject in the second phase. The manner of assessing drawings was as follows: the assessor assigned a decreasing order of preference to four drawings that conveyed each impression, of the eight drawings created for each impression. We then scored each subject who had created the drawings according to the evaluator's choice and the intended impression. We accorded points depending on the order of preference: four points when an image was accorded first choice, three points when it was chosen second, two points for third choice, and one point for fourth choice. Each subject who created the drawings had two scores: without/with recommendations.

Figure 13 shows the experimental result by assessing the impressions of the drawings. The vertical axis indicates the scores of the seven subjects who created the drawings according to the two methods (without/with recommendations).

The most important feature of the result is the difference in effect by the recommendations between the subjects who originally had high scores and those who originally had low scores. Our system notably supported subjects not having adequate drawing ability: subject4, subject5, subject6, and subject7. This means that the recommendations of our system conveyed design knowledge of the original images from one user to another. In contrast to results for these subjects, our system to some extent reduced the scores of subjects with high powers of representation: subject1, subject2, and subject3. This implies that our system is not universally sensitive to artistic capacity in supporting the creation of graphic designs.

In order to determine the effects of our system on users possessing high drawing ability, we selected the drawings by subject1, subject2, and subject3, and asked five subjects to compare the designability and novelty of each two groups of the drawings. Figure 14 shows examples of drawings created by subject1, and Figure 15 shows the results of this assessment. The vertical axis of Figure 15 indicates the number of times the drawings without/with recommendations were assessed more highly than the others.

Although assessing designability and novelty is exceedingly difficult, the results show that our system helps users render graphic design works more flexible and designable using existing design knowledge. The novelty and universality of impressions are contrary properties. Thus, the results of this experiment imply the hybrid nature of our system because it simultaneously helped subjects 1, 2, and 3, who had good drawing skills, to draw higher-quality images, as well as subjects 4, 5, and 6, who were novices, to draw images of acceptable or adequate quality.

VII. Conclusion and Future Works

In this study, we proposed a framework to share and reuse graphic design knowledge on the Web. We implemented the system using a novel Web-based graphic design system that utilizes data-mining techniques to abstract design features from existing image data. A unique feature of this system is a Web-based user interaction model for drawing graphics, where a user's operations are enhanced by extracting and reusing color schemes in a dynamic manner. We performed experiments to examine the effectiveness of our system by publishing it as a modern HTML5 application on a cloud infrastructure. Experimental result shows that our system can support users by conveying graphic design knowledge and rendering graphic design works more flexible. In future research, we plan to improve the design abstraction method and the color retrieval function in order to provide more effective recommendations according to the needs of the user.

REFERENCES

[1] K. Sasaki, Invitation to Aesthetics, Tokyo: Chuokoron-Shinsha, 2004.

[2] M. Tokumaru, and K. Yamashita, "Color Coordinate Evaluating System Using Fuzzy Reasoning," Trans. IEICE, Japan, vol. J83-D2, no. 2, Feb. 2000, pp. 680–689, ISSN: 0915-1923.

[3] M. Tokumaru, N. Muranaka, and S. Imanishi, "Color Design Support System Considering Color Harmony," FUZZ-IEEE'02, May. 2002, pp. 378–383, doi: 10.1109/FUZZ.2002.1005020.

[4] M. Tokumaru, and N. Muranaka, "An Evolutionary Fuzzy Color Emotion Model for Coloring Support System," FUZZ-IEEE'08, Jun. 2008, pp. 408–413, doi: 10.1109/FUZZY.2008.4630400.

[5] A. Jahanian, et al. "Recommendation System for Automatic Design of Magazine Covers," 18th International Conference on Intelligent User Interfaces (IUI'13), Mar. 2013, pp. 95–106, doi: 10.1145/2449396.2449411.

[6] S. Kobayashi, Color Image Scale, Tokyo: Kodansha Intern, 1992.

[7] Adobe Systems Incorporated. *Adobe Color*. [Online]. Available at: https://color.adobe.com/ 2015.01.19.

[8] S. Lin, and P. Hanrahan, "Modeling How People Extract Color Themes from Images," ACM Human Factors in Computing Systems (CHI 2013), Apr. 2013, pp. 3101–3110, doi: 10.1145/2470654.2466424.

[9] J. Delon, A. Desolneux, J. L. Lisani, and A. B. Petro, "AUTOMATIC COLOR PALETTE," Inverse Problems and Imaging (IPI), vol. 1, no. 2, May. 2007, pp. 265–287, doi:10.3934/ipi.2007.1.265.

[10] P. Obrador, "Automatic color scheme picker for document templates based on image analysis and dual problem," SPIE's International Symposium Medical Imaging 2006 (SPIE 2006), vol. 6076, Feb. 2006, pp. 64–73, doi: 10.1117/12.647075.

[11] T. Kagawa, H. Nishino, and K. Utsumiya, "A color design assistant based on user's sensitivity," IEEE International Conference on Systems, Man & Cybernetics (SMC 2003), Oct. 2003, pp. 974–979, doi: 10.1109/ICSMC.2003.1243941.

[12] Y. Chang, S. Saito, K. Uchikawa, and M. Nakajima, "Example-Based Color Stylization of Images," ACM TAP, vol. 2, pp. 322–345, July. 2005, doi: 10.1145/1077399.1077408.

[13] B. Wang, Y. Yu, T. Wong, C. Chen, and Y. Xu, "Data-Driven Image Color Theme Enhancement," SIGGRAPH ASIA '10, Dec. 2010, no. 146, doi: 10.1145/1866158.1866172.

[14] Dribbble LLC. *Dribbble*. [Online]. Available at: https://dribbble.com 2015.01.19.

[15] Adobe Systems Incorporated. *Behance*. [Online]. Available at: https://www.behance.net 2015.01.19.

[16] pixiv Inc. *pixiv*. [Online]. Available at: http://pixiv.net/ 2015.01.19.

[17] M. R. Luo, G. Cui, and B. Rigg, "The development of the CIE 2000 colour-difference formula: CIEDE2000," Color Research & Application, vol. 26, issue. 5, Oct. 2001, pp. 340–350, doi: 10.1002/col.1049.

Active Learning to Rank Method for Documents Retrieval

Faïza Dammak, Imen Gabsi, Hager Kammoun, Abdelmajid Ben Hamadou
MIRACL Multimedia, InfoRmation systems and Advanced Computing Laboratory,
Technology Center of Sfax, Tunis Road Km 10, B.P. 242 Sfax 3021.
Sfax, Tunisia
e-mail: faiza.dammak@gmail.com e-mail: imenmri@gmail.com e-mail: hager.kammoun@isd.rnu.tn e-mail: abdelmajid.benhamadou@isimsf.rnu.tn

***Abstract*—This paper presents a new active learning to rank algorithm based on boosting for active ranking functions. The main goal of this algorithm is to introduce unlabeled data in the learning process. Since this type of ranking is based on a phase of selection of the most informative examples to label, the proposed algorithm allows the cost of labeling to be reduced. In a first step, the algorithm proposed is going to select at each iteration the most informative query-document pair from unlabeled data using the "Query by Committee" strategy. It is this pair which maximizes the measure of disagreement between a representative committee model chosen randomly and the model generated by the supervised algorithm. In fact, the randomly chosen model is generated from the main labeled set. While the other model is generated from the labeled set which changes in each iteration, by using a supervised ranking algorithm. For the latter, we choose to use three algorithms of *boosting*: RankBoost belonging to the family of *the pairwise approach*; AdaRank and LambdaMART belonging to the family of the *listwise approach.* Our choise is meant to subsequently compare the performance of pairwise and listwise approaches. In a second step, once this pair is selected, it will be added to the labeled set. To evaluate the performance of the active model proposed, we hav carried out an experimental study using the benchmark Letor 4.0 dataset. The obtained results show that the active model has a significant improvement in Normalized Discounted Cumulative Gain and Mean Average Precision.**

Keywords-active learning, learning to rank, boosting ranking algorithms

I. INTRODUCTION

In front of the constant increase in the volume of information available electronically, a new field of research, dedicated to automatically optimize the ranking of results returned by systems and based on machine learning techniques, has emerged. This area of research, called learning to rank, has led to the development of many approaches and algorithms. By combining a number of existing ranking models within a single function, these approaches and algorithms have improved the quality of results lists [2].

There are three groups of learning to rank algorithms: pointwise, pairwise and listwise approaches [3]. The pointwise and pairwise approaches respectively transform ranking into (ordinal) regression or classification on single object and pairs object such as RankBoost [8]. The listwise approach [4] treats ranking lists of objects (e.g., ranking lists of documents in IR) as instances in learning, such as AdaRank [5] and LambdaMART [9], in which the group structure is considered. In our study, we focus attention on pairwise and listwise approaches: the two most successful approaches for learning to rank in IR [2].

In learning to rank, the performance of a ranking model is strongly affected by the number of labeled examples in the training set [2]. However; obtaining such information relies on human experts and hence is in general very expensive in time and in resources. Thus, we need to introduce the unlabeled data, which helps by reducing the version space size, in the training set [6].

In this article, we are interested first of all, in the problem of the reduction of the training cost of the labeled base by introducing a large unlabeled learning set as input.

We proposed an active learning to the rank algorithm which introduces a labeling process with Query-by-Committee (QBC) active learning strategy [7]. The latter has less computation than others strategies. In this method, the learner constructs a committee of classifiers based on the current training set. Each committee member then classifies the query/document pair and the learner measures the degree of disagreement among the committee members.

Nevertheless, this model used a supervised ranking algorithm to learn a ranking function. For this, we proposed three boosting algorithms using pairwise and listwise approaches: RankBoost [8] has the characteristic to consider a pair of documents as entry. While both AdaRank [5] and LambdaMART [9] use the listwise approach. This approach tries to directly optimize the value of one of the above evaluation measures, averaged over all queries in the training data. Thus, we are interested secondly in comparing the performance of pairwise and listwise approaches during the use of active learning to the rank algorithm. The applications concerned are related to the Documents

Retrieval (DR). Indeed, ranking of documents is a popular research area in IR and Web community.

The rest of the paper is organized as follows. Section 2 introduces related work. Section 3 discusses the basic principles of the proposed approach. Section 4 presents the various experiments conducted to adopt the most efficient active learning to the rank model. Section 5 presents the conclusion and perspectives of this work.

II. RELATED WORK

A. *Learning to Rank in DR*

The main idea of learning to rank is to learn ranking functions that achieve good ranking objectives on test data. Learning to rank can be used in large variety of applications in IR. Among the typical one, we cite the DR which we take as an example in this paper. Considering a set of data compounds of query-document pairs with known relevance, the learning to rank methods learn automatically from these data the best way to combine models for optimal results list [10]. By giving a query, the ranking function attributes a score to each pair query-document. Then, this function ranks the documents in descending order of these scores. The ranking order represents the relevance of documents according to the query. This type of ranking is known as ranking of alternatives [1]. It is based on a supervised learning. However, such learning methods require a large labeled data for training. The creation of this data is generally very costly in time and resources and requires efforts from the user because it requires the intervention of a human expert. So, it is advantageous to introduce unlabeled data into the training base. The semi-supervised and active learning makes it possible to solve this problem but with different perspectives [11]. These two types of learning used a small set of labeled data and a large set of unlabeled data. By assembling both types of data, called partially labeled data, the need for labeled examples can be reduced. In the following, we present the active learning to rank approaches.

B. *Active Learning to Rank*

In order to get better performances and unlike the semi-supervised learning [11] which uses the unlabeled data in addition with the labeled ones, active learning puts limited human resources on labeling the most informative examples among the unlabeled ones to label [12]. This type of active learning is known as selective sampling [12] and it becomes central to many areas of applications including ranking of alternatives. On the one hand, active learning consists in learning a ranking function from a training set built during the learning and this is done by interaction with an expert. The quality of the ranking function is highly correlated with the amount of partially labeled data used to train the function. On the other hand, it proposes to the user optimal selection strategies in order to build the training set of the model [13]. The typical one is the query-by-committee (QBC) algorithm [34] which is formed by two steps. The first consists in building a committee formed by a set of *diverse* hypotheses trained on currently labeled data. The second aims to select the optimal queries by measuring their informativeness and by calculating the disagreement among the committee members on their ranking [14] [15].

Although learning to rank has been widely studied, there are not a lot of works referring to active learning to rank [16]. Donmez and Carbonell [14] presented an active learning approach to ranking problem in the context of DR, which is in principle extensible to any other partially (or totally) ordered ranking task. The novelty of their approach lies in relying on expected loss minimization for rank learning via the use of a normalized ranking loss estimation. Long et al [17] integrate both query and document selection into active learning to rank, and propose a two-stage optimization that minimizes the expected DCG loss. Truong [18] proposed an active learning method suggested within the framework of the ranking of alternatives for the task of the text summarization. He proposed several strategies to select instances to label. Experiments have shown that they allowed to effectively forming the basis of learning by selecting the most informative instances.

III. PROPOSED APPROACH

As reported in [18], two declensions of the active ranking are cited. The first consists in selecting an entry and labeling all related alternatives. It is suitable for example for automatic summarization. The second declension seeks to select only one entry-alternative pair (query-document). The user specifies if the alternative is relevant or not in relation to this entry. This declension is particularly well adapted to applications such as the IR. In his approach, Truong [18] uses the first declension. We choose to use the second since we are interested in the field of IR, where document (alternative) and query (entry) are the components in our proposed algorithm. This algorithm uses the effective strategy to selective sampling QBC [6]. This strategy selects the element which puts in conflict most of the members of all models called committee. In our context, the most informative entry-alternative pair is the one which makes a maximum of disagreement between the committee model and the model induced on the set of alternatives by a supervised ranking algorithm. The effectiveness of this method depends on the construction of the committee which must be varied enough and representative of space of entry as well as the choice of the measure of disagreement.

A. *Notation*

Given a set of entry X and a set of alternatives A, we assume that each query x is associated with a subset of known alternatives $A_x \subset A$. We consider a training labeled set $S_L=\{(x_i , y_i); i \in \{1,..,m\}\}$ with x_i an input and y_i a set of labels associated with A_x. In addition, we consider another great set of inputs unlabeled $S_U = \{(x_i^{'}); i \in \{m+1,..,m+n\}\}$. The algorithm proposed begins initially with S_L, S_U, a supervised ranking algorithm, K the number of partitions of all labeled data and Nb the desired number of examples to be labeled.

B. Active learning to rank algorithm

On the one hand, the active learning to rank algorithm (Figure 1) consists in building a committee formed by a set of diverse hypotheses trained on currently labeled data. In fact, we firstly subdivided the labeled set of training in K partitions and then associated for each partition a model. Hence; each model generates a score function h_k^{cv} as well as a score file. On the other hand, the proposed algorithm learns a model h from the labeled set. This model, changes in each iteration with the addition of a new labeled pair, by using a supervised ranking algorithm. Then, the algorithm will randomly choose a model among the K models learned at each iteration. Thereafter, it will select the most informative query-alternative pair from unlabeled data with the "Query by Committee" strategy. It is this very pair that maximizes the measure of disagreement between a representative committee model h_k^{cv} chosen randomly and the model h generated by the supervised algorithm (Figure 2). This measure is defined as follows :

$$d_c(h,\ h_k^{cv}\ ,\ x)\ \overset{def}{=}\ \max_{l\in L_x}\ \{(c(h,\ x,\ l) - c(\ h_k^{cv}\ ,\ x,\ l)\} \qquad (1)$$

where $x \in X$, L_x is the set of possible labels on alternative A_x, c is a cost function, h_k^{cv} and h two score functions. The current model asks the user to label the selected pair.
Lastly, this algorithm withdraws the selected pair from S_U and adds it in S_L until reaching the desired number of labeled data. As an output, it provides the required score function.

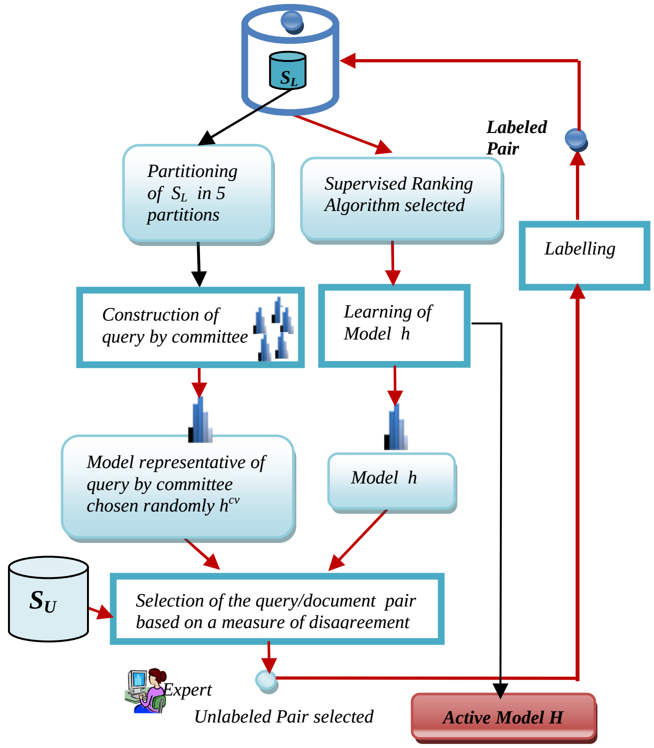

S_L
Labeled Pair
Partitioning of S_L in 5 partitions
Supervised Ranking Algorithm selected
Labelling
Construction of query by committee
Learning of Model h
Model representative of query by committee chosen randomly h^{cv}
Model h
S_U
Selection of the query/document pair based on a measure of disagreement
Expert
Unlabeled Pair selected
Active Model H

Figure 1. Approach proposed

For the supervised ranking algorithm, we have chosen three boosting algorithms: RankBoost [8], LambdaMART [9] and AdaRank [5]. RankBoost [8] is a powerful pairwise supervised learning algorithm that learns a real-valued (scoring) function, by optimizing a specific error measure suitable for ordering sets of objects. More precisely, at each round of boosting, the algorithm minimizes the weighted number of instances that are disordered.The pairs on which we have made mistakes (with respect to the weaker ranker chosen for that round) are given a higher importance weight for correct ordering in the next round. Thus, the goal of RankBoost is to produce an order, by a scoring function h_t for each document, which places as many relevant documents as possible at the top. LambdaMART [19] is a listwise method; it is the boosted tree version of LambdaRank [20]. It uses Gradient boosting [21] to optimize a ranking cost. It employs the MART (Multiple Additive Regression Trees) algorithm to learn a boosted regression tree as a ranking model. LambdaMART has been shown to be among the best performing learning methods based on evaluations on public data sets [22]. Readers can refer to [26] for details of this algorithm. AdaRank [5] is a listwise algorithm for learning ranking models in DR. It repeatedly constructs 'weak rankers' on the basis of re-weighted training data. Finally, it linearly combines the weak rankers for making ranking predictions. In contrast to the existing methods, AdaRank optimizes a loss function that is directly defined on the performance measures. It employs a boosting technique in ranking model learning.

In the following, we give the active ranking algorithm.

```
Algorithm. Active learning to rank algorithm of alternatives
Entry :
    - A small labeled data $S_L$={($x_i$ , $y_i$); $i\in${1,..,$m$}}
    - A large unlabeled data $S_U$ = {( $x_i^{'}$ );$i\in${$m$+1,..,$m$+$n$}}
    - A supervised ranking algorithm
    - $K$ : number of partitions of $S_L$
    - $Nb$ : number of labeled examples required
    - Learn models of the committee and obtain $h_k^{cv}$
    - $nbIter \leftarrow 0$
While  nbIter < = Nb  do
   - learn a ranking function $h$ with supervised algorithm on $S_L$
     - choose randomly $h^{cv}$
     - select the most informative query-document pair from $S_U$ which maximizes the measure of disagreement $d_c(h, h^{cv}, (x,d))$
     - ask the expert to label this pair
     - withdraw this pair of $S_U$ and add it in $S_L$
     - $nbIter \leftarrow nbIter + 1$
  End
Output : Active Model H
```

Figure 2. Active learning to rank algorithm of alternatives

The total cost is dominated by the step of selecting the input. In our case, it is the calculation of the disagreement measure, which requires taking into account all the possible values of labels for a given input. As well, QBC strategy is easy to implement. It has a low complexity. The algorithm requires training K+1models. The cost of learning is therefore multiplied by *K+1*.

IV. EXPERIMENTAL STUDY

We conducted a number of experiments in order to evaluate the importance of unlabeled data to learn an efficient ranking function. Once the ranking function is learned in the training phase, it will be used to order unlabeled examples from the test data. This training phase can be followed by a validation phase.

A. *Experimental tools*

Evaluating the quality of ranking functions is a core task in DR and other IR domains. For the realization of the algorithm (Figure 2), we propose to extend the library of learning to rank algorithms RankLib [23]. Currently, this library contains the implementation of eight ranking algorithms. It also provides the implementation of the evaluation measures based on the performance measures used in IR tools.

For the realization of the stage of selection of the query-document pair, we initially propose to evaluate the disagreement between a representative of committee model h_k^{cv} and the model h in order to select the unlabeled pairs. Then, we choose to use as measures of disagreement the Euclidean distance between the score given by model h_k^{cv} and the score obtained by the model h deduced from the supervised algorithm. It is the selected pair which has the maximum distance. By varying the number desired of data to label and by calculating the variation of the evaluation measures, we can deduce the suitable supervised boosting algorithm from the three chosen: RankBoost, LambdaMART, and AdaRank.

1) Data collections

Since the performance of a model depends on the quality of data used in the learning phase, we use the standard benchmark LETOR (*LEarning TO Rank*) [1] which constitutes a baseline in IR and evaluation measures [24]. We use specially the MQ2008-semi (*Million Query track*) collections in LETOR 4.0 as it contains both labeled and unlabeled data. There are about 2000 queries in this dataset. On average, each query is associated with about 40 labeled documents and about 1000 unlabeled documents.

MQ2008-semi [25] is conducted on the .GOV2 corpus using the TREC 2008, which is crawled from Web sites in the .gov domain. There are 25 million documents contained in the .GOV2 corpus, including HTML documents, as well as the extracted text of PDF and Word and postscript files [25].

Each subset of the collection MQ2008-semi is partitioned into five divisions, denoted as S1, S2, S3, S4, and S5, in order to conduct a five-fold cross validation. The results reported in this section are the average results over multiple folds. For each fold, three parts are used: The training part is used to learn the ranking model. The validation part is used to tune the parameters of the ranking model, like the number of iterations in RankBoost. The test part is used to report the ranking performance of the model.

Also, in this semi-MQ2008 collection, each training file contains a small number of pairs of labeled data and a large number of pairs of unlabeled data. We choose to extract pairs of labeled data in a first file for the training phase and the unlabeled pairs in a second file for the testing phase. In addition, the unlabeled pairs of data will be selected and labeled by the proposed algorithm in learning and added to the labeled file extracted.

2) Evaluation Measures

For the evaluation of the algorithms proposed (RankBoost_Active, LambdaMART_Active, AdaRank_ NDCG_Active and AdaRank_MAP_Active), we use a set of standard ranking measures such as Precision at position n, Mean Average Precision (MAP) and Normalized Discounted Cumulative Gain (NDCG) [24]. P@n measures the accuracy within the top n results of the returned ranked list for a query:

$$P@n = \frac{\#\mathit{relevant\ docs\ in\ top\ n\ results}}{\mathrm{n}}. \tag{2}$$

MAP takes the mean of the average precision values over all relevant documents:

$$MAP = \frac{\sum_{n=1}^{N} (P@n * rel(n))}{\#\mathit{total\ relevants\ \ docs\ \ for\ this\ query}}. \tag{3}$$

NDCG@k is widely used to handle multiple levels of relevance (whereas Precision and MAP are designed for binary relevance levels). The value of the NDCG to a position k of ordered list is calculated as follows:

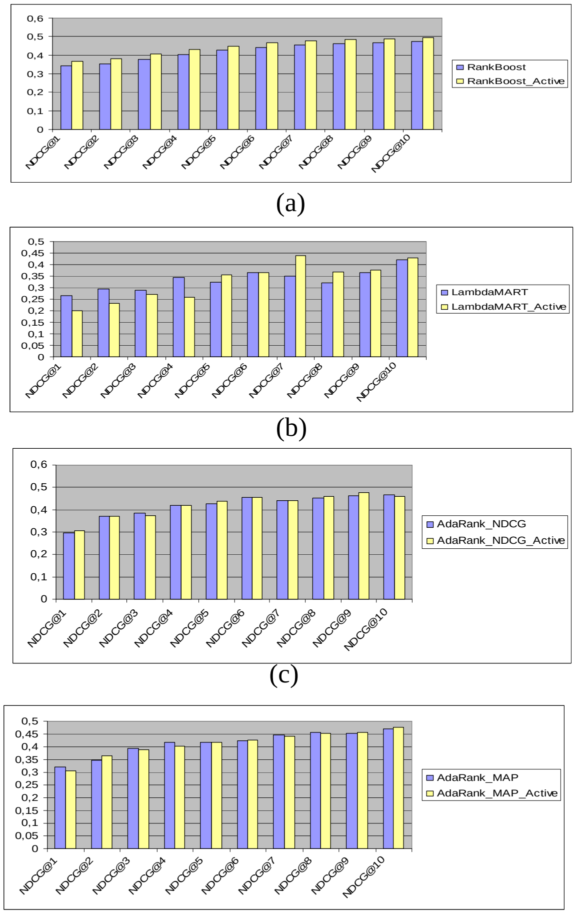
$$NDCG@k = \frac{1}{n}\sum_{k=1}^{n}\frac{1}{z_{K\ n}}\sum_{j=1}^{m_k}\frac{2^{r(j)}-1}{\log(1+j^k)} \tag{4}$$

3) Experimental results

These experimental results test how unlabeled data affect the ranking performance of the proposed algorithms.

RankBoost, LambdaMART, AdaRank_MAP and AdaRank_NDCG were selected as baselines in the experiments. For the proposed algorithms, the number of iterations was determined automatically during each experiment. Specifically, when there is no improvement in ranking accuracy in terms of the performance measure, the iteration stops. For both RankBoost_Active and LambdaMART_Active, we train the ranker for 500 rounds. For the others (AdaRank_NDCG_Active and AdaRank_ MAP_Active), the number of iterations was stoped at 200 rounds.

Then, we calculated the variation of NDCG for the three algorithms according to the number desired of data to label (Figure 3: (a), (b), (c) and (d)). Each group of bars corresponds to one NDCG@n. As shown in this figure and Table I, NDCG@n measures are better in RankBoost _Active algorithm, quite better in AdaRank_NDCG_Active and AdaRank_MAP_Active. But, they are variable in LambdaMART_Active algorithm.



(a)

(b)

(c)

(d)

Figure 3. Performance of RankBoost_Active (a), LamddaMART_ Active (b), AdaRank_NDCG_Active (c) and AdaRank_MAP_ Active (d) on training set : NDCG@n measures on the MQ2008-semi collection

We noticed also that, in the training and testing set, the pairwise RankBoost_Active has NDCG values slightly higher than AdaRank_NDCG_Active, AdaRank_MAP_ Active and LambdaMART_Active algorithms, which belong to listwise approach, (Figure 5). Although, compared with the other two types of approaches (pointwise and pairwise), the listwise approaches express the real sense of the learning to rank. The experimental studies have shown that the pairwise approach is better for the active algorithm proposed. Indeed, pairwise ranking methods have shown their performances by balancing the distribution of document pairs across queries [2]. These results illustrate how the unlabeled data affect the performance of ranking in the proposed algorithm. We notice a slight improvement by using the criterion NDCG@n for the fourth active algorithms.

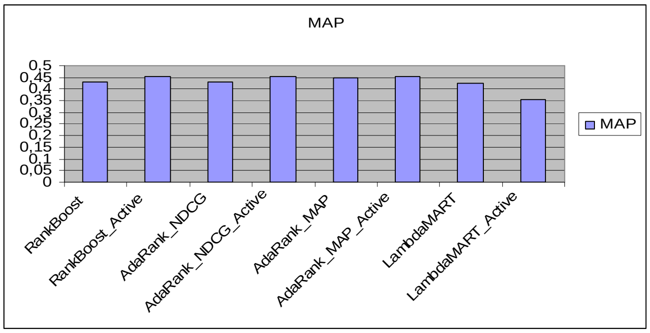


Figure 4. Performance on test set : MAP measures on the MQ2008-semi collection

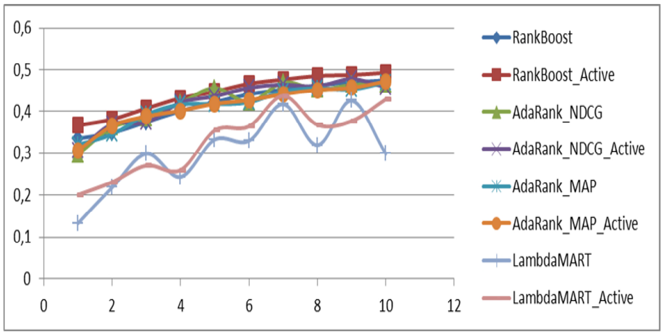


Figure 5. Performance on testing set : NDCG@n measures on the MQ2008-semi collection

TABLE 1. EVALUATION RESULTS IN TERMS OF NDCG@N ON MQ2008-SEMI DATA SET ON TRAINING SET

	NDCG@1	NDCG@2	NDCG@3	NDCG@5	NDCG@7	NDCG@8	NDCG@9	NDCG@10
RankBoost	0,3419	0,3543	0,3772	0,4267	0,4543	0,4640	0,4675	0,4731
RankBoost_Active	**0,3672**	**0,3817**	**0,4075**	**0,4474**	**0,4764**	**0,4852**	**0,4874**	**0,4930**
AdaRank_NDCG	0,2970	0,3704	0,3850	0,4256	0,4423	0,4522	0,4637	0,4655
AdaRank_NDCG_Active	**0,3056**	**0,3704**	**0,3745**	**0,4365**	**0,4422**	**0,4601**	**0,4777**	**0,4581**
AdaRank_MAP	0,3205	0,3458	0,3929	0,4174	0,4480	0,4561	0,4528	0,4702
AdaRank_MAP_Active	**0,3065**	**0,3660**	**0,3886**	**0,4183**	**0,4416**	**0,4516**	**0,4566**	**0,4755**
LambdaMART	0,2660	0,2951	0,2875	0,3221	0,3498	0,3194	0,3651	0,4208
LambdaMART_Active	**0,2009**	**0,2314**	**0,2710**	**0,3552**	**0,4374**	**0,3688**	**0,3776**	**0,4293**

Figure 4 demonstrates the results of MAP measures on the MQ2008-semi collection. The results of this figure show that RankBoost_Active, AdaRank_NDCG_Active and AdaRank_MAP_Active algorithms have an average precision (MAP) better than that found by RankBoost and AdaRank_NDCG and AdaRank_MAP (Figure 4).

These results prove the interest of integrating unlabeled data in ranking functions with active learning.

V. CONCLUSION

In this article, we have proposed an active learning to rank algorithm based on a supervised ranking one. The contribution of this algorithm is presented in the use of a very small number of labeled examples and a large number of unlabeled data preselected incrementally by the Query By Committee method. This method has been shown to be effective in different classification tasks. For supervised ranking algorithm, we have chosen three boosting algorithm for different approaches the pairwise and listwise. The training and the test phases were carried out with the collections of the benchmark standard LETOR 4.0. Basing on the measures of evaluation NDCG and MAP, the preliminary results show that the active algorithm using pairwise approach provides better results.

The performance of such a model lies in its ability to use, for training, unlabeled data and QBC method. The latter allows minimizing the version space. However, its performance degrades when the number of labeled data and the learning time increase. To solve this problem, we suggest integrating a semi-supervised learning method to label the selected pair instead of the expert to reduce the learning time.

REFERENCES

[1] T.-Y. Liu, J. Xu, T. Qin, W.-Y. Xiong, and H. Li, “LETOR: Benchmark dataset for research on learning to rank for Information Retrieval”. Proceedings of the Learning to Rank workshop in the 30th annual international ACM SIGIR conference on Research and development information retrieval, 2007

[2] T. -Y. Liu. “Learning to rank for information retrieval”. Springer-Verlag Berlin Heidelberg, 2011.

[3] Z. Cao, T.Qin, T.-Y. Liu, Tsai, M.-F., and Li, H. “Learning to rank: from pairwise approach to listwise approach”. ICML ’07, 2007, pp. 129-136.

[4] F. Xia, T. Liu, J. Wang, W. Zhang, and H. Li, “Listwise approach to learning to rank: theory and algorithm”. In ICML ’08, New York, NY, USA, ACM 2008, pp. 1192-1199.

[5] J. Xu and H. Li, “AdaRank: a boosting algorithm for information retrieval”. In Proceedings of the 30th Annual International ACM SIGIR Conference (SIGIR’07), Amsterdam, 2007, pp. 391-398.

[6] B. Settles, “Active learning literature survey”, Computer Sciences Technical Report 1648, University of Wisconsin-Madison, 2009.

[7] S. Jerzy and P. Mateusz, “Comparing Performance of Committee Based Approaches to Active Learning”. Recent Advances in Intelligent Information Systems, 2009, pp. 457-470.

[8] Y. Freund, R. Iyer, R. E Schapire, and Y. Singer, “An efficient boosting algorithm for combining preferences”. Journal of Machine Learning Research, 2003, pp. 933-969.

[9] C. Burges. “From RankNet to LambdaRank to LambdaMART: An overview. Technical report”, Microsoft Research Technical Report MSR-TR-2010-82, 2010.

[10] O. Chapelle and Y. Chang. “Yahoo! Learning to Rank Challenge Overview”. Journal of Machine Learning Research - Proceedings Track, vol. 14, 2011, pp. 1-24.

[11] K. Duh and K. Kirchhoff, “Learning to rank with partially-labeled data”. In Myaeng, S.-H. Oard, D. W. Sebastiani, F. Chua, T.-S., and Leong, M.-K., editors, SIGIR, ACM. 2008, pp. 251-258.

[12] Y. Freund, H. S. Seung, E. Shamir, and N.i Tishby, “Selective sampling using the query by committee algorithm,” Machine Learning, vol. 28, 1997, pp. 133-168.

[13] N. Ailon, “An active learning algorithm for ranking from pairwise preferences with an almost optimal query complexity”. Journal of Machine Learning Research, 2012, pp.137-164.

[14] P. Donmez and J. G. Carbonell, “Active samplings for rank learning via optimizing the area under the roc curve”. ECIR, volume 5478 of Lecture Notes in Computer Science, Springer. 2009, pp. 78-89.

[15] W. Shen and H. Lin, “Active Sampling of Pairs and Points for Large-scale Linear Bipartite Ranking”. In Proceedings of ACML'13 (JMLR W&CP 29), 2013, pp. 388-403.

[16] B. Qian, H. Li, J. Wang, X. Wang, and I. Davidson. “Active Learning to Rank using Pairwise Supervision”. Proceedings of the 13th SIAM International Conference on Data Mining, 2013, pp. 297-305.

[17] B. Long, O. Chapelle, Y. Zhang, Y. Chang, Z. Zheng, and B. Tseng. “Active learning for ranking through expected loss optimization”. In Proceedings of the 33rd international ACM SIGIR’10New York,,USA, 2010, pp. 267-274.

[18] T.-V. Truong, “Learning Functions ranking with little Labeled Examples”, PhD thesis, University Pierre and Marie Curie – Paris VI, 2009.

[19] Q.Wu, C.J.C. Burges, K. Svore, and J. Gao, “Adapting Boosting for Information Retrieval Measures”. Journal of Information Retrieval, 2007.

[20] J. H. Friedman. “Greedy function approximation: A gradient boosting machine”. Annals of Statistics, 29: 2000, pp. 1189-1232.

[21] Y. Ganjisaffar, R. Caruana, and C.V. Lopes, “Bagging Gradient-Boosted Trees for High Precision, Low Variance Ranking Models”, SIGIR’11, Beijing, China. 2011.

[22] C. Sawade, S. Bickel, T. Oertzen, T. Scheer, and N. Landwehr. “Active Evaluation of Ranking Functions based on Graded Relevance”. ECML PKDD’12 - Volume II. Springer-Verlag Berlin, 2012, pp. 676-691.

[23] http://people.cs.umass.edu/~vdang/ranklib.html

[24] K. Jarvelin and J. Kekalainen. “IR evaluation methods for retrieving highly relevant documents”. Special Interest Group on Information Retrieval (SIGIR), 2000.

[25] http://research.microsoft.com/en-us/um/beijing/projects/letor//

[26] A. J. Aslam, E. Kanoulas, V. Pavlu, S. Savev, and E. Yilmaz. “Document selection methodologies for efficient and effective learning-to-rank”.In Proceedings of the 32nd international ACM, SIGIR ’09, New York, USA. 2009, pp. 468-475.

Online Client-Side Bottleneck Identification on HTTP Server Infrastructures

Ricardo Filipe, Serhiy Boychenko, Filipe Araujo

CISUC, Dept. of Informatics Engineering
University of Coimbra
Coimbra, Portugal
{rafilipe, serhiy}@dei.uc.pt, filipius@uc.pt

***Abstract*—Ensuring short response times is a major concern for all web site administrators. To keep these times under control, they usually resort to monitoring tools that collect a large spectrum of system metrics, such as CPU and memory occupation, network traffic, number of processes, etc. Despite providing a reasonably accurate picture of the server, the times that really matter are those experienced by the user. However, not surprisingly, system administrators will usually not have access to these end-to-end figures, due to their lack of control over web browsers. To overcome this problem, we follow the opposite approach of monitoring a site based on times collected from browsers. We use two browser-side metrics for this: i) the time it takes for the first byte of the response to reach the user (request time) and ii) the time it takes for the entire response to arrive (response time). We conjecture that an appropriate choice of the resources to control, more precisely, one or two URLs, suffices to detect CPU, network and I/O bottlenecks. In support of this conjecture, we run periodical evaluations of request and response times on some very popular web sites to detect bottlenecks. Our experiments suggest that collecting data from the browsers can indeed contribute for better monitoring tools that provide a deeper understanding of the system, thus helping to maintain faster, more interactive web sites.**

Keywords–Cloud computing; Bottleneck; Virtualization.

I. Introduction

In the operation of a Hypertext Transfer Protocol (HTTP) server [1], bottlenecks may emerge at different points of the system often with negative consequences for the quality of interaction with users. To control this problem, system administrators must keep a watchful eye on a large range of system parameters, like CPU, disk and memory occupation, network interface utilization, among an endless number of other metrics, some of them specifically related to HTTP, such as response times or sizes of waiting queues. Despite being very powerful, these mechanisms cannot provide a completely accurate picture of the HTTP protocol performance. Indeed, the network latency and transfer times can only be seen from the client, not to mention that some server metrics might not translate easily to the quality of the interaction with users. Moreover, increasing the number of metrics involved in monitoring adds complexity to the system and makes monitoring more intrusive.

We hypothesize that a simpler mechanism, based on client-side monitoring, can fulfill the task of detecting and identifying an HTTP server bottleneck from a list of three: CPU, network, or disk input/output (simply I/O hereafter). The arguments in favor of this idea are quite powerful: client-side monitoring provides the most relevant performance numbers, while, at the same time, requiring no modifications to the server, which, additionally, can run on any technology. This approach can provide a very effective option to complement available monitoring tools.

To achieve this goal, we require two metrics taken from the web browser: i) the time it takes from requesting an object to receiving the first byte (request time), and ii) the time it takes from the first byte of the response, to the last byte of data (response time). We need to collect time series of these metrics for, at least, one or two carefully chosen URLs. These URLs should be selected according to the resources they use, either I/O or CPU. The main idea is that each kind of bottleneck exposes itself with a different signature in the request and response time series.

To try our conjecture, and create such time series, we resorted to experiments on real web sites, by automatically requesting one or two URLs with a browser every minute, and collecting the correspondent request and response times. With these experiments, we managed to discover a case of network bottleneck and another one of I/O bottleneck. We believe that this simple mechanism can improve the web browsing experience, by providing web site developers with qualitative results that add to the purely quantitative metrics they already own.

The rest of the paper is organized as follows. Section II presents the related work in this field and provides a comparison of different methods. Section III describes the online method to detect and identify the HTTP server bottlenecks. In Section IV we try a specific approach to show monitoring results from popular web sites, thus exposing different types of bottlenecks. Finally, in Section V we discuss the results and conclude the paper.

II. Related Work

In the literature, we can find a large body of work focused on timely scaling resources up or down, usually in N-tier HTTP server systems, [2–7]. We divide these efforts into three main categories: (i) analytic models that collect multiple metrics to ensure detection or prediction of bottlenecks; (ii) rule-based approaches, which change resources depending on utilization thresholds, like network or CPU; (iii) system configuration analysis to ensure correct functionality against bottlenecks and peak period operations.

First, regarding analytic models, authors usually resort to queues and respective theories to represent N-tier systems [8][9]. Malkowski *et al.* [10] try to satisfy service level objectives (SLOs), by keeping low service response times.

ISBN: 978-1-61208-412-1

They collect a large number of system metrics, like CPU and memory utilization, cache, pool sizes and so on, to correlate these metrics with system performance. This should expose the metrics responsible for bottlenecks. However, the analytic model uses more than two hundred application and system level metrics. In [11], Malkowski *et al.* studied bottlenecks in N-tier systems even further, to expose the phenomenon of multi-bottlenecks, which are not due to a single resource that reaches saturation. Furthermore, they managed to show that lightly loaded resources may be responsible for such multi-bottlenecks. As in their previous work, the framework resorts to system metrics that require full access to the infrastructure. The number of system metrics to collect is not clear. Wang *et al.* continued this line of reasoning in [7], to detect transient bottlenecks with durations as low as 50 milliseconds. The transient anomalies are detected recurring to depth analysis of metrics in each component of the system. Although functional, this approach is so fine-grained that it is closely tied to a specific hardware and software architecture.

In [2], authors try to discover bottlenecks in data flow programs running in the cloud. In [6], Bodík *et al.* try to predict bottlenecks to provide automatic elasticity. [5] presents a dynamic allocation of VMs based on SLA restrictions. The framework consists of a continuous "loop" that monitors the cloud system, to detect and predict component saturation. The paper does not address questions related to resource consumption of the monitoring approach or scalability to large cloud providers. Unlike other approaches that try to detect bottlenecks, [12] uses heuristic models to achieve optimal resource management. Authors use a database rule set that, for a given workload, returns the optimal configuration of the system. [13] presents a technique to analyze workloads using k-means clustering. This approach also uses a queuing model to predict the server capacity for a given workload for each tier of the system.

Other researchers have focused on rule-based schemes to control resource utilization. Iqbal *et al.* [3][14] propose an algorithm that processes proxy logs and, at a second layer, all CPU metrics of web servers. The goal is to increase or decrease the number of instances of the saturated component. [15] also scales up or down servers based on CPU and network metrics of the server components. If a component resource saturation is observed, then, the user will be migrated to a new virtual machine through IP dynamic configuration. This approach uses simpler criteria to scale up or down compared to bottleneck-based approaches, because it uses static performance-based rules.

Table I illustrates the kind of resource problem detected by the mentioned papers. The second column concerns the need to increase CPU resources or VM instances. The third column is associated to I/O, normally an access to a database. The network column represents delays inside the server network or to the client — normally browser or web services. It is relevant to mention that several articles [16][2][11] only consider CPU (or instantiated VM) and I/O bottleneck, thus not considering internal (between the several components) or external (client-server connection) bandwidth.

Finally, some techniques scan the system looking for mis-configurations that may cause inconsistencies or performance issues. Attariyan *et al.* [17] elaborated a tool that scans the system in real time to discover root cause errors in configuration. In [18], authors use previous correct configurations to eliminate unwanted or mistaken operator configuration.

TABLE I. BOTTLENECK DETECTION IN RELATED WORK.

Article	CPU/Threads/VM	I/O	Network
[2]	X	X	
[3]	X		
[10]	X	X	
[4]	X		
[7]	X	X	Internal
[11]	X	X	Internal
[15]	X		X

Our work is different from the previously mentioned literature in at least two aspects: we are not tied to any specific architecture and we try to evaluate the bottlenecks from the client's perspective. This point of view provides a better insight on the quality of the response, offering a much more accurate picture regarding the quality of the service. While our method could replace some server-side mechanisms, we believe that it serves better as a complementary mechanism.

It is also worth mentioning client-side tools like HTTPerf [19] or JMeter [20], which serve to test HTTP servers, frequently under stress, by running a large number of simultaneous invocations of a service. However, these tools work better for benchmarking a site before it goes online.

III. A CONJECTURE ON CLIENT-SIDE MONITORING OF HTTP SERVERS

We now evaluate the possibility of detecting bottlenecks based on the download times of web pages, as seen by a client. We conjecture that we can, not only, detect the presence of a bottleneck, something that would be relatively simple to do, but actually determine the kind of resource causing the bottleneck, CPU, I/O or network. CPU limitations may be due to thread pool constraints of the HTTP Server (specially the front-end machines), or CPU machine exhaustion, e.g., due to bad code design that causes unnecessary processing. I/O bottlenecks will probably be related to the database (DB) operation, which clearly depend on query complexity, DB configuration and DB access patterns. Network bottlenecks are related to network congestion.

To illustrate this possibility, we propose to systematically collect timing information of one or two web pages from a given server, using the browser side JavaScript Navigation Timing API [21]. Figure 1 depicts the different metrics that are available to this JavaScript library, as defined by the World Wide Web (W3) Consortium. Of these, we will use the most relevant ones for network and server performance: the request time (computed as the time that goes from the request start to the response start) and the response time (which is the time that goes from the response start to the response end). We chose these, because the request and response times are directly related to the request *and* involve server actions, which is not the case of browser processing times, occurring afterwards, or TCP connection times, happening before.

Consider now the following decomposition of the times of interest for us:

- Request Time: client-to-server network transfer time + server processing time + server-to-client network latency.

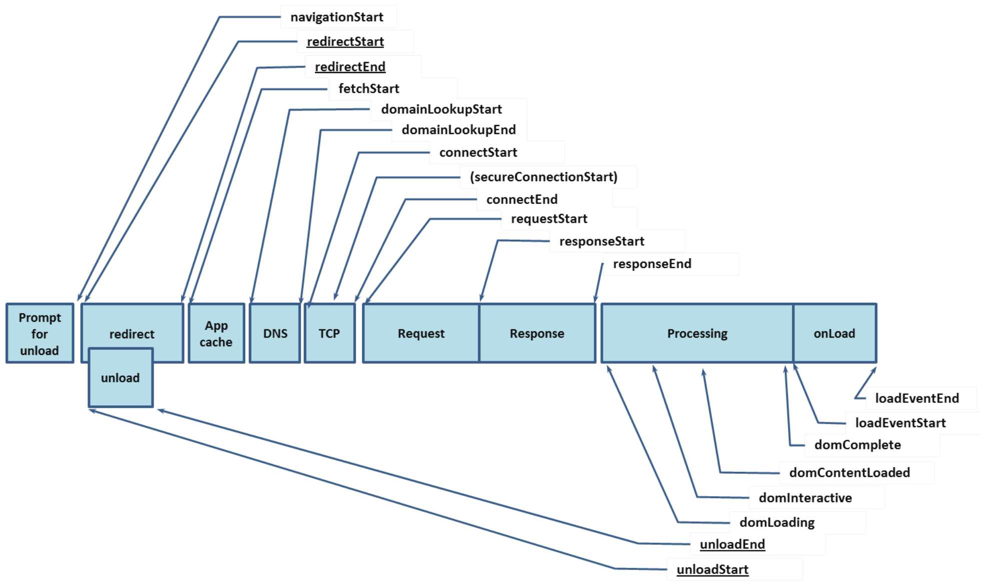


Figure 1. Navigation Timing metrics (figure from [21])

- Response Time: server-to-client network transfer time.

To make use of these times, we must assume that the server actions, once the server has the first byte of the response ready, do not delay the network transfer of the response. In practice, our analysis depends on the server not causing any delays due to CPU or (disk) I/O, once it starts responding. Note that this is compatible with chunked transfer encoding: the server might compress or sign the next chunk, while delivering the previous one.

We argue that identifying network bottlenecks, and their cause, with time series of these two metrics is actually possible, whenever congestion occurs in both directions of traffic. In this case, the request and response times will correlate strongly. If no network congestion exists, but the response is still slow, the correlation of request and response times will be small, as processing time on the server dominates. Small correlation points to a bottleneck in the server, whereas high correlation points toward the network. Hence, repeated requests to a single resource of the system, such as the entry page can help to identify network congestion, although we cannot tell exactly where in the network does this congestion occur. To this correlation-based evaluation of the request and response time series from a single URL, we call "single-page request" analysis.

Separating CPU from I/O bottlenecks is a much more difficult problem. We resort to a further assumption here: the CPU tasks share a single pool of resources, possibly with several (virtual) machines, while I/O is often partitioned. This, we believe, reflects the conditions of many large systems, as load balancers forward requests to a single pool of machines, whereas data requests may end up in separate DB tables, served by different machines, depending on the items requested. Since scarce CPU resources affect *all* requests, this type of bottleneck synchronizes all the delays (i.e., different parallel requests tend to be simultaneously slow or fast). Thus, logically, unsynchronized delays must point to I/O bottlenecks. On the other hand, one cannot immediately conclude anything, with respect to the type of bottleneck, if the delays are synchronized (requests might be suffering either from CPU or similar I/O limitations).

TABLE II. SOFTWARE USED AND DISTRIBUTION.

Component	Observations	Version
Selenium	selenium-server-standalone jar	2.43.0
Firefox	browser	23.0
Xvfb	xorg-server	1.13.3

The challenge is, therefore, to identify pairs of URLs showing unsynchronized delays, to pinpoint I/O bottlenecks. Ensuring that a request for an URL has I/O is usually simple, as most have. In a news site, fetching a specific news item will most likely access I/O. To have a request using only CPU or, at least, using some different I/O resource, one might fetch non-existing resources, preferably using a path outside the logic of the site. We call "independent requests" to this mechanism of using two URLs requesting different types of resources.

One should notice that responses must occupy more than a single TCP [22] segment. Otherwise, one cannot compute any meaningful correlation between request and response times, as this would always be very small.

We will now experimentally try the "single-page request" and the "independent requests" mechanisms, to observe whether they can actually spot bottlenecks in real web sites.

IV. EXPERIMENTAL EVALUATION

In this section we present the results of our experimental evaluation.

A. Experimental Setup

For the sake of doing an online analysis, we used a software testing framework for web applications, called Selenium [23]. The Selenium framework emulates clients accessing web pages using the Firefox browser, thus retaining access to the Javascript Navigation Timing API [21]. We use this API to read the request and response times necessary for the "single-page request" and "independent requests" mechanisms. We used a UNIX client machine, with a crontab process, to request a page each minute [24]. The scheduler launched the Selenium process (with the corresponding Firefox browser) each minute. We emulated a virtual display for the client machine using Xvfb [25]. Table II lists the software and versions used.

One of the criteria we used to choose the pages to monitor was their popularity. However, to conserve space, we only show results of pages that provided interesting results, thus omitting sites that displayed excellent performance during the entire course of the days we tested (e.g., CNN [26] or Amazon [27]) — these latter experiments would have little to show regarding bottlenecks. On the other hand, we could find some bottlenecks in a number of other real web sites:

- **Akamai/Facebook photo repository** — We kept downloading the same 46 KiloBytes (KiB) Facebook photo, which was actually delivered by the Akamai Content Delivery Network (CDN). During the time of this test, the CDN was retrieving the photo from Ireland. This experiment displays network performance problems.
- **SAPO [28]** — this webpage is the 5th most used portal in Portugal (only behind Google – domain .pt and .com, Facebook and Youtube) and the 1st page of Portuguese language in Portugal [29]. This web page shows considerable performance perturbations on the server side, especially during the wake up hours.
- **Record sports news [30]** — This is an online sports newspaper. We downloaded an old 129 KiB news item [31] and an inexistent one [32] for several days. The old news item certainly involves I/O, to retrieve the item from a DB, whereas the inexistent may or may not use I/O, we cannot tell for sure. We ensured a separation of 10 seconds between both requests. One should notice that having a resource URL involving only CPU would be a better choice to separate bottlenecks. However, since we could not find such resource, a non-existing one actually helped us to identify an I/O bottleneck.

B. Results

We start by analyzing the results of Facebook/Akamai and SAPO, in Figures 2, 3 and 4. These figures show the normal behavior of the systems and allow us to identify periods where response times fall out of the ordinary.

Figure 2 shows the response of the Akamai site for a lapse of several days. We can clearly observe a pattern in the response that is directly associated to the hour of the day. During working hours and evening in Europe, we observed a degradation in the request and response times (see, for example, the left area of the blue dashed line on September 19, 2014, a Friday). The green and the red lines (respectively, the response and the request times), clearly follow similar patterns, a sign that they are strongly correlated. Computing the *correlation coefficient* of these variables, $r(Req, Res)$, for the left side of the dashed line we have $r(Req, Res) = 0.89881$, this showing that the correlation exists indeed. However, for the period where the platform is more "stable" (between the first peak periods) we have $r(Req, Res) = -0.06728$. In normal conditions the correlation between these two parameters is low. This allows us to conclude that in the former (peak) period we found a network bottleneck that does not exist in the latter. However, our method cannot determine where in the network is the bottleneck. Interestingly, in Figure 3, we can observe that the bottleneck disappeared after a few days. On September 29^{th}, we can no longer see any sign of it.

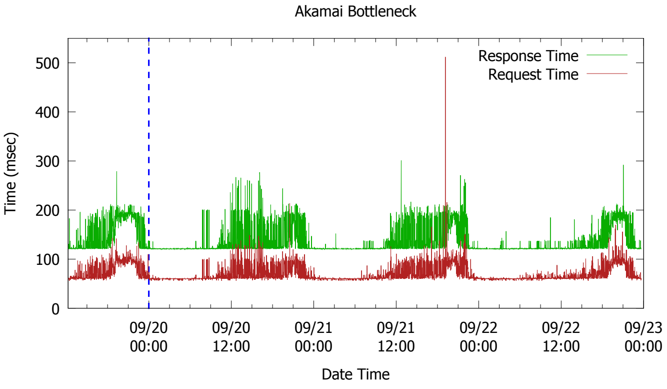


Figure 2. Akamai/Facebook bottleneck.

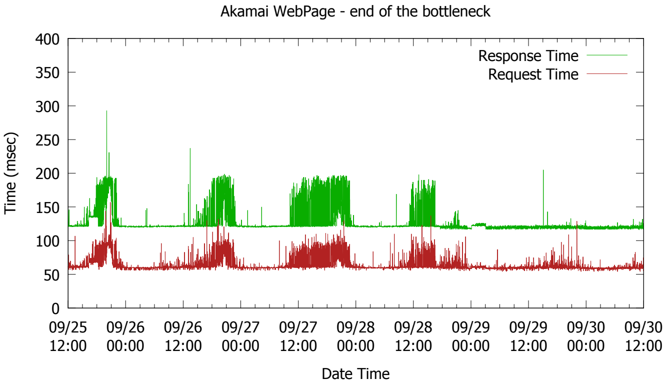


Figure 3. Akamai/Facebook - end of the bottleneck.

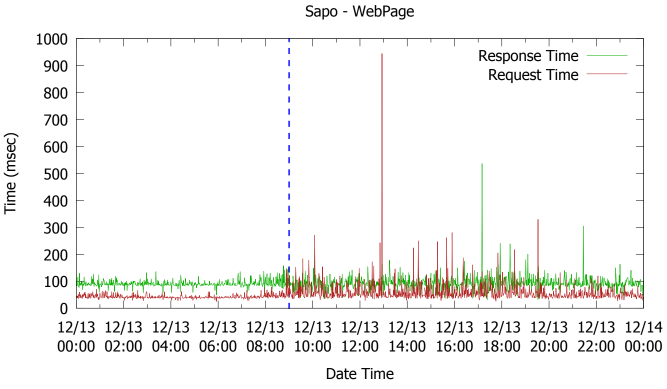


Figure 4. SAPO bottleneck.

Regarding Figure 4, which shows request and response times of the main SAPO page, we can make the same analysis for two distinct periods: before and after 9 AM (consider the blue dashed line) of December 13, 2013 (also a Friday).

Visually, we can easily see the different profiles of the two areas. The correlation for these two areas are:

- $r(Req, Res)_{before9AM} = 0.36621$
- $r(Req, Res)_{after9AM} = 0.08887$

The correlation is low, especially during the peak period, where the response time is more irregular. This case is therefore quite different from the previous one, and suggests that no network bottleneck exists in the system, during periods of intense usage. With the "single-page request" method only, and without having any further data of the site, it is difficult to precisely determine the source of the bottleneck (CPU or I/O).

To separate the CPU from the I/O bottleneck, we need to resort to the "independent requests" approach, which we followed in the Record case. Figures 5, 6, 7 and 8 show time series starting on February 18^{th}, up to February 21^{st} 2015. We do not show the response times of the inexistent page as these are always 0 or 1, thus having very little information of interest for us. In all these figures, we add a plot of the moving average with a period of 100, as the moving average is extremely helpful to identify tendencies.

Figures 5 and 6 show the request time of the old 129 KiB page request. The former figure shows the actual times we got, whereas in the latter we deleted the highest peaks (those above average), to get a clearer picture of the request times. A daily pattern emerges in these figures, as woken hours have longer delays in the response than sleeping hours. To exclude the network as a bottleneck, we can visually see that the response times of Figure 7 do not exhibit this pattern, which suggests a low correlation between request and response times (which is indeed low). Next, we observe that the request times of the existent and inexistent pages (refer to Figure 8) are out of sync. The latter seems to have much smaller cycles along the day, although (different) daily patterns seem to exist as well. For the reasons we mentioned before, in Section III, under the assumption that processing bottlenecks would *simultaneously* affect both plots, we conclude that the main source of bottlenecks in the existent page is I/O. This also suggests the impossibility of having the request time dominated by access to a cache on the server, as this would impact processing, thus causing synchronized delays. A final word for the peaks that affect the request time: they are weakly correlated to the response times. Hence, their source is also likely to be I/O.

V. Discussion and Conclusion

We proposed to detect bottlenecks of HTTP servers using client-side observations of request and response times. A comparison of these signals either over the same or a small number of resources enables the identification of CPU, network and I/O bottlenecks. We did this work having no access to internal server data and mostly resorting to visual inspection of the request and response times. If run by the owners of the site, we see a number of additional options:

- Simply follow our approach of periodically invoking URLs in one or more clients, as a means to complement current server-side monitoring tools. This may help to reply to questions such as "what is the impact of a CPU occupation of 80% for interactivity?".

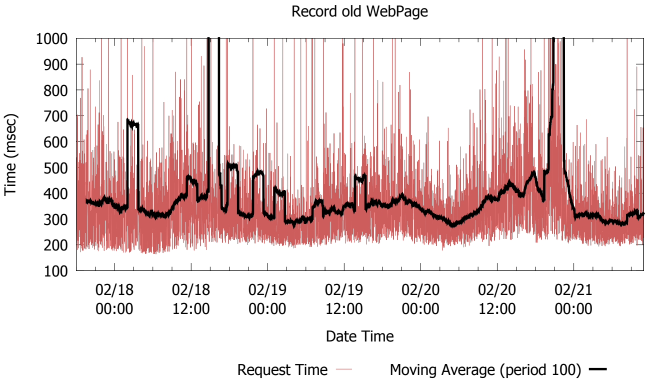


Figure 5. Record old page — request times.

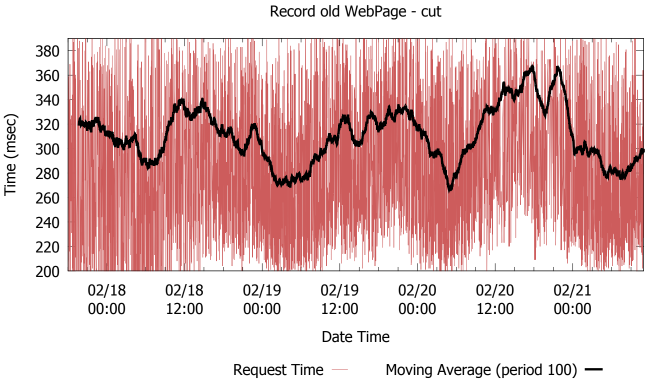


Figure 6. Record old page — response times with peaks cut.

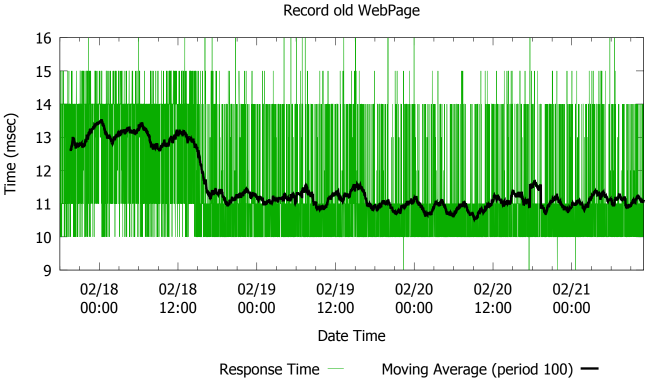


Figure 7. Record old page — response times.

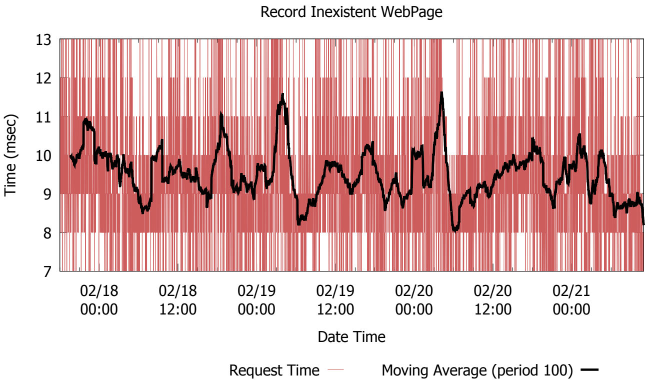


Figure 8. Record inexistent page — request times.

- A hybrid approach, with client-side and server-side data is also possible. I.e., the server may add some internal data to *each* request, like the time the request takes on the CPU or waiting for the database. Although much more elaborate and dependent on the architecture, instrumenting the client and the server sides is, indeed, the only way to achieve a fully decomposition of request timings.
- To improve the quality of the analysis we did in Section IV, site owners could also add a number of very specific resources, like a page that has known access time to the DB, or known computation time.
- It is also possible to automatically collect timing information from real user browsers, as in Google Analytics [33], to do subsequent analysis of the system performance. In other words, instead of setting up clients for monitoring, site owners might use their real clients, with the help of some Javascript and AJAX.

In summary, we collected evidence in support of the idea of identifying bottlenecks from the user side. Nonetheless, to unambiguously demonstrate the results we found, we recognize the need for further evidence, from a larger number of sites, and from supplementary monitoring data from the server.

REFERENCES

[1] RFC 2616 - Hypertext Transfer Protocol – HTTP/1.1, Internet Engineering Task Force (IETF), Internet Engineering Task Force (IETF) Std., June 1999. [Online]. Available: http://www.faqs.org/rfcs/rfc2616.html

[2] D. Battre, M. Hovestadt, B. Lohrmann, A. Stanik, and D. Warneke, “Detecting bottlenecks in parallel dag-based data flow programs,” in Many-Task Computing on Grids and Supercomputers (MTAGS), 2010 IEEE Workshop on, 2010, pp. 1–10.

[3] W. Iqbal, M. N. Dailey, D. Carrera, and P. Janecek, “Adaptive resource provisioning for read intensive multi-tier applications in the cloud,” Future Generation Computer Systems, vol. 27, no. 6, 2011, pp. 871–879.

[4] Y. Shoaib and O. Das, “Using layered bottlenecks for virtual machine provisioning in the clouds,” in Utility and Cloud Computing (UCC), 2012 IEEE Fifth International Conference on, 2012, pp. 109–116.

[5] N. Huber, F. Brosig, and S. Kounev, “Model-based self-adaptive resource allocation in virtualized environments,” in Proceedings of the 6th International Symposium on Software Engineering for Adaptive and Self-Managing Systems, ser. SEAMS ’11. New York, NY, USA: ACM, 2011, pp. 90–99. [Online]. Available: http://doi.acm.org/10.1145/1988008.1988021

[6] P. Bodík, R. Griffith, C. Sutton, A. Fox, M. Jordan, and D. Patterson, “Statistical machine learning makes automatic control practical for internet datacenters,” in Proceedings of the 2009 conference on Hot topics in cloud computing, ser. HotCloud’09. Berkeley, CA, USA: USENIX Association, 2009. [Online]. Available: http://dl.acm.org/citation.cfm?id=1855533.1855545

[7] Q. W. *et al.*, “Detecting transient bottlenecks in n-tier applications through fine-grained analysis,” in ICDCS. IEEE Computer Society, 2013, pp. 31–40. [Online]. Available: http://dblp.uni-trier.de/db/conf/icdcs/icdcs2013.html#WangKLJSMKP13

[8] Q. Zhang, L. Cherkasova, and E. Smirni, “A regression-based analytic model for dynamic resource provisioning of multi-tier applications,” in Autonomic Computing, 2007. ICAC ’07. Fourth International Conference on, June 2007, pp. 27–27.

[9] G. Franks, D. Petriu, M. Woodside, J. Xu, and P. Tregunno, “Layered bottlenecks and their mitigation,” in Quantitative Evaluation of Systems, 2006. QEST 2006. Third International Conference on, Sept 2006, pp. 103–114.

[10] S. Malkowski, M. Hedwig, J. Parekh, C. Pu, and A. Sahai, “Bottleneck detection using statistical intervention analysis,” in Managing Virtualization of Networks and Services. Springer, 2007, pp. 122–134.

[11] S. Malkowski, M. Hedwig, and C. Pu, “Experimental evaluation of n-tier systems: Observation and analysis of multi-bottlenecks,” in Workload Characterization, 2009. IISWC 2009. IEEE International Symposium on. IEEE, 2009, pp. 118–127.

[12] R. Chi, Z. Qian, and S. Lu, “A heuristic approach for scalability of multi-tiers web application in clouds,” in Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2011 Fifth International Conference on, 2011, pp. 28–35.

[13] R. Singh, U. Sharma, E. Cecchet, and P. Shenoy, “Autonomic mix-aware provisioning for non-stationary data center workloads,” in Proceedings of the 7th international conference on Autonomic computing, ser. ICAC ’10. New York, NY, USA: ACM, 2010, pp. 21–30. [Online]. Available: http://doi.acm.org/10.1145/1809049.1809053

[14] W. Iqbal, M. N. Dailey, D. Carrera, and P. Janecek, “Sla-driven automatic bottleneck detection and resolution for read intensive multi-tier applications hosted on a cloud,” in Advances in Grid and Pervasive Computing. Springer, 2010, pp. 37–46.

[15] H. Liu and S. Wee, “Web server farm in the cloud: Performance evaluation and dynamic architecture,” in Proceedings of the 1st International Conference on Cloud Computing, ser. CloudCom ’09. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 369–380. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-10665-1_34

[16] B. Singh and P. Nain, “Article: Bottleneck occurrence in cloud computing,” IJCA Proceedings on National Conference on Advances in Computer Science and Applications (NCACSA 2012), vol. NCACSA, no. 5, May 2012, pp. 1–4, published by Foundation of Computer Science, New York, USA.

[17] M. Attariyan, M. Chow, and J. Flinn, “X-ray: automating root-cause diagnosis of performance anomalies in production software,” in Proceedings of the 10th USENIX conference on Operating Systems Design and Implementation, ser. OSDI’12. Berkeley, CA, USA: USENIX Association, 2012, pp. 307–320. [Online]. Available: http://dl.acm.org/citation.cfm?id=2387880.2387910

[18] F. Oliveira, A. Tjang, R. Bianchini, R. P. Martin, and T. D. Nguyen, “Barricade: defending systems against operator mistakes,” in Proceedings of the 5th European conference on Computer systems, ser. EuroSys ’10. New York, NY, USA: ACM, 2010, pp. 83–96. [Online]. Available: http://doi.acm.org/10.1145/1755913.1755924

[19] “Papers — HP Web server performance tool,” http://www.hpl.hp.com/research/linux/httperf/, retrieved: May, 2015.

[20] “Performance tools — Apache JMeter^{TM},” http://jmeter.apache.org/, retrieved: May, 2015.

[21] “Papers — Navigation Timing,” https://dvcs.w3.org/hg/webperf/raw-file/tip/specs/NavigationTiming/Overview.html, retrieved: May, 2015.

[22] J. Postel, “Transmission Control Protocol,” RFC 793 (Standard), Internet Engineering Task Force, Sep. 1981, updated by RFCs 1122, 3168. [Online]. Available: http://www.ietf.org/rfc/rfc793.txt

[23] “Papers — Selenium Browser automation,” http://www.seleniumhq.org/, retrieved: May, 2015.

[24] “Crontab - quick reference — admin’s choice - choice of unix and linux administrators,” http://www.adminschoice.com/crontab-quick-reference, retrieved: May, 2015.

[25] “Xvfb,” http://www.x.org/archive/X11R7.6/doc/man/man1/Xvfb.1.xhtml, retrieved: May, 2015.

[26] “Breaking news, u.s., world, weather, entertainment & video news - cnn.com,” http://edition.cnn.com, retrieved: May, 2015.

[27] “Amazon.com: Online shopping for electronics, apparel, computers, books, dvds & more,” http://www.amazon.com, retrieved: May, 2015.

[28] “SAPO,” http://www.sapo.pt, retrieved: May, 2015.

[29] “Alexa — Top Sites in Portugal,” http://www.alexa.com/topsites/countries/PT, retrieved: May, 2015.

[30] “:.:Jornal Record:.:,” http://www.record.xl.pt, retrieved: May, 2015.

[31] “:.: Albiol: Relação arrefeceu entre Casillas e Arbeloa - Entrevistas - Jornal Record :.:,” http://www.record.xl.pt/Entrevistas/interior.aspx?content_id=826333, retrieved: May, 2015.

[32] “Inexistent record page,” http://www.record.xl.pt/naoexiste.

[33] B. Clifton, Advanced Web Metrics with Google Analytics. Alameda, CA, USA: SYBEX Inc., 2008.

A Domain-Specific Language for Modeling Web User Interactions with a Model Driven Approach

Carlos Eugênio Palma da Purificação / Paulo Caetano da Silva
Salvador University (UNIFACS)
Salvador, Brazil
email: carloseugenio@gmail.com / paulo.caetano@pro.unifacs.br

***Abstract*-Domain Specific Languages have many applications. They provide a way to abstract domain concepts and express these concepts in a more expressive way when compared with general-purpose languages. Recently the Object Management Group has released the beta version of its Interaction Flow Modeling Language standard to model user interaction in applications. We contribute with a proposal of a textual domain-specific language, to use this model as a base for modeling user interaction in web applications, along with Model Driven Software Development techniques and a set of tools and components, to propose a generative approach to model user interaction in web applications. The approach allows the definition of user interaction flow models using a textual form, architecture reuse, improved code quality and speed in development.**

Keywords-Model Driven Software Development; Domain Specific Languages; Web Applications.

I. INTRODUCTION

Researches in software reuse show that in order to achieve significant results in this field, a paradigm shift towards the use of software families rather than individual systems is required [18]. In order to achieve this goal, many technics have been proposed. In this work, we propose a tool-set for modeling user interactions in applications using a textual Domain-Specific Language (DSL), called EngenDSL [12], modified for accommodate most *Interaction Flow Modeling Language* (IFML) [11] concepts. The language is used along with a set of tools and libraries that realize the Model-Driven Software Development (MDSD) [10] generative software approach to create applications, to generate a small application as a proof of concept of the methodology.

We have used the IFML standard [11] to base our user interaction model, adding some important features such as the possibility to define and reuse the layout and style for views, specify presentation libraries, attaching behavior semantics and chaining to user actions. Whilst previous researches in this area [5] provide some kind of support to user interaction modeling, until recently, there was not a formal specification for modeling user interactions such as the OMG IFML Standard.

The rest of this paper is structured as follows. In the next section, concepts and background information related to main topics of this paper are presented. Section 3 describes the EngenDSL language meta-model and constructs. Section 4 discusses the model transformation problem. Section 5 provides an example application that shows language constructs and some development tools used in the approach. Section 6 presents some related work. Section 7 provides some conclusion and planning for future work.

II. CONCEPTS AND BACKGROUND

The MDD – *Model-Driven Development* is a technique that gives software models a high importance role in software development process.

Following the concept of prioritization of models and their use as key artifacts during all phases of software development (specification, analysis, design, implementation and testing), is the Model-Driven Software Development (MDSD) [10]. This concept uses an agile approach to software development, along with generative techniques to deliver software based on a software product lines – SPL approach.

As in all areas of science and engineering, there are always specific and generic approaches to solve a given problem [3]. The description of a problem in a language developed specifically for its domain tends to be a more optimized and direct solution, possibly caring greater expressiveness to describe the domain concepts. Domain-specific Languages are languages created for a particular domain. They are also called specialized languages, problem-oriented or special purpose languages [9]. Deursen et al [3], defines DSL as programming languages or executable specification languages that offer great power to express, with notations and abstractions usually focused on a restricted domain, a particular problem.

IFML [11] is a modeling standard developed by OMG designed for expressing the content, user interaction and control behavior of software applications front-end. It allows software practitioners to model and describe the main parts of an application front-end. Therefore, the specification divides these parts in the following dimensions: (i) **View** – the view part of the application composed of containers and view components; (ii) **State and Business Logic** – the representation of objects that carries application state and the business logic that can be executed; (iii) **Data and Event Binding** – the binding of the view components to data objects and events; (iv) **Event Control Logic** – the logic responsible to determining the action execution order after in response to an event; (v) **Architecture Distribution** – the distribution of control, data and business logic at the application tiers.

The general IFML meta-model is presented in [11]. The *IFMLModel* is the core container of IFML meta-model elements and contains a *InteractionFlowModel* which is the user view of an IFML application. It is composed of *InteractionFlowModelElements* – an abstract class that represents a generalization of every element of a *InteractionFlowModel* (the view elements); *DomainModel* – model elements that represents the business domain or data; *ViewPoints* – represents only specific aspects of the system to facilitate comprehension of referencing *InteractionFlowModelElements*.

The *DomainModel* represents the business domain of the application, the content and the behavior that is referenced by the *InteractionFlowModel*. The *DomainModel* contain *DomainElements*. These are specialized in terms of concepts (*DomainConcept*), properties (*FeatureConcept*), behaviors (*BehaviorConcept*) and methods (*BehaviorFeatureConcept*).

III. THE ENGENDSL

The EngenDSL extension was modeled to aid in developing applications by abstracting user interaction concepts based on IFML standard. The main objective was to define a textual DSL, which could be used to streamline user interaction modeling. The foundation for this was the IFML standard and in Figures 1, 2 and 3, we present the main concepts and extensions provided by EngenDSL to the IFML standard.

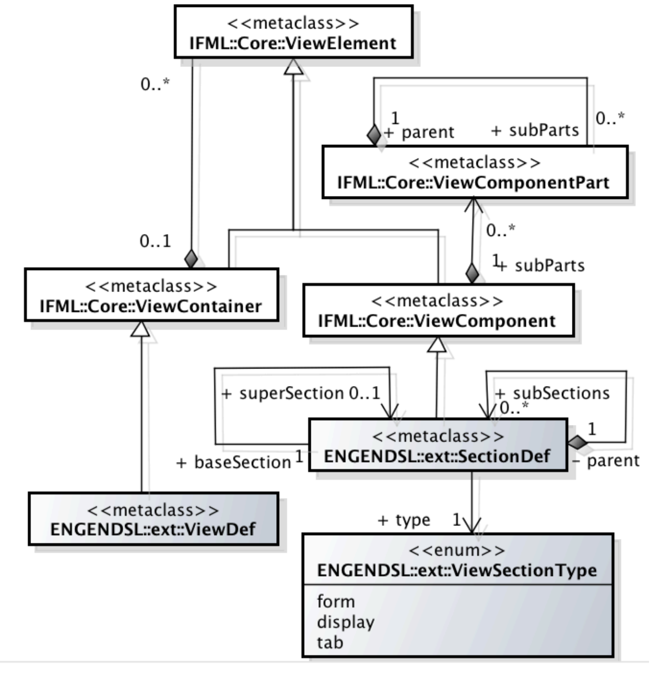


Figure 1. The EngenDSL IFML Extension Package for View Elements

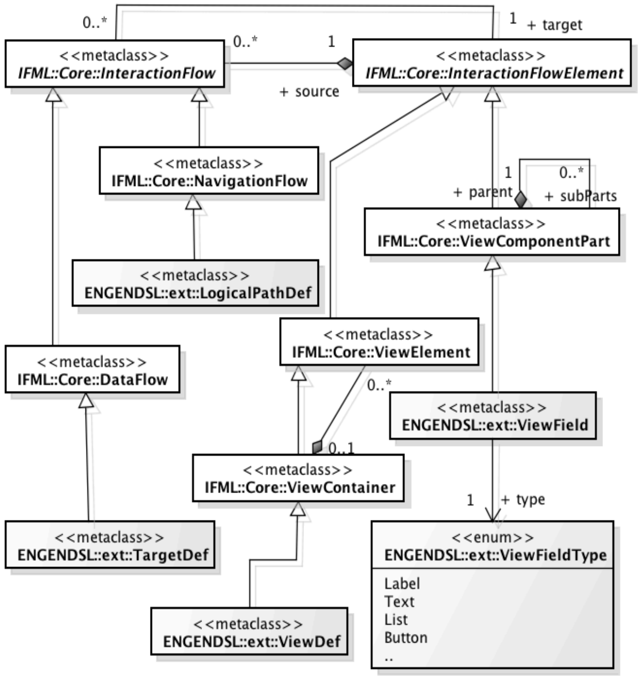


Figure 2. The EngenDSL IFML Extension package for Interaction Flow Elements

Therefore, this EngenDSL language extension was designed to comply with the IFML standard as much as possible, while giving the possibility to define other architectural aspects not covered by the standard. Figure 4 summarizes the mapping between IFML and EngenDSL main elements.

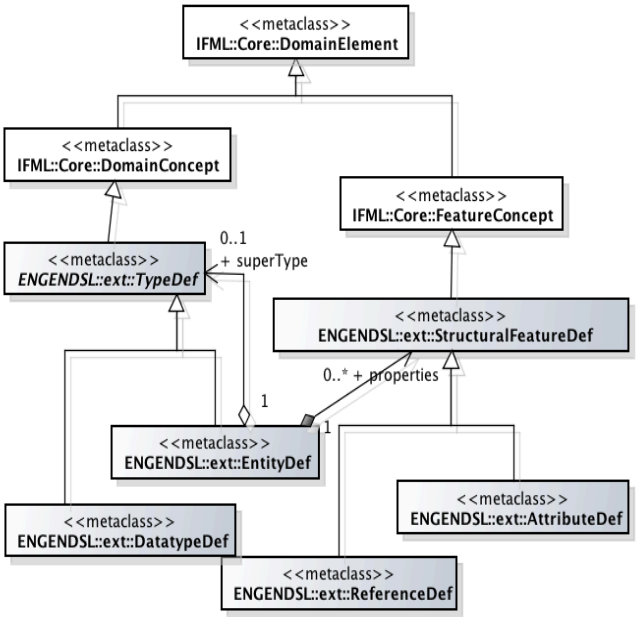
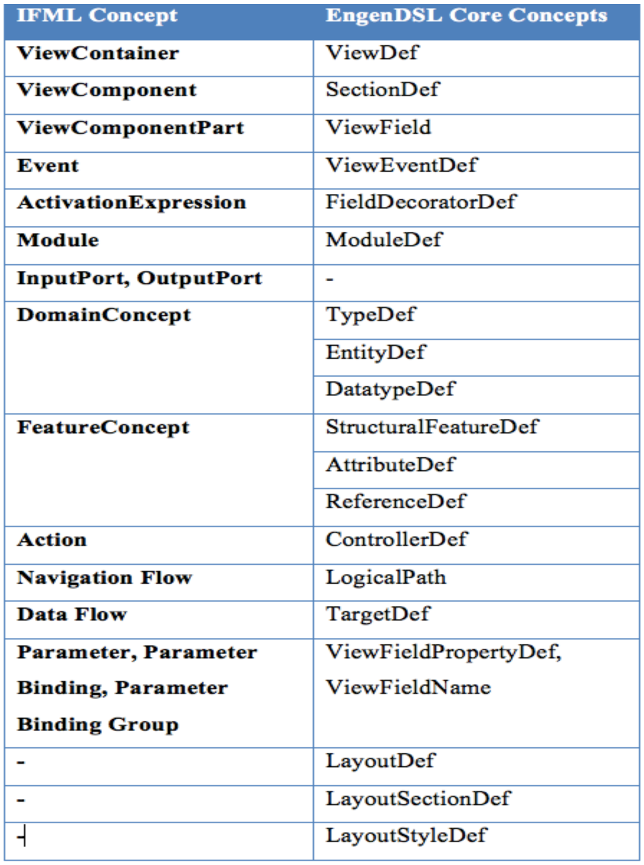


Figure 3. EngenDSL IFML Extensions Package for Domain Elements



IFML Concept	EngenDSL Core Concepts
ViewContainer	ViewDef
ViewComponent	SectionDef
ViewComponentPart	ViewField
Event	ViewEventDef
ActivationExpression	FieldDecoratorDef
Module	ModuleDef
InputPort, OutputPort	-
DomainConcept	TypeDef
	EntityDef
	DatatypeDef
FeatureConcept	StructuralFeatureDef
	AttributeDef
	ReferenceDef
Action	ControllerDef
Navigation Flow	LogicalPath
Data Flow	TargetDef
Parameter, Parameter Binding, Parameter Binding Group	ViewFieldPropertyDef, ViewFieldName
-	LayoutDef
-	LayoutSectionDef
-	LayoutStyleDef

Figure 4. EngenDSL-IFML Elements mapping table

We briefly describe these elements and show examples on how they are used in application models. The examples and further details can be found in Section 5, and the complete model for the approach can be seen in the GitLab public repository for this project [4].

The IFML standard defines *ViewContainers* as containers to other *ViewElements* like other *ViewContainers* and *ViewComponents*. In the proposed language, the same semantics apply. In the language, *Views* are model elements that represent the container IFML concept. Structurally, in the language, *Views* may be composed of other *Views* and *Sections* (see *Section* element). *Sections* are components that compose *Views* and display content. *Section* definitions are laired and composed inside *View* definitions and have a type - *ViewSectionType*. A *Section* definition, in the proposed language, represents the IFML concept of *ViewComponent*. Therefore, sections are used to present or capture any user data, which conforms to the IFML standard. They have a name, a type, a namespace, and may be composed of other *Sections* or *Fields* – which are called subsections. Within *Sections,* we have *Fields*. They represent the IFML *ViewComponentPart* concept. *Fields* have a name, a type – *ViewFieldType* and properties that define specific field semantics like its ID, Label, etc. The Field's type semantically defines its behavior and presentation characteristics such as *"button"* or *"select"* field types. The *EngenDSL* language extends the IFML *ViewComponent* concept, in that a *Section* definition can extend, or in-line include another *Section* definition.

When developing business applications, as a rule, we usually have present the idea of model entities that represent real-world concepts. In the *EngenDSL*, the *Entity* concept is derived from the Model concept in Model View Controller (MVC) [8] and the Domain Concept in IFML.

The EngenDSL directly uses the concept of *Controllers* representing, according to MVC standard, the components that mediate data input and output for both the MVCs *Model* and *View*. The concept of *LogicalPath* is directly associated with the concept of navigation between *Views* and *Controllers*. A *Controller*, at the end of processing, redirects the user to a path. Defining this path is the function of a *LogicalPath*, which is an abstraction for the concept of a redirect in a web application. However, as a *Controller* can resolve to trigger another *Controller* at processing end, *Views* and *Controller*s represent, and are extensions of *LogicalPath*s within *EngenDSL*. Therefore, a *Controller* may determine, in accordance with its implementation logic, redirect the user to a path – *LogicalPath*, which will ultimately render a web page to the application user, or send the application flow to another *Controller*. Note that, at some point, the last *LogicalPath* in a flow within the application logic will be a *View*.

Figure 5 shows an example of such a navigation flow, detailed later in this work in Section 5 using the proposed language. Here the *ListProductsCtrl*, a controller whose type is *Search* (will find data in repository and return the list found), has one of its *LogicalPaths* (*success* – the predefined default logical path) pointing to the View *ListProductsView*. After the selection event is triggered on an item in the list in this view, the *ProductListTable* field *target* property defines the next item in the interaction flow: the *ViewProductCtrl*, which is a controller of type *View*. These controllers will find an item in persistence storage and return the item data. After performing the action, the controller will use its default "*success*" *LogicalPath* to determine the next element in the interaction flow - *ViewProductVw*.

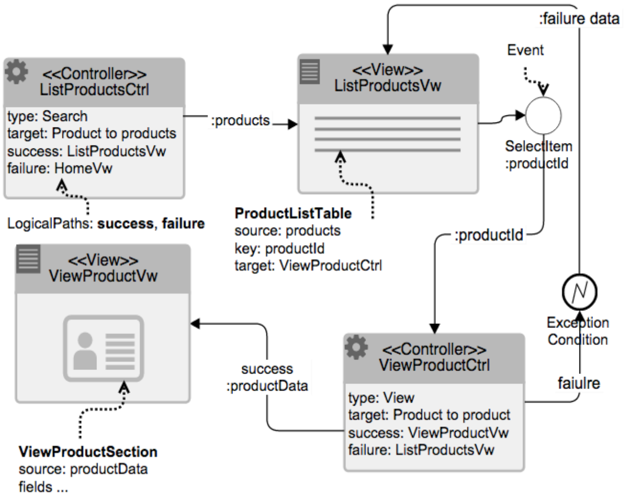


Figure 5. Navigation Flow Example

IV. MODEL TRANSFORMATIONS

In consonance with Model-Driven Software Development technics, the proposed approach uses intermediate model-to-model transformation phase, or phases, before final code generation and integration in the target platform architecture. We use this technique to enable the specification of requirements and accidental concerns, which are foreign to the proposed DSL. The proposed approach uses an intermediate model – *Engen Intermediate Model* (EIM), which can be partially seen in Figure 6. The entire model can be seen in the project GitLab repository [4].

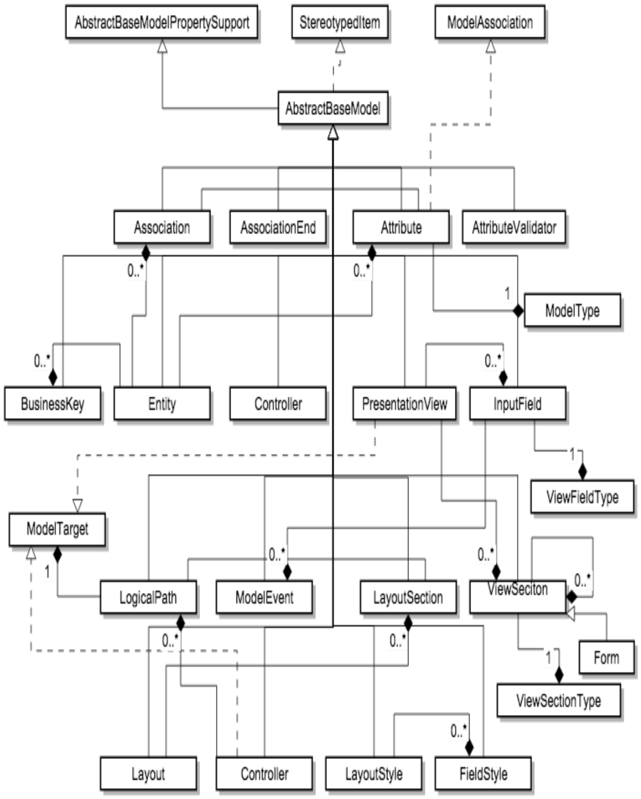


Figure 6. The Engen Intermediate Model (EIM)

The solution was implemented using the XText Framework [19] and other components. The overall elements included in the solution are shown in Figure 7. One of these components is the *EngenGenerator Plugin*. This component, in turn, uses a set of rules specified by the internal *MTMTransformer* interface.

The rule set is configured in a per project configuration file. Each line in this file defines a component (usually a Java class) that implements the *MTMTransformer* interface and its purpose is to transform one or more elements from the given EMF model to the EIM model already shown in Figure 6. The triggering of M2M transformation component can be done by any Java component.

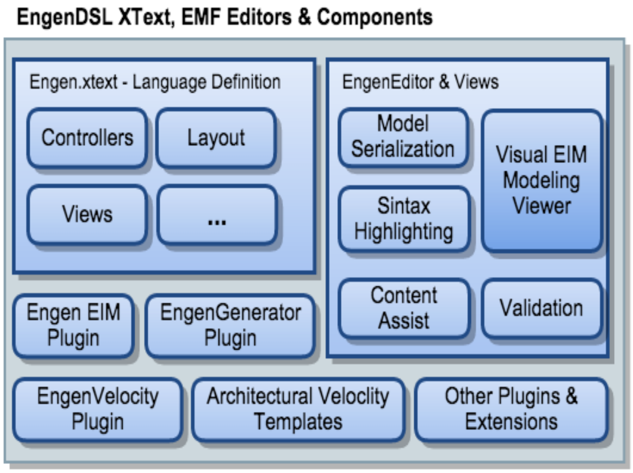
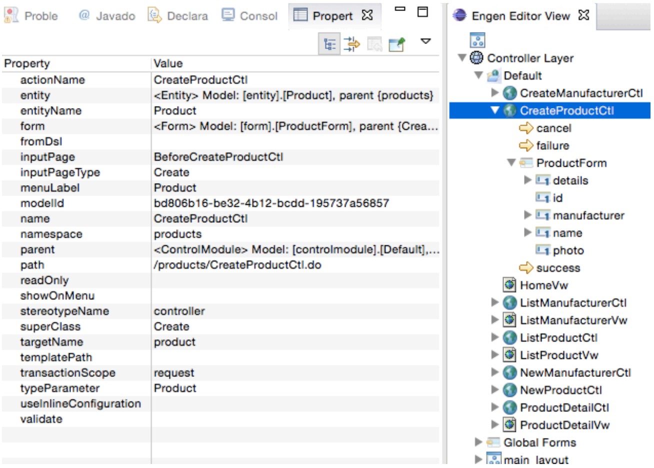


Figure 7. EngenDSL Components

The output from M2M transformation is a XML file that serializes the generated intermediate model. This model can then be further customized by using a technique called model weaving [14]. The proposed implementation of this concept consists in augmenting the original model by setting the value of some properties in separated configuration files.

The rationale for using this technique is that after each model transformation and serialization, attached custom information is not lost. This is automatically done by the provided tooling as shown in Figure 8, in which we can see some properties available for customization after a M2M transformation phase. These properties come from the previously defined intermediate model already shown in Figure 6, and usually have a default value that can be changed by setting the related property of the model.

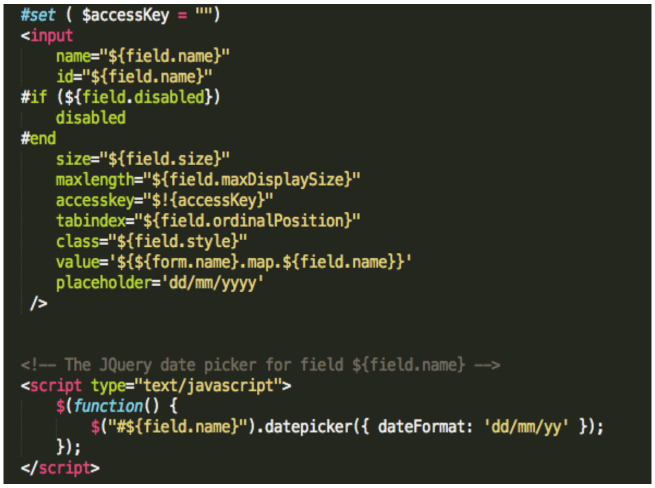


Property	Value
actionName	CreateProductCtl
entity	<Entity> Model: [entity].[Product], parent {products}
entityName	Product
form	<Form> Model: [form].[ProductForm], parent {Crea...
fromDsl	
inputPage	BeforeCreateProductCtl
inputPageType	Create
menuLabel	Product
modelId	bd806b16-be32-4b12-bcdd-195737a56857
name	CreateProductCtl
namespace	products
parent	<ControlModule> Model: [controlmodule].[Default],...
path	/products/CreateProductCtl.do
readOnly	
showOnMenu	
stereotypeName	controller
superClass	Create
targetName	product
templatePath	
transactionScope	request
typeParameter	Product
useInlineConfiguration	
validate	

Figure 8. Model weaving customization in tool

The system developer should use this model to fine-tune the created model representation of the DSL concepts. After this step, the M2T transformation can take place. This last step is similar to M2M transformation but require templates targeted to the final application specific platform and architecture. These templates are responsible to transform the augmented intermediate model into target platform code.

In Figure 9, we show a small portion of a template written in Apache Velocity [15] that receives a model element, identified as *$field* in the template, and outputs an HTML date text field with an associated action.



```
#set ( $accessKey = "")
<input
    name="${field.name}"
    id="${field.name}"
#if (${field.disabled})
    disabled
#end
    size="${field.size}"
    maxlength="${field.maxDisplaySize}"
    accesskey="$!{accessKey}"
    tabindex="${field.ordinalPosition}"
    class="${field.style}"
    value='${${form.name}.map.${field.name}}'
    placeholder='dd/mm/yyyy'
 />


<!-- The JQuery date picker for field ${field.name} -->
<script type="text/javascript">
    $(function() {
        $("#${field.name}").datepicker({ dateFormat: 'dd/mm/yy' });
    });
</script>
```

Figure 9. Apache Velocity template for a text field

Therefore, in conjunction with templates, tailored to generate the artifacts to the target architecture, this intermediate step, allows to complete the system configuration, build and generation of the final solution, which will be based on the target architecture specified by intermediate and final transformations, model weaving and templates.

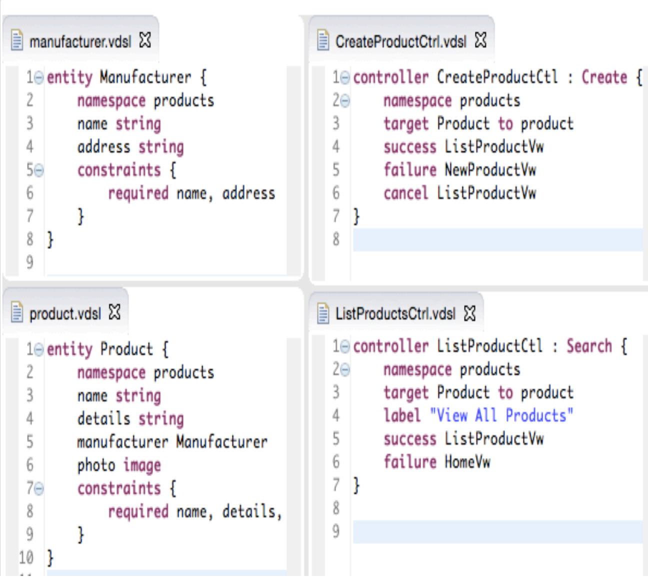
V. AN APPLICATION EXAMPLE

In this section, we show a portion of an application modeled using the proposed method. The full examples can be found on the project GitLab public folder [4].

The example is a small shopping application. Some of the *Entities*, along with properties and constraints are shown in the example in Figure 10.

The domain model represented by the *Entity* instances is straightforward. They define the domain model elements along with their properties and constraints. Following are the controllers, for example the *ListProductCtl*. The definition specifies a controller model element. In this case, the *Search* type is defined, areas in the *CreateProductCtl* a *Create* controller type is defined. The “*Search”* and “Create” controller types (after the colon in the controller definition) configuration binds the model element to an external component that performs a search in the persistence storage for the domain model element - specified by the *target* property, using the provided criteria. In this specific example, if this controller is accessed from the home page, all elements in persistence storage should be returned, since there are no criteria defined.

On the *ListProductVw* view element in Figure 11, we can see how *Sections* can be nested and how the search for the products can be further customized.



manufacturer.vdsl

```
entity Manufacturer {
    namespace products
    name string
    address string
    constraints {
        required name, address
    }
}
```

CreateProductCtrl.vdsl

```
controller CreateProductCtl : Create {
    namespace products
    target Product to product
    success ListProductVw
    failure NewProductVw
    cancel ListProductVw
}
```

product.vdsl

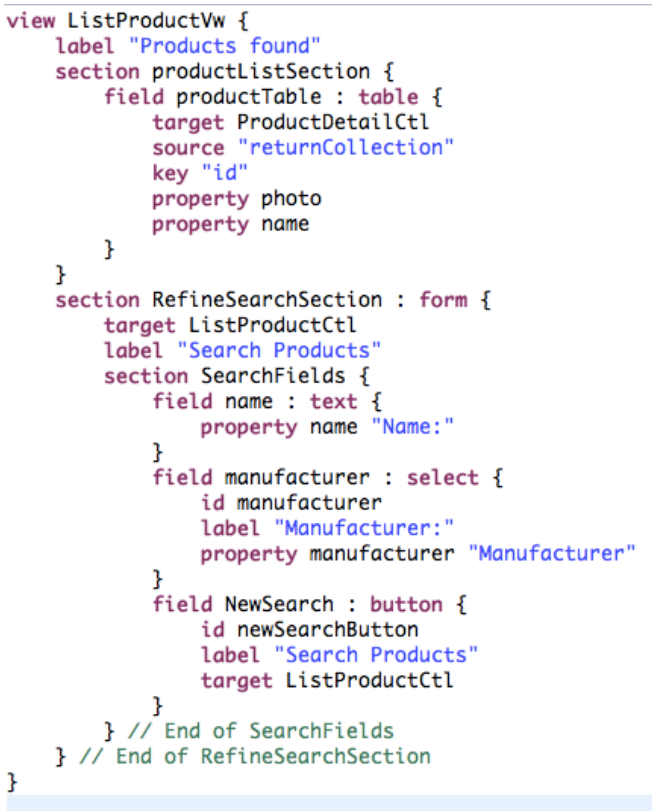
```
entity Product {
    namespace products
    name string
    details string
    manufacturer Manufacturer
    photo image
    constraints {
        required name, details,
    }
}
```

ListProductsCtrl.vdsl

```
controller ListProductCtl : Search {
    namespace products
    target Product to product
    label "View All Products"
    success ListProductVw
    failure HomeVw
}
```

Figure 10. Entities and Controllers definitions

While when the *ListProductsCtl* controller was accessed without any search context information the controller performed a database search for all products, without any search criteria, the search triggered by the *NewSearch* button, on the *RefineSearchSection* form, will use the context information from the input fields (*name* AND *manufacturer*) as the criteria for the controller to further filter the results.

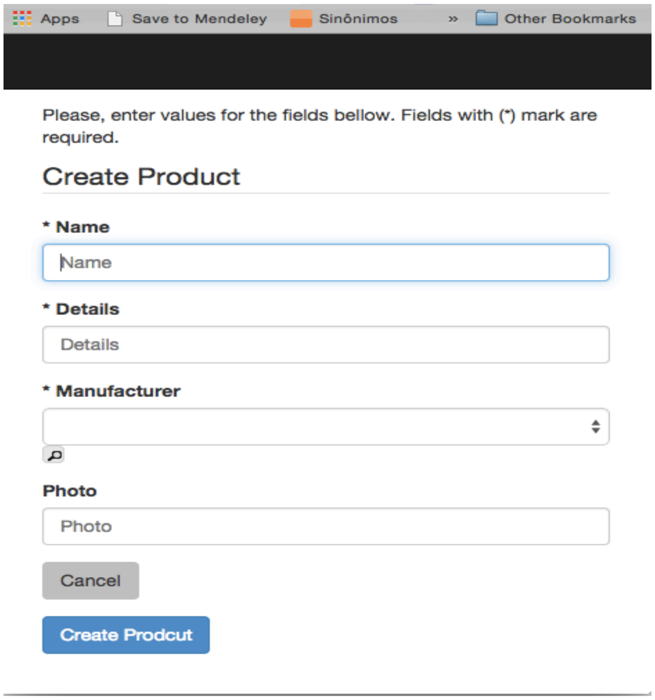


```
view ListProductVw {
    label "Products found"
    section productListSection {
        field productTable : table {
            target ProductDetailCtl
            source "returnCollection"
            key "id"
            property photo
            property name
        }
    }
    section RefineSearchSection : form {
        target ListProductCtl
        label "Search Products"
        section SearchFields {
            field name : text {
                property name "Name:"
            }
            field manufacturer : select {
                id manufacturer
                label "Manufacturer:"
                property manufacturer "Manufacturer"
            }
            field NewSearch : button {
                id newSearchButton
                label "Search Products"
                target ListProductCtl
            }
        } // End of SearchFields
    } // End of RefineSearchSection
}
```

Figure 11. The ListProductVw View Definition

This is automatically done by the ad-hoc *Web Framework* we are using. We can see from the *NewSearch* button configuration, that the *target* property points to the *ListProductCtl* controller. This information corresponds to the event model definition for this element. In the proposed model, this information is sufficient for templates to interpret the event model: the target controller for the action (along with its fields containing the criteria for the query entered by the user).

Figure 12 shows a screenshot for the generated HTML running in an application server for the *NewProductVw View*.



Please, enter values for the fields bellow. Fields with (*) mark are required.

Create Product

* Name

Name

* Details

Details

* Manufacturer

Photo

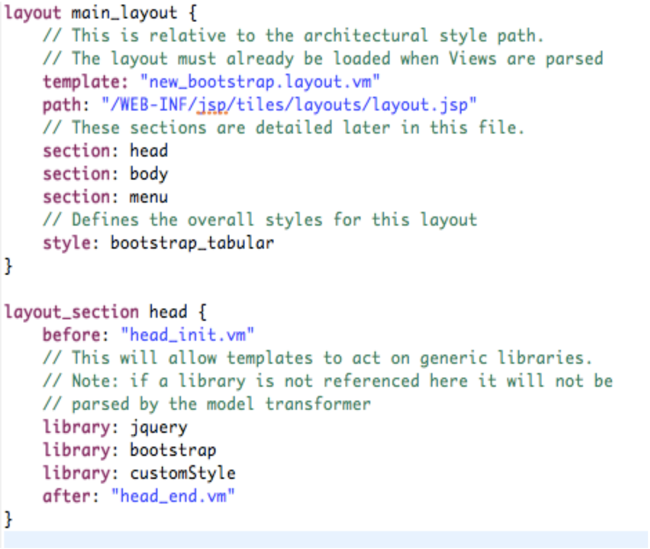
Photo

Figure 12. Running Application Screenshot for the Create Product View

As stated before, the proposed approach includes a way to define and reuse application layout. In the DSL, we declare the partial layout for the application as shown in Figure 13.

The layout name appears after the *layout* keyword. The layout specifies a *template* property that will be parsed by the M2T engine, and will produce the file defined by the *path* property. A small section for the specified template is shown in Figure 14.

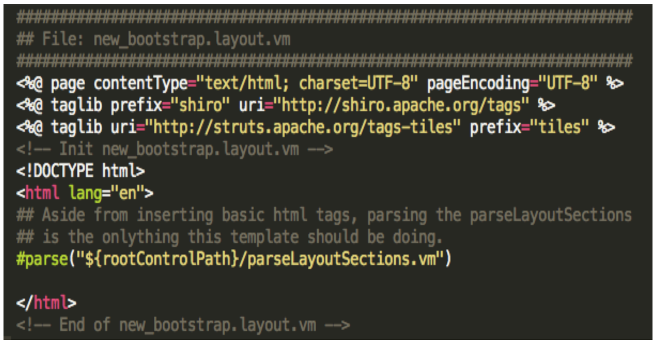
This template defines some markup instructions and delegates the layout subsections parsing to another template called *parseLayoutSections*. Each section defined by the main layout element is a composition of URLs, templates and libraries. For example, the *head* section shown in Figure 12, defines a template to be parsed before and other after the main template, for this section (which will be by default the *section* name if the *template* attribute is not defined). This information is just stored in the model, not hardcoded in the solution, so they can be used by the templates and M2T components as they find applicable.



```
layout main_layout {
    // This is relative to the architectural style path.
    // The layout must already be loaded when Views are parsed
    template: "new_bootstrap.layout.vm"
    path: "/WEB-INF/jsp/tiles/layouts/layout.jsp"
    // These sections are detailed later in this file.
    section: head
    section: body
    section: menu
    // Defines the overall styles for this layout
    style: bootstrap_tabular
}

layout_section head {
    before: "head_init.vm"
    // This will allow templates to act on generic libraries.
    // Note: if a library is not referenced here it will not be
    // parsed by the model transformer
    library: jquery
    library: bootstrap
    library: customStyle
    after: "head_end.vm"
}
```

Figure 13. Partial Layout definition



```
########################################################################
## File: new_bootstrap.layout.vm
########################################################################
<%@ page contentType="text/html; charset=UTF-8" pageEncoding="UTF-8" %>
<%@ taglib prefix="shiro" uri="http://shiro.apache.org/tags" %>
<%@ taglib uri="http://struts.apache.org/tags-tiles" prefix="tiles" %>
<!-- Init new_bootstrap.layout.vm -->
<!DOCTYPE html>
<html lang="en">
## Aside from inserting basic html tags, parsing the parseLayoutSections
## is the onlything this template should be doing.
#parse("${rootControlPath}/parseLayoutSections.vm")

</html>
<!-- End of new_bootstrap.layout.vm -->
```

Figure 14. Section Layout template

VI. RELATED WORK

Several works have been proposed for modeling and specification of Web systems. *WebML* [16] and *WebDSL* [17] are examples. This work starts from these approaches to specify a textual and declarative form of describing the various aspects involved in this type of application. However, none of them is based in an independently defined standard like IFML.

The *WebML* proposed in Ceri et al [16], also defines a model that uses other work as the basis for the definition of Web applications. *WebML* uses *UML* and *XML* in the data dimensions (structural model), pages (composition model), topology of links between pages (navigation model), the graphics and layout requirements for the pages (presentation model) and customization needs (customization model) for delivering content to the user.

Visser proposal [17] is more similar to the one presented in this work, called *WebDSL*. It defines a language for describing web applications, covering its various aspects. However, in a different direction than the one presented in this article, the language goes beyond the simple statement of these concepts to specify conditional structures, functions and methods, including algorithms, which are used to define the behavior of certain parts of the application architectural structure and components. While this approach has some advantages, it possibly creates a greater complexity in the development process.

VII. Conclusion and Future Work

Domain Specific Languages are a technological trend. Its application in Web systems is perfectly feasible and even recommended. Not only because of productivity that this application can achieve, but also by all other benefits that a development approach based on models and automatic generation of artifacts can bring to Web Systems Engineering. The benefits include rapid construction and prototyping, the reduction of failures, standard architecture and solutions, suitable and adaptable technology approach, in addition to allow a safe evolution between different technologies and frameworks that constantly arise for this type of application.

This work demonstrated a DSL, developed for web specific domain, in consonance with an adopted OMG standard for defining user interactions – IFML. This can improve reusability of models, view interactions modeling understanding and collaboration since the text nature of the solution, enables better handling of models, inclusive when related to version control repositories and tools.

We are now looking into Business Rule integration into the DSL to allow advanced rules verification, along with the basic constraint based rules. We also intend to implement a visualization interface to the model using standard IFML notation.

References

[1] J. Bettin, “Best Practices for Component-Based Development and Model-Driven Architecture”, 2003. Available from: http://s23m.com/publicwhitepapers/best-practices-for-cbd-and-mda.pdf. [retrieved: May, 2015]

[2] K. Czarnecki and S. Helsen. “Classification of Model Transformation Approaches". OOPSLA’03 Workshop on Generative Techniques in the Context of Model-Driven Architecture. 2003, pp. 1–17. Available from: http://s23m.com/oopsla2003/czarnecki.pdf. [retrieved: May, 2015].

[3] A. V. Deursen, P. Klint and J. Visser. “Domain-Specific Languages: An Annotated Bibliography.”. ACM SIGPLAN Notices. Volume 35, Issue 6. 2000, pp 26-36. http://www.st.ewi.tudelft.nl/~arie/papers/dslbib.pdf.

[4] Engen, 2015. “Engen Project Site.” Available at https://gitlab.com/engen/public/wikis/home [retrieved: May, 2015].

[5] V. Estêvão, S. Souza, R. D. A. Falbo, and G. Guizzardi. “A UML Profile for Modeling Framework-Based Web Information Systems.” In Proc of the 12th International Workshop on Exploring Modeling Methods in Systems Analysis and Design. 2007, pp. 149–58.

[6] JM. Favre. “Towards a Basic Theory to Model Model Driven Engineering.” In Procs. of the 3rd Int. Workshop in Software Model Engineering (WiSME). 2004, pp. 262-271.

[7] F. Fondement and R. Silaghi. "Defining Model Driven Engineering Processes". Third International Workshop in Software Model Engineering (WiSME). held at the 7th International Conference on the Unified Modeling Language (UML), 2004.

[8] E. Gamma, R. Helm, R. Johnson, and J. Vlissides. "Design Patterns: Elements of Reusable Object-Oriented Software". 2nd ed. Boston - 1995.

[9] M. Mernik, J. Heering and A. M. Sloane. “When and How to Develop Domain-Specific Languages.” Journal ACM Computing Surveys (CSUR), Volume 37, issue 4, 2005, pp. 316–44. DOI:10.1145/1118890.1118892.

[10] N. Moreno, J. R. Romero and A. Vallecillo. “An Overview of Model-Driven Web Engineering and the MDA". Web Engineering: Modelling and Implementing Web Applications Human-Computer Interaction Series 2008, pp. 353-382.

[11] OMG 2014. Interaction Flow Modeling Language. Available from: http://www.ifml.org. [retrieved: May, 2015].

[12] C. E. Purificação and P. C. Da Silva. “EngenDSL - A Domain Specific Language for Web Applications". 10th International Conference on Information Systems and Technology Management – CONTECSI 10. 2013. pp. 879–99. doi:10.5748/9788599693094-10CONTECSI/PS-474.

[13] T. Stahl, M. Voelter and K. Czarnecki. "Model-Driven Software Development: Technology, Engineering, Management". 2006, John Wiley & Sons. ISBN:0470025700.

[14] M. Völter. “Model-Driven Software Development", 2006. Available from: http://www.voelter.de/data/books/mdsd-en.pdf. [retrieved: May, 2015].

[15] Apache Velocity, 2009. Available from: http://velocity.apache.org. [retrieved: May, 2015].

[16] S. Ceri, P. Fraternali and A. Bongio, "Web Modeling Language (WebML): a modeling language for designing web sites". Computer Networks 33 (1-6), 2000, pp. 137-157.

[17] E. Visser. "WebDSL: A case study in domain-specific language engineering, generative and transformational techniques in software engineering". GTTSE, Lecture Notes in Computer Science, Springer (2008). Tutorial for International Summer School GTTSE, 2008, pp. 1-60.

[18] K. Czarnecki. "Overview of generative software development. In J.-P. Bantre et al,, Ed., Unconventional Programming Paradigms (UPP’04), Lecture Notes in Computer Science, vol. 3566, 2004, pp. 313–328, Mont Saint-Michel, France.

[19] XText, 2009. Available from: http:// www.xtext. org. [retrieved: May, 2015].

Model-Driven Business Process Analytics

Falko Koetter, Andreas Wohlfrom, Désirée Graf
Fraunhofer Institute for Industrial Engineering IAO
and University of Stuttgart IAT
Stuttgart, Germany
Email: firstname.lastname@iao.fraunhofer.de

***Abstract*—Business process analytics helps companies to optimize their processes by identifying shortcomings and improvement potentials. However, existing approaches require either a homogenous process execution environment or expert knowledge, both of which many companies lack. In previous work, we introduced aPro, a model-driven architecture enabling process monitoring even in heterogenous environments. Utilizing the model-driven approach, this work presents a new out-of-the-box approach for business process analytics, allowing the business user to analyze the process within the familiar context of the model without needing to know details of the implementation or data mining techniques. We evaluate this approach using both a simulated and a real-life process.**

***Keywords*–business process management; model-driven architecture; business intelligence; data mining**

I. Introduction

Continuously optimizing business processes is a necessity in todays' fast-changing market [1]. But to identify relevant changes deficits in business process performance need to be known. For this, business intelligence or business process analytics is used, which transforms data obtained during process execution into metrics, Key Performance Indicators (KPI) and finally insights [2].

In previous work, we developed aPro, a model-driven methodology for creating a service-based monitoring infrastructure for business processes [3]. In aPro, a process model is augmented by a so-called goal model containing data to monitor as well as KPI and process goals to be calculated. From this model all components of the infrastructure are automatically created, including monitoring data collection, processing [4], visualization [5] and storage [6]. Using automatically created stubs, these components can be integrated in any business process execution environment, even allowing monitoring of processes run in heterogenous environments, e.g., legacy systems. aPro uses complex event processing (CEP) to process monitoring data, calculate KPIs and detect goal violations in real-time. However, this monitoring only provides data, but does not give any insights in relationships and causes, e.g., for goal violations. A way to analyze process data and detect optimization potential is needed.

In this work, we present an approach for model-driven business process analytics. Based on the aPro architecture, business process analysis has to occur on a level familiar to the business user. As aPro is a model-based approach, the technical details of the implementation are hidden from the business user, who only works with the conceptual goal model. Thus, the analysis needs to present results on this level as well. to achieve this, we investigate exisiting approaches for data mining in regards to their suitability for model-driven business process analysis. In particular, no detailed configuration or expert knowledge of underlying techniques is to be required from the business user. We select two suitable techniques and implement them within the apro prototype. The approach is evaluated using a synthetic and real life use case.

The remainder of this work is structured as follows. In Section II, we give an overview of related work. In Section III, we detail the concept for model-driven business process analytics, including the architecture, extensions and requirements. Section IV contains the selection of data mining techniques sutiable for the requirements. We shortly describe the implementation in Section V. We evaluate the work using both a synthetic and real-life process in Section VI. In Section VII, we give a conclusion and outlook of future work.

II. Related work

In this paper, we describe a technique for model-driven business process analytics. This encompasses the following topics of related work: Model-driven business process

One of the design criterions is integration within the aPro methodology for model-driven process monitoring. In aPro, the process goal modeling (ProGoalML) language is used for modeling the monitoring of a business process, including measurements, KPIs and goals [3]. Measurements are taken at run-time and correlated by process instance using CEP. From the correlated measurements, KPIs and goals are calculated, either for a single process instance or for an aggregation of process (e.g., all instances of the last hour). While this monitoring allows for basic analysis by visualizing data in charts and giving alerts in case of goal violations, it is not sufficient for long-term analysis. As a preliminary work we extended aPro with a data warehouse, which is automatically configured from a ProGoalML model [6]. It provides data for business process analytics.

As model-driven business process analytics aims to be an out-of-the-box technique, related work exists in the areas of business process intelligence and business process analytics.

Castellanos et al. [7] describe one of the first tools for out-of-the-box business process intelligence. Business processes are modeled in a specific notation and executed in a process engine. Process metrics, similar to KPIs and goals in aPro, are modeled in forms and calculated from process audit logs during extraction to a data warehouse. Data mining techniques like decision trees and classification algorithms are used to generate insights, which are visualized in a cockpit, from which process execution can be controled as well. In comparison to this work, metrics are defined separately from the process model, process execution is limited to a process engine and result visualization is separated from the process model as well. In comparison, aPro aims at a high degree of integration between process modeling, goal modeling and analysis, as well as supporting a wide range of execution environments.

ISBN: 978-1-61208-412-1

The *deep Business Optimization Platform* (dBPO) is a platform for business process optimization before, during and after execution [8]. Before execution the process is optimized investigating the process model and finding structural optimization potentials, e.g., by cross-referencing patterns of best practices [9]. During execution optimization may be changing parameters, e.g., switching web-services. After execution, processes may be adapted using insights gained from execution data [8]. To gather this data, dBPO uses a semiautomated mapping connecting execution data from a process engine and operational data from heterogenous sources. Using this data, parts of the process with high potential for optimization are identified. Using customized data mining techniques, the applicability of patterns is tested, e.g., testing if an ability can be split into distinct variants. These patterns can be applied semi-automatically to the process model depending on the goals set by the analyst. The approach has been implemented using the Business Process Execution Language (BPEL). While dBPO offers sophisticated optimization techniques, it requires a BPEL compliant execution environment as well as operational data from an independent source. In comparison, aPro provides its own method for capturing operational data (i.e., monitoring stubs), which is integrated ex-ante instead of matching data ex-post. Data is then captured as part of process execution and stored in an automatically configured data warehouse, making independent data storage unnecessary.

Similarly to aPro, Wetzstein et al. [10] present a business process intelligence system for monitoring process goals, which are derived from process metrics. Processes are executed in BPEL, monitoring data is gathered by monitoring the services orchestrated within the process. Metrics and goals are stored in a metrics database, which provides data for an analysis step. During the analysis, a decision tree is generated for each process goal depicting the influential factors leading to goal fulfillment or violation. In comparison to aPro, the approach is limited to BPEL processes and the metrics, which can be derived from service calls.

Overall, the related work shows a high degree of maturity in data mining techniques as well as highly integrated approaches covering the complete business process lifecycle. However, in all cases expert knowledge in some domains of the lifecycle is needed. Therefore, we decided to use existing data mining techniques, but provide a new approach for analytics on the modeling level, hiding the implementations of other parts from the business user.

III. CONCEPT

In this section, we summarize the existing aPro architecture, define the requirements for analytics within this context and describe the concept for model-driven business process analytics.

A. aPro overview

Model-driven process analytics build on the established aPro architecture, which is shown in Figure 3. Conceptually, the system is divided. The *model layer* represents a business view of the process and its monitoring. The underlying implementation of the process is hidden. The *aPro layer* contains automatically configured components of the aPro architecture for monitoring. The *execution layer* contains the concrete implementation of the process, which consists of one or more *executing systems*.

On the model layer a *process model* and a *goal model* containing metrics, KPIs and goals are created in a graphical editor, as shown in Figure 1. The process model depicts a simplified real-life claim management proess, as is performed by an insurance service provider many times a day. When a damage claim (e.g., for a damaged vehicle) is received from an insurance company, it is checked using rules to independently calculate the claimed amount. It then is decided, if the claim is valid and what amount is paid out. In the final step, the results are sent back to the insurance company.

Below the process model a goal model is shown in the lower half of Figure 1. The most important elements of the ProGoalML notation are shown in Figure 2. Attached to each process activity is a *measuring point*, indicating a measurement is taken during execution of this activity. It contains the *parameters* to be measured, which are named values of a specific data type. For example, the measuring point *receive_mp* is attached to the activity *receive_claim* and contains three parameters: a *case_id* of type *ID* identifying the current case, a *timestamp (TS)* indicating the time the claim is received and a *claimed_amount* of type *Double (D)* in Euro. Other parameter types are *String (S)*, *enumeration (E)* and *Boolean (B)*.

KPIs are connected by arrows to the parameters (or other KPIs) they are calculated from. For example, the KPI *savings_percentage* is calculated from the parameters *claimed_amount* and *calculated_amount*. Similar to parameters, KPIs have a specific data type. Process goals may be imposed on KPIs or parameters and determine the success of process executions. A goal is a Boolean condition on a KPI or parameter. If it is not fulfilled, compensating actions may be triggered [11]. For example, the goal *five_percent_saved* is imposed on the KPI *savings_percentage*. A special type of goal is a *timing_goal*, which is fulfilled if the time between two measurements is below a specified value. In the example, the timing goal *time_sla* is fulfilled, if a claim is processed within 15 seconds.

These models are stored in a ProGoalML file, which is used as basis for the model transformation. In this step, the components of the aPro architecture are created and/or

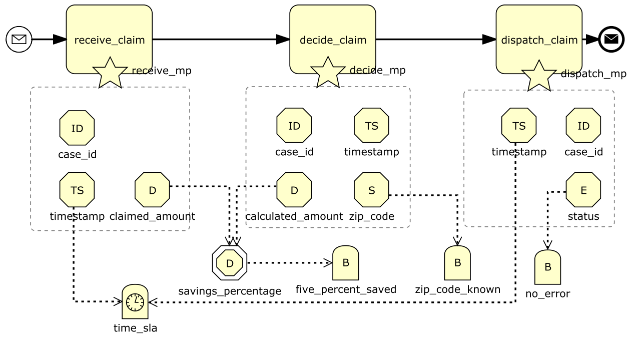


Figure 1. Example process and goal model: claim management process

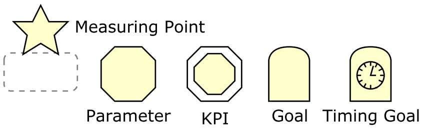


Figure 2. Overview of ProGoalML elements [3]

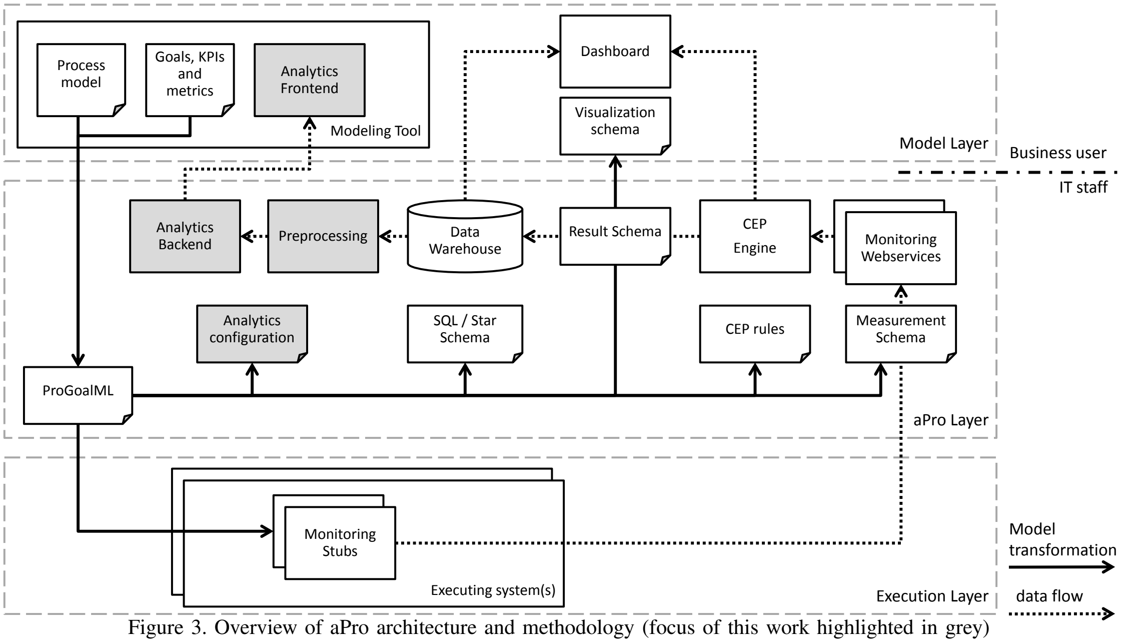


Figure 3. Overview of aPro architecture and methodology (focus of this work highlighted in grey)

configured. Monitoring stubs are created, which are integrated automatically (e.g., in a process engine) or manually (e.g., in a java webservice) in the executing systems to gather measurements. These measurement are sent to *monitoring webservices*, which in turn deliver them to a *CEP engine* for processing. *CEP rules* to correlate measurements of process instances as well as calculate KPIs and goals are generated automatically. One of the results of CEP is an XML file (*result schema*) containing all parameters, KPIs and goals of a process instance. These files are imported into a *data warehouse* for long-term storage. Short-term data from the CEP engine and long-term data from the data warehouse are visualized in a *dashboard*, which is configured using an automatically generated *visualization schema*.

While data visualization is sufficient to monitor system operations, to optimize the process a deeper survey of the data is needed. In the course of this section, we will formulate the requirements as well as the concept for model-driven process analytics.

B. Requirements

The main goal of business process optimization with aPro is the fulfillment of process goals. Thus, the goal of business process analytics is to determine which circumstances lead to the violation of a goal. These circumstances may not readily apparent and cannot be derived in an analytical way from the process model. This suggests the use of knowledge discovery [12] to derive these hidden dependencies from execution data. However, different data mining techniques used in knowledge discovery have several advantages and drawbacks and have to be selected for each problem individually. In this section, we detail the requirements for business process analytics in the context of aPro. These requirements will be used to (a) create a concept and (b) select suitable data mining techniques.

1) **Influential factors to goals** To get a deep insight into the process the influential factors of a goal need to be identified. Which factors aid or hinder goal fulfillment to which extent?
2) **Types of data dimensions:** The data dimensions of measurements and KPIs are not uniformly scaled. The more scales (nominal, ordinal, interval, ratio or absolute) the better the concept is suited.
3) **Model layer presentation:** Due to the model-driven approach, the execution and aPro layers are invisible to the business user. Thus, the analysis results must be contextualized within the process and goal model.
4) **Intuitive comprehensibility:** As a business user is not a data mining expert, the used techniques have to provide results which are comprehensible to him or her.
5) **Visualization of results:** Results need to be visualized in dashboards, diagrams, reports, etc. in a user frontend.
6) **Automatisation:** As the business user is not qualified to adjust parameters, the chosen techniques have to provide acceptable results with default or automatically derived parameters.
7) **Accuracy:** As the target attribute is Boolean the simple rate of correct classified samples is used as a measure for accuracy.
8) **Data warehouse compatibility:** The input data is provided by the data warehouse. The solution has to be compatible with this data structure with only automatic conversions.

C. aPro extension

According to [13], knowledge discovery consists of four steps. During *cleaning and integration (1)*, data is gathered from different sources, correlated and data sets with errors or gaps are corected or excised. During *selection and transformation (2)*, redundant or irrelevant attributes are removed and data is converted in a format fit for *data mining*. Especially, numerical values may be transformed to nominal values, for example by summarizing them to intervals. This process is called discretization and is used because some data mining techniques only work on nominal values [14]. During *data mining (3)*, patterns and insights within the data are discovered.

These are presented to the user during *knowledge presentation (4)*.

Adapting these steps for model driven process analytics within aPro shows that *cleaning and integration* is already performed by existing components. *Monitoring webservices* accept measurement data only in a defined schema sent by *monitoring stubs*. The *CEP* correlates measurements to results spanning process instances, generating complete and unified data sets.

To perform the other steps, additional components are added to the aPro architecture (Figure 3, highlighted in grey).

A *Preprocessing* component reads data from the data warehouse and performs the *selection and transformation* step. Ordinarily, during data selection incomplete datasets are excised. However, in aPro, missing data is not caused by a lack of record-keeping, but rather by a lack of measuring, e.g., if a measuring point is attached to an optional activity. In this case, a lack of measurement indicates the optional activity has not been performed. Thus, missing values are explicitly set to *no_value*. During implementation, we discovered another property of monitoring data in aPro. As goals are always derived from a single attribute (either a parameter or KPI), that attribute is a sufficient influential factor. It alone can be used to determine goal fulfillment. To avoid this problem, during selection all attributes the target attribute is directly derived from are excised. The attributes as well as their data types and dependencies are known from an *analytics configuration*, which is automatically generated from ProGoalML.

The preprocessed data is provided to the *Analytics Backend*. Here, the *data mining* step takes place. Taking into account the requirements defined above, suitable data mining techniques are selected in the following section.

The Analytics Backend offers the implemented data mining techniques as a webservice. That webservice is called by the *Analytics Frontend*, which is integrated in the modeling tool. This allows for a model layer presentation of results, for which the frontend offers a range of visualizations.These visualizations can visualize results from any technique, as long as they are in the correct data format, thus enabling extensibility.

1) **Colorization:** Colors elements of the process and goal model along a gradient. A red/green gradient ist used to visualize relative information gain of metrics, KPIs and goals in relation to a selected goal.
2) **Picture:** Displays a picture rendered by the backend, for example generated decision trees. A picture is linked to a model element. It will be shown when this model element is selected.
3) **InfluenceChart:** A bar chart displaying the top influence factors of a selected goal. In case of the naive Bayes classifier, influence factors are ranked by relative information gain.
4) **DistributionChart:** A bar chart showing the distribution of process instances among values of a selected parameter or KPI (i.e., how many instances have value x) or the distribution of a single value in regards to a selected goal (i.e., how many instances with value x fulfill a goal or not).

The *Analytics Frontend* offers an analysis mode, which locks the process model for edits. Now, one or more techniques for analysis can be chosen. For these techniques, calls to the Analytics Backend are created from the data. As soon as result data is provided, modeling elements can be selected to show results. For example, the selection of a goal can show a picture of a decision tree or color the goal model according to relative information gain from naive Bayes classifier. Visualizations are organized in a dashboard, which allows displaying multiple visualizations side by side and moving them around the model.

After analysis is finished, analytics mode can be deactivated and the model can immediately be adapted.

IV. Selection of data mining techniques

Modern computer science offers a wide variety of data mining concepts. As the target variable is a goal, it is known for all instances. Therefore, process analytics in aPro is a supervised learning problem. For the selection, we chose six data mining techniques to evaluate according to the requirements introduced in section III-B except model layer presentation. These techniques are widely used and described in several works [14][15].

A. Decision trees

Decision trees use the entropy of the different influential factors to build up a sequential classification. A drawback of decision tree is the intransparency of the dependencies between the attributes [16]. The most important advantage of decision trees is the intuitive understanding of decision trees and there high grade of accuracy and automatisation [17].

B. Support vector machines

Support vector machines seperate the data set in two classes by maximizing the distance of the support vectors. Support vector machines are easy to understand if the kernel function is linear. If a non-linear kernel function is used, understanding of results and dependencies is hard. However support vector machines provide high accuracy and can handle with all scales [18].

C. Bayesian networks

Bayesian networks are directed acylic graphs and for each vertice is a conditionally random distribution calculated. Bayesian networks provide a high transparency of dependencies and an intuitive visualization. However, the requirement for automatisation of the analysis is not fulfilled and thus not suitable for business users [18].

D. Naive Bayes classifiers

Naive Bayes classifiers classify the data due to their marginal distributions with the condition that the influential factors are stochastically independent. The assumption that the attributes are stochastically independent is leading to a classifier that fullfills all requirements. The misclassification increases with that assumption but we found the correct classification rate to be still sufficient. Another advantage of the naive Bayes classifier is the intuitive understanding [18].

E. Neural networks

Neural networks have few layers with perceptrons in each layer and the perceptrons transmitting signals through the neural network to activate other perceptrons. Neural networks with their black-box-system make it impossible to identify the dependencies between the target variable and the influential factors. On the other hand, neural networks have a high accuracy [19].

TABLE I. REQUIREMENT FULLFILLMENT OF EVALUATED MACHINE LEARNING CONCEPTS.

Requirements \ Concepts	Decision tree	Support vector machines	Baysian networks	Neural networks	Naive Bayes classifier	Lazy learner
Influence factor dependency	+	o	+	-	+	-
Types of data dimension	o	+	+	+	+	+
Intuitive understanding	+	-	o	-	+	-
Visualization	+	o	o	o	+	o
Automatisation	+	-	o	o	+	o
Accuracy	o	+	+	+	o	o
Data warehouse compability	+	+	+	+	+	+

F. *Lazy learners*

Lazy learner save the whole instances and compare every new instance with the saved instances. Since lazy learners classify instance-based the technique is comprehensible and all scales can be used. However, to identify dependencies between the attributes and the target variable lazy learning is unsuitable.

For each requirement and each classifier we evaluated if the classifier fulfills the requirements fully (+), partially (o) or not at all (-). The table I shows the results of this evaluation. Taking into account the results we chose decision trees and naive Bayes classifiers for the implementation. There are several algorithms that generate decision trees. We chose C4.5 [20] as an algorithm which creates small decision trees to aid intuitive understanding. However, we kept the approach extensible in case other techniques are needed.

V. Implementation

We extended the existing aPro prototype [4] to encompass model driven business process analytics as shown in Figure 3.

In the analytics backend, we utilize WEKA [21], which contains reference implementations of data mining techniques. The preprocessing component reads data from the data warehouse according to the structure implied by the ProGoalML file and converts it to WEKA-compliant input formats.

The frontend is based on the Oryx Editor [22] and existing dashboard components for visualization. Figure 4 and Figure 5 shows the analytics frontend with results from simplified real-life data.

VI. Evaluation

We will evaluate the approach using two business processes, evaluating (a) if the chosen algorithms provide suitable results and (b) if the visualizations can communicate these results. The first process is a blood donation process [23] and the second is a real-world claim management process. The first process was the training process evaluated using synthetically generated data. The data set of the blood donation process has seven dimensions and one target variable. Data was generated by a test data generator. For the second process we had 48 dimensions in the data set, one target variable and 5718 cases. We used analytics to find the dependencies hidden in the generated data sets. The second process we evaluated with anonymized real-life data as the test process. As a metric we used the plain accuracy and made a 10 cross-validation. The different techniques are validated with the differently prepared data sets III-C. Learning accuracy was evaluated by 10-fold cross validation.

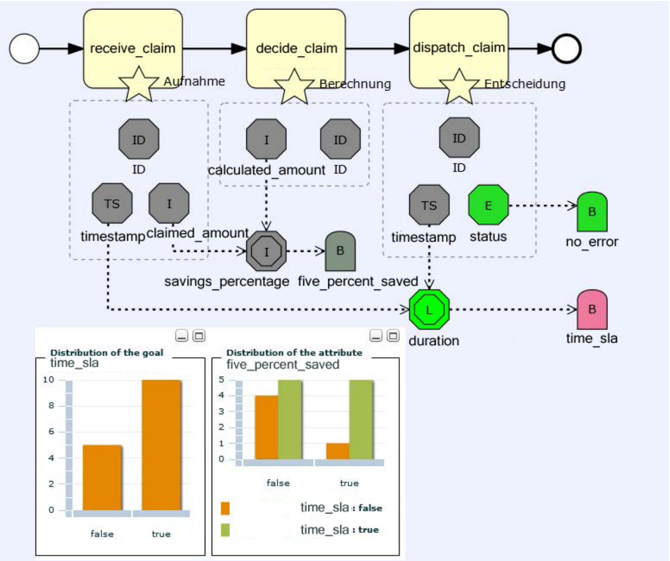


Figure 4. Analysis of example process with simplified real-life data (translated from German)

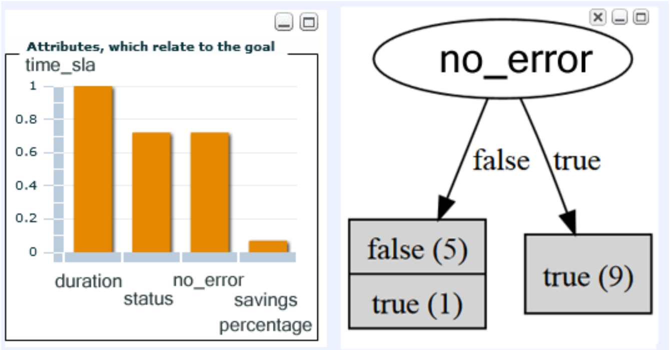


Figure 5. Influence factors and decision tree for the goal time_sla (translated from German)

A. *Blood donation process*

Most of the generated dependencies were found by the naive Bayes classifier. Other found dependencies were not explicitly generated but found to be emergent from the process simulation. The extreme cases (direct dependency or random numbers) were correctly classified. The accuracy of the classifier was high (87%). We modified the generation of the data set and the information gain reflected all modifications. It was possible to identify the strong dependencies with a high information gain. The increase or decrease of the information gain is a stable and robust measure. The decision tree found the direct KPIs or parameters for a goal. Therefore it was necessary to exclude these KPIs and parameters as described in Section III-C. After the exclusion of these KPIs and parameters the accuracy increased. Using discretization the tree size shrank without a significant decrease in the accuracy. The accuracy of the decision trees was less high (61%). Reasons for this low classification rate could be the low base line or the absence of high dimensions in the data set.

B. *Claim management process*

The discretization of the data set increased the accuracy of the naive Bayes classifier. If the data set was prepared the classification rate was slightly better. Overall the highest accuracy of 96,1% of the naive Bayes classifier occurs with a discretization without preselection. The best results with decision trees were with a discretization together with a information gain ratio data selection and a discretization without a selection. We use this configuration as a default in the prototype. The worse performance of the decision trees arised from the low number os attributes in the blood donation process data. The real world claim management process has hundreds of parameters, KPIs and goals. We evaluated rules for fraud and irregularity detection and retired over half of the rules with low information gain as well as found several new rules from decision trees, evaluating an anonymized dataset of 98167 cases.

The naive Bayes classifier fulfills all requirements and has an accuracy of 87% in the synthetically generated data set. The accuracy in the real-world data set even higher (96,1%). Furthermore the information gain is a robust and stable measure for dependencies in aPro. Decision trees as an analytics concept are only partially suitable for aPro. Decision trees represent the most strongest dependencies and neglect other influential factors because it is sufficient to classify the instances with these stronger dependencies. The accuracy of decision trees in aPro was 61% in the synthetical data set and 90,7% in the real-world data set.

Overall the combination of decision trees and naive Bayes classifier are suitable for model-driven process analytics because decision trees allow a quick overlook over the most strongest dependencies in the data sets and the naive Bayes classifier complements this overview with high accuracy and deeper insights into the dataset.

VII. Conclusion and outlook

In this work, we presented an approach for model-driven business process analytics using the aPro architecture. The main contribution is the support of diverse process execution environments, while still providing out-of-the-box analytics for business users, hiding both monitoring and implementation layers. We defined criteria for data mining techniques and selected two exemplary techniques. The concept proved sufficient during evaluation and real-world use. However, support for more data mining techniques and different configurations may provide better insights, while still preserving ease-of-use for business users.

In future work, we would like to extend the approach to process adaptation as well as compliance scenarios. As process goals can be used to monitor compliance requirements, reporting and root cause detection for compliance violations can be provided [24].

Acknowledgment

The work published in this article was partially funded by the Co.M.B. project of the Deutsche Forschungsgemeinschaft (DFG) under the promotional reference SP 448/27-1.

References

[1] N. Slack, S. Chambers, and R. Johnston, Operations management. Pearson Education, 2010.

[2] M. zur Mühlen and R. Shapiro, “Business process analytics,” in Handbook on Business Process Management 2. Springer, 2010, pp. 137–157.

[3] F. Koetter and M. Kochanowski, “Goal-oriented model-driven business process monitoring using progoalml,” in 15th BIS. Vilnius: Springer, 2012, pp. 72–83.

[4] ——, “A model-driven approach for event-based business process monitoring,” Information Systems and e-Business Management, 2014, pp. 1–32.

[5] M. Kintz, “A Semantic Dashboard Description Language for a Process-oriented Dashboard Design Methodology,” in Proceedings of 2nd International Workshop on Model-based Interactive Ubiquitous Systems (MODIQUITOUS 2012), Copenhagen, Denmark, 2012.

[6] F. Koetter, M. Kochanowski, M. Kintz, T. Renner, and P. Sigloch, “Model-driven data warehousing for service-based business process monitoring,” in Global Conference (SRII), 2014 Annual SRII. IEEE, 2014, pp. 35–38.

[7] M. Castellanos, F. Casati, U. Dayal, and M.-C. Shan, “A comprehensive and automated approach to intelligent business processes execution analysis,” Distributed and Parallel Databases, vol. 16, no. 3, 2004, pp. 239–273.

[8] F. Niedermann and H. Schwarz, “Deep business optimization: Making business process optimization theory work in practice,” in Enterprise, Business-Process and Information Systems Modeling. Springer, 2011, pp. 88–102.

[9] F. Niedermann, S. Radeschütz, and B. Mitschang, “Business process optimization using formalized optimization patterns,” in Business Information Systems. Springer, 2011, pp. 123–135.

[10] B. Wetzstein, P. Leitner, F. Rosenberg, I. Brandic, S. Dustdar, and F. Leymann, “Monitoring and analyzing influential factors of business process performance,” in Enterprise Distributed Object Computing Conference, 2009. EDOC’09. IEEE International. IEEE, 2009, pp. 141–150.

[11] F. Koetter, M. Kochanowski, M. Kintz, and I. Fraunhofer, “Leveraging model-driven monitoring for event-driven business process control,” in 1. Workshop zur Ereignismodellierung und verarbeitung im Geschaeftsprozessmanagement (EMOV), 2014, pp. 21–33.

[12] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth et al., “Knowledge discovery and data mining: Towards a unifying framework.” in KDD, vol. 96, 1996, pp. 82–88.

[13] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011.

[14] I. H. Witten and E. Frank, Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2005.

[15] S. Theodoridis and K. Koutroumbas, “Pattern recognition,” IEEE Transactions on neural networks, vol. 19, no. 2, 2008, p. 376.

[16] S.-M. Lee and P. A. Abbott, “Bayesian networks for knowledge discovery in large datasets: basics for nurse researchers,” Journal of Biomedical Informatics, vol. 36, no. 4, 2003, pp. 389–399.

[17] S. Tufféry, Data mining and statistics for decision making. John Wiley & Sons, 2011.

[18] C. M. Bishop et al., Pattern recognition and machine learning. springer New York, 2006, vol. 1.

[19] J. Schmidhuber, “Deep learning in neural networks: An overview,” CoRR, vol. abs/1404.7828, 2014, last accessed 29.01.2015. [Online]. Available: http://arxiv.org/abs/1404.7828

[20] J. R. Quinlan, C4. 5: programs for machine learning. Morgan kaufmann, 1993, vol. 1.

[21] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The weka data mining software: an update,” ACM SIGKDD explorations newsletter, vol. 11, no. 1, 2009, pp. 10–18.

[22] G. Decker, H. Overdick, and M. Weske, “Oryx — An Open Modeling Platform for the BPM Community,” in Proceedings of the 6th International Conference on Business Process Management. Heidelberg: Springer, 2008, pp. 382–385.

[23] D. Schleicher, C. Fehling, S. Grohe, F. Leymann, A. Nowak, P. Schneider, and D. Schumm, “Compliance domains: A means to model data-restrictions in cloud environments,” in Enterprise Distributed Object Computing Conference (EDOC), 2011 15th IEEE International. IEEE, 2011, pp. 257–266.

[24] F. Koetter, M. Kochanowski, A. Weisbecker, C. Fehling, and F. Leymann, “Integrating compliance requirements across business and it,” in Enterprise Distributed Object Computing Conference (EDOC), 2014 IEEE 18th International. IEEE, 2014, pp. 218–225.

Development of a Distributed Cryptographic System Employing Parallelized Counter Mode Encryption

Dev Dua, Sujala D. Shetty, Lekha R.Nair
Department of Computer Science
BITS Pilani, Dubai U.A.E.
Emails: {devdua@live.com, sujala@dubai.bits-pilani.ac.in, lekharnair@gmail.com}

***Abstract*— In the current era of Cloud Computing, Big Data and always-connected devices and apps, it is critical that the data that reside on networks be secured by robust techniques and protocols. As the data are highly distributed and voluminous, the security measures should employ methods that efficiently and rapidly encrypt data. This calls for scaling the model up to employ a distributed programming paradigm, so that the utilization of resources (computing and storage) on the network is high and channeled for processing in an optimum way. Exploring the idea of distributed cryptography might hold solutions to address this potential problem. We have tried to probe in this direction by building a distributed and flexible cryptosystem that employs a parallelized version of the AES algorithm. The results obtained emphasize the performance gains, agility, flexibility and scalability of the concept of cryptography using distributed systems.**

Keywords-Big Data; Cluster Computing; Distributed Systems; Security; Cryptography; MPI.

I. INTRODUCTION

In the past few years, the concept of Cloud Computing [13] has gained a lot of traction, and is seeing a high adoption rate. This concept, wherein a third party provides networked hardware, computing platforms and application software as a service accessible via the Internet to its customers is more commonly given the generic term of “cloud”. In most cases the exact location of data stored is unknown to the customers which raises concerns over the privacy of the data on the external network. Security of data and their integrity is a crucial concern that most organizations have when moving their data to a cloud. Since cloud computing is relatively a recent trend, encryption in the cloud is still in its early stages. Only a few cloud providers one of them being Microsoft Azure provision encryption of data stored in their data centers [12]. The primary objective of encryption in the cloud is to deter unauthorized access, as access to sensitive data (without the knowledge of the owner) by non-permitted entities is a privacy violation. The focus of this paper is on how encryption can be adapted to utilize the virtually infinite amount of resources that the cloud provides, and not Key Management, which is another crucial aspect of a cryptosystem.

The cloud model being considered here is a private/hybrid cloud. The public cloud model provides services and infrastructure over the Internet, from a location that is external to the customers’ computing environment. Since resources are shared in this model, this commonality of storage/processing space makes them more vulnerable to data protection issues than private clouds. Private clouds are basically the same in infrastructure and software as public clouds, and they differ in the aspect that the network that powers and interconnects the systems in the private cloud is exclusive to the organization that houses the cloud. Only authorized users belonging to the organization domain can make use of the private cloud. Hybrid clouds provide a balance between the other two models, as one can even combine services provided by different cloud providers to keep each aspect of the organization’s operation process efficient and streamlined. However, one of the drawbacks is the hassle to orchestrate security platforms of all providers to play with each other. Hence, it is recommended to keep critical data safe in the private side of the hybrid cloud, and use strong security policies to protect the data in-house.

Thus, this paper assumes that the data to be kept safe in the data storage facilities has already been transmitted there using a Secure Sockets Layer encrypted connection or a connection protected by Transport Layer Security. The cryptosystem aims to operate as soon as the data reaches the cloud, so that the window between the unencrypted and encrypted states of the data is as small as possible.

Parallelized encryption methods have been explored in the past [6][10][11] and researchers have implemented parallel encryption/decryption algorithms on single machines with multi-core CPUs. This leads to an overall speedup on the time spent on encryption/decryption on such machines when compared to single process, sequential execution of the cryptography technique. However, the parallelizable algorithms can be scaled to the next level, by allowing them to run on distributed systems, so that the number of processes employed during the computation can be drastically increased. This would allow for noticeably faster and efficient encryption and decryption of voluminous, stagnant files. Since the data being encrypted will be distributed in segments over a net-work, it has a higher level of security by virtue of the randomness of its location.

The objective of this paper is to meet the following:

- Build and setup a distributed homogenous computing cluster with 3 compute nodes
- Create a native cryptographic platform that makes use of the nodes to encrypt/decrypt files in a parallelized manner.

ISBN: 978-1-61208-412-1

- Analyse the performance of the parallelized encryption and decryption code on the cluster by varying the number of slave processes operating on varying file sizes.
- Further explore possible extensions to the project aims to efficiently ensure security of the platform by deploying user access control, more control in terms of options and features to encrypt data, and scheduling regular refreshes to the encrypted state.

II. MPI AND PARALLELIZATION OF CODE

A. Message Passing and Message Passing Interface

Message Passing [1][3][4] is one of the types of inter-process communication, which is employed in object oriented programming and parallel programming. Message passing involves sending functions, signals and data packets from one process to either all other processes or a specific process executing the Message Passing Interface (MPI) job. The invoking program sends a "message" to the object. This standard contrasts from routine programming [2] wherein a method, process or a subroutine is specifically called by name. One of code libraries that have been consented by the MPI Forum is the Message Passing Interface Standard (MPI) [3], and is widely supported by users, software library developers, researchers and even vendors. The Message Passing Interface was built to develop a flexible, standard and portable message passing standard to be extensively used in programs that employ message passing.

B. Parallelization of the AES algorithm

To adapt the cryptosystem to make full use of the cluster, the most apt cryptography mode had to be determined so that the cipher can optimally encrypt/decrypt data. The Counter (CTR) mode [9] was chosen, due to the reason that unlike traditional encryption modes, the CTR mode encrypts data by converting the underlying block cipher into a stream cipher. It generates subsequent key stream blocks by encrypting successive values of a counter, which can be any sequence generating function that produces a sequence which is guaranteed not to repeat for a long time. The most commonly used counter, due to its popularity and simplicity is a simple increment-by-one counter.

The CTR mode encrypts blocks of data in a way that encryption of a block is totally independent of the encryption done on the previous block. This enables efficient use of hardware level features like Single Instruction Multiple Data (SIMD) [14] instructions, large number of registers, dispatch of multiple instructions on every clock cycle and aggressive pipelining. This makes CTR mode encryption to perform much better and faster than a CBC (Cipher Block Chaining) mode encryption. Using the CTR mode, encryption can be easily parallelized as each block can be independently encrypted, thus using the entire power of the underlying hardware.

After the plaintext is received by the cipher, encryption can then be carried out by using the pre-computed number sequence (if pre-processing of Initialization Vectors is used). This can lead to massive throughput, of the order Gigabits per second. As a result, random access encryption is possible using this mode, and is particularly useful for encryption of Big Data, where the plaintext are chunky blocks of huge amounts of data. A unique characteristic of the CTR mode is that the decryption algorithm need not be designed, as the decryption has the same procedure as encryption, differing only in terms of initialization of the Initialization Vector (IV). This fact holds most weightage in case of ciphers such as Advanced Encryption Standard (AES), used in this project, because the forward "direction" of encryption is substantially different from traditional inverse mapped "direction" of decryption, which has no effect on the CTR mode whatsoever. An additional benefit of the CTR mode is the lack of need to implement key scheduling. All these factors contribute to a significant yield in hardware throughput, by making the design and implementation simpler.

III. PROPOSED DESIGN FOR THE CRYPTOSYSTEM

The CTR mode of cryptography has similar characteristics to Output Feedback (OFB) [8], but also allows a random access property during decryption, which is quite suitable for this project. CTR mode is well suited to operate on a multi-processor machine where blocks can be encrypted in parallel. However, since the data and processing in a cluster is distributed, the way in which distributed encryption can be achieved needs to be determined. One way in which the CTR mode is easily parallelizable across nodes is by letting each process have its own counter, and by virtue of this, its own Initialization Vector (IV). The approach in our case is different compared to [6] in terms of the extent to which the processing power has been utilized. However the file splitting is somewhat similar in approach.

The IVs and the counters should be initialized at runtime, depending on the number of processes attached to the MPI job, which is accessed using MPI_COMM_SIZE. The file to be encrypted is read in parallel as *n* equal and distinct parts, where n is equal to MPI_COMM_SIZE. Each process, after reading the portion of the file that has been assigned to it, encrypts it by passing substrings of size AES_BLOCK_SIZE to the AES_ctr128 method of the OpenSSL library, and writes the encrypted data to a file that contains the block of data encrypted by that process. Thus, after encryption, *n* encrypted blocks along with *n* IVs are created, which reside in the machine that generated them.

A. Setup of the experimental environment

To implement and test the cryptosystem, a Beowulf cluster [5] built using 3 compute nodes, which are regularly indistinguishable, was setup and connected in a 100M network over password-less SSH [7] and libraries and frameworks required by the cryptosystem like unison, libcrypto, openssl-dev and OpenMPI 1.6 were installed. Each compute node in the Beowulf Cluster created to collect results had the following specifications:

CPU : Intel Core i7 4770 (4 cores X 3.40 Ghz)
Memory : 8GB
Operating System : Ubuntu 14.04 LTS

BASH scripts were used to run the scripts used to control the cluster. All programs that were run on the MPI ORTE were written in C. OpenMPI was used to provide a parallel environment to the code. Unison was used to synchronize the source code and executables across the compute nodes. The 128bit CTR implementation of the AES algorithm available in the OpenSSL library was used in the cryptosystem.

For encryption, the file path and the number of processes are passed as command-line arguments to the cryptosystem. The script then synchronizes the file among all available nodes, so that a copy of the file under encryption is available locally on the HDD of every node. This is done to minimize I/O overhead, and the network overhead involved in this approach is minimal compared to the performance gains expected. The master process evaluates the file and pads the end of the file with bytes if the file size is not a multiple of the number of processes specified for use. Each node then operates on the file with 4 processes per node, with each process generating an encrypted block and iv as separate files. The display names of the files themselves are generated in a way that the file names of related encrypted blocks (blocks generated from the same unencrypted file) seem totally unrelated to each other.

During decryption, the cryptosystem recognizes the blocks to be decrypted by means of the file name passed via command line arguments and then attempts to decrypt those blocks. Decryption succeeds if the number of processes specified matches the number of processes that were used to encrypt the file. The master process then accumulates the decrypted blocks and creates a decrypted file with the padding removed (if any).

B. *Advantages of the cryptosystem and Impact on security*

While thinking of the above design, there were concerns about the security and integrity of the approach. The advantages outweigh the issues, which can be improved upon as explained in Section V.

- Counters in the CTR mode are usually limited to the value 2^{64}-1, which then has to be reset once the counter of the code reaches this value to avoid overflow. However, in the proposed design, each process has a counter of its own with the above limit, so an increase in the number of processes that are running the engine simultaneously linearly increases the amount of data that can be encrypted/decrypted. Thus with *n* processes, the amount of data that can be encrypted/decrypted increases *n* fold.
- Data encrypted by *n* processes can be decrypted if and only if *n* processes are used to decrypt it. Any mismatch in the number of processes will lead to an Exception or files with junk characters. This ensures that the data is undisclosed and confidential as only the party that encrypts the data would know the number of processes that were chosen to encrypt the data. This adds an extra security parameter apart from the secret key. Brute-forcing will take longer as well, as the master key is 10 bytes long (8 bytes for the AES key and 2 bytes for the number of processes).
- Since process ranks are arbitrarily decided at runtime, the splitting and transmission of data that occurs during encryption is completely random. Thus, data being encrypted is spread across the compute nodes, which increases entropy of location of the data. One cannot determine with surety the exact location of every block being generated by the cryptosystem.
- As the blocks being generated are much smaller than the original unencrypted file, the network overhead is also minimized.
- The system built is highly portable and extensible, with very less setup involved. This scalability makes the cryptosystem highly compatible with the scaling capabilities offered by the cloud environment it is hosted on.
- Since OpenSSL is being at the very core of the cryptosystem, the security of the entire system in general is enforced. The core crypto library of OpenSSL provides basic cryptographic functions that are highly supported by the network security community and provides an abundantly accepted set of encryption implementations. Any vulnerabilities in the library are updated fairly often, and thus the cryptosystem is free of core security issues.

C. *Description of the test files*

The files used to measure the total execution time of the algorithm were generated by using the native *dd* command provided by Linux, and contained randomized data with the text file having a fixed sizes. A random stream was generated by Linux, which was subsequently captured by the *dd* command till the stream filled in the file of the specified size. The sizes of the aforementioned files were 100 KB, 200 KB, 500 KB, 1MB, 2.5MB, 5MB, 10MB, 20MB, 40MB, 50MB, 100MB, 500MB, and 1GB.

IV. Performance Analysis

As can be seen from Figure 1, the most inefficient use of the cluster is in the first case where the number of processes (nP) is 1. It can be safely assumed that the time taken by the cluster to execute parallel code using just 1 process will be close to executing a serial version of the code on a uniprocessor. Also, as the file size increases, there is a sharp increase in execution time, which is largely due to high process idle times. Hence it is meaningless to use just 1 process to encrypt a big file. On the other end, the most efficient use of the cluster is when it is used to maximum capacity, i.e., 12 processes. A maximum performance increase of 6.7x is obtained when the cluster is optimally used, in the case of encryption of the 1GB file. The throughput obtained here is 4.233 Gbps. It can also be seen from the graph as the number of processes allocated to the encryption sequence is increased, the time vs. size graph tends to become flatter and linear.

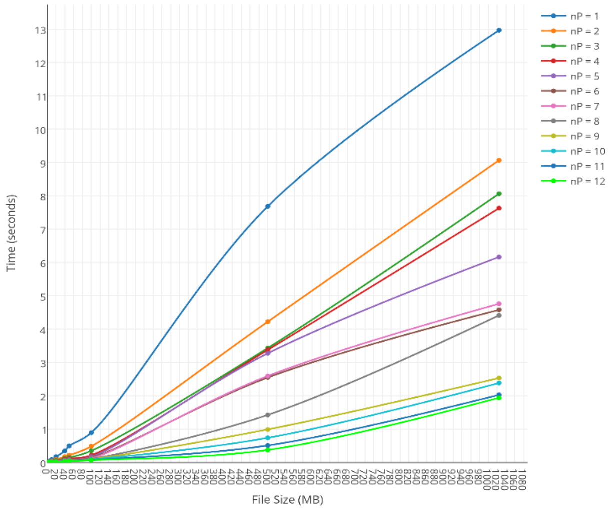


Figure 1. Execution time (s) vs. File size (MB) for encryption

The similarity between Figure 1 and Figure 2 is apparent, courtesy of the nature of CTR mode cryptography. Most of the code in the decryption process is the same as the encryption process, and only slight variations in execution times are observed. The performance speedup obtained upon operation on the 1 GB file by using 12 processes in this case is only 6.378x compared to 6.7x in the case of encryption. This is solely due to the conglomeration operation at the end of the process, which consumes time in stitching the decrypted blocks together. Another pattern observed is that the curves obtained when the number of processes are increased tend to group together, indicating that O(n) times are possible as the number of processes goes up.

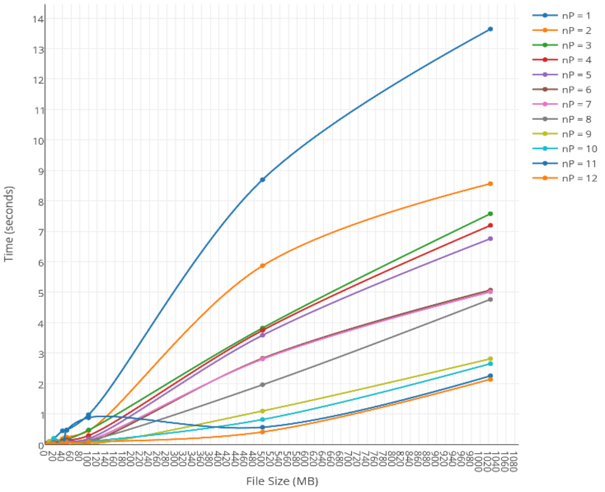


Figure 2. Execution time (s) vs. File size (MB) for decryption

V. Possible Improvements to the Cryptosystem

Several hardware level and implementation improvements can be made to the cryptosystem as the one designed here is not production ready, and is a prototype to validate the design concepts. Some improvements include:

- Using an SSD array for storage of the files would lead to drastically improved performance, as the I/O capabilities of SSDs are much better compared to conventional HDDs
- The processors used for testing were consumer machine grade, and not high performance processors typically used by server farms. Hence, better processors could be used
- The network on which the system was tested was a 100M LAN, however the network can be upgraded to a Gigabit connection and a fast Switch can be used exclusively for the compute nodes in the cluster
- The Scatter/Gather pattern of communicative computation provided by Open can be employed to

ensure robustness of the system, and allow for better handling of processes to avoid process idle times

- In-memory storage can be used to store the files and restrict hard drive reads and writes. This can lead to significant improvements in execution times, and reduce I/O overhead
- Reduction in network data transmission and file generation to disclose as little data as possible, as well as avoid chances of snooping and eavesdropping

VI. CONCLUSION

Clusters are easy to setup using OpenMPI. However, the network being used to connect the compute nodes has to be reliable and sufficiently fast to reduce the communication time. The security of the platform can be further increased by scheduling encryption at periodic intervals, in a real world scenario, so that chances of attacks are less, and the safety of data is uncompromised. The computing power of the cloud can also be harnessed to create virtualized clusters, with a fixed number of nodes, but scalable in terms of configuration of the nodes. So, this could result in a low performing cluster with 3 compute nodes with just 3 logical processors and 3GB of RAM, to a high performance 3 node cluster with 48 logical processors and 168 GB of RAM. Furthermore, the cluster can be connected to a REST Web API that can accept files to be encrypted, run encryption on them, and return a link to the encrypted versions.

This shows the flexibility of the platform designed in this project, and also proves the ease by which a cluster can be created in just a few hours. Also, since the tools used are open source and are industry leading solutions with a lot of community support, the maintenance of the cluster is minimal. This makes the cluster fault tolerant in a way, as well as highly extensible.

REFERENCES

[1] Goldberg, Adele, David Robson (1989). *Smalltalk-80 The Language*. Addison Wesley. pp. 5–16. ISBN 0-201-13688-0.

[2] Orfali, Robert (1996). *The Essential Client/Server Survival Guide*. New York: Wiley Computer Publishing. pp. 1–22. ISBN 0-471-15325-7.

[3] Blaise Barney, Lawrence Livermore National Laboratory, *Message Passing Interface,* URL : https://computing.llnl.gov/tutorials/mpi/, Retrieved: October 2014

[4] Extreme Science and Engineering Discovery Environment (XSEDE) Project, National Center for Supercomputing Applications (NCSA), *Introduction to MPI*, University of Illinois at Urbana-Champaign. URL: https://www.xsede.org/high-performance-computing, Retrieved: November 2014

[5] Donald J Becker, Thomas Sterling, Daniel Savarese, Johne E Dorban,; Udaya A Ranawak and Charles V Packer, "*BEOWULF: A parallel workstation for scientific computation"*, in Proceedings, International Conference on Parallel Processing vol. 95, (1995). URL: http://www.phy.duke.edu/~rgb/brahma/Resources/beowulf/papers/ICPP95/icpp95.html, Retrieved: October 2014

[6] Ozgur Pekcagliyan and Nurdan Saran, *Parallelism of AES Algorithm via MPI*, 6th MTS Seminar, Cankaya university, April 2013 URL: http://zgrw.org/files/mpi_AES.pdf, Retrieved: November 2014

[7] Hortonworks, *HOWTO: Generating SSH Keys for Passwordless Login,* URL: http://hortonworks.com/kb/generating-ssh-keys-for-passwordless-login/, Retrieved: October 2014

[8] ISO/IEC 10116:2006 - Information technology -- Security techniques – *"Modes of operation for an n-bit block cipher"*, URL: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=38761, Retrieved: December 2014

[9] Helger Lipmaa, Phillip Rogaway, Chiang Mai and David Wagner, Helsinki University of Technology (Finland) and University of California at Davis (USA) and University of Tartu (Estonia) University (Thailand), University of California Berkeley (USA), *"CTR-Mode Encryption",* URL: http://csrc.nist.gov/groups/ST/toolkit/BCM/documents/workshop1/papers/lipmaa-ctr.pdf, Retrieved: December 2014

[10] Deguang Le, Jinyi Chang, Xingdou Gou, Ankang Zhang and Conglan Lu, *"Parallel AES algorithm for fast Data Encryption on GPU"*, 2nd International Conference on Computer Engineering and Technology (ICCET), 16-18 April 2010 vol.6, no., pp.V6-1,V6-6, doi: 10.1109/ICCET.2010.5486259
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5486259&isnumber=5485932, Retrieved: November 2014

[11] M. Nagendra and M. Chandra Sekhar, Department of Computer Science & Technology, Sri Krishnadevaraya University, Anantapuramu, India, "*Performance Improvement of Advanced Encryption Algorithm using Parallel Computation",* International Journal of Software Engineering and Its Applications Vol.8, No.2 URL: http://www.sersc.org/journals/IJSEIA/vol8_no2_2014/28.pdf, Retrieved: November 2014

[12] Microsoft Azure Trust Center: Security URL: http://azure.microsoft.com/en-us/support/trust-center/security/, Retrieved: October 2014

[13] Mladen A. Vouk, Department of Computer Science, North Carolina State University, Raleigh, North Carolina, USA, *"Cloud Computing – Issues, Research and Implementations"*, Journal of Computing and Information Technology - CIT 16, 2008, 4, 235–246, doi:10.2498/cit.1001391, URL: http://hrcak.srce.hr/file/69202 Retrieved: October 2014

[14] William Gropp, Mathematics and Computer Science Division Argonne National Laboratory Argonne, IL 60439, USA, *"Tutorial on MPI: The Message-Passing Interface"*, URL: https://www.idi.ntnu.no/~elster/tdt4200-f09/gropp-mpi-tutorial.pdf, Retrieved: October 2014

On the Power of Combinatorial Bidding in Web Display Ads

Khaled Elbassioni[1] and Mukesh Jha[2]

Masdar Institute
Abu Dhabi, UAE
Email: kelbassioni@masdar.ac.ae[1],
Email: mjha@masdar.ac.ae[2]

***Abstract*—Web display advertisement is occupying a major part of the marketing industry. With the vastly increasing number of competing advertisers, allocating web advertising space becomes a challenge, due to the need to run a large-scale auction in order to determine the winners and payments. As a result, some of the most desired properties, such as *truthfulness* (a.k.a. *strategy proofness*), and *social welfare maximization* are typically sacrificed by the auction mechanisms used in practice, in exchange of computational efficiency. Furthermore, those mechanisms typically assume a *fixed partition* of the advertising space, and use a simple mechanism, such as *Generalized Second Price (GSP)* auction to avoid the combinatorial explosion of the size of the problem when users are allowed to bid on *arbitrary regions of the space*. In this paper, we go beyond this non-combinatorial approach, and investigate the implementation of strategy–proof mechanisms which are truthful–in–expectation and approximately socially efficient, with an attempt to understand their computational efficiency, social welfare and revenue, when applied to Web display Ads. Those mechanisms were proposed recently as a theoretical engineering of the computationally less efficient mechanisms of Lavi and Swamy. Our experimental results show that allowing combinatorial bidding can offer substantial improvements in both social welfare and revenue as compared to slot-based methods, such as the GSP mechanism.**

Keywords*—*Combinatorial auctions; algorithmic mechanism design; Generalized Second Price; truthfulness.

I. INTRODUCTION

Internet advertising is one of the most important marketing tools due to its growing number of audience. Internet advertising revenues in the United States totaled $42.8 billion for 2013, with an increase by 17% over the previous year [1]. One can basically distinguish two types of online advertisement: (I) Search advertising, which typically appears on search web pages, usually based on keywords searched by user, e.g., Google, Yahoo!; (II) Display advertising, which typically appears on non-search web pages, usually based on content and type of web-page, e.g., news sites, airlines sites, social networks sites, etc. [2].

In both types of advertisement, auctions are typically used to determine which advertisement will be displayed. In contrast to Search advertising, display advertising does not need a real-time auction mechanism and can thus be done offline, allowing for a substantially larger processing time. This paper will be concerned only with the second type.

Algorithmic mechanism design (AMD) studies optimization problems in which part of the input is not directly available to the algorithm; instead, this data is collected from self-interested agents. who can manipulate the algorithm by misreporting their parts of the input, if that would improve their own objective functions. It is therefore desirable to design a protocol or a *mechanism* which interacts with the agents so that their selfish behaviour yields a globally desirable outcome. Adding to this the requirement of *computational efficiency*, AMD quests for *efficient* algorithms that *(approximately) optimize* a global objective function (usually called *social welfare*), subject to the *strategic requirement* that the best strategy of the agents is to *truthfully* report their part of the input. Such algorithms are called *incentive compatible* or *truthful mechanisms*.

If the underlying optimization problem can be efficiently solved to optimality, the celebrated *VCG mechanism* [3] achieves truthfulness, social-welfare optimization, and polynomial running time. In general, and more specifically in the display Ad auctions with a relatively large number of advertisers, the underlying optimization problem can only be solved approximately. Lavi et al. [4][5] showed that certain linear programming based approximation algorithms for the social welfare problem can be turned into randomized mechanisms that are truthful-in-expectation, i.e., reporting the truth maximizes the expected utility of an agent. The Lavy-Swamy (LS)-reduction is powerful [4]–[7], but unlikely to be efficient in practice because of its use of the *Ellipsoid method* for linear programming. In fact, we are not aware of any attempt to apply the LS-approach in practice or at least to perform a systematic study of its applicability and effectiveness, compared to the mechanisms which are currently being used.

Presently Google and Yahoo! are using Generalized-Second-Price (GSP) auction mechanism to auction off slots. GSP looks somewhat similar to VCG but its properties and equilibrium behavior are quite different. Unlike the VCG mechanism, GSP is *not* truthful [8], but is by far computationally more efficient. On the other hand, a major drawback of GSP when applied to Display Ad auctions is that it is inherently *slot-based*, that is, the advertising space has to be apriori partitioned into fixed slots, which are auctioned off in a way similar to keyword auctions. This has the obvious disadvantage of limiting the bidding power of the agents,

ISBN: 978-1-61208-412-1

which could be otherwise exploited to increase the total social welfare of the agents and/or the revenue due to the auctioneer. Allowing combinatorial bidding, where agents can bid on different regions (bundles) of the advertisement space can offer substantial improvements in both social welfare and revenue, provided that there are computationally efficient mechanisms that can deal with this kind of bidding.

Very recently, Elbassioni et al. [9] gave an efficient implementation of the LS-approach based on the simpler *multiplicative weights update (MWU) methods*. The simplification comes at the cost of slightly weaker approximation and truthfulness guarantees. In this paper, we investigate the effectivity of their method when applied to display advertisement auctions. We carry out extensive experiments to compare these methods with GSP in terms of truthfulness, social welfare and revenue. Our experiments show that the proposed implementation can handle auctions involving hundreds of bidders and thousands of bundles, and offer substantial improvements in both social welfare and revenue as compared to slot-based methods, such as the GSP mechanism.

In Section II, we discuss about Internet advertisement, auction, drawbacks of non-truthful auction mechanisms and its impact on online advertisement. In Section III, we explain the outline of the problem and algorithmic solutions. We also define relevant terminology and definitions. In IV, we elaborate our experimental set-up. We discuss our evaluation process in Section V. We present our results and analysis for social-welfare, revenue and running time in Section VI. We compare the VCG-based mechanism with GSP mechanism in Section VII. Finally we conclude our work in Section VIII.

II. Related Work

Online advertisement dates back to at least 1994 when HotWire started to sell banner ads to advertisers [10]. After a small hiatus during the period of dot-com crash, online advertisement along with online auctions became one of the major form of exchange of items and services over the Internet. GoTo.com introduced the use of auction mechanisms for allocating advertising space on web pages showing results of the query to generate revenue [10]. Presently, most of online advertisement uses some form of auction based on keywords, combination of keywords, or space [8][10][11].

Overture (now a part of Yahoo!) operated a first-price auction in early 2000s for search advertisement. Since this form of auction is not truthful, bidders used to bid strategically to increase their utility [12]. Furthermore, there was a loss in revenue and the market was unstable due to frequently lying bidders. Hence, VCG-based mechanisms were suggested to stabilize auction outcomes. Since, these mechanisms require solving an NP-hard winner determination problem, a variant of second-price auction came in practice, the so-called Generalized Second Price (GSP) auction. Yahoo! and Google AdWords use the GSP mechanism. This mechanism is non-truthful for multiple items, but it has envy-free equilibrium [8]. Although it is better than the first-price auction and widely used, various researches showed that the GSP mechanism has its own flaws. In particular, the bidders maybe forced to undertake complicated tasks of choosing a bidding strategy to increase their utility. Asdemir [13] and Edelman et al. [12] showed that the GSP mechanism might result in bidding wars cycles and static bid patterns are frequently observed. The strategy of bidders to outbid each other until one of them drops their bid and the other one follows by dropping its bid to just by very small ϵ above the competitor's bid is known as Bidding war cycle. Static bid patterns is defined as the bidder's fixed pattern of bidding based on their expectation to win. This might result in unstable markets. Matthew et al. [14] showed that for some strategic bidding the GSP mechanism does not converge for 3 or more slots/items. These results show that GSP is not strategy-proof and there is a requirement for strategy proof mechanisms for a stable market.

The term AMD was first coined by Nisan et al. [15] in 1999. AMD combines the idea of utility maximization of independent selfish agents and mechanism design from economics, Nash equilibrium and individual rationality from game theory, with the analytic tools of Theoretical Computer Science, such as computational constraints, worst-case analysis and approximation ratios which are not addressed in classical economic mechanism design. From practical implementation point of view, the most important aspect of AMD is the analysis of computational efficiency. If a mechanism cannot be implemented to scale well with respect to number of items and bidders, it cannot be considered as a viable solution. This rules out many classic economic mechanisms which satisfy the mechanism designer's requirements but are computationally inefficient in general to implement. The celebrated classical mechanisms like, Vickrey-Clarke-Groves auction (VCG) and Generalized Vickrey Auction, involve the solution of NP-hard winner determination problems and, in spite of their rich game theoretic properties, are impractical to implement.

Combinatorial auctions are market mechanisms in which bidders can bid on bundles of multiple items, i.e., combination of items. The common example of combinatorial auctions are Federal Commission for Communications (FCC) auctions of spectrum licenses, course registration, airport take-off and landing time slots, job shop scheduling and electricity markets etc [16]. There are various mechanisms proposed for combinatorial auctions, such as AUSM auction, RAD mechanism, PAUSE mechanism and iBundle mechanism etc [16].

Sandholm [17] presented an approximate search algorithm for solving combinatorial auctions. He showed that dynamic programming and exhaustive enumeration methods are too complex to scale for large number of items (20-30 items). Restricting the combinations might result in polynomial time winner determination problem for combinatorial auctions but it has economic inefficiency, since imposing restrictions on certain combinations of items bars bidders from bidding on the combination they might prefer. Furthermore, the bids in [17] were generated by randomly taking values from either $[0;1]$ or from $[0;g]$, where g is the number of items in the bid. This method of bid generation does not consider various economic aspects, such as highly valued/preferred combinations, sub-additive bids, super-additive bids, etc. Hence, [18] claimed that this type of bid generation results in computationally trivial winner determination problem.

In [19], an optimal auction mechanism for multi-unit combinatorial auctions for single-minded bidders was presented. If a bidder is only interested in a single specified bundle (or any of its supersets), it is known as single-minded bidder; a single-minded bidder values at zero any other (non-superset) bundle. In real life, a bidder might be interested in buying multiple bundles with varying preferences. Thus, this setting is not practically applicable in real life. Furthermore, [19] does not consider *sub-additive* and *super-additive* scenarios for combinatorial auctions bids. Archer et al. [20] gave an approximate truthful *randomized* mechanism for combinatorial auctions with single parameter agents. In a single parameter setting, bidders provide one single input to specify their preference, e.g., single-minded bidder. They provided a general technique to modify linear programming-based randomized rounding algorithms to make them monotone with high probability, giving an efficient truthful-in-expectation (TIE) mechanism that approximately maximizes social-welfare. However, single parameter mechanisms are rarely used in practice [21], as the present development in online transactions, Internet markets, Internet advertising and online auctions mandate multi-parameter setting mechanisms. Lavi et al. [4][5] extended the results in [20] to non-single parameter settings and showed that certain linear programming based approximation algorithms for the social welfare problem can be turned into randomized mechanisms that are truthful-in-expectation with the same approximation ratio. Dughmi et al. [22] suggested the first black-box reduction from arbitrary approximation algorithms to truthful approximation mechanisms problems for a non-trivial class of multi-parameter problems. In particular, they proved that every social-welfare-maximization problem that admits an FPTAS and can be encoded as a packing problem, also admits a truthful-in-expectation randomized mechanism which also an FPTAS.

In display Ad auctions, the social welfare maximization problem amounts to finding a maximum-weight independent set (that, is a pairwise-disjoint collection) of squares (or more generally ”fat” rectangles), among a given set of weighted squares. Auctions of this type have been considered in [23][6], but only theoretical results were provided without considering the efficient implementation of the proposed mechanisms.

In this paper, we will focus on the efficient implementation of Lavi-Swamy mechanism, as proposed in [9], and its application to display Ad auctions as suggested in [6]. To the best of our knowledge, this is the first attempt to implement a VCG-based mechanism, which scales with the number of bidders and items.

III. The Setting

A. Items and Bundles

The Advertising space is divided into small units of unitary squares which we refer to as *items*. A combination of items is called a *bundle*. In our case, bundles are restricted to be also squares; see Figure 1 where bundles are represented by dotted squares.

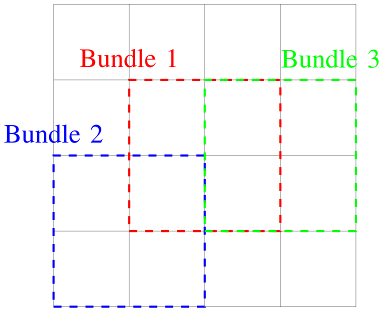


Figure 1. Items and bundles

B. Bidding Mechanism

This refers to the manner in which the value of bid for items and bundles are offered or quoted.

We distinguish the following types of valuations (bids) v_i, for bidder i:

1) Sub-additive setting: For a bundle S consisting of unitary squares $S_1, \ldots, S_k$, $v_i(S) \leq \sum_{j=1}^{k} v_i(S_j)$. This takes into account the bundles/items that are substitutabilities.
2) Super-additive setting: For a bundle S consisting of unitary squares $S_1, \ldots, S_k$, $v_i(S) \geq \sum_{j=1}^{k} v_i(S_j)$. This takes into account of bundles/items that are complementarities.
3) Arbitrary-setting: In arbitrary case, we provide bidders the capability to choose between sub-additive and super-additive valuations.

We call a bidder *k-minded* if (s)he bids a positive value on at most k bundles.

C. The Social Welfare Maximization Problem

Given the vector of bids $v = (v_1, \ldots, v_n)$, the social welfare maximization (SWM) problem is to find the optimum *integer* solution of the following linear program:

$$z^*(v) = \max \sum_{i,S} v_i(S) x_{i,S} \tag{1}$$

$$\text{s.t.} \sum_{i,S:\ j \in S} x^*_{i,S} \leq 1 \qquad \forall \text{ items } j \tag{2}$$

$$\sum_{S} x^*_{i,S} \leq 1 \qquad \forall \text{ bidders } i \tag{3}$$

$$x_{i,S} \geq 0 \qquad \text{for all } i, S.$$

Informally, we want to find an allocation that maximizes the total valuations (i.e., the social welfare) while making sure that (2) each item is only assigned to one bidder and (3) each bidder is only assigned at most one bundle. In our implementation, we used CPLEX [24], to solve the relaxation LP.

For $\alpha \in (0, 1]$, we say that an algorithm $\mathcal{A}$ is an α-*integrality-gap-verifier* for the LP (4) if for any vector of bids v and any (fractional) feasible solution x^* of the LP, $\mathcal{A}$ returns an *integral* solution x^I of social welfare $\sum_{i,S} v_i(S)x^I_{i,S} \geq \alpha \cdot \sum_{i,S} v_i(S)x^*_{i,S}$. A randomized $\frac{1}{16}$-integrality gap verifier for (4) was given in [6] and works as follows:

Algorithm 1 Integrality-Gap-Verifier (x^*)

```
Require: A fractional allocation $x^*$
Ensure: An integral solution $x^I$ s.t. $\sum_{i,S} v_i(S)x^I_{i,S} \geq \frac{1}{16}\sum_{i,S} v_i(S)x^*_{i,S}$
 1: for each bidder $i$ do
 2:     choose a bundle $S_i$ as follows
            $S_i = \begin{cases} S & \text{with prob. } \frac{1}{8}x^*_{i,S} \\ \emptyset & \text{with prob. } 1 - \frac{1}{8}\sum_S x^*_{i,S} \end{cases}$
 3: end for
 4: Let $W = \emptyset$; $x^I = \mathbf{0}$
 5: Let $S_1, S_2, \ldots, S_\ell$ be the bundles selected in Step 2 in
    non-increasing order of size
 6: for $i = 1, \ldots, \ell$ do
 7:     if $S_i$ does not intersect any range $S_j$ with $j \in W$ then
 8:         add $i$ to $W$; $x^I_{i,S_i} = 1$
 9:     end if
10: end for
11: return $x^I$
```

In our implementation, we run Algorithm 1 a constant number of times and take the solution with largest social welfare, to ensure with high probability that it returns a solution with social welfare at least $\frac{1}{16}$ of the optimal.

D. Randomized Truthful-in-expectation Mechanisms

A mechanism consists of an allocation rule and a payment scheme. Given the vector of bids v, the allocation rule determines a feasible allocation, that is, a feasible integral solution to the LP (4), while the payment scheme determines the payment p_i to be charged to bidder i. The utility of a bidder is defined as her valuation over her allocated bundle minus her payment: $U_i(v) := v_i(S) - p_i$. In a randomized mechanism, all these (allocation, payments, and utility) are random variables. Such a mechanism is said to be truthful-in-expectation (TIE) if for all i and all $v_i, \bar{v}_i, v_{-i}$, it guarantees $\mathbb{E}[U_i(\bar{v}_i, v_{-i})] \geq \mathbb{E}[U_i(v_i, v_{-i})]$, where the expectation is taken over the random choices made by the algorithm. Here, we denote by $\bar{v}_i$ the *true* valuation of bidder i, and use the standard terminology $v_{-i} := (v_1, \ldots, v_{i-1}, v_{i+1}, \ldots, v_n)$. In other words, the expected utility is maximized when the bidder reports her valuation truthfully.

E. The Lavi-Swamy Mechanism

We next review the LS-reduction. It consists of three steps:

1) Find an optimal solution x^* for the LP-relaxation (4), and determine the VCG prices p_1 to p_n. The price for the i^{th} agent is $p_i = \sum_{j \neq i}\sum_S v_j(S)(\hat{x}_{j,S} - x^*_{j,S})$, where $\hat{x}$ is an optimal solution for LP (4) with input $(0, v_{-i})$, that is, when bidder i is removed from the auction. The allocation x^* and the VCG-prices are a truthful mechanism for the fractional problem.
2) Write $\alpha \cdot x^*$ as a *convex combination of integral solutions*, i.e., $\alpha \cdot x^* = \sum_I \lambda_I x^I$, $\lambda_I \geq 0$, $\sum_{I \in \mathcal{N}} \lambda_I = 1$, and x^I is an integral solution to (4). This step requires an *α-integrality-gap-verifier* for (4) for some $\alpha \in (0, 1]$.
3) Pick the integral solution x^I with probability λ_I, and charge the i-th agent the price[1] $p_i \cdot (\sum_S v_i(S)x^I_{i,S} / \sum_S v_i(S)x^*_{i,S})$.

F. The EMR Implementation

If one considers the implementation of the LS-mechanism in the display Ad setting, step 2 stands as the major bottleneck as it requires solving a linear program with an *exponential* number of variables. A direct solution of this would require the use of the Ellipsoid method for linear programming which is typically highly inefficient in practice. Elbassioni et al. [9] proposed a solution using the simpler multiplicative weights update methods, which were used for speeding-up convex optimization [25]–[31]. In particular, it was shown that a variation of the approach by Garg et al. [27] can be used to obtain a convex combination that dominates $\alpha \cdot x^*$. Then the packing property of the polytope can be used to covert this into an exact equality algorithm (see Algorithm 3 below). The details of this are given in the following sections.

1) Finding a Dominating Convex Decomposition: Such a decomposition is equivalent to finding an optimal solution of the following LP:

$$\begin{aligned} \min \sum_I \lambda_I & & & (4) \\ \text{s.t.} \sum_I \lambda_I x^I_{i,S} & \geq \alpha \cdot x^*_{i,S} & \text{for all } (i,S) \in L & \quad (5) \\ \sum_{x \in \mathcal{N}_I} \lambda_I & = 1 & & \quad (6) \\ \lambda_I & \geq 0 & \text{for all } I. & \end{aligned}$$

Here, $L = \{(i,S) : \; x^*_{i,S} > 0\}$.

By turning (C2) into an inequality $\sum_{x \in \mathcal{N}_I} \lambda_I \geq 1$, the authors in [9] reduced the problem to a *covering* linear program, which can be solved via the approach in [27][28], at the cost of losing a factor of $(1 + 4\varepsilon)$ in the approximation ratio. The procedure is given as Algorithm 2 and works by maintaining a set of weights $p_{i,S}$ that can be though of as a penalty for the violation in the constraint corresponding to the pair (i,S) in (C1). As long as a scaled version of (C1) is not satisfied for some (i,S), the algorithm uses the weights $p_{i,S}$ (in step 5) to construct a valuation vector v' that is fed to the α-integrality gap verifier $\mathcal{A}$ in step 6, which in turn returns an integral solution x^I. This solution and its multiplier λ_I (computed in steps 7-9) are added to the list of solutions $\mathcal{I}$. Then the set of ”active” constraints L (which are not yet satisfied) is updated and a new iteration is started.

For $y \geq 0$, we define the function

$$h(y) = \begin{cases} y & \text{if } y < 1 \\ -\infty & \text{otherwise.} \end{cases}$$

[1] If $\sum_S v_i(S)x^*_{i,S} = 0$, the price is zero.

```
Algorithm 2 Dominating($x^*, \mathcal{A}$)
Require: A feasible fractional solution to (4), an $\alpha$-integrality
    gap verifier $\mathcal{A}$ for (4), and an accuracy parameter $\varepsilon \in (0, 1/2]$
Ensure: A collection of integral feasible solutions $\{x^I\}_{I\in\mathcal{I}}$
    to (4) and convex multipliers $\{\lambda_I\}_{I\in\mathcal{I}}$ s.t. $\sum_{I\in\mathcal{I}} \lambda_I x^I \geq \frac{\alpha}{1+4\varepsilon} x^*$
$\mathcal{I} := \emptyset$; $M = 0$
$L := |\{(i,S) :\ x^*_{i,S} > 0\}|$; and $T := \frac{\ln|L|}{\varepsilon^2}$
while $M < T$ do
    $p_{i,S} := (1-\varepsilon)^{h(\sum_{I\in\mathcal{I}} \frac{\lambda_I x^I_{i,S}}{\alpha x^*_{i,S}})}$, for $(i,S) \in L$
    Set
        $v'_i(S) := \begin{cases} \frac{p_{i,S}}{\alpha x^*_{i,S} \cdot \sum_{i',S'} p_{i',S'}} & \text{for } (i,S) \in L \\ 0 & \text{otherwise.} \end{cases}$
    Let $x^I := \mathcal{A}(x^*, v')$;
    $\lambda_I := \min_{(i,S)\in L:\ x^I_{i,S}=1} \alpha x^*_{i,S}$
    if $\sum_{I'\in\mathcal{I}} \lambda_{I'} < T$ then
        $\lambda_I := \min\{\lambda_I, 1\}$
    end if$\mathcal{I} := \mathcal{I} \cup \{I\}$
    $L := L \setminus \{(i,S) :\ \sum_{I\in\mathcal{I}} \frac{\lambda_I x^I_{i,S}}{\alpha x^*_{i,S}} \geq T\}$
    $M := \min\left\{\min_{(i,S)\in L} \sum_{I\in\mathcal{I}} \frac{\lambda_I x^I_{i,S}}{\alpha x^*_{i,S}}, \sum_{I\in\mathcal{I}} \lambda_I\right\}$
end while
$\lambda_I := \lambda_I / \sum_{I'\in\mathcal{I}} \lambda_{I'}$, for $I \in \mathcal{I}$
return $(\{x^I\}_{I\in\mathcal{I}}, \{\lambda_I\}_{I\in\mathcal{I}})$
```

At the end of the procedure, the λ_I's are normalized to get a collection of integral feasible solutions $\{x^I\}_{I\in\mathcal{I}}$ to (4) and convex multipliers $\{\lambda_I\}_{I\in\mathcal{I}}$ s.t.

$$\sum_{I\in\mathcal{I}} \lambda_I x^I \geq \frac{\alpha}{1+4\varepsilon} x^*; \tag{7}$$

see [9] for details.

2) Getting an Exact Decomposition: Given the dominating convex decomposition (7), Algorithm 3 can be used to modify it into an exact decomposition. The idea is to use the *packing* property of the feasible set[2] to add more feasible solutions with smaller convex multipliers to offset the difference between the L.H.S and R.H.S. of (7).

We refer the interested reader to [9] for the details as well as the running time analysis of Algorithms 2 and 3, and only summarize the results here.

Theorem 1. *Consider a Display Ad auction on n k-minded bidders and m items. Then, for any $\varepsilon > 0$, there is a TIE which approximates the optimum social welfare within a factor of $\frac{1}{16}(1-4\varepsilon)$ and whose running time is $O(n \cdot T_{LP}(n(k+m), n+m, V_{\max}) + \frac{n^2 m(n+m)\log(n+m)}{\varepsilon^2})$, where $T_{LP}(\ell, r, V_{\max})$ is the time to solve an LP (4) with ℓ variables, r constraints and maximum objective function coefficient $V_{\max} := \max_{i,S} v_i(S)$.*

```
Algorithm 3 Equality($\widehat{x}, \{x^I\}_{I\in\mathcal{I}}, \{\lambda_I\}_{I\in\mathcal{I}}$)
Require: A feasible solution $\widehat{x}$ to (4), a collection of integral
    feasible solutions $\{x^I\}_{I\in\mathcal{I}}$ to (4) and convex multipliers
    $\{\lambda_I\}_{I\in\mathcal{I}}$ s.t. $\sum_{I\in\mathcal{I}} \lambda_I x^I \geq \widehat{x}$
Ensure: A collection of integral feasible solutions $\{x^I\}_{I\in\mathcal{I}}$ to
    (4) and convex multipliers $\{\lambda_I\}_{I\in\mathcal{I}}$ s.t. $\sum_{I\in\mathcal{I}} \lambda_I x^I = \widehat{x}$
Create a new integral solution $x^{I_0} = y^0 := \mathbf{0}$
$\mathcal{I} := \mathcal{I} \cup \{I_0\}$; $\lambda_{I_0} := 0$
while $\exists I \in \mathcal{I}, (i,S)$ s.t. $x^I_{i,S} = 1$ and $\sum_{I\in\mathcal{I}} \lambda_I x^I_{i,S} - \lambda_I \geq \widehat{x}$ do
    $x^I_{i,S} := 0$
end while
while $\exists (i,S) : \Delta_{i,S} := \sum_{I\in\mathcal{I}} \lambda_I x^I_{i,S} - \widehat{x}_{i,S} > 0$ do
    Let $I$ be s.t. $x^I_{i,S} = 1$
    $B := \{(i',S') :\ x^I_{i',S'} = 1$ and $\Delta_{i',S'} > 0\}$; $b = |B|$
    Index the set of pairs $\{(i,S)\}_{i,S}$ s.t. $B = [1..b]$ and
    $\Delta_1 \leq \ldots \leq \Delta_b$
    For $\ell \in [0..b-1]$ define a vector $y^\ell$ by
        $y^\ell_j = \begin{cases} 1 & \text{for } j \leq \ell, \\ 0 & \text{for } j > \ell \end{cases}$
    $\lambda_I := \lambda_I - \Delta_b$
    for $1 \leq \ell < b$ do
        Create a new integral solution $x^{I'} := y^\ell$
        $\mathcal{I} := \mathcal{I} \cup \{I'\}$; $\lambda_{I'} := \Delta_{\ell+1} - \Delta_\ell$
    end for
    $\lambda_{I_0} := \lambda_{I_0} + \Delta_1$
end while
return $(\{x^I\}_{I\in\mathcal{I}}, \{\lambda_I\}_{I\in\mathcal{I}})$
```

IV. Experimental Setup

A. Data Set

Since general combinatorial auctions have never been widely implemented, it is hard to find realistic data sets for them. In view of this, it is natural to rely on artificially generated data sets which can capture the economic and combinatorial issues and successfully represent the sort of scenarios one might expect to encounter [18][32] [33].

The first ever known attempt for generating test data sets for combinatorial auctions (CAs) was analyzed by Leyton-Brown et al. in [18]. They discussed guidelines for generation of CA's bid data and provided a method for bid generation which can successfully capture the economic and underlying structural factors of items and their combinations in CAs. Although they considered as many factors as possible, they overlooked a number of economic issues which should be considered for CA bid data generation [33].

For our experimental data, we considered various aspects as suggested in the literature for generating the data. The aim of generated data was to capture real-life competitive bidding along with ideas of economic substitutability and complementarity of items. The items, bundles and bidders were randomly generated. The bidding value on each item/bundle was made competitive by enforcing a range on bid values

[2] that is, if x is feasible solution then any $\mathbf{0} \leq \mathbf{x'} \leq \mathbf{x}$ is also feasible.

of each item such that bidders can randomly bid within that range. Furthermore, the users were allowed to bid on bundles of items, and the bid values of the bundles could be sub-additive, super-additive or arbitrary.

The following aspects were considered in data generation to capture economic parameters of realistic CAs:

1) *Number of bidders n*: The number of bidders is a good measure of the size of the problem. It essentially depends on whether the auction is between firms or individuals. For firms, a realistic number is substantially $n < 100$ [33]. In case of individuals, n can be up to 10,000 for a typical business-to-consumer auction [33].
   For our experimental setup, we considered $50 \leq n \leq 400$. Further, we also considered the number of bidders relative to number of items (m), reasonably being $0 < \frac{n}{m} \leq 2$.
2) *Number of items m, number of bundles k*: For m unique items, there are $O(m^2)$ possible bundles (each is a square region). This shows the input complexity of CAs. In realistic settings, not all combinations are preferred and the total number of bundles is $\ll O(m^2)$, as there are certain combinations which are preferred by most of the agents. In case of advertisement auctions, typically central regions of the screen, are more valued by bidders. We imposed this restriction in our experimental setting. In our experiments, the number of items/bundles ranges from a few dozens to a few hundreds. In particular, we considered k-minded bidders for $k \in [0, 200]$. In realistic settings, the total number of items $m > 100$ [33].
3) *Bidders valuations v_i*: In realistic settings, bidders go for competitive bidding, i.e., they have a speculation of the possible value of the item or combination of items which is close to its market value, and it is likely that they bid within a certain range of the market value [33]. To capture this idea we associated a range with each item, and bidders can randomly bid within that range for the item.
   If the set of items are structured, other considerations also come into picture, such as location. Our case of online auction also represents such a case. The assumption that all bundles of the same size are equally likely to be requested is clearly violated in real-world auctions [18]. Most of the time, advertisers would like to get slots or combination of slots at particular location or be indifferent to certain locations.
   The following are important parameters that affect the bidding mechanism:
   - Stretch Factor (S-factor): To make the bidding competitive, an S-factor is associated with each item/bundle such that bidders can randomly bid within the stretch factor of the item/bundle value. This captures the real-life setting where bidding values of bidders for a particular item remain close to the estimated or assumed market value.
   - Additive/Subtractive factor (E-factor): To regulate the bid price in case of bundles, we associated an E-factor to the bid value of bundles. The E-factor determines the factor of value that can be added (super-additive) or subtracted (sub-additive) from the total sum of bids of individual items in the bundle.

   If bidding values are not regulated or chosen carefully, then even a hard distribution can become computationally easy [18]. For example, if one particular bidder is randomly bidding very high as compared to others, it can make the optimization problem an easy matching and does not capture the idea of competitive bidding [18]. Boutilier et al. [34] considered the values of bids from the normal distribution with mean 16 and standard deviation 3 and Sandholm [17] generated bid values randomly from either $[0, 1]$ or from $[0 : g]$, where g is the number of items in the bid. Since the bid values are not related to number of items in a bundle, these methods are unreasonable (and computationally trivial) [18].
   In our experimental setting, we provided a stretch factor (S-factor) which ensured that the bid value for bigger bundle is within a certain range of the cumulative sum of bids of smaller bundles in the bigger bundle. We considered the super-additive and sub-additive cases along with the arbitrary case.

V. Evaluation Process

We evaluated our algorithm in terms of truthfulness, running time, social welfare and revenue generated.

A. Truthfulness

To test experimentally the truthfulness of the VCG-based or the GSP-mechanisms, we modified the bid of a particular agent to see if (s)he can increase its expected utility by lying. In particular, we conducted experiments such that one particular, randomly selected bidder is lying, i.e., changes her (his) bid values in small increments while all other bidders fix their bid values. We computed the maximum utility the lying bidder achieves by the mechanism, and computed the ratio to the actual utility that bidder would have got by truthfully bidding. We took the average ratio over thousands of experiments and used this as a measure of how truthful is the mechanism.

B. Running Time

In the display Ad setting, where hundreds of agents are typically participating simultaneity in the auction, it is mandatory for an auction designer to consider its running time. The practical application of any mechanism depends crucially on the fact that it should meet the time requirement.

We have considered the running time of our VCG-based mechanism with respect to increasing the number of items/bundles and the number of bidders.

C. *Social Welfare*

Theorem 1 states that our implemented mechanism ensures that the resulting allocation is almost $\frac{1}{16}$-socially efficient, i.e., the resulting social welfare is almost at least $1/16$ times the optimal social-welfare. While the GSP-mechanism does not ensure a similar guarantee, we experimentally tested the gap in social welfare between the two mechanisms. Since GSP does not allow combinatorial bidding, we expect the behavior to be dependent on the type of valuations used; for supper-additive (resp., sub-additve) valuations, the mechanism VCG-based mechanism would perform better (worse) than GSP with respect to social welfare.

D. *Revenue*

Revenue can be defined as the amount of value generated by the auction in the market which can be taken by seller. It is, in essence, the summation of all the bidders' payments made to the seller. One of the important aspects of any auction mechanism is also the revenue generated.

VI. Results And Analysis

In this Section, we analyse the truthfulness, running time, revenue, and social welfare of the VCG-based mechanism. We further present results comparing this mechanism to GSP. Our experimental results show that the VCG-based mechanism is superior in many aspects to the GSP mechanism. In this Section, we mainly consider the setting where $0.001 <$ S-factor < 0.1 and $1 <$ E-factor < 100 and grid size is 10×10, unless otherwise stated. The vertical bar shows the 95% confidence in the results obtained.

A. *VCG-based Mechanism analysis*

We conducted our experiments with CPLEX [24] and approximate packing algorithm. All the experiments in this Section were run independently for 100 times and their average was calculated for every-parameter.

1) Social-Welfare: We observed that the social welfare obtained increases with the number of bidders and then comes to a saturation value for super-additive and sub-additive case, as shown in Figure 2.

2) Revenue: The revenue generated by our mechanism showed its applicability in real life as shown in Figure 3. Although the VCG-mechanism does not theoretically provide any guarantee on revenue, our experiments show that it provided a good revenue for the market maker. Any commercial auction market cannot be subsided, so it must generate enough revenue for itself to sustain.

3) Running Time: As Theorem 1 claims, the running time for our mechanism is acceptable with respect to number of bidders, items, and bundles. An empirical verification of this can be seen in Figure 4. As the number of bidders increases, the running time also increases almost linearly. We observe that the running time mostly depends on the amount of fractional components in the fractional allocation; if this number is high, the time needed to decompose the fractional allocation into an integral one increases.

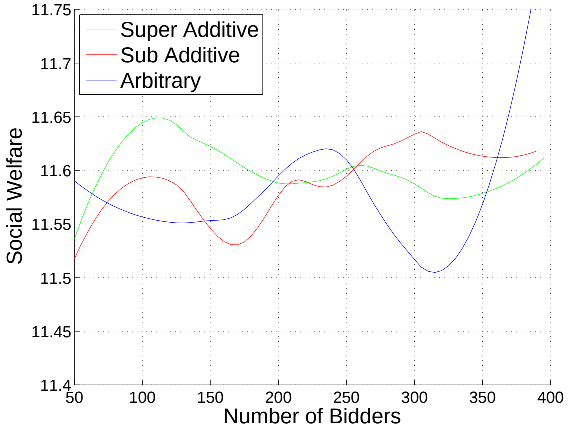


Figure 2. Log curve for Social Welfare of VCG-based mechanism

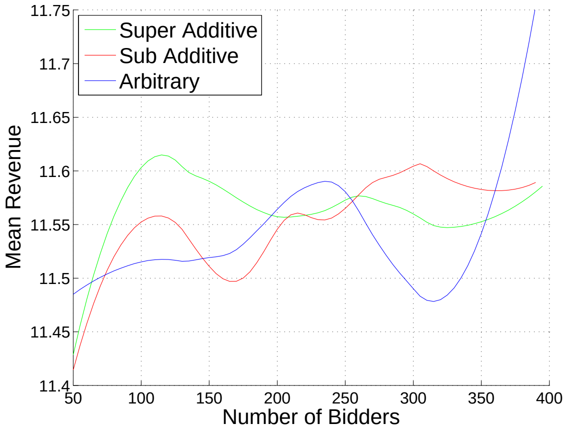


Figure 3. Log curve for Revenue of VCG-based mechanism

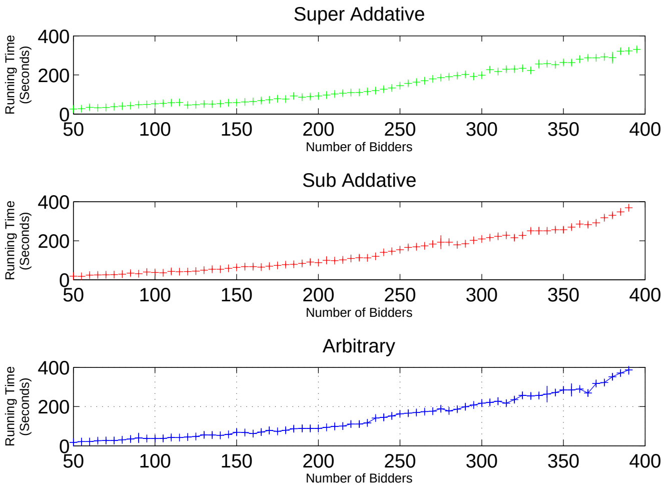


Figure 4. Running Time

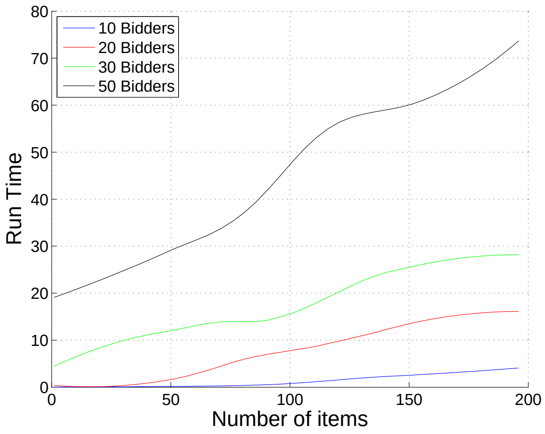


Figure 5. Run Time with varying bundles

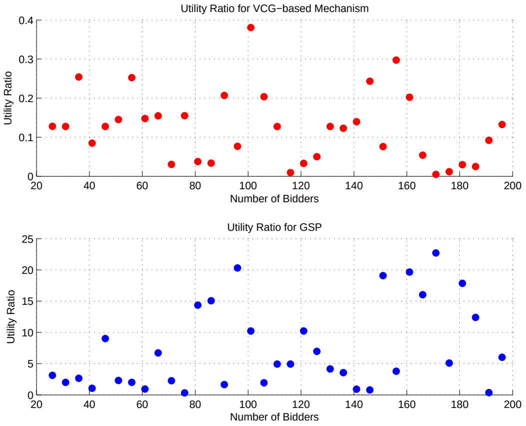


Figure 6. Utility Ratio

We also conducted experiments to evaluate the running time by fixing the number of bidders and increasing the number of bundles successively. We discovered that as the number of bundles increases, the running time increases almost linearly. With a higher number of bidders, the running time tends to increase faster. The same can be seen in Figure 5.

VII. VCG-BASED MECHANISM AND GSP MECHANISM

To compare the truthfulness, social-welfare and revenue of our VCG-based versus GSP, we conducted the experiments with same inputs for the both. However, since the GSP mechanism cannot handle bundles, it was not considered in its input.

A. Truthfulness

To quantify the truthfulness of our approach as explained in sub-section V-A, we conducted many experiments and took the average gain/loss in utility of a lying bidder. Although in some particular instances the lying bidder was able to increase her (his) utility in our mechanism, in overall multiple iterations, the lying bidder was not able to increase the average utility. At the same time, for GSP we observed that there were random increases in utility. In particular, most of the time the lying bidder was able to increase the utility because even if (s)he bids very high as compared to market valuation of the item, (s)he ended up winning the item and still paying the bid-value near to market price as quoted by other losing bidders. In GSP, the items are arranged in specific order and bid-values also have similar order. Hence, a lying bidder who was earlier winning a lower order item or no item at all, might win higher value item and gain substantially high utility by paying something close to market value as quoted by other bidders. In case of GSP, one single lying bidder can change the complete allocation pattern; this might account for the random gain in utility we see.

For the VCG-based mechanism, we observed that over many iterations, the lying bidder either got either zero or less utility for lying. Over a large number of experiments (about 1000 for a given number of players), we found that overall the lying bidder did not achieve substantially more than he could have achieved by telling the truth. In our truthfulness experiments, we used a grid size of 5×5, E-factor $= 1$ and S-factor $= 0.01$. In Figure 6, the curves shown in blue represent the results for the VCG-based mechanism and the ones in red represent those for the GSP mechanism. We can observe that the utility ratio for true valuation and lying valuations is less than one for our mechanism. Thus, overall, the lying bidder is not able to increase her (his) utility by lying. But, in case of GSP, the lying bidder is able to increase the utility by a substantial amount.

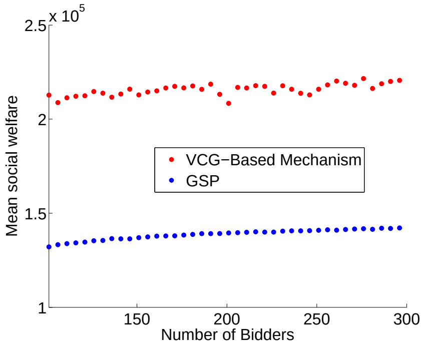


Figure 7. Social Welfare (Arbitrary Setting)

B. Social-Welfare

The magnitude of difference between the social-welfare of GSP and VCG depends on the parameters like E-factor and S-Factor. The results are presented in Figures 7, 8 and 9. In case of super-additive, our VCG-based mechanism is able to generate higher social-welfare as compared to GSP. This can be attributed to the higher economic capacity of combanitorial auction. The social-welfare in arbitrary setting is also observed to be higher than GSP. In case of sub-additive the social-welfare of VCG-based mechanism is lower than GSP. This can be attributed to the fact that sub-additive bids do not increase the overall valuation of bundle as a bid.

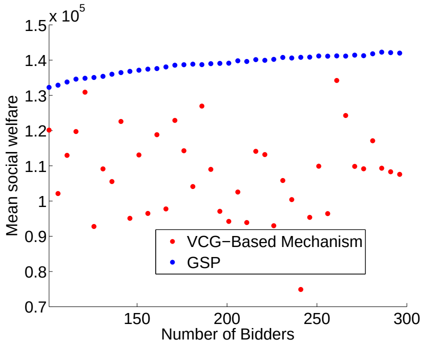


Figure 8. Social Welfare (Sub-Additive Setting)

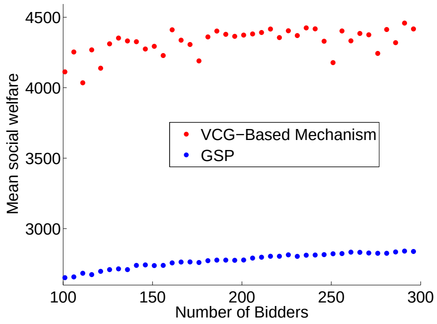


Figure 9. Social Welfare (Super-Additive Setting)

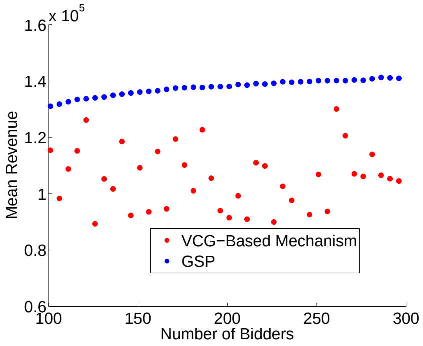


Figure 11. Revenue (Sub-Additive Setting)

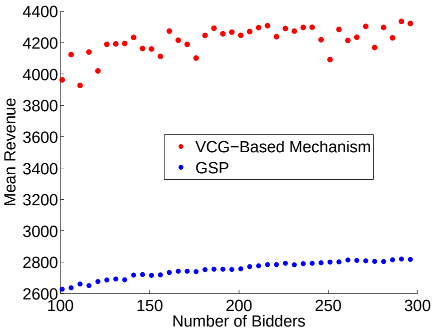


Figure 12. Revenue (Super Additive Setting)

C. Revenue

The revenue generated by our mechanism was higher than that of GSP in super-additive and arbitrary settings. It follows similar pattern as social-welfare. The magnitude of difference between the revenue of GSP and VCG-based mechanism depends on the E-factor and S-Factor. The results are presented in Figures 10, 11 and 12 for various settings.

VIII. Conclusion

In this paper we provided, for (what we believe to be) the first time, an implementation of a VCG-based mechanism for display Ad auctions. Our experiments show that this mechanism is more resilient to lying bidders as compared to GSP, and has reasonable time requirements for expected problem sizes. We also found out experimentally that the implemented mechanism can offer substantially better revenue and social welfare than GSP in many cases. One of the reasons for this is that the combinatorial setting allows for expressing valuations over bundles and generally, bundles have more economic values than single items.

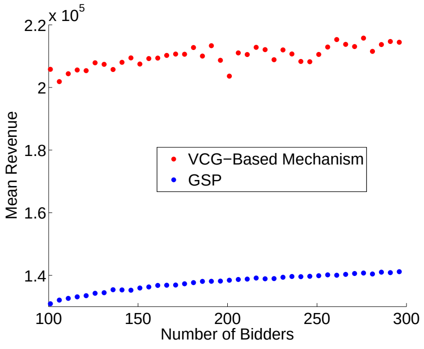


Figure 10. Revenue (Arbitrary Setting)

References

[1] "Iab internet advertising revenue report 2012 full year results," http://www.iab.net/media/file/IAB_Internet_Advertising_Revenue_Report_FY_2012_rev.pdf, retrieved: 05, 2015.

[2] "Google ads display network," http://www.google.ae/ads/displaynetwork/, retrieved: 05, 2015.

[3] N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani, Algorithmic Game Theory. Cambridge University Press, 2007.

[4] R. Lavi and C. Swamy, "Truthful and near-optimal mechanism design via linear programming," in FOCS, 2005, pp. 595–604.

[5] R. Lavi and C. Swamy, "Truthful and near-optimal mechanism design via linear programming," Journal of the ACM (JACM), vol. 58, no. 6, 2011, p. 25.

[6] G. Christodoulou, K. Elbassioni, and M. Fouz, "Truthful mechanisms for exhibitions," in WINE, 2010, pp. 170–181.

[7] M. Hoefer, T. Kesselheim, and B. Vöcking, "Approximation algorithms for secondary spectrum auctions," in SPAA, 2011, pp. 177–186.

[8] B. Edelman, M. Ostrovsky, and M. Schwarz, "Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords," American Economic Review, vol. 97, no. 1, 2007, pp. 242–259, retrieved: 05, 2015. [Online]. Available: http://www.aeaweb.org/articles.php?doi=10.1257/aer.97.1.242

[9] K. M. Elbassioni, K. Mehlhorn, and F. Ramezani, "Towards more practical linear programming-based techniques for algorithmic mechanism design," CoRR, vol. abs/1408.1577, 2014. [Online]. Available: http://arxiv.org/abs/1408.1577

[10] D. S. Evans, "The online advertising industry: Economics, evolution, and privacy," The journal of economic perspectives, 2009, pp. 37–60.

[11] Z. Wei and M. Lin, "Auction vs. posted-price: Market mechanism, lender behaviors, and transaction outcomes in online crowdfunding," Posted-Price: Market Mechanism, Lender Behaviors, and Transaction Outcomes in Online Crowd-Funding (September 1, 2013), 2013.

[12] B. Edelman and M. Ostrovsky, "Strategic bidder behavior in sponsored search auctions," Decision support systems, vol. 43, no. 1, 2007, pp. 192–198.

[13] K. Asdemir, "Bidding patterns in search engine auctions," in Second Workshop on Sponsored Search Auctions. Citeseer, 2006.

[14] M. Cary, A. Das, B. Edelman, I. Giotis, K. Heimerl, A. R. Karlin, C. Mathieu, and M. Schwarz, "Greedy bidding strategies for keyword auctions," in Proceedings of the 8th ACM conference on Electronic commerce. ACM, 2007, pp. 262–271.

[15] N. Nisan and A. Ronen, "Algorithmic mechanism design (extended abstract)," in Proceedings of the Thirty-first Annual ACM Symposium on Theory of Computing, ser. STOC '99. New York, NY, USA: ACM, 1999, pp. 129–140. [Online]. Available: http://doi.acm.org/10.1145/301250.301287

[16] P. Cramton, Y. Shoham, and R. Steinberg, "Combinatorial auctions," 2006, retrieved: 05, 2015.

[17] T. Sandholm, "Algorithm for optimal winner determination in combinatorial auctions," Artificial Intelligence, vol. 135, no. 12, 2002, pp. 1 – 54. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S000437020100159X

[18] K. Leyton-Brown, M. Pearson, and Y. Shoham, "Towards a universal test suite for combinatorial auction algorithms," in Proceedings of the 2Nd ACM Conference on Electronic Commerce, ser. EC '00. New York, NY, USA: ACM, 2000, pp. 66–76. [Online]. Available: http://doi.acm.org/10.1145/352871.352879

[19] S. Gujar and Y. Narahari, "Optimal multi-unit combinatorial auctions," Operational Research, vol. 13, no. 1, 2013, pp. 27–46.

[20] A. Archer, C. Papadimitriou, K. Talwar, and É. Tardos, "An approximate truthful mechanism for combinatorial auctions with single parameter agents," in SODA '03: Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2003, pp. 205–214.

[21] L. Ausubel and P. Milgrom, "The lovely but lonely vickrey auction," 2006, p. 1740.

[22] S. Dughmi and T. Roughgarden, "Black-box randomized reductions in algorithmic mechanism design," in Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on, 2010, pp. 775–784.

[23] M. Babaioff and L. Blumrosen, "Computationally-feasible truthful auctions for convex bundles," in In RANDOM+APPROX, 2004, pp. 64–75.

[24] "IBM ILOG CPLEX Optimizer," urlhttp://www-01.ibm.com/software/integration/optimization/cplex-optimizer/, Last 2010.

[25] D. Bienstock and G. Iyengar, "Approximating fractional packings and coverings in o(1/epsilon) iterations," SIAM J. Comput., vol. 35, no. 4, 2006, pp. 825–854.

[26] M. D. Grigoriadis and L. G. Khachiyan, "A sublinear-time randomized approximation algorithm for matrix games," Operations Research Letters, vol. 18, no. 2, 1995, pp. 53 – 58.

[27] N. Garg and J. Könemann, "Faster and simpler algorithms for multicommodity flow and other fractional packing problems," in 39th Annual IEEE Symposium on Foundations of Computer Science (FOCS), 1998, pp. 300–309.

[28] R. Khandekar, "Lagrangian relaxation based algorithms for convex programming problems," Ph.D. dissertation, Indian Institute of Technology, Delhi, 2004.

[29] C. Koufogiannakis and N. E. Young, "Beating simplex for fractional packing and covering linear programs," in 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS), 2007, pp. 494–504.

[30] S. A. Plotkin, D. B. Shmoys, and É. Tardos, "Fast approximation algorithms for fractional packing and covering problems," in FOCS, 1991, pp. 495–504.

[31] N. E. Young, "Sequential and parallel algorithms for mixed packing and covering," in FOCS, 2001, pp. 538–546.

[32] N. Nisan, "Bidding and allocation in combinatorial auctions," in Proceedings of the 2nd ACM conference on Electronic commerce. ACM, 2000, pp. 1–12.

[33] C. Gallo, "A data set generation algorithm in combinatorial auctions," in Computer as a Tool, 2005. EUROCON 2005.The International Conference on, vol. 1, Nov 2005, pp. 744–747.

[34] C. Boutilier, M. Goldszmidt, and B. Sabata, "Sequential auctions for the allocation of resources with complementarities," in Proceedings of the 16th International Joint Conference on Artifical Intelligence - Volume 1, ser. IJCAI'99. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999, pp. 527–534. [Online]. Available: http://dl.acm.org/citation.cfm?id=1624218.1624294

LDaaSWS: Toward Linked Data as a Semantic Web Service

Leandro José S. Andrade and Cássio V. S. Prazeres
Computer Science Department
Federal University of Bahia
Salvador, Bahia, Brazil
Email: {leandrojsa, prazeres}@dcc.ufba.br

***Abstract*—The Web was originally created to link HTML documents. Nowadays, the Web has improved its potential, and heterogeneous applications, resources, data and users can interact with each other. Two proposals for improvement of the current Web, Semantic Web and Web Services, have established standards that make the interoperability between heterogeneous Web applications possible. Another way to improve the current Web is the Web of Data, which provides guidelines (Linked Data) about how to use Semantic Web standards for publication and definition of semantic links on diverse data sources. However, there is a gap in the integration between Web Service based applications and Web of Data applications. Such a gap occurs because Web Services are "executed" and Web of Data applications are "queried". Therefore, this paper introduces the LDaaSWS (Linked Data as a Semantic Web Service), in order to provide Web Services for data sources from the Web of Data. The LDaaSWS can fulfill the current gap between Web Services and Web of Data applications by making the Web of Data "executed" through Web Services. In order to compare this work with current approaches for Web Services, this paper also presents an evaluation of LDaaSWS by comparing with SOAP Web Services.**

***Keywords*–Web of Data; Semantic Web Services; OWL-S; Linked Data.**

I. Introduction

The initial purpose of the Web was to create hyperlinks between Hypertext Markup Language (HTML) documents [1]. From this initial purpose, the capability of the Web has been improved extensively, for example, now making user collaboration (Web 2.0) [2] and applications interoperability (Web API and Web Services) possible [3]. According to Martin et al. [4], standards should be developed for the Web and, furthermore, Web Services have to produce and consume data through a common protocol to make data interchange between heterogeneous applications possible.

In this case, standards are means to describe Web Services with languages such as Web Services Description Language (WSDL) and Web Application Description Language (WADL), which present service syntactical description of Simple Object Access Protocol (SOAP) services [5] and RESTful services [6], respectively. After Berners-Lee's [7] first article about the Semantic Web, several works have introduced different approaches for the semantic description of Web Services, in order to automate tasks, such as discovery, composition and invocation of Web Services [8] [9]. As a result, several approaches were proposed as standards for semantic description of Web Services, among which the Semantic Markup for Web Services (OWL-S) used in this work.

On the other hand, Berners-Lee [10] introduced a set of guidelines (Linked Data) to publish data on the Web. These rules indicate to use URI for identify resources, the HTTP protocol to access resources, Resource Description Framework (RDF) and SPARQL Query Language for RDF (SPARQL) for description, query, and hyperlinks to other resources. Such guidelines were inspired by a project to publish open data on the current Web: Linking Open Data [11]. Furthermore, the Web of Linked Data is being called the Web of Data [1].

Developers of applications from the current Web want to make their applications functionalities and/or data available to be accessed by other applications [12]. According to O'Reilly [2], there are several different data sources that demand applications through combining such data sources to offer composite services. These kinds of applications are known as mashups, which integrate Web resources to create new applications [13]. However, developing such mashups demands programming efforts for program developers, because they have to discover and compose the available Web resources [14] .

Therefore, there is a gap in the integration between these two Web evolution trends: Web Service based applications and Web of Data applications. Such a gap occurs because Web Services are "executed" through Hypertext Transfer Protocol (HTTP) requests, and Web of Data applications are "queried" through SPARQL queries. This issue is current and relevant, as several authors presented works [15] [16] [17] with approaches to address it. Some of these works introduce approaches to describe Web Services with Linked Data and others introduce approaches to produce Linked Data through Web Services or Web APIs. In order to overcome this gap, this paper introduces the Linked Data as a Semantic Web Service (LDaaSWS), in order to provide Semantic Web Services of data sources from the Web of Data. In summary, this paper proposes to make access on Linked Data sources through Web Services possible.

In order to establish the Linked Data cloud as a Web Service provider, LDaaSWS implements Semantic Web Services described with OWL-S. This approach enables the automatic integration of data from the Web of Data with others types of OWL-S based services (for instance, SOAP and RESTful). Furthermore, LDaaSWS also enables the automatic generation of Linked Data queries from service requests described on OWL-S.

Figure 1 presents the overall view of our proposal by explaining a possible scenario of a tourism application, which needs a Web Service that receives as input a city and gives as outputs some information about this city (latitude and longitude, hotels, description, population size and phone code). In this scenario, there is not a Web Service that fulfill the request, however some parts of the resquest can be attend. In Figure 1 shows two OWL-S/WSDL Web Services (Services 2 and 5 in Figure 1) that fulfill output about hotels and phone code of city. It has other Web Service (Service 1 in Figure 1), now a LDaaSWS one, that suplement information about latitude and longitude of the city. However, in this scenario there are not Web Services for outputs about description of city and population size (Services 3 and 4 in Figure 1). Then, we present an extension (Section III) to language OWL-S in order to allow the description of LDaaSWS proposed in this article, used in Service 1 in Figure 1.

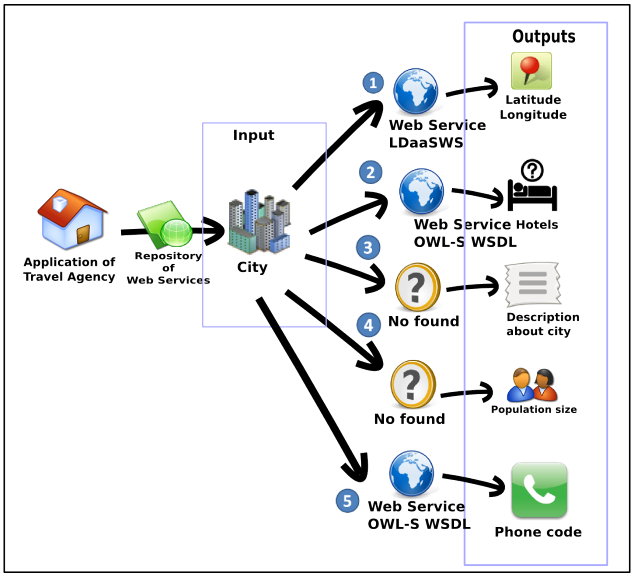


Figure 1. LDaaSWS overall view.

We developed a module for discovery of new LDaaSWS services to fulfill parts of services requests without a match (showed in Section IV). In Figure 1 the outputs 3 and 4 do not have services to attend, so we can use the module for discovery LDaaSWS to do it.

In this sense, this paper introduces the following results: i) OWL-S ontology extension to provide support for services derived from the Web of Data; ii) the usage of the OWL-S API [18], in order to enable the execution of the LDaaSWS in the same way as traditional (SOAP or RESTful) Web Services are executed; iii) automatic generation of Web of Data queries from OWL-S service requests, in order to enable the automatic requests and execution of LDaaSWS.

In this paper, Section II presents related works. Linked Data as a Semantic Web Service (proposal of this work) is described in Section III. Section IV introduces the automatic request and execution of LDaaSWS. Section V describes the results of the evaluation performed in our approach. Finally, Section VI presents the final remarks and recommendations for future works.

II. Related Work

The literature reports works that use the Web of Data as Web Services, or even use Linked Data for describing Semantic Web Services (SWS). According to Pedrinaci et al. [19], SWS and Web of Data together can solve some problems that limit the use of both. The combination of these two features, in addition, can increase use of SWS, as this will add to your field a growing multidisciplinary information base, serving as a complete and complementary method in the discovery and composition of Web Services.

Following a different line of research, but within the same context, Taheriyan et al. [16] identified the need for the composition of services from different sources to improve application development. Thus, the authors proposed an approach to integrate Web API's to a Linked Data cloud with the use of semantic description, using RDF and SPARQL. Norton and Stadtmller [20], [21] underscore the need for composing RESTful services by reducing the effort of the manual programming developer; it proposes a description of the services using Linked Data principles and semantically describing its inputs and outputs with SPARQL and RDF.

Paolucci et al. [22] propose the integration of SAWSDL (Semantic Annotations for WSDL) with Semantic Web Services described in OWL-S. The authors point out advantages in describing Semantic Web Services in OWL-S language and propose an extension of the ontology description of executions of OWL-S Web Services (Grounding) to support services with SAWSDL descriptions in OWL-S.

III. Linked Data as a Semantic Web Service

This paper proposes Linked Data as a Semantic Web Service (LDaaSWS), in order to provide Web Services from Linked Data sources. The motivation for this proposal mainly focuses on three points: i) the Web of Data is a database where access is restricted to SPARQL, which limits its potential queries, because it hinders integration with other data sources that have no Linked Data; ii) LDaaSWS makes the usage of the Linked Data cloud as a service provider possible; iii) LDaaSWS enables Linked Data to be integrated automatically with other Web Services supported by the language OWL-S (for instance, SOAP and RESTful Web Services), enabling interoperability of data and reducing the programming effort for service discovery and composition.

Thus, this paper presents SPARQLGrounding, which is an extension to the language OWL-S in order to allow the description of LDaaSWS. The OWL-S ontology is composed of three sub-ontologies: `Profile`, `Process` and `Grounding`. The first two are abstract, generic and can include any implementation of service (SOAP-WSDL, RESTful-WADL, etc.). The `Grounding` ontology had been defined to be the concrete part of the OWL-S. Whereas it does not define type of service implementation, the `Grounding` is responsible for describing how the service will be performed. OWL-S can and should be extended to any type of service through the extension of the `Grounding` ontology.

Thus, the `SparqlGrounding` ontology proposed in this work is an extension of the OWL-S `Grounding` used to allow the Web of Data be executed as a service, i.e., to actually implement execution of LDaaSWS proposed in this paper.

In this context, the extension proposed in this paper followed the following requirements: i) allow execution of SPARQL queries based services that enable the mapping of input and output of services to SPARQL triples; ii) be in line with other OWL-S ontologies, that is, its inputs and outputs are correctly associated with the elements of `Process` and `Profile` ontologies; iii) not to be dependent on anything other than document OWL-S and its sub-ontologies for a full description; iv) offer a semantic description, unambiguously, for automated processes of discovery, selection, composition and execution of services, performed by software agents.

The mapping of OWL-S `Grounding` for SPARQL extends the abstract layer composed by `Grounding` and `AtomicProcessGrounding` classes, both defined by OWL-S. Figure 2 shows a UML class diagram that displays such extension.

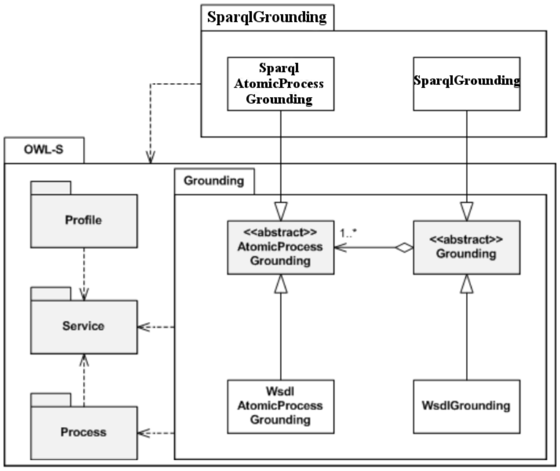


Figure 2. SparqlGrounding model.

Figure 2 illustrates the `Service`, `Profile` and `Process` ontologies which details were omitted to highlight the specialization of `Grounding` ontology. Classes `SparqlGrounding` and `SparqlAtomicProcessGrounding` in Figure 2 are not part of the OWL-S `Grounding` ontology. These classes are proposed in this paper with the aim of describing the execution of LDaaSWS.

The model presented in Figure 2 proposes grouping the new classes in a dedicated ontology, called `SparqlGrounding`. This approach avoids any change in the specification of OWL-S, which will facilitate the adoption of our new ontology, without interfering with existing services described in OWL-S prior to incorporation of `SparqlGrounding`.

In order to formalize `SparqlGrounding`, it is necessary to provide a description of all its elements following the syntax of the OWL language. Figure 3 describes the formalization of the `SparqlGrounding` class as a subclass of the `Grounding` class previously defined by OWL-S. Note also that the definition has restrictions related to the existence of `AtomicProcessGrounding` (line 5 of Figure 3) and that all such should be of type `SparqlAtomicProcessGrounding` (line 6 of Figure 3).

The `SparqlAtomicProcessGrounding` class, shown in Figure 4, is a subclass of `AtomicProcessGrounding` and has the restriction of a data property (`Datatype-Property`) called `sparqlEndPoint`, which stores the URI endpoint that a SPARQL query must be submitted on completion of the service. Other elements are associated with LDaaSWS belonging to this class.

```
1 <owl:Class rdf:ID="SparqlGrounding">
2   <rdfs:subClassOf rdf:resource="&grounding;Grounding"/>
3   <rdfs:subClassOf>
4     <owl:Restriction>
5       <owl:onProperty rdf:resource="&grounding;hasAtomicProcessGrounding"/>
6       <owl:allValuesFrom rdf:resource="#SparqlAtomicProcessGrounding"/>
7     </owl:Restriction>
8   </rdfs:subClassOf>
9 </owl:Class>
```

Figure 3. SparqlGrounding Class.

```
1 <owl:Class rdf:ID="SparqlAtomicProcessGrounding">
2   <rdfs:subClassOf rdf:resource="&grounding;AtomicProcessGrounding"/>
3   <rdfs:subClassOf>
4     <owl:Restriction>
5       <owl:onProperty rdf:resource="#sparqlEndPoint"/>
6       <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
7     </owl:Restriction>
8   </rdfs:subClassOf>
9 </owl:Class>
```

Figure 4. SparqlAtomicProcessGrounding Class.

A SPARQL query can have prefixes that are eventually used in triplets, so in `SparqlGrounding` a property of type *ObjectProperty* (lines 1 to 4 of Figure 5) is defined, followed by the OWL class `SparqlPrefixMap` (lines 6 to 19 of Figure 5), which defines the prefix name (`PrefixName`) and the associated URI (`PrefixUri`). Figure 5 presents an excerpt of this definition; note that `SparqlPrefixMap` defines the existence of only one elements `PrefixName` and `PrefixUri`, ensuring integrity for the formation of prefixes.

```
 1 <owl:ObjectProperty rdf:ID="SparqlPrefixes">
 2   <rdfs:domain rdf:resource="#SparqlAtomicProcessGrounding"/>
 3   <rdfs:range rdf:resource="#SparqlPrefixMap"/>
 4 </owl:ObjectProperty>
 5
 6 <owl:Class rdf:ID="SparqlPrefixMap">
 7   <rdfs:subClassOf>
 8     <owl:Restriction>
 9       <owl:onProperty rdf:resource="#PrefixName"/>
10       <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
11     </owl:Restriction>
12   </rdfs:subClassOf>
13   <rdfs:subClassOf>
14     <owl:Restriction>
15       <owl:onProperty rdf:resource="#PrefixUri"/>
16       <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
17     </owl:Restriction>
18   </rdfs:subClassOf>
19 </owl:Class>
```

Figure 5. SparqlPrefixes Property and SparqlPrefixMap Class.

Figure 6 (lines 1 to 4) shows the definition of input parameter of `SparqlGrounding`. The `SparqlIntputParamMap` class (lines 6 to 21 of Figure 6) is responsible for the mapping between elements of `Process` (`Input`) through of property `owlsParameter` (line 17 of Figure 6), and data associated with the input of the service. In WADL and WSDL groundings, the `owlsParameter` is used to connect the input ID of syntactic service document. However, in the context of LDaaSWS, it is used to indicate which variable in the query will be mapped with the input data.

Finally, Figure 7 describes the definition of triple belonging to the clause `WHERE` of SPARQL queries, mapping the subject, predicate and object of the triple, respectively represented by the properties `TripleSubject` (line 9 of Figure 7), `TriplePredicate` (line 15 of Figure 7) and `TripleObject` (line 21 of Figure 7). In

```
<owl:ObjectProperty rdf:ID=”SparqlInputParam”>
  <rdfs:domain rdf:resource=”#SparqlAtomicProcessGrounding”/>
  <rdfs:range rdf:resource=”#SparqlInputParamMap”/>
</owl:ObjectProperty>

<owl:Class rdf:ID=”SparqlIntputParamMap”>
  <rdfs:subClassOf rdf:resource=”#SparqlDataParamMap”/>
  <rdfs:subClassOf rdf:resource=”#InputMessageMap”/>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource=”#SparqlDataParam”/>
      <owl:cardinality rdf:datatype=”&xsd;nonNegativeInteger”>1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource=”&grounding;owlsParameter”/>
      <owl:cardinality rdf:datatype=”&xsd;nonNegativeInteger”>1</owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>
```

Figure 6. OWL elements to describe service input.

`SparqlTripleMap` there is a restriction on the amount of elements that make up the triple to ensure its syntactic integrity.

```
<owl:ObjectProperty rdf:ID=”SparqlTriples”>
  <rdfs:domain rdf:resource=”#SparqlAtomicProcessGrounding”/>
  <rdfs:range rdf:resource=”#SparqlTripleMap”/>
</owl:ObjectProperty>

<owl:Class rdf:ID=”SparqlTripleMap”>
<rdfs:subClassOf>
  <owl:Restriction>
    <owl:onProperty rdf:resource=”#TripleSubject”/>
    <owl:cardinality rdf:datatype=”&xsd;nonNegativeInteger”>1</owl:cardinality>
  </owl:Restriction>
</rdfs:subClassOf>
<rdfs:subClassOf>
  <owl:Restriction>
    <owl:onProperty rdf:resource=”#TriplePredicate”/>
    <owl:cardinality rdf:datatype=”&xsd;nonNegativeInteger”>1</owl:cardinality>
  </owl:Restriction>
</rdfs:subClassOf>
<rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource=”#TripleObject”/>
      <owl:cardinality rdf:datatype=”&xsd;nonNegativeInteger”>1</owl:cardinality>
    </owl:Restriction>
</rdfs:subClassOf>
</owl:Class>
```

Figure 7. OWL elements to describe SPARQL triples.

IV. AUTOMATIC REQUEST AND EXECUTION OF LDAASWS

LDaaSWS makes automatic service requests (Section IV-A) possible through automatic generation of SPARQL queries from OWL-S requests. Moreover, it is also possible to perform automatic execution (Section IV-B) of LDaaSWS from the queries generated. This section presents such two features (automatic request and execution) of LDaaSWS, which enable software agents to access services based on Linked Data automatically.

A. Automatic Request

In order to explore the Web of Data, it is necessary to use SPARQL, which makes the need to map the semantics described by the OWL-S language to a SPARQL query indispensable. Figure 8 presents an overview of the generation of the SPARQL queries module from service requests described in OWL-S: OWL-S to SPARQL. From the service request (part 1 of Figure 8), one SPARQL query of type *ASK* (returns *true* or *false*, respectively, exist or not one or more data according to the request). (part 2 of Figure 8) is developed, which is generated from the ontology of the inputs and outputs. We choose this type of query, because it has a lower cost of implementation, given that, at this stage, no one wants to return data but rather validate the query.

After the generation of an ASK query, the next step is the application of this query in an *endpoint* (part 3 of Figure 8) to check for data that matches the request. In the case of ASK query return positive, an OWL-S service can be created and can use this database for the answers (part 4 of Figure 8), which one can follow to implement the `SparqlGrounding` (described in Section III).

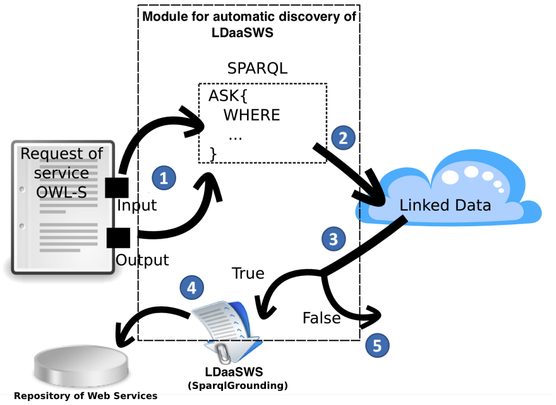

Module for automatic discovery of LDaaSWS
SPARQL
ASK{
WHERE
...
}
Request of service OWL-S
Input
Output
1
2
Linked Data
3
True
False
4
5
LDaaSWS (SparqlGrounding)
Repository of Web Services

Figure 8. OWL-S to SPARQL: Automatic Request.

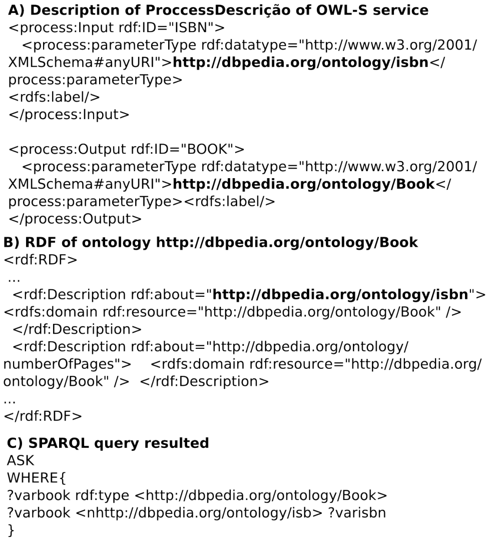
There are some challenges associated with this mapping OWL-S to SPARQL shown in Figure 8. Initially, the semantic expressiveness of OWL-S language is greater than the expressiveness of SPARQL. This implies the need to identify what types of requests are enabled to perform the query expression. Therefore, for an experimental evaluation, we selected a profile of simple OWL-S request, which allows the generation of queries. Thus, requests for services where the input element is owned (is a property) by the output element (or otherwise) were selected. For example, a service request where the input is the latitude and longitude of a city and the output is an ontology class that represents city.

Figure 9 presents the description of a service request (part “a” of Figure 9), which has the ISBN of a book as input and a Book (a class) as output. The ISBN is part of the domain Book (part “b” of Figure 9), thus this type of service allows to generate a SPARQL query similar to the query displayed in part “c” of Figure 9.

We can map other types of queries through using techniques for similarity of ontologies, or even through inferences, which allow the resources of inputs to be related with the resources of outputs. We can also generate queries that partially meet the requests, where the generated service can later be combined with other services to meet the initial request. As a result, `SparqlGrounding` allows the execution (Section IV-B) of any service described with SPARQL query, in other words, it does not restrict the execution of services with queries automatically generated. Thus, if a developer designs a SPARQL query manually and wants to set it as an OWL-S service, the `SparqlGrounding` can be implemented for such a query. However, the mapping of queries is out of the scope of this paper.

B. Automatic Execution

Figure 10 illustrates an overview of the LDaaSWS execution. The starting point is the OWL-S request (step 1 in



```
A) Description of ProccessDescrição of OWL-S service
<process:Input rdf:ID="ISBN">
   <process:parameterType rdf:datatype="http://www.w3.org/2001/
XMLSchema#anyURI">http://dbpedia.org/ontology/isbn</
process:parameterType>
<rdfs:label/>
</process:Input>

<process:Output rdf:ID="BOOK">
   <process:parameterType rdf:datatype="http://www.w3.org/2001/
XMLSchema#anyURI">http://dbpedia.org/ontology/Book</
process:parameterType><rdfs:label/>
</process:Output>
B) RDF of ontology http://dbpedia.org/ontology/Book
<rdf:RDF>
 ...
  <rdf:Description rdf:about="http://dbpedia.org/ontology/isbn">
<rdfs:domain rdf:resource="http://dbpedia.org/ontology/Book" />
  </rdf:Description>
  <rdf:Description rdf:about="http://dbpedia.org/ontology/
numberOfPages">    <rdfs:domain rdf:resource="http://dbpedia.org/
ontology/Book" />  </rdf:Description>
...
</rdf:RDF>

C) SPARQL query resulted
ASK
WHERE{
?varbook rdf:type <http://dbpedia.org/ontology/Book>
?varbook <nhttp://dbpedia.org/ontology/isb> ?varisbn
}
```

Figure 9. Service request and SPARQL query.

Figure 10) and from analysis of the ontologies that describe the inputs and outputs, a SPARQL query equivalent to the request is developed (OWL-S to SPARQL module in Figure 10 corresponds to Figure 8). At this stage, not all service requests produce SPARQL queries, since not all of them refer to information available on the Web of Data (this issue is treated in Section IV-A). After that, the resulting query is validated in the Linked Data cloud (step 2 in Figure 10), at which time whether or not there is data available to fulfill this request is reviewed.

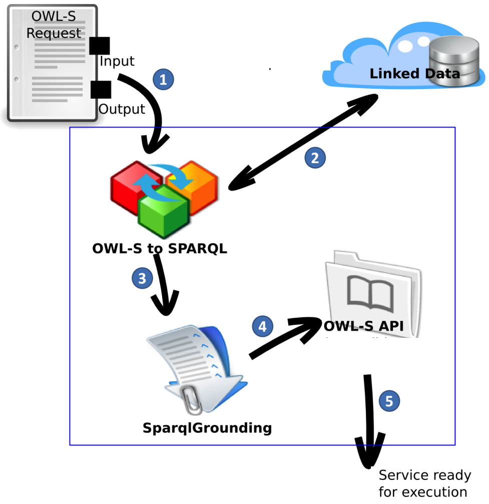


Figure 10. LDaaSWS Automatic Execution.

In step 3 in Figure 10, if there is information in the Web of Data for the OWL-S request, it has been a Grounding OWL-S (described in Section III), which describes how to run a service. Finally, the resulting `Grounding` is sent to the OWL-S API (step 4 in Figure 10) to be executed (step 5 in Figure 10).

Importantly, once a whole process to request execution of LDaaSWS is held (steps 1-5 of Figure 10), this procedure should not be repeated for subsequent requests, because the new discovered service may be stored in a database of services that can be accessed in the future, reducing the time for discovery and generation of service.

Figure 11 shows a snippet of a code to perform a service that uses the `SparqlGrounding`. Initially, the service is loaded into knowledge base (line 3 of Figure 11) to allow access to the `Process`, which, consequently, indicates the data input (lines 5 to 7 Figure 11). Finally, in line 12 of Figure 11, the service is performed, which, from `Process`, is called the class `SparqlGroundingProvider`, which starts the whole foundation class `SparqlGrounding` implemented by the service.

```
 1  // loading of ontologies
 2  URI uri = new URI("http://localhost/services/isbn_book_sparql_grounding.owls");
 3  Service service = kb.readService(uri);
 4
 5  Process process = service.getProcess();
 6  // Creating input parameters
 7  ValueMap<Input, OWLValue> inputs = new ValueMap<Input, OWLValue>();
 8  inputs.setValue(process.getInput("ISBN"), kb.createDataValue("0-375-50137-1"));
 9
10  // Creating output parameters and executing the service
11  ValueMap<Output, OWLValue> outputs = new ValueMap<Output, OWLValue>();
12  outputs = exec.execute(process, inputs, kb);
```

Figure 11. Code snippet to perform a service.

V. LDaaSWS Evaluation

In the LDaaSWS evaluation, we used an Intel Core i5 computer with four cores of 1.80GHz, 6GB of RAM memory, with operation system Debian/Linux 8.0, Java environment with J2SE 1.7 and Eclipse 3.8.1. Additionally, we used the Apache Web Service for access and storage of Web Services and the DBPedia [23] for access to a Linked Data base. The services used to execute the tests were extracted from the *OWL-S Test Collection* [24] package, which have been adapted to use DBPedia ontology and the updated version 1.2 of OWL-S.

Experiments have been performed to evaluate the performance (execution time) of our proposal. Toward greater consistency of results, for each evaluation 30 tests were executed and the execution time in each test was observed. As described in Sections III and IV, the development of LDaaSWS has three important contributions, which were separately analyzed for better evaluation. Therefore, Section V-A shows the evaluation of the `SparqlGrounding` ontology; Section V-B presents the evaluation of automatic generation of SPARQL queries – request and execution.

A. *SparqlGrounding ontology*

In the ontology evaluation, the OWL-S API was used. Indicating the correct functioning of `SparqlGrounding` ontology is necessary in order to run a Semantic Web Service using `SparqlGrounding`; in other words, it must use the OWL-S API with support for LDaaSWS for execution of a Web Service.

Therefore, in order to evaluate the performance of `SparqlGrounding`, the execution time was fully measured in the following points: i) reading of ontologies before starting the execution Process; ii) mapping and reading of `Grounding` classes; and iii) preparation time for Semantic Web Service execution. This paper did not deal with the Web Service execution, because it was beyond the scope of contributions and its execution time is generally associated with the performance of the service itself.

Because the `WSDLGrouding` is a built-in grounding of the OWL-S API, we related measurements taken with `SparqlGrounding` in comparison to equivalent measurements of `WSDLGrouding`. Figures 12, 13 and 14 show the results of measurements of `SparqlGrounding` in comparison with `WSDLGrouding`. We can see that there are no substantial differences in performance; however, `SparqlGrounding` has better performance because in the query preparation it does not need access to a syntactic document, which reduces the execution time of `SparqlGrounding`. As a result, through an analysis of the graphs, it appears that the `SparqlGrounding` gives satisfactory performance with the OWL-S API.

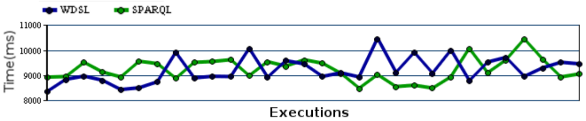


Figure 12. Mapping and reading classes WSDLGrounding and SPARQLGrounding

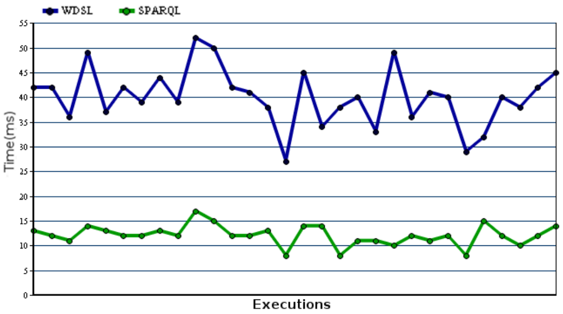


Figure 13. Reading ontologies related to WSDLGrounding and SPARQLGrounding

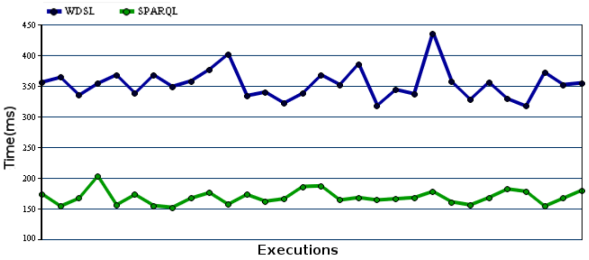


Figure 14. Preparation time for execution of the service with WSDLGrounding and SPARQLGrounding

B. Request and Execution

In order to evaluate the automatic generation of SPARQL queries, we applied a scenario where the main module for SPARQL queries generation was subjected to executions using OWL-S service requests. Figures 15 and 16 show a piece of the code of service requests highlighting elements of input (lines 1 to 6 of Documents 15 and 16) and output (lines 8 to 13 of Figure 15 and lines 8 to 20 of Figure 16) of the `Process`, which are the main features for the generation of SPARQL queries.

```
 1 <process:Input rdf:ID="_ISBN">
 2   <process:parameterType rdf:datatype="http://www.w3.org/2001/XMLSchema#anyURI">
 3     http://dbpedia.org/ontology/isbn
 4   </process:parameterType>
 5   <rdfs:label>isbn</rdfs:label>
 6 </process:Input>
 7
 8 <process:Output rdf:ID="_BOOK">
 9   <process:parameterType rdf:datatype="http://www.w3.org/2001/XMLSchema#anyURI">
10     http://dbpedia.org/ontology/Book
11   </process:parameterType>
12   <rdfs:label>book</rdfs:label>
13 </process:Output>
```

Figure 15. Part of the service request ISBN-BOOK

```
 1 <process:Input rdf:ID="_CITY">
 2   <process:parameterType rdf:datatype="http://www.w3.org/2001/XMLSchema#anyURI">
 3     http://dbpedia.org/ontology/City
 4   </process:parameterType>
 5   <rdfs:label></rdfs:label>
 6 </process:Input>
 7
 8 <process:Output rdf:ID="_LAT">
 9   <process:parameterType rdf:datatype="http://www.w3.org/2001/XMLSchema#anyURI">
10     http://www.w3.org/2003/01/geo/wgs84_pos#lat
11   </process:parameterType>
12   <rdfs:label></rdfs:label>
13 </process:Output>
14
15 <process:Output rdf:ID="_LON">
16   <process:parameterType rdf:datatype="http://www.w3.org/2001/XMLSchema#anyURI">
17     http://www.w3.org/2003/01/geo/wgs84_pos#long
18   </process:parameterType>
19   <rdfs:label></rdfs:label>
20 </process:Output>
```

Figure 16. Part of the service request City-Latitude/Longitude

We split the evaluation into three parts, for better measurement of the performance of the SPARQL query generator module: i) time of reading of ontologies associated with the request OWL-S service; ii) the building time of the SPARQL query; and iii) query execution time (DBPedia endpoint[25]).

Figures 17, 18 and 19 show graphs with execution time measurements for OWL-S service requests of Figures 15 and 16, noting the time spent in three situations: reading of the OWL-S service request (Figure 17), creation of the SPARQL query (Figure 18) and execution of the SPARQL query created (Figure 19). It is important to note that over 50% of total execution cost is associated with the reading of ontologies, a point that is not connected with the implemented solution, but rather the access of resources of the OWL-S service request.

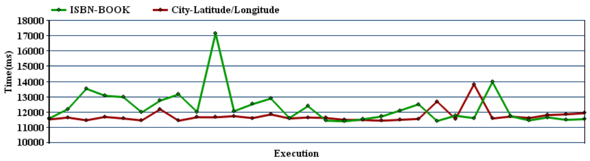


Figure 17. Time of reading of ontologies

Thus, the results reported for measuring execution time of the creation and execution of SPARQL queries show that total time for automatic generating SPARQL queries was about 5 seconds. This time is quite acceptable considering that the developer of a Web Service would not need to develop a query manually and this process will not be repeated in cases of OWL-S services requests already converted into SPARQL queries.

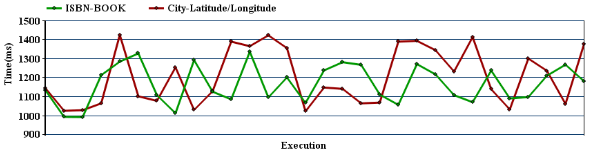


Figure 18. Time of creation of SPARQL query

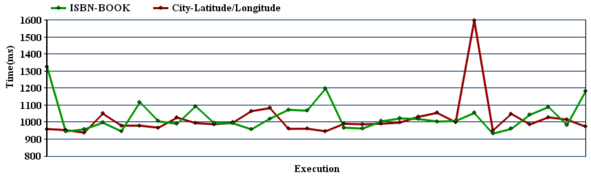


Figure 19. Time of execution SPARQL query

VI. Final Remarks

The evolution of the Web presents a scenario with data coming from various sources and applications that can provide its functionality by merging information from different sources. This trend motivates researches efforts for the development of techniques that create environments and techniques for automatic discovery of data and services in the Web.

Therefore, this paper aims at providing Semantic Web Services using semantic descriptions with OWL-S language from Linked Data: the LDaaSWS. This proposal presents important contributions to the area of Semantic Web Services, especially regarding the discovery, because LDaaSWS allow automatic generation of Web Services from the Linked Data cloud. Furthermore, it enables the development of more elaborate applications, which require less expertise of developers and enable more integration and reuse of data from the Web of Data.

Improvements that can be made from the reported work include: i) implementation of other mappings of complex OWL-S requests to automatically generated SPARQL queries; ii) conducting experiments to evaluate the composition of LDaaSWS with other already established Web Services types (such as SOAP and RESTful), based on approaches to Semantic Web Service composition reported in literature [26].

References

[1] C. Bizer, “The emerging web of linked data,” IEEE Intelligent Systems, vol. 24, no. 5, Sep. 2009, pp. 87–92.

[2] T. OŔeilly, “What is web 2.0. design patterns and business models for the next generation of software,” Communications and Strategies, September 2005.

[3] C.-C. Tsai, C.-J. Lee, and S.-M. Tang, “The web 2.0 movement: mashups driven and web services,” W. Trans. on Comp., vol. 8, no. 8, aug 2009, pp. 1235–1244.

[4] D. Martin, M. Burstein, D. Mcdermott, S. Mcilraith, M. Paolucci, K. Sycara, D. L. Mcguinness, E. Sirin, and N. Srinivasan, “Bringing semantics to web services with owl-s,” World Wide Web, vol. 10, no. 3, Sep. 2007, pp. 243–277.

[5] D. Box, D. Ehnebuske, G. Kakivaya, A. Layman, N. Mendelsohn, H. F. Nielsen, S. Thatte, and D. Winer, “Simple Object Access Protocol (SOAP) 1.1,” World Wide Web Consortium, W3C Note, 2000.

[6] L. Richardson and S. Ruby, Restful web services, 1st ed. OŔeilly, 2007.

[7] T. Berners-Lee, J. Hendler, and O. Lassila, “The semantic web,” Scientific American, vol. 284, no. 5, May 2001, pp. 34–43.

[8] S. McIlraith, T. C. Son, and H. Zeng, “Semantic web services,” Intelligent Systems, IEEE, vol. 16, no. 2, 2001, pp. 46–53.

[9] P. Larvet, B. Christophe, and A. Pastor, “Semantization of legacy web services: From wsdl to sawsdl,” in Internet and Web Applications and Services, 2008. ICIW ’08. Third International Conference on, June 2008, pp. 130–135.

[10] T. Berners-Lee, “Linked data - design issues,” W3C, no. 09/20, 2006. [Online]. Available: http://www.w3.org/DesignIssues/LinkedData.html

[11] C. Bizer, T. Heath, D. Ayers, and Y. Raimond, “Interlinking open data on the web,” www4.wiwiss.fu-berlin.de/bizer/pub/LinkingOpenData.pdf, 2007, stand 12.5.2009. [Online]. Available: www4.wiwiss.fu-berlin.de/bizer/pub/LinkingOpenData.pdf

[12] D. Benslimane, S. Dustdar, and A. Sheth, “Services mashups: The new generation of web applications,” Internet Computing, IEEE, vol. 12, no. 5, 2008, pp. 13–15.

[13] S. Makki and J. Sangtani, “Data mashups & their applications in enterprises,” in Internet and Web Applications and Services, 2008. ICIW ’08. Third International Conference on, June 2008, pp. 445–450.

[14] G. Di Lorenzo, H. Hacid, H.-y. Paik, and B. Benatallah, “Data integration in mashups,” SIGMOD Rec., vol. 38, no. 1, jun 2009, pp. 59–66.

[15] S. Roy Chowdhury, C. Rodríguez, F. Daniel, and F. Casati, “Baya: assisted mashup development as a service,” in Proceedings of the 21st international conference companion on World Wide Web, ser. WWW ’12 Companion. New York, NY, USA: ACM, 2012, pp. 409–412.

[16] M. Taheriyan, C. A. Knoblock, P. Szekely, and J. L. Ambite, “Rapidly integrating services into the linked data cloud,” in Proceedings of the 11th international conference on The Semantic Web - Volume Part I, ser. ISWC’12. Berlin, Heidelberg: Springer-Verlag, 2012, pp. 559–574.

[17] S. Stadtmüller and B. Norton, “Scalable discovery of linked apis,” Int. J. Metadata Semant. Ontologies, vol. 8, no. 2, Sep. 2013, pp. 95–105.

[18] M. D. Evren Sirin and T. Mller. Owl-s api. Available on the internet at http://on.cs.unibas.ch/. Last access in 21 October 2014. [Online]. Available: http://on.cs.unibas.ch/ (2012)

[19] C. Pedrinaci, J. Domingue, and R. Krummenacher, “Services and the web of data: An unexploited symbiosis.” in AAAI Spring Symposium: Linked Data Meets Artificial Intelligence. AAAI, 2010.

[20] B. Norton and S. Stadtmüller, “Scalable discovery of linked services,” in Proceedings of the Fourth International Workshop on Resource Discovery, vol. 737, RED Workshop. Heraklion, Greece: CEUR-WS, Mai 2011.

[21] S. Stadtmuller, “Composition of linked data-based restful services,” in Proceedings of the 11th international conference on The Semantic Web, ser. ISWC’12. Berlin, Heidelberg: Springer-Verlag, 2012, pp. 461–464.

[22] M. Paolucci, M. Wagner, and D. Martin, “Grounding owl-s in sawsdl,” in Service-Oriented Computing - ICSOC 2007, ser. LNCS, B. Kramer, K.-J. Lin, and P. Narasimhan, Eds. Springer Berlin Heidelberg, 2007, vol. 4749, pp. 416–421.

[23] Dbpedia. Available on the internet at http://www.dbpedia.org. Last access in 24 December 2014. [Online]. Available: http://www.dbpedia.org (2014)

[24] M. Klusch and P. Kapahnke. Owls-tc is a owl-s service retrieval test collection to support the evaluation of the performance of owl-s semantic web service matchmaking algorithms. [Online]. Available: http://projects.semwebcentral.org/projects/owls-tc/ (2010)

[25] Sparql endpoint dbpedia. Available on the internet at http://www.dbpedia.org/sparql. Last access in 27 October 2014. [Online]. Available: http://www.dbpedia.org/sparql (2014)

[26] T. Weise, S. Bleul, D. Comes, and K. Geihs, “Different approaches to semantic web service composition,” in Internet and Web Applications and Services, 2008. ICIW ’08. Third International Conference on, June 2008, pp. 90–96.

An Empirical Study For Investigating How Politeness in Virtual Commercial Contexts Influence Customer Satisfaction and Loyalty

I-Ching Chen
Department of Information Management
Chung Chou University of Science and Technology
Chang Hua 51003, Taiwan, ROC
e-mail: jine@dragon.ccut.edu.tw

Shueh-Cheng Hu
Department of Computer Science & Comm. Engineering
Providence University
Taichung 43301, Taiwan, ROC
e-mail: shuehcheng@gmail.com

***Abstract*—Politeness exhibited in a commercial context influences a business. E-commerce emerges a major way to conduct business; by contrast, politeness issues in virtual commercial contexts receive rare attention. This work aims to investigate whether politeness influence customer satisfaction and loyalty in online storefronts. The present work extended the American customer satisfaction index (ACSI) model by taking the politeness construct into account. The instrument's reliability and validity were confirmed through empirical data analysis. By using the extended model, business can examine to which extent the politeness will influence their customers' satisfaction and loyalty. Besides its practical applications, this work sets a stage for future studies trying to investigate the relationships between the politeness construct and other constructs interesting business administrators.**

Keywords - E-commerce; online storefronts; politeness; ACSI model; SEM.

I. INTRODUCTION

Politeness broadly refers to legitimate and considerate interactions among persons, which was found as a foundation of modern civilization [1] and a key factor upholding prosperous and peaceful societies [2]. Particularly, politeness is significant within commercial contexts. A merchant will lose its customers gradually if it cannot treat them politely; even it has other merits such as competitive pricing, plentiful product choices, advanced facilities, convenient layout, etc. Impoliteness in commercial contexts often hurts people's feelings and faces, thus will overshadow the above merits, and leave customers negative impression and words-of-mouth. Based on practical experiences and rationales, politeness in commerce contexts influences peoples' perceptions, satisfaction, and loyalty. Many prior academic studies [3]-[5] confirmed the influence of politeness on customer satisfaction, which is a key driver of customer loyalty [6], sustainable revenue [7][8], and successful business.

In addition, according to prior studies that developed measurements for measuring service quality in different segments, politeness was treated as one of the determinants of business' service quality [9][10], which in turn has been proved as a significant influence on customer satisfaction [11][12], and on buyers' re-purchasing and referral behaviors [13], which is called customer loyalty.

In light of its significance in operating a successful business and the latent relationships with customer satisfaction and loyalty, the present work aims to formally investigate how politeness will influence customer satisfaction and loyalty in online storefronts and to which extent the influence will be.

The remaining parts of this article are organized as follows: Section II briefs prior studies regarding the politeness and the ACSI model; Section III describes the research method; Section IV analyzes the research findings; and the concluding remarks, implications, and future directions were provided in Section V.

II. PRIOR WORKS REVIEW

A. Politeness and Business Administration

Prior study found that people expect politeness from computers reciprocally, just like they treat their computers with politeness [14]. The findings indicate that people do care about the politeness of computers with which they interact. Another study indicated that the politeness shown by computers will make users behave reciprocally with more politeness [15]. Besides, a number of prior studies [16]-[19] also confirmed the influence of politeness on human-computer interactions.

Regarding the commercial contexts, Berry [20], Reynolds and Beatty [21] found that rapport consisting of enjoyable interactions and personal connections, is a major determinant affecting customers' satisfaction and loyalty, which contribute to a successful business. Kim and Davis [22] further pointed out that politeness plays a key role in early stage of nourishing rapport between sales representatives and customers. The implication of the above studies is that merchants not likely to build a satisfying and loyal customer base without paying attention to the politeness issues in their commercial contexts.

When waves of computer and Internet keep on permeating into various aspects of our daily life, customers eventually will well recognize the politeness issues in online storefronts, just like they do in physical commercial contexts. Whitworth [23] stated that impolite software is one kind of social error, which likely to drive away users. In light of the significance of politeness in widely-computerized societies, Whitworth established a "polite computing" framework [24] that took a multi-facet viewpoint to examine cyberspace's politeness beyond linguistic strategies. The framework consists of five principles for judging whether computer-initiated actions in five different facets are polite or not, based on users' perceptions. The 5 principles for judging politeness are summarized as follows:

1. Respect user's rights; polite software respects and thus does not preempt users' rights. Besides, polite software does not utilize information before obtaining the permission from its owner.

2. Behave transparently; polite software does not change things in secret, in contrast, it clearly declares what it will do or is doing, the real purpose of the action, and who it represents.

3. Provide useful information; polite software helps users make informed decisions by providing useful and comprehensible information, in contrast, they avoid providing information that distract or even mislead users.

4. Remember users; polite software memorize its past interactions with a specific user, thus can bring that user's choices and preferences to future interactions.

5. Respond to users with fidelity; polite software must respond to users' requests faithfully rather than trying to pursue its own agenda.

B. Customer Satisfaction and the ACSI

According to prior studies, customer satisfaction plays key role in improving revenue [25]-[27] and increasing profit [28]-[30]. Furthermore, because it also positively affects stock investment returns [31][32], smart investors incline to those enterprises with higher customer satisfaction. In view of its significance, enterprises must be concerned about how to satisfy their customers, in effective and efficient ways.

The ACSI [33], is a benchmark for measuring customer satisfaction with the quality of products and services available to household consumers in the United States. The ASCI periodically reports customer satisfaction scores ranging from 0 to 100 on four different levels: national, 10 economic sectors, 47 major industries, and more than 230 companies/agencies, according to the perceived experience of consumers. To collect data, roughly 70,000 customers are randomly picked and surveyed annually.

Many research works have been conducted based on the rationales of the ASCI model, some used the original ACSI model, while many others applied variant models that were adjusted according to specific requirements. By using the ACSI, profitability and firm value in the hospitality and tourism industry were proved to be related with customer satisfaction [34]. The reliability of ACSI was studies and confirmed in different industries of other countries [35]. Antecedents of aggregate customer satisfaction were investigated by analyzing the relationships between cross-country economic indicators and national customer satisfaction data [36]. A model derived from the ACSI was successfully applied to identify factors which most significantly affect customer satisfaction of low-priced housing industry in Beijing, China [37]. An index for gauging customer satisfaction in online re-tailing in Taiwan (e-CSI), was developed based on the ACSI and was found to be effective in measuring customer satisfaction and predicting customer loyalty accordingly [38]. Overall speaking, the ACSI methodology have been proved to be a reliable and valid instrument for gauging customer satisfaction in national, sector, industry, and company levels.

III. Research Method

A. Hypothesis Model Development

Responses from surveyed customers are fed into the extended ACSI model, which is a multi-equation econometric model developed by the University of Michigan's Ross School of Business, American Society for Quality, and the CFI group in 1994. As Figure 1 illustrates, the extended ACSI model is a cause-and-effect model with 4 constructs for representing antecedents of customer satisfaction on the left side: customer expectations, perceived quality, perceived value, and politeness; construct of customer satisfaction in the center; while two constructs for representing consequences of satisfaction on the right side: customer complaints and customer loyalty [39]. Customer loyalty consists of the re-purchase intention and the price tolerance; the former gauges customer's professed likelihood to repurchase from the same supplier in the future, while the latter one gauges customer's likelihood to purchase a company's products or services at various price points. Customer loyalty is a critical construct in the model since it is a key determinant of firm profitability.

Each construct is a multivariable component, which could be measured by several questions that are weighted within the model, and the questions assess customer evaluations of the determinants of each construct. Since the present study adopted the extended ACSI model to investigate the antecedents and consequences of customer satisfaction. Being consistent with prior studies adopting the similar model, the following 11 hypotheses are made about customers' perceptions in the context of online commerce:

H_1: Customer expectations (CE) will have a positive impact on perceived quality (PQ).

H_2: Perceived quality (PQ) will has a positive impact on perceived value (PV).

H_3: Perceived quality (PQ) will has a positive impact on customer satisfaction (CS).

H_4: Customer expectations (CE) will have a positive impact on perceived value (PV).

H_5: Customer expectations (CE) will have a positive impact on customer satisfaction (CS).

H_6: Perceived value (PV) will has a positive impact on customer satisfaction (CS).

H_7: Customer satisfaction (CS) will has a negative impact on customer complains (CC).

H_8: Customer satisfaction (CS) will has a positive impact on customer loyalty (CL).

H_9: Customer complains (CC) will have a negative impact on customer loyalty (CL).

H_{10}: Politeness (PL) will has a positive impact on customer satisfaction (CS).

H_{11}: Politeness (PL) will has a positive impact on customer loyalty (CL)

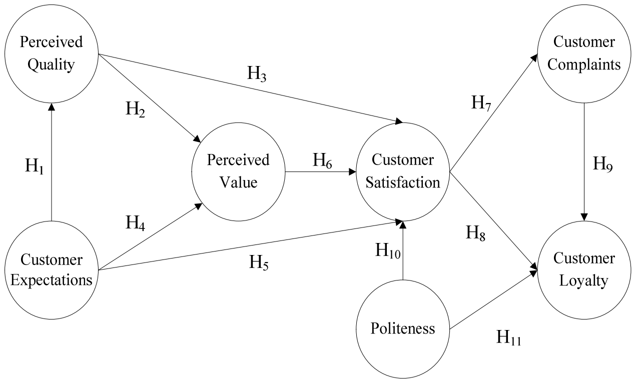


Figure 1. The hypothesis model of the extended ACSI

B. *Instrument*

To verify the hypothesis model, a field study technique was employed through a survey. A structured questionnaire was used to survey customers' perceptions. The questionnaire contains total 21 items as Table. I shows; each construct (dimension) have number of corresponding items reflecting the manifest variables. The items basically came from the methodology report of the ACSI [40] and the polite principles proposed by Brian Whitworth and his colleagues [41], all these question items were devised according to the relevant studies and theories. All items in the survey were on a seven-point scale, ranging from strongly disagree (1) through neutral (4) to strongly agree (7). A pretest of the survey was conducted to check if there exist any ambiguous loadings before administration of the survey.

C. *Participants*

An online questionnaire was used to collect participants' opinions; the participants were, in part, recruited from information management majored college students in Taiwan. Besides, to broaden the sampling population, friends and family members of the recruited students were also invited. Before answering the questionnaire, a short instruction was provided for guiding the participants to assess online storefronts. After the orientation, 536 participants filled the online survey in May and June 2014, and 346 completed the survey effectively. The subjects whose responses were considered to be effective must have more than 5-year experience in online shopping. 182 (52.6%) out of 346 effective respondents were female, respondents were aged between 18 to 63 year-old, and their average age is 45.2.

TABLE I. CONSTRUCT AND INSTRUMENT ITEMS

Latent	Variable	Manifest Variable (Question) Description
Perceived Quality	PQ1	Overall evaluation of quality experience with service (post-purchase)
	PQ2	Evaluation of customization experience, or how well the service fits the customer's personal requirements (post-purchase)
	PQ3	Evaluation of reliability experience, or how often things have gone wrong with service (post purchase)
Perceived Value	PV1	Rating of price given quality
	PV2	Rating of quality given price
Customer Satisfaction	CS1	Overall satisfaction
	CS2	Expectancy disconfirmation (performance that falls short of or exceeds expectations)
	CS3	Performance versus the customer's ideal product and service in the category
Customer Complaints	CC1	Has the customer complained to the company regarding the product/service quality
	CC2	Has the customer complained to the company regarding the service encounter
Customer Loyalty	CL1	Repurchase likelihood rating
	CL2	Price tolerance (increase) given repurchase
	CL3	Price tolerance (decrease) to induce repurchase
	CL4	Say good things about the merchant to other people
Customer Expectations	CE1	Overall expectation of quality (pre-purchase)
	CE2	Expectation regarding customization, or how well the product and service fits the customer's personal requirements (pre-purchase)
	CE3	Expectation regarding reliability, or how often things would go wrong (pre-purchase)
Politeness	PL1	Merchants do not display disturbing but irrelevant messages
	PL2	Merchants use member information only after notification and getting permission
	PL3	Merchants provide well-organized catalogues and/or search engines, so patrons can find particular products with ease
	PL4	Merchants remember my preferred choices

IV. FINDINGS & ANALYSIS

The analysis of collected data was conducted with the Statistical Product and Service Solutions (SPSS). After that, an advanced statistics method - structured-equation model (SEM) was employed to carry out the subsequent analysis by applying the LISREL 9. The LISREL takes into account all co-variances in the data set and thus allows users to simultaneously examine the correlations, shared variances, the casual relationships between constructs (hypothesis), and the significance level and coefficient of the lines.

A. *Reliability of the Instrument*

Reliability of the questionnaire, which comprises 7 constructs, was evaluated using Cronbach's alpha. As Table II shows, the Cronbach's alpha values of all constructs were close to 0.6, except the customer loyalty (CL), which composite reliability value is 0.53. These values indicated the instrument has a moderate reliability. Besides, other measurement model fit indices all exceed the common threshold values recommended by domain experts [42][43]. The figures also indicated that all items load significantly on their corresponding construct demonstrating adequate convergent validity.

TABLE II. MEASUREMENT MODEL FIT INDICES FOR CONVERGENT VALIDITY

Variable	Standardized item loading	Measure error	Indicator reliability (SMC)	Composite reliability (CR)	Variance extracted (VE)
PQ1	0.78	0.4	0.61	0.60	0.775
PQ2	0.79	0.37	0.62		
PQ3	0.76	0.43	0.58		
PV1	0.87	0.24	0.76	0.74	0.860
PV2	0.85	0.28	0.72		
CS1	0.59	0.65	0.35	0.58	0.763
CS2	0.84	0.29	0.71		
CS3	0.83	0.31	0.69		
CC1	0.85	0.28	0.72	0.65	0.806
CC2	0.76	0.42	0.58		
CL1	0.80	0.37	0.64	0.53	0.728
CL2	0.78	0.4	0.61		
CL3	0.81	0.34	0.66		
CL4	0.47	0.78	0.22		
CE1	0.78	0.39	0.61	0.62	0.790
CE2	0.78	0.4	0.61		
CE3	0.81	0.34	0.66		
PL1	0.75	0.43	0.56	0.68	0.824
PL2	0.86	0.26	0.74		
PL3	0.85	0.28	0.72		
PL4	0.83	0.31	0.69		

B. *Discriminant validity and goodness-of-fit*

Discriminant validity was assessed according to the Holmes-Smith [44] stating that variance extracted estimates should exceed square of the correlation between the two constructs. In this work, correlation matrix approach and factor analyses were applied to examine the convergent and discriminant validity. As summarized in Table III, the smallest within-factor correlations are adequate. Besides, each smallest within-factor correlation was considerably higher among items intended for the same construct than among those designed to measure different constructs. These data suggest that adequate convergent and discriminant validity of the survey.

TABLE III. INTER-CONSTRUCT CORRELATIONS MATRIX

Latent	PQ	PV	CS	CC	CL	CE	PL
PQ	0.775*						
PV	0.40	0.860*					
CS	0.46	0.39	0.763*				
CC	0.00	0.00	-0.37	0.806*			
CL	0.00	0.00	0.58	-0.03	0.728*		
CE	0.68	0.18	0.17	0.00	0.00	0.790*	
PL	0.00	0.00	0.09	0.00	0.02	0.00	0.824*

*. THE SQUARE OF VE

The eight common goodness-of-fit indexes, summarized in Table IV, exceed their respective common acceptance levels, suggesting that the research model exhibited a good fit with the collected data.

TABLE IV. GOODNESS-OF-FIT MEASUREMENTS

Goodness-of-Fit Measure	Level of Acceptable fit	Model Result
Chi-square statistic	$P \geq 0.05$ [12]	388.23 (p=0.0)
x^2/df	<3 [2]	388.23/177=2.193
RMSEA	< 0.08 [13]	0.059
CFI	≥ 0.9 [11]	0.97
GFI	≥ 0.9 [10, 21]	0.90
AGFI	≥ 0.8 [11, 21]	0.87
NFI	≥ 0.9 [12]	0.95
NNFI	≥ 0.9 [12]	0.97

C. *Influential Effects Analysis*

The LISREL was used to calculate the coefficients (factor loadings) indicating the extent to which the latent variables affect the measured variables. In summary, Figure 2 and Table V show the standardized LISREL path coefficients and corresponding t-values. They show that 9 out of the 11 original hypotheses (the corresponding relationships between construct nodes) are significant, except the two: one is between politeness and customer loyalty; another is between customer complaints and loyalty.

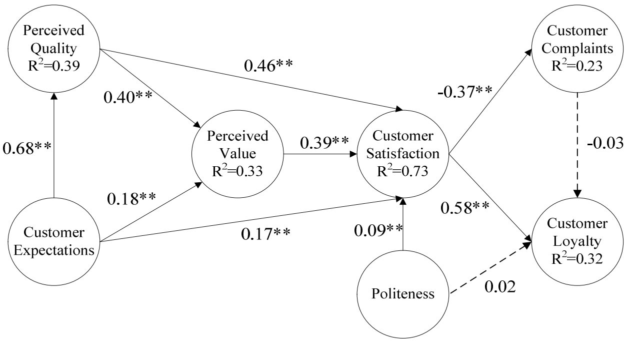

Perceived Quality R^2=0.39
Customer Complaints R^2=0.23
0.46**
0.40**
-0.37**
0.68**
Perceived Value R^2=0.33
0.39**
Customer Satisfaction R^2=0.73
-0.03
0.18**
0.58**
Customer Expectations
0.17**
0.09**
Customer Loyalty R^2=0.32
Politeness
0.02

Figure 2. Standardized LISREL solution (*:p< 0.05 ; **:p< 0.01)

TABLE V. HYPOTHESES RESULTS OF RESEARCH MODEL

Hypothesis	Path coefficient	t-value	Acceptable
H_1: Customer Expectations →Perceived Quality	0.68^{**}	10.57	Yes
H_2: Perceived Quality →Perceived Value	0.40^{**}	4.38	Yes
H_3: Perceived Quality →Customer Satisfaction	0.46^{**}	5.74	Yes
H_4: Customer Expectations →Perceived Value	0.18^{**}	2.08	Yes
H_5: Customer Expectations →Customer Satisfaction	0.17^{**}	2.48	Yes
H_6: Perceived Value →Customer Satisfaction	0.39^{**}	6.33	Yes
H_7: Customer Satisfaction →Customer Complaints	-0.37^{**}	-5.21	Yes
H_8: Customer Satisfaction →Customer Loyalty	0.58^{**}	7.36	Yes
H_9: Customer Complaints →Customer Loyalty	-0.03	-0.51	No
H_{10}: Politeness →Customer Satisfaction	0.09^{**}	2.25	Yes
H_{11}: Politeness →Customer Loyalty	0.02	0.45	No

*.p<0.05; **.p<0.01

Table VI summarizes the total causal effects on latent independent variables.

TABLE VI. ANALYSIS OF INFLUENTIAL EFFECTS

Independent Latent	Dependent latent	Total Effects	Indirect Effects	Direct Effects
Customer Expectations	Perceived Quality	0.68	--	0.68
Customer Expectations	Perceived Value	0.45	0.27	0.18
Customer Expectations	Customer Satisfaction	0.65	0.49	0.16
Politeness	Customer Satisfaction	0.09	--	0.09
Politeness	Customer Loyalty	0.08	0.06	0.02
Perceived Quality	Perceived Value	0.40	--	0.4
Perceived Quality	Customer Satisfaction	0.62	0.15	0.47
Perceived Value	Customer Satisfaction	0.39	--	0.39
Customer Satisfaction	Customer Complaints	-0.37	--	-0.37
Customer Satisfaction	Customer Loyalty	0.59	0.01	0.58
Customer Complaints	Customer Loyalty	-0.03	--	-0.03

--:no path; *:p<0.05; **:p<0.01

V. DISCUSSION AND CONCLUSIONS

A. *Managerial Implications*

The research findings provide 3 major implications for online business administrators as follows:

1) The customer expectations, perceived quality, perceived value, and politeness will influence customer satisfaction, but at different scales. Among the 4 antecedents, the perceived quality influence satisfaction most significantly, which means if customers cannot get products or service with good quality, they will be unsatisfied. This is rational for online customers because they usually spend some time on doing research before they purchasing particular items online, and the research work dilutes the impact of perceived expectation and value.

2) The politeness in virtual contexts positively influences customer satisfaction. Thus, to construct a satisfactory virtual commercial environment; online merchants need to take politeness into account, besides those factors including visual design, functionality, operational procedure, and performance of Web sites. Although the findings did not support the direct the causal relationship between politeness and customer loyalty, but customer satisfaction does influence customer loyalty, which still implied the indirect impact brought by politeness on customer loyalty that is a key factor affecting company's performance.

3) There was no significant and negative relationship between customer complaints and loyalty. That means customers who complained about an agent/vendor during the course of a prior transaction still might shop with the same agents/vendors in the future, or they will not incline to the same agent/vendor that they did not complain about. This is not in line with most prior studies adopting the ACSI model. A rational explanation is that customers can find new online merchants with ease, comparing with finding a substitute merchant in physical context. Thus, complaining toward an online merchant looks time-consuming since customers can switch to a new merchant easily, not to mention the processing duration and responses might be unpredictable in virtual contexts. In addition, unhappy online patrons usually tend to file complains toward a customer servant rather than to fill a Web form [45] since they usually can expect to obtain more instant and concrete responses from real persons.

B. *Conclusion and Contribution*

In a civilized society, people dislike verbal and behavioral impoliteness, regardless of contexts. Obviously, various forms of impoliteness in virtual storefronts that customers tend to avoid will be harmful to online merchants. Both prior studies and rationales told us that politeness in e-commerce contexts are well worth notice and consideration.

In view of the politeness issue's significance in the contexts of commerce, this work developed a model and an instrument for examining the effects of the politeness. The findings confirmed that the instrument is reliable and valid, also indicated that 9 out the 11 hypotheses are accepted in the extended ACSI model. This extended model could be used by business to measure the impact of politeness on their customers' satisfaction and loyalty.

C. *Limitation and Future Directions*

One major limitation of this work is that there might exist geographical and cultural factors contributing to the research findings; most surveyed subjects are domestic customers who possess different perspectives toward the politeness from customers in other regions or countries. However, e-business models spread over the globe nowadays. Therefore, broader sampling of subjects is necessary to study the same issue from a global viewpoint. Besides, demographic aspects of subjects including gender, age, income, and occupation may result in some of the differences in customer satisfaction, its antecedents and consequences, and is a worthy topic for further research. Furthermore, this work could be extended by adopting a more delicate research model that take dimensions that are associated with e-commerce, such as Web site usability, service encounter, and trust into consideration.

ACKNOWLEDGMENT

This research work has being funded by the grant from the Ministry of Science and Technology, Taiwan, ROC, under Grant No. MOST 103-2410-H-235-001-. We deeply appreciate their financial support and encouragement.

REFERENCES

[1] B. Whitworth and A. De Moor, "Legitimate by Design: Towards Trusted Socio-Technical Systems," Behaviour & Information Technology, vol. 22, pp. 31-51, 2003.

[2] F. Fukuyama, The End of History and the Last Man. New York: Free Press, 1992.

[3] K. Matzler, E. Sauerwein, and K. Heischmidt, "Importance-Performance Analysis Revisited: The Role of the Factor Structure of Customer Satisfaction," The Service Industries Journal, vol. 23, pp. 112-129, 2003.

[4] Á. Millán and A. Esteban, "Development of a Multiple-Item Scale for Measuring Customer Satisfaction in Travel Agencies Services," Tourism Management, vol. 25, pp. 533-546, 2004.

[5] M. Zineldin, "The Quality of Health Care and Patient Satisfaction: An Exploratory Investigation of the 5qs Model at Some Egyptian and Jordanian Medical Clinics," International Journal of Health Care Quality Assurance, vol. 19, pp. 60-92, 2006.

[6] C. Fornell, M. D. Johnson, E. W. Anderson, J. Cha, and B. E. Bryant, "The American Customer Satisfaction Index: Nature, Purpose, and Findings," the Journal of Marketing, pp. 7-18, 1996.

[7] R. N. Bolton, "A Dynamic Model of the Duration of the Customer's Relationship with a Continuous Service Provider: The Role of Satisfaction," Marketing science, vol. 17, pp. 45-65, 1998.

[8] R. Hallowell, "The Relationships of Customer Satisfaction, Customer Loyalty, and Profitability: An Empirical Study," International journal of service industry management, vol. 7, pp. 27-42, 1996.

[9] A. Parasuraman, V. A. Zeithaml, and L. L. Berry, "A Conceptual Model of Service Quality and Its Implications for Future Research," the Journal of Marketing, pp. 41-50, 1985.

[10] S. L. Nelson and T. R. Nelson, "Reserv: An Instrument for Measuring Real Estate Brokerage Service Quality," Journal of Real Estate Research, vol. 10, pp. 99-113, 1995.

[11] E. Sivadas and J. L. Baker-Prewitt, "An Examination of the Relationship between Service Quality, Customer Satisfaction, and Store Loyalty," International Journal of Retail & Distribution Management, vol. 28, pp. 73-82, 2000.

[12] F. Olorunniwo, M. K. Hsu, and G. J. Udo, "Service Quality, Customer Satisfaction, and Behavioral Intentions in the Service Factory," Journal of Services Marketing, vol. 20, pp. 59-72, 2006.

[13] V. L. Seiler, J. R. Webb, and T. W. Whipple, "Assessment of Real Estate Brokerage Service Quality with a Practicing Professional's Instrument," Journal of Real Estate Research, vol. 20, pp. 105-117, 2000.

[14] C. Nass, "Etiquette Equality: Exhibitions and Expectations of Computer Politeness," Communications of the ACM, vol. 47, pp. 35-37, 2004.

[15] A. von der Pütten, C. Reipen, A. Wiedmann, S. Kopp, and N. Krämer, "The Impact of Different Embodied Agent-Feedback on Users´ Behavior," in Intelligent Virtual Agents. vol. 5773, Z. Ruttkay, M. Kipp, A. Nijholt, and H. Vilhjálmsson, Eds., ed: Springer Berlin Heidelberg, 2009, pp. 549-551.

[16] A. Cooper, The Inmates Are Running the Asylum: Why High Tech Products Drive Us Crazy and How to Restore the Sanity. Indianapolis, Indiana: Sams, 1999.

[17] R. Parasuraman and C. A. Miller, "Trust and Etiquette in High-Criticality Automated Systems," Communications of the ACM, vol. 47, pp. 51-55, 2004.

[18] J. Preece, "Etiquette Online: From Nice to Necessary," Communications of the ACM, vol. 47, pp. 56-61, 2004.

[19] W. G. Skogan, "Citizen Satisfaction with Police Encounters," Police Quarterly, vol. 8, pp. 298-321, 2005.

[20] L. L. Berry, "Relationship Marketing of Services—Growing Interest, Emerging Perspectives," Journal of the academy of marketing science, vol. 23, pp. 236-245, 1995.

[21] K. E. Reynolds and S. E. Beatty, "A Relationship Customer Typology," Journal of retailing, vol. 75, pp. 509-523, 1999.

[22] K. S. Campbell and L. Davis, "The Sociolinguistic Basis of Managing Rapport When Overcoming Buying Objections," Journal of Business Communication, vol. 43, pp. 43-66, 2006.

[23] B. Whitworth, "Politeness as a Social Software Requirement," International Journal of Virtual Communities and Social Networking (IJVCSN), vol. 1, pp. 65-84, 2009.

[24] B. Whitworth, "Polite Computing," Behaviour & Information Technology, vol. 24, pp. 353-363, 2005.

[25] M. Terpstra, T. Kuijlen, and K. Sijtsma, "An Empirical Study into the Influence of Customer Satisfaction on Customer Revenues," Service Industries Journal, vol. 32, pp. 2129-2143, 2012.

[26] C. Clara Xiaoling, "Who Really Matters? Revenue Implications of Stakeholder Satisfaction in a Health Insurance Company," Accounting Review, vol. 84, pp. 1781-1804, 2009.

[27] E. Babakus, C. C. Beinstock, and J. R. Van Scotter, "Linking Perceived Quality and Customer Satisfaction to Store Traffic and Revenue Growth," Decision Sciences, vol. 35, pp. 713-737, Fall2004 2004.

[28] H. Pickle, R. Abrahamson, and A. Porter, "Consumer Satisfaction and Profit in Small Business," Journal of Retailing, vol. 46, p. 38, Winter70/71 1970.

[29] R. F. Gault, "Managing Customer Satisfaction for Profit," Management Review, vol. 82, p. 22, 1993.

[30] C. Fongemie, "Buyer Satisfaction Bolsters Insurer Profits," National Underwriter / Property & Casualty Risk & Benefits Management, vol. 103, p. 3, 1999.

[31] L. Aksoy, B. Cooil, C. Groening, T. L. Keiningham, and A. Yalçın, "The Long-Term Stock Market Valuation of Customer Satisfaction," Journal of Marketing, vol. 72, pp. 105-122, 2008.

[32] D. O'Sullivan, M. C. Hutchinson, and V. O'Connell, "Empirical Evidence of the Stock Market's (Mis)Pricing of Customer Satisfaction," International Journal of Research in Marketing, vol. 26, pp. 154-161, 2009.

[33] C. Fornell, M. D. Johnson, E. W. Anderson, J. Cha, and B. E. Bryant, "The American Customer Satisfaction Index: Nature, Purpose, and Findings," Journal of Marketing, vol. 60, pp. 7-18, 1996.

[34] K.-A. Sun, "Customer Satisfaction, Profitability, and Firm Value in the Hospitality and Tourism Industry: An Application of American Customer Satisfaction Index (Acsi)," 1521021 M.S., University of Missouri - Columbia, United States -- Missouri, 2011.

[35] N. S. n. m. s. a. z. Terblanche, "An Application of the American Customer Satisfaction Index (Acsi) in the South African Motor Vehicle Industry," South African Journal of Business Management, vol. 37, pp. 29-38, 12// 2006.

[36] M. Ogikubo, S. J. Schvaneveldt, and T. Enkawa, "An Empirical Study on Antecedents of Aggregate Customer Satisfaction: Cross-Country Findings," Total Quality Management & Business Excellence, vol. 20, pp. 23-37, 2009.

[37] S. Yang and Y. Zhu, "Customer Satisfaction Theory Applied in the Housing Industry: An Empirical Study of Low-Priced Housing in Beijing," Tsinghua Science & Technology, vol. 11, pp. 667-674, 2006.

[38] S. H. Hsu, "Developing an Index for Online Customer Satisfaction: Adaptation of American Customer Satisfaction Index," Expert Systems with Applications, vol. 34, pp. 3033-3042, 2008.

[39] E. W. Anderson and C. Fornell, "Foundations of the American Customer Satisfaction Index," Total Quality Management, vol. 11, pp. S869-S882, 2000.

[40] "American Customer Satisfaction Index Methodology Report," The Regents of the University of Michigan, Ann Arbor, Ml2001.

[41] B. Whitworth and A. Ahmad, "Polite Computing," in The Social Design of Technical Systems: Building Technologies for Communities, ed: The Interaction Design Foundation, 2013, pp. 77-105.

[42] J. F. Hair, R. E. Anderson, R. L. Tatham, and W. C. Black, Multivariate Data Analysis with Reading (3rd Ed.). New York: Macmillan Publishing Company, 1998.

[43] J. C. Nunnally and I. H. Bernstein, Psychometric Theory. New York: McGraw-Hill, 1994.

[44] P. Holmes-Smith, "Introduction to Structural Equation Modelling Using Lisreal," Perth: ACSPRI-Winter training Program, 2001.

[45] C. McEleny, "Social Media Users Complain Little but Want Fast Response," New Media Age, pp. 04-04, 2011.

Does the Right to be Forgotten Work in Social Machines Context?

Camila Lima, Amanda Nuines, Célio Santana, Fabiola Queiroz
Science Information Department. Federal University of Pernambuco (UFPE). Recife. Brazil
{camila.oalima, almeidanunes23, celio.santana, fabiolaqroz}@gmail.com

Cristine Gusmão
Biomedical Engineering Department. Federal University of Pernambuco (UFPE). Recife. Brazil
cmgg@cin.ufpe.br

***Abstract*— This paper presents a reflection about the right to be forgotten in the context of social machines operating in Web 3.0 and present this "new" distribution of information environment. The reflection was based on a literature review and suggests the ineffectiveness of how the right to be forgotten is being applied due the complex structures existing connection between users, social software and devices (hardware) designed to work together. Thus, disconnecting more than a right, becomes a duty for those who choose to be forgotten.**

Keywords-Social machines; Right to be forgotten; Connection; forgetting; Internet.

I. Introduction

On May 13th, 2014 the Court of Justice of the European Union (CJEU) has taken a decision that caused a huge impact, not only in the legal aspect, but also in the relationship between users and the content related to them on the Internet. The decision referred to search engines, which, from that moment, must enable users, in European Union (EU) territory, to delete their personal information. The court considered that any person "has the right to be forgotten" on the Internet under certain conditions [1].

This decision triggered a worldwide debate about the right to be forgotten and their implications in the current Internet scenario. The central subject was the existing conflict between legal aspects related to the right to be forgotten against freedom of expression and law about data protection in EU [2].

Another issue was about the request made by a citizen requiring “only” that his name did not appear anymore in Google results. But the fact Google hides this information means that they were indeed forgotten? EU legislation is not valid in other countries and the search results cannot be hidden outside EU. This omission indicates that forgetting is more superficial than real [1].

But, even if Google and Facebook were forced to delete all data of a specific citizen, would he/she be forgotten? In this light, this paper presents a reflection on the “right to be forgotten” and how it will work, if there is no change in the way of how it will be implemented, in this social machine context.

Besides this introductory session, this paper presents the following structure: Section 2 will present concepts about forgetting. Section 3 will present the concerns about the “Right to be Forgotten”. Section 4 will present connection concepts. Section 5 will present social machines concepts. Section 6 will present a reflection of this whole scenario. And in section 7 will present final considerations.

II. Forgetting (Oblivion)

When investigating forgetting, we soon realized its inseparability of the concept of memory. It seems more logical to begin this Section explaining concepts related to memory and its connection to oblivion.

Ricoeur [3] suggests that memory can be observed from two approaches: (i) the cognitive one, which refers to the ambition to reproduce or forgotten the past and (ii) the pragmatic approach referred to memory operative side.

Levy [4] also proposes three categories to memory: (i) biological memory: which is that all knowledge was transmitted orally to individuals through narrations, rites and myths. (ii) Support memory: means that facts could be recorded in physical objects. The human memory is not the unique support to retain and preserve information. And (iii) digital memory: these are stored in electronic format using bits and bytes.

When analyzing these categories proposed by Ricoeur, we observed that cognitive approach considers important to determine what are the "traces" left and perceived by individuals in the reconstitution of the memory. In the pragmatic approach, Ricoeur states that there are three types of forgetting: (i) the deep oblivion: which is the one from the deletion of tracks; (ii) forgetting of reserve: which is "necessary" for the proper functioning of memory; and (iii) the manifest oblivion that is exercised intentionally.

Deep oblivion occurs when tracks needed to rebuild memories are not found. Ricoeur suggests that memory in the process of "remembering" follows four steps: (i) persistence, (ii) the remanence, (iii) revival and (iv) detailing. The deep oblivion occurs when there is a lack of “traces” at least in one of these four steps.

The forgetting of reserve is characterized by the deliberate forgetfulness in the everyday life of our memory. Ricoeur himself says: "there is no memory that nothing forget ..." and "forgetfulness would not, in all aspects, be an enemy of memory. The memory should negotiate with forgetting to find, blindfold, the balance between both". The lack of balance led to the "Societal Forgetting" [5].

The last category of forgetting suggested by Ricoeur is the manifest oblivion, which is exercised with some level of intentionality. The manifest oblivion presents another three categories that are: (i) hindered memory, (ii) manipulated memory manipulated and (iii) controlled forgetting also called Amnesty.

To address the hindered memory Paul Ricoeur refers to clinical and therapeutic categories mainly from Freudian psychoanalysis, seeking to link this "pathology" to human and fundamental historical experiences.

The manipulated memory is in the field of power relations. The balance of power, memory and forgetting is forged suggesting a kind of instrumentalization of the memory.

Controlled Forgetting, also called amnesty, has in it something of the reversibility order. The preservation of memory happens through mechanisms of latency and the control of physical supports. The controlled forgetting is related to what the author considered the small miracle of happy memory that is “the recognition”.

Observing the memory classification of Levy, we should begin with the investigation of biological memory that according to Rignano [6] is derived from the connections between neurons and the contact points receiving the denomination of synapses. According to Levy, this memory is more susceptible to forgetfulness.

To minimize the inherent and constant forgetfulness of biological memory, the humans began to use some kind of objects to keep records of their memories. This model was not reflected only in simple transformation of how memories are preserved, but the constitution of a new way of thinking about memory. Coulmas [7] states that memory supports not only mean preservation, but the conditions of memory creation.

Levy [8] states “inscription supports (clay, wax tablets, parchment, papyrus or paper), represented an extension of human biological memory. Thus, writing extended the biological memory transforming it into large long-term memory semantic network.

Finally, we have the digital memory, and on it Garde-Hansen *et al*. [9] suggests that digital mind is susceptible to oblivion: “scanning our memories and the production of new information already in digital media together with the fragility and complexity of maintenance of the files in a virtual environment leads us to create a new concept that threatens the modern world, called digital amnesia.

Cerf [10] suggests that the memory stored in cyberspace is constantly a threat: Cyberspace is a fickle and virtual environment in which the data are in endless movement, succeed, change, interact and mutually exclusive. In cyberspace the issue of preservation of information and knowledge is questioned because, being in the virtual environment, there is no guarantee that this information is available after a certain time, or, if it is available in which format or conditions.

Levy [8] suggests that communication networks and digital memories will incorporate most of the representations and messages produced on the planet thus becoming the main form of human memory.

III. Right To Be Forgotten

Ambrose [11] states that any citizen has the right to not belong to a particular memory, whether collective or individual. In this context, the right to be forgotten, through the right to informational self-determination begins to be exercised all over the world, in view of the many violations committed daily by the media, such as the rights related to honor, privacy and intimacy, all of them, results of constitutional protections given to human dignity. The right to be forgotten comes to guarantee that no one need to be forced to live forever with a past that no longer represents the current condition of an individual.

Hornung, and Schnabel [12] state that the first case in which the informational self-determination was related to digital data processing was observed in 1983 in Germany. The German government, after conducting a general population census, was target of several constitutional complaints that the census directly violates some fundamental rights, particularly the right to free development of personality. The German Supreme Court considers, given the conditions of the automatic processing of data, it is needed an effective protection of the free right of personality, since with electronic data processing, detailed information about personal relationships can be stored indefinitely and consulted at any time.

The right to be forgotten in the EU legislative framework was proposed by the European Parliament on 25 January 2012. Viviane Reding, Vice-President of the European Commission and responsible for Justice, Fundamental Rights and Citizenship areas, announced a reform of legislative framework reserved for personal data protection in EU [13].

Since then, search engines, more specifically Google, become the center of controversy in the EU. The discussion lies in the distinction between data storage services and search engines, and the consequent legal position to which they are subject. For the EU justice, Google is not just a storage service that maintains particular content without liability. Thus, both are similar search engines and data producers, exercising control over the content presented [13].

Peter Fleicher [14] states that Google is nothing more than a tool that promotes facility to finding content, but merely redirects users to the provisions content elsewhere. In his view, the responsibility to eliminate inappropriate content published should lie, above all, to the source of information and not to the search engines.

The EU members believes that any search engines or social networks has the same legal responsibilities of the original sources of information when it the right to be forgotten. This understanding suggests a kind of connection between the sources of information and Internet services.

IV. Connection

It is not known exactly when, and in what context, the word “connection” was created. The earliest reference dates from the second century AD, in the Chinese book “I Ching, the Book of Changes”. In it, the phrase "watch what connects and separate people". This text is applied to a complex context of social capital.

Barabási and Frangos [15] state that a connection between two elements happens when they could be represented by a graph and any one of them can reach the other following a path. These objects do not need to know the existence of these connections to make them real.

Christakis and Fowler [16] suggest that a connection is a set of links, as well as, particular patterns that provide meaning to links. These ties are more important than the people themselves because they determine the existence of networks that are more complex than a "flat" collection of "disconnected" people. Connections affect every aspect of everyday life of an individual. These are links that explain why the whole (network) is greater than the sum of the parts (individuals).

Shaviro [17] states that the connections are important for people to remain visible. If no action is taken to connect, the tendency is that person disappears, after all, someone disconnected is someone who is not part of the system.

Castells [18] relates the concept of connection with network and places. “The network itself cannot suggest a sense of space, but there is a whole series of connections and disconnections of “places” on net. These places are connected globally and physically locally disconnected and socially. Megacities are discontinuous constellations of spatial fragments, functional parts and social sectors, which are all networked.

Levy [19] states that "The human mind works to connect" and suggests: “grasp the development of perception, memory, communication, general connection as a single organic movement that tends to develop a collective intelligence of humanity ... The growing connection between the men is the other side of the world growth”.

Frigyes Karinthy conducted the first, documented; scientific study about connection and it was called "chains". It suggests that people are far apart, on average, in a degree of separation of six. This theory was first confirmed by Stanley Milgram in 1967 and ratified by Duncan Watts [20] that after receiving data collected from 48,000 members of 157 different countries, found the same number “six”. This theory is known as the 'six degrees of separation ".

Christakis and Fowler [21] state that our connections do not end in the people we know, but beyond our social horizon, friends of friends are part of the chain reaction that eventually passes us within a network.

V. SOCIAL MACHINES

The term “Social Machines” was coined by Wade Roush [21] where it was highlighted the role that Internet exerted on people's lives. Almost all interactions between people and electronic devices have Internet involved. Roush pointed the mobile nature of the connection via smartphones and now, the "network" follows the individual to “where” and “when” he wants to. Roush defines a social machine as a mechanism operated by a human who is responsible for the socialization of information between different communities.

Meira *et al.* [22] suggested a more complex definition of social machine: "A social machine is an entity" pluggable "containing an internal processing unit and an interface that waits for requests and responses to other social machines. Its processing unit receives inputs, produces outputs, have states and their connections, intermittent or permanent, define its relations with other social machines ".

Meira [23] suggests that social machines enable any person to "set" his own network, creating their connections and deciding who participates, and how to get involved. Social machines are programmable platforms in the network, whose function and purpose can be, largely extended and redefined by those who hold the knowledge to do it.

Now, instead programming computers as in the past, users will increasingly programming the Internet itself. Programming social machines, each user will be able to create their own applications and provide new forms of articulation and expression network [23].

Burégio *et al*. [24] suggest that social machine has its origins in social computing. It would be an evolution of social software based on the Internet, referred collectively as "Web 2.0". Social machines are based on three pillars that are (i) the social software, (ii) software as sociable entities and (iii) persons as computational unit as shown in Figure 1.

(i) Social Software are those with working on social data, based on Application Programs Interface (API), Web Services or Mashups and have the ability to collaborate with other services and enable users to program "their own Internet".

(ii) Software as sociable Entities means that software has the ability not only store social data, but that they are able to “socialize” autonomously and automatically and thus have "Social Relations" with other software, people and even devices (Internet of Things).

And (iii) People as Computational refer to the effort to integrate people and software in the task of processing of social data. If the software is able to create, process and store data, people also are, and this integration promotes greater social capacity machines. Social machines represent the intersection of these three categories.

Shadbolt *et al.* [25] suggest that the power of the metaphor of social machines comes from the notion that a machine is not just a computer used by users, but rather something purposely designed in a socio-technical system comprising equipment and personnel. Thus, one can view this ecosystem as a set of social interaction machinery and studying each machine becomes only part of the story.

Semmelhack [26] states that transformation promoted by Web 3.0 is deeper than the simple attribution of meaning to the digital objects. This new Internet generation allows that several entities could be created online and from there various services/resources can access these organized and modularized data. An example of this reality portrayed by Semmelheck is the interoperability standard called Microformats that has a specific pattern called H-card, which stores information about people and can be reused on multiple websites.

VI. THE RIGHT TO BE FORGOTTEN AND SOCIAL MACHINES

Before turning our attention to the questions about right to be forgotten in the current structure of social machines, it is worth to mention two aspects of the current EU decision in

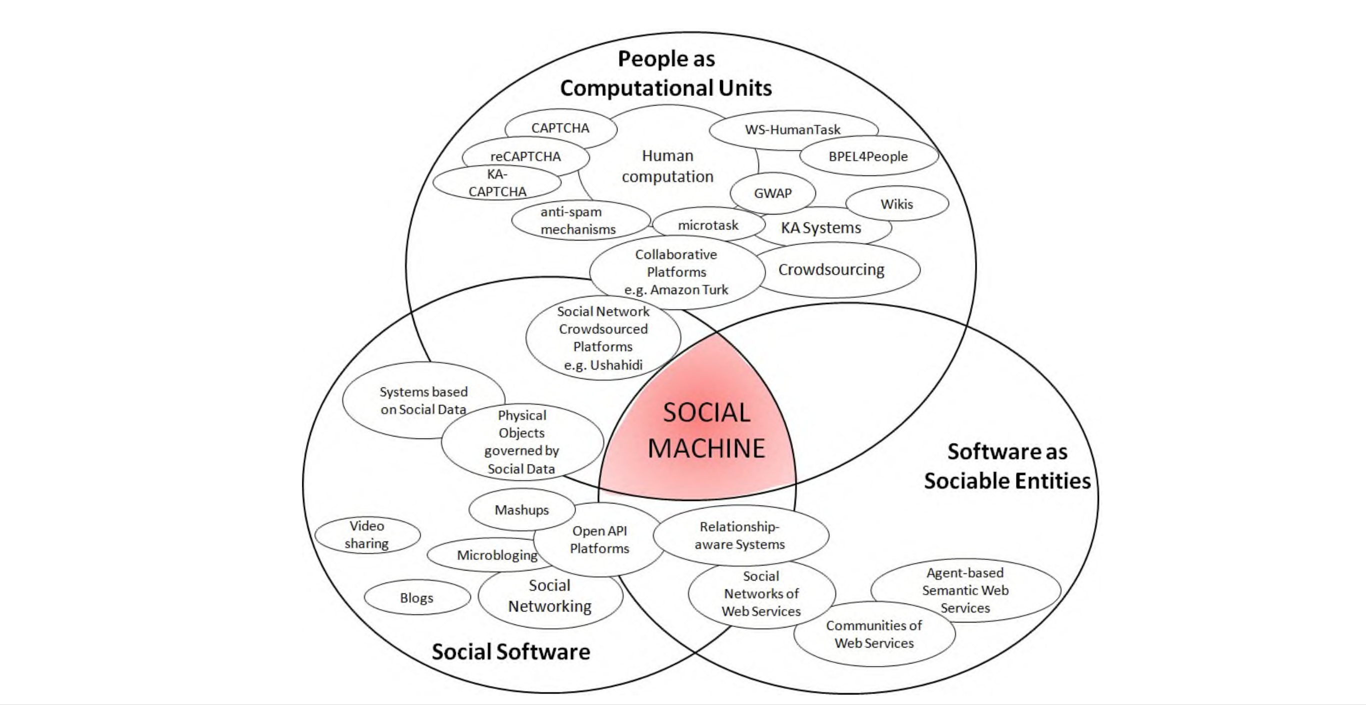


Figure 1. Social Machines Definition - Retrieved from Burégio *et al.* [24]

the case of Google that will not be discussed in this section that are (i) The question of the location of the decision and (ii) the question of the sources of information do not respond jointly to Google in the decisions. These matters are related to legal aspects and not with the structure of the Internet itself.

Following Ricoeur`s [3] statement about forgetting, the right to be forgotten, in this context, falls in manifest oblivion and, more specifically, the manipulated memory since that one organism (CJEU) exercises its power as a justice court that refers about legal aspects. In this case, the manipulation does not have the intention to create a new image from an individual or even reduces the memories for any suitable purpose. This manipulation towards the direction of deep oblivion considering the real intention that is to erase, deliberately, the tracks.

If we consider the structure of web 1.0, where the content was only available to users, data erasing, similar to what has been requested by the CJEU, of a particular service, determines that the user is in fact forgotten by that company since the information contained therein do not suffer any correlation or semantic treatment.

When we consider the web 2.0, also known as "social web", some social networking services store information about users, their connections and their behavior on their networks. Figure 2 shows a fictitious scenario of a user that uses the services Facebook (a), Twitter (b), LinkedIn (c) and Whatsapp (d) and their respective, different, connections on each.

The deletion of data implies, certainly, in forgetting static data and connections. However, the traces inherent to collective memory, such as friends' referrals, demographic data and others that do not belong necessarily to the deleted user will be maintained. Most likely, traces that could restore or reassemble the memory of the deleted user will not exist.

In Web 3.0, the social software enables to use a new kind of service when using "enter using Facebook" option. Now, that service collects from Facebook, the necessary information to carry out the registration. This type of data sharing, known as data skimming, in most cases, is based on the transfer (copy) of the data from the original service (e.g., Facebook) to the requesting service. The available data that could be shared with other services changes the connection structure from several individual networks observed in Figure 2 to a single network shown in Figure 3.

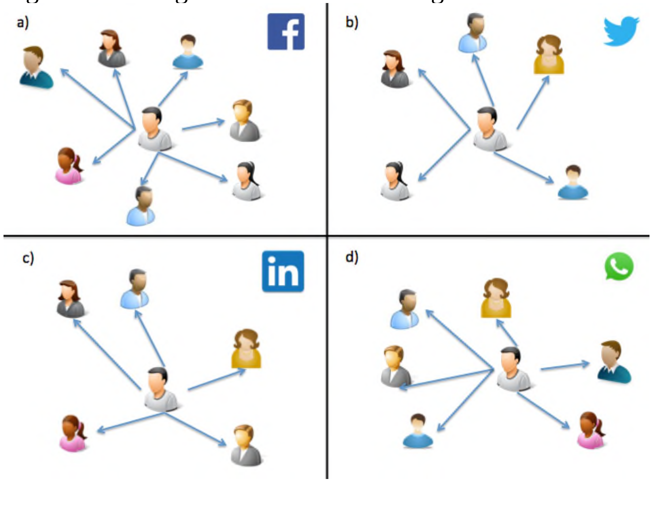


Figure 2. Social Networks of an User in Web 2.0

In this new scenario, we can observe that even Facebook keeps information about User A. Due to the amount of "places" where he exists on Internet, he becomes, in practice, an entity on the Internet because his data, as well as his connections, are spread over the Internet, becoming structures of "information redundancy" scattered in various services.

In scenario (a), User A is connected to User B through Facebook, Twitter and Candy Crush Saga application running on Facebook. If Google does not find data of this relationship on Facebook, these can be found or in Twitter or in Candy Crush Saga. A consideration about this connection is the situation of Candy Crush Saga. Both users run the application from Facebook, but it exists out of Facebook. This means that the Candy Crush Saga also has in its database information about users A and B, and every time that they play the game on Facebook there is an exchange of information between services.

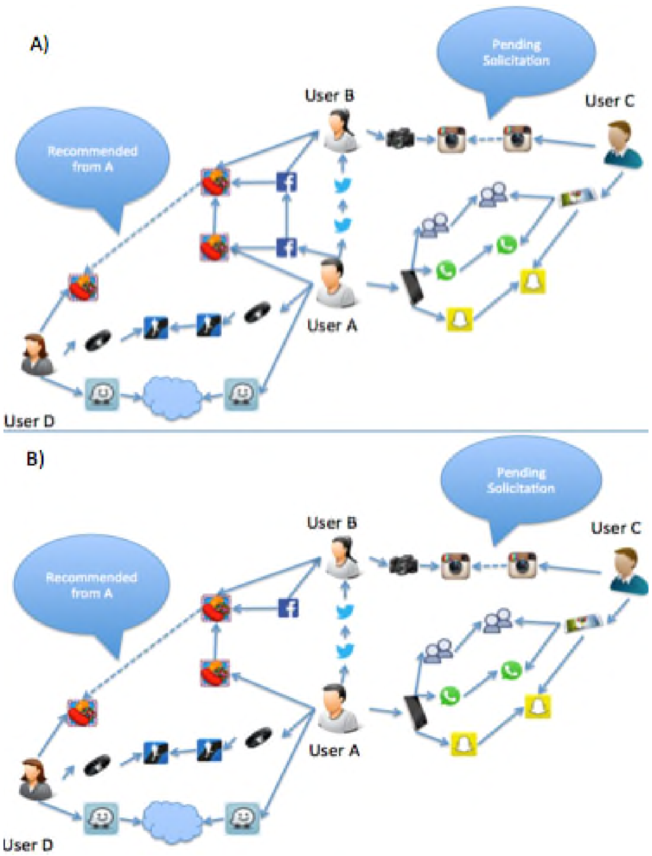


Figure 3. The Web 3.0 Network from an user A, before (a) and after (b), being forggoten by Facebook.

If we consider the connection with User C, we now have a new element part of network of both users: the smartphones. Before the emergence of the Internet of Things, devices such as Tablets, Laptops, Smartphones, Google Glass among others, were considered only “means” to access Internet. However, when we consider that Social machines are also software as entities sociable, the software running on smartphone also contributes to the network.

VII. Final Considerations

The structures of connections and stakeholders in context of social machines suggest a complex flow of information scheme in Internet, particularly as regards the distribution and use of information from social software. Although this article presents only the reflection of the social machines and its relation to the right to be forgotten, is still needed discussion of privacy, connectivity, distribution, use and other issues relating to information in this context.

The connections of an individual are as important as the information itself since they determine, not just for who, but, for what and where information regarding this individual may move. In part, these connections are invisible or transparent to users causing a false sense of control and the naive thought that the data are restricted to that service.

If everything is connected, the right to be forgotten should not be restricted to the exclusion of data on one, or more particular services, but suggests the user disconnection from Internet. This new reality changes the point of view that to be forgotten is a right, but actually, forgotten should be the conscious, intentional and voluntary duty of any citizen, who wants to be forgotten and not return to the Internet. The oblivion, in this network of social machines, without disconnection is, at least, paradoxical, since when an individual connects to the Internet, the network should remember that he/she is a forgotten one.

More than apparent, this forgetfulness is naive and imposes penalties on precisely those who would be the main allies in this task, search engines, which are able to identify the sources of information and their connections. If, we want to reflect about the effectiveness of actions taken toward oblivion we should rethink the role of Internet, as whole, and the role of its services.

Finally, we believe that the right to be forgotten, in the way of how it is being promoted, only transformed, in a very subtle way, how the information is used by Internet services. The data will no longer be visible to regular users of the search engines, but may continue to be collected and interpreted freely to other types of business transactions. They move the actors, but not change the actual information flows and the memory built about a particular user. This information remains unscathed and connected, but invisible.

Aware of all these conclusions we recognize that further researches should be performed on: how and how much “our information” is spread on Internet and which are the sources of this information? This future work is related to the degree of influence of “robots” that are sharing and collecting our information. Is our information private? Or they start a viral process on Internet? Further research is still necessary on how the user behavior impacts this process of information sharing on Internet.

REFERENCES

[1] D. Sullivan, “How Google’s New ‘Right To Be Forgotten’ Form Works: An Explainer”. 2014. Available in: <http://searchengineland.com/google-right-to-be-forgotten-form-192837>. Last access in: Dec, 16th, 2014.

[2] C. Preece and R. Clark, “Google ‘right to be forgotten’: Everything you need to know”. 2014. Available in: <http://www.itpro.co.uk/security/22378/google-right-to-be-forgotten-everything-you-need-to-know>. Last access on Jan. 12th, 2015.

[3] P. Ricoeur, Memory, history, forgetting. University of Chicago Press, 2004.

[4] P. Lévy, Cyberculture. University of Minnesota Press, 2001.

[5] V. Mayer-Schönberger, Delete: the virtue of forgetting in the digital age. Princeton University Press. 2011.

[6] E. Rignano, Biological memory. Routledge, 2013.

[7] F. Coulmas, The writing systems of the world. Blackwell. 1989.

[8] P. Levy. Les Technologies de l’intelligence. L’avenir de la pensée à l’ère informatique. La Découverte. 1990.

[9] J. Garde-Hansen, A. Hoskins and A. Reading, Save as... digital memories. Palgrave Macmillan. 2009.

[10] The Guardian. “The Digital Blackhole: Will it Delete your Memories?”. 2015. Available in: < http://www.theguardian.com/technology/2015/feb/16/digital-black-hole-delete-memories-information-lost-google-vint-cerf Last access on Feb. 20th. 2014.

[11] M. L. Ambrose, It’s about time: Privacy, information lifecycles, and the right to be forgotten. Stanford Technology Law Review, Vol.16. 2012

[12] G. Hornung and C. Schnabel. Data protection in Germany I: The population census decision and the right to informational self-determination. Computer Law & Security Review Vol.25. No. 1. 2009, pp. 84-88.

[13] European Union, “The EU Data Protection Reform 2012: Making Europe the Standard Setter for Modern Data Protection Rules in the Digital Age”. Available in: <http://europa.eu/rapid/press-release_SPEECH-12-26_en.htm>. Last Access: Jan. 12th, 2015.

[14] P. Fleicher, “Our thoughts on the right to be forgotten”. 2012. Available in: < http://googlepolicyeurope.blogspot.com.br/2012/02/our-thoughts-on-right-to-be-forgotten.html>. Last Access in. Jan. 16th, 2015.

[15] A. Barabási and J. Frangos. Linked: the new science of networks science of networks. Basic Books. 2014.

[16] A. Christakis and J. Fowler. Connected: The surprising power of our social networks and how they shape our lives. Little Brown. 2009.

[17] S. Shaviro, Connected: Or what it means to live in the Network Society. University of Minnesota Press. 2003.

[18] M. Castells,“The rise of the network society: The information age: Economy, society, and culture. Vol. 1. John Wiley & Sons. 2011.

[19] P. Levy, World Philosophie: le marché, le cyberespace, la conscience. 2000.

[20] D. Watts, Six degrees: The science of a connected age. WW Norton & Company. 2004.

[21] W. Roush, Social machines: Computing means connecting. Technology Review-Mancheser, Vol. 108. No. 8. 2005.

[22] S. Meira, et al., “The emerging web of social machines” Proc. of Computer Software and Applications Conference (COMPSAC), IEEE Press, 2011, pp. 26-27.

[23] S. Meira, “O Meio é... Programável!” Available in <http://www1.folha.uol.com.br/fsp/mercado/me2207201026.htm.> Last access Jan. 10th, 2014.

[24] V. Buregio, S. Meira and N. Rosa. Social machines: a unified paradigm to describe social web-oriented systems. Proc. 22nd international conference on World Wide Web companion. 2013.

[25] N. Shadbolt, D, Smith, E. Simperl, M. Van Kleek, Y. Yang and W. Hall, Towards a classification framework for social machines. Proc, 22nd international conference on World Wide Web companion. 2013, pp. 905-912.

[26] P. Semmelhack, Social Machines: How to Develop Connected Products that Change Customers' Lives. John Wiley & Sons. 2013.

Crawling and Mining Social Media Networks: A Facebook Case

Abd El Salam Al Hajjar, Haissam Hajjar, Mazen El Sayed, Mohammad Hajjar
Institute University of Technology
Lebanese University
Lebanon
e-mail: abdsalamhajjar@hotmail.com, haissamh@ul.edu.lb, mazen_elsayed@yahoo.fr, m_hajjar@ul.edu.lb

***Abstract*—Social media is computer-mediated tool that allows people to create, share or exchange information, ideas, and pictures/videos in virtual communities and networks. The most popular social media in these days is Facebook. This paper treats the problems of web crawling and mining. More precisely, we focus our intention on Facebook information extraction, and its importance in gathering data and tracking people. Also, we focus on the crawling mechanism of several Facebook pages and extracting the data contained in it and finally storing it in our database. Specifically, we extract the basic information, such as home town, age, work, education, friends list, events and other information. At first, we describe the content of Facebook web pages and the principle of web crawling and mining. Second, we propose the architecture of our system, which allows extracting the Facebook information, for a specific user logged in. The result of our work is an automated system that takes a Facebook user as an input, extracts recursively the list of friends for this user, and returns the friends information (name, university, etc.).**

Keywords-Web data crawling and mining; social media network; Facebook; HTML; PHP; MySQL; database.

I. INTRODUCTION

The “Social media” axiom is widely used these days; it is a computer-mediated tool that allows people to create, share or exchange information, ideas, and pictures/videos in virtual communities and networks [2]. Social media is defined as "a group of Internet-based applications that build on the ideological and technological foundations of web and that allows the creation and exchange of user-generated content". Furthermore, social media depends on mobile and web-based technologies in order create highly interactive platforms through which individuals and communities share, co-create, discuss, and modify user-generated content [4].

Social media is different from other traditional or industrial media in many ways, including quality, usability, immediacy, and permanence [3]. Internet users spend more time with social media sites than any other type of site. Furthermore , the total time spent on social media in the U.S. increased by 99 percent to 121 billion minutes in July 2012 compared to 66 billion minutes in July 2011 [4][5].

In 2014, the most popular used social network was Facebook, comparing to other networks such as Twitter, Instagram, LinkedIn, Pinterest, etc. [6].

Facebook is an online social network service. Its website was launched on February 4, 2004, by Mark Zuckerberg with his college roommates and fellow Harvard University students Eduardo Saverin, Andrew McCollum, Dustin Moskovitz and Chris Hughes [7]. The founders had initially limited the website's membership to Harvard students, but later expanded it to colleges in the Boston area, the Ivy League, and Stanford University. It gradually added support for students at various other universities and later to high-school students. Facebook now allows anyone who claims to be at least 13 years old to become a registered user of the website [8]. Its name comes from a colloquialism for the directory given to it by American universities students [9]. After registering to use this type of site, users can create a user profile, add other users as "friends", exchange messages, post status and photos, share videos and receive notifications when others update their profiles. Additionally, users may join common-interest user groups, organized by workplace, school or college, or other characteristics, and categorize their friends into lists, such as "people from work" or "close Friends". Facebook had over 1.3 billion active users as of June 2014 [10]. Due to the large volume of data collected about users, the service's privacy policies have faced scrutiny, among other criticisms. Facebook held its initial public offering in February 2012 and began selling stock to the public three months later, reaching a peak market capitalization of 104 billion.

A social media network has a large volume of data; therefore, it will provide a big amount of useful information (people pictures, relations, etc.). So, it is very important to follow a methodology to extract and analyze information from social media web site.

In this work, we collect the data from the social media network, specifically Facebook, organize them in a database, analyze these data, and show the information on screen. The presented information can be a list of all friends at many levels (e.g., level1 gives the only friends of the selected person, level2 presents also the friends of each friends. etc.), and it can be events as like/comment/share for each friends already extracted, etc.

In the next section, we present the web crawling and mining definition and concept. Section 3 presents the social media and Facebook. Section 4 presents the Facebook web crawling and mining methodology, and presents the Facebook content with the architecture of the system. Section 5 presents the result. Section 6 describes the conclusion and the future work.

II. WEB CRAWLING AND MINING

A web crawler is an Internet bot (a bot is an automated application used to perform simple and repetitive tasks that would be time-consuming, impossible for a human to perform), that systematically browses the World Wide Web, typically, for the purpose of web indexing. A Web crawler may also be called a web spider [11]. Web search engines and some other sites use web crawling to update their web content or indexes of others sites' web content. Web crawlers can copy all the pages they visit for later processing by a search engine that indexes the downloaded pages so that users can search them much more quickly. Web crawlers can validate hyperlinks and HTML (HyperText Markup Language) code. They can also be used for web scraping (see also data-driven programming) [12]. Within the past few years there has been an increase of free web crawler datasets [13]. The challenges become increasingly difficult when doing this on a larger scale. However, capturing the content of a single web page is quite easy. Below, we present the main steps of the web crawler:

i. Select a URL (Uniform Resource Locator).
ii. Fetch and parse the corresponding page.
iii. Save the important content into database.
iv. Extract URLs from this page.
v. Add URLs to queue.
vi. Select a URL and repeat.

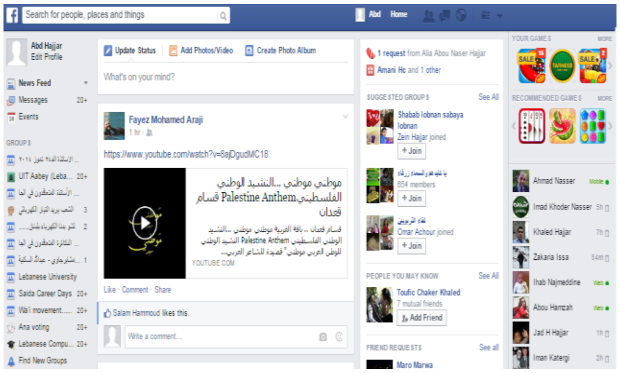
Data-mining is the analysis step of the Knowledge Discovery in Databases (KDD) process [14]; it is allowed to process big amounts of data to provide meaningful and relevant information. The collected information are in an unstructured form, must be transformed into a structured format to be suitable for processing. The data mining technology is coming from a huge evolution; the new and better technique is made available continually to gather whatever information is required [15]. The term “web data mining” is a technique used to crawl through various web resources to collect required information, which enables organizations and individuals to gather information, and to utilize this information in the best interest [16]. The advantage of the web data mining can be shown in the following general example: a company is thinking about launching a new product of cotton shirts, through the client databases founded on web, so they can clearly determine how many clients have placed orders for cotton shirts over the last year and how much profit such orders have brought to the company. The disadvantage is resumed by losing the user privacy when individual information is obtained, used, and distributed, especially if this happens without the user knowledge.

III. SOCIAL MEDIA NETWORK, FACEBOOK CASE

Social media network is the cooperation of online communications channels dedicated to community-based input, interaction, content-sharing, and collaboration [17]. Social networking is the practice of growing the number of one's business and social contacts by making connections through peoples. While social networking becomes societies themselves, the unparalleled potential of the Internet to promote such connections is only now being fully recognized and exploited, through web-based groups established for that purpose[1][2] (see Figure 1).

Figure 1. Social media networking.

The most popular social media networking sites in the world are Facebook, Google+, LinkedIn, Instagram, and Twitter, etc.



Fayez Mohamed Araji
1 hr ·

https://www.youtube.com/watch?v=8ajDgudMC18

موطني موطني ...النشيد الوطني الفلسطيني Palestine Anthem قسام قعدان

YOUTUBE.COM

Salam Hammoud likes this.

1 request from Alia Abou Naser Hajjar

Amani Hc and 1 other

Shabab lobnan sabaya lobnan
Zein Hajjar joined

654 members

Omar Achour joined

Toufic Chaker Khaled
7 mutual friends

Maro Marwa

Figure 2. Facebook general layout

Facebook represents a potentially useful tool in educational contexts (see Figure 2). It allows for both an asynchronous and synchronous, open dialogue via a familiar and regularly accessed medium, and supports the integration of multimodal content, such as student-created photographs, video, and URLs to other texts, in a platform that many students are already familiar with. Further, it allows students to ask more minor questions that they might not feel motivated to visit a professor personally during office hours. It also allows students to manage their own privacy settings, and often work with the privacy settings they have already established as registered users [18].

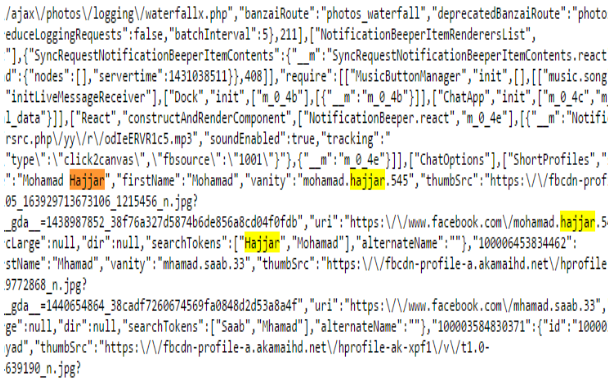
IV. FACEBOOK CRAWLING AND MINING METHODOLOGY

In this section, we present the crawling and mining methodology for a social media networking, specifically, a Facebook case. We describe the Facebook web pages content and the system architecture that allows crawling and mining automatically the Facebook information for a specific user logged in.

A. *Facebook Content*

The first page in Facebook is the login page that allows authenticating a user. If this user is authenticate, the user can navigate different pages in the Facebook website, starting from the profile that includes the following: basic information, messages, photos, friends, notes, status, comments, groups, pages, and the wall. A user is able to search for friends by e-mail address, or just by typing a name of the friend. When people become friends, they are able to see all of each other’s profiles including contact information. The user also can create groups. This group allows members who have common interests to find and interact with each other. Also, the user can create Facebook pages that contain many members (other Facebook users). Beside all that, Facebook contains a set of games which allows users to play online, etc.

The target is crawling and extracting data from the Facebook pages, then, saving the data into a database, using a specific web programming language. When, Facebook is not anymore a secure network, we can get the html source of the page using any browser. However, some data, maybe for privacy reasons cannot be viewed. So, we only collect public data organize and view them in the simplest desired form.



```
/ajax\/photos\/logging\/waterfallx.php","banzaiRoute":"photos_waterfall","deprecatedBanzaiRoute":"photo
educeLoggingRequests":false,"batchInterval":5},211],["NotificationBeeperItemRenderersList",
"],{"SyncRequestNotificationBeeperItemContents":{"__m":"SyncRequestNotificationBeeperItemContents.react
d":{"nodes":[],"servertime":1431038511}},408]],"require":[["MusicButtonManager","init",[],[["music.song
"initLiveMessageReceiver"],["Dock","init",["m_0_4b"],[{"__m":"m_0_4b"}]],["ChatApp","init",["m_0_4c","m
l_data"}]],["React","constructAndRenderComponent",["NotificationBeeper.react","m_0_4e"],[{"__m":"Notifi
rsrc.php\/yy\/r\/odIeERVR1c5.mp3","soundEnabled":true,"tracking":"
"type\":\"click2canvas\",\"fbsource\":\"1001\"}"},{"__m":"m_0_4e"}]],["ChatOptions"],["ShortProfiles","
":"Mohamad Hajjar","firstName":"Mohamad","vanity":"mohamad.hajjar.545","thumbSrc":"https:\/\/fbcdn-prof
05_16392971367310б_1215456_n.jpg?
_gda__=1438987852_38f76a327d5874b6de856a8cd04f0fdb","uri":"https:\/\/www.facebook.com\/mohamad.hajjar.5
cLarge":null,"dir":null,"searchTokens":["Hajjar","Mohamad"],"alternateName":""},"100006453834462":
stName":"Mhamad","vanity":"mhamad.saab.33","thumbSrc":"https:\/\/fbcdn-profile-a.akamaihd.net\/hprofile
9772868_n.jpg?
_gda__=1440654864_38cadf7260674569fa0848d2d53a8a4f","uri":"https:\/\/www.facebook.com\/mhamad.saab.33",
ge":null,"dir":null,"searchTokens":["Saab","Mhamad"],"alternateName":""},"100003584830371":{"id":"10000
yad","thumbSrc":"https:\/\/fbcdn-profile-a.akamaihd.net\/hprofile-ak-xpf1\/v\/t1.0-
639190_n.jpg?
```

Figure 3. Facebook HTML generated code

In general, most Facebook data is very important, since it allows getting information about someone (e.g., who likes swimming or whom wearing jeans), collecting data for investigations purposes (e.g., crimes, retrace criminals), and for statistics and analysis purpose (e.g., number of likers for a specific event). PHP (Hypertext Preprocessor) presents the main programming language of the Facebook frontend, which is suited for web development and can be embedded into HTML. PHP is an open source, support object-oriented, powerful built in functions. PHP can works and connects with several databases, such as MySQL (My Structured Query Language), Oracle, etc., and it can manipulate XML (Extensible Markup Language) documents. For that, we can use the PHP programming language for the data crawling and mining reasons (see Figure 3).

B. *Architecture of Facebook Crawling and Mining system*

The login page presents the entrance into Facebook according to a specific user account (username or email, and password); if this user is authenticated, then we can directly access the full information about it, such as news feed, group information, friends, photos, events, comments, etc. All this information is presented in several HTML generated pages; knowing that, the HTML code is generated from many others programming languages. Inside this HTML code, we have all the data viewed on the web browser, so, we can crawl and extract the data from the opened Facebook user account page. This data may be the list of friends, information about each friends, events and notifications.

The data crawling procedure will automatically occur according to an automated system. Our developed automated system allows takes taking as an input a Facebook user, fetching the Facebook HTML pages, splitting the HTML code according to specifics delimiters (rules) into a more manageable portion, removing the unwanted HTML tags, reformatting HTML, adjusting spaces, removing entities, matching content with regular expressions, and storing the pertinent content into a structured MySQL database for future data mining use. The system database structure and algorithm will describe in detail below.

1) Database design

The parsing presents a main step in the data crawling process; it is based on specific characters and symbols that must be defined according to the text to be analyze. In the Facebook crawling, the parsing process must split the HTML code according to specific tags that can be used later as rules. For that, we save these rules in the database (in the “rules” table). Later, for each parsing process, we will select the corresponding rules from the database.

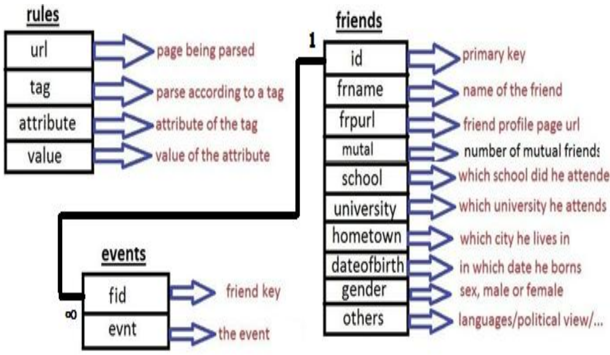


Figure 4. Database diagram

For a specific Facebook account, we can recursively extract a list of friends, and information about these friends (friend name, URL of the friend profile page, number of mutual friend according to the authenticated account, school, university, hometown, dateofbirth, gender, others). For each extracted friend, we can extract their events which saved on the table events (Figure 4).

2) Algorithms

In this section, we will present an algorithm that describes several operations on Facebook crawling and mining. Firstly, the operation allows extracting recursively the list of friends for a specific Facebook user, e.g., for a given Facebook user, we will extract the list of their friends L1, and, for each one in the list L1, we will extract their friends, and so on. Also, in this section, we will present the other operations that allow extracting other information, such as: events, comments, etc.

Figure 5 shows the architecture of the system algorithm for Facebook data crawling automatically starting from the home page of a specific Facebook authenticated user.

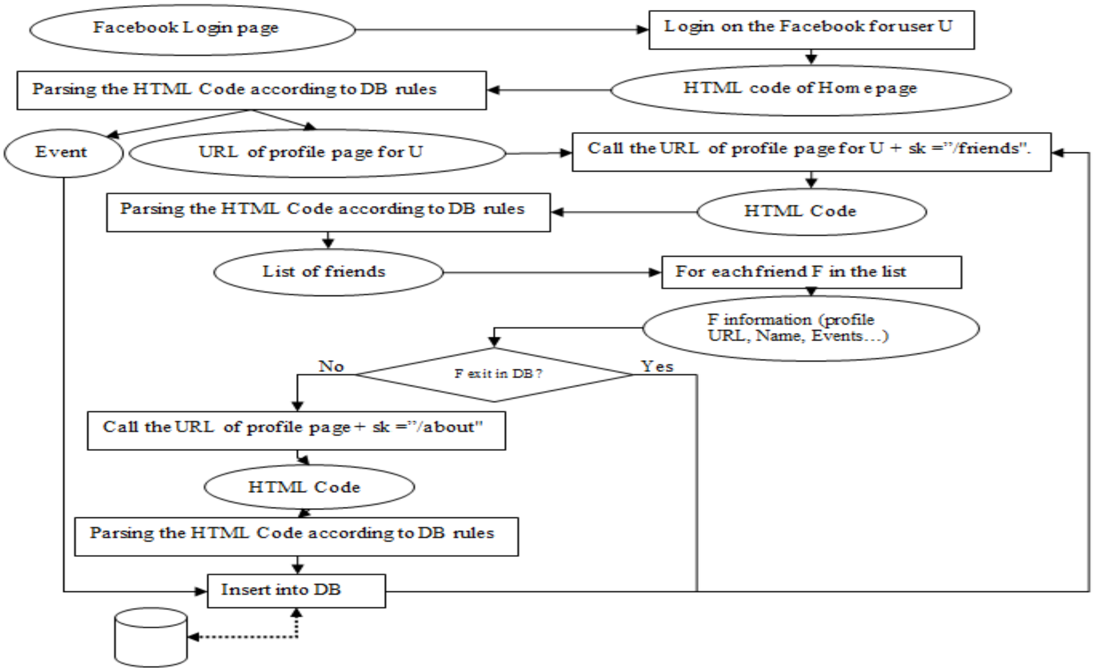

Facebook Login page
Login on the Facebook for user U
Parsing the HTML Code according to DB rules
HTML code of Home page
Event
URL of profile page for U
Call the URL of profile page for U + sk ="/friends".
Parsing the HTML Code according to DB rules
HTML Code
List of friends
For each friend F in the list
F information (profile URL, Name, Events…)
No
F exit in DB?
Yes
Call the URL of profile page + sk ="/about"
HTML Code
Parsing the HTML Code according to DB rules
Insert into DB

Figure 5. General architecture of Facebook Crawling and Mining system.

This algorithm is composed of several operations, as follow: firstly, we get the HTML of the home page for a specific Facebook user U and parse this page according to the database rules; the objective of parser is to transform an HTML code into data and information. Parsing an HTML code is done in several stages. Firstly, we determine the main delimiters that allow extracting some information; these delimiters are HTML tags. We determine these delimiters by navigation and analysis of the Facebook HTML code, then extract manually these delimiters, and save them as rules in the database; for example in the Facebook HTML code: friends name exist between the two tags <Friendtag1> FriendName </ Friendtag1>, then these 2 tags (<Friendtag1> and </Friendtag1>) must be saved in the table rules. Then, the table rules present the reference of parsing stage in the Facebook crawling system. For example, to extract friend from a Facebook HTML code, firstly, we must get the specific rules of friends from the table rules in our database. The parsing stage allows extracting all the events and the URL of this user profile page, the list of all friends for the user U are presented in a page that have as URL the concatenation between the URL of the user U profile page and "/friends". From this new URL, we can extract, after parsing, the list of all friends. For each friend in list, we can follow the same procedure to getting their information, such as friend name, friend events, friend profile URL page, etc., according to their information (name, date of birth, etc.). We can test whether this friend is already added into DB in the table friends; if it is new, it must go to the About page, the About page URL can be done by concatenating the friend profile URL and "\About"; then, we can extract other information, such as school, university, etc. Finally, all the extracted information (friend profile page URL, friend name, friend events, etc.) will save in the database (in the 2 tables friends and events). Otherwise, if this friend already exists in the database; then, we must get the URL of their profile page, and we will follow the same procedure to get their friends; in this case, we extract all the friends of friends for the user U, and so on.

V. RESULT

The main result of this work is an automated Facebook crawling and mining system. This system take as input a Facebook user and give as output big amount of information about this user such as friends, comments, likers, etc. It is based on set of rules saved manually in the database. This system allows extracting recursively the list of friends for a specific Facebook user, and storing the entire extracted friends' information into a MySQL database for future use. The start point of this system is the home page for a specific Facebook authenticated user. After that, we apply the system operations (see Figure 5) in order to collect the pertinent information for all friends, friends of friends, friends of friends of friends, etc., of the logged user. We evaluate the system on several Facebook users, for each one it returned the demanded information. For that, we will describe the result as an example of the user "Joe", in order to explain that the result of our work is a system that return the pertinent information about a given. Let us consider that we have a Facebook user named

"Joe@hotmail.com", presented as input to our system. The system takes the HTML code of the home page, and selects from the database the rules (or tags). These rules are already saved manually in the rules table. Next, the system allows extracting the profile page link for Joe, according to the selected rules (for example :< a class="_2dpe _1ayn">, <title="Profile">). Then, we apply the parsing operation according to the extracted tags in order to deliver the profile page link "https://www.facebook.com/profile.php?id=1013"; from this URL, we can explore the HTML code in order to extract the friends list, by applying the parsing operation according to the stored rules of friends list. For each friend F in the extracted list, we can apply the same procedure to extract its profile page (profile page of friend F). The profile page includes pertinent information such as friends, photos, posts, etc.

We have several purposes for gathering the information from the system, which focused on extracting many facts about a person. For example, we can view that (Joe) is not friend with (George), but he is related to him on FaceBook because (Joe) is friend with (Mario), (Mario) is friend with (Elie), and (Elie) is friend with (George). Also, we can analyze the stored data and conclude that (Joe)'s hobby is politics, because the collected data describe that (Joe) likes several politic pages, and he likes the politic events.

VI. CONCLUSION AND FUTURE WORK

Social media is becoming an integral part of online life, as social websites and applications proliferate. Most traditional online media include social components, such as comment fields for users. In business, social media is used to market products, promote brands, and connect to current customers and foster new business.

Social media analytics is the practice of gathering data from social media websites and analyzing that data to make business decisions. The most common use of social media analytics is to mine customer sentiment to support marketing and customer service activities.

In this paper, we are interested to study the problem of gathering information from Facebook pages and storing them in a database. More specifically, one gathers the basic information, such as home town, age, work, education, friends, events and much other information. First, we describe the content of Facebook web pages and the principle of web crawling and web data mining. Second, we proposed the architecture of our automated system that allows crawling and mining the Facebook information for a specific user logged in.

In future works, we plan to study the problem of web crawling and web data mining in the cases of other social media websites like YouTube, Instagram, etc. We also plan to study the way in which we use gathering data in the database.

ACKNOWLEDGMENT

This work has been done as a part of the following project: "Automatic information extraction form Arabic texts" supported by the Lebanese National Centre of Scientific Research (CNRSL).

REFERENCES

[1] A. Kaplan and M. Haenle, "Users of the world, unite! The challenges and opportunities of social media", Business Horizons, vol.53 (1), pp. 61, 2010.

[2] J. Kietzmann, K. Hermkens, I. McCarthy, and B. Silvestre, "Social media? Get serious! Understanding the functional building blocks of social media", Business Horizons, vol. 54, no. 3, 2011.

[3] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. "Finding high-quality content in social media", WISDOM – Proceedings of the 2008 International Conference on Web Search and Data Mining: 183–193, 2008.

[4] Nielsen Holdings,"State of the media: The social media report 2012", Featured Insights, Global, Media + Entertainment. Nielsen. Retrieved 9 December 2012.

[5] Q. Tang, B. Gu, Bin, and A. Whinston, "Content Contribution for Revenue Sharing and Reputation in Social Media: A Dynamic Structural Model", Journal of Management Information Systems, vol. 29, no. 2, pp. 41-76, Fall 2012 .

[6] Nielsen Holdings, "The U.S. Digital Consumer Report". Featured Insights, Global, Media + Entertainment. Nielsen [Retrieved November 25, 2014].

[7] N. Carlson,"At Last – The Full Story Of How Facebook Was Founded", Business Insider, http://www.businessinsider.com/, March 5, 2015.

[8] R. E. Cash, "Depression In Young Children: Information For Parents And Educators". Facebook Retrieved, Social/Emotional Development, November 22, 2011.

[9] E. Eldon, "2008 Growth Puts Facebook In Better Position to Make Money", VentureBeat(San Francisco). [Retrieved December 19, 2008].

[10] M. Zuckerberg, "Company Info | Facebook Newsroom", Facebook newsroom, http://newsroom.fb.com/company-info/. September 30, 2014.

[11] J. Wu, P. Teregowda, M. Khabsa, S. Carman, D. Jordan, J. Wandelmer, X. Lu, P. Mitra, and C. Giles "Web crawler middleware for search engine digital libraries: a case study for citeseerX", In proceedings of the twelfth international workshop on Web information and data management pp. 57-64, Maui Hawaii, USA, November 2012.

[12] Y. Sun, "A comprehensive study of the regulation and behavior of web crawlers". A comprehensive study of the regulation and behavior of web crawlers, Publisher: Pennsylvania State University, 2008.

[13] OutWit Technologies, "OutWit Hub - Find, grab and organize all kinds of data and media from online sources". Outwit.com. 2014-01-31. [Retrieved March 20, 2014].

[14] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in Databases". American Association for Artificial Intelligence, fall 1996.

[15] S. Chakrabarti, "Data Mining Curriculum: A Proposal (Version 0.91)". Group of ACM SIGKDD Curriculum Committee, August 5, 2004.

[16] K. Wahlstrom, J. Roddick, R. Sarre, V. Estivill-Castro, and D. de Vries, "Legal and Technical Issues of Privacy Preservation in Data Mining". Legal and Technical Issues of Privacy Preservation in Data Mining, 2007.

[17] D. Boyd and N. Ellison, "Social Network Sites: Definition, History, and Scholarship". Journal of Computer-Mediated Communication, vol.13, pp. 210–230, 2008.

[18] M. Moody, "Teaching Twitter and Beyond: Tip for Incorporating Social Media in Traditional Courses". Journal of Magazine & New Media Research, vol. 11(2): pp. 1-9, 2010.

A Semantic Risk Management Framework for Digital Audio-Visual Media Preservation

Vegard Engen, Galina Veres, Simon Crowle, Maxim Bashevoy, Paul Walland, Martin Hall-May
IT Innovation Centre
University of Southampton
Southampton, United Kingdom
Email: {ve, gvv, sgc, mvb, pww, mhm}@it-innovation.soton.ac.uk

Abstract—**Digitised and born-digital Audio-Visual (AV) content presents new challenges for preservation and Quality Assurance (QA) to ensure that cultural heritage is accessible for the long term. Digital archives have developed strategies for avoiding, mitigating and recovering from digital AV loss using IT-based systems, involving QA tools before ingesting files into the archive and utilising file-based replication to repair files that may be damaged while in the archive. In this paper, we focus on dealing with issues resulting from system errors, rather than random failure or corruption; issues caused by the people, systems and processes that handle the digital AV content. We present a semantic risk management framework designed to support preservation experts in managing workflow risks, combining workflow and risk specification within a risk management process designed to support continual improvement of workflow processes.**

Keywords–Risk management; semantic modelling; business processes; workflows; media preservation; digital archives.

I. Introduction

Digital preservation aims to ensure that cultural heritage is accessible for the long term. From the 20th century onwards, AV content has provided a significant record of cultural heritage, and increasing volumes of AV content that have been digitised from analogue sources or produced digitally present new preservation challenges. The focus is no longer on reducing damage to the physical carrier by maintaining a suitable environment; rather, archives must ensure that the significant characteristics of the content, represented digitally, are not lost over time. Digital data enables easier transfer, copying, processing and manipulation of AV content, which is at once a boon but also a problem that requires continuous and active management of the data.

Digital damage is defined here as any degradation of the value of the AV content with respect to its intended use by a designated community that arises from the process of ingesting, storing, migrating, transferring or accessing the content. The focus here is on strategies that can be used to minimise the risk of loss. In particular, we focus on dealing with issues resulting from system errors, rather than random failure or corruption, considering the risks to the AV content as it is being manipulated by various activities in a workflow process. This includes the people, systems and processes put in place to keep the content safe in the first place.

Archival processes dealing with digital AV content are underpinned by IT systems. In the few years that archives have been working with digitised and born-digital content, best practice in terms of digital contents management has rapidly evolved. Strategies for avoiding, reducing and recovering from digital damage have been developed and focus on improving the robustness of technology, people and processes. These include strategies to maintain integrity, improve format resilience and interoperability, and to combat format obsolescence.

A business process risk management framework (BPRisk) has been developed in the EC FP7 DAVID project [1], which combines risk management with workflow specification. BPRisk has been designed to support a best practice approach to risk management of digital AV processes (and thus the content itself). In this paper, we will give an overview of this framework, but focus on the semantic modelling and risk specification aspects. Within the DAVID project, this research and development has been conducted to provide a tool to help prevent damage to digital AV content. In addition to this, the DAVID project focuses on understanding damage (how it occurs and its impact), detecting and repairing damage, and improving the quality of digital AV content.

The BPRisk framework is generic in nature, supporting risk specification for Business Process Modelling Notation (BPMN) 2.0 [2] workflows in any domain. The framework utilises a novel semantic risk model developed in the project that encapsulates domain knowledge generated in the DAVID project on known risks (and controls) associated with activities in a controlled vocabulary for the domain of digital preservation (also developed in the project). This enables the framework to be an effective support tool to users who are typically not familiar with formal risk management. The semantic risk modelling provides the domain experts with a starting point for doing risk analysis, but semantic reasoning is utilised to enable suggestions on risks and controls for the activities in the workflows at design time.

In the remainder of this paper, we will further discuss the challenges and related work on digital preservation in Section II and risk management in this domain in Section III. Thereafter, in Section IV, we present the BPRisk framework, followed by details of the semantic modelling adopted in the framework in Section V. Section VI discusses the implementation status of BPRisk and a real application from within the DAVID project. Section VII concludes this paper and discusses further work.

II. Digital Preservation

AV content is generated in vast quantities from different sources such as film, television and online media, environmental monitoring, corporate training, surveillance and call recording. Some content needs to be retained and archived to enable content re-use, e.g., for cultural heritage, or due to regulatory compliance for security, health and safety. Historically, the preservation of analogue content has been intrinsically linked to its method of production; specifically, the media that is used to carry the signal (the carrier). This means that archives preserved ‘masters’ on magnetic tape, film and even phonograph cylinders [3]. Where masters no longer exist or content was not professionally produced, archives needed to preserve ‘access’ copies on media such as vinyl records, VHS/Betamax tapes, and audio cassettes. To reduce the risk of damage, archives had to consider the physical characteristics of the media and care for the physical environment to which the media was sensitive (e.g., light, heat, humidity and dust) and to look after the machines that read the media. To increase the chances of being able to read the content again, archives often created copies of the artefact, in case one copy was damaged.

Nowadays, AV content is commonly born-digital and archives such as INA (the French national archive) and ORF (the Austrian broadcaster) in the DAVID project undergo digital migration projects to digitise the older, analogue, content [4]. Digital content (digitised or born digital) can be copied, transferred, shared and manipulated far more readily than its analogue equivalent. In a world of digital AV content, preservation is largely agnostic to the carrier that is used to store and deliver the content. Therefore, preservation and archiving is about making sure that the digital data is safe and that processes that manipulate the data do not cause damage. When referring to ‘digital damage’ in this paper, it is worth noting the following definition:

> *“Digital damage is any degradation of the value of the AV content with respect to its intended use by a designated community that arises from the process of ingesting, storing, migrating, transferring or accessing the content.”* [4]

The above definition may seem broad. Indeed, it covers damage arising from failure of the equipment used to store and process digital content, as well as that arising from human error or from ‘failure’ of the process. The challenge for digital preservation is to keep the AV content usable for the long-term, which is threatened by format obsolescence, media degradation, and failures in the very people, processes and systems designed to keep this content safe and accessible [5], [6], [7].

Therefore, the core problem is greater than the potential for a digital file already in the archive to become damaged over time due to, e.g., bit rot [5], which can effectively be addressed by keeping multiple copies of each file [4], [6]. We also need to consider the future challenges for digital preservation as some analyses [8] predict that as ever more 8K AV content is ingested into archives, the growth in data volumes will, with all likelihood, outstrip the growth in storage capacity and increase in data write rate, such that it becomes impossible to store and replicate all content as it is produced. Therefore, strategies such as file-level replication may not be feasible in the future, and managing risk to the entire workflow process becomes essential.

III. Risk Management for Digital Preservation

Risk management, in a broad sense, can be understood as *“the coordinated activities to direct and control an organisation with respect to risk”* [9]. Risk, as defined by ISO 31000 [9], is the *“effect of uncertainty on objectives”*. In this context, *uncertainty* arises from random or systematic failure of preservation systems and processes (that may involve manual human activities). The *effect* of which is to cause damage to AV content. In general terms, we can say that the key *objective* is to ensure long-term preservation of digital AV content, i.e., avoid damage and ensure that it can be accessed in the future.

Current archives such as the French national archive, INA, and the Austrian broadcaster ORF typically deploy a number of IT based strategies for avoiding, preventing or recovering from loss [4]. These archives are engaged in a process of long-term Digital Asset Management (DAM) [10], specifically Media Asset Management (MAM), which focuses on storing, cataloguing and retrieving digital AV content. Several commercial tools exist to support the MAM process, some of which support risk treatment strategies such as keeping multiple copies of each file (redundancy). However, these tools do not include a model of risk. The archive must decide on risk indicators and define the way in which these can be measured in order to monitor them, often using separate tools to do so.

Workflows are often used to describe business processes and, increasingly often, are used to automate some or all of the process. Automated workflow execution is possible if the process is specified in a machine-interpretable fashion, such as using BPMN. In Hazard and Operability Studies (HAZOP), risks are seen as inherent in processes, as individual steps may fail, causing consequences for later parts of the process, or if the process is not executed correctly. Risk-aware business process management is critical for systems requiring high integrity, such as archives.

A recent review of business process modelling and risk management research has been conducted by Suriadi et al. [11], identifying three parts to risk-aware business process management:

- Static / design-time risk management: analyse risks and incorporate risk mitigation strategies into a business process model during design time (prior to execution).
- Run-time risk management: monitor the emergence of risks and apply risk mitigation actions during execution of the business process.
- Off-line risk management: identify risks from logs and other post-execution artefacts, such that the business process design can be improved.

Several approaches have been proposed to model business processes and risk information such that it enables risk analysis. Rosemann and zur Muehlen propose integrating process-related risks into business process management by extending Event-driven Process Chains (EPC) [12]. Risks are classified according to a taxonomy including structural, technological and organisational risks.

Analysis of process risks is difficult given that operational risks are highly dependent on the specific (and changing) business context. Many risks are caused by business decisions (e.g., preservation selection strategy or migration path), so large volumes of data required for statistical methods are often not available for analysis. Those who subscribe to this thesis use structural approaches, such as Bayesian networks, HAZOP and influence diagrams. For example, Sienou et al. [13]

present a conceptual model of risk in an attempt to unify risk management and business process management using a visual modelling language.

In contrast to the above thesis, some believe that run-time analysis of risks is possible with a suitably instrumented execution process. Conforti et al. [14] propose a distributed sensor-based approach to monitor risk indicators at run time. Sensors are introduced into the business process at design time; historical as well as current process execution data is taken into account when defining the conditions that indicate that a risk is likely to occur. These data can be used for run-time risk management or off-line analysis.

Given that analysis of business processes using structured and/or statistical approaches can reveal vulnerabilities, it is important to control the risk that these vulnerabilities lead to loss. Bai et al. [15] use Petri nets (a transition graph used to represent distributed systems) and BPMN to model business processes and to optimise the deployment of controls, such that the economic consequences of errors (measured as Conditional Value at Risk - CVaR) are minimised.

Using BPMN, the PrestoPRIME project described the preservation workflows that were implemented in the preservation planning tool iModel [16]. It has shown that tools are required to model such generic preservation workflows in such a way that they can be related to specific preservation processes and augmented with information concerning risks.

IV. Business Process Risk Management Framework

Here, we present a Business Process Risk management framework (BPRisk) developed in the DAVID project (Section IV-C), designed to support the aims and risk management process discussed below in Sections IV-A and IV-B.

A. Aims of Risk Framework for Digital Preservation

Above, we have discussed the motivations for a risk management of business processes, according to the wider challenges in the domain of digital preservation. For digital preservation / archive management, the key actor we are addressing with the proposed risk framework is the preservation expert / specialist, who is responsible for designing workflows for managing and processing digital AV content. We can summarise here some key value-added aims of a risk management framework in the context of digital preservation:

1) Helping preservation experts develop new workflows, especially the early stages of development. Note that the purpose of the framework is not to replace MAM tools (discussed in Section III, above), nor the preservation experts, but to be a value-added tool to assist them.
2) Helping preservation experts optimise workflows (in terms of cost effectiveness and security), considering also trade-offs where too many corners are cut (to reduce cost), which may lead to increased risk.
3) Helping preservation experts communicate and justify decisions about choices for elements in workflows. This may be related to arguing expected financial Return On Investment (ROI) of putting in place certain risk mitigations, for example. By risk mitigation, we here refer to reducing the likelihood or impact of risk.
4) Helping organisations change their processes, as the risk of change is typically seen as very high, which inhibits change. However, change is necessary to address the issue of format obsolescence.

From an organisational point of view, some of the key reasons to perform risk management can be summarised as follows:

1) Workflows can be large and complex. Therefore, there can be too many variables and options for preservation experts to consider simultaneously to accurately estimate the potential impact of risk.
2) Risk information is typically in experts' heads, which is itself a risk from the organisation's point of view. The risk framework ensures that the knowledge is captured and retained, and is readily available should the organisation be subject to an audit or the expert is unavailable or leaves the organisation.
3) Improve cost-benefit by a) identifying and understanding key vulnerabilities and b) targeting investments to address those vulnerabilities.
4) Move away from "firefighting". That is, organisations may spend more time dealing with issues rather than preventing them in the first place. Risk management is key to prevention, i.e., spending more time in the planning stages to save time and cost on dealing with issues in the future that could have been avoided.

It is important to note that the end users of the risk management framework in this context are unlikely to be risk experts. They are domain (preservation) experts, and they will be acutely aware of a wide range of potential issues concerning the preservation workflows they manage. However, the term risk and explicitly managing risk may be entirely unfamiliar and it is important that the risk management framework is suitably designed to aid the domain experts (rather than simply being a risk registry).

B. Risk Management Process

The risk framework should support a process that promotes best practices to address the aims discussed above in order to reduce the risks to long-term preservation. There is a natural focus on the planning aspects regarding risk management, but we do need to consider the wider context as well.

Several risk standards and methodologies exist, but it is not within the scope here to discuss them in detail. However, we will make reference to one in particular here, ISO 31000 [9], to show how it relates to a risk management approach proposed here based on the Deming cycle. The Deming cycle is a four-step iterative method commonly used for control and continuous improvement of processes and products. The four steps are: Plan, Do, Check and Act. For this reason it is also commonly referred to as the PDCA cycle, and is key to, for example, ITIL Continual Service Improvement [17]. In general terms, risk management is a part of continual improvement of processes – preservation workflows in this context.

The ISO 31000 [9] risk management methodology is depicted in Figure 1, below, which depicts the various stages from 'establishing the context' to 'treatment' (of risk) that is also cyclic. Supporting continual improvement of workflow processes is imperative in digital preservation, as discussed in Section II, as one of the key challenges in this domain

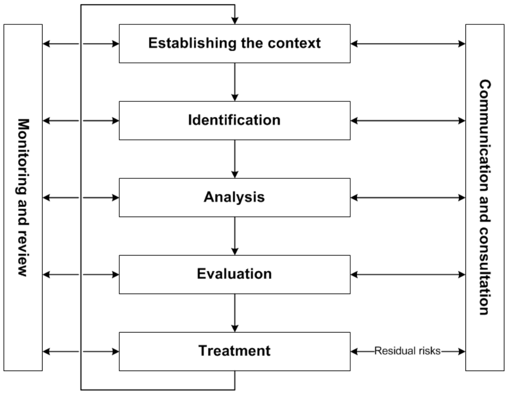


Figure 1. ISO 31000 risk management process.

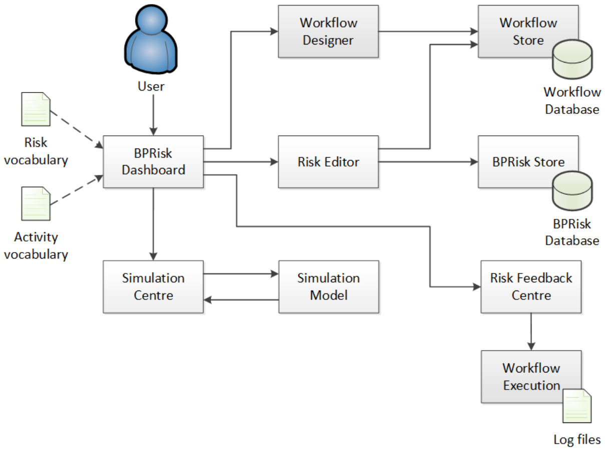


Figure 2. BPRisk framework high level component view.

is obsolescence and one of the key current risk strategies involving file-replication may not be feasible in the future.

Given the aims discussed above, each of the four stages of the Deming cycle is covered below from the perspective of what a user (preservation expert) would do, with reference to the related stages of the ISO 31000 methodology).

Plan (‘establishing the context’ and ‘identification’ stages of ISO 31000): build workflows, capture risk information, simulate workflow execution scenarios to identify key vulnerabilities and estimate impact of risk, and make decisions.

Do (‘analysis’ stage of ISO 31000): execute business process, orchestrate services, and record execution meta-data.

Check (‘evaluation’ stage of ISO 31000): analyse workflow execution meta-data and process analytics, calibrate simulations and trigger live alerts.

Act (‘treatment’ stage of ISO 31000 as well as feedback and loop-back to the previous stages): adapt workflows and manage risk. Re-run simulations (Plan), enacting the offline changes in the real business process and continues execution (Do) and monitoring (Check).

Note also how this relates to the three risk-aware business processes discussed above from Suriadi et al. [11]; static/design-time risk management (Plan), run-time risk management (Do) and off-line risk management (Check). The final step in the Deming cycle, Act, covers multiple processes.

C. *Risk Components*

Based on the above aims, a high level component view of the BPRisk framework developed in the DAVID project is depicted in Figure 2. This framework is implemented as a RESTful web application, integrating both new components developed in the DAVID project as well as existing open source technologies, which is discussed below.

BPRisk Dashboard: The main entry point for the user from which the user can access the functionalities of the framework, e.g., to create workflows, specify risks, run and view risk simulation results, etc. Figure 2 also shows two vocabularies used, one for known domain-specific risk and one for domain specific activities. This is discussed further below.

Workflow Designer: There are several existing, mature, tools for this, supporting the well-known BPMN 2.0 standard, such as Signavio Decision Manager [18] and the jBPM Designer [19]. The latter has been adopted in the BPRisk framework as it is available as open source.

Workflow Store: This is a component to persist any workflows created, updated or imported. Existing tools, such as jBPM come with multiple persistence options and a RESTful API for accessing and managing the workflows.

Risk Editor: As described above, this component is responsible for allowing users to specify risks. As discussed earlier in this paper, the end-users of this system are not likely to be risk experts. Therefore, the Risk Editor utilises the two vocabularies mentioned above in a semantic risk model, which is used to aid users in specifying risks. See Section V for further discussion.

BPRisk Store: This is a component for persisting risk specifications and risk simulation results (a connection from the Simulation Centre has not been depicted in Figure 2 for the sake of simplifying the diagram).

Simulation Centre: This is a component for managing the running of simulation models for workflows annotated with risk information. This component deals with configuring different simulation scenarios and allows users to visualise and compare the results.

Simulation Model: A stochastic risk simulation model that the Simulation Centre can execute. This component simulates executions of the workflow process and the occurrences of risks defined for the workflow activities. As output, the simulation model gives information on, for example, risk occurrences, time and cost spent on risk, and impact of risk.

Risk Feedback Centre: A component for getting data from real workflow executions that can be used to a) analyse the workflow execution meta-data and b) to modify/adapt/calibrate the workflows (e.g., risk details) and simulation configurations to improve the accuracy for future simulation scenarios.

Workflow Execution: An external software component to the BPRisk framework, which would be invoked to execute a workflow process. This is a source of workflow execution data for the Risk Feedback Centre.

V. Semantic Risk Modelling

The BPRisk framework utilises a semantic risk model for specifying and reasoning about risks associated with workflow activities. The modelling approach is generic in nature, utilising a multi-level ontology to include domain specific workflow activities and risks.

A. Modelling Approach

The BPRisk ontology represents information related to risks, controls and activities. This representation allows flexibility and extensibility of the risk model. It can be easily published (e.g., as a set of OWL files), can be extended in unexpected ways, and it can be combined with other ontologies such as W3C PROV [20].

The approach to building the ontology is based on work done in SERSCIS project [21]. The authors use a layered, class-based ontology model to represent knowledge about security threats, assets and controls. Each layer inherits from the layer above. The CORE layer describes the relationships between a central triad (threat, asset, control). A domain security expert creates sub-classes for each of these core concepts to create a GENERIC layer. A system expert further subclasses the generic concepts to specialise them for the system of interest, creating the SYSTEM layer. Note that this ontology was used in the context of modelling systems and interactions between system components, where it is assumed that a system of a particular type is always subject to the threats identified by the security and system experts. This expert knowledge, therefore, helps the system designer who may not have this expert knowledge themselves when they are designing new systems.

The same, layered, ontological approach has been taken here, but the core ontology is slightly different. While, in SERSCIS, the triad in the CORE layer includes Asset, there is only one asset of value in this context – the digital AV object, which can be affected by different Activities in a workflow process (e.g., ingest, storage and transcoding). The term Threat used in SERSCIS can be understood as Risk in this context. Therefore, the CORE layer in BPRisk comprises a triad of Risk, Activity and Control.

B. Model Definition

The model focuses on the Activities in the preservation lifecycle and the Risks that are inherent in their execution. Controls can be put in place to block or mitigate these Risks. The CORE layer comprises risk, activity and control, as well as basic relationships such as ‘Risk threatens Activity’ and ‘Control protects Activity’. However, the relationship between Control and Risk is established via SPIN rules (see the following section), to determine the appropriate relationship. That is, a Risk is only considered Mitigated if an appropriate Control is in place. This is illustrated below in Figure 3.

The DOMAIN layer has been developed in the DAVID project for digital preservation, which describes common preservation activities, risks and controls. These are modelled as sub-classes, which can be quite hierarchical. As an example, the DOMAIN level classes in Figure 3 include two sub-classed Activities, ‘Migration’ and ‘Digital Migration’, with an associated risk ‘Migration Fails’.

The SYSTEM layer is a further extensible part that would be populated by the users of the BPRisk framework when they build a workflow of specific Activities and associate Risk to them. ‘FFmpeg Migration’ is given as an example in Figure 3 (dark grey), which is a subclass of ‘Digital Migration’. This digital migration risk is specific to using the tool FFmpeg [22], which is a common AV media transcoding tool. This sub-classing is important, as we can reason about risks throughout the hierarchy, which we go further into below.

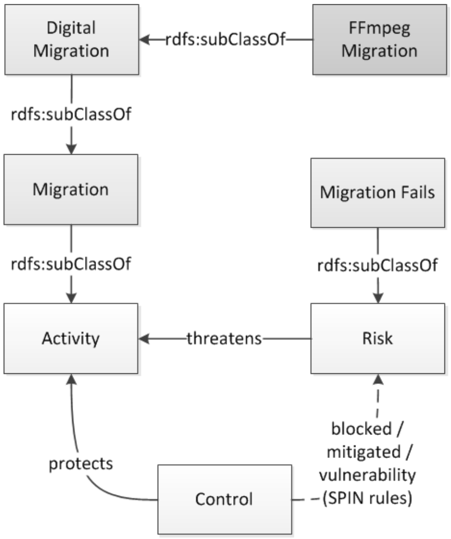


Figure 3. BPRisk ontology with sub-classing examples.

C. Semantic Reasoning

The relationships between risks, controls and activities are encoded as risk classification rules using SPIN [23]. Running inferencing over the model automatically does the classification. For example, the following SPIN rule classifies an instance of the risk *FieldOrderIssues*, which threatens the activity *Transcoding*, as blocked, if the control *ChangeTranscodingTool* is present:

```
CONSTRUCT {
    ?r a dom:BlockedRisk .
} WHERE {
    ?a a act:Transcoding .
    ?r a dom:FieldOrderIssues .
    ?c a dom:ChangeTranscodingTool .
    ?r core:threatens ?a .
    ?c core:protects ?a .
}
```

As noted above, the SYSTEM layer is developed so that it subclasses the DOMAIN layer for a specific organisation using the BPRisk framework, as seen above in Figure 3. This should specify the kind of activity in the preservation workflow of interest, e.g., subclass *Scanning* as *35mmToJPEG2kScanning*. Workflow-specific risks can then be automatically generated using SPIN. For example, the following is a generic SPIN rule to generate all risks:

```
CONSTRUCT {
    ?uri a owl:Class .
    ?uri rdfs:subClassOf ?gr .
    ?uri rdfs:subClassOf _:b0 .
    _:b0 a owl:Restriction .
    _:b0 owl:onProperty core:threatens .
    _:b0 owl:someValuesFrom ?sa .
} WHERE {
    ?sa (rdfs:subClassOf)+ act:Activity .
    ?sa rdfs:subClassOf ?ga .
    ?gr rdfs:subClassOf core:Risk .
    ?gr rdfs:subClassOf ?restriction1 .
    ?restriction1 owl:onProperty core:threatens .
```

```
   ?restriction1 owl:someValuesFrom ?ga .
   FILTER NOT EXISTS {
      ?uri rdfs:subClassOf _:0 .
   } .
   FILTER STRSTARTS(str(?sa),
      "http://david-preservation.eu/bprisk#") .
   BIND (fn:concat(STRAFTER(str(?gr), "#"),
      "_", STRAFTER(str(?sa), "#")) AS ?newclass) .
   BIND (URI(fn:concat(fn:concat(STRBEFORE(str(?sa),
      "#"), "#"), ?newclass)) AS ?uri) .
}
```

This rule finds all activities in the SYSTEM layer and creates a workflow-specific risk for each of the DOMAIN layer risks that threaten the activities' parent class. The name of the workflow-specific risk in this example is generated by concatenation of the DOMAIN layer risk name and the workflow-specific activity name.

Encapsulation of media preservation knowledge (linking activities, risks and controls) using SPIN rules provides a flexible and extensible representation of knowledge based reasoning in our architecture. Specifically, we extend the SPIN templates rule hierarchy into which we insert groups that can contain rules to be called upon, for example, in the construction of new risk instances in the presence of particular activities. Using this approach, it is possible to progressively refine the core preservation knowledge base (or augment it with additional, domain specific rules) without necessarily updating system code with new SPARQL [24] queries.

VI. BPRisk Implementation and Application

At the time of writing, the BPRisk framework prototype has been developed within the DAVID project [1]. It has been implemented as RESTful web service using Java Spring [25]. As noted in Section IV-C, above, the jBPM Designer [19] has been integrated for workflow design. A risk simulation model has been implemented in Matlab Simulink [26]. The Risk Feedback Centre is under development.

Within the DAVID project, the BPRisk framework has been developed with use cases from INA and ORF, such as planning for migration of old, analogue, content into new, digital, formats (digital migration). Here, we include an example of the use of BPRisk in the planning of an MXF Repair workflow at ORF, which has been used within the DAVID project for validation purposes. MXF is an abbreviation for a file format, Material eXchange Format. The standard for its use is ambiguous in places and some tool implementations are inconsistent. The result is format compatibility issues, i.e., the files are not standard compliant and may not be possible to play in the future. After the workflow design (planning) was completed, the workflow has been executed and the results of the planning could be compared with the monitoring data collected during its execution.

The MXF Repair workflow is depicted below in Figure 4. Due to space restrictions, a description of the workflow activities is not included here, nor are some application/modelling details. The aim here is to clarify aspects of the workflow risk specification and the role of workflow simulation. However, interested readers are referred to specific parts of [4], below, for further details.

Firstly, in the DAVID project, the DOMAIN layer of the BPRisk ontology has been created based on controlled vocabularies for preservation activities, tools and risks. Interested readers can refer to Annex C in [4] for details. The first activity in the workflow, TSM (Tivoli Storage Manager, a data backup system from IBM) Retrieve, maps to 'Acquisition/Recording' in the preservation vocabulary, for example. And two risks have been identified for this activity: a) wrong file selection and b) retrieve fails. The semantic reasoning rules discussed above, in Section V, enables the BPRisk framework to prompt users with such risks at design time.

After specifying risks for the different activities, workflow simulation scenarios were set up with ORF for this workflow. To simulate workflow execution, additional parameterisation is required, such as estimates for how often the risks are likely to occur, and the expected time and costs for dealing with any issues that may occur. Values were set based on the experiences the workflow and technical experts at ORF have of the tools and activities used in the workflow, as well as observations from monitoring data where available. In the future, these estimates are intended to be updated and improved via the Risk Feedback Centre, as discussed above in Section IV-C.

The simulation results on this workflow showed very clearly that the activity most affected by risk is the Upload activity (upload fails). Not just in terms of frequency of occurrence, but it affects the most media files and accrues the most significant financial cost. At this stage of the planning phase for designing new workflows, it is such observations that are important in terms of highlighting the key vulnerabilities and start to quantify their impact and potential cost savings by addressing the problems differently. As discussed in Section IV-B, different versions of a workflow may be designed and simulated as part of the planning before making a final, informed, decision and moving to executing the workflow in the real environment. Interested readers are referred to Section 8.4 and 9 in [4] for details on the simulation modelling, results and validation of these results based on the observations made after executing the workflow.

VII. Conclusions and Further Work

We have presented a Business Process Risk management framework (BPRisk) that allows users to manage workflow processes with regards to risk. The framework is generic in nature, but has been discussed here in the context of digital preservation, where the objective is to avoid damage to the digital content and ensuring that the content can be accessed in the future. Long-term digital preservation is threatened by format obsolescence, media degradation, and failures in the very people, processes and systems designed to keep the content safe and accessible.

The BPRisk framework combines workflow specification (and adaption) and risk management. It has been designed in accordance to a risk management process presented in this paper, based on the Deming (PDCA) cycle and we have shown how it relates to the stages of the ISO 31000 risk methodology. Key to the process is continual improvement, as risk management is not only a static exercise performed at design time [11], but is also imperative during process change.

A layered semantic risk model has been presented, which a) enables reasoning about threats in a workflow and b) assists end-users (who are typically not risk experts) by automatically suggesting risks and respective controls for workflow activities. The framework helps end-users develop and optimise workflows, and improve cost-benefit by identifying (and address-

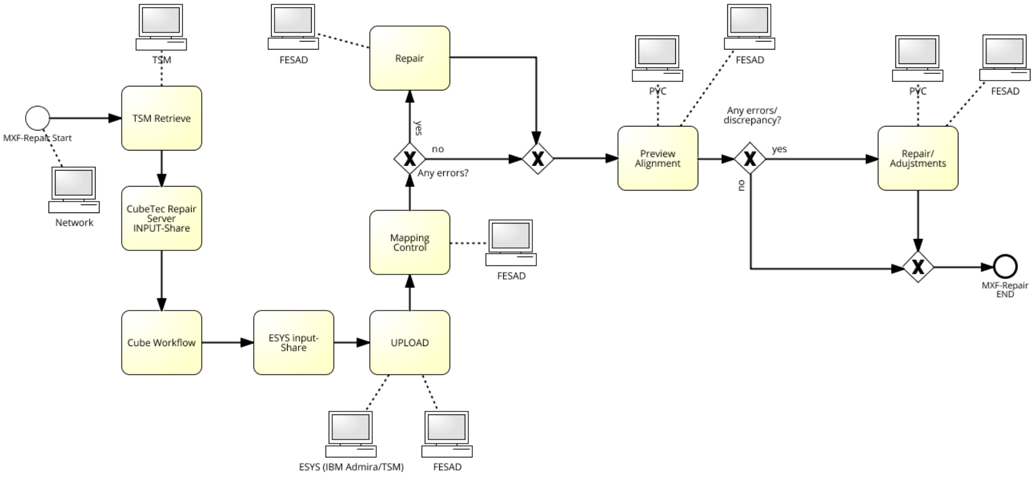


Figure 4. MXF Repair workflow.

ing) key vulnerabilities by simulation workflow executions to estimate the impact of risk.

A prototype is developed in the DAVID project at the time of writing. Further research involves mechanisms for automatically updating risk models and respective simulation configurations according to observed workflow execution data.

ACKNOWLEDGEMENTS

This work has been carried out in the DAVID project, supported by the EC 7th Framework Programme (FP7-600827).

REFERENCES

[1] EC FP7 DAVID Project, “Digital AV Media Damage Prevention and Repair,” http://david-preservation.eu/, [retrieved: May 2015].

[2] Object Management Group, “Business Process Model and Notation (BPMN) Version 2.0,” http://www.omg.org/spec/BPMN/2.0/PDF/, [retrieved: May 2015].

[3] Department of Special Collections, Donald C. Davidson Library, University of California, “Cylinder Preservation and Digitization Project,” http://cylinders.library.ucsb.edu/, [retrieved: May 2015].

[4] V. Engen, G. Veres, M. Hall-May, J.-H. Chenot, C. Bauer, W. Bailer, M. Höffernig, and J. Houpert, “Final IT Strategies & Risk Framework,” EC FP7 DAVID Project, Tech. Rep. D3.3, 2014, available online http://david-preservation.eu/wp-content/uploads/2013/01/DAVID-D3.3-Final-IT-Strategies-Risk-Framework.pdf [retrieved: May 2015].

[5] M. Addis, R. Wright, and R. Weerakkody, “Digital Preservation Strategies: The Cost of Risk of Loss,” SMPTE Motion Imaging Journal, vol. 120, no. 1, 2011, pp. 16–23.

[6] J.-H. Chenot and C. Bauer, “Data damage and its consequences on usability,” EC FP7 DAVID Project, Tech. Rep. D2.1, 2013, available online http://david-preservation.eu/wp-content/uploads/2013/10/DAVID-D2-1-INA-WP2-DamageAssessment_v1-20.pdf.

[7] D. Rosenthal, “Format Obsolescence: Assessing the Threat and the Defenses,” Library Hi Tech, vol. 28, no. 2, 2010, pp. 195–210.

[8] M. Addis, “8K Traffic Jam Ahead,” PrestoCentre Blog, April 2013, available online: https://www.prestocentre.org/blog/8k-traffic-jam-ahead [retrieved: May 2015].

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

[10] D. Green, K. Albrecht, and et al, “The NINCH Guide to Good Practice in the Digital Representation and Management of Cultural Heritage Materials,” National Initiative for a Networked Cultural Heritage, Tech. Rep., 2003.

[11] S. Suriadi, B. Weiß, A. Winkelmann, A. Hofstede, M. Adams, R. Conforti, C. Fidge, M. La Rosa, C. Ouyang, A. Pika, M. Rosemann, and M. Wynn, “Current Research in Risk-Aware Business Process Management - Overview, Comparison, and Gap Analysis,” BPM Center, Tech. Rep. BPM-12-13, 2012.

[12] M. Rosemann and M. zur Muehlen, “Integrating Risks in Business Process Models,” in ACIS Proceedings, 2005.

[13] A. Sienou, E. Lamine, A. Karduck, and H. Pingaud, “Conceptual Model of Risk: Towards a Risk Modelling Language,” in Web Information Systems Engineering, ser. LNCS 4832, 2007, pp. 118–129.

[14] R. Conforti, G. Fortino, and A. t. M. La Rosa, “History-Aware, Real-Time Risk Detection in Business Processes,” in On the Move to Meaningful Internet Systems, ser. LNCS. Springer, 2011, vol. 7044, pp. 100–118.

[15] X. Bai, R. Krishnan, R. Padman, and H. Wang, “On Risk Management with Information Flows in Business Processes,” Information Systems Research, vol. 24, no. 3, 2013, pp. 731–749.

[16] M. Addis, M. Jacyno, M. H. Hall-May, and S. Phillips, “Tools for Quantitative Comparison of Preservation Strategies,” EC FP7 PrestoPRIME Project, Tech. Rep. D2.1.4, 2012, available online: http://eprints.soton.ac.uk/349290/.

[17] V. Lloyd, ITIL Continual Service Improvement – 2011 Edition. The Stationary Office, 2011, iSBN: 9780113313082.

[18] Signavio GmbH, “Signavio Decision Manager,” http://www.signavio.com/products/decision-manager/, [retrieved: May 2015].

[19] JBoss, “jBPM,” http://www.jboss.org/jbpm, [retrieved: May 2015].

[20] L. Moreau and P. Missier, “PROV-DM: The PROV Data Model,” W3C Recommendation, 2013, available online: http://www.w3.org/TR/2013/REC-prov-dm-20130430/ [retrieved: May 2015].

[21] M. Surridge, A. Chakravarthy, M. Hall-May, C. Xiaoyu, B. Nasser, and R. Nossal, “SERSCIS: Semantic Modelling of Dynamic, Multi-Stakeholder Systems,” in 2nd SESAR Innovations Days, 2012.

[22] F. Bellard, “FFmpeg,” https://www.ffmpeg.org/, [retrieved: May 2015].

[23] W3C, “SPARQL Inferencing Notation (SPIN),” W3C Submission, 2011, available online: http://www.w3.org/Submission/2011/02/ [retrieved: May 2015].

[24] ——, “SPARQL 1.1,” W3C Recommendation, 2013, available online: http://www.w3.org/TR/sparql11-overview/ [retrieved: May 2015].

[25] Pivotal Software, “Spring,” https://spring.io/, [retrieved: May 2015].

[26] Mathworks, “Simulink,” http://uk.mathworks.com/products/simulink/, [retrieved: May 2015].

ICT - based Approaches for Entrepreneurship Education

Ileana Hamburg, Sascha Bucksch
Institut Arbeit und Technik, Westfälische Hochschule
Gelsenkirchen, Germany
hamburg@iat.eu

Emma O Brien
University of Limerick
Limerick, Ireland
emma.obrien@ul.ie

Abstract **– Businesses are organised or operated by an entrepreneur. It implies creativity, innovation, risk taking and the competence to plan and manage projects in order to achieve objectives. The role of entrepreneurship education is to offer students the tools to be creative, to solve problems efficiently, to analyse a business idea objectively, and to communicate, cooperate, lead, develop and evaluate projects. Approaches like mentoring and Information and Communication Technologies (ICT) supported forms of learning like Problem Based Learning (PBL), could be used in entrepreneurship education. Mentoring supports professional development and increases the mentees opportunities. PBL is suitable for entrepreneurship education i.e., by presenting properly real problems like "starting a business" and creates motivation in the students. ICT could improve the efficiency of PBL, but this aspect was not taken into consideration until now. Mentoring, ICT and PBL are used in the on-going European project Erasmus+ "Supporting PBL in entrepreneurial education and in Small and Medium sized Enterprises (SMEs) through ICT facilitated mentoring – Archimedes". The authors developed an ICT platform in frame of this project to support PBL, which has been tested with SMEs and is shortly described in this paper.**

Keywords: ICT; Entrepreneur; Entrepreneurship education; Mentor; Problem Based Learning; Platforms; TikiWiki.

I. Entrepreneurship Education

Entrepreneur usually means an individual who organises or operates businesses. Entrepreneurship is the art of being entrepreneur, so to be able to turn ideas into action. This implies creativity, innovation and risk taking, and the competence to plan and manage projects in order to achieve objectives [1] [2].

Some of the qualities of entrepreneurs (http://under30ceo.com/10-qualities-of-a-successful-entrepreneur/) should be:

- Discipline to follow the business steps of the established strategy to achieve the proposed objectives and eliminate obstacles.
- Confidence in own ability
- Open minded for new ideas
- Competitive.
- Creative and problem solving identifying solutions
- Determination, not believing that something cannot be done.
- Communication skills to motivate people to work and to sell products.
- Passion, loving work to be done.

Entrepreneurship education programmes should offer students the tools to be creative, to solve problems efficiently, to analyse a business idea objectively, and to communicate, cooperate, lead, develop and evaluate projects. Students can learn to set up their own businesses if they can test their ideas in an educational, supportive environment. Many European countries included entrepreneurship in the national curricula for vocational education training (VET) programmes and they are very different. Reports show that there are some gaps in most of these programmes [3] i.e., teaching methods are ineffective, student participation is limited, teachers are not fully competent, business people are not involved, the practical element is missing, entrepreneurship is not linked to specific training subjects or professions, education is not linked with labour market demands. It is important that entrepreneurship education takes these gaps into consideration.

Mentoring within entrepreneurship education can address some of these gaps as it brings in expertise from business; it is practical and can assist in linking the training to particular professions and labour market demands [4].

Mentoring is a human resource development approach and a vital aspect of knowledge management which needs to be looked by all organizations and education institutions wishing to improve their efficiency [5]. Educators and practitioners have noted the importance of mentorship in promoting leader development and career opportunities [6].

According to Kram's mentor role theory [5], mentors provide career development in order to integrate and prosper within the organization, and social advancement, contributing to the mentee personal growth and their professional development. The literature has found that receiving mentorship has been associated with positive career outcomes [6].

The functions of the mentoring, career advancement for beginners, professional development and social integration (particularly of mentees with special needs) increase the mentees opportunities. Many of these methods can be used for mentoring in entrepreneurial education. For example, experienced entrepreneur-mentors could help their mentees to understand that a failed business is an important part of their entrepreneurial training and that they can continue a successful career.

Copyright (c) IARIA, 2015. ISBN: 978-1-61208-412-1

Mentorship from an entrepreneur can provide students with a greater level of security and inspiration. It can help students to know how a business was developed directly from its founder, and can be more effective than being mentored by an employee or an investor in this case. Also the story of an unsuccessful business venture is useful for students, particularly if it was a courageous idea, or the entrepreneur would like to create other interesting ventures.

For mentors supporting young /future entrepreneurs (19-25 years), it is important to focus on developing life plans and passion for a career, helping these young mentees to keep their vision in sight and to reflect what is happening [7]. Softer skills such as listening, communicating as well as some including the review of business plans and meeting objectives are necessary. Mentors should increase mentees motivation, encouraging them to try to implement their ideas.

Particularly supporting students/starters in small and new business creates a contribution to the local community, more jobs and a more attractive place to do business. Mentors could gain a better understanding of challenges facing small business which could enhance their working life or their retirement period.

Another aspect is that many education institutions and companies offer diversity initiatives to support collaboration, understanding and the use of different competences and cultures, but most diversity initiatives, which are important in a global environment, do not go far enough to promote real diversity and improve firm´s competitiveness. Particularly within vocational education such initiatives are missing.

Entrepreneurship learning does not relate to a single occupation; it covers a variety of occupational skills and learners. Students engaged in entrepreneurship education should acquire different competences according to the focus of their learning [8].

The implementation of efficient entrepreneurship teaching and learning methods, particularly in schools and VET, requires structural changes in most countries. In many institutions of higher education and VET, where learning approaches are not driven by national policy, introduction of entrepreneurial teaching and learning depends on the institution which should also make a cultural change including diversity approaches. Knowledge about diversity as well practical training should be offered in entrepreneurial education and these will be more efficient than large, abstract diversity lectures. The main objectives of such training include awareness, education and positive recognition of the differences among people in the workforce.

Information and Communication Technologies (ICT) affects the entrepreneurship education because new technologies support the development of new entrepreneurship forms. ICT have the potential to improve student competences and skills, to motivate and engage students, to help them to link school knowledge to work practices. ICT contribute to change and improve VET practices. Technology becomes quickly obsolete requiring new skills and knowledge and also changes in entrepreneurship education.

In the following we describe shortly Problem Based Learning (PBL) as a suitable form for Entrepreneurial education and an approach for ICT support developed by the authors.

II. Problem Based Learning

Problem Based Learning (PBL) has been proven to develop higher order thinking and critical thinking skills. There are many different approaches to PBL [9], however little research has been done into the most effective methods in terms of learner success [10]. PBL should be adopted outside academic contexts i.e., as an excellent method of training for SMEs, because the staff learns solving real problems. It allows the learner to develop skills relevant to the needs of the company, it is conducted in a work based environment, it provides them with the skills to sustain the company beyond the initial training, it is low cost and it directly solves problems for the SME providing an immediate return [11]. Donnelly [12] highlighted that little is known about the use of technology in PBL. However after conducting a study in an academic context of the use of Communities of Practice (CoPs) [13] [14] for PBL it was found that CoPs provide an opportunity to enhance collaboration and extend face to face time with mentors and peers. In a business environment PBL, mentoring, CoPs and social media can be used to provide an opportunity for the communication between the mentor and mentee and to work with peers (or experts inside and outside the company) to find potential solutions to the problem or approaches to solve the problem.

PBL is suitable for entrepreneurship education i.e., by presenting properly real problems like “starting a business”. It creates motivation in the students.

It is important to have a structured way in PBL, because at the beginning the students feel like they know nothing but after a short introduction and the guidance from the trainer/teacher they realise that they themselves can be the drivers in creating their own business.

In the following list we present steps which could be used by teaching PBL, based on methods described in PBL step by step [16]:

- Clarifying the task – The purpose of the first step is to explain the task, to agree on the meaning of the various words and terms and on the situation described in the problem
- Defining the problem
- Brainstorming - This should result in ideas to structure the problem. Each individual may express his or her ideas free and without immediate discussion
- Rating of Brainstorming outcomes
- Formulating learning objectives to cover knowledge deficits

- Self-study
- Rating of possible solutions and working out a final solution
- Reflection.

III. EXAMPLE

The European Erasmus+ project "Supporting PBL in entrepreneurial education and in small and medium sized enterprises (SMEs) through ICT facilitated mentoring – Archimedes" will develop a framework for organisational problem-based learning and supports the use of this form of learning. It is expected that these approaches will be widely adopted in entrepreneurial education and SMEs.

PBL will be supported by an ICT platform taking into consideration the PBL steps described above. The platform should help the tutor and the participants during the PBL seminars. Figure 1 describes a Flow Chart for platform.

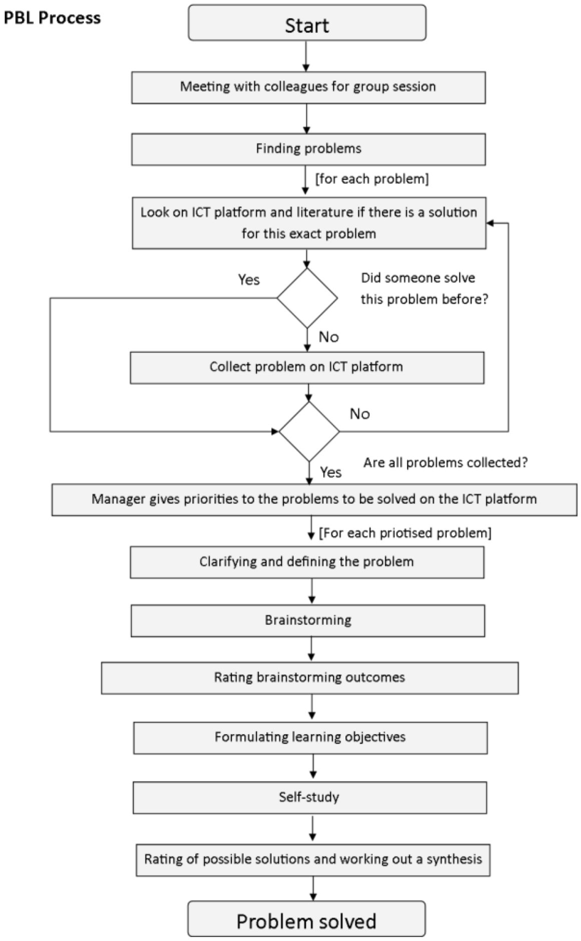
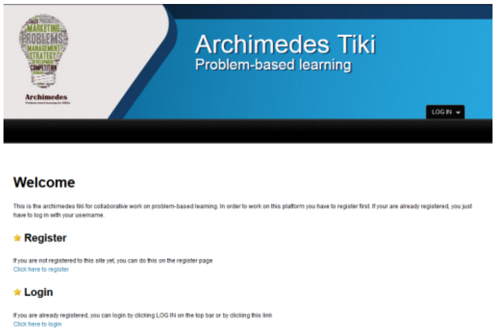

PBL Process
Start
Meeting with colleagues for group session
Finding problems
[for each problem]
Look on ICT platform and literature if there is a solution for this exact problem
Yes
Did someone solve this problem before?
No
Collect problem on ICT platform
No
Are all problems collected?
Yes
Manager gives priorities to the problems to be solved on the ICT platform
[For each priotised problem]
Clarifying and defining the problem
Brainstorming
Rating brainstorming outcomes
Formulating learning objectives
Self-study
Rating of possible solutions and working out a synthesis
Problem solved

Figure 1. Flow chart Source: IAT

The platform is in development by using TikiWiki [17]. TikiWiki, also known as Tiki, is an open source Content Management System (CMS). It provides many rich features like websites, forum, chat, wiki, blogs, quiz, calendar, document management, social software and many more. It is highly configurable and is mainly used in companies to organise tasks and to work collaboratively.

Tiki was used in some of our former project and has proven to be a good ICT solution for collaborative working and will be used to support PBL now. The following figures show screenshots of the Archimedes ICT platform supporting PBL.



Welcome

This is the archimedes tiki for collaborative work on problem-based learning. In order to work on this platform you have to register first. If your are already registered, you just have to log in with your username.

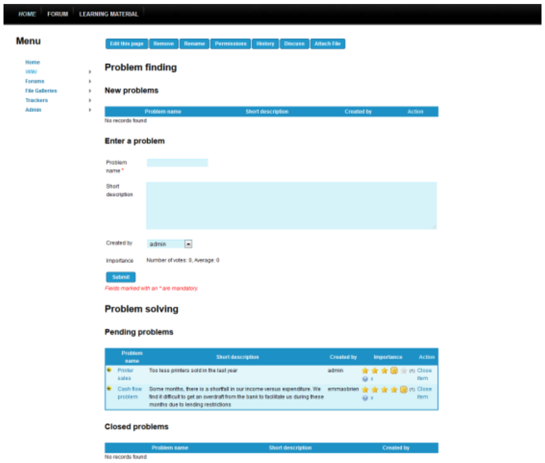
★ **Register**

If you are not registered to this site yet, you can do this on the register page
Click here to register

★ **Login**

If you are already registered, you can login by clicking LOG IN on the top bar or by clicking this link
Click here to login

Figure 2. Homepage of the Archimedes ICT platform [18]



Problem finding

New problems

Problem name	Short description	Created by	Action

No records found

Enter a problem

Problem name *
Short description
Created by: admin
Importance: Number of votes: 0, Average: 0

Fields marked with an * are mandatory.

Problem solving

Pending problems

Problem name	Short description	Created by	Importance	Action
Printer sales	Too less printers sold in the last year	admin	★★★★☆ (1) x	Close item
Cash flow problem	Some months, there is a shortfall in our income versus expenditure. We find it difficult to get an overdraft from the bank to facilitate us during these months due to lending restrictions	emmaobrien	★★★★★ (1) x	Close item

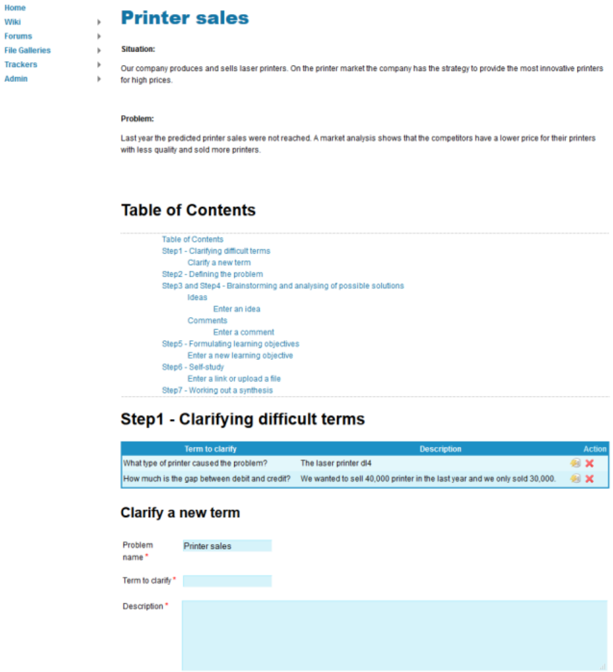
Closed problems

Problem name	Short description	Created by

No records found

Figure 3. Problem overview on the ICT platform [18]

ISBN: 978-1-61208-412-1



Printer sales

Situation:

Our company produces and sells laser printers. On the printer market the company has the strategy to provide the most innovative printers for high prices.

Problem:

Last year the predicted printer sales were not reached. A market analysis shows that the competitors have a lower price for their printers with less quality and sold more printers.

Table of Contents

Table of Contents
Step1 - Clarifying difficult terms
Clarify a new term
Step2 - Defining the problem
Step3 and Step4 - Brainstorming and analysing of possible solutions
Ideas
Enter an idea
Comments
Enter a comment
Step5 - Formulating learning objectives
Enter a new learning objective
Step6 - Self-study
Enter a link or upload a file
Step7 - Working out a synthesis

Step1 - Clarifying difficult terms

Term to clarify	Description	Action
What type of printer caused the problem?	The laser printer dl4	✕
How much is the gap between debit and credit?	We wanted to sell 40,000 printer in the last year and we only sold 30,000.	✕

Clarify a new term

Problem name * Printer sales

Term to clarify *

Description *

Figure 4. PBL steps visualised on the ICT platform [18]

The first version of the platform has been tested with students and staff from SMEs. The results are positive. Both students and SME staff consider PBL as a suitable form for learning and solving real problems. At two academic cooperation partners PBL has been introduced in the courses for entrepreneurs. Some SMEs would like to have its own platform (a copy of the Archimedes ICT platform for solving and saving own problems). The improvements, proposed by the users, are taken into consideration for further project developments.

CONCLUSIONS

Development of entrepreneurial attitudes is a complex process, an important goal of education and requires cooperation of all actors involved. Within the project Archimedes Focus Group Discussions have been organised with education experts, students, entrepreneurs to discuss about suitable methods in education in order to achieve these goals. Intensive cooperation between companies, higher and VET institutions its one of the future activities of the authors.

Implementation of PBL requires some changes in the curriculum of entrepreneurship education and trainers/teachers with special knowledge. Rooms should be available for group discussions and the libraries should contain references which allow students to research for their PBL cases. Until now it was not successfully realised. Projects should be developed in this context.

ACKNOWLEDGMENTS

This paper describes work within the on-going European project Archimedes

REFERENCES

[1] D.A. Shepherd, Educating entrepreneurship students about emotion and learning from failure. Academy of Management Learning & Education, 3(3), pp. 274-287. 2004.

[2] I. Hamburg, Improving young entrepreneurship education and knowledge management in SMEs by mentors. In: World journal of education 4, no. 5, pp. 51-57, 2014.

[3] European Commission. Enterprise and Industry, Entrepreneurship in Vocational Education and Training. Final report of the Expert Group. ec.europa.eu/.../sme/.../vocational/entr_voca_en.pdf, 2009.

[4] E. O'Brien and I. Hamburg, Supporting sustainable strategies for SMEs through training, cooperation and mentoring. Higher education studies 2014, 4(2), pp. 61-69, 2014.

[5] K. Kram, Mentoring at work. Developmental relationships in organizational life. Scott, Foresman & Company, Glenview, ISBN 0-673-15617-6. 1985.

[6] S.B. Srivastava, “Network Intervention: A Field Experiment to Assess the Effects of Formal Mentoring on Workplace Networks”. University of California, Berkeley, Working Paper. 2013.

[7] J. Cull, Mentoring Young Entrepreneurs: What Leads to Success? International Journal of Evidence Based Coaching and Mentoring, 4(2), pp. 8-18, 2006.

[8] Aarchus Technical College, Standards for Qualifications in. Entrepreneurship Learning. An EU-funded project managed by the European Agency for Reconstruction http://www.masht-gov.net/advCms/documents/Standards_for_Qualifications_in_Entrepreneurship_Learning.pdf, 2013.

[9] H.S. Barrows, A taxonomy of problem-based learning methods. Medical Education 20, pp. 481-486, 1986.

[10] W. Huag, Theory to reality: a few issues in implementing problem-based learning, Education Tech Research Dev (2011), vol. 59, 2011.

[11] S. Bell, Project-Based Learning for the 21st Century: Skills for the Future, The Clearing House: A Journal of Educational Strategies, Issues and Ideas, 83:2, pp. 39-43, 2010.

[12] R. Donnelly, Blended problem-based learning for teacher education: Lessons learnt learning. Media and Technology, 31(2), pp. 93-116, 2006.

[13] E. Wenger, “Communities of Practice: Learning, Meaning and Identity”. Cambridge MA: Cambridge University Press, 1998.

[14] I. Hamburg and E. O'Brien, Using strategic learning for achieving growth in SMEs. Journal of information technology and application in education 3(2), pp. 77-83, 2014.

[15] I. Hamburg and E. O'Brien, Engaging SMEs in cooperation and new forms of learning. In: Computer and information science 7, no. 1, p. 9, 2014.

[16] PBL step by step | UM PBL PREP www.umpblprep.nl/pbl-step-by-step

[17] TiKiWiki cms groupware http://www.tikiwiki.org

[18]Archimedes Tiki http://archimedes-tiki.eu

CobWeb Multidimensional Model: Filtering Documents using Semantic Structures and OLAP

Omar Khrouf, Kaïs Khrouf
University of Sfax, MIR@CL
Laboratory, Tunisia
Omar.khrouf@yahoo.fr,
Khrouf.Kais@isecs.rnu.tn

Abdulrahman Altalhi
Faculty of Computing and IT,
King Abdulaziz University,
Jeddah, Saudi Arabia
ahaltalhi@kau.edu.sa

Jamel Feki
Faculty of Computing and IT,
University of Jeddah,
Jeddah, Saudi Arabia
Jamel.Feki@gmail.com

***Abstract*—Today, the documents constitute a capitalization of knowledge in the Information Systems of companies. For the decision-makers, analyzing the contents of documents represents a real challenge. This paper proposes an approach based on the *CobWeb* model to filter semantic structures in order to find documents relevant to the decision-makers' needs. In order to validate our approach, we have developed a GUI for the multidimensional queries and we have applied the Online Analytical Processing (OLAP) analysis on 250 documents taken from the academic domain.**

Keywords-XML documents; standard facet; OLAP; multidimensional model.

I. Introduction

The information systems of companies accumulate, over time, an important volume of data. With the web applications, the users can improve the internal communication within the company by creating, sharing, and modifying real-time work files. The information sharing and the professional work are essential for communication and business productivity. Faced with the rapid development of data (particularly in Web applications), the decision-making process has become an essential activity and an important research area, which requires the implementation of efficient systems called Decision Support Systems (DSS). In addition to the classical DSS systems, which handle numeric data, several studies have been interested in the exploitation of documentary information in order to extract semantic knowledge, for example, the multi-representation of documents using a set of "Facets" [8], or the OLAP of documents [13].

For the multi-representation of documents, some authors, such as Hernandez and al. [8] and Charhad and al. [4] have proposed to use a set of facets in order to describe the useful aspects of documents. These facets take into account not only the semantic aspect, but also other factors related to the exploitation context of documents in order to better satisfy the users' needs. However, the various proposed facets vary according to the application field. Therefore, it would be interesting to define the standard facets, which enable the representation of documents in any research area.

For the OLAP of documents, two categories of works can be distinguished: (1) Those which have adopted the classical multidimensional model, i.e., the star, the snowflake and the constellation models by enriching them with extensions for textual processing ([6] and [7] for data-centric documents; [15] for document-centric documents); and (2) Those which have proposed specific models for the OLAP of documents, such as *galaxy model* [13] and *diamond model* [1]. However, these studies did not treat the heterogeneity of structures and hence, require the definition, in advance, of parameters and hierarchies.

In order to give more flexibility to the user in OLAP analysis tasks, we have proposed a multidimensional model called "*CobWeb model*", as an extension of the galaxy model dedicated to the OLAP of documents based on standard facets [9]. Each facet includes a set of data and is considered as a means of expression for the user's needs. For this reason, we have transformed every facet into a dimension. In multidimensional modeling, each dimension has a structure composed of a set of attributes called parameters, arranged hierarchically from the finest to the highest granularity (e.g., the Time dimension is composed of: Day < Week < Month < Quarter < Semester < Year). The dimension can be considered as an analysis axis; a parameter represents an analysis level and may be associated with one or several descriptive attributes, commonly called weak attributes. However, the integration of facets into an OLAP model creates a set of new problems, for which the classical models of the literature are unable to solve. Such problems arise from the multiple use of the same dimension within the same analysis, and from the concept of recursivity for a parameter of a given hierarchy, etc. To overcome these problems, we have proposed a set of extensions in the *CobWeb* model such as the exclusion constraint between two dimensions, which doesn't allow using these two dimensions in the same analysis. The recursive parameters are used when the hierarchy parameters are not known in advance. The duplicated dimension allows the use of the same dimension twice in the same analysis, whereas the correlated dimension enables the movement between dimensions in the same analysis.

This paper introduces the *CobWeb* model concepts and presents our approach of document filtering by using Semantic Structures. It is organized as follows. Section 2 presents the related works dealing with the representation and exploitation of facets of documents and the OLAP of documentary information. Then, we define in Section 3 the set of five proposed standard facets of documents. Section 4 describes the *CobWeb* multidimensional model focusing on its specificities. Section 5 presents the filtering of documents and the OLAP querying. Finally, Section 6 is reserved for the conclusion.

II. Related works

In this section, we first overview the related works dealing with the multidimensional modeling of documents. Then, we examine the major works dealing with the representation and the exploitation of facets extracted from documents.

For the multidimensional modeling of documents, most works have adopted the three proposed models in the literature for the factual data (star model, snowflake model and constellation model [10]) and have suggested some approaches or functions for the analysis of textual content.

Tseng and al. [14] have used the star schema in order to analyze documents. This schema distinguishes between three types of dimensions: metadata (describing the document, e.g., author, language), ordinary (an ordinary dimension contains keywords extracted from the document), and category (external data for the document description as issued from Wordnet). However, it is limited to a simple count of general documents (e-mails, articles, Web pages, etc.) according to dimensions.

Boussaid and al. [3] have proposed a modeling in snowflake of multidimensional XML data with data mining methods. These studies allow the analysis of complex data, but are not adapted for the analysis of textual data from XML documents.

Azabou and al. [1] have proposed a *diamond model*, which is the star model enriched with a central dimension that attempts to represent the semantics of the document. The parameters of this semantic dimension are linked to parameters of other dimensions. The main disadvantage of this work is that it proposes a model made by a collection of documents with the same structure. Ravat and al. [13] have proposed a multidimensional model, called *Galaxy*, which is adapted to the analysis of XML documents. A galaxy schema is uniquely based on the dimension concept; it connects several dimensions by nodes instead of facts. A connecting node denotes compatible dimensions for analysis. However, this work does not take into account the heterogeneity of document structures.

Zhang and al. [15] have proposed a new model called *Topic Cube,* based on the star schema which extends the traditional data cube by integrating a hierarchy of topics as an analytical dimension. It is a new cube model using a topic dimension and a text content measure which uses parameters of a probabilistic model. However, *Topic Cube* supports only a predefined set of themes.

We notice that the studies dealing with the OLAP of documents provide the analysis of the documents having the same or similar structures.

For the representation of documents, the concept of facet has been used in several domains and with different types of documents.

For tweets, Kumar and al. [11] have proposed a navigation system by facet called Navigating Information Facets on Twitter *(NIF-T)* based on three facets: the Geo Facet showing the location of tweets in a map. Subject facet is a word showing the different thematic exchanges by the tweets. Time facet presents the number of tweets in a given date.

For the video documents, Charhad and al. [4] have proposed to widen the Extended Model for Image Representation and Retrieval *(EMIR²)* created by Mechkour [12] in order to include audiovisual documents. They have added two facets: the temporal facet and the event facet. These two facets characterize the dynamic aspect, which is specific for this type of document. This new model allows the synthetic and integrated consideration of information about the image, text and sound elements.

For textual documents, Hernandez and al. [8] have proposed a model based on a multi-facet representation of documents in order to associate several facets into the same document. They have defined two types of facets: the first one represents the semantics of the contents and the second one includes parameters aiming to improve the research results of documents, such as the description of the educational theories, the description by metadata, etc.

As a conclusion, we notice that some of the studies which have used various facets vary depending on the application domain. However, for other approaches, the facets are fixed for a specific application domain.

The purpose of this paper is to integrate the notion of facet in the OLAP model because it is interesting to represent documents from several points of view. Then we propose an approach based on the *CobWeb* model to filter semantic structures in order to determine the documentary information for the user's needs.

III. Standards facets of documents

To the best of our knowledge, the concept of facet has not been addressed in the decision domain. In order to provide a facet-based OLAP model, we define a set of five facets to represent one or many documents according to a given viewpoint. These facets must be standard, i.e., independent of any specific domain of application and must give the user the ability to consider the same document or set of documents from multiple views (Metadata, Keyword, etc), so that he can have a more targeted access to information as needed [9].

- The *Metadata* Facet: this facet aims to provide the users with a structured collection of the data describing a document (such as: title, rights, format, etc.). In our work, we use the metadata defined by the Dublin Core [5].
- The *Keyword* Facet: this facet constitutes a set of the most important keywords describing the content of the document. These keywords can be determined, by using the indexing techniques of information retrieval, or they come from the document itself when they exist explicitly.
- The *Content* Facet: this facet aims to present the information contained in the document (image, text,

etc.) by removing everything about the comments, structure, etc.

- The *Semantic* Facet: this facet describes the semantics of the content of the document. It is used in the classification of all or parts of the documents in order to facilitate the retrieval /analysis of these documents. For the determination of this semantics, we have relied on the work in [2] which defines a method for the determination of a semantic structure for a given document.
- The *Structural* Facet: this facet is a viewpoint of the structure of a document. It aims to focus on parts of the document (section, subsection, etc.) and not the whole document.

Based on the previously defined facets, we present, in the following sections, the *CobWeb* multidimensional model devoted to the OLAP of documents. Then, we present an approach to filter documents by using semantic structures and OLAP querying.

IV. COBWEB MULTIDIMENSIONAL MODEL

In this section, we present the *CobWeb* multidimensional model (Fig. 1), which is an extension of the galaxy model based on standard facets in order to provide more opportunities for the expression of analytic queries and a more targeted vision of the data to decision makers. To build this model, the main idea consists in transforming every facet into a dimension since these facets may represent a means of expressing the users' viewpoints and therefore, describe their requirements. Besides, we have added the dimension *Document* in order to link the information from different facets to their documents. *CobWeb* differs from the existing models by the following extensions:

- **Duplicated Dimension:** The classical multidimensional modeling does not allow using the same dimension twice in the same analysis. Let us suppose that we want to analyze the documents by two parameters belonging to the same Metadata dimension (namely, *Date* and *Editor*). This type of query is not possible. In order to give more flexibility to the user in the task of OLAP analysis, we propose the duplicated dimension, which can participate many times in the same analysis. Graphically, a duplicated dimension is symbolized by the letter **D** in the concerned dimension. In the *CobWeb* model, we have only one duplicated dimension, called *Metadata* (Fig. 1).
- **Recursive Parameter:** In the classical schema of data warehouses, the parameters and dimension hierarchies are known in advance. However, in our work:
    - ○ The structure of documents may differ from one collection to another.
    - ○ The semantic structure of documents is determined from taxonomies (hierarchical representation of the concepts) and helps describe the textual content of documents. Specifically, the concepts of taxonomies will be assigned to different parts of the documents. Therefore, the number of concepts and levels varies from a semantic structure to another. For the representation of these two dimensions, we will use a new type of parameters, called recursive parameter, since the documents and the taxonomies are represented in a hierarchical manner.
    - ○ The structural dimension helps us to move from one level to another (Content →Section →SubSection →Paragraph) using the conventional OLAP operators namely *RollUp* and *Drill Down*.
    - ○ The semantic dimension allows the movement between concepts (Information System →Data Warehouse →Cube, etc.).

  Graphically, a recursive parameter is schematized by a directed loop (Fig. 1).
- **Correlated Dimensions:** In the classical multidimensional modeling, the movements between the dimensions cannot be achieved because of the absence of inter-dimensional relationships. To solve this problem, we propose the concept of correlated dimensions which allows, for the same query, to move between dimensions. Graphically, the correlation that can be possible between the dimensions of our multidimensional model is represented by dashed arrows between dimensions. The transition from one dimension to another is accepted when we respect the direction of the arrow. For example, it is possible to move from the Content dimension to the Semantic dimension.
- **Exclusion Constraint between Dimensions:** The exclusion constraint requires that a couple of dimensions cannot be used simultaneously in the same analysis. In *CobWeb*, the exclusion constraint concerns the *Content* and the *Structural* dimensions because an analysis must concern the content or parts of the documents (title, section, paragraph, etc.), but not both at the same time. Graphically, this exclusion constraint is denoted by a circle containing the letter **X** connected to the involved dimensions, such as: *Document* and *Structural* in Figure 1.

V. FILTERING AND OLAP QUERING

In this section, we describe our approach of filtering documents using Semantic Structures and OLAP querying in order to find the documentary information relevant to the decision-makers' needs according to several analysis axes, as shown in Figure 2.

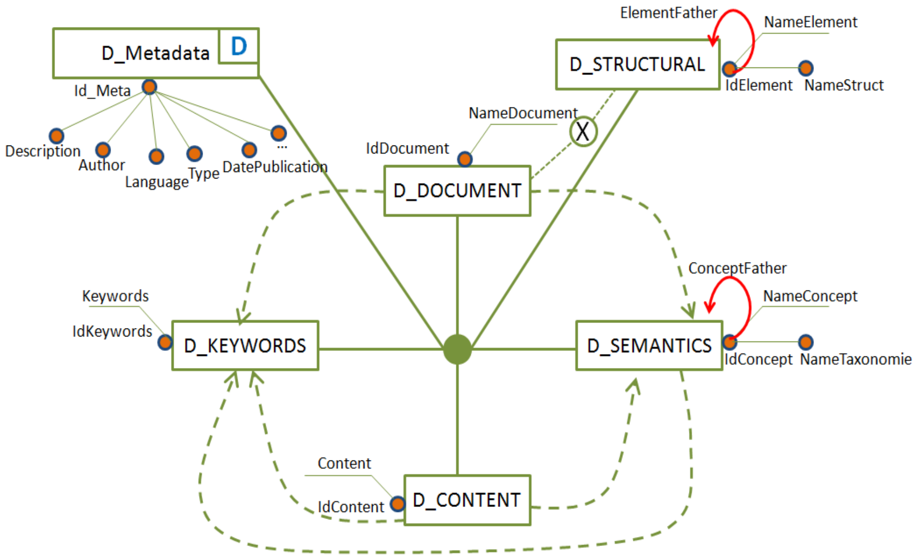


Figure 1. *CobWeb* multidimensional model.

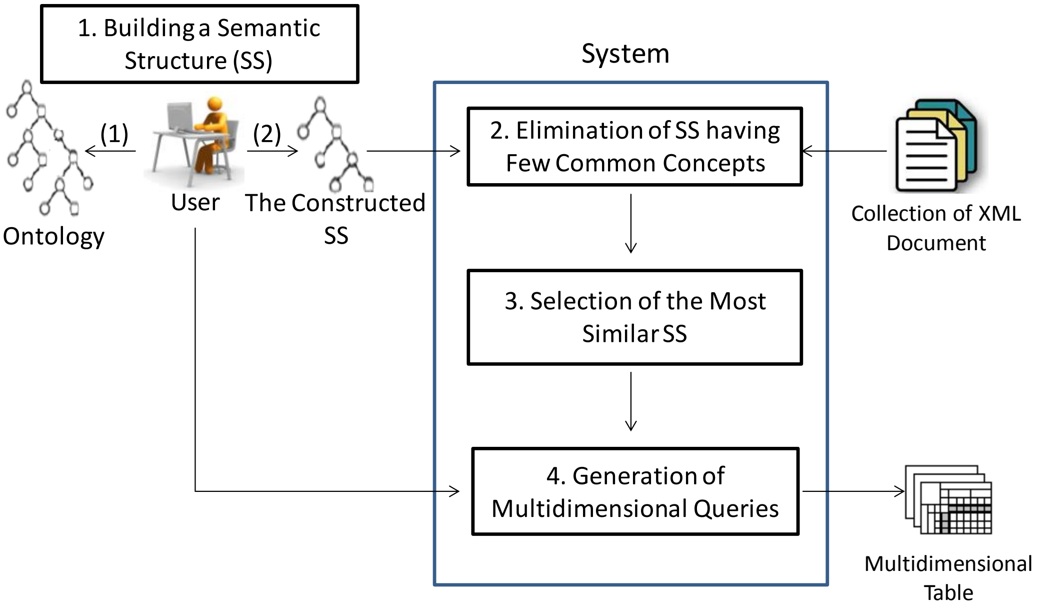

1. Building a Semantic Structure (SS)
System
(1)
(2)
Ontology
User
The Constructed SS
2. Elimination of SS having Few Common Concepts
Collection of XML Document
3. Selection of the Most Similar SS
4. Generation of Multidimensional Queries
Multidimensional Table

Figure 2. The proposed process of querying

.

This approach is based on four fundamental stages:

- The first step is building a semantic structure from a selected ontology. The user defines the closest concepts to his needs and puts them together into a semantic structure according to a set of predefined rules.
- The second step consists in a first filtering of the documents to be analyzed by the elimination of the semantic structures of documents having few common concepts with the built semantic structure defined by the user.
- The third step is a second filtering of documents in order to keep those having the most similar semantic structures compared to the semantic structure defined by the user.
- The last step in our approach is the generation of the multidimensional queries and the result will be displayed as a multidimensional table.

We note that we have automatically generated, in our previous works, a semantic structure for every XML document; this semantic structure is superposed on its logical structure [2] (Fig. 3).

The purpose behind the use of semantic structures, in this phase of querying, is to keep only the relevant documents for the user's need. In what follows, we explain the different steps of our approach.

A. *Step 1:Building a Semantic Structure (SS)*

In order to build his semantic structure, the user chooses, through a web application, a semantic resource to select a set of concepts depending on his needs and organizes them in a hierarchical way. This web application allows communicating and exchanging the semantic structures between systems or applications in order to determine the documentary information for the user's needs. The user will be assisted by the system that displays error messages for the incorrect manipulations and it suggests one or more solutions (Fig. 4).

A semantic resource (ontology, taxonomy, thesaurus, etc.) serves to represent the semantics of a given domain in a generic and reusable way in order to share knowledge and data.

To build his semantic structure, the user must respect the following rules:

- Rule 1: No reverse hierarchical order between concepts. Example: the ontology of Figure 4 shows that the *OLAP* is the father concept of the *Dimension*, in the semantic structure built by the user which it is prohibited to represent the *Dimension* as the father concept of the *OLAP*.
- Rule 2: The conceptual father of the concepts selected by the user represents the common ancestor of these concepts in ontology. Example: Figure 4 shows that the user has selected the *OLAP* and *Dimension* concepts. The conceptual father attributed to these concepts is *Data warehouse* because, in ontology, it represents the common ancestor of selected concepts.

B. *Step 2: Elimination of SS with Few Common Concepts*

Once the semantic structure is built by the user, we compare this structure with the semantic structures of documents, in order to keep the pertinent structures, i.e., eliminate the structures having few common concepts with the semantic structure built by the user. For this reason, we propose the measure of similarity (1).

$$\mathrm{Sim}_{(SSC,\,SSU)} = \frac{\left| C_{SSC} \right|}{\left| C_{SSU} \right|} \qquad (1)$$

$|C_{SSC}|$: Number of common concepts between the semantic structure of the user and the semantic structures of documents.

$|C_{SSU}|$: Number of concepts in the semantic structure of the user.

Sim $_{(SSC,\ SSU)}$: The similarity degree between the two semantic structures.

We define a threshold for selecting structures. This threshold is determined by experiments and may be modified by the user according to his need.

C. *Step 3: Selection of the most Similar SS*

In our work, the order of concepts (father-son relation) is very important so, we propose to start comparing branches. A branch is a path composed of all the concepts between the root and the leaf of the semantic structure. Example: the branches of the semantic structure built by the user (Fig. 4) are:

Branch 1: Information System → Graph.

Branch 2: Information System → Data Base.

Branch 3: Information System → Data Warehouse → Dimension.

Branch 4: Information System → Data Warehouse → OLAP.

The measure of similarity (2) allows comparing two branches of both semantic structures.

$$\mathrm{SimB}_{(SSC,\,SSD)} = \left(\frac{\left| CA_{(SSC,\,SSD)} \right|}{\left(\left| CA_C \right| + \left| CA_D \right| \right) / 2} + \frac{\left| AA_{(SSC,\,SSD)} \right|}{\left| AA_C \right|} \right) / 2 \qquad (2)$$

$|CA_{(SSC,\ SSD)}|$: Number of common concepts between the two mapped branches.

$|CA_C|$: Number of concepts in the branch of the built semantic structure.

$|CA_D|$: Number of concepts in the branch for the semantic structure of the document.

$|AA_C|$: Number of arches in the branch of the built semantic structure.

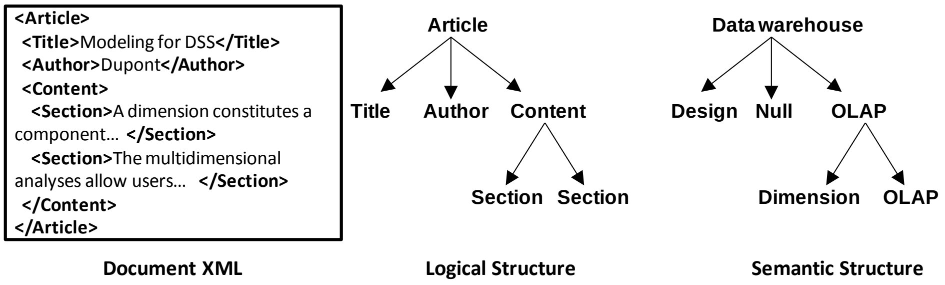
$|AA_{(SSC,\ SSD)}|$: Number of common arches between the two mapped branches.

Table I presents an example of comparison between two branches of two different documents with a branch of the semantic structure built by the user (Information System (IS) → Data warehouse (DW) → Dimension (DIM)).

TABLE I. A COMPARISON TABLE BETWEEN TWO BRANCHES OF TWO SEMANTIC STRUCTURES

<table>
<tr><th>Branche of SSC</th><th>Branche of SSD</th><th>Degree of similarity <b>SimB_(SSC, SSD)</b></th></tr>
<tr><td rowspan="2">IS
|
DW
|
DIM</td><td>IS
|
X
|
DW
|
DIM</td><td>✓3 common concepts(<b>IS</b>, <b>DW</b>, <b>DIM</b>)
✓One arch aligned: <b>DW</b> ⟶ <b>DIM</b>
$\mathrm{SimB}_{(SSC,\,SSD)} = \left(\frac{3}{(3+4)/2} + \frac{1}{2}\right)/2$
$= (0.85 + 0{,}5)/2$
$= 0.67$</td></tr>
<tr><td>IS
|
DIM</td><td>✓2 common concepts(<b>IS</b>, <b>DW</b>)
✓ 0 arch aligned
$\mathrm{SimB}_{(SSC,\,SSD)} = \left(\frac{2}{(3+2)/2} + \frac{0}{2}\right)/2$
$= 0.8 + 0\ /2$
$= 0.4$</td></tr>
</table>

ISBN: 978-1-61208-412-1



```
<Article>
  <Title>Modeling for DSS</Title>
  <Author>Dupont</Author>
  <Content>
    <Section>A dimension constitutes a component... </Section>
    <Section>The multidimensional analyses allow users... </Section>
  </Content>
</Article>
```

Figure 3. Logical and semantic structure of XML document.

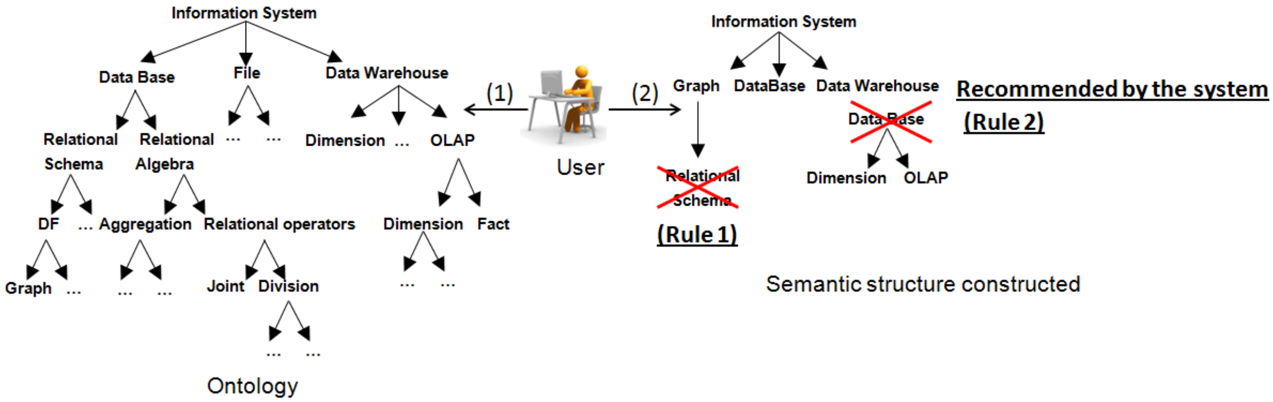


Figure 4. Example of building a semantic structure.

At this level, we need to calculate the weight sum of the different branches. The documents having a similarity degree above a threshold (fixed by experiment and may be modified by the user) will be selected for the OLAP Query.

D. *Step 4: Generation of Multidimensional Queries*

For the multidimensional querying, the user should specify his query by indicating the fact and its measures and the various dimensions. Then, the system automatically generates the needed queries.

We have developed a GUI for the multidimensional querying. In the left part of the interface, the user specifies his request by indicating the dimensions and the fact. The right part is devoted to the results of the query as a multidimensional table. To validate our work, we test and evaluate our approach on 250 documents taken from the academic domain.

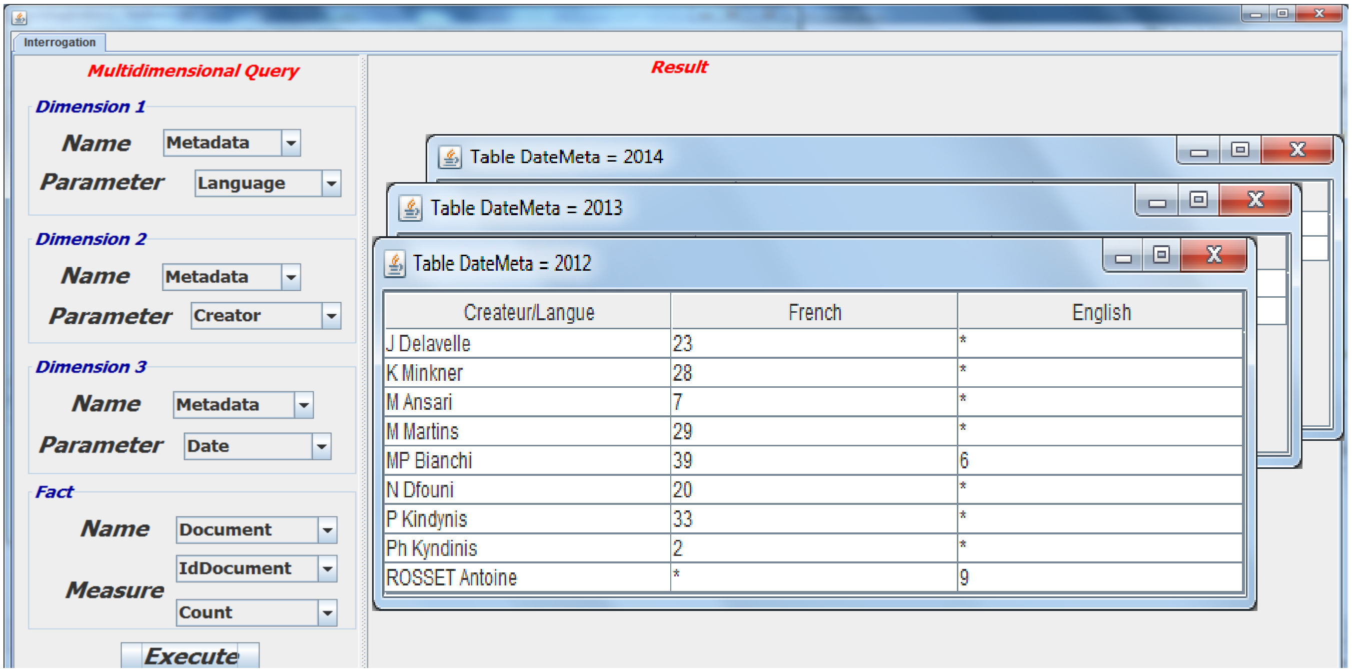
Figure 5 shows the results of the previous query in a multidimensional table, where the columns and the lines represent the first two dimensions (*Author* and *Language*), and where the plans represent the third dimension (*Date*).

The measures are placed in the intersection of a line and a column for a given plan. The symbol * indicates that there is no value for the measure. In this experiment, we observe that most documents are written in French (181 documents in 2012). So we can note that the majority of authors are francophone.

VI. Conclusion

In this paper, we have proposed an approach based on the *CobWeb* model to filter the documents using Semantic Structures in order to determine the documentary information for the user's needs. In addition, we have developed a GUI for the multidimensional querying.

The main limitation of our approach is that the user builds his semantic structure by using a single semantic resource; it would be interesting to offer more opportunities to the user in order to build his semantic structure from several resources and not just to one. We also intend to propose new OLAP operators that take into consideration the specificities of the *CobWeb* model, for example an operator for the correlation of dimensions. These operators will facilitate the interpretation of the results of the multidimensional analyses. In the long run, we plan to introduce the personalized OLAP analysis which takes into account the needs and skills of the users, based on their profiles.



Dimension 1
Name: Metadata
Parameter: Language

Dimension 2
Name: Metadata
Parameter: Creator

Dimension 3
Name: Metadata
Parameter: Date

Fact
Name: Document
Measure: IdDocument, Count

Table DateMeta = 2014

Table DateMeta = 2013

Table DateMeta = 2012

Createur/Langue	French	English
J Delavelle	23	*
K Minkner	28	*
M Ansari	7	*
M Martins	29	*
MP Bianchi	39	6
N Dfouni	20	*
P Kindynis	33	*
Ph Kyndinis	2	*
ROSSET Antoine	*	9

Figure 5. Graphical multidimensional querying.

REFERENCES

[1] M. Azabou, K. Khrouf, J. Feki, C. Soulé-Dupuy, and N. Vallès, "A Novel Multidimensional Model for the OLAP on documents: Modeling, Generation and Implementation," International Conference on Model & Data Engineering, Larnaca, Cyprus, September 2014, pp. 258–272.

[2] S. Ben Mefteh, K. Khrouf, J. Feki, and C. Soulé-Dupuy, "Semantic Structure for XML Documents: Structuring and pruning," Journal of Information Organization, vol. 3, issue 1, March 2013, pp. 37-46.

[3] O. Boussaid, R. Ben Messaoud, R. Choquet, and S. Anthoard, "X-Warehousing : An XML-Based Approach for Warehousing Complex," Proc. The 10 th East-European Conference on Advances in Databases and Information Systems (ADBIS 06), vol. 4152, September 2006, pp. 39-54.

[4] M. Charhad and G. Quénot, "Semantic Video Content Indexing and Retrieval using Conceptual Graphs," Proc. IEEE Information and Communication Technologies: From Theory to Applications (ICTTA 04), IEEE Press, April 2004, pp. 399-400, doi: 10.1109/ICTTA.2004.1307800.

[5] Dublin Core Metadata Initiative (DCMI), Dublin Core Metadata Element Set, Version 1.1, ISO Standard 15836, 2007. http://dublincore.org/documents/dces/.

[6] J. Feki, I. Ben Messaoud, and G. Zurfluh, "Building an XML Document Warehouse," Journal of Decision Systems (JDS). Ed. Taylor & Francis, vol. 22, issue 2, April 2013, pp. 122-148, doi: 10.1080/12460125.2013.780322.

[7] Y. Hachaichi and J. Feki, "An Automatic Method for the Design of Multidimensional Schemas from Object Oriented Databases," International Journal of Information Technology and Decision Making, vol. 12, issue 12, November 2013, pp. 1223-1259, doi : 10.1142/S0219622013500351.

[8] N. Hernandez, J. Mothe, B. Ralalason, B. Ramamonjisoa, and P. Stolf, "A Model to Represent the Facets of Learning Objects," Interdisciplinary Journal of E-Learning and Learning Objects, Information Science Institute, Santa Rosa - USA, vol. 4, January 2008, pp. 65-82.

[9] O. Khrouf, K. Khrouf, and J. Feki, "CobWeb Multidimensional Model: From Modeling to Querying," International Conference on Model & Data Engineering, Larnaca, Cyprus, September 2014, pp. 273–280.

[10] R. Kimball and M. Ross, The Data Warehouse Toolki: The Definitive Guide to Dimensional Modeling, 3rd ed., John Wiley & Sons, New York, July 2013.

[11] S. Kumar, F. Morstatter, G. Marshall, H. Liu, and U. Nambiar, "Navigating Information Facets on Twitter (NIF-T)," Proc. The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 12), August 2012, pp. 1548-1551.

[12] M. Mechkour, "A multifacet formal image model for information retrieval," Proc. The Final WorkShop on Multimedia Information Retrieval (MIRO 95), September 1995, pp. 18-20.

[13] F. Ravat, O. Teste, R. Tournier, and G. Zurfluh, "Designing and Implementing OLAP Systems from XML Documents," Proc. Annals of Information Systems, Springer, Special issue New Trends in Data Warehousing and Data Analysis, vol. 3, November 2008, pp. 1-21, doi: 10.1007/978-0-387-87431-9_15.

[14] F. S. C. Tseng and A. Y. Chou, "The concept of document warehousing for multidimensional modeling of textual-based business intelligence," Journal Decision Support System (DSS), vol. 42, issue 2, November 2006, pp. 727-744.

[15] D. Zhang, C. Zhai, J. Han, A. Srivastava, and N. Oza, "Topic modeling for OLAP on multidimensional text databases: topic cube and its applications," Journal Statistical Analysis and Data Mining, vol. 2, issue 5-6, December 2009, pp. 378-395.