

**Mimoun Malki
Salima Benbernou
Sidi Mohamed Benslimane
Ahmed Lehireche (Eds.)**

ICWIT 2012

Web and Information Technologies

**4th International Conference on Web and Information Technologies
ICWIT 2012, Sidi Bel-Abbes, Algeria, April 29-30 2012**



Proceedings

Mimoun Malki
Salima Benbernou
Sidi Mohamed Benslimane
Ahmed Lehireche (Eds.)

Web and Information Technologies

4th International Conference on Web and Information Technologies
ICWIT 2012, Sidi Bel-Abbes, April 29-30 2012



Proceedings

Preface

Welcome to the fourth edition of the International Conference on Web and Information technologies, ICWIT 2012. This year the ICWIT conference continued the tradition that has evolved from the inaugural conference held in 2008 in Sidi Bel-Abbes and since then has made its journey around the Maghreb: 2009 Sfax (Tunisia) and 2010 Marrakech (Morocco).

This year we were happy to hold the event in Sidi Bel-Abbes, a city of 300,000 inhabitants in western Algeria. Sidi Bel-Abbes's geographical location has predestined the city to be a significant scientific, cultural and economic center with more than just regional influence.

The ICWIT 2012 Conference provided a forum for research community and industry practitioners to present their latest findings in theoretical foundations current methodologies and practical experiences. ICWIT 2012 will focus on new research directions and emerging applications in Web and Information Technologies. The submitted contributions address challenging issues of Web Technologies, Web Security, Information Systems, Ontology Engineering and Wireless Communications.

The 136 papers submitted for consideration for publication originated from 7 countries from all over the world: Algeria, Brazil, Belgium, France, Morocco Saudi Arabia, and Tunisia. After a thorough reviewing process, 30 papers were selected for presentation as full papers – the acceptance rate was 22%. In addition 15 papers were selected for presentation as posters, yielding an overall acceptance rate of 33%. The papers published in these proceedings are included in CEUR-WS.org Workshop Proceedings service and indexed by DBLP. Best Papers will be recommended for publication in special issues of journals like: International Journal of Information Technology and Web Engineering (IJITWE), International Journal of Metadata, Semantics and Ontologies (IJMSO) and International Journal of Reasoning-based Intelligent Systems (IJRIS).

We believe that this volume provides an interesting and up-to-date picture of what are the last trends and new ideas fermenting right now in the Web and information technologies community. Some of the papers included in this volume unveil unexpected, novel aspects and synergies that we think will be taken up in the future and may become main-stream research lines.

This conference was made possible through the efforts of many people. We wish to thank everyone involved, including those who worked diligently behind the scenes and without formal recognition. First, we would like to thank the ICWIT Steering Committee for selecting the Djillali liabes University of Sidi Bel-Abbes to hold ICWIT 2012 conference. Great thanks to Conference Honorary President Abdenacer Tou, head of Djillali liabes University, for all his encouragement and financial support to make sure the success of this conference.

Without enthusiastic and committed authors this volume would not have been possible. Thus, our thanks go to the researchers, practitioners, and PhD students who contributed to this volume with their work. We would like to thank the Program Committee members and reviewers, for a very rigorous and outstanding reviewing process. Our thanks should also reach the Organizing Committee of the conference, for their dedication and hard work in coordinating the organization of a wide array of interesting papers presentation, tutorials, posters and panels that completed the program of the conference, and providing an excellent service in all administrative and logistic issues related to the organization of the event

Special thanks go to the various sponsors – Djillali Liabes University Evolutionary Engineering and Distributed Information Systems Laboratory National Administration of Scientific Research and National Agency of University Research Development – who kindly support this 4th edition of ICWIT 2012 and make these proceedings available.

We wish to thank Aris M. Ouksel (University of Illinois at Chicago, USA), and Mourad Ouzzani (QCRI Doha, Qatar), for graciously accepting our invitations to serve as keynote speakers.

April 2012

Mimoun Malki
Salima Benbernou
Sidi Mohamed Benslimane
Ahmed Lehireche

Organization

Conference Honorary President

Prof. Abdenacer Tou Head of Djillali Liabes Unviversity

Conference General Chair

Mimoun Malki Djillali Liabes University of Sidi Bel-Abbes, Algeria

Steering Committee

Ahmed Lehireche	Djillali Liabes University of Sidi Bel-Abbes, Algeria
Boualem Benatallah	CSE Sydney, Australia
Djamal Benslimane	University of Lyon1, France
Faiez Gargouri	ISIMSF Sfax, Tunisia
Ladjel Bellatreche	ENSMA Poitiers, France
Mimoun Malki	Djillali Liabes University of Sidi Bel-Abbes, Algeria

Program Committee Chair

Salima Benbernou Paris Descarte University, France

Program Committee members:

El Hassan Abdelwahed	UCAM University, Morocco
Mustapha Kamel Abdi	Es-Sénia University of Oran, Algeria
Driss Aboutajdine	FSR Mohammed V University, Morocco
Réda Adjoudj	University of Sidi Bel-Abbes, Algeria
Mohamed Ahmed Nacer	USTHB University, Algeria
Rachid Ahmed-Ouamer	University of Tizi Ouzou, Algeria
Yamine Ait Ameer	IRIT-ENSEIHT Toulouse, France
Otmane Ait Mohamed	Concordia University, canada
Idir Aitsadoune	SUPELEC Gif Paris, France
Fahad Ahmed Al-Zahrani	UQU University, Saudi Arabia

Zaia Alimazighi	USTHB University, Algeria
Djamel Amar Bensaber	University of Sidi Bel-Abbes, Algeria
Youssef Amghar	INSA Lyon, France
Abdelmalek Amine	University of Saida, Algeria
Baghdad Atmani	University of Oran, Algria
Nadjib Badache	CERIST Algiers, Algeria
Hassan Badir	ENSA Tanger, Morocco
Youcef Baghdadî	Sultan Qaboos University, Oman
Karim Baina	ENSIAS Rabat, Morocco
Amar Balla	ESI Algiers, Algeria
Kamel Barkaoui Barkaoui	CNAM Paris, France
Ghalem Belalem	University of Oran, Algeria
Bouziane Beldjilali	University of Oran, Algeria
Abdelghani Bellaachia	George Washington University, USA
Ladjel Bellatreche	ENSMA Poitiers, France
Fatima Zahra Belouadha	EMI Mohamadia, Morocco
Boualem Benatallah	CSE Sydney, Australia
Nabila Benharkat	INSA Lyon, France
Mohamed Benmohamed	University of Constantine, Algeria
Djamel Bennouar	University of Blida
Kamel Bensalem	Manar University of Tunis, Tunisia
Djamal Benslimane	University of Lyon1, France
Sidi Mohamed Benslimane	University of Sidi Bel-Abbes, Algeria
Abdelkader Benyettou	USTO University, Algeria
Fatiha Boubekour	IRIT Toulouse, France
Noureddine Boudriga	SUPCOM Tunis, Tunisia
Mahmoud Boufaïda	University of Constantine, Algeria
Zizette Boufaïda	University of Constantine, Algeria
Kamel Boukhalifa	USTHB University, Algeria
belahJalil Boukhobza	UBO-University of Occidental Brittany, France
Azedine Boulmakoul	FST Mohammedia, Morocco
Omar Boussaid	University of Lyon2, France
Lotfi Bouzguenda	ISIMSF Tunis, Tunisia
Allaoua Chaoui	University of Constantine, Algeria
Chihab Cherkaoui	ENCG Rabat, Morocco
Azzedine Chikh	University of Tlemcen, Algeria,
Mohamed Amine Chikh	University of Tlemcen, Algeria

Salim Chikhi	University of Constantine, Algeria
Samir Chouali	University of Franche-Comté, France
Abdellah Chouarfia	USTO University, Algeria
Souad Chraïbi	UCAM University, Morocco
Alfredo Cuzzocrea	ICAR-CNR and University of Calabria, Italy
Jerome Darmon	University of Lyon2, France
Abdelouahid Derhab	CERIST, Algiers
Noureddine Djedi	University of Biskra, Algeria
Djamel Djenouri	CERIST Algiers, Algeria
Habiba Drias	USTHB University, Algeria
Abdelaziz El Fazzikki	UCAM University, Morocco
Essaid Elbachari	UCAM University, Morocco
Zakaria Elberrichi	University of Sidi Bel-Abbes, Algeria
Mohammed Erradi	ENSIAS, Morocco
Kamel Mohamed Faraoun	University of Sidi Bel-Abbes, Algeria
Jamel Feki	FSEGS Sfax, Tunisia
Andre Flory	INSA Lyon, France
Abdelkader Gafour	University of Sidi Bel-Abbes, Algeria
Momo Gammoudi	University of Tunis, Tunisia
Faiez Gargouri	ISIMSF Sfax, Tunisia
Khaled Ghedira	ISG Tunis, Tunisia
Herve Guyennet	University of Franche-Comté, France
Fatima Zohra Hadjam	University of Sidi Bel-Abbes, Algeria
Hafid Haffaf	University of Oran, Algeria
Ahmed Hammad	University of Franche-Comté, France
Zahi Jarir	Ucam University, Morocco
Wassim Jaziri	ISIMSF, Sfax, Tunisia
Stéphane Jean	ENSMA, Poitiers, France
Anis Jedidi	ISIMS, Sfax, Tunisia
Okba Kasar	University of Biskra, Algeria
Bouabdellah Kechar	University of Oran, Algeria
Abdelaziz Khadraoui	University of Geneva, Switzerland
Hamamache Kheddouci	University of Lyon1, France
Mohamed-Khireddine Krolladi,	University of Constantine, Algeria
Mouloud Koudil	ESI, Algiers
Azzeddine Lazrek	UCAM University, Morocco

Yahia Lebbah	University of Oran, Algeria
Ahmed Lehireche	University of Sidi Bel-Abbes, Algeria
Sofian Maabout	LABRI, University of Bordeaux, France
Zakaria Maamar	Zayed University, UAE
Walid Mahdi	ISIMSF, Sfax, Tunisia
Djoudi Mahieddine	University of poitiers, France
Qusay H. Mahmoud	University of Guelph, Canada
Mimoun Malki	University of Sidi Bel-Abbes, Algeria
Patrick Marcel	University of Tours, France
Belhadri Messabih	USTO, Oran, Algeria
Mohamed Mezghiche	University of Boumerdes, Algeria
Abdellatif Mezrioui	INPT, Casablanca, Morocco
Rokia Missaoui	UQO, Montréal Canada
Abdelillah Mokkedem	ST , Morocco
Hassan Mountassir	University of Franche-Comté, France
Abdelouahab Moussaoui	Ferhat Abbas University of Setif, Algeria
Safia Nait Bahloul	University of Oran, Algeria
Kazumi Nakamatsu	SHSE, University of Hyogo, Japan
Tho Nguyen Manh	ITS Vienna, Austria
Rachid Nourine	University of Oran, Algeria
Aris M. Ouksel	University of Illinois at Chicago, USA
Mourad Oussalah	University of Nantes, France
Said Raghay	FSSTG, Marrakech, Morocco
Abdellatif Rahmoun	University of Sidi Bel-Abbes, Algeria
Mustapha K. Rahmouni	University of Oran, Algeria
Chantal Reynaud	LRI Orsay, France
Ounsa Roudies	EMI of Mohamadia, Morocco
Mohammed Sadgal	UCAM of Marrakech, Morocco
Djamel Eddine Saïdouni	University of Constantine, Algeria
Michel Schneider	ISIMA Audiere, France
Larbi Sekhri	University of Oran, Algeria,
Mokhtar Sellami	NASR, Algeria
Sid-Ahmed Selouani	UMCS Moncton, Canada
Mohamed Senouci	University of Oran, Algeria
Hassina Seridi	Badji Mokhtar University of Annaba
Michel Simonet	TIMC-IMAG, Grenoble, France
Zohra Slama	University of Sidi Bel-Abbes, Algeria

Yahya Slimani	Manar University of Tunis, Tunisia
Kamel Tari	University of Bejaia, Algeria
Thouraya Tebibel	ESI, Algiers
Mohamed Tmar	ISIMSF, Tunisia
Farouk Toumani	Blaise Pascale University, clermont-ferrant, France
Robert Wrembel	Poznan University of Technology, Poland
Abbes Yagoubi	University of Oran, Algeria
Yuhang Yang	Shanghai Jiao Tong University
Kokou Yetongnon	University of Bourgogne Dijon, France
Abdelrahmane Yousfate	University of Sidi Bel-Abbes, Algeria
Djamel Eddine Zegour	ESI Algiers, Algeria
Djelloul Ziadi	Rouen University, France

Organization Committee Chair

Sidi Mohamed Benslimane University of Sidi Bel-Abbes

Secretariat

Djamel Amar Bensaber	University of Sidi Bel-Abbes
Sofiane Boukli hacene	University of Sidi Bel-Abbes,
Kamel Mohamed Faraoun	University of Sidi Bel-Abbes

Members

Reda Adjoudj,	University of Sidi Bel-Abbes
Mohamed Benhamouda	University of Sidi Bel-Abbes
Abdelouafi Bouamama	University of Sidi Bel-Abbes
Salim Chiali	University of Sidi Bel-Abbes
Zakaria Elberichi	University of Sidi Bel-Abbes
Mohamed Ismail Arrar	University of Sidi Bel-Abbes
Affaf Mérazi	University of Sidi Bel-Abbes
Mohamed Taieb Brahim	University of Sidi Bel-Abbes
Toumouh Adil	University of Sidi Bel-Abbes

Table of Contents

Abstracts of the Invited Talks

Towards Automated Information Factories	2
<i>Aris M. Ouksel</i>	
Data Quality Not Your Typical Database Problem	3
<i>Mourad Ouzzani</i>	

Full Papers

Context driven mediation service in Data-as-a-Service composition.	4
<i>Idir Amine Amarouche and Djamel Benslimane</i>	
Service Substitution Analysis in Protocols Evolution Context	12
<i>Ali Khebbizi, Hassina Seridi-Bouchelaghem, Imed Chemakhi, Hychem Bekakria</i>	
Dynamic Web Service Composition. Use of Case Based Reasoning and AI Planning	22
<i>Fouad Henni and Baghdad Atmani</i>	
A collaborative web-based Application for health care tasks planning	30
<i>Fouzi Lezzar, Abdelmadjid Zidani and Chorfi Atef</i>	
Building Semantic Mashup	40
<i>Abdelhamid Malki and Sidi Mohammed Benslimane</i>	
An approximation approach for semantic queries of naïve users by a new query language	50
<i>Ala Djeddaï, Hassina Seridi and Tarek Khadir</i>	
Semantic annotation of web services	60
<i>Djelloul Bouchiha and Mimoun Malki</i>	
Semantic multimedia search: the case of SMIL documents	70
<i>Mounira Chkiwa and Anis Jedidi</i>	
A Muti-Representation and Generalisation Based Webmapping Approach Using Multi-Agent System	83
<i>Khalissa Derbal, Kamel Boukhalfa and Zaia Alimazighi</i>	
Numerical modeling for an urban transportation system	93
<i>Karim Bouamrane, Hadj Ali Beghdadi and Naima Belayachi</i>	
Urbanization of information systems with a service oriented architecture according to the PRAXEME approach - Application to the Information System of the National Social Insurance Fund (CNAS)	102
<i>Boussis Amel and Nader Fahima</i>	
Using Vector Quantization for Universal Background Model in Automatic Speaker Verification	112
<i>Djellali Hayet and Laskri Mohamed Tayeb</i>	
The Use of WordNets for Multilingual Text Categorization: A Comparative . . . Study.	121

<i>Mohamed Amine Bentaallah and Mimoun Malki</i>	
Enhanced Collaborative Filtering to Recommender Systems of Technology Enhanced Learning	129
<i>Majda Maatallah and Hassina Seridi</i>	
Meta-Learning for Escherichia Coli Bacteria Patterns Classification	139
<i>Hafida Bouziane, Belhadri Messabih and Abdallah Chouarfia</i>	
Ontology-based gene set enrichment analysis using an efficient semantic similarity measure and functional clustering.	151
<i>Sidahmed Benabderrahmane and Hayet Mekami</i>	
Theoretical Overview of Machine translation	160
<i>Cheragui Mohamed Amine</i>	
Effective Ontology Learning : Concepts' Hierarchy Building using Plain Text Wikipedia	170
<i>Khalida Ben Sidi Ahmed and Adil Toumouh</i>	
Security Ontology for Semantic SCADA	179
<i>Sahli Nabil and Benmohammed Mohamed</i>	
Automatic construction of ontology from arabic texts	193
<i>Ahmed Cherif Mazari, Hassina Aliane and Zaia Alimazighi</i>	
Model driven approach for specifying WSMO ontology	203
<i>Djamel Amar Bensaber and Mimoun Malki</i>	
Foundations on Multi-Viewpoints Ontology Alignment	214
<i>Djakhdjakha Lynda, Hemam Mounir and Boufaïda Zizette</i>	
A Flexible Integration of Security Concern in Rule based Business Process modeling	222
<i>Bekki Khadhir and Belbachir Hafida</i>	
Security Requirements Analysis of Web Applications using UML	232
<i>Salim Chehida and Mustapha Kamel Rahmouni</i>	
Development of RSA with random permutation and inversion algorithm to secure speech in GSM networks	240
<i>Khaled Merit and Abdelazziz Ouamri</i>	
Spam Detection System Combining Cellular Automata and Naïve Bayes Classifier	250
<i>Fatiha Barigou, Naouel Barigou and Baghdad Atmani</i>	
Clustering-based data in ad-hoc networks	261
<i>Bakhta Meroufel and Ghalem Belalem</i>	

Short Papers

A Recommendation-based Approach for Communities of Practice of E-learning	270
<i>Lamia Berkani, Omar Nouali and Azeddine Chikh</i>	
AMSI: An Automatic Model-Driven Service Identification from Business Process Models	276

<i>Mokhtar Soltani and Sidi Mohammed Benslimane</i>	
Relations extraction on patterns lacking of Resulting Context	282
<i>Asma Hachemi and Mohamed Ahmed-Nacer</i>	
Reverse Engineering Process for Extracting Views from Domain Ontology. . . .	288
<i>Soraya Setti Ahmed</i>	
Multi-Agents Model for Web-based Collaborative Decision support systems	294
<i>Abdelkader Adla and Bakhta Nacet</i>	
Agent-based Approach for Mobile Learning using Jade-LEAP	300
<i>Khamsa Chouchane, Okba Kazar and Ahmed Aloui</i>	
New Web tool to create educational and adaptive courses in an E-Learning platform based fusion of Web resources	306
<i>Mohammed Chaoui and Mohamed Tayeb Laskri</i>	
Complete and incomplete approaches for graph mining	312
<i>Amina Kemmar, Yahia Lebbah, Mohammed Ouali and Samir Loudni</i>	
Alignment between versions of the same ontology	318
<i>Ahmed Zahaf</i>	
Discovery of similar blocks from very large-scale ontologies	324
<i>Boubekeur Aicha and Abdellah Chouarfia</i>	
From UML class diagrams to OWL ontologies: A Graph transformation based Approach.	330
<i>Belghiat Aissam and Bourahla Mustapha</i>	
Automatic composition of semantic Web services-based alignment of OWL-S. .	336
<i>Boukhadra Adel, Benachtba Karima and Balla Amar</i>	

Keynotes



Towards Automated Information Factories

Aris M.Ouksef

University of Illinois at Chicago
aris@uic.edu

Abstract.

There has been a growing trend toward the automated generation of massive data at multiple distributed locations, leading to a future of computing that is data-rich, heterogeneous, distributed, and rife with uncertainty. Examples include systems to monitor the physical world, such as wireless sensor networks, and systems to monitor complex infrastructures, such as distributed Internet monitors. This trend will likely continue. Most information available today on the Internet is fabricated by human data entry. While such this type of information will continue to be produced, it will be only a small fraction of the volume of information generated by automated factories. This trend raises a number of key questions: How to fuse, process, reason with and analyze this tremendous amount of automated data streams? How to integrate raw information with high-level information available in traditional media and reason about uncertainty? How to recognize emergent communities of users in this new scenario? How to reason about security in an uncertain data environment?

Our talk will focus on information-generating factories in networks of fixed and mobile heterogeneous smart sensing devices. Our goal in this area is to develop a unified model, which captures the characteristics of both the new information generating factories and the traditional information available in the cyberspace, including distribution, heterogeneity, self-emergence, dynamic resource management, reaction to complex chains of events, continuous evolution, context-awareness, and uncertainty.

Data Quality – Not Your Typical Database Problem

Mourad Ouzzani

Qatar Computing Research Institute
mouzzani@qf.org.qa

Abstract.

Textbook database examples are often wrong and simplistic. Unfortunately Data is never born clean or pure. Errors, missing values, repeated entries, inconsistent instances and unsatisfied business rules are the norm rather than the exception. Data cleaning (also known as data cleansing, record linkage and many other terminologies) is growing as a major application requirement and an interdisciplinary research area.

In this talk, we will start by discussing some of the major issues and challenges facing creating effective and efficient data cleaning solutions. Then we will discuss some challenges and criticize current conservative approaches to this very critical problem. Finally we will discuss some of our work at QCRI in this area.

Internet and Web Technologies I



Context driven mediation service in Data-as-a-Service composition

Idir Amine Amarouche¹ and Djamel Benslimane²

¹ Université des Sciences et de la Technologie Houari Boumediene
BP 32 El Alia 16111 Bab Ezzouar, Algiers, Algeria

² Université Lyon 1, LIRIS UMR5205
43, bd du 11 novembre 1918, Villeurbanne, F-69622, France
i.a.amarouche@gmail.com, Djamel.Benslimane@liris.cnrs.fr

Abstract. Data as a Service (DaaS) builds on service-oriented technologies to enable fast access to data resources on the Web. Many approaches are proposed to achieve the DaaS composition task which is reduced to query rewriting problem. In this context, DaaS is described as Parametrized-RDF View (*PRV*) over Domain Ontology (*DO*). However, the *DO* is unable to capture the different perspectives or viewpoints for the same domain knowledge. This limitation raises semantic conflicts between pieces of data exchanged during the DaaS composition process. To face this issue, we present a context-driven approach that aims at supporting semantic mediation between composed DaaS. The semantic reconciliation based on mediation service is performed through the execution of rule mapping which achieves the transformation between contexts.

Keywords: DaaS composition, mediation service, context, semantic conflict.

1 Introduction

Nowadays, modern enterprises are using Web services for data sharing within and across the enterprise's boundaries. This type of Web service is known as Data-as-a-Services (DaaS) which return collections of Data for a given set of parameters without any side effects. DaaS composition is a powerful means to answer users' complex queries. Semantic-based approaches are proposed to enable automatic composition by describing the Web services properties over ontology. In fact, many ontology languages (e.g., OWL-S³, WSMO⁴) and extension mechanisms (e.g., WSDL-S⁵) provide standard means by which WSDL⁶ document can be related to semantic description. However, this means do not provide a way to relate semantically the Web service parameters (i.e., input and

³ <http://www.w3.org/Submission/OWL-S/>

⁴ <http://www.wsmo.org/TR/d2/v2.0>

⁵ <http://www.w3.org/Submission/2005/SUBM-WSDL-S-20051107/>

⁶ Web Service Description Language

output) which hampers their applicability to DaaS composition. The automation of DaaS composition requires the specification of the semantic relationships between inputs and outputs parameters in a declarative way. This requirement can be achieved by describing DaaS as views over a *DO* following the mediator-based approach [8]. Thereby, the DaaS composition problem is reduced to a query rewriting problem in the data integration field. In this context, several works [2, 9, 7] consider DaaS as Parametrized RDF⁷ Views (*PRVs*) with binding patterns over a *DO*, to describe how the input parameters of the DaaS relate to the data it provides. Defined views are then used to annotate DaaSs description files (e.g., WSDL files) and are exploited to automatically compose DaaSs. However, there are several reference ontologies which formalize the same domain knowledge. Thus, the construction of a *DO* unifying all existing representations of real-world entities in the domain is a strong limitation to interoperability between DaaS, this essentially raises semantic conflicts between pieces of data exchanged during DaaS composition. To this end, the applicability of previously cited DaaS composition approaches is not practical. Therefore, considering the semantic conflict detection and resolution during the composition process is crucial as service providers' contexts are practically different. In this regard, the approaches discussed in [4] and [5], have used the context representation for semantic mediation in Web service composition. In fact, they propose an extension of *DO* by a lightweight ontology which needs a small set of generic concepts to capture the context. However, these representations assure only simple mapping between semantically equivalent context parameter (price, unit, etc.). Further, the technical transformation code assuring the conversion from one context to another makes harder the maintainability of the semantic mediation between service composition components.

Motivating example: Let us consider an e-health system where the information needs of health actors are satisfied with DaaS Composition System (DCS), as proposed by [2, 9], which exports a set of DaaSs to query patient data. We assume that a physician submits the following query Q_1 : "What are the states indicated by the recent Blood Pressure Readings (*BPR*) for a given patient". We assume that the DCS will automatically generate DaaS composition, as response to physician query, including respectively S_1 , S_2 and S_3 as depicted in figure 1.(a). The DCS invokes automatically in the following order: 1) " S_1 " that provides the recent Vital Sign Exam (*BPR*, etc.) performed on his patient; 2) " S_2 " to retrieve the *BPR* measure⁸; 3) " S_3 " to retrieve the "BPR" state from the *BPR* value returned by S_2 . However, the DCS exports DaaSs expressed over *DO* does not take into account the context. By the context we mean the knowledge allowing to compare DaaS parameters values when there is a conflict (i.e, measurement unit, codification system, classification system, *BPR* value structure, etc.). Indeed, the physician has to manually detect the existing conflict in generated DaaS composition. For that, he has to select and to invoke

⁷ RDF: Resource Description Framework

⁸ *BPR* is represented by two concatenated values. eg., 120/80 where 120 is *BPR* Diastolic (*BPR.D*) value and 80 *BPR* Systolic (*BPR.S*) value

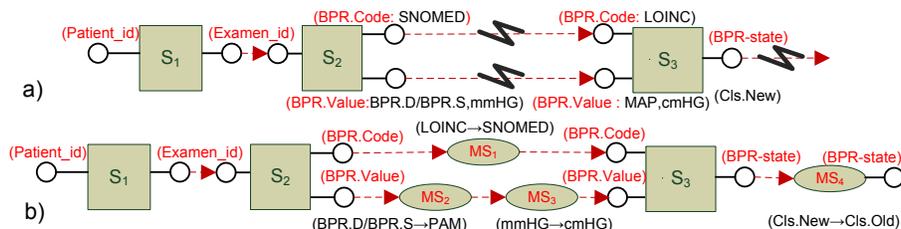


Fig. 1. Physician query scenario: a) DaaS composition generated by the DCS; b) The DaaS composition with the appropriate mediation services.

the appropriate mediation services, in the right order, to make the generated composition executable as depicted in figure 1.(b). The physician has to invoke: 1) “ MS_1 ”: to map the BPR code returned by S_2 (LOINC⁹) to code acceptable by S_3 (SNOMED¹⁰); 2) The composition of “ MS_2 ” and “ MS_3 ” where : “ MS_2 ” aggregates the two values expressing BPR measure returned by “ S_2 ” to MAP ¹¹ value acceptable by S_3 and “ MS_3 ” converts the MAP value expressed with the measurement unit (“mm/Hg”) returned by MS_2 to the MAP value expressed with the measurement unit acceptable by S_3 (“cm/Hg”); “ MS_4 ” : to map the BPR state returned by “ S_3 ” represented according to the new classification BPR value table (e.g., stage 1,2,3,4) to the state acceptable by the physician represented according to the old classification (e.g., severe, moderate, mild). This is a rather demanding task for non expert users (e.g.physicians). Thus, automating conflict detection and resolution in DaaS composition is challenging.

Contributions: In this paper we propose a context driven approach for automatically inserting appropriate mediation services in DaaS compositions to carry out data conversion between interconnected DaaS. Specifically, we propose 1) a context model expressed over Conflicting Aspect Ontology(CAO) which is an extension of “ DO ”; 2) an extension of PRV based DaaS model based on context to express more accurately the DaaS parameters semantic; 3) a mediation service model behaving as a mapping rule to perform the transformation of DaaS parameters from one context to another.

Outline: The rest of this paper is organized as follows. Section 2, presents the overview of our approach. In Section 3, we leverage different proposed models. In Section 4, we present a global view on our conflict detection and resolution algorithm and our implementation. Finally, section 5 provides a conclusion and future works.

⁹ LOINC : Logical Observation Identifiers Names and Codes

¹⁰ SNOMED: Systematized Nomenclature of Medicine, Clinical Terms

¹¹ Mean Arterial Pressure is BPR value, $MAP = \frac{2}{3}(BPR.D) + \frac{1}{3}(BPR.S)$

2 Approach overview

Figure 2 gives an overview of our approach. Our proposal aims to provide a framework for automatic conflict detection and resolution in DaaS composition. Our approach takes into account the context of the service components in DaaS composition and the context of the query. DaaS services are modeled as *PRV* over a *DO* and contextualized over Conflicting Aspect Ontology (*CAO*). The mediation services are modeled as mapping rule over *CAO* specifying the DaaS parameters transformation from one context to another. The contextualized *PRV* and the mapping rule are incorporated within correspondent WSDL description files as annotation. The DaaS service registry includes business services while the mediation service are organized in other registry to keep the mediation concerns orthogonal from functionalities of DaaS.

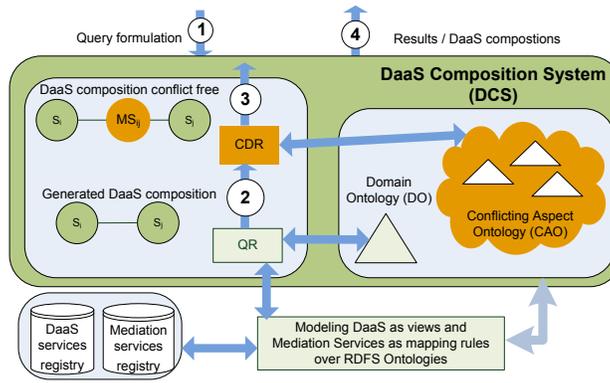


Fig. 2. Approach overview

The DaaS composition process starts when the user specifies a query over *DO* and *CAO* using SPARQL¹² query language (see circle 1 in figure 2). The DCS uses the query rewriting algorithm proposed by [2] and existing *PRV* to select the DaaS that can be combined, to answer the query (see circle 2 in figure 2). After that, our Conflict Detection and Resolution Algorithm (CDR) takes place for conflict verification in each generated DaaS composition. Then, in case where a conflict is detected between output/input operation (i.e., subsequent services in DaaS compositions, query and DaaS compositions) our algorithm insert automatically calls to appropriate mediation services to resolve semantic conflict (see circle 3 in figure 2). Then, the DCS translates a composite DaaS conflict free into query execution plan describing data and control flow. The plan will be executed and returns data to the user (see circle 4 in figure2). In this paper, we will focus only on Conflict Detection and Resolution process.

¹² We adopt SPARQL: <http://www.w3.org/TR/rdf-sparql-query/>, the de facto query language for the Semantic Web, for posing queries.

3 Modeling issues

We leverage in this section different models used through the paper. The definition of the basic concepts such as the Domain Ontology (DO), the Parametrized RDF view (PRV) and the Conjunctive Query (CQ) are presented formally in [1]. Due to space limitations, we will not present their corresponding figures. In the sens of the present work, the DaaS Composition $cs = \{s_1..s_n\}$ represents the set of ordered services into DaaS composition ; $First(cs)$ (e.g, s_1) and $Last(cs)$ (e.g, s_n) denote the first and the last DaaS in “ cs ”. We mean by the “ CSs ” the set of compositions, generated by the query rewriting algorithm of “DCS”, requiring testing and resolution of conflicts.

3.1 Conflicting Aspect Ontology:

Conflicting Aspect Ontology (CAO) is a family of a lightweight ontology, specified in RDFS. CAO extends the DO entities with a taxonomic structure expressing different DaaS parameters semantic conflict¹³. The CAO is a 3 tuple $\langle AC_g, AC_i, \tau \rangle$, where: 1) “ AC_g ” is a set of classes which represents the different conflicting aspects of a DO entities. Each ac_g class in AC_g has one super-class and a set of sub-classes. Each ac_g class has a name representing a conflicting aspect, such as, CAO:Measurement-Unit as depicted in Figure 3; 2) “ AC_i ” is a distinct set of instanceable classes having one super-class in AC_g . By definition, ac_i is not allowed to have sub-classes. For instance “ mm/HG ” and “ cm/HG ” are two instanceable classes from the CAO:BPR-Unit class; 3) “ τ ” refers to the sibling relationships on AC_i and AC_g . The relationships among elements of AC_g is disjoint. However, elements of AC_i of a given ac_g can be related by the *Peer relationship* which indicates similar data semantics. *Part-Of relationship* which relates ac_i entity and its components (e.g., BPR.D and BPR.S values are Part-Of BPR).

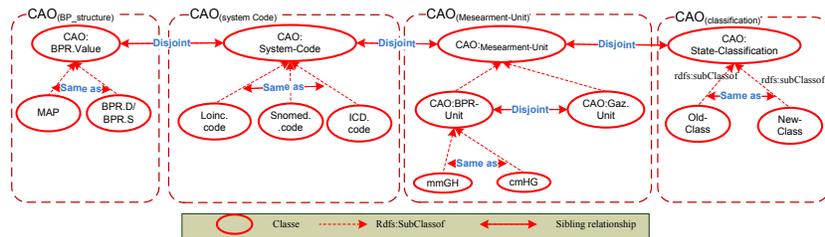


Fig. 3. Conflicting Aspect Ontology

¹³ For the classification of the various incompatibility problems in web service composition see [6]

3.2 Context model:

The context has the form: $C = \{(D_i, V_i) | i \in [1, n]\}$ where D_i , represents an ac_g class whose values are from a value-set (V_i) where $V_i \in AC_i$. For instance, the context $C_{MU} = \{BPR - Unit : mm/HG\}$ indicates that the BPR measurement unit is “mm/HG”. The proposed context model is used to express more precisely the query formulated by the user, the DaaS published by the provider and the semantic conflict occurring in each O/I ¹⁴ operation in given $cs_K \in CS$. **1) “Contextualized Conjunctive Query model”** is $CCQ(X) : - \langle CQ(X) | CCQ_{(X,CO)} \rangle$ where $CQ(X)$ is the conjunctive query expressed over DO , and $CCQ_{(X,CO)}$ is the context of the distinguished variable X and the query constraint CO expressed over CAO ; **2) “Contextualized DaaS model”**: The C-DaaS is $S_j(\$X_j, ?Y_j) : - \langle V_{DO} \rangle | \langle Ext_{CAO} \rangle$ where V_{DO} is the PRV of S_j and Ext_{CAO} is a tuple $\langle C_{X_j}, C_{Y_j} \rangle$ where C_{X_j} and C_{Y_j} are respectively the input and the output DaaS parameter contexts. C_{X_j} and C_{Y_j} are described by a set of RDF triples over CAO in form of 2-tuple $\langle AC_g, AC_i \rangle$; **3) “Context and semantic conflict”**: In the sense of the present work, semantic conflict occurs in O_n/I_m operation having respectively O_n and I_m as an output and an input parameter which refer to the same DO entity. However, their contexts represented respectively by C_{O_n} and C_{I_m} refer to different “ ac_i ” entities from the same “ ac_g ”. Then we say that a parameter semantic conflict “ ac_g ” exists in O_n/I_m .

3.3 Mediation service model

Mediation Services MS assures the semantic reconciliation in the case where the O/I operation causes a conflict. The MS model consists of mapping rule having the form $MS(\$O_J, ?I_J) : G_O \rightarrow G_I$, where $\$O_J$ and $?I_J$ are the sets of input and output variables of MS_j respectively. G_O and G_I represent the set of RDF triples representing contextualized DaaS /query parameters. We deem appropriate to use the SPARQL’s construct statement (i.e., CONSTRUCT G_I WHERE G_O) as a rules language to define rule mapping as proposed by [3]. For

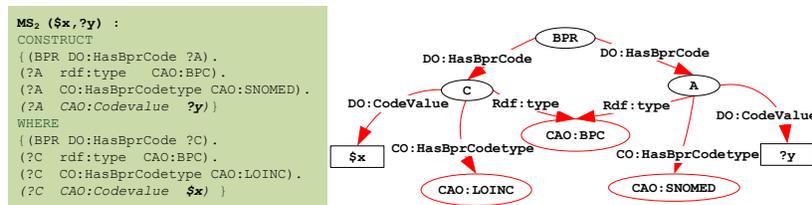


Fig. 4. Mediation service model

¹⁴ i.e., two subsequent DaaS “ S_n ” and “ S_m ” in “ cs ”, $First(cs)$ and $CCQ_{(CO)}$, $Last(cs)$ and $CCQ_{(X)}$.

each conflicting aspect AC_g we define a mapping rule template. For instance, the mediation service MS_2 assuring the same-as mapping one-to-one of BP code value from “LOINC” code to “SNOMED” code is presented in figure 4. In the same manner, we define the mapping many-to-one, one-to-many and many to many. To the best of our knowledge, this work is the first to use SPARQL construct statement to model mediation services.

4 Algorithm and implementation

In the following, we present the details of our Conflict Detection and Resolution Algorithm (CDR) depicted in figure 5. The inputs to the CDR is a set of

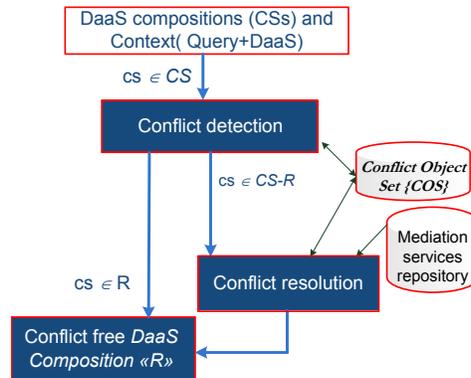


Fig. 5. CDRM architecture

“ CSs ” generated by the QR algorithm as explained in section 2. The outputs of CDR are “ CSs ” conflict free. The desired mediation service is found and called automatically using the CDR algorithm which is two phases : Detection and Resolution. In the first phase each composition “ cs ” is examined to detect potential conflicts. Thus, if “ cs ” is without conflicts then it is inserted into the set of compositions without conflicts R ; else the conflicts of “ cs ” are added into the conflict object set “ COS ”. Finally, the set of composition without conflict R is removed from CS . Thus CS consists of the composition with conflicts. In the second phase, each detected conflict is resolved by performing the matching between the required context transformation to the mapping rules defining the mediation services. The matching is obtained, the automatic calls to the correspondent mediation services are inserted in “ cs ” to resolve conflicts. Then, the new set of composition CS (i.e, composition without conflict) are added into R and returned to DCS for query plan execution. In order to test test our proposal, we have implemented a Java Based application and test it with multiple examples, including the motivating example ¹⁵. Each Web services is deployed

¹⁵ The implementation test are available in <http://sites.google.com/site/ehrdaas/home>

on top of a GlassFish web server. Each DaaS is annotated by the contextualize *PRV* and each Mediation service is annotated by SPARQL construct statement. In the evaluation phase we have considered a set of queries through which we identify the following : 1) During the detection phase, we can detect the set of conflict aspect identified in “*AC_g*”. 2) During the resolution phase, according to the number of conflict detected in each *O/I* operation: when there is a conflict including one aspect *ac_g* (e.g., BPR-code) or a conflict including several aspects *ac_g* (e.g., BPR-value), our solution provides automatically the appropriate mediation service. When we have a several mediation services allowing to resolve the same conflict, our algorithm returns randomly one of them as long as they achieve the same functionality.

5 Conclusion and future work

In this paper, we propose an extension of PRV based DaaS model based on context. The proposed context model expressed over Conflicting Aspect Ontology aims to handle semantic conflict in DaaS composition. Our model allows to specify the mediation service as mapping rule performing the simple or complex transformation of DaaS parameters from one context to another. Our future perspective will to deal with the performance issues of our algorithm and how to resolve a given conflict for which there is no appropriate mediation service.

References

1. Amarouche, I.A, Benslimane, D., Barhamgi, M., Mrissa, M., Alimazighi, Z. : Electronic Health Record DaaS services Composition based on Query Rewriting. Transactions on Large-Scale Data and Knowledge-Centered Systems. 4, 95–123 (2011)
2. Barhamgi, M., Benslimane, D., Medjahed, B. : A Query Rewriting Approach for Web Service Composition. IEEE Transactions Services Computing. 3, 206–222 (2010)
3. Euzenat, J., Polleres, A., Scharffe, F. : Processing Ontology Alignments with SPARQL. International Conference on Complex, Intelligent and Software Intensive Systems. 913–917 (2008)
4. Li, X., Madnick, S., Zhu, H., Fan, Y. : Reconciling Semantic Heterogeneity in Web Services Composition. ICIS 2009 Proceedings. 20 (2009)
5. Mrissa, M., Ghedira, C., Benslimane, D., Maamar, Z. : A Context Model for Semantic Mediation in Web Services Composition. ER. 12–25 (2006)
6. Nagarajan, M., Verma, K., Sheth, A.P., Miller, J.A. : Ontology Driven Data Mediation in Web Services. Int. J. Web Service Res. 104–126 (2007)
7. Vaculín, R., Chen, H., Neruda, R., Sycara, K. : Modeling and Discovery of Data Providing Services. ICWS. 54–61 (2008)
8. Wiederhold, G. : Mediators in the Architecture of Future Information Systems. Computer. 25, 38–49 (1992)
9. Zhou, L., Chen, H., Wang, H., Zhang, Y. : Semantic Web-Based Data Service Discovery and Composition. SKG. 213–219 (2008)

Service Substitution Analysis in Protocols Evolution Context

Ali Khebizi¹, Hassina Seridi-Bouchelaghem², Imed Chemakhi³, and Hychem Bekakria³

¹ LabStic Laboratory, 08 May 45 University, Guelma -Algeria-
ali.khebizi@gmail.com

² LABGED Laboratory, University Badji Mokhtar Annaba, Po-Box 12, 23000,
Algeria seridi@labged.net

³ Computer science Institute, 08 May 45 University, Guelma -Algeria-
Chemakhi.imed@gmail.com,hychem.bekakria@gmail.com

Abstract. As Web services become the dominant technology for integrating distributed information systems, enterprises are more interested by these environments. However, enterprises socio-economic environments are more and more subject to changes which impact directly business processes published as Web services. In parallel, if at change time some instances are running, the business process evolution will impact the equivalence and substitution classes of the actual service. In this paper, we present an equivalence and substitution analysis in dynamic evolution context. We suggest an approach to identify residual services that can substitute a modified one, where ongoing instances are pending. Our analysis is based on protocol schema matching and on real execution traces. The proposed approach has been implemented in a software tool which provides some useful functionalities for protocol managers.

Keywords: Service protocol, Protocol equivalence, Protocol substitution, Dynamic evolution, Execution path, Execution trace.

1 Introduction

Web services are the new generation of distributed software components. They generate a lot of enthusiasm among different socio-economic operators's which favourite these environments to deploy applications at a large scale. Standardization is a key concept, so actors uses standards like WSDL [1], UDDI [2] and SOAP [3] to publish, discover, invoke and compose distributed software. In this context, intra and inter enterprises applications integration is more flexible, easy and transparent. Moreover, integration process is accelerated among internet stakeholders.

In Web services technology, two elements are fundamental for providing a high interactivity level between service providers and service requesters. The first one is service interface, described via the standard WSDL. The second element

is service protocol (Business Protocol), which describes the provider's business process logic. A Business process consists of a group of business activities undertaken by one or more organizations in pursuit of some particular goal [4],[5]. For example, booking flight tickets and B2B transactions. In addition, Business process specifies the service external behaviour by providing constraints on operations order, temporal constraints [6] and transactional constraints [7], in order to promote correct conversations, as service operations can't be invoked in an aleatory order. However, if a service protocol is published in the Web (its interface and its protocol), at a moment during its life cycle, it can be invoked by some clients. Furthermore, in large public applications (e-commerce, e-government, electronic library, . . .), thousand of clients are invoking the service at the same time and every one has reached a particular execution level. In parallel, as enterprises are open systems, changes are permanent and inevitable. Consequently, business processes may evolve to adapt to environment changes that affect real world. In this case, related service protocols must be updated, otherwise services execution may produce incoherences when they are invoked. This context is called **dynamic protocol evolution**.

In dynamic protocol evolution, the evolved service protocol may not be able to satisfy initial customer requirements. Furthermore, some services may fails and clients must find new services that can replace actual one. Services substitution analysis deal with checking if two services satisfy the same functionalities; if they support the same conversation messages [5]. This concept is very useful in case of service failure, in order to search an other one to replace it. In some cases, this analysis can serve to search and locate a new service with the same functionalities but with a higher quality of service (Qos). It can also be used to test whether a new proposed version, that expresses evolution or maintenance requirements, is yet, equivalent to an obsolete one and for finding new services that can support conversations required by standards like ebXML [8], Rosetnet [9] and xCBL [10]. Service evolution is expressed through the creation and de-commission of its different versions during its lifetime. [11]

Service protocol update induces challenges for filtering which services, already identified and compared to old version, remain equivalents or can replace the evolved one. The major constraint is related to active instances that have already executed some operations based on old version. In this context, we must deal with historical executions in substitution analysis process.

In this paper we are interested to dynamic evolution and we focus on change impact analysis on service protocol substitution. A set of methods are exposed to check if new service version, can be yet substituted by the hole (or partially) set of services that were discovered corresponding to the obsolete version.

The remainder of the paper is structured as follow. We start by describing the problem and exposing our motivations, in section 2. In section 3, we propose our formal approach and algorithms for managing substitution aspects in dynamic evolution context. Section 4, describes system architecture and software tool implementation. We expose related works in section 5 and conclude with a summary and directions for future works in section 6.

2 Problem and Motivations

Every organisation (enterprises, administrations, banks, ...etc.) is an open system which is, eventually, impacted by environment changes. In order to survive, organization must adapt there business processes. Today's organizations information systems -reflecting business processes- are exposed on the Web as services (interfaces and protocols) and every business process changes induce, immediately, these two descriptions update. The challenge in dynamic protocol evolution context is to identify, among the set of already identified class substitution services, the subset of those that can, yet, replace an actual service after its specification changes, with respect to past interactions. Addressing service protocols substitution analysis, after protocol evolution, responds to the following motivations:

1. Ensuring service execution continuation for active instances.
2. Ensuring correct interactions between customers and providers by specifying the new service substitution class.
3. In dynamic environments, like Web services, transactions are long duration and resources consumer. It's not conceivable to restart execution from scratch because loss of work is catastrophic for customers.
4. In real time systems and critical applications (aeronautics, e-commerce, medical systems, control systems, manufacture, . . .), brutal service stop is catastrophic for organizations. It is imperative to treat with precision and accuracy services that can substitute an evolved or failed one.
5. In large public applications (e-libraries, e-government, e-learning. . .), a large number of active instances are pending, at a time. Manual management of these instances is cumbersome and an automatic support is required to ensure that only pertinent services are proposed for substitution process.

The main issue is to manage protocol substitution with respect to historical traces. Starting a new search query for locating new services, based on the new version, is expensive and in addition, returned services that can be inconsistent with old version.

To illustrate our motivations, we present in **Fig.1** a real world scenario.

In this scenario, service protocol P have some equivalent services (P_1, P_2, \dots, P_n) and other services (P_3, P_4, \dots, P_m) can substitute it. However, service P has evolved to a new version P' (for different reasons). Evolution operations added a new message *cancelOrder* and removed the message *Order validated*. At evolution time, active instances (instance 1, instance 2, . . .), are running and have reached a particular execution level. In order to be able to substitute service P in case of problems, protocol manager want to know: **Which protocols remain in conformance with the new protocol specification and can replace it ?**

3 Analysing Substitution in Dynamic Protocol Evolution

One of the most challenging issues in dynamic protocol evolution context is to find potential protocols for substitution, where instances are running accord-

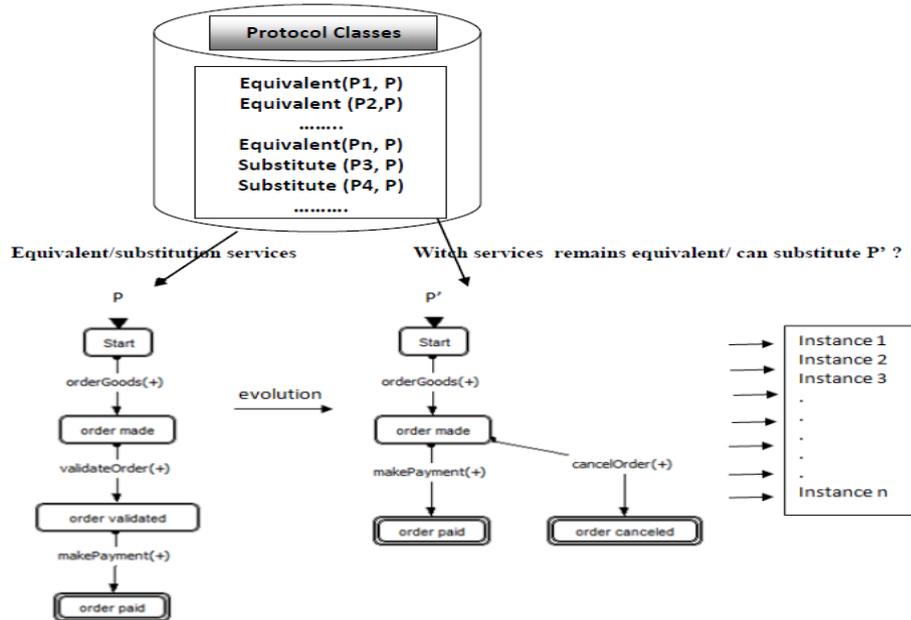


Fig. 1. After Protocol evolution of P to P' which services can substitute P' ?

ing to old protocol. To address this analysis, we introduce three fundamental concepts: **service protocol model**, **execution path** and **execution trace**.

- **A service protocol:** We use finite state machine to represent service protocols. In this model states represent different phases that a service may go through during its interaction with a requester. Transitions are triggered by messages sent by the requester to the provider or vice versa [4], [5]. A message corresponds to operation invocation or to its reply, as shown in **Fig 1**. A finite state machine is described by the tuple: $P = (S, s_0, F, M, R)$, consisting of:
 - S : A finite set of states;
 - $s_0 \in S$: is the protocol initial state;
 - F : Set of final states machine, with $F \subset S$;
 - M : Finite set of messages;
 - $R \subset (S \times S \times M)$: Transitions set. Each one involves a source state to a target state following the message receipt.
- **Execution trace:** Service behaviour traces is a finite sequence of operations (a, b, c, d, e, \dots) . It represents events that service have invoked, from its beginning to the actual state. We note : $trace(P, i)$ to express the execution trace performed by an active instance i in a protocol P .
- **Complete Execution path:** Represents the sequence of states and messages from an initial state to a final one. We note : $expath(P)$.

3.1 Structural approach based protocol schema

Let P and P' respectively, old and new service versions, after operating changes. $E_P = \{P_i (i = 1 \dots n)\}$: Is the services set **equivalent** to P . We note $Equi(P_i, P)$, the equivalence relationship between services.

Two service are equivalent if they can be used interchangeably and they provide the same functionalities in every context. Every service $P_i \in E_P$ can replace P . $Equi(P_i, P) \Leftrightarrow \forall (i = 1 \dots n)(expath(P_i) \subset expath(P)) \wedge (expath(P) \subset expath(P_i))$. (\wedge is the logic **and** operator). (1)

Let $S_P = \{P_j (j = 1 \dots m)\}$: The services set that can substitute P . We note $Subst(P_j, P)$, the substitution relationship.

A service can substitute an other one if it provides, at least, all the conversations that P supports [5] (complete execution paths).

$Subst(P_i, P) \Leftrightarrow \forall (i = 1 \dots m)(expath(P) \subset expath(P_i))$ (2).

Based on this formalization we notice that if protocol P has evolved to a new version P' , equations (1) does not remain valid. So we want to identify the protocols subset that satisfy equation (2), in order to provide services that can replace the evolved protocol.

From equation (2): $Subst(P_i, P') \Leftrightarrow \forall (i = 1 \dots m) (expath(P') \subset expath(P_i))$. (3) . We conclude:

Lemma 1: If the changes related to protocol evolution are reductionist, all protocols (P_i) that can substitute P , can substitute the new reduced version P' . Reducing Protocol description is an application of change operations including:

- Loops removal.
- Final sub-paths removal.
- Operations and messages removal.
- Complete paths removal and sub-protocols removal.

This change operation goals are motivated by procedures cancellation, reducing tasks, business processes alignment, and so one. However, when changes are additive, substitution analysis must consider the protocol difference. Protocol difference between two protocols P' and P describe the set of all complete execution paths of P' that are not common with P [5]. We note P'/P this difference. Substitution analysis consists to examine each protocol in the class S_P , with the aim to identify possible protocols that can substitute P' . Because equation (1) no longer holds, we must comply with equation (2).

In order to replace P' , each protocol $P_i \in S_P$ must be able to replace the new requirements (the difference P'/P).

$Subst(P_i, P') \Rightarrow Subst(P_i, P'/P) \Leftrightarrow \forall (i = 1 \dots m) (expath(P'/P) \subset expath(P_i))$ (4). We conclude:

Lemma 2: If changes are additive, protocols subset $\subset S_P$ which are containing the difference P'/P can substitute the new extended version P' . Additive changes are operations performing:

- Adding loops.
- Adding sub-paths
- Adding messages and operations .
- Adding new complete paths and sub-protocols.

3.2 Execution traces based analysis

Protocol schema based analysis is rigid and does not take into account the actual execution for active instances. Really, it's possible that a protocol $P_i \in S_P$ can't substitute an evolved one in general, but by taking into account execution traces, it can do that for specific instances. As an example, let a protocol P and its active instances i_1, i_2, \dots, i_n , as mentioned in **Fig.1**. In parallel, protocol changes have added new states and messages to a particular path: *part-path*. After analysing active instances execution traces, we see that all instances have't executed this unexpected path *part-path*. In this case, even if we can't replace P' with a protocol $P_i \in S_P$, basing on protocol schema analysis, we can substitute it basing on real execution traces, because changes do not impact real instances. We notice that execution traces may inform protocol managers on how to proceed with substitution analysis. We propose two substitution analysis based execution traces: **Historical execution paths and state execution paths**.

Historical execution paths substitution analysis:

Let *histpath* a protocol p historical execution path executed by an active instance i , during its execution, instance i has invoked an operations sequence : a, b, c, d, e, \dots . And, let *futurpath*: future paths not yet executed by this instance. If P' is the new version of P , after changes and S_P is the protocol set that can substitute P . We are interested by filtering instances that are not concerned with changes. We consider protocol changes as the difference between P' and $P : P'/P$. In this situation, if protocol $P_i \in S_P$ can't substitute P' , contrarily, it can substitute it for the instances subset that have not executed this difference. We note: $Subst(P_i, P')/Occur_j$: The substitution of P' by P_i for occurrence j . $Subst(P_i, P')/Occur_j (i = 1 \dots n), j = 1 \dots m)$ is possible if : $(histpath(occur_j) \notin allpaths(P'/P))$. **(5)**, where $Allpaths(P'/P)$ is the hole possible paths set generated by protocol difference P'/P . This means that, substitution is possible if actual instance i had executed an old path not affected by changes.

State Based Substitution Analysis:

Historical execution path analysis is more general and based on the hole historical execution paths. Although, protocols $P' \in S_P$ can't replace P in the general case, substitution is possible for some states. We need to compute which states are not affected by changes. Substitution analysis must deal with this kind of traces by selecting protocol services that substitute active service by considering actual state and future execution path.

As an example, consider the execution path from **Fig. 1**: If a subset of actual instances are in the state: *Order made*, so their execution trace are : *begin.order made*. A service $P_i \in S_P$ can substitute P' if it can replace it from the actual state and future execution paths. We don't consider past execution paths because changes occurs after the state (*Order made*). We formalize this analysis as follows:

Let *futurpaths* the future execution path set of an active instance (all future paths), and *s* the actual instance *i* state.

$Subst(P_i, P')/state(s) : (i = 1 \dots n)$ if :
 $(futurpaths(state(s)) \subset allpaths(P'/P)) \wedge (trace(P, i) \in histpath(P_i))$.(6)

3.3 Algorithms

We present here **Substitution-schema-based** algorithm to calculate schema protocol based substitution analysis presented in section 3.1.

Algorithm 1: Substitution-schema-based

Input: $P = (S, s_0, F, M, R), P' = (S', s'_0, F', M', R'), P_i = (S_i, s_{0i}, F_i, M_i, R_i)$.

Output: Decision on Substitution.

Begin

1. *Substitution* := *True*
2. *Path* = ϕ , *Completpath* = ϕ , *Completpath'* = ϕ
3. *Completpath* := *RecursivePaths*(*S*, *s*₀, *F*, *M*, *R*)
4. *Completpath'* := *RecursivePaths*(*S'*, *s'*₀, *F'*, *M'*, *R'*)
5. For *i* = 1 to *n* // *n* is protocol number
6. While (*path* ∈ *Completpath*) Do
7. If (*path* ∉ *Completpath'*) Do
8. *Substitution* := *False*
9. break
10. EndIf
11. EndWhile
12. EndFor
13. Return (*Substitution*)
14. End **Substitutions-schema-based**.

RecursivePaths Algorithm, computes recursively all possible paths in a protocol definition, from an initial state to a final one. (is not exposed by lack of space).

4 System Architecture and Software tool presentation

We have implemented the software performing substitution analysis in Java-Eclipse environment. Service protocols are implemented as automata and saved in XML files. The software performs some operational functions useful for protocol manager like protocol description and correctness specification. Furthermore, it allows final users performing different changes operations. The system kernel provides checking static equivalence and substitution. **Fig. 3.** Based on schema definitions or on execution traces, the system strength is dynamic analysis. This analysis allows user to select a particular protocol, operate changes and then proceed to change impact analysis on protocol substitution. The system filters the protocol database, analysis logs directory and searches for the remaining service protocols which can substitute the evolved version **Fig. 4.** We visualize, below some screen-shots of the the software tool.

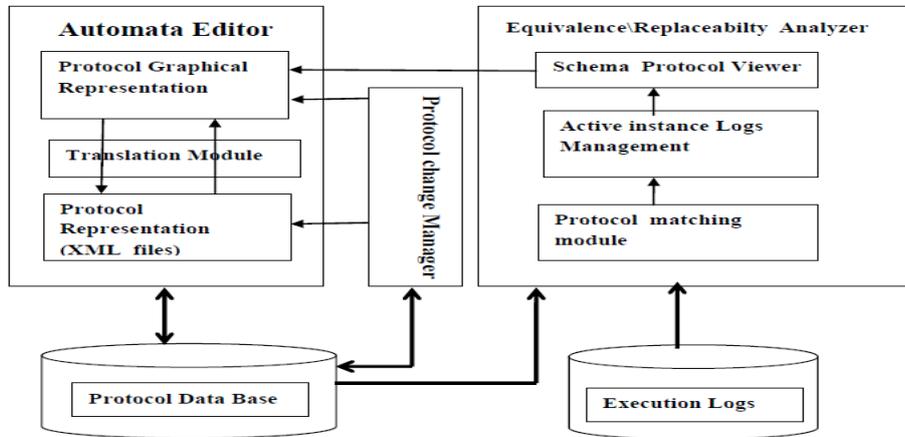


Fig. 2. System Architecture for managing dynamic substitution

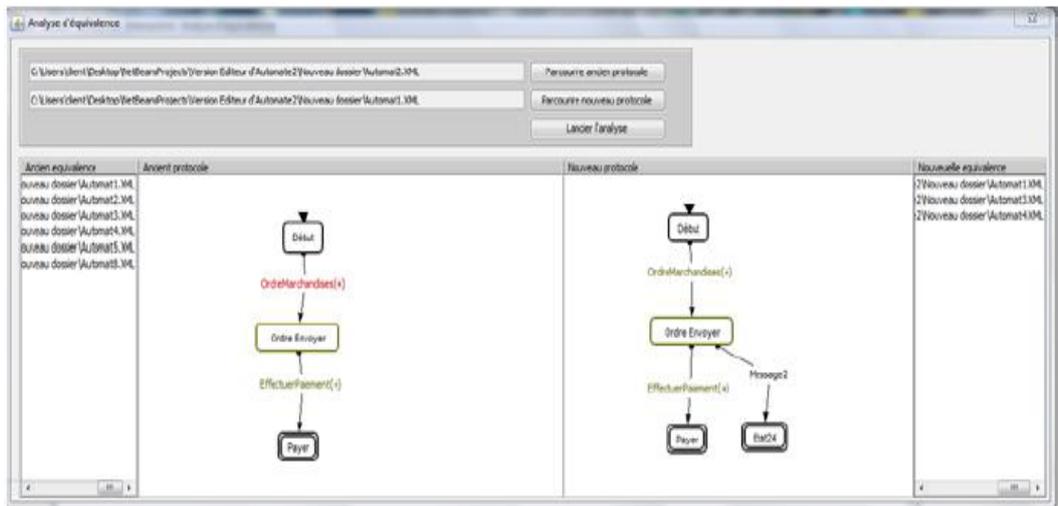


Fig. 3. Protocol specification, evolution, and static equivalence and substitution

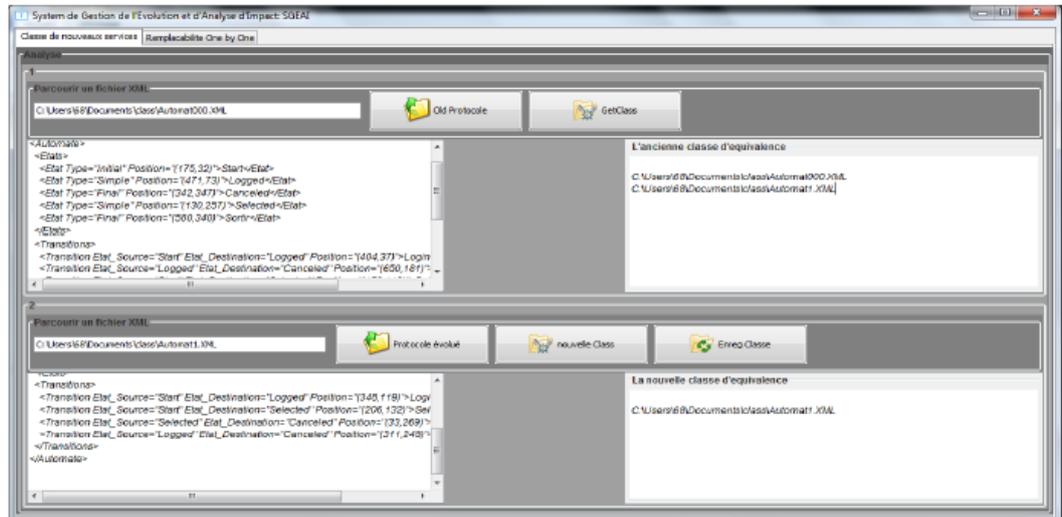


Fig. 4. Substitution analysis provide services substitution class for an evolved protocol

5 Related Work

Protocol management and analysis had benefited for a lot off contributions, from protocol schema matching to static evolution. But, dynamic protocol analysis did not receive all the interest it deserves. In [4],[5] authors presents a general framework for representing, analysing and managing Web service protocols, however this analysis is restricted to static context. In [6], protocol description is enriched with temporal constraints and the same static analysis was performed. In [12], authors had proposed some change operators and patterns specification for changes, but change impact analysis was not presented. In [13], dynamic replaceability analysis had been presented in therns of compatibility only. Authors studies the compatibility between old and new protocol version only. No comparison with other services was made. Our work responds in a consistent manner to the previous deficiencies.

6 Conclusion and future Work

In this article we have formalized substitution problem inherent to dynamic protocol evolution. We have proposed an approach and a software tool for providing service protocols that can, yet replace an evolved one.

As future work, we plan to address protocol substitution analysis for richer protocols descriptions, such as timed and transactional constraints automata. In addition, we aim to specify protocol changes more formally by identifying evolution patterns and by their classification with respect to induced impact on protocol substitution and compatibility.

References

1. R. Chinnici and al. Web Services description Language (WSDL) version 2.0 June 2007. <http://www.w3.org/TR/wsdl20/>
2. T. Bellwood and al. UDDI Version 3.0.2 UDDI Spec Technical Committee Draft, 2004. <http://uddi.org/pubs/uddi-v3.htm/>
3. M. Gudgin and al. SOAP version 1.2, July 2001. <http://www.w3.org/TR/2001/WD-soap12-20010709/>
4. B. Benatallah and al : Web Service Conversation Modeling A cornerstone for E-Business automation, IEEE Internet computing 8 (1) (2004) 46-545 WSC
5. B. Benatallah and al : Representing, Analysing and Managing Web Service Protocols. Data Knowledge Engineering. 58 (3): 327-357, 2006.
6. J. Ponge and al: Fine-Grained Compatibility and Replaceability Analysis of Timed Web Service Protocols. ER 2007: 599-614
7. A. Khebizi: External Behavior Modeling Enrichment of Web Services by Transactional Constraints, ICSOC PhD Symposium, December 2008. <http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-421/paper12.pdf>
8. Technical Architecture Specification v1.0.4 February 2001 <http://ebxml.org/specs/ebTA.pdf>.
9. Rosetanet; <http://www.rosettanel.org/>.
10. xCBL; <http://www.xcbl.org/>.
11. Gustavo Alonso, Fabio Casati, Hurumi Kuno, Vijay Machiraju : Web services concepts Architectures and applications, Edition Springer Verlag Berlin 2004.
12. Barbara Weber and al : Change Patterns and Change Support Features - Enhancing Flexibility in Process-Aware Information Systems , 2008
13. Ryu, S. H. and al, 2008. Supporting the dynamic evolution of Web service protocols in service-oriented architectures. ACM Trans. Web 2, 2, Article 13, 46 pages. DOI = 10.1145/1346237.1346241 <http://doi.acm.org/10.1145/1346237.1346241>.

Dynamic Web Service Composition: Use of Case Based Reasoning and AI Planning

Fouad HENNI

Baghdad ATMANI

Mostaganem University – Algeria

fouad.henni@gmail.com

Equipe de recherche Simulation, Intégration et Fouille de données (SIF)

Laboratoire d'Informatique d'Oran (LIO)

Oran University – Algeria

atmani.baghdad@univ-oran.dz

Abstract: Web services have emerged as a major technology for deploying automated interactions between distributed and heterogeneous applications. The main advantage of web services composition is the possibility of creating value-added services by combining existing ones to achieve customized tasks. How to combine these services efficiently into an arrangement that is both functionally sound and architecturally realizable is a very challenging topic that has founded a significant research area within computer science. A great deal of recent web-related research has concentrated on dynamic web service composition. Most of proposed models for dynamic composition use semantic descriptions of web services through the construction of domain ontology. In this paper, we present our approach to dynamically produce composite services. It is based on the use of two AI techniques: Case-Based Reasoning and AI planning. Our motivating scenario concerns a national system for the monitoring of childhood immunization.

Keywords: semantic Web services, dynamic composition, OWL-S, CBR, AI planning, immunization system

1 Introduction

A Web service is a software component identified by a URL, whose public interfaces and bindings are defined and described using XML. Web services provide a standard means of interoperating between different software applications, running on a variety of platforms and/or frameworks [1]. This has led to the emergence of Web services as a standard mechanism for accessing information and software components programmatically [2].

Service composition refers to the technique of composing arbitrarily complex services from relatively simpler services available over the Internet. Composition of Web services enables businesses to interact with each other and facilitates seamless business-to-business or enterprise application integration. Applications are to be assembled from a set of appropriate Web services and no longer written manually [3]. For example, a composite Web service for an online order from a retailer Web site

could bring together a number of internal and external services such as credit checking, inventory status checking, inventory update, shipping, etc.

Web Service Composition is currently one of the most hyped and addressed issues in the Service Oriented Computing. Several models, techniques and languages have been proposed to achieve service composition.

The construction of a composite Web service can be made up in three main steps (not necessarily in this order): (a) Creation of the *process model* specifying control and data flow among the activities. (b) *Discovery, selection and binding* of concrete Web services to every activity in the process model. (c) *Execution* of the composite service by a coordinating entity (e.g. a process execution engine) [4].

In *static* composition the process model is created manually and the bindings of concrete Web services to the process activities are done at design time. *Semi-dynamic* composition strategies actively support the user with the creation of the process model and/or in the services selection and bindings. Finally, in *Dynamic composition* the creation of the process model and the services selection and bindings are made at runtime. In this paper, the focus will be done on dynamic composition of services.

The remainder of this paper is organized as follows: Section 2 presents the main ideas in dynamic composition of Web services and particularly the use of Case-Based Reasoning (CBR) and AI planning. Our proposal of using both CBR and AI planning is described in section 3, while section 4 presents a scenario as a direct application of our proposal. The paper is concluded by a discussion of the solution, some limitations, and future works.

2 Dynamic Web service composition

In dynamic composition, automated tools are used to analyze a user query, and select and assemble Web service interfaces so that their composition will solve the user demand. From a user perspective, the composite service will continue to be considered as a simple service, even though it is composed of several Web services.

In order to support greater automation of service selection and invocation, recognition is growing of the need for richer semantic specifications of Web services, so as to enable fuller, more flexible automation of service provision and use, support the construction of more powerful tools and methodologies, and promote the use of semantically well-founded reasoning about services [5]. As a result, Web services have semantic descriptions in addition to their traditional standard syntactic description (WSDL). This is referred to as semantic Web services.

Semantic Web services solve Web service problems semantically and address Web services descriptions as a whole [6]. Semantic markup languages such as OWL-S [5, 7], WSDL-S [8] and SAWSDL [9] describe Web service capabilities and contents in a computer-interpretable language and improve service discovery, invocation, composition, monitoring, and recovery quality.

Several methods and tools have been proposed for dynamic Web service composition [2, 3, 10, 11, 12]. The majority of researches conducted in dynamic composition have their origins in the realm of artificial intelligence [10].

It is not in the scope of this paper to present an exhaustive list of all methods and techniques proposed for dynamic composition. In this work, we are particularly interested in the use of CBR and AI planning in order to achieve a dynamic composition.

IA Planning is certainly the area that offers the most operational solutions in dynamic composition of services [13-16]. Several tools are available for research use and many studies still try to improve the performances, in particular by proposing AI planners dedicated to dynamic generation of composite Web services plans.

On the other hand, recent research used CBR efficiently in dynamic (or semi-dynamic) Web service composition. We aim to apply CBR over an AI Planner. The idea is that the used plans are generated by an AI planner and whenever a new query is given the system first attempts to get a solution from the stored cases. If no similar case is found, or in case of unsatisfactory solution: a planner is used to generate a new solution from scratch.

The following subsections present the main ideas in using AI planning and CBR for dynamic Web service composition.

2.1 AI Planning for Web service composition

Let's recall that a planning problem can be defined as a five tuple $\langle S, S_0, G, A, \Gamma \rangle$ where, S is the set of all possible states of the world, $S_0 \in S$ denotes the initial state of the world, $G \in S$ denotes the goal state of the world that the planning system attempts to reach, A is the set of actions the planner can perform in attempting to change one state to another state in the world and the transition relation $\Gamma \subset S \times A \times S$ defines the precondition and effects for the execution of each action [10].

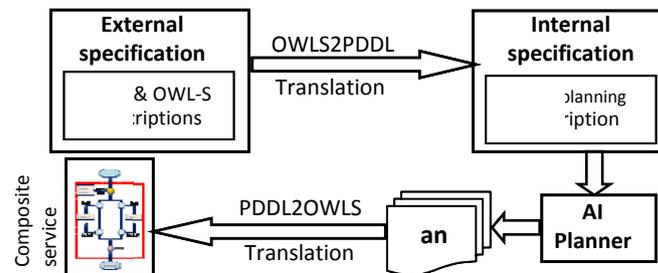


Figure 1: Applying AI planning to Web service composition

A simple analogy can be made between a Web service composition problem and a planning problem as follows: consider the user query as the initial state (S_0) of the world; the set of available Web services represents the set (A) of actions; Web service inputs (resp. outputs) represent the precondition (resp. effects) of the corresponding action. This correspondence makes it possible to transform a Web service composition problem into a planning problem. Then, an AI planner can be used to derive a plan to offer an acceptable solution to the user query.

This transformation can be done by translating the original description of the problem into a description which corresponds to a planning problem. A Web service composition problem is often described using the OWL-S language [7]. This

description is referred to as the *external specification*. On the other hand, the PDDL language [17] is most often used for the description of a planning problem. This description is referred to as the *internal specification*. Figure 1 depicts the overall principle of resolving a Web service composition problem by using AI planning.

Many research works [13, 15, 16] used the principle of figure 1 to generate a composition plan automatically. However, there are some limits in translating OWL-S descriptions into PDDL. These restrictions concern some complex plan structures allowed by OWL-S (such as unordered and iterations) but not permitted in PDDL.

2.2 Case based reasoning for Web service composition

Case-based reasoning is a problem solving paradigm that in many respects is fundamentally different from other major AI approaches [18]. In CBR, the primary knowledge source is a memory of stored cases (case base) recording specific prior episodes. The processes involved in CBR can be described by: A new problem is matched against cases in the case base and one or more similar cases are *retrieved*. A solution suggested by the matching cases is then *reused* and tested for success. Unless the retrieved case is a close match the solution will probably have to be *revised* producing a new case that can be *retained* [19] (figure 2).

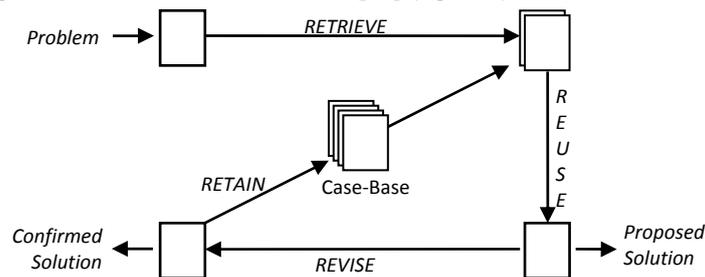


Figure 2: The CBR Cycle [19]

During the last few years, many research works used CBR in Web service composition. We present in the following the main ideas published in this area.

Lajmi et al. [22] propose an approach called WeSCo CBR that aims at enhancing the process of Web service composition by using a CBR technique. Web services are annotated using OWL-S and grouped into communities to facilitate the search process. In order to improve the search of the most relevant case (for a new case), a classification of the existing cases is proposed. The proposed solution is intended to respond to a request for a medical diagnosis of the early detection of cardiac ischemia and arrhythmia.

Osman et al. [20] present an approach that uses CBR for modeling dynamic Web service discovery and matchmaking. The framework considers Web services execution experiences in the decision making process and it is sensitive to rules issued by the service requester. The framework also uses OWL semantic descriptions extensively to implement the components of the CBR engine, as well as the services selection profiles. In addition, the proposal uses a classification of user groups into profiles that have standard set of constraint rankings.

Recently, Lee et al. [6] build a prototype that combines planning and CBR for dynamic service composition. The work accepts a service request from a user through intent analysis producing a goal model by extending the service request with keywords representing the user intent. CBR is used to provide composite services quickly. The tool JSHOP2 [14] is used to generate composition plans. The work used simulated Web services for transport, including airline tickets and other services. It also proposed merging internal and external services to meet user needs.

3. Our Proposal

Our approach to dynamically produce composite services is based on the use of case-based reasoning and AI planning. We apply CBR to store planning and related information in a case base to create planning much faster when users have similar needs. The overall architecture of the system is depicted in figure 3.

A case is a triplet consisting of the goal model extracted from the query, the corresponding OWL-S solution and an outcome. The goal model is used as features for case searching and matchmaking.

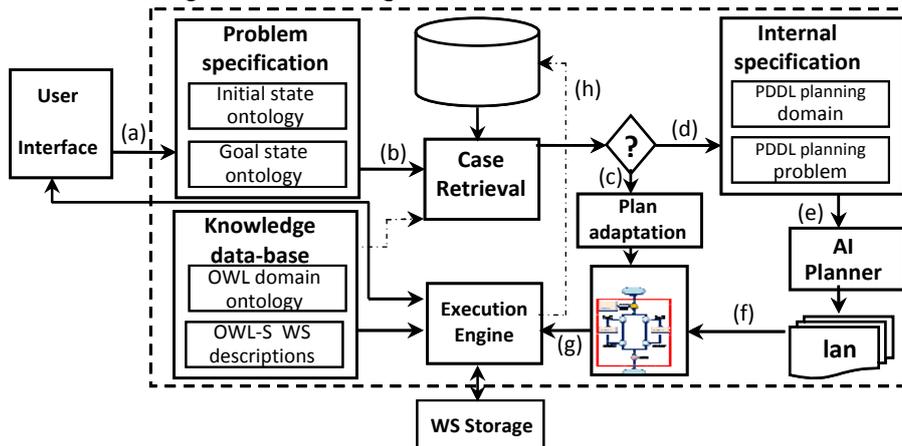


Figure 3: Architecture of the proposed solution

- a) A new query is introduced via the user interface. This query is considered as a new case and is semantically annotated using OWL-S.
- b) The case retrieval module tries to find a match for the new case in the case base.
- c) Unless the retrieved case is an exact match, an adaptation of the corresponding solution is necessary.
- d) When no matches exist, or in case of an unsatisfactory solution, the new problem is translated into a planning problem.
- e) An AI planner is used to derive a new plan for the translated problem. Our system uses OWLS-Xplan2.0 [23] to generate a new AI composition plan.
- f) In order to be executed, the generated plan is translated into OWL-S.

- g) The execution engine binds the composite service activities to concrete Web services (by querying service registries) and returns the resulting composite service to the user. An evaluation of the proposed solution is then made.
- h) Depending on the evaluation the new case can be stored in the case base.

4. Motivating scenario

Our prototype for dynamic Web service composition is currently applied in a national research project (PNR 12/u310/65) [24] that concerns the Monitoring of Childhood Immunization (MCI) in Algeria. The system presently underway aims to have total immunization coverage and an access to the immunization status of every child from any department all over the country. In order to insure that every child is immunized according to a fixed calendar a vaccination notebook (VN) is established and maintained by the immunization monitoring service (IMS). This notebook is generated by the IMS of the municipality where the child was born (city of birth CB). Every municipality is attached to an IMS which in turn monitors several immunization services (IS). Children are dispatched into different ISs according to their parents' address (PA at birth date).

The information manipulated by the MCI system comes from many sources:

- a) The birth registry located at the municipality: Information about the child's name, date of birth, parents' names, hospital of birth, name of the doctor, etc.
- b) The address registry located at the municipality: Information about the IS a child is assigned to according to his PA and to the urban cutting.
- c) The vaccination notebooks registry located at an IMS: The history of previous vaccinations for a given child and the schedule of incoming immunizations.

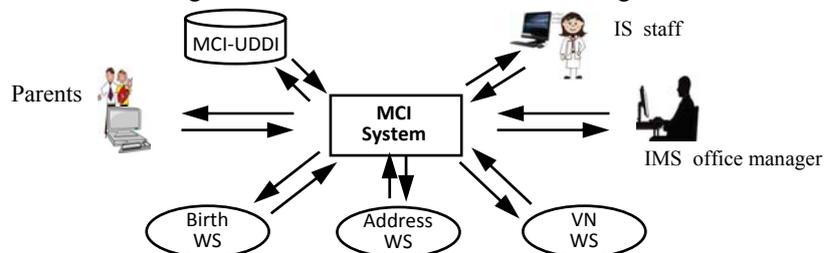


Figure 4: A Service Oriented Architecture for the MCI-System

Web services are used to access each registry. Every municipality and every IMS has its own registries. And even though the structure of information stored in different municipalities or IMSs is roughly the same (e.g. the birth registry), different Web services should be implemented because of particular considerations (e.g. use of different DBMS). It means that the activities are exactly the same for all municipalities and IMSs, but each of which may rely on a different technological platform. All Web services are advertised in a private UDDI called MCI-UDDI. Figure 4 depicts the overall functional structure of the MCI. Domain ontology is developed which allows giving OWL-S annotations for published Web services.

Queries to the MCI system come from different types of users and each query triggers a composition of services depending on the information given by the user (CB, PA, ..), the type of user, and the desired result.

5. Conclusion and discussions

We presented a solution that combines CBR and AI planning for dynamic composition of services. Instead of testing the solution on simulated Web services we have chosen to apply our proposal on a real example. The use of CBR gives a way to memorize past experiences in order to reuse previous successful solutions. As a result, a solution is provided quickly. On the other hand, the use of AI planning allows proposing a solution when no previous similar cases exist or when the proposed solution does not satisfy the user. AI planning also allows populating the case base when applying our solution in a new domain. The advantage of using PDDL is to pave the way toward the use of a wide range of planners. Moreover, in addition of using an existing planner, we are implementing a new AI planner that utilizes the principle of the cellular machine [25]. The objective is to produce faster and more efficient plans.

A few issues in the use of CBR are still under examination. In particular we are experiencing the use of decision trees to improve the similarity calculus as in [26]. The other issue is the adaptation of a solution. We are still working on a satisfactory approach to adapt an existing solution.

6. References

1. Web Services Architecture. <http://www.w3.org/TR/ws-arch/wsa.pdf>.
2. Agarwal, V., Chafle, G., Mittal, S., Srivastava, B.: Understanding Approaches for Web Service Composition and Execution. Proc. Of the 1st Bangalore Annual Compute Conference (COMPUTE'08), 2008.
3. Srivastava, B., Koehler, J.: Web Service Composition Current Solutions and Open Problems. In Workshop on Planning for Web Services (ICAPS), 2003.
4. Prasath Sivasubramanian, S., Ilavarasan, E., Vadivelou, G.: Dynamic Web Service Composition: Challenges and Techniques. Int. Conf. on Intelligent Agent & Multi-Agent Systems (IAMA'09), 2009 (1-8).
5. Martin, D., Paolucci, M., McIlraith, S., Burstein, M., McDermott, D., McGuinness, D., Parsia, B., Payne, T., Sabou, M., Solanki, M., Srinivasan, N., and Sycara, K., Bringing Semantics to Web Services: The OWL-S Approach. Proc. 1st Int. Workshop on Semantic Web Services and Web Process Composition (SWSWPC'04), California, USA, 2004.
6. Lee, C.L., Liu, A., and Huang, H.: Using Planning and Case-Based Reasoning for Service Composition. In J. Advanced Computational Intelligence and Intelligent Informatics, Vol. 14, Issue 5, 2010 (540-548).
7. OWL-S: Semantic Markup for Web Services. <http://www.w3.org/Submission/OWL-S/>.

8. Miller, J., Verma, K., Shelth, A., Aggarwal, R., Sivashanmugan, K.: WSDL-S: Adding Semantics to WSDL white paper. Technical Report, Large Scale Distributed Information Systems (2004).
9. Semantic Annotations for WSDL and XML Schema – Usage Guide. Available at: <http://www.w3.org/TR/sawSDL-guide/>.
10. Rao, J., Su, X.: A Survey of Automated Web Service Composition Methods. In Proc. 1st Int. Workshop on Semantic Web Services and Web Process Composition, July 2004.
11. PonHarshavardhan, Akilandeswari, J., Manjari, M.: Dynamic Web Service Composition Problems and Solution -A Survey. In Proc. 2nd Int. Conference on Information Systems and Technology (ICIST), India, May 2011 (1-5).
12. Milanovic, N., Malek, M.: Current Solutions for Web Service Composition. IEEE Internet Computing, Vol. 8, No.6, 2004 (51-59).
13. Ďurčik, Z.: Automated Web Service Composition with Knowledge Approach. Information Sciences and Technologies Bulletin of the ACM Slovakia, Vol. 2, No. 2, 2010 (35-42).
14. Wu, D., Sirin, E., Hendler, J., Nau, D., Parsia., B.: Automatic Web Services Composition using SHOP2. In Proc. 2nd Int. Semantic Web Conference (ISWC), 2003.
15. Klusch , M., Gerber, A.: Semantic Web Service Composition Planning with OWLS-XPlan, 1st Int. AAAI Fall Symposium on Agents and the Semantic Web, USA, 2005.
16. Peer, J.: A PDDL Based Tool for Automatic Web Service Composition. Lecture Notes in Computer Science, Vol. 3208, 2004 (149-163).
17. Ghallab, M., Howe, A., Knoblock, C., McDermott, D., Ram, A., Veloso, M., Weld, D., Wilkins, D.: PDDL - The Planning Domain Definition Language, Version 1.2. Yale Center for Computational Vision and Control, Tech. Report CVC TR-98-003/DCS TR-1165, October, 1998.
18. Aamodt , A., Plaza, E.: Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. AI Communications, Vol.7 No.1, 1994 (39-59).
19. Watson, I., Marir, F : Case-Based Reasoning: A Review. The Knowledge Engineering Review, Vol. 9, No. 4, 1994 (355-381).
20. Osman, T., Dhavalkumar Thakker, D., Al-Dabass, D.; Semantic-Driven Matchmaking of Web Services Using Case-Based Reasoning. IEEE Int. Conference on Web Services (ICWS'06), 2006 (29-36).
21. Leake, B.: CBR in Context: The Present and Future. In Leake, D., ed., 1996, Case-Based Reasoning: Experiences, Lessons, and Future Directions. Menlo Park: AAAI Press/MIT Press, 1996.
22. Lajmi, S., Ghedira, C., Ghedira, K., Benslimane, D.: CBR Method for Web Service Composition. In Advanced Internet Based Systems and Applications, Lecture Notes in Computer Science, Vol. 4879, 2009 (314-326).
23. Sem WebCentral. Available in: <http://www.semwebcentral.org/projects/owls-xplan/>.
24. Accepted PNR projects. Available in: http://www.nasr-dz.org/dprep/pnr2/projets-pnr/PNR_a.htm.
25. Atmani, B., Beldjilali., B.: Knowledge Discovery in Database: Induction Graph and Cellular Automaton. J. Computing and Informatics, Vol. 26, No. 2, 2007, (1001-1027).
26. Benbelkacem, S., Atmani,B., Mansoul, A.: Planification Guidée par Raisonnement à Base de Cas et Datamining : Remémoration des Cas par Arbre de Décision. Accepted paper in EGC 2012 : http://eric.univ-lyon2.fr/~aide/?page=accepted_papers.html, Bordeaux, France, Jan 2012.
27. Kolodner, J. L.: Case-Based Reasoning. Morgan Kaufmann, San Mateo, 1993.

A collaborative web-based Application for health care tasks planning

Fouzi Lezzar¹, Abdelmadjid Zidani², Atef Chorfi³

^{1,2,3} Laboratoire de sciences et technologies de l'information et de la communication, University of Batna, Algeria

¹Lezzar.fouzi@gmail.com

²abdelmadjid.zidani@univ-batna.dz

³Atefchorfi@yahoo.fr

Abstract. Hospital emergency wards such as Gynaecology and obstetrics maternities are extremely complex to manage and pose serious health risks to patients. Related tasks which are mainly focused around patient management are basically achieved through a cooperative way that involves several health care professionals. Such team members with separate skills and roles should work together during patients' management. In this paper, we firstly discuss our study of work in-situ within an Algerian maternity ward to better understand the usual way under which tasks are effectively achieved and identify the artefacts used. Such observation allows us to highlight the vital collaborative medical tasks that need to be modelled. The following sections outline basic design concepts of our collaborative planning system, which is designed to provide a flexible group interaction support for care coordination and continuity.

Keywords: Healthcare tasks modelling, cooperative work, shared artefacts, synchronous/asynchronous interaction, social planning and coordination.

1 Introduction

In Algeria, many healthcare institutions across the country suffer from multiple dysfunctions. Despite the reforms initiated by the authorities to improve the quality and effectiveness of patients care, the changes promised by the reforms initiators slow in coming and health care wards still do not meet the expectations of the patients. Certainly, no one can ignore the achievements made in this sensitive area that is public health such as rehabilitation of old infrastructures and reception of new ones, opening of new services, training of more physicians, medical specialists and skilled staff, etc., but problems persist and severely affect the medical activity and globally the health system.

It is here about an observation that cannot be restricted to the isolated counties within the country like south for example, which use old medical equipment and lack in terms of qualified personnel, but also concerns hospitals in major northern cities which are nevertheless materially well equipped with the availability of the required

skills. Indeed, the study¹ that we led on this vital question revealed various reasons that are mainly linked to mismanagement of the related activities, equipment, human and material resources.

First, it is necessary to note that the artefacts used during work are essentially restricted to paper sheets which are often not updated and sometimes even get lost between the different services because of the infernal work load imposed on the personnel. The observation study has also explicitly showed that most of the medical activities we supervised were group-based. Likewise the main deficiencies in patients monitoring precisely arose from the lack of coordination between the various members of the involved medical team, which thus constitutes a key factor as it has been so well confirmed several numerous studies carried out on this issue [20] [21].

Based on this observation, we believe that the task management should require from us a special attention. We must therefore address the issue of the targeted maternity ward under a new perspective, that of the medical staff needs, taking into account the economic and performance constraints as well as socio-health hospitals mission objectives of providing optimal care and well-being of patients. Providing a technological answer through cooperation, coordination and communication facilities seems to be the most appropriate initiative. However the past experiences reported in this area that work in-situ [19] should be first carefully analyzed from a social point of view [4], and through a structurally opened cooperation vision that enables users to build their cooperation workspace structure in order to interact within it [18]. Consequently, we focused our interest on collaborative practices of patient care teams [14] as well as their organization [16] to better understand the usual manner with which tasks are actually performed.

Several collaborative medical care needs have been identified by a wide body of researches in informatics and medical science fields. There are common processes that are more difficult and complex in collaborative situations, because they need to integrate many parties. Such as decision making process that needs involvement of several persons to arrive to a decision, which can take long time [15]. In [13] authors showed that the collaborative nature of the executed process, determines the type of information management necessary for this process [15]. Though, a poor structure of information can lead to coordination and communication breakdowns [15].

Maternity services are highly risky and still very hard to manage. They require coordination among several teams whose tasks achievement most of the time confront them to conflict situations [19]. The exploitation of information and communication technologies proves to be an effective approach if it is appropriately used [11]. It will enable us reduce the effects generated by the coordination problems that directly disrupt the patient's care chain and degrade their quality as noted Scupelli [10] through his study. Consequently, coordination breakdowns among the medical staff members inevitably that have an impact on the quality of care provided to patients and put them in a potentially vulnerable and dangerous situations should be significantly reduced with the availability of a medium of communication, cooperation, and coordination.

¹We led an investigation mainly based on observations and interviews with a medical staff within a gynecology and obstetrics emergency unit.

Such an approach will provide collaborative tools that may effectively address medical staff vital needs and improve the quality of patients' care. Our research work falls then within the CSCW area (Computer Supported Cooperative Work). Thus, with a CSCW-based management strategy [8] we wish to provide an effective support of these activities enabling by the way finer planning features of the related tasks as well as providing real time mutual awareness around the occurring events within the maternity unit which constitutes a priority of our work.

The main objective of this paper is to outline basic design concepts of our cooperative planning system CPlan. In the following sections we will first discuss our observational study achieved within an Algerian maternity ward to better understand the usual way under which tasks are effectively completed as well as identify the used artefacts. We will attempt to analyze the healthcare process to highlight the appropriate design guidelines. Section 3, exposes our conceptual methodology and discusses the choices made as well as the software architecture designed for CPlan. We consider the main components of the different architecture levels as well as the main supported features. We will explicitly attempt to show that CPlan design is mainly focused on concepts of data sharing and exchange to favour coordination between participants. To provide details on our design approach, section4 discusses its deployment issue. Finally, perspectives of the accomplished work are presented in the conclusion of the paper.

2 Targeted context study

Our study of work in-situ led us to consider an Algerian maternity ward to better understand the usual way under which medical staff such as gynaecologists-obstetricians, anaesthetists, midwives, nurses, ..., effectively achieve their tasks and identify the main used artefacts to coordinate the work. The maternity targeted is about 200 beds and comprises 4 operating rooms, 4 labour rooms, an analysis laboratory, an imagery service, an emergency service, etc.

We started, therefore, by analyzing the interactions among the medical staff members and attempted to understand how they may interact and collaborate while dealing with patients' cases, and what happens when this work is done with a team of collaborators. Such understanding will undoubtedly allow us to provide the adequate design by addressing the following interrogations:

- How medical staff members' collaboration naturally takes place?
- Which artefacts are used to coordinate work? And how?
- Which impact has the spatiotemporal dimension on staff members' interactions and on the collaboration process among them?
- What means are required to improve the care process within a maternity ward?
- Which computer tools may provide the required assistance for the medical staff members and get them to work collaboratively?

— From a collaboration point of view, what are the specific characteristics of collaborative medical activities?

It is practically impossible to design a computer tool addressing all users' needs. Nevertheless, group work experiences provide us with pertinent information to clarify some useful development ideas about the suitable support tools. The experimentation of these tools, thereafter, will unveil obstacles to overcome as well as perspectives to follow. Our approach is drawn in a direction which aims to favour collective work and enables coordination. Therefore, as we will show it in the following sections, the care tasks analysis will bring us an understanding to concretely increase the commitment of participants that may have a great impact on the whole chain care process.

2.1 Collaboration process

The meticulous analysis of healthcare activities reveals that patient care chain planning is a complex task that has an important impact on their quality and consequently on patients' safety. Such care process must be carefully managed since the patients' admission to the hospital until they recover and leave it. This includes ongoing care chain planning of a pregnant woman since her admission to the maternity until her delivery that can occur naturally (labour room) or through a caesarean surgery (operating room).

When there is a coordination breakdown between team members, this can affect directly the patients care activity. In this study we noticed that there are many sources of coordination breakdowns, that have to be taken in consideration. A change in the patient's physical condition either for the worse or for the better can require a changing in the schedule. For example if a doctor decides that a pregnant woman can need a caesarean operation, this needs an immediate allocation of an operating room. Some coordination problems can come from surgeons. Surgeons do not have often one obligation, but many; such as carrying out a surgery, seeing their patients, or working in other hospitals. When the amount of tasks is big, and there is a lack of awareness and coordination; this can be a source of delays. A not experienced nurse, who is not accustomed with the work in the maternity ward, can make some mistakes, what can affect the schedule. Team members can affect and slow down each other, with unexpected events and requests for information which require updating and adjustment in the plan.

When coordination breakdowns occur, schedule has to be adjusted: reallocations of resources, update of priorities, notifying of the involved medical staff... Negative consequences can happen when the medical staff fails to act collaboratively to adjust breakdowns. Our analysis has revealed that medical care is often administrated with a delay, and unfortunately even for critical cases, what can be sometimes dangerous for patients' life. These coordination breakdowns, can lead to delays, what leads to more work hours and additional costs what reduces profits. Also, trying to coordinate every time between team members; can generate stress and workload. Sometimes delays can oblige patients to come back another day, what disturbs their personal plans.

Our study, reveals us that putting artefacts in some specific positions inside the maternity ward can increase awareness, improve the collaboration process, reduce the costs of sharing and gathering information and decrease coordination problems.

The planning process should take into account for any task the availability of the associated medical team members (such as gynaecologist-obstetrician, anaesthetist, midwife, and nurse), the location, the period of time, etc. The collaborative planning tool shows immediately the old scheduled tasks and easily allows planning new ones while visualization provides for specific periods information on availability of current working staff as well as locations (labour and operating rooms).

2.2 Work analysis

Designing group work support features requires first a better work in situ analysis, and particularly identifying the implied participants, their roles, prerogatives as well as the used artefacts. Such way will without doubt enable us to understand how to satisfy both individual and group requirements within a shared environment.

Our design approach is intended to enable medical staff members to cooperate and share responsibility of a patient. We insist here on the necessity for an effective groupware tool to take into account the procedural, intellectual and social complexity of the cooperative care process planning. Indeed the diversity of opinions inside the staff, often generate a great intellectual activity that should be gathered and made available to the community rather than neglected until it becomes a source of conflicts or misunderstandings.

In addition to the obstetricians-gynaecologists, anaesthesiologist, paediatrician, etc., the gynecology unit functioning is mainly based on the chief midwife, who is in charge of the care organization, of their quality and their ongoing as well as the motherhood monitoring and her staff management (usually other midwives and assistant nurses). Among the other professionals involved in the service we also distinguish the anaesthetist nurse (a specialized nurse) who assists the anaesthesiologist and supervises the postoperative recovery room. Finally, the staff also involves a social assistant who mediates between patients meeting personal problems and administrative agencies, a psychologist who offers listening, support and advice to patients and families, a physiotherapist for functional rehabilitation and massage therapy, and a nutritionist who tailors the appropriate diet to health problems.

2.3 Used Artefacts

During patients' management, the involved team usually resorts for scheduling to a classical plan board or paper sheets to specify who does what, and when? As well as to coordinate the work with people who are not available, it is usual to use the telephone, email and short messages. Such way to achieve work promotes creativity and information sharing that suitably works and allows the group to get at an on-time objective. Thus, the whole process requires from the team members to take part to the planning process and do nothing else at this moment. Because the planning process works well when under a face to face way, while the detailed tasks are discussed with

the whole team. However, because of their Ad-hoc nature and emergencies the medical activities require not only continuous availability but also a high level of vigilance from the medical staff which should constantly be focused on the evolution of patients' conditions. Therefore, these meetings which are necessary to ensure coordination should be minimized as much as possible. Just as it would be necessary to constantly maintain mutual awareness on the occurring events, even for the busy group members while dealing with emergencies. That is how coordination problems arise and lead to the disruption of the balance within the group leading to tension, nervousness, tiredness, anxiety, etc.

The most used artefact as we said above is the paper medical record. It contains much information about the patient (observations, plan of the day, and dosage of drugs). There is a new paper each day that is placed on the top of the old ones. With time, the consultant need more time to consult a patient state because of the big number of papers added every day. The use of papers has to be reduced to minimum, to avoid some problems such as lose of papers, the need to move to the patient's room to consult its state,... The use of electronic medical records (EMRs) can improve awareness among the group members, and improve collaboration process [1]. The use of EMRs, must be coupled with the correct display device [3]. Because a poor development of the ergonomic design can lead to a difficulty using [7].

Likewise, the strategy used by these institutions has often emphasized on a management that attempts to deal with the massive affluence of patients rather than the quality of their care. Furthermore, we currently see in the emergency service the admission of more and more complex health cases, whose take-in-charge remains an extremely hard task.

3 Software architecture

The developed system is a synchronous web-based groupware accessible through a browser that enables real-time collaboration among collocated or geographically separated group members. The proposed architecture is illustrated on Figure 1. It is developed under AJAX Push also known as Server Push or Comet [2].

The first layer contains the system database which is mainly characterized by its capacity to provide reliable data for long time, concurrency control management, data storage, and security capabilities.

The second layer contains all the defined software components. In the case of a real-time groupware, sharing data and events constitutes the most important aspect. Thus enabling data sharing requires that any event or data generated by one user has to be immediately notified and delivered to all the other users (in real-time). For better workspace awareness, fault tolerance, responsiveness, and replication of shared data, objects are often used together with other operations on them like creation, updating, deletion and reading.

Some web 2.0 technologies such as AJAX and Comet, allow creation of rich internet applications (RIAs). In an Ajax application, servers respond to each request in sequence, just as in classical web, but in the browser only a part of the user interface

is updated, rather than updating the whole page and refresh the whole display. However, the user must send a request to the server to see the updates.

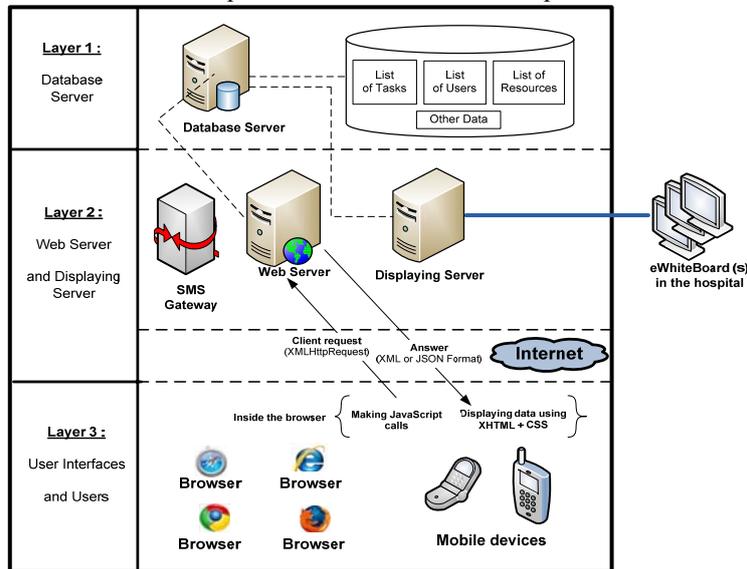


Fig. 1. Software architecture.

The problem of AJAX which is the absence of two-way communication that is needed for synchronous groupware implementation is partially solved with a set of technologies commonly called Comet. They allow a server to push data to the browser ('server push') without requiring a new connection for each update. This capability allows a server to notify data to clients at any moment. Comet is ideal for collaborative and real time applications, because of its abilities for improving responsiveness of collaborative systems, without causing any throughput problems [2].

In [12], authors carried out a study on groupware-based framework requirements to assess the performance of different networking approaches including Comet. They found that web-based networking approaches perform well and can support the communication requirements of many types of real-time groupware. The results suggest that web technologies can support a wide variety of network requirements, including highly interactive shared workspaces and systems for large groups.

The second layer defines two servers: Web-Server and Displaying-Server [5] [6]. The web server contains rich web pages that may be loaded on users' browsers. The Displaying Server is intended to display the schedule on eWhiteBoards (screens disposed on the appropriate locations in the hospital). After every modification of the schedule, all the eWhiteBoards are automatically updated. Also to significantly reduce users' cognitive overload such as nurses, surgeons..., the eWhiteBoard(s) can be configured to restrict the display only for pertinent information needed by each group [19] and decrease the amount of data on screens.

Finally, the third part consists of the client machines, which may be a laptop or a desktop. However, recent years have seen a wide variety of computer devices includ-

ing mobile telephones, smartphones and tablets that can be considered as an alternative for traditional computers.

3.1 Software Architecture Components

The following components are loaded in the browser from the Web Server. Our architecture is composed of several modules which are important for the collaborative scheduling task:

- Interface component: this module plays the role of a medium between the user and the system.
- Session Manager: This component is intended to manage users' work sessions, like rights on the schedule list (read/write), users join/leave within the shared workspace, latecomers...
- Collaboration: to allow users to simultaneously work together, we designed several appropriate tools, such as the tasks shared table. Users share the same display which instantly shows any event that may occur in consequence of a user action. Such way provides users with real time awareness capabilities and enhances coordination.
- Communication: CPlan supports both synchronous (instant messaging) and asynchronous (post comments, and give valuable suggestions to their colleagues).
- Scheduling: This is the most important system component; it provides all the necessary tools for managing resources and tasks. The list of the tasks is displayed to the users in a table that showing all related tasks information (starting time, priority, location...). Once a task is created, it will immediately appear on the other participants (users) screens.
- Collaborative diagnostic: This is a component intended for diagnostics elaboration of a given case under a collaborative way.

The server extracts the required information from the database and uses an SMS gateway to send messages to the staff members. There are two kinds of messages: reminders, notifiers to notify new events or a new created task. To allow users to connect to our web interface with their mobile devices, an adapted version of the web application is developed. The developed system allows an authorized physician to access at any location to the electronic patient record data, using a hand held device or a desktop, and can for example remotely access the patient medical images.

3.2 Events notification

Our system uses an event notification mechanism. When any action is executed within the shared workspace [17], the web server notifies the other users to inform them (XML messages) about the different actions in the shared workspace. Such mechanism keeps the whole group member aware of their mutual actions [9].

4 Conclusions and perspectives

To objectively measure the efficiency of CPlan, we have implemented a first prototype of the collaborative tool, and the evaluation of the current version on a local network brings rich ideas on collaboration and coordination opportunities provided.

In this paper, we have discussed basic design concepts of our groupware application CPlan. We have attempted to show that it allows several participants to collaborate within a shared workspace. It allows the execution of individual and collective actions on a common patient case as the elaboration of the planning. The sharing of the planning scheduling has been widely discussed, because its use allows us to concretely inform participants on their mutual actions. At the visual level, the simplified and rich web interface shows explicitly the shared plan phases and significantly reduces participants' cognitive loads and enables them to intuitively understand what is currently going on and get knowledge about their patients' states evolution as well as the next actions that should be achieved.

During a work session, medical staff members may act on the shared plan under specific role, dynamically exchange messages and interact through natural cooperation way. Such flexibility is motivated by the necessity to enable CPlan to support the dynamics implied by the care process.

Being conscious of the great interest of CPlan experimentation in effective context situations, we plan in the next step of our research work to collect information about efficient activities from medical staff. It is of an extreme importance for us and represents a double objective. First, we can validate or forsake some technical choices among those we made for implementation. Second, we will be able to determine with more precisions the appropriated adaptations we should apply to the supports provided in CPlan. To this end such as any software project we designed modular and extendable software architecture, in the sense that it allows design and integration of new modules through an incremental way.

References

1. Craig, E. Kuziemy, James, B. Williams., Jens, H. Weber-jahnke.: Towards Electronic Health Record Support for Collaborative Processes. SEHC '11 Proceedings of the 3rd Workshop on Software Engineering in Health Care (2011)
2. Russell, A.: Comet: Low Latency Data for the Browser. Continuing Intermittent Incoherency (2006)
3. Cecily Morrison., Geraldine Fitzpatrick., Alan Blackwell.: Multi-disciplinary collaboration during ward rounds: Embodied aspects of electronic medical record usage. International Journal of Medical Informatics Vol. 80, Issue 8, Pages e96-e111 (2011)
4. J. Cummings., S. Kiesler.: Coordination and success in multidisciplinary scientific collaborations. International Conference on Information Systems (ICIS), Seattle, WA: Association for Information Systems (2003)
5. Wong, H.J., Caesar, M., Bandali, S., Agnew J., Abrams H.: Electronic inpatient whiteboards: improving multidisciplinary communication and coordination of care. Int J Med Inform.;78(4):239-47 (2009)

6. M. Hertzum.: Electronic emergency-department whiteboards: A study of clinicians' expectations and experiences. *I. J. Medical Informatics*, Vol. 80, No. 9, pp. 618-630 (2011)
7. C. Tang, S. Carpendale., An observational study of information flow during nurses' shift change, *Proceedings of CHI 2007* 219–228. (2007)
8. K. Schmidt., C. Simone.: Coordination mechanisms: Towards a conceptual foundation of CSCW systems design. *Journal of Computer Supported Cooperative Work: The Journal of Collaborative Computing*, 5(2-3), 155-200(1996)
9. Dourish, P.: What we talk about when we talk about context. *Personal and Ubiquitous Computing*, 8(1), 19–30 (2004)
10. Scupelli, P., Xiao, Y., Fussell, S. R., Kiesler, S., Gross, M.: Supporting coordination in surgical suites: Physical aspects of common information spaces. *Proceedings of the Conference on Human Factors in Computing Systems CHI10*, NY: ACM Press (2010)
11. Z. Niazkhani., et al.: Evaluating the medication process in the context of CPOE use: The significance of working around the system, *Int. J. Med. Inform.* (2011).
12. Carl, A. Gutwin, Lippold,M.: Real-Time Groupware in the Browser: Testing the Performance of Web-Based Networking, T. C. Nicholas Graham March 2011 CSCW '11: Proceedings of the ACM 2011 conference on Computer supported cooperative work (2011)
13. Reddy., M.S., Spence, P.R.: Collaborative information seeking: A field study of a multidisciplinary patient care team. *Information Processing and Management*;44:242–255(2008)
14. Seffah A, Forbrig P, Javahery H Multi-devices “multiple” user interfaces: development models and research opportunities. *J SystSoftw* 73(n 2):287–3001 (2004)
15. Kuziemsky, C.E., Varpio, L.: A Model of Awareness to Enhance Our Understanding of Interprofessional Collaborative Care Delivery and Health Information System Design to Support it, *International Journal of Medical Informatics*, doi:10.1016/j.ijmedinf.2011.01.009, forthcoming (2011)
16. S. R. Barley., W. H. Dutton., S. Kiesler., P. Resnick., R. E. Kraut., J. A. Yates.: Does CSCW Need Organization Theory ?,” *Proceedings of the 2004 ACM conference on Computer supported cooperative work (CSCW'04)*, ACM Press, November 6–10, 2004, Chicago, Illinois, USA, pp.122-124 (2004)
17. U. K. Wiil, :Using Events as Support for Data Sharing in collaborative Work,” *Proceedings of the International Workshop on CSCW*, Berlin (1991)
18. M. Zacklad.: Communities of Action: a Cognitive and Social Approach to the Design of CSCW Systems,” *GROUP'03*, November 9–12, 2003, Sanibel Island,Florida, USA, pp.190-197 (2003)
19. Ren, Y., S. Kiesler, S. Fussell, P. Scupelli.: Multiple Group Coordination in Complex and Dynamic Task Environments: Interruptions, Coping Mechanisms, and Technology Recommendations. *Journal of Management Information Systems / Summer 2008*, Vol. 25, No. 1, pp. 105–130. (2008)
20. Bardram, J.E., Hansen, T.R: Context-based workplace awareness concepts and technologies for supporting distributed awareness in a hospital environment. *Computer Supported Cooperative Work*. 2010;19:105–138 (2010)
21. Kuziemsky, C.E., Varpio, L.: A Model of Awareness to Enhance Our Understanding of Interprofessional Collaborative Care Delivery and Health Information System Design to Support it, *International Journal of Medical Informatics*, doi:10.1016/j.ijmedinf.2011.01.009, forthcoming (2011)

Internet and Web Technologies II



Building Semantic Mashup

Abdelhamid Malki, Sidi Mohammed Benslimane

EEDIS Laboratory , University of Djilali Liabes , Sidi Bel Abbes, Algérie
abdelhamid.malki@gmail , Benslimane@univ-sba.dz

Abstract. Mashups allowed a significant advance in the automation of interactions between applications and Web resources. In particular, the combination of Web APIs is seen as a strength, which can meet the complex needs by combining the functionality and data from multiple services within a single Mashup application. Automating the process of building Mashup based mainly on the Semantics Web APIs facilitate to the developer their selection and matching. In this paper, we propose SAWADL (Semantic Annotation for Web Application Description Language), an extension of the WADL language that allows the semantization of the REST Web Service. We introduce a reference architecture with five layers representing the main functional blocks for annotating and combining web APIs, and therefore make the engineering process of Mashup applications more agile and more flexible.

Keywords: Semantic Mashup, Matching, API, SOAP, REST, SAWADL, SAWSDL.

1 Introduction

Dynamics, agility and efficiency are concepts of the future. The World Wide Web is undergoing an evolution from a static environment to a dynamic world in which mashups will play a central role. The Mashups are web applications developed by the combination of data, business logic, and/or user interfaces of web sources published and reused via APIs [8]. Thus, Mashups are designed to reduce the cost and development time of web applications.

Despite these advantages, engineering of Mashups applications requires the intervention of the developer which needs not only programming skills but also to understand the structure and semantics of APIs that wants to integrate. Currently, several tools Mashup (e.g. IBM WebSphere¹, Yahoo-pipes², etc.) are used by end-users (i.e. with less programming skills) to facilitate the building of Mashup applications. However, the intervention of the professional developer is required when the application Mashup is complex, thing that has prompted researchers to find effective solutions for creating Mashups, So that end users can build an application with a tool Mashup that guarantees the discovery, selection, and automatic or dynamic superposition of APIs

¹<http://www-01.ibm.com/software/webservers/>

²<http://pipes.yahoo.com/pipes/>

based on the semantic approach, the so-called ‘‘Semantic Mashups’’. The semantic Mashups is a Mashup whose combined APIs are supported (or annotated) by a semantic layer that allows to select and compose them in an automatic way (unambiguous).

We propose in this work SAWADL, a novel language for the semantization of REST web services [1]. SAWADL uses WADL³ description to enrich RESTful APIs with a semantic layer that allows the discovery and automatic superposition of APIs in order to automatically build Mashup applications. SAWADL is more flexible and adaptive with respect to other approaches of semantization such as SAWSDL [2] which is used to annotate the WSDL⁴ description of SOAP web services with ontological concepts.

The rest of the paper is organized as follows. Section 2 presents briefly the semantic Mashup, and presents some related work for the semantization of REST web services. In Section 3, we introduce SAWADL, a semantic annotation language for REST APIs. Our approach to build Semantic Mashup is described in Section 4. Finally we conclude and give some perspectives in Section 5.

2 Related Works

Web services enable applications to call remote procedures and to exchange data by passing well-defined messages. This can easily be used for Mashup application as a way to orchestrate different web applications. For instance, *Amazon Web Service*⁵ allows users to access most of the features of Amazon.com by using SOAP-based web services and REST-based web services. The semantic Mashup is Mashup whose combined APIs are annotated by a semantic layer that allows to select and compose them in an automatic way. In order to build an automatic Mashup, it is necessary to semantize these APIs.

For SOAP-based Web services there are two types of semantization approaches. The first (service ontology) consists of developing a complete language that describes Web services and their semantics in a single block (e.g. OWL-S, WSMO, etc.). The second approach (semantic annotation) consists of annotating existing web services with semantic information. WSDL-S, SAWSDL used to manually annotate a WSDL description with elements referring to ontologies.

As SOAP-based Web services, semantic REST-based Web services can be classified in two approaches. The first approach consists of developing an ontology that describes the REST-based Web services and their semantics in a single block. The second approach consists of annotating existing languages with semantic information. In the following, we present different propositions for the second approach.

- ***SOOWL-S advertisements (a social-oriented version of OWL-S advertisements)***

The SOOWL-S advertisements [6] proposes an extension of the OWL-S ontology in order to semanticizes the different types of APIs (e.g. SOAP, REST, JS, RSS, etc.) used in the construction of Mashup applications.

SOOWL-S ontology annotates just the I/O parameters and non-functional properties of a Web service (using the service-Profile module of the OWL-S ontology).

³ <http://www.w3.org/Submission/wadl/>

⁴ <http://www.w3.org/TR/wsdl>

⁵<http://aws.amazon.com/>

Thus, SOOWL-S ontology allows searching and automatic selection of APIs, but not their combination owing to the absence of the extension of service-Model module of the ontology OWL-S.

- **SA-REST**(*semantic annotation for REST*)

According to J. Lathem [4], most of RESTful web services use HTML pages to describe to users what the service does and how to invoke it. However, HTML is designed to be human legible but not machine readable. In order to solve this problem, [4] have used the RDFa micro formats¹⁰ which allows the integration of RDF triples above HTML description in order to add semantics to REST service and make it visible and interpretable by the machine.

- **SWEET** (*Semantic Web Services Editing Tool*)

Maleshkova et al [5] propose an integrated approach to formally describe the semantics of RESTful web services. The approach enables both the creation of machine-readable RESTful service descriptions using the hRESTS (HTML for RESTful Services) Microformat [3], and the addition of semantic annotations by the MicroWSMO Microformat⁶, in order to better support discovering services, creating mashups, and invoking them.

Table 1. shows a comparison between the different approaches of semantics REST web services.

Table 1. Comparison between the different approaches of semantics REST Web services

	SOOWL-S	SA-REST	SWEET
Type of semantization	Service Ontology	Annotation	Annotation
Publication of services	+	-	+/-
Discovery of services	+	+	+
Combinaton of services	-	+	+
Annotated description	Absent, is a Service Ontology	HTML	HREST
Type of accepted ontology	Owl	All	All
Type of API semantized	SOAP, REST, RSS	REST	REST

3 SAWADL

In this section we propose an annotation language that allows the semantization of RESTful web services to strengthen the selection and superposition of these services in Mashups applications.

SAWADL, the extension of WADL language that we propose is part of those approaches that add semantic annotations above the service description while most approaches are based on a semantic annotation above a description based on HTML which gives less homogeneity between semantized REST web services. SAWADL does not specify a language for representation of semantic models. Rather, it provides mechanisms for referencing ontological concepts defined in the external of WADL document.

⁶ <http://www.w3.org/TR/rdfa-syntax/>

The methods of annotation in SAWADL are summarized in two mechanisms: `modelReference` and `SchemesMapping`. This is done by the attribute "sawadl" followed by the appropriate extension.

`ModelReference` attribute used to associate a WADL's component to a concept of a semantic model. The items annotated a REST web service described by WADL description are the methods (`<method id="method1" name="GET">`) and parameters of input / output (`<param name = "name" type="xsdtype"/>`) of the service. The semantic concept (ontological) associated to elements of WADL through the `modelReference` attribute is represented by zero or more URLs separated by spaces, which are references to ontological concepts.

The mechanism of `schemesMapping` is achieved through two other attributes `liftingSchemesMapping` and `loweringSchemesMapping`. These attributes are used to specify the mappings between semantic data and WADL elements. The mechanism of `schemaMapping` is very interesting to understand. In fact, we employ the `loweringSchemesMapping` attribute when an element annotated in the WADL description matches more than one ontological concepts, and the URIs of the `loweringSchemesMapping` attribute point to files containing SPARQL⁷ queries and XSLT⁸ transformations. While we use `liftingSchemesMapping` when several elements annotated in the WADL description represent a single ontological concept, and URIs can point to files that contain XQuery⁹ queries or XSLT transformations.

3.1 Annotation of methods

SAWADL provides mechanisms to annotate methods in a WADL documents. To illustrate these mechanisms, we use a domain ontology of tourism `TravelOnto` (which we implemented in OWL) to annotate the `BookFlight` operation of Flight API. Although traditionally the inputs and outputs provide an intuitive semantics of an operation, a simple semantic annotation can be helpful. Thus we will annotate the `BookFlight` operation by associating through the `modelReference` attribute with a `BookFlight` concept in the `TravelOnto` ontology (Figure.1).

3.2 Annotation of Inputs/Outputs parameters

In SAWADL, the Input/Output parameters annotation is done in two different ways:

Internal Annotation. This annotation consists in associating each parameter input/output "`<param...>`" of a method to a concept in an ontology. This supposes that for each parameter input/output of a method there exist a corresponding concept in the ontology. For example, the input of the operation `BookFlight` is composed of name and age of the passenger, and the number and class of Flight. We suppose that for each attribute, there exists a concept that corresponds to it in the `TravelOnto` ontology. In the case where there is no match, the semantics of the input/output parameters is not specified. Figure 2 show an example of internal annotation.

⁷ <http://www.w3.org/TR/rdf-sparql-query/>

⁸ <http://www.w3.org/TR/xslt>

⁹ <http://www.w3.org/TR/xquery>

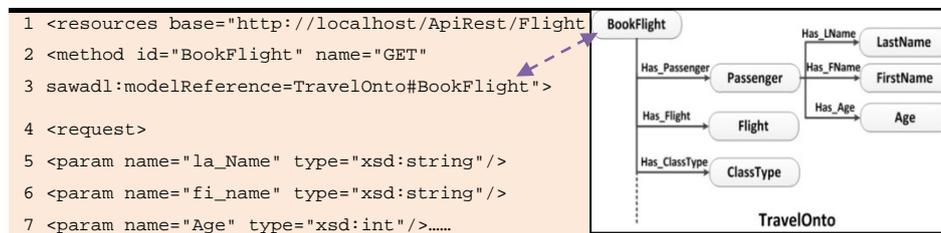


Fig. 1. Annotation of methods with SAWADL

```

1 <resources base=" http://localhost/ApiRest/Flight " >
2 <method id="BookFlight" name="GET">
3 <request>
4 <param name="la_name" type="xsd:string" sawadl:modelReference="TravelOnto#LastName" />
5 <param name="fi_name" type="xsd:string" sawadl:modelReference="TravelOnto#FirstName" />
6 <param name="Age" type="xsd:int" sawadl:modelReference="TravelOnto#Age" />
7 <param name="NFlight" type="xsd:string" sawadl:modelReference=" TravelOnto#Flight" />
8 <param name="Class" type="xsd:String" sawadl:modelReference="TravelOnto#ClassType" />
9 </request>...

```

Fig. 2. Internal Annotation

External Annotation. In this case, the parameters are annotated globally via the tag “<request>”, however, it must create a schémaMapping for specifying the transformation rules between the input’s/output’s parameters and the domain ontology.

As an illustration, we take the example of `credit card` defines in WADL and the OWL ontology `TravelOnto` (see Figure 3). In this ontology there is no individual correspondence for the two attributes `last_name` and `first_name`. However, the `Owner` concept of ontology is the merger of these two attributes. To establish the correspondence between the inputs of the `credit card` API and `CreditCard` concept, it must first associate using `sawadl:modelReference` and then define a transformation scheme using an XSL style sheet via the attribute `sawadl:liftingSchemaMapping` (see Figure 4).

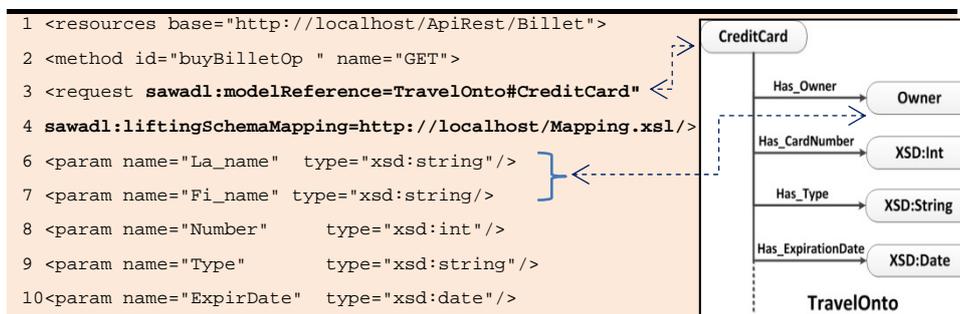


Fig. 3. External Annotation

```

<xsl:transform version="2.0"
  xmlns:Travel=http://localhost/ApiRest/Billet#  xmlns:TravelOnto="http://localhost/TavelOnto#"
  <xsl:output method="xml" version="1.0" encoding="iso-8859-1" indent="yes"/>
  <xsl:template match="/">
    <rdf:RDF>
      <TravelOnto:CreditCard>
        <hasOwner  rdf:resource="#Owner">
          <xsl:value-of select="concat(Travel:./param[@name='La_name'],Travel:./param[@name='fi_name'])"/>
        </hasOwner>
        <hasCardNumber  rdf:datatype="xs:Int"><xsl:value-of select="Travel:./param[@name='Number']">
        </hasCardNumber>
        <hasType  rdf:datatype="xs:string"> <xsl:value-of select="Travel:./param[@name='Type']">
        </hasType>
        <hasExpritionDate  rdf:datatype="xs:Date"> <xsl:value-of select="Travel:./param[@name='ExpirDate']">
        </hasExpritiondate>
      </TravelOnto:CreditCard>
    </rdf:RDF>
  </xsl:template>
</xsl:transform>

```

Fig.4. XSL style sheet via the attribute sawadl: lifting Schema Mapping

4 Building semantic Mashup

The construction of automatic Mashups necessarily requires a semantic layer on top of APIs (web services). As the dynamic composition of standard web services, the semantic Mashup allows a more rapid development and transparent composition to the user. But unlike to that of traditional web services, the Mashups are composed of APIs of different nature which makes their combination process more difficult.

Figure 5 shows reference architecture for Semantic Mashup. This architecture consists of five layers. The layers represent the main functional blocks for automatic generation of Mashup. The ontology is used to enrich the engineering process by a semantic layer that allows him an automatic selection and a combination of APIs included in the Mashup application.

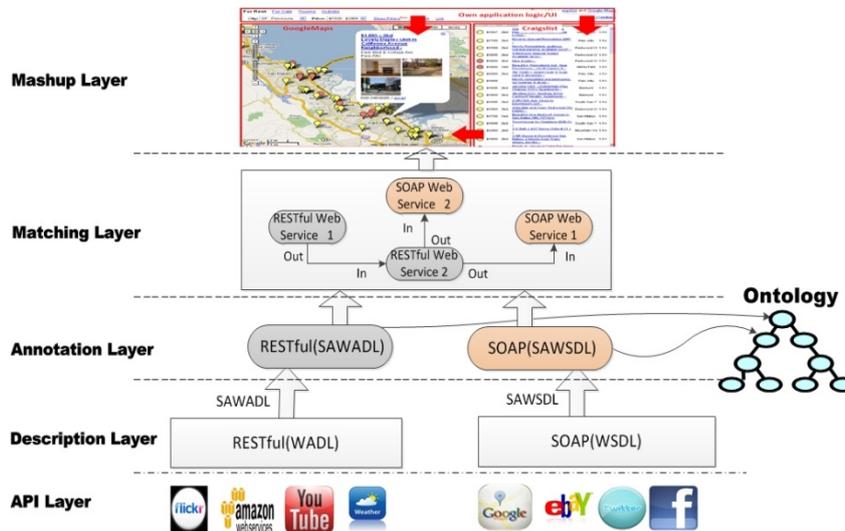


Fig. 5. Reference architecture of a Semantic Mashup .

4.1 API Layer:

At this level several types of APIs are concerned. In particular APIs based on SOAP and RESTful which are the most widely used in engineering applications Mashups.

4.2 Description Layer:

At this layer, WADL and WSDL languages are used respectively to describe SOAP and REST APIs.

4.3 Annotation Layer:

In addition to the SAWADL language that we propose in this paper, several languages of web services annotation are considered at this level. In particular, SAWSDL which is used to semanticize SOAP-based web services by annotating the input/output of WSDL file with ontological concepts. This layer will be used in the automatic construction of Mashups, by allowing discovery, selection and combination of unambiguous way of the various APIs.

4.4 Matching Layer:

The heterogeneities between different annotation languages are resolved at this layer. In the following, we propose four rules to match SAWSDL and SAWADL annotation languages.

Rules1. A method described by the tag "<method>" of a resource or a sub-resource "<resource>" of a SAWADL file corresponds to an operation described by the tag "<operation>" of a SAWSDL file.

Rules2. An input described by the tag "<param>" for a set of inputs "<request>" of a SAWADL file corresponds to an entry described in the web service's XML schema by the tag "<element>" of a "<complexType>" of an operation's Input described in SAWSDL file.

Rules3. An output described by the tag "<response>" of a SAWADL file corresponds to an output described in the web service's XML schema by the tag "<element>" of a "<complexType>" of an operation's output described in SAWSDL file.

Rules4. The "modelReference", "liftingSchemaMapping", "loweringSchemaMapping" attributes of a SAWADL file correspond to the "modelReference", "liftingSchemaMapping", "loweringSchemaMapping" attributes of a SAWSDL file.

Correspondences between APIs are established based on of semantic similarity [7] which allows calculating a distance between the ontological concepts of Input/Output. This distance will be compared with a predefined threshold in order to know if an API could be combined with another or not.

The matching score between a pair of matching services S_{in} and S_{out} is calculated using the following formula:

$$Match(S_{in}, S_{out}) = 2 * hi * \sum (1 - dist(i, j)) / (n_{in} + n_{out})$$

Where n_{in} is the number of query attributes of the service S_{in} And n_{out} is the number of annotated attributes present in service S_{out} , hi is the number of annotated attributes of services S_{out} that have been matched out of n_{in} , and finally $dist(i, j)$ is the ontological distance score between the j^{th} term in service S_{out} and a corresponding query term.

4.5 Mashup Layer

At this layer, an application mashup is really created based on the results obtained by matching layer. The Mashup layer integrates APIs that have a matching value greater than or equal to a threshold predefined by domain experts. The combination of APIs can be made using different technologies (e.g. Ajax, PHP, JSP, etc.).

5 Conclusion and perspective

The Mashups are web applications developed by combining data, business processes, and/or user interfaces of web sources published and reused via APIs. Thus, Mashups aimed at reducing the cost and development time of web applications. However in order to address the shortcomings of existing languages and protocols established by the IT community, we saw that the work related to engineering the Mashups applications are particularly oriented towards the semantic level.

The aim through the use of semantics is to enable machines to interpret the processed data and seize their significance in an automatic way in order to automate the selection and combination of APIs used to build the Mashup application.

Many languages and semantic annotations have been proposed for the semantic description of RESTful APIs. However, they did not give a great success and are not simple to implement. For example, SA-REST and SWEET approaches require an HTML web page that describes the API and that will be later transformed into a machine readable description to add semantic annotations. One thing that is not always true and that makes it more difficult especially if the REST API does not have a web page that describes it. In order to respond to these problems, that we conducted our research. Our work focuses on the semantics, and more particularly towards a proposal for an annotation language for semantic REST Web services. Our language SAWADL is one of the approaches that add semantic annotations on top of the service description. Unlike approaches that annotate on top of an HTML description, we use the WADL description which is used to describe syntactically REST web services. Semantization APIs is not sufficient to design and implement an automatic Mashup. Thus a process of matching is necessary to find correspondences between the different APIs, and to discover automatically the Mashable components followed the needs of users.

Finally, several perspectives can be considered in order to contribute more to the agility and flexibility of the semantic Mashup building. We cite as an examples:

- The Semantization of other Web APIs such as javascript or RSS / ATOM that represent Mashable components widely used in the development of Mashups. However, the absence of a structured and modular description of these APIs makes this task a big challenge.
- The use of ontological resource and service like OWL-S and WSMO.
- The use of the semantic approach in the construction of process-oriented enterprise Mashups that allows a user to automate her tasks.

REFERENCES

- [1] R.Fielding, Architectural Styles and the Design of Network-based Software Architectures, PhD thesis, University of California, 2000.
- [2] J.Kopecký, T.Vitvar, C.Bournez, J.Farrell: SAWSDL: Semantic Annotations for WSDL and XML Schema, IEEE Internet Computing, vol. 11, no. 6, pp. 60–70, November-December 2007.
- [3] J.Kopecky , K.Gomadani, T.Vitvar: hRESTS: an HTML Microformat for Describing RESTful Web Services , Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence (WI-08), 2008.
- [4] J.Latham, K.Gomadani, P. Sheth; SA-REST and (S)mashups Adding Semantics to RESTful Services , Proceedings of the First IEEE International Conference on Semantic Computing (ICSC 2007), September 17-19, 2007, Irvine, California, USA. IEEE Computer Society 2007.

- [5] M.Maleshkova, C.Pedrinaci, J.Domingue, Supporting the Creation of Semantic RESTful Service Descriptions, 2009, In: 8th International Semantic Web Conference (ISWC 2009), 25-29 Oct 2009, Washington D.C., USA.
- [6] G.Meditskos, N. Bassiliades , A combinatory framework of Web 2.0 mashup tools, OWL-S and UDDI, Expert Systems with Applications, vol. 38, no. 6, pp. 6657–6668, June 2011.
- [7] H.Ngu Anne, P.Carlson Michael, Z.Quan Sheng. Semantic-based Mashup of Composite Applications, IEEE Internet Computing, vol. 3, no. 1, pp. 2–15, January-March 2010.
- [8] J.Yu, B.Benatallah, F.Casati, F.Daniel. Understanding Mashup Development and its Differences with Traditional Integration, , IEEE Internet Computing, vol. 12, no. 5, pp. 44–52, September-October 2008.

An approximation approach for semantic queries of naïve users by a new query language

Ala Djedjai, Hassina Seridi-Bouchelaghem and Med Tarek Khadir

LABGED Laboratory, University Badji Mokhtar Annaba, Po-Box 12, 23000, Algeria

{djedjai, seridi, khadir}@labged.net

Abstract. This paper focuses on querying semi structured data such as RDF data, using a proposed query language for the non-expert user, in the context of a lack knowledge structure. This language is inspired from the semantic regular path queries. The problem appears when the user specifies concepts that are not in the structure, as approximation approaches, operations based on query modifications and concepts hierarchies only are not able to find valuable solutions. Indeed, these approaches discard concepts that may have common meaning, therefore for a better approximation; the approach must better understand the user in order to obtain relevant answers. Starting from this, an approximation approach using a new query language, based on similarity meaning obtained from WordNet is proposed. A new similarity measure is then defined and calculated from the concepts synonyms in WordNet, the measure is then used in every step of the approach for helping to find relations between graph nodes and user concepts. The new proposed similarity can be used for enhancing the previous approximate approaches. The approach starts by constructing a graph pattern (*GP*) from the query and finalized by outputting a set of approximate graph patterns containing the results ranked in decreasing order of the approximation value level.

Keywords. Graph matching, RDF, Naïve user, Graph pattern, Semantic Queries, Regular Path Queries, Approximation, Similarity, Ranking and WordNet

1 Introduction

In recent years, the amount of information on the web grows increasingly and the classic information retrieval is not able to find the answer which satisfies the user queries, therefore, the semantic search may be a proposed solution for such situations. Most users have not much knowledge about the querying language in the semantic web, they are not aware of target knowledge base; so the user query does not match necessary the data structure. It is very hard and difficult to understand intend of naïve users.

In this paper we propose an approach for answering a new query language inspired from the conjunctive regular path queries [1], the user query is transformed to a graph pattern. We use a new method to calculate the approximation level between the paths of the graph data and the query paths; approximation is enhanced using the

WordNet database so the method is based on a proposed meaning similarity between concepts from WordNet

We consider the problem of querying the semi-structured data such RDF data which is modeled by a graph $G = (V, E)$ and an ontology $O = (V_o, E_o)$. Where each node in V is labeled with a constant and each edge e is labeled with a label drawn from a finite set of symbols S , V contains nodes representing entity classes or instances or data values (values of properties), the blank nodes are not considered, the edges between the class nodes and the instance nodes is labeled by 'type', E represents the relations between the nodes in V , $V \subset V_o$ and $E \subset E_o$.

Users specify their request by a proposed language inspired from the conjunctive regular path queries CRP which have the next format:

$$Q : q(X_1 \dots X_n) :- (Y_1 R_1 Z_1), \dots, (Y_n R_n Z_n) \quad (1)$$

- Each Y_i or Z_i is a variable or a constant. The variable is specified by? , we make a simple modification to the constants for specifying the choices so the user is able to specify constants which are not necessarily appearing in G and he is able to use many constants by using the symbol '|', so Y_i or Z_i is a variable or a constant or expression (in our approach).
- Regular path expressions $\{R_1, \dots, R_n\}$, which are defined by the grammar:

$$R: = \epsilon \mid a \mid _ \mid L \mid (R_1.R_2) \mid (R_1|R_2) \mid R^*, \quad (2)$$

Where ϵ is the empty string, "a" is a label constant, " _ " denotes any label and L is a label variable.

- $X_1 \dots X_n$ are head variables and the result is returned in these variables.

In this paper, for helping the naïve users, we propose a new simple query language, we focus on the regular expression which has a simple format (using only the '.' and the '|'), the query Q_1 is an example of the proposed language, We construct from the user query a graph patterns GP for finding a set of sub graphs in G (approximate graph patterns) whose nodes matches the nodes in GP and its paths have a level of approximation to the paths in GP .

Example1. We assume that a user writes a query Q_1 for finding the publications and the authors in 'California' university or 'Harvard' university in the 'ESWC 2012' conference:

$$\begin{aligned} (? pub, ? author) : - & (? pub, writer, ? author), \\ & (? author, his_univ.name|location, California|Harvard), \\ & (? pub, conf.name, ESWC2012). \end{aligned}$$

Figure 1 shows a GP constructed from Q_1 , the separate points between symbols represented by non-labeled nodes, the query paths 1 2 3 correspond to user paths 1 2 3 of Q_1 . The variable nodes are specified with '?' to indicate that only these nodes are shown in the answer. In our work, the answers for the query is a set of approximated graph patterns ranked in order of decreasing the approximation level value, every one contains nodes that correspond to the user variables, the paths in every approximate graph pattern are an approximation of the paths in GP (every path in GP is corresponding to a single conjuncts query [4]). We use the graph patterns as answers, for

giving to the user the ability to explore the results for more information about the result nodes.

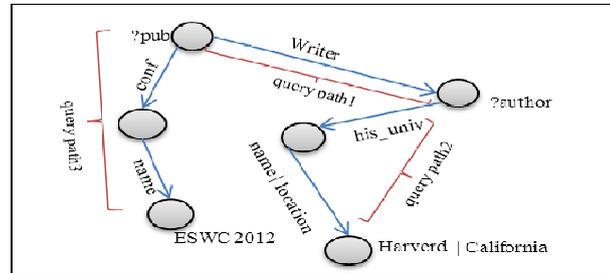


Fig.1. A graph pattern GP constructed from Q1

In section 2 related works are discussed and Section 3 presents WordNet and the new proposed similarity meaning. In section 4 the approximation approach is detailed. Section 5 is dedicated to the approach implementation and experimentation, whereas the conclusion and future works are presented in section 6.

2 Related works

Many approaches, methods and query language are proposed for the search in the semantic web search, and may be classified as follows:

1. Approaches consider structured query languages, such as: Corese [9], Swoogle [11] and ONTOSEARCH2 [4].
2. Approaches for naive users, these approaches can themselves be divided into:
 - ✓ Keyword-based approaches, such as QUICK [8], where queries consist of lists of keywords;
 - ✓ Natural-language approaches, where users can express queries using natural language, such that PowerAqua [10].

In this work we are interested by using the regular path queries with simple regular expression, this helps the naive users to use the query language as they are able to write simple regular expression. Our approach combines the two previously cited classes, so the naïve user queries the system using simple structure and the user constants are seen as keywords. Many works are proposed for the approximation such as [1] and [2], where the approximation is applied to the conjuncts queries. The ISPARQL [3] is a similarity based approach which added the notion of similarity to SPARQL queries, where another technique in [7] calculates the approximate answer from RDF graph using an evolutionary algorithm. Despite their efficiency, the approaches discard the user influence and opinion. The obtained results do not, therefore, often satisfy the latter. In addition to the above approaches, our work propose a new query language inspired from conjunctive queries, using a technique for the approximation based on meaning similarity from WordNet for a better understanding of the user query as well as finding the correspondences between its concepts and the graph data. The answers are a set of approximate graph patterns ranked in decreasing

order approximation level so the user can explore these results in order to acquire more knowledge.

3 Using WordNet

WordNet [5] is a lexical resource for the English language; it groups terms (name, verbs, adjectives etc.) in sets of synonyms called Synsets. Approaches based on characters strings become insufficient when concepts are systematically close to each other and when their names are different (example: « car » and « automobile »), the interrogation of a linguistic resource such as WordNet may indicates that two concepts are similar . For the calculation of the linguistic similarity, the function $Syn(c)$ calculates the set of WordNet Synsets related to the concept c .

3.1 Definition of a new WordNet Meaning Similarity

In this section we define a new WordNet meaning similarity, this measure is used in the process of discovering the nodes mapping from the user query and graph data.

Let $S_{com} = Syn(c1) \cap Syn(c2)$ the set of common senses between $c1$ and $c2$ to be compared, the cardinality of S_{com} is : $|S_{com}| = |Syn(c1) \cap Syn(c2)|$, we use the following measure:

Let $\min(|Syn(c1)|, |Syn(c2)|)$ be the minimum cardinality between the two sets $Syn(c1)$ and $Syn(c2)$ for the concept $c1$ and $c2$ respectively, thus our similarity measure is constructed from analyzing of the next metric [7]:

$$Sim1(c1, c2) = \frac{|S_{com}|}{\min(|Syn(c1)|, |Syn(c2)|)} \quad (3)$$

This metric based on common senses of $c1$ and $c2$, it return 1.0 if $c1$ is synonym of $c2$, but if the set of senses for $c1$ (or $c2$) are including in the set of senses of $c2$ (or $c1$) so this metric return again 1.0, for example the concept “machine” has 8 senses and “motorcar” have 1 sense (included in the 8 sense of “machine”), utilizing this metric:

$Sim1(machine, motorcar) = \frac{1}{\min(8,1)} = \frac{1}{1} = 1$, so “machine” is the synonym of “motorcar” but this is wrong because “machine” is the generalization of “motorcar”, so from this idea we propose the next new measure which is based on the different senses between two concepts:

Let $S_{dif} = ((Syn(c1) \cup Syn(c2)) - (Syn(c1) \cap Syn(c2)))$: the set of different senses between $c1$ and $c2$, so $S_{Dif} = union - intersection$. $|S_{dif}| = |(Syn(c1) \cup Syn(c2))| - |(Syn(c1) \cap Syn(c2))|$, the set of union is defined as: $U = Syn(c1) \cup Syn(c2)$, our metric is:

$$Sim2(c1, c2) = 1 - \frac{|S_{dif}|}{|U|} = 1 - \frac{|union| - |intersection|}{|union|} \quad (4)$$

$$Sim2(c1, c2) = 1 - \frac{|S_{dif}|}{|U|} = 1 - \frac{|U| - |S_{com}|}{|U|} \quad (5)$$

If $Syn(c1) = Syn(c2)$ (no different senses $c1$ is synonym of $c2$) then $Sim2(c1, c2) = 1 - 0 = 1$. $Sim2(machine, motorcar) = 1 - \frac{7}{8} = 1 - 0.87 = 0.13$ (7 common senses).

In this paper we use the next measure which takes advantage of Sim1 (common senses) and Sim2 (deferent senses):

$$Sim_senses(c1, c2) = \grave{u}1 * Sim1 + \grave{u}2 * Sim2 \quad (6)$$

where, $\grave{u}1$ and $\grave{u}2$ are the widths associated to Sim1 and Sim2 respectively, $\grave{u}1 = 0.5$, $\grave{u}2 = 0.5$ by default i.e. same importance, we adjust $\grave{u}1$ and $\grave{u}2$ according the preference of the user.

Example 2. Table 1 shows values of similarity for some pair of concepts. We cannot find a significant similarity between these concepts if we use a metric based on syntax only, the *Levenshtein* similarity indicates that “house” and “mouse” are similar but this is wrong, this highlights the importance of the proposed measure as it is used to find relationships between terms of the semantic regular path queries and the nodes of the graph data.

<i>Concept1</i>	<i>Concept2</i>	<i>Sim1</i>	<i>Sim2</i>	<i>Sim_sense</i>	<i>Levenshtein</i>
<i>Car</i>	<i>Automobile</i>	0.5	0.16	0.33	0.0
<i>Location</i>	<i>Placement</i>	0.33	0.16	0.245	0.22
<i>House</i>	<i>mouse</i>	0.0	0.0	0.0	0.86

Table 1. Some similarity values calculated using *Sim_sense* and *Levenshtein*

4 Approximating the naïve user queries

We start by defining the problem that is: how to satisfy the user in case if he specifies concepts that do not exist in the graph data? This is a big difficulty, as the approximation is the solutions for finding results and approximating the user query. However, it must take into account the concept meaning, this is the goal of the new proposed query language and the meaning similarity. This helps to better understand the user and helps the discovery a set of concepts in the structure which are relevant to user concepts in order to begin the process of exploration and finding the responses for the variables.

The proposed approach may be divided in three steps:

- 1- Discovering nodes which correspond to discovering user concepts in *GP*.
- 2- Finding for every query path its approximate paths in the graph data.
- 3- Generation of the results which are a set of approximate graph patterns with its approximation level value, these graph patterns contain the nodes results corresponding to the projection of the user variables.

The procedure is based on the following objectives:

- ✓ Giving to ability to the naïve user to take advantage from the power of semantic search, in this case we let him specify his needs by writing simple regular paths.
- ✓ Understanding the naïve user query by finding relationships between the user paths and the knowledge base (RDF graph). Most user concepts do not appear in the structure, for this reason, we propose a new query language and a meaning similarity leading to a better understanding of the user needs on one hand and discovering the correspondences between the query concepts and the graph nodes on the other hand. The user, however, still plays an important role in the query answer paradigm.
- ✓ The outputted answer must be understandable for the user and it should be simple.

We make clear the procedures have been omitted, in the rest of the paper, because of pages limitation; we cannot describe the approach in detail so only the main steps are mentioned.

4.1 Mapping from Nodes in GP to Nodes in G

The mapping process is necessary to find the correspondences of the nodes in GP (variables and constants in the conjuncts query); these nodes are used for finding the set of the approximate paths in G . Because the user have lack knowledge of the graph data structure so he is able to use concepts not necessarily appearing in the graph and the process of mapping is important for discovering the nodes matches these concepts using WordNet. In order to enhance the matching we use a similarity metrics based on syntax (characters strings) (like: Levenshtein, NGram, JaroWinkler) and our meaning similarity (using the WordNet ontology) for discovering the senses (the meaning) commons between the concepts.

Definition 1. Two concepts $c1 \in GP$, $c2 \in G$ are similar if $Sim_senses(c1, c2) > T$ (WordNet similarity), T is predefined threshold, if $Sim_senses(c1, c2) = 0$ then we test $sim_synt(c1, c2) > T$, the values of Sim_senses , Sim_synt and T is defined in $[0, 1]$.

Sim_senses and Sim_synt (any syntax similarity) use the labels of nodes and edges. In the rest of the paper we use $sim(c1, c2)$ for the value returned by Sim_senses or Sim_synt .

For finding the sets of node mapping the procedure *get_nd_map* returns for every node $n_i^j \in GP (i \in \{1, 2\})$ i.e. the first or the last node in the query path QP_j , the set $NdMap_i^j$ contains the nodes in G which are similar to n_i^j using its label by the similarity based senses (or based syntax), in addition this procedure use a strategy for discovering another nodes in G from the first and last edge in the query path QP_j .

4.2 Computing the Approximate Paths

In this section we introduce the notion of approximation level between two paths and describe the method for calculating its values apx_lev , this section is for the computation of the approximate paths from G , the finale answers (approximate graph patterns) are calculated in the next sub section. This calculation is started after the generation of the set of nodes mapping $NdMap_i^j$ for every node $n_i^j \in QP_j$. The procedure *get_apx_paths* take as input a query path QP_j and outputting the set of tuples answer tup_ans_j , every tuple (V_1, V_2, p, apx_lev) containing two node : V_1 is first node in the approximate path p , V_2 is the last node and ap_lev is the value of the approximate level between QP_j and p , the sets of tuples answer are used for constructing the approximate graph patterns for GP .

We consider the next points in the calculation process of ***apx_lev*** :

- The number of edges in p similar to the edges in QP_j (similarity $\neq 1$), each similar edge in p is a non-direct substitution for its corresponding edge in QP_j so we added the value of substitution to apx_lev .

- The number of additional edges in p ($cost_{ad}$), not appearing in QP_j , each additional edge in p is an insertion.
- We also take into account the two values: similarity value between the first node in p and the first in QP_j , similarity value between the last node in p and the last in QP_j .
- The order of edges in the query path QP_j for respecting the preference of the user.

Our approach considers common and similar edges, therefore common edges are not associated to a value of '0' but '1', as well as the similarity values for similar edges. Before starting the process of finding the approximate local answer, the procedure get_apx_paths generates the set of all paths $ApxPaths_j$ from the two sets of nodes mapping $NdMap_1^j$ and $NdMap_2^j$.

Definition 2. Let p a path in G , QP a query path in GP , p is an approximate path for QP if the value of the approximation level between p and QP is higher than T_{apx} (predefined threshold of approximation), $T_{apx} \in [0,1]$.

The procedure get_apx_path use the similarity obtained between two nodes v_1 and $v_2, v_1 \in QP_j, v_2 \in path p$. If v_1 is labeled with more than one term by the symbol ']' so all terms are compared to the label of v_2 and only one value of similarity is returned i.e. the *MAX* value.

Example 3. Figure 2 shows the computation of the approximation level apx_lev for the paths $P \in G$ and $P' \in G$ for the query path QP_3 . For $QP_3 \in GP$: the first node n_1^3 is labeled with the variable '?pub' and has the set of nodes mapping $NdMap_1^3 = \{publication, pub1, pub2\}$, the last node is labeled with constant 'ESWC2012' and has the set of nodes mapping $NdMap_2^3 = \{ESWC2012, ISWC2012\}$. The similar edges by discontinued line, additional edges by double line, first and last nodes by the dark circle; the values of similarity between edges and nodes are in italic,). The common edges are represented by single continued line. In the path P , number of similar or common edges is 2 (with two values of similarity: 0.95, 1), $sim1 = sim(?pub, pub2) = 0.90$, $sim2 = sim(ESWC2012, ESWC2012) = 1$, the approximate level associated with P is:

$$apx_lev = \left(\frac{\sum val_{similarity}}{nb\ val_{similarity}} + \left(1 - \frac{cost_{ad}}{p.length}\right) + sim1 + sim2 \right) / 4 \quad (7)$$

$$apx_lev = \left(\frac{0.95+1}{2} + \left(1 - \frac{0}{2}\right) + 0.90 + 1 \right) / 4 = 0,97 \quad (8)$$

The tuple answer corresponding to P is: $(pub2, ESWC2012, P, 0,97)$.

In the path P' there is one additional edge the (the edge *type*) so $cost_{ad} = 1$ and, number of similar or common edges is 2 (with two values of similarity: 0.95, 1), $sim1 = sim(?pub, publication) = 0.70$, $sim2 = sim(ESWC2012, ISWC2012) = 0.20$, so the approximate level associated with P' is :

$$apx_lev = \left(\frac{0.95+1}{2} + \left(1 - \frac{1}{3}\right) + 0.70 + 0.20 \right) / 4 = 0,64 \quad (9)$$

The tuple answer corresponding to P is: $(pubpublication, ISWC2012, P', 0,64)$.

apx_lev for P is greater than apx_lev for P' so the path P have a good approximation than P' .

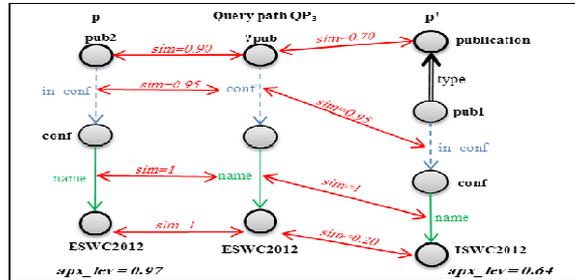


Fig. 2. Computing the approximation level apx_lev for the paths P and P'

4.3 Computing the Approximate Graph Patterns

In this section we describe how the final answers (Approximate graph patterns) are computed from the approximate paths discovered in the previous step, The final answers for the approximation is returned in form of tuples and every tuple represented by $(v_1, v_2, \dots, v_n, apx_gp, apx_gp_lev)$, where v_1 to v_n are the nodes corresponding to the nodes variable in GP (the nodes answers corresponding to the variables in the user query). apx_gp is the approximate graph pattern constructed from the approximate paths returned in tup_ans_j and apx_gp_lev is the approximation level between GP and $ApxGP$ i.e. the mean of the values apx_lev of the approximate paths used for the construction of apx_gp .

In [1] and [2], the final answers are a set of nodes corresponding to the variable in the query, in addition; as our approach based on graph patterns, a graph pattern with each nodes result is returned for a better answers understanding.

For computing the final answer we must generate the set of tuples answer tup_ans_j for every path $QP_j \in GP$, exploring the paths in every tuple and combining same paths for the generation of the graph patterns answer $\in G$.

Definition 3. Let GP a graph pattern constructed from a regular conjunctive query Q , Let GP' a graph pattern $\in G$. GP' is an approximate graph pattern for GP , if the value of approximate level apx_gp_lev between GP and GP' is higher than T_apx_gp (pre-defined threshold of approximation for GP), $T_apx_gp \in [0,1]$.

The procedure $final_ans$ is called with the set tup_ans_1 and its first tuple. The procedure $final_ans$ explores all tuples in any tup_ans_j to generate all approximate graph patterns, added them in the final set $final_ans$ with its nodes variables and its approximation level. For the process of ranking, the value apx_gp_lev is used to rank the tuples in $final_ans$, the tuples are outputted ranked in a decreasing order.

5 Implementation and Experimentation

Our approach is implemented in Java and Jena API, we use JAWS (Java API For WordNet Searching) for the implementation of the proposed meaning similarity. The RDF data set used is a sub set from the *SwetoDblp* ontology which is large size ontology focused on bibliography data of Computer Science publications where the main data source is DBLP, it has 4 millions of triples. The used subset contains a collection

of books and its book chapters. For making the execution faster, an offline phase which contains: RDF triples normalization, (getting triples that are closely to the natural language), building 2 indexes, is computed in order to allow quick finding of the approximate paths. The thresholds T, T_{apx}, T_{apx_Gp} are automatically initialized and updated according the query structure, this update allows the reduction of the found answers number.

For experimentation purposes and because our query language is inspired from the conjunctive path queries for helping the naïve (non-expert) users, a query benchmark is created. The benchmark contains a set of queries, with different intends that are executed over the RDF subset. For every query, from the subset; we computed, manually, the set of the relevant solutions (RS) for evaluating *Precision* and *Recall*:

$$Precision = \frac{\text{the number of relevant solutions returned (included in RS)}}{\text{number of solutions found by the system}} \quad (10)$$

$$Recall = \frac{\text{the number of relevant solutions returned (included in RS)}}{\text{the number of solutions in RS}} \quad (11)$$

In comparison with SPARQL, our work can be used by a non-expert users and it allows specifying a query paths between variables and constants for a better understanding of the user intend. It is difficult for the naïve user to use SPARQL efficiently because its complexes structure. Table 2 includes some queries, used for the evaluation whereas Figure 4 shows the precision and recall for some queries, proving the effectiveness of the approach.

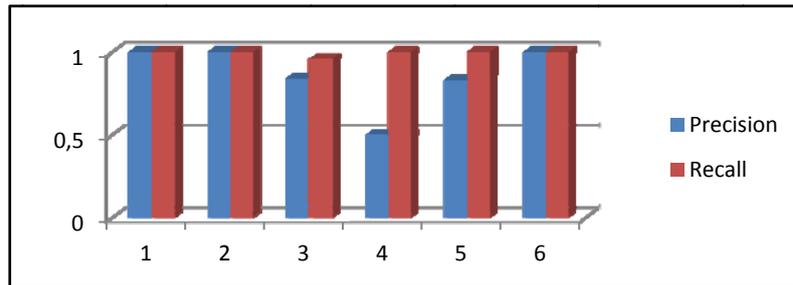


Fig. 3. Evaluation results for some queries

user query	Nb answers in RS	Nb system answers	Nb Relevant answers	Precision	Recall
(?Book_Chapter , title.contains, web) <i>User intend: Find all Book_chapters that have title contains «web »</i>	52	59	50	0.84	0.96
- (?Book_Chapter, book chapter included in the book, Prolog and Databases) - (?Book_Chapter, pages number, ? pages) <i>User intend: Find all Book_Chapters included in the book: «Prolog and Databases », associated with the pages number.</i>	20	20	20	1.0	1.0

- (?Book, year of publication, 2000) - (?Book, book isbn, ?isbn) - (?Book, has publisher, ?publisher) <i>User intend: Find Books published in 2000, associated with its isbn and the publisher.</i>	5	6	5	0.83	1.0
--	---	---	---	-------------	------------

Table 2. Some user queries used for the evaluation

6 Conclusion and Future Works

In this paper a novel approach for query approximation based on meaning similarity from WordNet is proposed, using a proposed query language inspired from the conjuncts queries. Using this technique, the naive users are able to write simple queries that not necessarily match the data structure. Our approach can be used as an extension to other approaches for a better understanding of the user query and obtaining results that satisfies the user's needs. It has been shown that the answers are a set of graph patterns ranked following the approximation level decreasing order. The work, is not considering only RDF graph but it can be seen as a general approach which may be applied to any semi-structured data that is modeled as graph, Future work will consist in applying the proposed approach to specific domains such as geographic, medical, biologic and bibliography, using query interface and building new indexes for scaling a huge number of triples.

References

1. A. Poulouvasilis and P. T. Wood Combining Approximation and Relaxation in Semantic web Path Queries. In Proc. ISWC, 2010.
2. C. A. Hurtado, A. Poulouvasilis, and P. T. Wood. Ranking approximate answers to semantic web queries. In Proc. ESWC, pages 263–277, 2009.
3. C. Kiefer, A. Bernstein, and M. Stocker. The fundamentals of iSPARQL: A virtual triple approach for similarity-based semantic web tasks. In Proc. ISWC, pages 295–309, 2007.
4. E. Thomas, J. Z. Pan, and D. H. Sleeman. ONTOSEARCH2: Searching ontologies semantically. In Proc. OWLED-2007, CEUR Workshop Proceedings 258. CEUR-WS.org, 2007.
5. Eyal Oren, Christophe Guéret, Stefan Schlobach. Anytime Query Answering in RDF through Evolutionary Algorithms, International Semantic Web Conference pp.98-113, 2008.
6. Fellbaum, C.: WordNet, an electronic lexical database. MIT Press, Cambridge (1998)
7. Fella, A., Malki, M and Zahaf, A., « Alignement des ontologies : utilisation de WordNet et une nouvelle mesure structurelle CORIA 2008 - Conférence en Recherche d'Information et Applications, Trégastel, France, 2008.
8. G. Zenz, X. Zhou, E. Minack, W. Siberski, and W. Nejdl. From keywords to semantic queries -Incremental query construction on the Semantic Web. J. Web Sem., 7(3):166–176, 2009.
9. O. Corby, R. Dieng-Kuntz, and C. Faron-Zucker. Querying the Semantic Web with Corese search engine. In Proc. ECAI-2004, pp. 705–709. IOS Press, 2004.
10. Lopez, V., Fernandez, M., Motta, E., Stierler, N.: PowerAqua: Supporting Users in Querying and Exploring the Semantic Web Content. Semantic Web Journal. IOS Press (2011).
11. T. W. Finin, L. Ding, R. Pan, A. Joshi, P. Kolari, A. Java, and Y. Peng. Swoogle: Searching for knowledge on the Semantic Web. In Proc. AAAI-2005, pp. 1682–1683.

Semantic Annotation of Web Services

Djelloul Bouchiha & Mimoun Malki

EEDIS Laboratory, Djillali Liabes University of Sidi Bel Abbes, Algeria.
bouchiha.dj@gmail.com, malki@univ-sba.dz

Abstract. Web services are the latest attempt to revolutionize large scale distributed computing. They are based on standards which operate at the syntactic level and lack semantic representation capabilities. Semantics provide better qualitative and scalable solutions to the areas of service interoperation, service discovery, service composition, and process orchestration. SAWSDL defines a mechanism to associate semantic annotations with Web services that are described using Web Service Description Language (WSDL). In this paper we propose an approach for semi-automatically annotating WSDL Web services descriptions. This allows SAWSDL Semantic Web Service Engineering. The annotation approach consists of two main processes: Categorization and Matching. Categorization process consists in classifying WSDL service description to its corresponding domain. Matching process consists in mapping WSDL entities to pre-existing domain ontology. Both categorization and matching rely on ontology matching techniques. A tool has been developed and some experiments have been carried out to evaluate the proposed approach.

Keywords. Annotation; Engineering; Web Service; Semantic Web Services; Ontology; SAWSDL; Ontology Matching Techniques; Similarity Measures.

1 Introduction

Web services are the latest attempt to revolutionize large scale distributed computing. They provide the means to modularize software in a way that functionality can be described, discovered and deployed in a platform independent manner over a network (e.g., intranets, extranets and the Internet). The representation of Web services by current industrial practice is predominantly syntactic in nature lacking the fundamental semantic underpinnings required to fulfil the goals of the emerging Semantic Web Services. SAWSDL defines a mechanism to associate semantic annotations with Web services that are described using Web Service Description Language (WSDL) [20]. The annotation process consists in relating and tagging the WSDL descriptions with the concepts of ontologies.

In this paper we propose an approach for semi-automatically engineering SAWSDL Semantic Web service from an existing Web Service and domain ontology. The proposed approach relies on an annotation process which consists in two phases: (1) Categorization phase, which allows classifying WSDL documents into their corresponding domain (2) Matching phase, which allows associating each entity from WSDL documents with their corresponding entity in the domain ontology. The annotation process relies on ontology matching techniques which in turn use some

similarity measures. An empirical study of our approach is presented to help evaluate its performance.

The remainder of paper is organized as follow: In section 2, we discuss some other efforts that describe adding semantics to Web services. In section 3, we present the proposed approach and its underlying concepts and techniques. An empirical study of our approach is presented in section 4 to help evaluate its performance. Finally, section 5 draws some conclusions.

2 Related Works

Several proposals have already been suggested for adding semantics to Web services, such as [18], [5], [6] and [4]. Other approaches concentrate on the Web service annotation: In a preliminary work Bouchiha and al., propose to annotate Web service with ontology using ontology matching techniques [21]. However, they focus on WSDL-S [1] instead of SAWSDL [20].

Table 1. Summary of Web service annotation approaches.

Approach	Considered elements	Annotation resource	Techniques	Tool
[22]	Operation parameters	Workflow	Parameter compatibility rules	Annotation Editor
[21]	Complex types and operations names	Domain ontology	Ontology matching	SAWSDL Builder
[8]	Operations, message parts and Data.	Domain ontology	Text classification techniques	ASSAM
[14]	Data (Inputs and Outputs of services)	Domain ontology	Schema matching techniques	MWSAF tool
[24]	Natural-language query	Domain Ontology	Text mining techniques	Visual OntoBridge (VOB)
[25]	Data (Inputs and Outputs of services)	Meta-data (WSDL)	Machine learning techniques	Semantic labelling tool
[23]	Annotation & Query	Workflow	Propagation method	Prolog Implementation
[26]	Datalog definitions	Source definitions	Inductive logic search	EIDOS

Table 1 summarizes the characteristics of the Web service annotation approaches as follow: (1) The "Approach" column corresponds to the approach in question; (2) The "Considered elements" column describes the considered elements in the annotation process; (3) The "Annotation resource" column indicates the model from which semantic annotations are extracted; (4) The "Techniques" column presents the used techniques for the annotation; (5) The "Tool" column indicates the tool supporting the approach.

3 Annotation approach

As shown in Fig 1, the annotation approach consists of two main processes: Categorization and Matching. Both categorization and matching rely on ontology matching techniques. The goal of ontology matching is to find the relations between entities expressed in different ontologies. Very often, these relations are equivalence relations that are discovered through the measure of the similarity between the entities of ontologies.

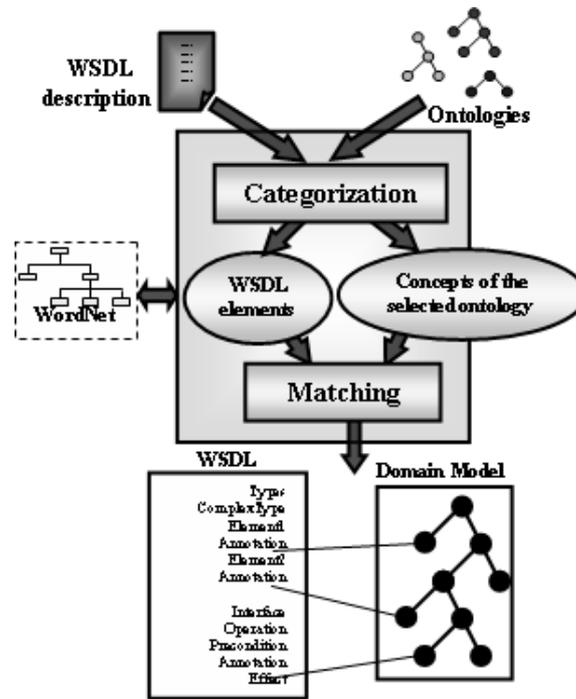


Fig. 1. The annotation approach.

To be accomplished, the ontology matching process uses similarity measures between entities. A similarity measure aims to quantify how much two entities are alike. Formally, it is defined as follow:

Definition 1 (Similarity): Given a set O of entities, a similarity $\sigma : O \times O \rightarrow \mathbb{R}$ is a function from a pair of entities to a real number expressing the similarity between two objects such that:

$$\forall x, y \in O, \sigma(x, y) \geq 0 \quad (\text{positiveness})$$

$$\forall x \in O, \forall y, z \in O, \sigma(x, x) \geq \sigma(y, z) \quad (\text{maximality})$$

$$\forall x, y \in O, \sigma(x, y) = \sigma(y, x) \quad (\text{symmetry})$$

In our approach, we use WordNet based similarity measures [16]. WordNet is an online lexical database designed for use under program control [13]. So, these measures are computed, and then normalized. Normalisation consists generally in inverting the measure value to obtain a new value between 0 and 1. The value 1 indicates that there is a full semantic equivalence between the two entities.

Similarity measures relying on WordNet can be classified into three categories: (1) Similarity measures based on path lengths between concepts: lch [11], wup [19], and path; (2) Similarity measures based on information content: res [17], lin [12], and jcn [7]; and (3) Relatedness measures based on relations type between concepts: hso [9], lesk [3], and vector [15].

When a set of ontologies are available, similarities between two sets have to be computed by comparing the set of entities of the WSDL file and the set of entities of each ontology. On the basis of such measures, systems will decide between which ontologies to run a matching algorithm. The chosen domain ontology determines the WSDL file category. This process is called the categorization process.

Our approach considers an ontology as a set of entities (concepts), and a WSDL file also as a set of entities (XSD data types, interface, operations, messages). Several strategies can be adopted for computing similarities between two sets. Next we define Single linkage, Full linkage and Average linkage strategies:

Definition 2 (Single linkage): Given a similarity function $\sigma : O \times O \rightarrow R$, the single linkage measure between two sets is a similarity function $\Delta : 2O \times 2O \rightarrow R$ such that:

$$\forall x, y \subseteq O, \Delta(x, y) = \max_{(e1, e2) \in x * y} \sigma(e1, e2)$$

Definition 3 (Full linkage): Given a similarity function $\sigma : O \times O \rightarrow R$, the complete linkage measure between two sets is a similarity function $\Delta : 2O \times 2O \rightarrow R$ such that:

$$\forall x, y \subseteq O, \Delta(x, y) = \min_{(e1, e2) \in x * y} \sigma(e1, e2)$$

Definition 4 (Average linkage): Given a similarity function $\sigma : O \times O \rightarrow R$, the average linkage measure between two sets is a similarity function $\Delta : 2O \times 2O \rightarrow R$ such that:

$$\forall x, y \subseteq O, \Delta(x, y) = \frac{\sum_{(e1, e2) \in x * y} \sigma(e1, e2)}{|x| * |y|}$$

Next we detail the two processes involved in our approach.

Categorization process. The categorization process aims to classify WSDL service description to its corresponding domain. For this end, the service description is broken down into its fundamental WSDL elements (XSD data types, interface, operations and messages). A list of concepts is also extracted from each ontology. Similarities between two sets based on similarity measure between two entities will be computed to identify which ontology concepts will be kept for the next process. The selected ontology indicates the WSDL domain or category.

We have developed an algorithm (see Listing 1) that implements the categorization process. The algorithm computes the similarity between a WSDL document and a set

of domain ontologies. A WSDL document belongs to the category of the domain ontology for which it gives the best similarity (the nearest ontology).

Listing 1. The Categorization algorithm.

```

Algorithm Categorization
Input
WSDL document
A set of domain ontologies
A similarity measure SM between two entities
A Similarity SD between two sets
Threshold
Output
An assigned WSDL document to a particular category
Begin algo
Filling a vector VE with the WSDL document elements
For each domain ontology Do
  Filling a vector VC with the domain ontology concepts
  For each element E of the vector VE Do
    For each element C of the vector VC Do
      // Next, Vector Sim is used to store the
      //Similarity between the two vectors VE and VC
      Switch SD of
        Single linkage : If (SM(E,C) > Vector Sim)
                          then Vector Sim • SM(E,C) End if
        Full linkage : If (SM(E,C) < Vector Sim) then
                       Vector Sim • SM(E,C)
                       End if
        Average linkage : Vector Sim • Vector Sim + SM(E,C)
      End switch
    End for
  End for
  If SD is Average linkage
    then Vector Sim • Vector Sim / (|VC| * |VE|)
  End if
  // Next, Final Sim is used to store Similarity
  //between VE and the nearest ontology
  If (Final Sim < Vector Sim )
    then Final Sim • Vector Sim
  End if
End For
If (Final Sim > Threshold )
  then the WSDL document is assigned to the corresponding
  ontology to the Final Sim
End if
End Algo

```

Matching process. The matching process aims to map WSDL elements to ontology concepts. Similarities between a WSDL element and the concepts of the selected ontology will be computed to identify which concept will be attached to the initial WSDL element. This operation is repeated for all WSDL elements.

We have developed an algorithm (see Listing 2) that implements the matching process. The algorithm computes the semantic similarities between WSDL document elements and domain ontology concepts. Each WSDL document element will be annotated by the nearest domain ontology concept.

Listing 2. The Matching algorithm.

```
Algorithm Matching
Input
WSDL document
A domain ontology
A similarity measure SM between two entities
Threshold
Output
An annotated WSDL document with a domain ontology concepts
Begin algo
Filling a vector VE with the WSDL document elements
Filling a vector VC with the domain ontology concepts
For each element E of the vector VE Do
  For each element C of the vector VC Do
    //Next, Entity Sim is used to store Similarity
    //between a WSDL element and the nearest
    //ontology concept
    If (SM(E,C) > Entity Sim) then Entity Sim • SM(E,C)
  End if
End for
If (Entity Sim > Threshold )
  then assign the element E to the corresponding concept
  of the domain ontology
End if
End for
End Algo
```

As result of the two algorithms, an annotated WSDL document will be generated.

4 Results and empirical testing

The algorithms presented above are generic and can be adapted to most domain model languages. The domain model language we have used is the OWL, but we believe that our results could be applied to any similar language. To evaluate and validate our approach a tool, called SAWSDL generator¹, has been developed. SAWSDL generator can be used to do semi-automatic annotations. It takes in a WSDL document which has to be annotated with a set of ontologies. It selects the best ontology for annotating the WSDL document and suggests most appropriate mappings for the XSD data types, interface, operations and messages in the WSDL file. The classification and matching are performed using ontology matching techniques. The tool produces annotated WSDL 2.0 file using extensibility elements and according to the SAWSDL recommendation [20].

To test our categorization algorithm we first obtained a corpus² of 424 Web services [8]. Although our initial intention was to test our algorithm on the whole corpus, we have limited our testing to one domain, due to lack of relevant domain specific ontologies. We are in the process of creating new domain ontologies and plan to extend our testing for remaining Web services in the future.

¹ <http://www-inf.univ-sba.dz/wsdl/>

² <http://www.andreas-hess.info/projects/annotator/ws2003.html>

The domain we have selected for testing is Business domain³. Although the ontology used is not comprehensive enough to cover all the concepts in this domain, they are sufficient enough to serve the purpose of categorization. We have taken a set of 31 services out of which 13 are from business domain, 13 from weather domain and 5 from the games domain.

As similarity measure, the path method has been used. It is defined as follow: For two entities e1 and e2, the similarity measure SIM can be given using the WordNet synsets (i.e. term for a sense or a meaning by a group of synonyms) based on the formula: $SIM(e1, e2) = 1 / \text{length}(e1, e2)$, where length is the length of the shortest path between two entities e1 and e2 using node counting.

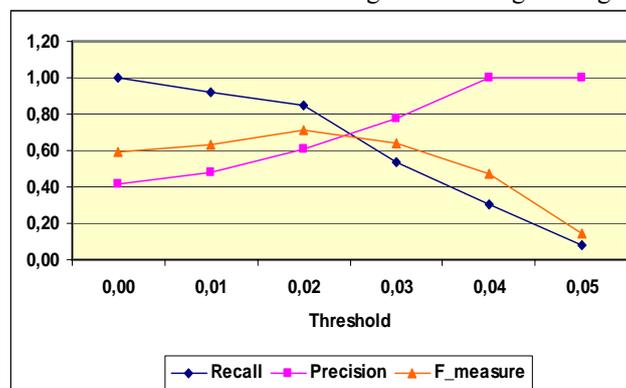
As in information retrieval [2], we use two metrics, Precision and Recall⁴, to evaluate the results of our algorithm of categorization.

- Recall (R): proportion of the correctly assigned WSDL documents of all the WSDL documents that should be assigned.
- Precision (P): proportion of the correctly assigned WSDL documents of all the WSDL documents that have been assigned.

Usually, Precision and Recall scores are not discussed in isolation. Instead, they are combined into a single measure, such as the F-measure [10], which is defined as follow: $F_measure = (2 * recall * precision) / (recall + precision)$.

The services are categorized based on the categorization threshold, which decides if the service belongs to a domain. If the best average service match calculated for a particular Web service is above the threshold then the service belongs to the corresponding domain.

Graph 1 depicts the corresponding curves to the precision, recall and f-measure statistics obtained by applying our categorization algorithm on this set of 31 Web services for different threshold values according to the average linkage strategy.



Graph 1. Precision, recall and f-measure curves for the categorization algorithm.

It is very important to choose the threshold value correctly. We can see from Graph 1 that for threshold = 0.02, which corresponds to the topmost value of the f-measure

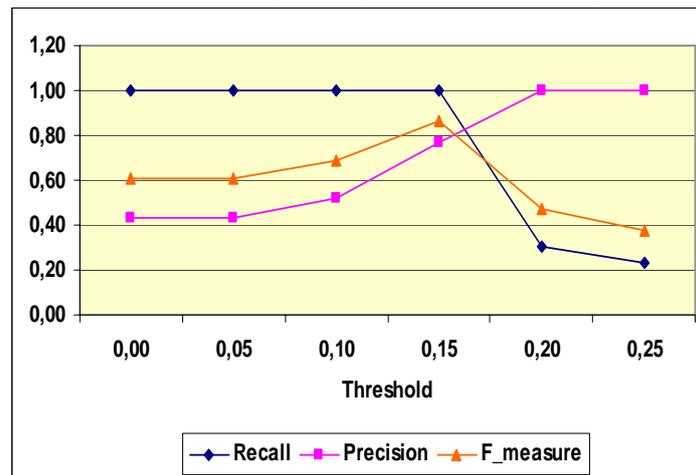
³ <http://www.getopt.org/ecimf/contrib/onto/REA/index.html>

⁴ http://en.wikipedia.org/wiki/Precision_and_recall

curve, gives the best categorization. However, even with the best threshold, some problems can appear. For example, The Web service "BasicOptionPricing" has not been rightly classified into the business domain, because it includes operations which have not meaningful names. Also, the two Web services "Weather Forecast By Zip Code" and "World Weather Forecast by ICAO" have been wrongly classified into business domain, although they belong to the weather domain. The reason behind this is that the two services include "Forecast" operations which can be shared between both business and weather domain.

To verify the fitness of the obtained result, a reference annotated WSDL document is considered as a valid. The chosen WSDL document was "TrackingAll". Now, to evaluate the quality of the matching algorithm, we compare the match result returned by our automatic matching process with manually determined match result in the reference WSDL annotated document. We determine the true positives, i.e. correctly identified matches.

Graph 2 depicts the corresponding curves to the precision, recall and f-measure statistics obtained by applying our matching algorithm on the chosen Web service for different threshold values according to the path measure similarity.



Graph 2. Precision, recall and f-measure curves for the matching algorithm.

Graph 2 shows that best results of the matching algorithm are obtained with threshold = 0,15. However, even with this threshold, a system user intervention is suggested for withdrawing some matching, or validating the result as it is generated. For example the WSDL elements "update_Company", "update_Customer", "update_Status" and "update_Tracking" have been matched wrongly to the concept "Agreement". The reason behind this is that the WSDL element names include the term "update" which has been treated by the system as name and not as a verb. As a name "update" means "news that updates your information". With a small threshold (<0,15), the user intervention is always necessary for keeping only right matching.

5 Conclusion

In order to harvest all the benefits of Web services technology, an approach has been proposed for annotating WSDL syntactic descriptions of Web services by ontological models. The benefits of such approach are twofold: Firstly, the approach provides a way to map WSDL descriptions to domain ontologies. Secondly, the approach enables the migration of syntactically defined Web services toward Semantic Web Services.

The proposed annotation approach consists of two main processes: Categorization and Matching. At the first process, WSDL service description is classified to its corresponding domain. At the second process the WSDL entities are mapped to pre-existing domain ontology. Both categorization and matching use WordNet based similarity measures.

A tool has been developed to implement the proposed approach. Some validation experiments have been carried out and they showed the usefulness of the proposed approach and highlighted possible areas for improvement of its effectiveness.

The developed approach provides very satisfactory and encouraging results and supports the potential role that this approach can play in providing a suitable starting point for SAWSDL semantic Web services development.

References

1. Akkiraju R., Farrell J., Miller J., Nagarajan M., Schmidt M-T., Sheth A., and Verma K., "Web service semantics – WSDL-S". Tech. rep., W3C. <http://www.w3.org/Submission/WSDL-S/>. 2005.
2. Baeze-Yates R., and Ribeiro-Neto B., "Modern information retrieval", Addison-Wesley, ACM Press, Reading, MA. 1999.
3. Banerjee S., and Pedersen T., "Extended gloss overlaps as a measure of semantic relatedness". In Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence. Pages: 805-810. 2003.
4. Bell D., de Cesare S., Iacovelli N., Lycett M., and Merico A., "A framework for deriving semantic web services". Information Systems Frontiers. Volume 9, Number 1, Pages: 69-84. 2007.
5. Bouchiha D., and Malki M., "Towards re-engineering Web Applications into semantic Web services". The first International IEEE Conference on Machine and Web Intelligence (ICMWI'2010). Algeria, Algiers. 2010.
6. Buitelaar P., and Gmbh D., "Ontology learning for semantic Web services". In Proceedings of ONLINE2003, Düsseldorf, Germany. 2003.
7. Jiang J., and Conrath D., Semantic similarity based on corpus statistics and lexical taxonomy. In Proceedings on International Conference on Research in Computational Linguistics, Pages: 19-33. 1997.
8. Hess A., Johnston E., and Kushmerick N., "ASSAM: A tool for semi-automatically annotating semantic Web services". International Semantic Web Conference. Hiroshima, Japan. Pages: 320-335. 2004.
9. Hirst G., and St-Onge D., "Lexical chains as representations of context for the detection and correction of malapropisms". In Fellbaum, C., ed., WordNet: An electronic lexical database. MIT Press. Pages: 305-332. 1998.

10. Larsen B., and Aone C., "Fast and effective text mining using lineartime document clustering", Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Pages: 16-22. 1999.
11. Leacock C., and Chodorow M., "Combining local context and WordNet similarity for word sense identification". In Fellbaum, C., ed., WordNet: An electronic lexical database. MIT Press. Pages: 265–283. 1998.
12. Lin D., "An information-theoretic definition of similarity". In Proceedings of the International Conference on Machine Learning. 1998.
13. Miller G-A., "WordNet: An on-line lexical database". International Journal of Lexicography. Pages: 235-312. 1990.
14. Patil, S. Oundhakar, A. Sheth, and V. Kunal. "METEOR-S Web service annotation framework". WWW 2004, ACM Press. Pages: 553-562. 2004.
15. Patwardhan S., "Incorporating dictionary and corpus information into a context vector measure of semantic relatedness". Master's thesis, Univ. of Minnesota, Duluth. 2003.
16. Pedersen T., Patwardhan S., and Michelizzi J., "WordNet::Similarity - measuring the relatedness of concepts". Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI-04). Pages: 1024-1025. 2004.
17. Resnik P., "Using information content to evaluate semantic similarity in a taxonomy". In Proceedings of the 14th International Joint Conference on Artificial Intelligence, Pages: 448-453. 1995.
18. Sabou M., Wroe C., Goble C., and Stuckenschmidt H., "Learning domain ontologies for semantic Web service descriptions". Journal of Web Semantics. Volume 3, N 4. Pages: 340-365. 2005.
19. Wu Z., and Palmer M., "Verb semantics and lexical selection". In 32nd Annual Meeting of the Association for Computational Linguistics, Pages: 133–138. 1994.
20. Farrell J., and Lausen H., "Semantic Annotations for WSDL and XML Schema". W3C Recommendation, 28 August 2007. Available at <http://www.w3.org/TR/sawSDL/>. 2007.
21. Bouchiha D., Malki M., Alghamdi, A., and Alnafjan, K., "An Empirical Approach for Annotating Web Services". The 24th International Conference on Computer Applications in Industry and Engineering. Hawaii, USA. November 16-18, 2011.
22. Belhajjame K., Embury S-M., Paton N-W., Stevens R., and Goble C-A., "Automatic annotation of web services based on workflow definitions". ACM Transactions on the Web (TWEB journal). Number 2, Volume 2. 2008.
23. Bowers S., and Ludäscher B., "A calculus for propagating semantic annotations through scientific workflow queries". Query Languages and Query Processing workshop (QLQP-2006) anised in conjunction with the 10th International Conference on Extending abase Technology, pages 712-723. 2006.
24. Grear M., and Mladenic D., "Visual OntoBridge: Semi-automatic Semantic Annotation Software". In ECML PKDD 2009, Bled, Slovenia, September 7-11, 2009, Proceedings, Part II. LNAI 5782, pages 726-729, Springer-Verlag Berlin, Heidelberg. 2009.
25. Lerman K., Plangprasopchok A., and Knoblock C-A., "Automatically labeling the inputs and outputs of web services". In Proceedings of the National Conference on Artificial Intelligence (AAAI-2006). Boston, Massachusetts, USA. July 2006.
26. Carman M-J., and Knoblock C-A., "Learning Semantic Definitions of Online Information Sources". Journal of Artificial Intelligence Research. Volume 30, pages 1-50. 2007.

Semantic multimedia search: the case of SMIL documents

CHKIWA Mounira*, JEDIDI Anis*

** Université de Sfax, MIRACL Multimedia, InfoRmation systems and Advanced Computing Laboratory, B.P 242 Sakiet Ezziat 3021 Sfax, Tunisie*

m.chkiwa@gmail.com, jedidianis@gmail.com

Abstract. Since the first implementations of its principles, the Semantic Web presented a field of free work to ensure its integration and adaptation to the various domains of research. The application of Semantic Web technologies into the process of search in a collection of SMIL documents appears a promising initiative seen the evolution of this language through its various versions. In this paper we propose a semantic search tool in a collection of SMIL documents; this tool adopts a procedure composed of three modules: description, interrogation and representation of the results. We employ for the first module metadata commonly used to annotate information semantically, and for the second we solicit languages of Semantic Web such as RDF, OWL and SPARQL and seen the importance of collaboration of ontologies in the semantic description of multimedia resources, we also present the technique of concepts connection allowing to extend an initial ontology.

Key words: Semantic Web, SMIL document, ontology, metadata, SPARQL, RDF.

1 Introduction

Faced with a variety of types of multimedia resources universally existing in the web, the importance of a SMIL document is that it provides a new vision more structured, organized and controlled of diversified multimedia contents. A SMIL document could enhance a classic multimedia presentation by integrating tags and attributes that provide a recommended overlap of resources in order to facilitate the understanding of a specific idea. Seen the valued prospect of this multimedia language and the spread of Semantic Web technologies come the need to generate a semantic tool to search multimedia contents in a collection of SMIL documents. By the contribution of its theoretical principles and standardized technologies, the Semantic Web offers a new vision for the search operation considering the meaning of things more than

its syntactic shapes. The reminder of this paper is as follows; in the three first sections we present the SMIL language and the Semantic Web framework. Section 5 describes the related works and section 6 presents our contribution in this topic followed by the technical view of the developed tool and its functionalities in section 7. Finally, section 8 underlines some conclusions and future research lines.

2 The SMIL multimedia presentation

Primarily, multimedia is everything dealing with the combination of two or more of the following media: image, sound, text and video. The presentation of multimedia contents is based on three fundamental axes:

- Time axis which defines the temporal ordering and synchronization of different objects by a script or scenario time predefined.
- Spatial axis which defines the spatial distribution of different media (with the exception of audio component).
- Logical axis considering hierarchical decomposition of a multimedia document into parts and subparts, with one or more media for each part.

SMIL (Synchronized Multimedia Integration Language) is a W3C specification which allows creating structured multimedia presentations. Another axis is considered in SMIL document: hypermedia axis; but this depend on the fact that such document offer or not a way to interact. This axis offers to the user the possibility to control the temporal, spatial and logical dimension making a personalized execution of a document according to the user preferences. The scope covered by the SMIL language is above websites and offers a range of possibilities such as:

- Collecting in a single presentation contents may come from different servers.
- Creating multimedia documents with very small size unlike the conventional multimedia presentations thanks to its simple textual structure.
- Insert controls events (play, stop, go to ...) to create customized presentations based on user interaction which allows many ways to present the same document.

A SMIL document is structured in two main parts: head and body; figure1 shows more details about these parts:

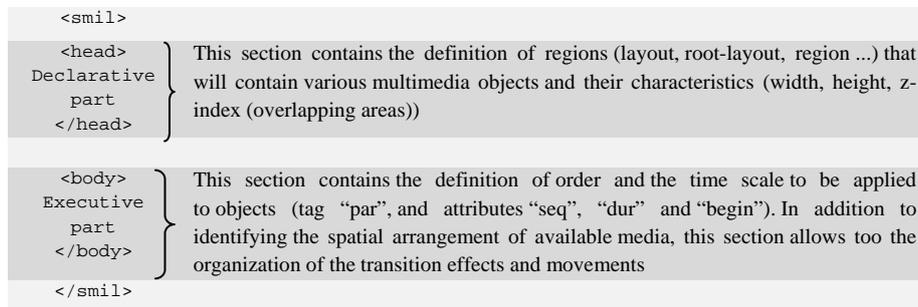


Fig.1. Structure of a SMIL document

3 Semantic Web

Accessible resources (images, text, audio, video ...) are at an earlier stage, formed by a set of documents, formatted in specific languages. These languages allow expressing the links between an object in the source document and another in a destination document. The Semantic Web is operated by software agents (browsers and search engines) browsing the links encountered. Metadata is the semantic descriptions of linked web contents; it's the global concept of Semantic Web which aims to yield semantic annotations to items accessible on the web even it is not a resource (i.e. image, text ...); it can be persons or associations... The overall vision of Semantic Web could be summarized in three fundamental points:

- Identifying resources universally (URI: Uniform Resource Identifier): We use URIs to identify pieces of information across the Web. The URI includes the "Uniform Resource Locator" (URL), the "digital Object Identifier (DOI) and the "International Standard Book Number" (ISBN).
- Describing the relationship between resources (RDF [1]: Resource Description Framework). It is a model for describing data on the web making automatic the access to sense of contents available on the web. Development of RDF has been motivated for several perspectives such as handling and defining semantic relationships between data (unlike primitive source/destination relationship).
- Extending the description of the properties of relations (OWL [2]: Web Ontology Language). The OWL provides to the Semantic Web syntax and semantics for automated reasoning about the inferences and implications of knowledge. In brief it's used generally to structure, share and exchange knowledge in universal format.

A major characteristic of a SMIL document compared with the rest of multimedia presentations is that it offers a structure clearly decomposable: components of a SMIL document (text, image, audio...) are each identified by URIs. Note that the decomposition is a fundamental operation preparing the annotation issue. Hence, the

components of a SMIL document are distinct, even pretend homogeneous during its presentation. This decomposability ensures the annotation of each element separately and we get rid of the classic problem of intricate partition of multimedia objects.

4 Semantic SMIL

The integration of Semantic Web in the information retrieval process has seen a great success expressed by the user satisfaction to the relevance of information returned after a typical search. This justified success allows to this technology to be larger than laboratories research and seek achieving prospects of general public in different fields. Hence the classic information retrieval process has been changed: the use of metadata become fundamental to annotate searchable resources. Thanks to the Metadata module, the SMIL language, performed many changes ensuring its integration to the Semantic Web view:

- The 1.0 version [3]: the “meta” element is used to define document properties (i.e. author, expiration date, key word list...) and provide values to these properties.
- The 2.0 version [4]: SMIL 2.0 extend SMIL 1.0 functionalities by the new element “metadata” which allow the use of RDF statement and make easier and more general the processing of metadata seen the ability of RDF to combine several standards of annotations as FOAF [5] and DC [6] in a single presentation.
- The 3.0 version [7]: the metadata module could be included in the body section of a SMIL document instead to be limited in the head section (as the previous versions). By this innovation we could make the description of an element right close to the definition of that element.

In figure 2 we present a set of metadata annotating an exemplar SMIL document containing the sections of this paper. In addition to the evolution of the language side to consider the semantic side of objects, further improvements are essential to a full exploitation of the principles of the Semantic Web in the context of search of SMIL documents: the use of ontology as a base of concepts composing the metadata set.

```
<rdf:Description about=http://exemple.com/article.smi
  dc>Title="Semantic multimedia search: the case of SMIL documents"
  dc>Date="2011-11-04"
  dc:Format="text/smil">
  <dc:Creator>
    <rdf:Seq ID="writen_by">
      <rdf:li>CHKIWA Mounira</rdf:li>
      <rdf:li>JEDIDI Anis</rdf:li>
    </rdf:Seq>
  </dc:Creator>
</rdf:Description>
```

Fig.2. Example of metadata set

5 Related Works

In the context of integration of Semantic Web technologies in the multimedia search process, many contributions are presented. Audiovisual documents cover a large range of multimedia contents commonly available such as television programs. In this topic, [8] propose a way to annotate semantically audiovisual documents by using Semantic Web languages in different levels:

- Using RDF to produce descriptions like: "the TV program" could have a "presenter" and the presenter is a "person". These descriptions seem more adequate to describe the structure and the content than the general conventional image annotation using low level techniques restricted on shapes of objects of "key frames" in an audiovisual sequence.
- Using the ontology of the audiovisual in order to formalize knowledge form descriptions, to express document patterns and to reuse those patterns in the description of documents process.

[8] uses also MPEG-7 describing technically the audiovisual resources to enrich semantic descriptions. Adopting MPEG-7 is suitable in this approach seeing its event-for features i.e., it can give details of the moment where something happens, people and even relations between objects in an emission.

In [9] we find an approach which aims to integrate a multimedia ontology into structured rich multimedia presentations such as SMIL, SVG, and Flash. The Multimedia Metadata Ontology M3O bases on Semantic Web technologies for representing sophisticated multimedia annotations. This ontology is represented in OWL; the annotations can therefore be represented in RDF, which can be directly embedded within formats such as SMIL or SVG. Note that these formats already provide appropriate means for embedding XML-based metadata.

The integration of SMIL documents in a "semantic" framework seems to [10] a way to present these type of multimedia content according to the user's preferences. Indeed, the semantics discussed in this context is the adaptation of SMIL documents in order to respect the limitations of the hardware platform display. [10] treated separately the spatial and temporal adaptation for SMIL documents whose textual structure allows any kind of software manipulation. Thus we can redistribute the components of a multimedia document in order to change their spatial arrangements or their moments display. We can say that the semantics discussed in this context seems more user-oriented than system-oriented: the "multimedia product" is packaged according to user preferences whereas the Semantic Web technologies promote the role of engines to automatically treat semantic information.

Although the studied reflections are close to our context (semantic search of SMIL documents), multimedia documents handled in some studies are unstructured unlike SMIL documents. In the topic of semantic multimedia search, some contributions [11] treat the multimedia issue as a vague item “collection of multimedia documents” whereas some others deals with multimedia types as distinct components such as the semantic search of images or semantic search of audio sequences. In the context of processing SMIL documents, other reflections remain restricted to a technical level as in the case of spatial/temporal adaptation of SMIL documents.

6 Our contribution

In the context of semantic search in a collection of SMIL documents, we propose a search procedure composed of three modules: the description of multimedia components, querying and reporting the results. In figure 3 we describe from a technical perspective, the proposed research process.

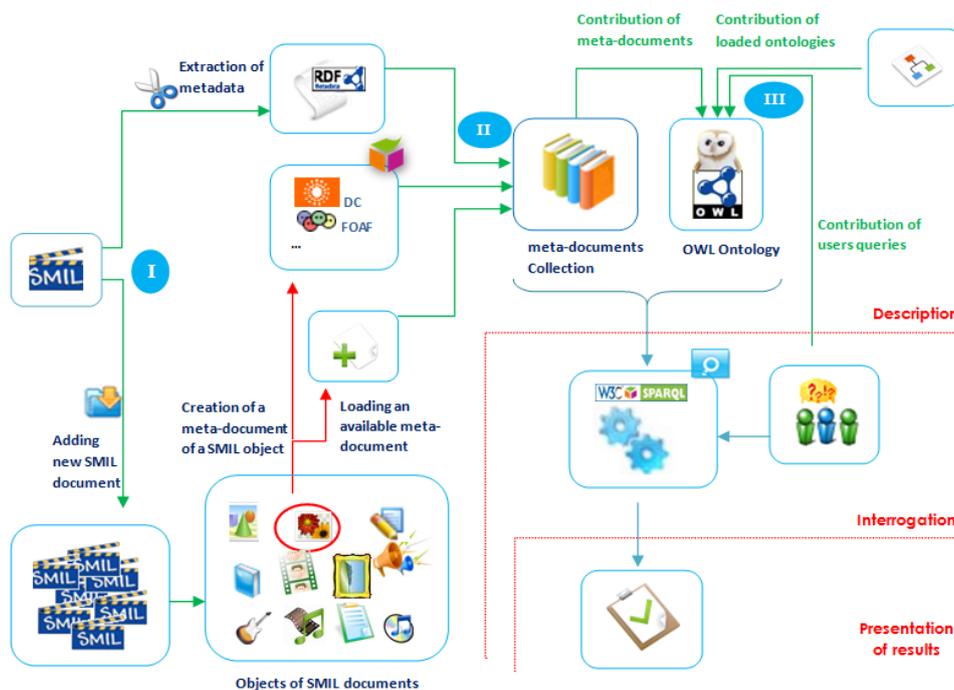


Fig. 3. Overall application architecture

The description of multimedia components is based on the transfer of a new SMIL documents to the collection, this operation is followed by an automatic process of control to check the document type and the eventually existence of an integrated

metadata. If a metadata part is found, it is extracted and assigned to a separated structure called a meta-document [12] describing the SMIL document or its components. The add operation (Part I of Figure 3) of the new SMIL is along with:

- Enumerating components existing in the SMIL document and the identification of its technical information (format, size, duration...).
- Duplicating components of the SMIL document using its URIs (specified in the code of SMIL document) and its transfer to the multimedia collection.
- A creation of a text copy of the SMIL document for further treatments.
- A record of all such information in the database for use in subsequent operations (description and the composition of a result of a query).

After adding a new SMIL document, the basic operation in the phase of description is the assignment of meta-documents [12] to the various components of SMIL documents. A meta-documents (part II of Figure 3) consists of a set of metadata, each bringing a different indication concerning the same multimedia component, take for example: the title, subject, creation date and creator of a piece of text existing in a SMIL document, all of this indications are encapsulated within the same structure : the meta-document. Compared to a traditional search process, the use of meta-documents in the phase of description brings several clear advantages:

- The description is selective: only significant items expressing the general idea of the document are described: (i.e. links images or background music are omitted).
- Provide to different types of multimedia components the same chance to be described and in this way the videos, sounds and images have the same level of expressiveness as a media text.
- A created meta-document could describe the same multimedia component existing in two or more different SMIL document which ensures its reuse.
- Offer a unified structure to annotate all multimedia components regardless their types.

The interrogation is the procedure triggered when submitting a query; user queries are categorized into three types:

- Simple query: a set of keywords query designed for the non-expert users.
- Advanced query: a set of parameters to be selected designed for more specific details and restrictions concerning the results.
- Experts query: Queries using SPARQL language oriented to the users knowing to use such language.

The interrogation allows extracting relevant information from a metadata set by comparing the query and the collection of annotations in meta-documents. The interrogation also aims to formulate and classify well the result satisfying a user need specified by the query. The classic interrogation way can consider items which are not reprehensive of a multimedia component for example when the description step extracts all the multimedia objects regardless of their value (sky, street, trees... in the image). The interrogation of SMIL documents set requires a unified structure describing multimedia components in order to perform fairly the same research process on the mixed contents. The use of meta-documents gives the privilege of querying only useful data strictly may reflect what a given component wants to express. The match meta-documents/query is performed thanks to a retrieval algorithm which takes into account the query regardless its type, turn it in SPARQL language, interrogates all of meta-documents written in RDF, retrieve relevant multimedia components (through its meta-document), assign to them a relevance score, rank multimedia items based on these scores, and finally show results.

Obtaining results starts with the selection of components / documents matching a query and followed by the classification and representation of these entities in an interactive way making easy the access to all of them. SMIL documents set presented in a given result have necessarily multimedia objects which respond to existing needs expressed in the user query. This relevance explains the degree of similarity between a query and multimedia components annotated by meta-documents. Representation of the results is the last part in the search procedure of SMIL documents. The way to display a given output could be set by the user when submitting the query: the user can choose the type of multimedia components to display [image, piece of text...] and how to display it, thus the result could be:

- Result composed by the same type of multimedia object (i.e. images only)
- Result composed by SMIL documents.
- Result grouping the two already mentioned types.
- Result composed by the same type of multimedia object grouped by SMIL document (i.e. all pieces of text in each SMIL document responding to a given query)

In our context we use ontology to retrieve relations between terms in the querying phase and to propose new queries to the user considering those relations. Independently to the progress of the three fundamental search modules, the extension of ontology is a continuous phase which aims to enrich ontology by concepts already used in the description module. For the enrichment of ontologies we propose a semi-automatic method of connecting concepts to extend an initial ontology with consideration of its meaning. The connection process (Part III of Figure 3) aims to choose a given term,

give it a type (class, property or individual in OWL), find a proper relationship with an existing term in the ontology and join the two by this relationship. The connection technique that we propose to enrich an initial ontology is based on three sources:

- From the meta-document annotating a multimedia component or a SMIL document, an automatic extraction of concepts is done using the anti-dictionary structure which removes not-meaningful terms, such as possessive pronouns or demonstrative Pronouns. Manual selection from the resulting concepts is performed in order to enrich the ontology base. After selecting a concept, we can set the connection parameters such as the type of the new concept, the relation of an existing concept in order to join the new concept to the ontology.
- From loaded ontology: the tool can automatically extract and categorize from an ontology file the constituent concepts, this extraction may drive the connection technique. To end the process of connection we should specify parameters concerning the new concept. By this type of connection we can connect even a complete OWL sub-arborescence to our initial ontology.
- From the user queries, a quantification frequency of occurrence of terms is carried out and a cloud of words based on these frequencies is established grouped by domain; the size of a term in a cloud is depending of the number of its occurrences in users' queries, finally, a selection of candidate concepts and an ordinary connection procedure could be applied.

7 Functionality

In our work we deal with a collection of SMIL documents and ontology concerning the LMD (License, Master and Doctorate) domain. The LMD Reform started in Tunisia in 2006. It aims to create flexible and efficient trainings, both fundamental and applied, offering to students wider opportunities for professional integration. We choose this domain in order to clarify some intricate notions to students using a semantic search engine based on standards of annotation which could be combined in an RDF code such as DC, FOAF and others. The functionality of our application becomes accessible through its interfaces. In this section we choose four basic interfaces among many others. The first application interface is shown at Figure 4 and it consists of three main parts designed as a flower. The first petal (blue) designed to add a new SMIL document to the collection, the second (green) is designed to trigger the search process, by the last part (orange) we can begin an annotation process in order to annotate a multimedia component.



Fig. 4. A screenshot of the application's first interface

In The orange part of figure 4 we select the SMIL document in order to annotate one of its multimedia components. This leads us to a new interface which is composed into 9 zones as we see in the figure 5. Those zones are explained subsequently:

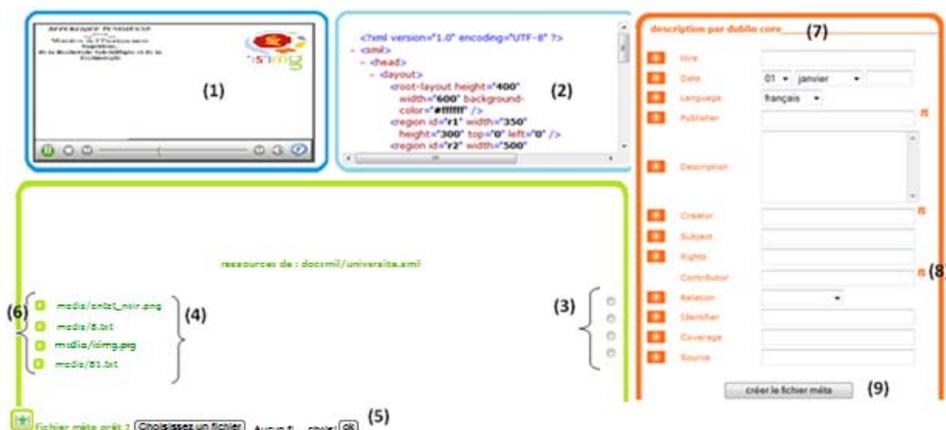


Fig. 5. form of annotation process

- (1) In this area the SMIL document is played to make an idea about the overall presentation and the temporal/spatial position of the multimedia component to annotate.
- (2) In this area, we find the source code of SMIL document from which the user could make a different kind of idea as the technical features of the multimedia component to annotate (the time, format, durations ... are picked up automatically).
- (3) Radio buttons for selecting the component to be described.
- (4) List of existing components in the SMIL document and which have not yet meta-document describing them: i.e. in the document "universite.smil" There are four types of components (two text and two images).
- (5) The user could load a file (.rdf or.txt extension) as a meta-document (instead of filling the form).

- (6) When clicking on a green squares a window appears displaying or playing the correspondent multimedia object (image, video, animation swf, text, textstream, audio sequence).
- (7) The following form contains the elements of the DC to fill in order to annotate the selected media. (Other forms could be displayed according the chosen namespace [orange petal of the previous figure] here we use DC to annotate the component).
- (8) The orange "n" ensures multiple descriptions for only one item; it could create RDF sequences i.e. several authors of a single text.
- (9) The check of information filled in the form and the generation of a new meta-document are done by pressing this button.

The importance of concept connection technique is that it allows making richer an ontology so we present in the next figure an example of this technique. Figure 6 shows the common window appearing when we choose a concept in order to connect it to the ontology. In our case we present a connection technique based on users' queries. The cloud of terms behind the window represents the candidates terms of connection, those terms are the most frequently used in queries concerning the LMD domain.



Fig. 6. Common window of connection technique

Our last chosen interface shows a typical presentation of results. Here, the type of multimedia picked in order to be searched is image, the form in the top of this figure represents two types of queries (advanced and expert query) while the other type of query (simple one) is presented in the green part of the first application interface (figure 4). Icons in the right side of this interface represent links to others interfaces of the application (clouds of queries terms, extending and loading ontologies, turn back to simple query interface ...). Small blue icons right on the bottom of each image shows more details about the annotation and the rank of the correspondent image.



Fig. 7. Typical screenshot of results presentation

8 Conclusion

In this article we developed a tool for the semantic search in a SMIL documents collection. Based on a simple text, SMIL allows creating rich interactive multimedia presentation where the multimedia components are uniquely identified by URIs ensuring an easily decomposition usable in the annotation issue. The metadata annotating web resources are fundamental to join the Semantic Web principles. We use meta-documents to annotate SMIL multimedia components by a unified structure. In addition to the use of meta-documents structure in the querying module, we use also ontology which is primary in a “semantic” context. In order to extend ontology, we develop a semi-automatic connection technique considering the user queries, meta-documents and ontologies loaded to this purpose. For our short-term outlook, we wish to extend our work to be usable in a collection of multimedia documents as HTML or PDF. As for long-term prospects we hope to restrict semantic results by exploiting in deep ontology’s relationships.

References

1. Resource Description Framework (RDF) <http://www.w3.org/RDF/>
2. Web Ontology Language (OWL) <http://www.w3.org/2004/OWL/>
3. Synchronized Multimedia Integration Language (SMIL) 1.0 Specification W3C Recommendation 15-June-1998 <http://www.w3.org/TR/REC-smil/>
4. Synchronized Multimedia Integration Language (SMIL 2.0) - [Second Edition] W3C Recommendation 07 January 2005 <http://www.w3.org/TR/2005/REC-SMIL2-20050107/>
5. The Friend of a Friend (FOAF) project <http://www.foaf-project.org/>
6. Dublin Core Metadata Initiative <http://dublincore.org/>
7. Synchronized Multimedia Integration Language (SMIL 3.0) W3C Recommendation December 2008 <http://www.w3.org/TR/smil/>
8. Troncy R. « Nouveaux outils et documents audiovisuels : les innovations du web sémantique » in « Documentaliste - Sciences de l’information » DSI 05, vol. 42, n°6. 2005.
9. Saathoff, Carsten ; Scherp, Ansgar. « Unlocking the Semantics of Multimedia Presentations in the Web with the Multimedia Metadata Ontology », Raleigh, North Carolina, USA. ACM 978-1-60558-799-8/10/04. April 26–30, 2010.

10. Sébastien Laborie, Antoine Zimmermann. « A Framework for Media Adaptation using the Web and the Semantic Web ». The Second International Workshop on Semantic Media Adaptation and Personalization (SMAP), Londres. 17-18 Decembre 2007
11. Laborie S, Manzat A., Sèdes F. « Création et utilisation d'un résumé de métadonnées pour interroger efficacement des collections multimédias distribuées » in 27th « Informatique des Organisations et Systèmes d'Information et de Décision » (INFORSID 2009), pages 227-242, Toulouse France. 26-29 May 2009.
12. Jedidi A. « modélisation générique de documents multimédia par des métadonnées : mécanismes d'annotation et d'interrogation » Thesis of « Université TOULOUSE III Paul Sabatier », France. July 2005.

Information Systems and Databases



Muti-Representation and Generalisation Based Webmapping Approach Using Multi-Agent System

Khalissa Derbal¹, Kamel Boukhalifa¹, and Zaia Alimazighi¹

LSI Laboratory, Computer Science Department, Faculty of Electronic and Computer Science, USTHB, El Alia BP 32, Bab Ezzouar, Algiers, Algeria
(kderbal,kboukhalifa,zalimazighi)@usthb.dz

Abstract. Over the last decade, an enormous demand for digital maps in different disciplines and fields was stated. Geographical information is currently available at anytime, from anywhere on the surface of the earth, by any person connected to internet. Some applications of design, implementation, generation and dissemination of maps on the Web are recognized as *Webmapping*. It uses among other things, a Geographic Data Base (GDB) and must be able to provide a fast response time (quasi-real time) and a high quality of visualized information. We propose in this paper, a Webmapping approach which is based on two principles; (1) exploiting an hybrid approach *Multiple Representation and Generalisation* in storing, handling and generating geographic data and (2) integrating *Multi-Agent technology*, in all steps of the Webmapping process. The effectiveness assessment of our webmapping approach is performed in *ArcGIS environnement 9.3*. We present some results of our experimentation which focused on the *road network theme*.

Keywords: Geographical Information, Webmapping, Multi Representation Data Base (MRDB), Automatic Generalisation Process, Multi Agent Systems.

1 Introduction

The large amount of handled geographical information comes mostly, from various GDBs designed independently of each other, although they relate to the same location. They are developed according two factors: (1) *Level of Detail (LoD)* which corresponds to map scale concept and (2) *Point of View (PoV)* that expresses the perceiving way of a real entity located on the surface of the earth. Producers and suppliers of cartographic data have deemed useful to exploit these different GDBs acquired with a very high cost. Thus, the same phenomenon may have multiple representations. One of developed approaches to model and manage such information is the integration of these DBs associated with the same location into an integrated one [1, 2] called for this purpose *Multiple Representation Data Base (MRDB)*. In the MRDB, representations associated with a same geographic phenomenon are linked by explicit relationships. We emphasize that

in this context, we consider two multiplicity factors; LoD and PoV. We have also, distinguished the concept of *Multiple Representation* (MR) associated to a MRDB as previously described, from that corresponding to results of generalisation process. We characterize the first by relevant because they are close to the real representation. Also, automatic generalisation process allows generating as many representations as expressed needs from a very detailed GDB (high LoD). It is concerned with the transformation of a representation of a part of the world. Despite the efforts, the automation of this process doesn't achieve, it keeps improving [3–5] since its inception thirty years ago. The agent-based approaches [6, 7] have attempted to imitate the cartographer reasoning who considers objects in their global context that is the purpose of the map. It represents the common goal to agents that interact by coordinating their actions and cooperating to achieve this goal. Webmapping is so, an application in which the web represent an important platform in dissemination of geographical information and offers several advantages such as accessibility and timeliness. However it requires a real-time map delivery.

Pre-designed and stored within a MRDB or generated by triggering a map generalisation process, the contents of these maps must be adapted and personalized according to a given user query and context. But *are users pleased with their displayed maps?! Have they felt any impatience in waiting the visualization of required maps? How about its quality?* Many researchers have addressed these issues with the aim to develop a Webmapping applications devoted to the management and delivery of geographical information on the web via generalisation services or geographic web services [8] and webmapping application [9–11]. In [12], the authors present a clear distinction between webmapping applications and geographic web services. [13] provides a synthesis of research orientations in this area based on the use of multiple representation and generalization, stating that the web already occupies a place which is developing all the time.

We propose in this paper, a Webmapping approach based on multiple representation and generalisation which remains an active research area [13, 14]. We use Multi-Agent Technology in all steps of our approach in order to reduce the map generalization process complexity by exploiting firstly, the autonomy of agents and secondly, the communication between agents to resolve conflicts over space use. We were inspired by some works developed in this context that we introduce in the next section. In section three we focuses on our contribution, it's organized on some subsections, in which we highlight our approach principles. A tool implementing our approach is presented in section four. Finally, we conclude the paper with a summary of the essential addressed points and make suggestions for future progress.

2 Related Works

Many research works have addressed webmapping globally or partly by focusing on specific tasks, which once integrated into the process, they ensure its proper performance, such, is the case of the generalisation process. It is mainly

the reason which has led researchers to develop variant of generalisation process according to various approaches. The agent-based generalization approach has been developed and improved during last years [14]. In the web context [15] talk about *on-the-fly-generalisation* which denotes the use of automated generalisation techniques in real time. Multiple representation, automatic generalisation and *Multi-Agent System* (MAS) are so, the three basic pillars of most works in webmapping. All these works have the same objective which consists of developing an automatic generalisation system adapted to the web. It must reduce the complexity and the cumbersome of the earlier systems based on different approaches (algorithmic approaches, knowledge based approaches, etc). It is therefore necessary to exploit the powerful features of agents such as, autonomy and communication. Thus, they have started from the basic idea, which is assigning a software agent to each object and/or group of objects. They are differentiated by the number and types of used agents according to the addressed themes such as *meso*, *macro*, *micro* and *submicro* agents in [7, 16] and agents or groups of agents which act upon three levels of data (the initial map, layers of interest and final generated map)in [11].

However, these approaches face agents number explosion in the case of a dense Area (urban area). Indeed, the number of created agents becomes huge, what makes communication between agents very complex and increases the likelihood of having a deadlock. These approaches are so, considered useful and efficient in low dense Area processing (rural area). In our approach we overcome this problem in a potential way. On the one hand, the generalisation process used is conditioned by search in the R-MRDB which allows moving towards the level of detail requested if it is available. On the other hand, we introduce the *map-area* concept which leads to process a dense map. *Map-area* represent a sublocation of a map. Each sublocation is handled by an agent *meso*. More details on our approach are presented in next section.

3 Webmapping proposed approach

To carry out our research work, we have set some assumptions; the Relevant MRDB(R-MRDB) is associated to an Area. It contains exclusively geographical data acquired through reliable process; it isn't the result of generalisation process. We use vector data with several LoDs and PoVs. The implemented generalisation process depends on some constraints like resolution interval and generalisation rate. Due to space limitations we don't address this aspect in this paper; more details are in [2, 5]

3.1 General description of proposed approach

Our approach is entirely based on a multi-agent system which supports the principle tasks as described in figure1: query Analysis, selection of the layers of interest and generation of the final map. A query initiated by a user is primarily analyzed in order to extract the defining features of the requested map that is the

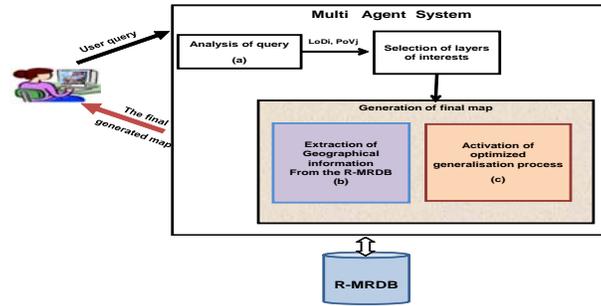


Fig. 1. General description of proposed approach

LoDi and PoVj (module (a) in figure 1). Some layers of interest are so selected and a driven search process through the web platform is triggered. This process begins with a search in the R-MRDB; if requested information associated with LoDi and PoVj is explicitly stored in the R-MRDB, it will be directly returned in a real response time (module(b)in figure 1). Otherwise, a generalization process is enabled to produce the requested Map (module(c)in figure 1). We also note that the terminology used in the description of different types of agents (*Coordinator*, *Macro*, *Meso* and *Micro*) of our system is inspired by [11] and [14]. The role and functionality of each of them is detailed in the following section.

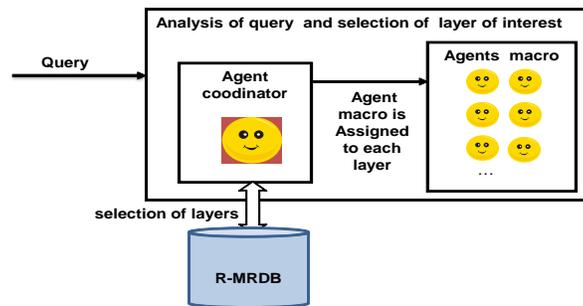


Fig. 2. Role of Coordinator Agent

3.2 Functionalities of the developed MAS

Our MAS consists of different types of agents, specific tasks are assigned for each of them. These agents interact with each other to ensure a smooth running of process. The sequencing tasks to be performed in generating final map is directly related to the triggering of hierarchical types of agents: *Coordinator*, *Macro*, *Meso* and *Micro*. Restitution of result is done in the reverse (see figure 4).

Agent coordinator : The *agent coordinator* is responsible of two tasks :(1) analyzing a user query in order to identify the layers of interest according to a LoD and a PoV(aim of the map) and (2) assigning agent *macro* to each layer of interest. Each agent *macro*, must decide the suitable processing for its layer. It so, initiates its inference engine while having as input parameters the LoD and the PoV. This is either a direct extraction of R-MRDB (module (b)in figure 1), or a triggering of a generalisation process (module (c) in figure 1) by triggering other type of agent (*meso, micro*) in a hierarchical way. And so on the processing is completed. The result (final map) is delivered to the coordinator agent (see figure 2).We emphasize that in developing the query analysis module, we restricted to two classes of users: (1) *professional user* who have depth knowledge in cartography and, (2) *occasional user* who shall be assisted during the process. We also manage a list of keywords related to the application domain (urban design) that we have chosen during the experimentation phase.

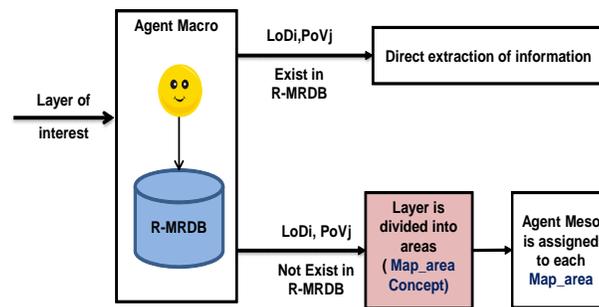


Fig. 3. Processing performed by an agent macro

Agent Macro : The *Agent Macro* continues the path of the process by accessing to the R-MRDB. If the requested map content corresponds to a LoDi and PoVj explicitly stored in the specified DB, it proceeds by direct extraction, the response time would be very efficient. Otherwise a generalisation process is triggered(see figure 3). As stated previously, our approach overcomes the map density problem (explosion of agents number), by introducing the concept of *map_area*. The layer is so partitioned into areas, to each of them is assigned an agent *meso*. Thus, our approach may be adapted to the processing of any geographical location with a high density (large number of objects) or low density. Thus, the preliminary search in the R-MRDB, *map_area* concept and parallel processing provided by developed MAS allow leading to a great process with a real-time response.

Agent Meso/Micro : The agent *Meso* assigns to each object in its own *map_area*, an agent *Micro*. Micro Agents are responsible for the accomplish-

ment of the generalisation process. We state that in the context of this work, we focused on the *road network theme*, so we have relied on the operator of simplification for conflicts resolution. In [5], we have described the various conflicts around this theme and constraints to satisfy in their resolution. Intra-conflict is the result of racing agents for the space occupation. They communicate in order to preserve the overall harmony of the map. Communication between agents is based on blackboard technique. It is a space in which each agent has a record (Id -agent, current geometry of the object, state of the agent) visible to other agents in the same area.

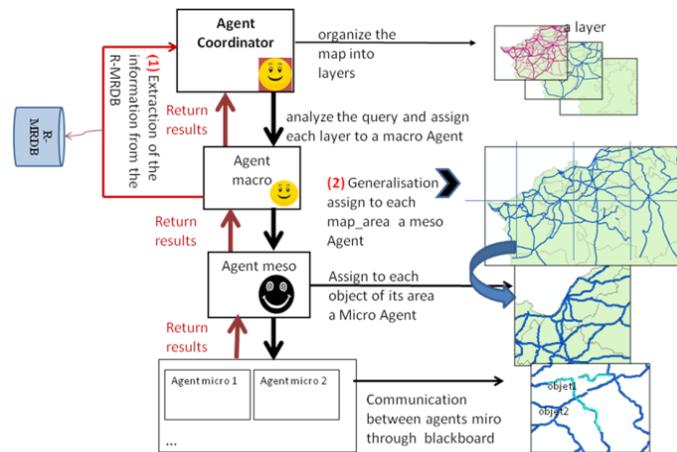


Fig. 4. Illustrative example

3.3 An illustrative example

In figure 4, we present a recapitulative of our approach through an example of application on the road network of an area in north of Algeria with illustration of all processing phases. In this example, we consider the road network theme with its different LoDs (national roads, departmental roads secondary roads, etc) according to the nomenclature of the National Institute of Cartography and Remote Sensing (INCT) of Algeria. We assume that a query initiated by an occasional user contains the keyword 'highway' such as: *I got lost on the highway Algiers-Blida*. The query analysis module will extract the location and level of detail associated with the national roads which is directly extracted from the R-MRDB in a real time (case (1) in figure 4). If against, a civil protection officer (professional user) looks for quick access to a location, we invite him to enter information such as visualization scale of requested area which is associated to the highest LoD in R-MRDB. The displayed map will be congested. So, we

proceed by generalisation in order to keep only useful information to the officer (case (2) in figure 4).

4 Experimentation

Our Webmapping prototype was developed in the ArcGIS server 9.3 environment. We have also used the platform JADE for the implementation of our MAS. We consider in this paper only simplification operator in generalisation process. Our experimental method and the software tools used are described in figure 5.

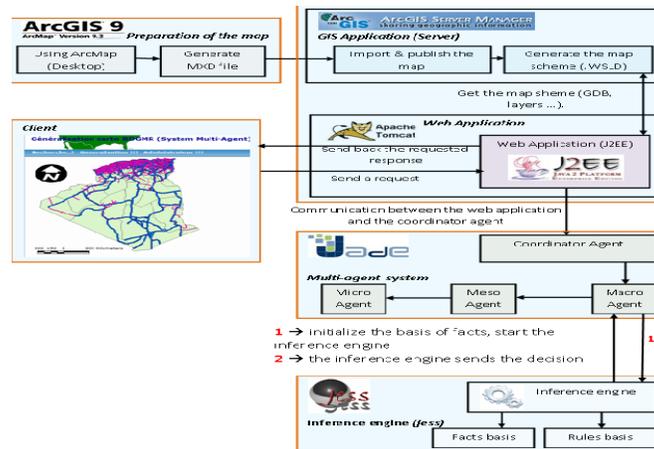


Fig. 5. Global scheme of different parts and steps of our application

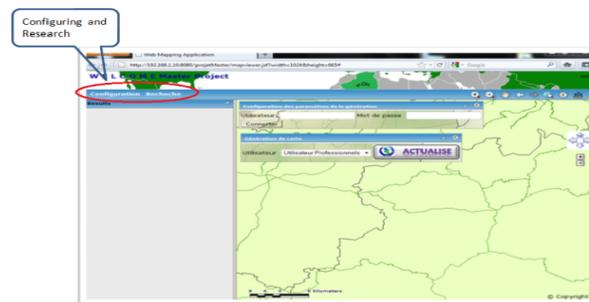


Fig. 6. The main interface

4.1 Main interface of our Webmapping Application

This interface is run through a web browser which allows two tasks; configuring, and search (figure 6). The first task presented in figure 7, is secured because it allows access to map generation parameters. The second task concerns customers (professional or occasional). We have configured two user interfaces, one for professional user (see figure8) who can provide valid information and occasional user(see figure 9) who hasn't depth knowledge in the field. We note that in figure 8, the field scale is activated, the user is identified as professional one (See illustrative example above).The LoD of the requested data is determined from the input map scale. However in figure 9 (occasional user) the same field is deactivated. The developed system must be able to define this entity as showed in the example below.

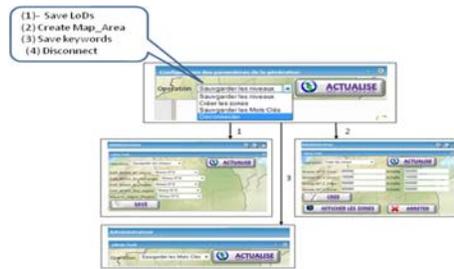


Fig. 7. Administrator Tasks



Fig. 8. Professional user interface

A first result of a running example is presented in figure 10. From a query of an occasional user, two keywords RN6 and RN13 are implicitly or explicitly expressed. The area which contains the specified roads is so identified. The requested LoD corresponding to the layer national roads is available in R-MRDB. Our webmapping system proceeds by direct extraction and the result is delivered in a real time.



Fig. 9. Interface for occasional user

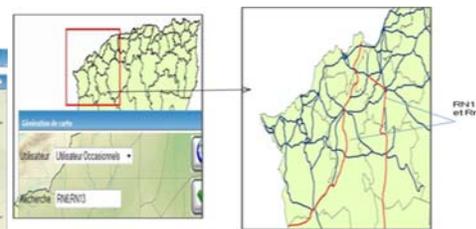


Fig. 10. Direct extraction from R-MRDB

5 Conclusion and future Work

The developed Webmapping approach is based on multiple representation and generalisation, in order to take advantages of the relevance of the first and flexibility of the second. The utilization of multi-agent system technology has provided our approach with parallel processing in automatic generation maps. Indeed, in our developed MAS, an agent is assigned to an object and /or group of object according to a hierarchical organization. These agents act independently while adapting to environment changes and communicate with other agents via the blackboard technique in order to provide the result (final map). Therefore the developed system allows reaching a real-time response required in this web context. As future issues, we suggest firstly to provide some improvements to the current solution specially, the development of spatial query module analysis based on a domain ontology and secondly to improve the supporting map customization preferences and user profiles. This can be done by collecting information on the web and mobile users through questionnaires, or by using a learning system that is able to distinguish between professional and occasional user.

Acknowledgement

The authors are grateful to Bouchenine Yakoub and Abada Lyes Ph.D. students in USTHB University, for the considerable effort carried in achieving this work.

References

1. Vangenot, C.: Multi-representation in spatial databases using the mads conceptual model. ICA Workshop on Generalisation and Multiple Representations, Leicester (August 2004)
2. Derbal, K.A., Lejdel, B., Alimazighi, Z.: A hybrid approach to modeling and managing multiple representation of spatial data (application theme :road network). International Conference on Information & Communication systems ICICS'2009, NYIT JORDAN (2009)
3. McMaster, R.: Automated line generalization. in *Cartographica* 2 (24) (2006) 74 –111
4. Ruas, A.: Modle de gnralisation de donnes gographiques base de contraintes et d'autonomie. Ph.Dthesis, Universit de Marne-la-Valle (1999)
5. Derbal, K.A., Alimazighi, Z.: Approche de rsolution de l'auto-conflit dans un processus de gnralisation automatique du linaire routier. Veille Stratgique Scientifique et Technologique VSST'2010, Toulouse (2010)
6. Duchne, C.: Coordination multi-agents pour la gnralisation automatique. Bulletin d'Information de l'IGN (74) (Mars 2003)
7. Ruas, A., Duchne, C.: A prototype generalisation system based on the multi-agent system paradigm. *Generalisation of Geographic Information: Cartographic Modelling and Applications Elsevier ltd chapitre 14* (2007) 269 – 284
8. Bergenheim, W., Sarjakoski, L.T., Sarjakoski, T.: A web processing service for grass gis to provide on-line generalisation. 12th AGILE International Conference on Geographic Information Science Leibniz Universitt Hannover, Germany
9. Cecconi, A.: Integration of cartographic generalization and multi-scale databases for enhanced web mapping. Ph.D. Thesis, University of Zurich (2003)

-
10. Bernier, E., Bdard, Y., Hubert, F.: Umapit: An on-demand web mapping tool based on a multiple representation database. 8th ICA WORKSHOP on Generalisation and Multiple Representation, A Corua (July 2005)
 11. Jabeur, N.: A multi-agents system for on-the-fly web map generation and spatial conflict. PHD thesis. Universit LAVAL, Qubec (Janvier 2006)
 12. Pornon, H., Yalamas, P., Pelegris, E.: Services web gographiques, tat de l'art et perspectives. Gomatique Expert (65) (octobre-novembre 2008)
 13. Burghardt, D., Petzold, I., Bobzien, M.: Relation modelling within multiple representation databases and generalisation services. The Cartographic Journal **47**(3) (August 2010) 238–249
 14. Duchne, C.: Automatisation de la gnralisation. Sminaire quipe SMA - LIP6
 15. Weibel, R., Burghardt, D.: On-the-fly generalization. Encyclopedia of GIS . Springer science & Business media LLc.ISBN :978-0-387-35973-1 (2008)
 16. Gaffuri, J.: Deformation using agents for map generalisation - application to the preservation of relationships between fields and objects. AutoCarto'06, Vancouver, United-States, 2006. ACMS. (2006)

Towards a numerical model for the representation of an urban transportation system

Karim Bouamrane, Beghdadi Hadj Ali, Naima Belayachi

Computer Science Department
Faculty of science, University of Oran
Oran, Algeria
{kbouamranedz, h_beghdadi}@yahoo.fr, bnaima2@hotmail.fr

Abstract. The paper discusses a method for modeling the operation of an urban transportation system. The proposed model models the bus operation in an urban transportation system with an equation system. The objective of the proposed model is to simulate the operation of an urban transport network. The network consists of a set of lines, a number of vehicles (buses) circulate on each line, where we consider some variables such as traffic conditions on the sections (bottling), the distribution of travelers on breakpoints, to see the impact of external environment on the network. The coordinates of the breakpoints of a given line are defined with respect to the filing. These points can define by the polynomial interpolation the displacement model in the line in question. The movement of vehicles is subject to a variable commercial speed and acceleration fixed in each segment. The speed depends entirely on a set of constraints related to traffic in the urban space. Finally, with an estimate of the filling of vehicles and the costs of travel related to spare parts, fuel, and payroll, we can deduce the economic profitability.

Keywords: urban transportation system (UTS); interpolation; displacement; commercial speed; filling; economic profitability.

1 Introduction

In the domain of Urban Transportation System (UTS), modeling is a complex task that requires the development of appropriate models to ensure customer satisfaction, namely, to propose an urban transportation service taking into account the operational constraints such as on-time theory, guarantee letters, reducing wait times, etc... This naturally led researchers to look at this problem and propose appropriate models. Several models have been compared in [2, 4]. Modeling of *UTS* was proposed by means of Petri Nets (*PN*). The *PN* were exploited to model the flow of travelers [5], where a model based on stochastic *PN* has been used for a bus network whose tokens represent the travelers. The Stochastic *PN* are an extension of "time" of ordinary *PN*. Stochastic Petri nets have been used for the modeling of matches in [1]. The work

proposed in [10] describes an approach based on modeling of the flow. It presents the rates to be assigned to different routes of the road network in accordance with criteria representing the cost of roads (toll ticket for public transport) and the duration of the course. *UTS* is represented by a graph, on each arc, the flow and travel time are marked, and the thickness is the importance of the flow. Ngamchai [9] models the every route by series of nodes represented the sequence of the stop points, using genetic algorithms. Another model of urban transport systems is given using the Multi-Agent Systems (*MAS*). Several studies were conducted in [4, 8], the first work in this area were made to model urban traffic by F.A. Bomarius [3] where he proposed a multi-agent modeling scenarios of urban traffic at the intersections. Subsequently, F. C. Besma [2] has used *MAS* for monitoring and diagnosis of a transit system operator with a hybridization of evolutionists' algorithms. Another approach with *MAS* has been published in [6] where an application designed to monitor long-term users' information system in *UTS* has been proposed.

A comparison between the modeling approaches mentioned previously was presented in [4]. The modeling by *PN* mentioned before does not take into account the influence of the external environment on the rate of vehicles displacements during the journey, in addition, the system does not evolve in a continuous manner as the system needs an event to cross a transition in the *PN*. The modeling of travelers flow considers only the flow of users with the travel time and neglecting the economic charges of the system. The modeling of a route models the route by a sequence of nodes in a static manner, this implies that this model does not include the movement of vehicles with continuous function of time from beginning to end of service. Modeling using *MAS* takes into account scenarios of urban traffic at intersections, however, this model is a little heavy compared to the number of messages that slow circular between the agents of the system. However, none of these models is suitable for modeling the operation of a transit system in a continuous function of time along a path that can be not linear knowing only its position on some stop points and taking into account the constraints of movement, means and resources available with a mastery of traffic loads in order to provide an estimate of the economic viability of *UTS*. Hence, the work proposed in this paper presents a numerical modeling for the operation of *UTS*. For a better understanding of the parameters that determine the evolution of the transport system, the modeling is a very effective means. Indeed, a model provides a better vision which allows us to provide a first step, the evolution of the phenomenon from the initial conditions, and subsequently to make predictions about the system and order it. For a simplified and usable representation of urban transportation systems in order to exploit it, we define the flow of a bus in an urban space by a system of equations. The displacement model on a given line is defined by an interpolation of the coordinates of the breakpoints of this line. The resolution of the proposed system of equations can predict the distribution and coverage of lines of operations and the economic profitability. The paper is organized as follows. The problem is described in Section 2. Section 3 is reserved for the proposed model. The section 4 presents the implementation of the numerical proposed model. Finally, we end with a conclusion.

2 Positioning of the problem

Travel in an urban transportation system paths are trajectories traveled between the contexts of social activities, they are becoming more complex and difficult to control because it depends on several phenomena that determines the traffic flow and causes a significant lengthening of journey times, for example: congestion, intersections, inappropriate timing of traffic signals, public works, weather and so on.

The impact of the external environment and road conditions make monitoring and control of the transport system more hard hence the difficulty to model the behavior of a bus in a transit system throughout the journey and locate its position on the route and see that its speed depends on various constraints of urban space and to estimate the filling deduce its economic viability. Among the constraints that affect the functioning of the urban transport network, we cite the road paths that can be not linear, the traffic constraints affecting the traffic on the sections, and the distribution of travelers on the breakpoints that we consider in this work.

For this reason, we propose a numerical modeling for the urban transportation system mono modal (bus) to obtain a simplified visual representation of reality, and to study the network behavior with respect to various disturbances (public works, passage of a VIP motorcade, congestion, ...) that can divert the system to normal running on one side, and in order to estimate the economic yield of the network on the other side. The numerical model proposed can better represent reality by modeling the running of an urban network continuously with time.

The behavior of a vehicle is modeled by a system of equations. The vehicle is moving according to an equation with time unlike other models such as modeling by Petri nets or the multi-agent systems where the system must wait for the crossing of transactions, or the presence of an event and sending decisions between agents to operate.

In addition, the proposed model can provide a view of the urban network, namely the movement of vehicles (bus), speed, and filling every moment during the journey, and consequently the economic profitability of the urban transportation system.

3 Numerical modeling proposed

Modeling of urban transport provides a help in developing appropriate policies in terms of planning and programming. For this reason, the numerical model proposed enable to model the behavior of a bus in the urban transportation system, namely the network operated by the carrier of Oran city. The resolution of the proposed system of equations can predict the distribution and the coverage of operating lines. It also helps to optimize the running of the network to minimize the costs of rolling, and therefore, it offers a means of estimating the economic profitability of the urban transportation system. The urban network is characterized by: A set of lines: each line is characterized by a well defined length, it has a specific number of bus stops distributed along the line between two extreme cases (terminus); The travels between the breakpoints are provided by a number of vehicles affected for each line of the network; The bus of each line circulate in a certain frequency of passage; A distance deadhead (HLP):

between the deposit where the buses are parked and the point stop starting; Distribution of Judgments: The distances between the breakpoints of each line are not equidistant; The lines on the network must be projected in a coordinate system that ignores the urban fabric where a projection using the software (MapInfo) which is a geographic information system to extract and display of geographic data; The values of displacement are determined by taking as a basis only the values observed in the buses stop points.

The diagram shown in Fig.1 represents the nonlinear path of a bus on a given line. On this journey, we have n breakpoints which are known.

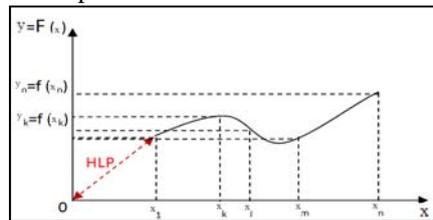


Fig. 1. Modeling of movement for a vehicle v on a line l

Considering this set of n breakpoints positions $(x_i, p(x_i))$ and a function F unknown, we determine a polynomial $P(x)$ of degree $(n-1)$ which interpolates F in the considered points. The n breakpoints that are known can define by the proposed model of interpolation, the displacement of the buses on the line in question. Therefore, the movement of vehicles (buses) of a given line is represented by (1).

$$P(x) = \sum_{k=1}^n [p(x_k) * \prod_{j=1, j \neq k}^n \frac{x-x_j}{x_k-x_j}] \quad (1)$$

$Y = P(x)$ is an interpolation function to estimate the values of the displacements $(x_i, p(x_i))$ at each time t from start to finish by having an estimates of speed rolling $V(t)$. Due to the proposed model, we know the displacement $Y = P(x)$ of a vehicle (bus) v and locate it on the line along the route which is not linear as shown in Fig.1.

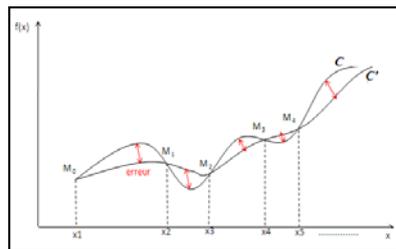


Fig. 2. Polynomial interpolation error

The landmark "0" of the Fig.1 corresponds to the deposit where the buses are parked initially. The HLP distance represents the distance between deposit and breakpoint of

departure. The distances between the breakpoints $(x_i, p(x_i))$ are not equidistant, they are defined taking into account the density of the flow of travelers on the journey. The goal is to approximate as closely as possible the unknown function of the displacement of the urban reality in order to provide more accurate values as possible with minimum error. The Fig.2 illustrates the curve "C" of the polynomial obtained $P(x)$ and the curve "C'" of the unknown function $F(x)$ of the displacements in the real urban environment.

We notice in Fig.2 a small gap between the two curves, the latter represents the interpolation error that we control in order to observe the quality of this approximation. For this purpose, equations (2, 3) show how we calculate the error of interpolation. Since F is a function $(n + 1)$ differentiable on the interval $[a, b]$ and $x \in [a, b]$ and let $I = [\min(x_1, x_n), \max(x_1, x_n)]$, then:

$$\exists \xi \in I / E(x) = P(x) - F(x) = \frac{F^{(n+1)}(\xi)}{(n+1)!} \Phi(x) \quad (2)$$

$$\text{Where } \Phi(x) = (x - x_1) * (x - x_2) * \dots * (x - x_n) \quad (3)$$

To minimize this error, we propose to decompose the nonlinear path of a given line a set of small linear sections in order to trace the curve in a more realistic with a minimal error (see Figure 3).

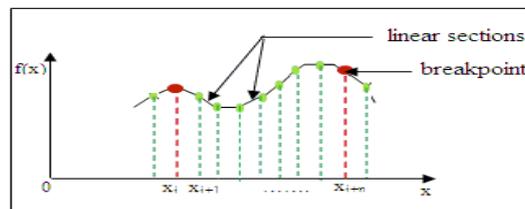


Fig. 3. Decomposition of the way into a set of sections

In the previous figure, we define a step "H" expressing the length of each section and we vary it not to have a minimum interpolation error.

In urban areas, several constraints affect the speed of the circulation of the bus. The speed varies from one point to another with a fixed acceleration by segment. The equation (4) shows how we model the change in speed during the journey.

$$V(t) = A t + V_0 \quad (4)$$

The speed $V(t)$ of a bus that moves on a line from one point to another depends on several constraints of the external environment. " V_0 " represents the commercial speed the start point. "A" symbolizes the vehicle's acceleration.

The vehicle acceleration is variable along the way. It is experimentally fixed on each linear segment from of the starting point to the ending point according to the state of the vehicle traffic on the sections.

On each linear segment of the trip, we have a speed $V(t)$ varies with a fixed acceleration 'A'. The function of displacement on this segment can be modeled by the function $h(t) = \frac{1}{2} A t^2 + V(t) + h_0$. Differentiating this function $h(t)$, we obtain the speed $V(t) = A t + V$ where V approaches V_0 .

The speed depends on the acceleration and the nature of the line. The acceleration is defined experimentally as a function of variables called disruptive that can happen wind happen in a race, for example: the density of traffic, bad weather, an accident in the road, large flow of passengers, a failure or temporary immobilization (damage) of a vehicle, peak hours, special events (VIP passage of a procession or parties), etc..

In our context, race is defined as travel between two extreme terminuses (forward or reverse direction). As the proposed model can model the operation of a bus all the way from the initial stop to the terminal stop of arrival, we just need to increment the number of strokes each time the bus in question happens to stop terminal in order to know the number of race made by him during the day. The objective of the proposed model is to model the operation of *UTS* also with an optimization of the economic profitability. For this reason, we try to calculate the formula *RC* of a bus during its commissioning in a day according to (5), where *NPM* is the number of people going up every time the bus comes to a stop-point and *PU* indicates the unit price that a traveler has to pay.

$$RC = RC + (NPM * PU) \quad (5)$$

The number of passengers on board a vehicle that travels from one point to another, is generated by the system according to a survey that was done on ground to see the distribution of passengers on the bus's breakpoints throughout the path of a given line, where we affect a rate of climb and a another of descent of passenger for each breakpoint.

Knowing the costs of running *CdR* (spare parts for vehicles, fuel, and payroll) to a given bus, and knowing his recipe *RC* during the day, we can deduce its economic profitability *RE* according to (6).

$$RE = RC / CdR \quad (6)$$

Using the proposed model, we can know at any time (t):

The position of the bus on the way: The polynom $P(x)$ is an interpolation function of the positions of a given bus in n breakpoints known throughout the journey, with an estimated speed of the roll at this moment, we can know its position on the route; The running speed is influenced by the constraints of movement (Will disruptive variables); The number and direction of strokes (or return); Filling to infer the economic profitability.

4 Implementation of the proposed model

The goal of any modeling and design is to produce a software tool to prove our statements of theoretical departure. We have developed our simulator mono post,

operator interface Microsoft Windows in the programming language C++ Builder 6 and using a geographic information system (MapInfo) to extract and display map data.

4.1 Treatment of mapping and collection of geographic data

From the map of Oran city (see Fig.4 (A)), we determine the roads and then we identify the paths of the lines of the bus transport company of Oran where we position the buses stop pointes of each line.

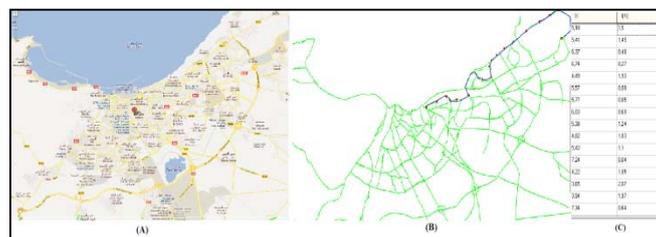


Fig. 4. Treatment of the Map of Oran

After the screening of the main roads of the city of Oran and the identification of the route with the positioning of the stops of each line of the Urban Transport Company of Oran, we generate a new map using MapInfo software. Figure.4 (B) illustrates the mapping generated, where the green lines represent the major routes of Oran. To simplify and clarify the picture, we have shown schematically in blue the way of one line “P1” on the urban transportation system of Oran, and symbolized with red points the bus stops points on this line.

To illustrate the results of the proposed model, we consider the path of the line “P1” where we have a vehicle “ v_i ” circulating on this line with 17 breakpoints (see Fig.4(B)), which means that we know the position “ $(x_i, p(x_i))$ ” bus to each of these 17 points (see Fig.4(C)).

4.2 Execution of the proposed model

Having located the stops and find their positions $(x_i, p(x_i))$, we turn to the performance of the model to find the displacement function $Y = P(x)$ that locates the bus “ v_i ” to each point $(x_i, p(x_i))$ for the journey. By applying the proposed model, we obtain the polynomial function of displacement whose curve is shown in Fig.5.

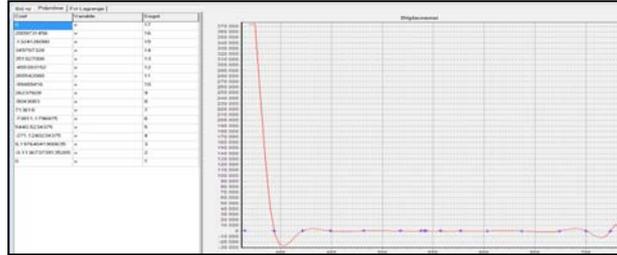


Fig. 5. Curve of displacement function $Y = P(x)$

To locate a bus and find its position in the path, it suffices to know the velocity can be estimated using (2). If we have accelerations of $A = 2\text{km/h}^2$, then $v(t) = 2t + v_0$, where " v_0 " is the initial speed of 5 km/h , so after 4 hours ($t = 4:00$), speed is 13km/h . Also, as we have broken the journey in a series of small sections then the linear distance "D" is linear and each section can be estimated by " $D = \frac{1}{2} At^2 + vt + d_0 = 56\text{km}$ " so the bus will be in the 4th race because the distance of the line "P1" is 15km , so the bus drove 11km in the 4th race and he is 4km to arrives at terminus. By replacing this value in the $x = 11\text{km}$ displacement function obtained, we can deduce the position $(x_i, p(x_i))$ of the bus on the non-linear way of the line in question after 4 hours of service. After finding the polynomial that corresponds to the displacement function per-putting to locate the position $(x_i, p(x_i))$ of the bus, we turn to the calculation error of interpolation according to (2).

Regarding the filling, every time that the vehicle (bus) comes to a stop point, we update the number of persons on board (the number of people coming down and getting on is random, but according to the periods in the course of a day, for example during peak hours: 08:00, 17:00, we have more people on board and hang the other periods of a day) and each time we calculate the following formula RC (5) as $PU = 15DA$, and at the end of service we calculate the economic profitability RE of the bus with an estimation of its traffic loads CdR during the day according to (6).

5 Conclusion

Like any model, the modeling of urban transport system responds primarily to the need for knowledge. The model gives a representation of a complex phenomenon, allowing a better understanding of its internal mechanisms and parameters that determine its evolution. The proposed numerical modeling defines the behavior of the bus in an observed line of the urban transportation system. It allows us to study the exogenous variables that are not part of the system (the original from the outside) and that affect the system variables such as urban traffic, and the endogenous variables that are part of the system such that: the frequency of passage, the number of vehicles, commercial speed etc.

To go further in the development of this model applied to an urban transportation system, we plan to integrate a module responsible for the regulation of this transporta-

tion network in case of perturbations in schedules and also we think to enrich the proposed model to move in a multimodal modeling (bus, tram ...).

References

1. Abbas Turki, Grunder O., Bouykher R. et Elmoudni A., "Modular controlled stochastic Petri nets for the connection monitoring", Proceeding of the World Automation Congress, 4th International Symposium on Intelligent Automation and control, ISIA030, Florida, 2002.
2. Bema FayeCh Chaar, "Régulation des réseaux de transport multimodal: algorithmes évolutionnistes et systèmes multi-agents", Thèse de doctorat, université de Lille 1 et Ecole Centrale de Lille, octobre 2003.
3. Bomarius F. A., "Multiagent approach towards modeling urban traffic scenarios", Tech. Rep. RR-92-47, DFKI-GmbH, Germany 1992.
4. Karim. Bouamrane, Fouzia. Amrani : "Un système d'aide à la régulation pour un réseau de transport urbain collectif : vers une approche à base de cas". Journal of Decision Systems (JDS) Vol 16 N°4, pp 469-504, (ISSN1246-0125), Hermès Science Publications, DOI: 10.3166/JDS.16.469-504, Lavoisier, Paris, 2007.
5. Castelain E., Mesghouni K., "Regulation of public transport network with consideration of the passenger flow: modeling of system with high level Petri Nets", Proceeding of the IEEE SMC Conference, WA2C3 Hammamet Tunisia, 6-9 October 2002.
6. Ezzedine Houcine, Christophe Kolski, "A. Péninou, Agent-oriented design of human-computer interface: application to supervision of an urban transport network", Engineering Applications of Artificial Intelligence, (18) 255–270, Elsevier, 2005.
7. Giulio Erberto Cantarella, Stefano de Luca, "Multilayer feedforward networks for transportation mode choice analysis: an analysis and a comparison with random utility models", Transportation Research Part C 13 (2005) 121–155.
8. Hakim Laichour, Modélisation "Multi-agents et aide à la décision : application a la régulation des correspondances dans les réseaux de transport urbain", Thèse de doctorat, Université des sciences et technologie de Lille, France, Décembre 2002.
9. Somnuk Ngamchai, Lovell, D.J. "Optimal time transfer in bus transit route network design using genetic algorithms", 8th conference CASPT, Computer Aided Scheduling of Public Transportation, Berlin, Germany, 21-23 June 2000.
10. Wynter Laura, Lolito P., "Boite à outils affectation du trafic de Scilab : étude sur l'analyse de transport en région périurbaine", 10eme rencontre INRIA Industrie, Application de l'informatique et le l'automatique aux transports, Paris-Rocquencourt, 16 janvier 2003.

Urbanization of Information Systems with a Service Oriented Architecture according to the PRAXEME Approach

Amel Boussis, Fahima Nader
LMCS (Laboratory of Systems Design Methods)
ESI (National School of Computer Science) ex (INI)
Algiers, Algeria
{a_boussis, f_nader}@esi.dz

Abstract: This article addresses the problem of urbanization of information systems. The development of an information system company is certainly a complex task. Hence the choice for organizations to opt for an urbanization approach of their information. In our work we are interested in a comprehensive approach to recast Information System (IS). This approach is oriented service (Service Oriented Architecture), based on a mapping and an orchestration of business processes related to the IS.

Keywords. SOA, Business Processes, Urbanization, Web Services.

1 Introduction

Everyone is willing to say that the information system is now at the center of running a business or an organization. Its operation and its efficiency are of an utmost importance.

An information system is a set organized of resources including hardware, software, personnel, data, and procedures to acquire, process, store and transmit information in companies. During the life of the company and its development, the information system is caused to change both in its structure and in its operation.

In response to these permanent developments, the idea of urbanism has been integrated within modern companies. The basic principle of urban planning in IT is, through rules and fundamentals principles, to follow these developments and its impact on the overall system.

The development of an IS company is certainly a complex task. Hence the choice for organizations to opt for an urbanization approach of their IS. Such an approach becomes necessary when the organization has accumulated a large number of projects over several years. Urbanization of IS designed to meet several objectives: the streamlining, modularity and more innovation. It is nevertheless a concept to simplify it, to use a term extension.

This paper is organized as follows: principles of urbanization IS are presented in section 2. Section 3 discusses basic elements of urbanization IS approach. Section 4 presents the context of our application case. Our urbanization approach is described in section 5 and the architecture system in section 6. Finally, we conclude and bring out some perspectives in section 7.

2 Principles of Urbanization Information Systems

Urbanization of IS has been studied by many authors [1-3] [5] [14]. The work of these authors complement the work on enterprise architecture. All these authors use metaphors to justify the notion of architecture and urbanization of IS. In particular, the metaphor of the city is used as the basis of urbanization of IS.

Club Urba-EA¹ offers the following definition: « Urbanization is to organize the gradual and continuous transformation of information system to simplify it, to optimize its added value and to make it more responsive and flexible towards strategic business changes, while relying on technological opportunities of the market. Urban planning defines rules and a coherent, stable and modular context, in which different stakeholders are referring to any investment decision in the Information System. »

The mapping is the set of studies and scientific, artistic and technical operations involved from the results of observations or the operation of documentation, to the development and the establishment of maps, plans and other expression patterns, and then their use [2].

Mappings are at the heart of the approach to follow for a project of urbanization of IS. We distinguish four types of mapping (business mapping, functional mapping, application mapping and technical mapping) that can be made to describe the existing system or the target system. As with city, the mapping of an IS is to time [1]:

- Scientific: isn't it a metamodel?
- Artistic: aesthetics is also a mean to facilitate communication.
- Technical: implementation is based on a number of techniques.

The process of urbanization is based on three key areas that feed each other [1]:

- Modeling strategy
- Mapping of existing systems
- Determination of target systems

The process of urbanization of the IS includes:

- Set a target IS, aligned to business strategy,
- Determine the path to follow to achieve this target IS.

¹ www.urba-ea.org

3 Basic Elements of Urbanization Information Systems Approach

3.1 Process

Process Notion: A process is a collection of related, structured activities or tasks that produce a specific service or product (serve a particular goal) for a particular customer or customers. It often can be visualized with a flowchart as a sequence of activities with interleaving decision points or with a Process Matrix as a sequence of activities with relevance rules based on the data in the process [13].

There are three types of processes [13]:

- *Management processes*, the processes that govern the operation of a system. Examples include Corporate Governance and Strategic Management.
- *Business (Operational) processes*, processes that constitute the core business and create the primary value stream. Examples include Purchasing, Manufacturing, Advertising and Marketing, and Sales.
- *Support processes*, which support the core processes. Examples include Accounting, Recruitment, Call center, Technical support.

A business process begins with a mission objective and ends with achievement of the business objective. Process-oriented organisations break down the barriers of structural departments and try to avoid functional silos.

Process Mapping: Before we focus on improving efficiency of an organization, it should be first to know it, therefore first establishing a mapping process component of this organization in order to know how it works. According to [11]: "The mapping process of a business or an organization is a graphical way to restore identification processes and their interaction."

According to [6], the development of a processes mapping and control interfaces meet perfectly the requirements of the version 2000 of the ISO standard and can provide solutions to many questions. It is the basis of the identification of important processes, it is useful to prepare the internal audit programs, it assistance in setting up for measuring systems and monitoring processes and can be used to set implement improvement programs.

To map, it is useful to proceed as follows: [7]

- Present the mapping of production process and control process.
- Mapping the support process.
- Define the flow interface between these three mappings.

3.2 Business Models

The company is a complex structure. In order to better understand the operation, organization, resources and exchanged information in a company, today we need abstract but manipulated representations: models. To model, it represented the “reality” of an object or a system [3]. A business model is used to represent different views and aspects of it [4]. A business model is not static but existing research work on defining structured methodological approaches for business model evaluation is rather fragmented. Several tools, languages and standards to model certain views of the company have emerged.

With a few exceptions [4], most literature has taken a static perspective on business models, implicitly assuming them to remain stable over time. However, in reality organizations often have to reinvent their business model continuously to keep aligned with fast-changing environments in some sectors. As a result, instantiations of business model dynamics may be found in any component of the business model, such as redefining or extending the service concept, replacing technologies. The UML is used today to model certain views of the company. UML is not a method but a technical representation because it does not permit to know precisely what to model. This is according to the methodology.

3.3 A logical Service-Oriented Architecture (SOA)

The concept of SOA (Service Oriented Architecture) defines an architectural style based on the assembly services offered by the applications. In this architecture style, the various software components are connected by a loose coupling (services are independent one from another in order to change easily the order about their orchestration to form the process).

A "service" within the meaning of SOA, is a connection to an application, providing access to certain of its functionalities. The functions provided by a service can be treatments, information researches. For example, an application of customer management can offer a service returning the contact information of a client. In an SOA architecture, we are interested However, to several different aspects of designing an IS. The PIM4SOA project [8] defines four views to define SOA architecture:

- **Informational view**: the information is related to messages and business objects exchanged between services.
- **Process view**: the process described sequencing and coordination of services in terms of interaction and control flow of processes.
- **Service view**: services present an abstraction and an encapsulation functionality provided by an autonomously entity.
- **Quality of service view (QoS)**: is interested on other non-functional aspects such as: safety and performance of services.

These views involve a logical architecture. The implementation of a solution of urbanization need to rely on a technical platform. A model on a technical architecture must be used. This architecture must be a technology framework on which the logical model is projected.

3.4 Oriented Architecture Technology - ESB

Service Oriented Architecture (SOA) is implemented using an ESB (Enterprise Service Bus). This technology platform integration is now developed as part of ObjectWeb community through the Petals² project. In [9] an ESB is defined as a platform to manage the joint use of applications shared by the partners.

The control of processes associated with this partnership can also be provided through the bus and its workflow management tools. The bus plays finally the role of mediator between the partners performing the functions of connection and access management. ESB is primarily a tool of asynchronous exchange with standardized interfaces (SOAP, JMS...) or integrated (JBI...). It can also provide added value services (translation, processing ...) activated by events. Currently, the challenge is the construction of UML profiles as a technical architecture. Some works has been done in this Optical (PIM4SOA project) and a UML profile for SOA was performed [8].

4 Application Case

The National Fund of Social Insurance (CNAS) was created by Executive Decree No. 92-07 of January 4, 1992 to reorganize the social security system. Throughout the national territory, it is represented by CNAS agency. In terms of services, to each agency, are attached payment structures, named: Paying centers, which insured persons, are affiliated. The benefits provided for the reimbursement covers the following risks: *Sickness, Maternity, Disability, Accident and Occupational Disease and Death*.

The fund aims to modernize gradually its IS because the logistics hardware, human resources, programs training, rules, procedures and regulations, in a word, all the IS was mobilized to ensure the quality of the service provided against population of insured persons , which is the essence of the existence of an insurance fund.

It turns out that the current CNAS IS, is characterized by the availability of certain useful information, updated through the web portal, but do not provide information on its business results. A functional partitioning slowing making any decision was being noted on existing applications.

Lead the activity by focusing on business processes from the beginning to the end requires a cross approach beyond the borders of the departments. These processes involve multiple actors and systems, which are actually divided into different functional zones, but often interact in procedures belonging to the same chain of value of the company. This kind of management is that the fund intends to undertake, a mode which highlights the idea that the fund may be a business oriented enterprise.

Note the existence of two main components of the macro business processes of two branches of the functional fund namely:

- The recovery of dues through the population of employers.
- Benefits for reimbursement of the insured population.

² <http://petals.objectweb.org>

The latter is divided into two basic business processes:

- Reimbursement of medical expenses.
- Reimbursement of work stoppages.

Both of them represent the main business in payment structures of the fund. They include the following steps: control rights to benefits/services, liquidation, validation and payment of the file. A mapping of these processes is illustrated in (Fig.1).

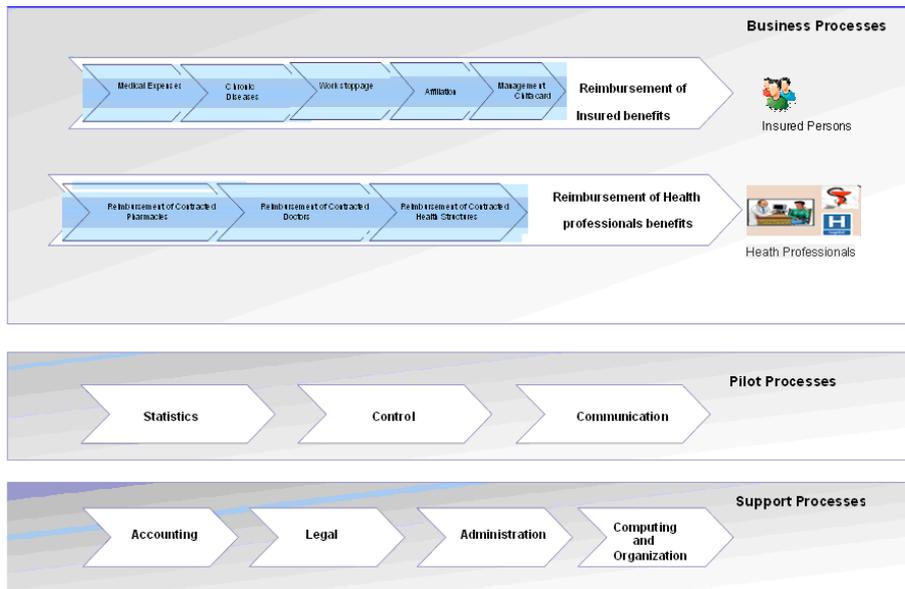


Fig. 1. Mapping of business processes relating to the operation of the fund.

In (Fig.1), all processes relating to the operation of the fund are mapped. Business processes include the macro processes: reimbursement of benefits for insured persons and health professionals. Pilot processes include statistics, control and communication. Supporting processes like accounting and legal.

However, we claim that rethinking the architecture of the IS can be done progressively. The business process “*Reimbursement of Medical Expenses*” was taken as pilot driver business process throughout the organization that accompanies because it is a major element in the right functioning of the fund. It is also a fundamental indicator to the services provided quality. And through it, the urbanization process of the fund is presented below.

5 Our Urbanization Approach

We adopt the **PRAXEME**³ method in our process of urbanization. It is an enterprise methodology, which aims to be open public source, for design or redesign of IS, covering all aspects of the business from strategy to software development. To represent the company and embrace all angles of appreciation, the method is based on a diagram identifying and articulating eight aspects (Fig. 2).

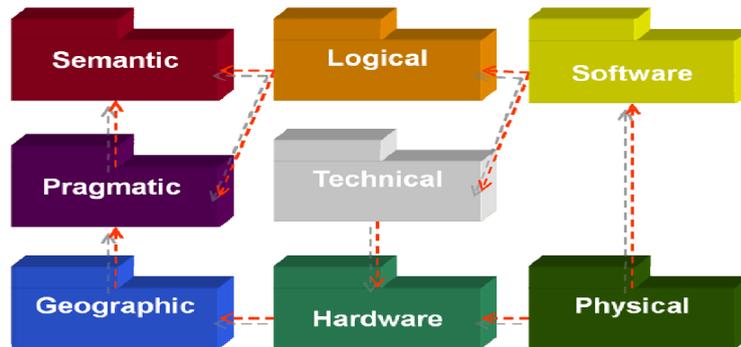


Fig. 2. Enterprise System Topology [10]

The aspect or facet is a view where the system is seen as a type of particular concern. The aspect, which is a component of system, has a relative nature, it is linked to a point of view, a kind of concern, a specialization.

Step 1: (Semantic Aspect) in this step, the objects and concepts at the heart system are described. We express in this what phase is most stable in our IS. We to find the semantic class (object class), the informative properties (attributes), and active properties (operations), such as: Class of Insured persons in our case. The semantics business model is developed in terms of packages, modeled through diagrams: field, class, transition and states of UML. A data repository is designed, on the studied business.

Step 2: (Pragmatic Aspect) at this second stage, the actors of IS are delineated and then their types and functions. The pragmatic modeling will define the management style, command structure and operation in organizational entities, business processes and user profiles, such as business process studied in our case. In other words, in terms of use, it is to identify functional areas and in terms of organization, it is to identify the business processes.

³ www.praxeme.org

Step 3: (Logical Aspect) the result of the previous two steps is projected on an SOA model. This last is compatible with the reference architecture in IBM's SOA [12]. The SOA model is composed of three layers:

- *SOA business layer*: is the business services component process “Reimbursement of Medical Expenses”.
- *SOA applicative logic layer*: consists of bus service with ESB Service Registry allows services to be published, sought and relied upon.
- *SOA component layer*: invokes the service components involved, which run at the server level where they are, the methods implemented by objects grouped in the components.

These are transcriptions by layer of semantic and pragmatic descriptions, transcriptions guided by the architectural structure rules.

Step 4: (Technical Aspect) in this step we move from an SOA logical model to an ESB technical model. For this we need to define a UML model for USB. This model will contain the basic elements that : define an ESB as that "Bus", "XML Message", "Directory services ", " service address ", etc.. The material aspect is closely related to the technical aspect because the choice of the hardware architecture used is set in this level.

Step 5: (Physical Aspect) in this step we focus on a transformation from a model to a text. The goal is generated from the ESB model set of text representations necessary for setup and the implementation of the ESB platform. We distinguish BPEL representation (Business Process Execution Language) for orchestration of services, XML representation for presenting messages exchanged between web services and structure, a representation of WSDL (Web Services Description Language) for describing web services, their addresses, and so on. . And thus set the rules for locating software components namely web services on the hardware architecture and the covering geographical aspect.

6 System Architecture

The IS architecture proposed below (Fig. 3) is a layered architecture, consistent with the SOA reference architecture of the IBM [12]. It distinguishes the business from the application, which the application architecture is an SOA implemented by the bus service (ESB) and packaged web service.

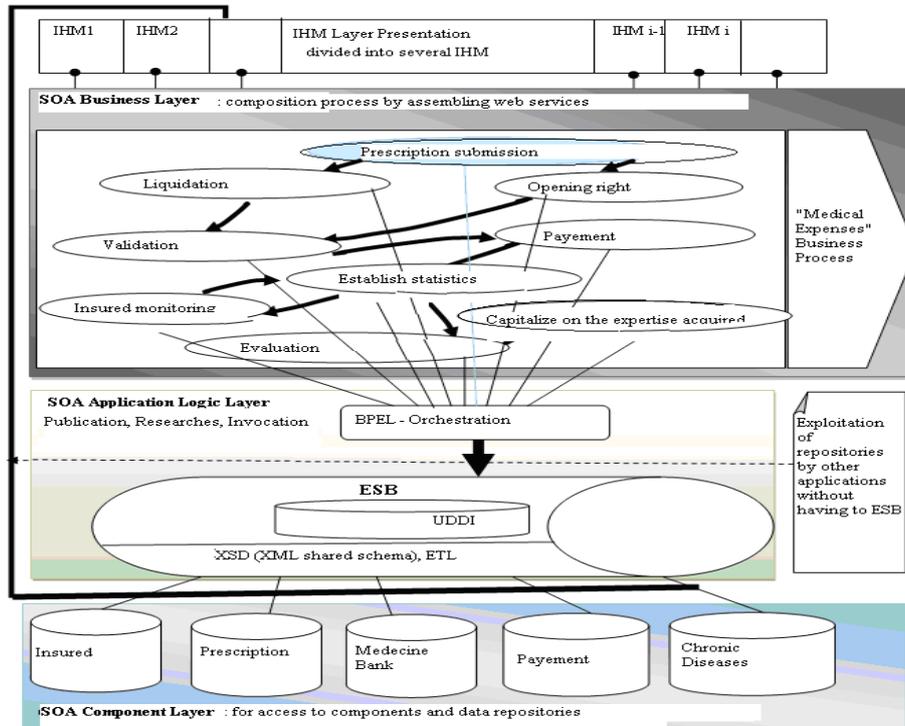


Fig. 3. Information System Architecture based on SOA

The business layer is the business services up our process "Reimbursement of Medical Expenses". Business services invoke services made in the register of services via an orchestration calls through the engine process BPEL.

Once a service is called, in his turn invokes services affected components that run in the servers where they are located the implemented methods by the objects grouped in the components. The implementation is fully distributed, synchronization is provided by bus services features which is doted of the transport function.

The proposed architecture (Fig. 3) provides substantial improvement in IS of the fund to support the business processes studied by:

Promoting agility: because it allows structure in a dynamic relative IS. Indeed, any change that can be made on the process, in the future, is easily manageable, because of the separation between the business and application.

Improving accessibility: facilitate communication between the fund and other companies or partners such as health professionals, through the use of Web services. And thus ensures the sustainability of the business solution implemented.

The implementation is fully distributed, synchronization is provided by the bus services (ESB). Therefore, the proposed solution offers a significant time saving for any extension of the process itself or generalization of SOA to the other processes.

7 Conclusion

We have presented in this paper an urbanization approach of IS. This approach is oriented services. We are inspired by the Praxeme methodology with their different aspects in order to stop the steps for our approach. This was validated by the application of this approach to our practical case of the IS on the field of social insurance benefits, specifically business process: reimbursement of medical expenses.

The work of the methodological point of view could be a cornerstone in the edifice of opens development in a progressive business-oriented enterprise, such as: CNAS. Contributions of SOA application in this case, manifested themselves explicitly via the flexibility of its business processes and the opening of its IS about the outside world through the Web services exposed so that it can be interactive on the web and allow businesses, employers and insured persons to be contributors. In the near future, we intend to generalize the application of SOA in the enterprise, by its propagating on other processes, following the approach presented. Also, think about designing a specific ontology to the social insurance field that would be beneficial in the semantic aspect of our approach. The use of domain ontology can facilitate the reuse of information in the IS. Reuse is a strategic problem for reducing costs and improving methods of design and development in our IS. Finally, the combination of SOA and Web 2.0 could be a useful extension for our work and for all service oriented IS.

References

1. Longép , C.: The Urbanization Project of Information System (2006)
2. Choay, F., Merlin, P.: Dictionary of Planning and Development, PUF (1996)
3. Proceedings of Spring School of Enterprise Modeling. GDR-MACS-Albi (2002)
4. G nzel, F., Wilker, H.: Patterns in Business Models : a Case Survey (2009)
5. Alami's, A.: The Service-Oriented Architecture and J2EE. Modeling the Process of Urbanization. <http://soaj2ee.blogspot.com/urbanisme> (2010)
6. Mougin, Y.: Mapping Process (2004)
7. Gypsies, D., Ouzennou., F.: Identification Process Processes-Approach. Engineering School Mohammadia Rabat, CPI'2007 - Rabat, Morocco (2007)
8. Benguria, G., Larrucea, X., Elveseater, B., Neple, T., Beardsmore, A., Friess, M.: A Platform Independent Model for Service Oriented Architectures. 2nd International Conference on Interoperability, IESA'06-Bordeaux (2006)
9. Chappell, D.A.. Enterprise Service Bus. O'Reilly (2004)
10. Bonnet, P. Vauquier, D., Desfray, D.: SOA and Praxeme Method. In : [http://praxeme.club-blog.\(2006\)](http://praxeme.club-blog.(2006))
11. H. Brandenburg, JP. Wojtyna, "The Process Approach Mode Employment (2003)
12. Dodani, MH: SOA 2006: State of the Art, JOT, Vol. 5, No. 8 (2006).
13. Sienou, : Proposition d'un Cadre M thodologique pour le Management Int gr  des Risques et des Processus Entreprise. Th se de Doctorat, Toulouse (2009)
14. Sassoon, J. Urbanisation des SI, Herm s. (1998)

Using Vector Quantization for Universal Background Model in Automatic Speaker Verification

Djellali Hayet¹, Laskri Mohamed Tayeb²

^{1,2} Badji Mokhtar University Annaba Algeria, Computer Science Department^{1,2},
LRS Laboratory¹, LRI Laboratory²
Badji Mokhtar University, P-O Box 12, 23000 Annaba, Algeria

Abstract. We aim to describe different approaches for vector quantization in Automatic Speaker Verification. We designed our novel architecture based on multiples codebook representing the speakers and the impostor model called universal background model and compared it to another vector quantization approach used for reducing training data. We compared our scheme with the baseline system, Gaussian Mixtures Models and Maximum a Posteriori Adaptation. The present study demonstrates that the multiples codebook gives more verification accuracy called equal error rate but this improvement also depends on the codebook size.

Keywords: Vector Quantization, Speaker Verification, Codebook, false Acceptance, False reject, Universal Background Models, Linde Buzo Gray.

1 Introduction

The speaker verification is a field of speaker recognition which the main objective is to authenticate a person's claimed identity. The speaker voice is used to recognize him (her), we create two models, the first one is the speaker model and the second is the impostor model called universal background model UBM. The recorded speech is preprocessed, compared to speaker and UBM model in order to compute the score and finally compared to threshold.

It has been proved that the variation factors like speaker identity, utterance length, gender, session, transmission channel, speaking, affect the system performance [1][2][3]. Intra speaker variability influences the verification performance system. Thus, it is important to record each speaker at different time but also means the huge speech data.

The state of the art of text independent speaker recognition is Gaussian mixture model and Maximum a posteriori adaptation. Speaker dependent GMM are derived from the speaker independent model called universal background model (UBM) and Maximum a posteriori adaptation MAP using target speaker speech data.

Vector Quantization (VQ) model was introduced in 1980's used in data compression [4]. VQ is one of the simplest text independent speakers model, and often used for computational technique. It also provides good accuracy when combined with background model adaptation [4][5].

In VQ based speaker recognition, each speaker is characterized with the set of code vectors and is referred to as that speaker's codebook. Normally, a speaker's codebook is trained to minimize the quantization error for the training data from that speaker. The most commonly used training algorithm is the Linde-Buzo-Gray (LBG) algorithm [6].

When the speaker speech data becomes huge, it involves the time consuming problem. Gurmeet replaced the EM algorithm with LBG algorithm. Experimentally, they found that the complexity of calculation can be reduced by 50% compared to the EM algorithm. The reason is the LBG algorithm utilize apart of feature vectors for classification [7].

We applied Vector Quantization in Automatic Speaker Verification; usually, each target speaker had his own codebook, when usually the speaker independent models had two gender dependent codebook originates from impostor speakers (male, female).

Our approach aim to select the best universal background model UBM, we try another way to model VQ UBM with set of sub UBM. We divide the features vectors extracted from processing step (Mel cepstral coefficients: MFCC) in a equal size and applied for each of them the LBG algorithm to obtain its codebook (cd1,cd2,...cdK).. The aim is to get the best sub model with LBG algorithm for impostors (UBM) and then compute the distortion error from optimal Sub UBM. We aim to reduce EER in the presence of small training data of each client and select the best sub UBM.

We organized paper as follows, modeling speakers based on vector quantization and MAP adaptation is introduced in Section 2, and the ASV architecture proposed in Section 3 followed experiments in Section 4 and conclusion in section 5.

2 Vector Quantization and MAP Adaptation

We introduce vector quantization and Maximum a posteriori adaptation in Automatic Speaker Verification:

2.1 Vector Quantization

Vector Quantization (VQ) is a pattern classification technique applied to speech data to form a representative set of speaker features. It was introduced to speaker recognition by Soong [8]. In speaker verification, Vector quantization (VQ) model were applied in Soong and Rosenberg, It is one of the simplest text-independent speaker models and usually used for computational speed-up techniques, it also provides competitive accuracy when combined with background model adaptation [5][8][9][10].

In the training phase, a speaker-specific VQ codebook is generated for each known speaker by clustering his training acoustic vectors. The distance from a vector to the closet codeword of a codebook is called a VQ distortion [4][11].

In the Test phase, an input utterance of a known voice is vector-quantized using trained codebook from proclaimed identity and the speaker independent model codebook (Universal Background Model). The total VQ distortion is computed.

In principle, when we get a large amount of training vectors representing speaker in the training vectors. We should reduce it by vector quantization. Suppose there are N vectors, to be quantized, the average quantization error is given by

$$\mathbf{E} = \frac{1}{N} \sum_{t=1}^N \mathbf{e}(\mathbf{x} \mathbf{t}) \quad (1)$$

The task of designing a codebook is to find a set of code vectors so that E is minimized. However, the commonly used method is the LBG algorithm [6].

In speaker verification, the codebook is used for classification and minimizing the quantization error. We selected LBG algorithm defined as the iterative improvement algorithm or the generalized Lloyd algorithm. Given a set of N training feature vectors, {t1, t2,..,tn} characterizing the variability of a speaker, we search a partitioning of the feature vector space, {S1, S2,.., SM}, for that particular speaker where S, the whole feature space, is represented as S = S1 U S2 U...U SM.

The performance of a quantizer is designed by an average distortion between the input vectors and the final vectors, where E represents the expectation operator (equation 1).

2.2 Gaussian Mixture Models & MAP Adaptation

GMM-UBM-Maximum Likelihood Modeling: this approach is based on training UBM male model with Gaussian mixture model and the other female UBM (from female speech). The model parameters (mean, covariance and weight of the Gaussian) are trained with the EM algorithm (Expectation-Maximization).

Maximum a Posteriori approach MAP resolve the problem of maximum likelihood ML(can't generalize well to unseen speech data in low training data). MAP use prior knowledge of the distribution of the model parameters and insert it in modeling process[12][13]. The Maximum A Posteriori MAP approach is to use the world model and client training data to estimate the client model on the basis of these data and MAP Adaptation [12][13] [14][15][16].

The client model is derived from the world model by adapting the GMM parameters (mean, covariance, weights) estimated. However, experimentally, only the averages of GMM are adapted [13].

3 Speaker Verification Architecture Based on Vector Quantization

We proposed two VQ-UBM models, the first one is the baseline system, the second is VQ Sub-UBM. We describe our new modeling UBM:

3.1 Training Phase

VQ Sub UBM

The acoustics vectors obtained in features extraction were split in subset of data with the same dimension and served to create codebook {CDU1, CDU2, ..., CDUk} for world model UBM. We divide UBM speech data in N subsets instead of one global UBM, in figure 1, after feature extraction, the MFCC vectors were the input of L.B.G algorithm which provide K codebook.

Codebook

There are several different approaches to finding an optimal codebook. The idea is to begin with a vector quantizer and a codebook and improve upon the initial codebook by iterating until the best codebook is found. We aim to reduce redundancy in UBM data by clustering, to do that, we implement this algorithm:

Algorithm 1: VQ Sub-UBM

Training Phase

Input : MFCC vectors; Output: Codebook CDU(1..M).

We divide MFCC vector in equal sub matrix and applied LBG algorithm for each of them.

Input [C] = MFCC vector (Feature Extraction).

Split C in M equal sub matrix Ci;

Train UBM of each Ci for different size of codebook (k=16,32,64,128,256); Result= CDU (i=1..M).

Test Phase

In recognition phase, we compute Euclidean distance and evaluate quantization error from each codebook and test vector,

We choose the best codebook with minimal quantization error.

The quantization square error ESQ

$$MSE(X, Y) = \frac{1}{|X|} \sum_{x_i} \min_k ||x_i - y_k||^2 \quad (2)$$

Where $y_k \in Y$; Xi :vector data; yK : centroid

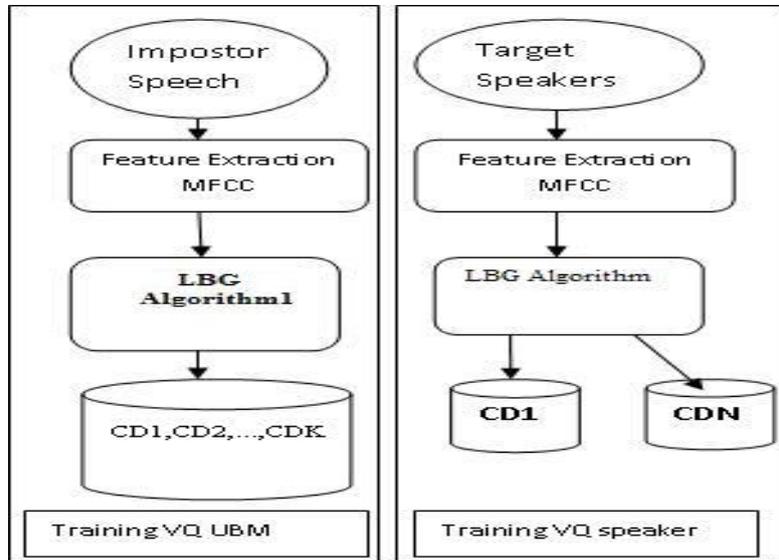


Fig 1. Vector Quantization Architecture: VQ Sub-UBM is applied for UBM only

3.2 Test Phase

We compute the threshold (CDT) from 8 male and 8 females' speakers others than UBM speakers and trained by LBG algorithm.

✓ Test Algorithm

CDU: UBM codebook; CDS: Speaker codebook;

VQdist : VQ distorsion

Input : X=speaker speech ,claimed identity,

MFCC= Feature Extraction(X)

For i=1 to M VQcdu=VQdistorsion(X,CDUi) End

CDUoptimal=CDU best cobdebook UBM where
 Argmin(VQdistorsion(X,CDUi))

VQdist(speaker) = VQdist(X,CDS) - Vqdist(X,CDUoptimal)

If VQdist(speaker)> VQdist(CDT) then client acces

Else reject

4 Protocol Experiment

In this section, we describe a set of experiments designed to evaluate the performance of the proposed system under a variety of condition and compare it to baseline system GMM MAP and standard VQ UBM.

4.1 Database and Baseline System

The Arabic database is recorded in Goldwave frequency 16KHz for a period of 60s for each speaker when training and 30s in the testing phase. The UBM population is 15 men's and 15 women. Four sessions are recorded for each speaker at an interval of 1 month. Ten clients are registered in the database (5 men and 5 women).

4.2 VQ Sub-UBM Model

We extract MFCC vector for all acoustics data allowed to UBM training and applied LBG algorithm for it. We obtain one centroid ($N \times T$) by gender, where we try different value of $N=k=16, 32, 64, 128, 256$. In recognition phase, we compute Euclidean distance and evaluate quantization error (equation 1) from centroid and test vector, we computed codebook for each target speaker and finally evaluate the score.

TABLE I. VQ-SUB UBM PERFORMANCES

CodeBook Size	FA(%)	FR(%)
CD32	22,86	23,86
CD64	25,71	23,86
CD128	7,14	22,73
CD256	14,29	22,73

4.3 Baseline VQ UBM Model

We compute one codebook for the Baseline VQ UBM and evaluate LBG algorithm for $k=16, 32, 64, 128$. We built UBM models from 30 Arabic speakers; UBM male with 15 male speakers and UBM female from 15 female speakers. The global threshold is computed from other database: 8 male and 8 female speakers.

TABLE II. BASELINE VQ UBM PERFORMANCES

CodeBook Size	FA(%)	FR(%)
CD32	14,29	73,03
CD64	12,86	4,89

4.4 Baseline GMM MAP system

We train universal background model UBM gender dependent(male, female) under expectation maximization algorithm EM and create each target speaker model with GMM MAP approach, we try different sizes of GMM (8, 16, 32,64,128) and evaluate the value of false acceptance and false rejection.

TABLE III. GMM MAP SYSTEM RESULTS

#Gaussians	8	16	32	64	128
GMM MAP EER %	19.16	36.12	35.2	35	36.04

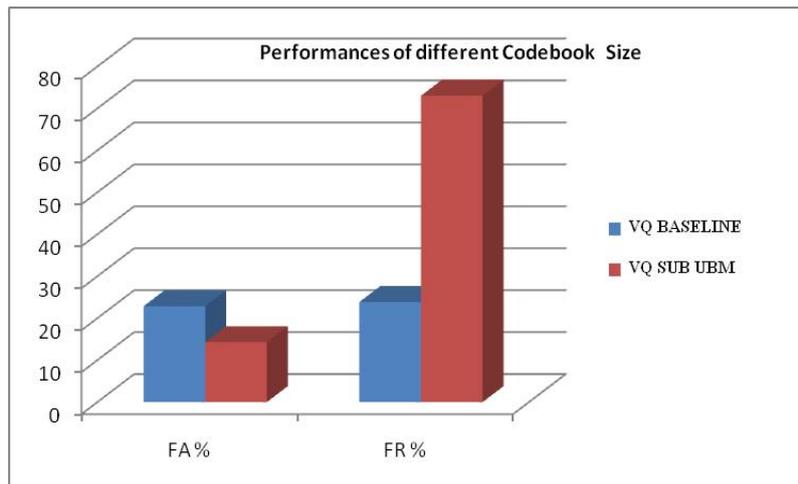


Fig 2. Comparaion of VQ Baseline and VQ SUB UBM

5 Discussions

We compare different modeling speaker techniques: VQ Sub-UBM, Baseline VQUBM and GMM MAP their performances were evaluated using the same data and front end processing.

Table I shows the value of false acceptance and false rejection for different codebook size (32, ..., 256) in VQ SUB UBM approach and observe that the best value is designed for 128 codebook size (7.14% and 22.73%). The result in table 1 provide

more accuracy recognition than table II for codebook size=32(FA=22.86%; FR=23.86%) and worst for codebook size 64. We observe that the size of codebook influences the performance and the multiple UBM provide better result.

Figure 2 demonstrate the performance of VQ SUB UBM is worst than VQ UBM in false rejection, however we tested only VQ UBM with 32 and 64 codebook size.

In Baseline GMM MAP system, Equal error rate is 19.16% for 8 mixtures and between [35%-36.12%] for model order M=16...128. The performances decreases because the reduced speech data and didn't apply normalization technique like Tnorm.

6 Conclusions

VQ SUB UBM achieved (FA=7.14% and FR=22.73%) for 128 codebook size and improved the performance of vector quantization applied in speaker verification compared to baseline vector quantization. The codebook size influences the verification accuracy. The size of speech data should be increased in order to validate our experiments in large database.

References

1. Campbell, J.: Speaker Recognition, A Tutorial. Proc. IEEE 85 (9), pp. 1437--1462 (1997)
2. Reynolds D. A, Rose R. C. : Robust Text Independent Speaker Identification Using Gaussian Mixture Speaker Models, IEEE Trans. Speech Audio Processing, vol. 3, pp. 72-- 83 (1995)
3. Doddington, G., Liggett, W., Martin, A., Przybocki, M., Reynolds, D.A.. : Sheeps, goats, lambs and wolves., : A Statistical Analysis of Speaker Performance in the NIST 1998 Speaker Recognition Evaluation. In Proc. of ICSLP (1998)
4. Wan-Chen C., Ching-Tang H., Chih-Hsu H., : Robust Speaker Identification System Based on Two-Stage Vector Quantization, Tamkang Journal of science and engineering, Tamkang Journal of Science and Engineering, Vol. 11, No. 4., pp. 357-- 366 (2008)
5. Jialong H, Li L, Gunther P, : A Discriminative Training Algorithm for VQ-Based Speaker Identification. IEEE Transactions on Audio and Signal Processing, vol 7, (1999)
6. Linde Y., Buzo A., Gray R.M., : An Algorithm for Vector Quantizer Design," IEEE Trans. Commun., vol. 20, pp. 84 --95 (1980)
7. Gurmeet S, Panda S, Bhattachryya S. Srikanthan S., : Vector Quantization Technique for GMM Based Speaker Verification. IEEE International conference on acoustics speech and signal processing, USA. pp. 65 -- 68 (2003)
8. Soong F.K., Rosenberg A.E, Rabiner L.R, Juang B.H 1985, : A Vector Quantization Approach to Speaker Recognition. IEEE International Conference on Acoustics speech and signal Processing, pp. 387 -- 390 (1985)
9. Rosenberg A. E., Soong F. K. : Evaluation of a Vector Quantization Talker Recognition System in Text Independent and Text Dependent Modes. Comput. Speech Lang, vol .22, pp. 143 -- 157 (1987)
10. Jenq-Shyang P, Thesis: Improved Algorithms For VQ Codeword Search, Codebook Design and Codebook Index Assignment. University of Edeinburgh (1996)

11. Kinnunen, T., Saastamoinen, J., V., Hautomaki, M., Vinni, P., Franti, : Comparing Maximum a Posteriori Vector Quantization and Gaussian Mixture Models in Speaker Verification, Pattern recognition letters, (2008)
12. Preti, A. : Thesis, Surveillance de Réseaux Professionnels de Communication par la Reconnaissance du Locuteur. Académie d'Aix Marseille, Laboratoire d'informatique d'Avignon (2008)
13. Bimbot, F., Bonastre, J.F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., Merlin, T., Ortega-Garcia, J., Petrovska-Delacretaz, D., Reynolds, D.A. : A tutorial on Text- Independent Speaker Verification. J. Appl. Signal Process. 4. pp. 430 -- 451 (2005)
14. B Vesnicer, F Mihelic., : The Likelihood Ratio Decision Criterion for Nuisance Attribute Projection in GMM Speaker Verification, Hindawi Publishing Corporation Eurasip Journal On advances in Signal Processing volume (2008)
15. Reynolds, D.A. Quatieri, T.F. Dunn, R.B. Speaker Verification Using Adapted Gaussian Mixture Models. Digital Signal Processing. 10. pp. 19 -- 41 (2000)
16. Furui S. : Speaker Dependent Feature Extraction, Recognition and Processing Techniques .NTT Human interface Laboratories, Japan Speech Communication , Elsevier Science Publishers North-Holland. pp. 505--520 (1991)

Learning and classification



The Use of WordNets for Multilingual Text Categorization: A Comparative Study

Mohamed Amine Bentaallah and Mimoun Malki

EEDIS Laboratory, Department of computer sciences
Djillali Liabes University
Sidi Bel Abbas, 22000. ALGERIA
mabentaallah@univ-sba.dz
malki-m@yahoo.com
<http://www.univ-sba.dz>

Abstract. The successful use of the Princeton WordNet for Text Categorization has prompted the creation of similar WordNets in other languages as well. This paper focuses on a comparative study between two WordNet based approaches for Multilingual Text Categorization. The first relates on using machine translation to access directly the Princeton WordNet while the second avoids machine translation by using the WordNet associated for each language.

Key words: Multilingual, Text Categorization, WordNet, Ontology, Mapping

1 Introduction

With the rapid emergence and proliferation of Internet and the trend of globalization, a tremendous number of textual documents written in different languages are electronically accessible online. Efficiently and effectively managing these textual documents written in different languages is essential to organizations and individuals. This necessity gave birth to a new domain of research that is the Multilingual Text Categorization.

The growing popularity of the Princeton WordNet as a useful resource for English and its incorporation in natural language tasks has prompted the creation of similar WordNets in other languages as well. Indeed, WordNets for more than 50 languages are currently registered with the Global WordNet Association¹. In this paper we try to answer the question: "*Will the use of these WordNets in Text Categorization guarantee good results better than those obtained by the Princeton WordNet ?*".

The rest of the paper is organized as follows. In section 2, we review some related works for Multilingual Text Categorization. In section 3, we describe the two approaches to be compared. Section 4 presents the experiments and the results. Finally, conclusion and future works are reported in section 5.

¹ <http://www.globalwordnet.org>

2 Multilingual Text Categorization

Multilingual Text Categorization(MTC) is a new area in Text categorization in which we have to cope with two or more languages (e.g English, Spanish and Italian).

MTC is a relatively new research topic, about which not much previous work in the literature appears to be available. Most approaches have mainly addressed different translation issues to solve the problem. R.Jalam et al. presented in [1] three approaches for MTC that are based on the translation of documents toward a language of reference. Rigutini et al. used in [2] a machine translation system to bridge the gap between different languages. The major disadvantage of Machine translation based approaches is the absence of machine translation systems for many language pairs and the wide gap between the translated documents and original documents.

In order to overcome the disadvantage of using machine translation systems, many researches have been working on using linguistic resources such as bilingual dictionaries and comparable corpora to induce correspondences between two languages. A.Gliozzo and C.Strapparava propose in [4] a new approach to solve the Multilingual Text Categorization problem based on acquiring Multilingual Domain Models from comparable corpora to define a generalized similarity function (i.e. a kernel function) among documents in different languages, which is used inside a Support Vector Machines classification framework. The results show that the approach largely outperforms a baseline. K.Wu et al. proposed in [3] a novel refinement framework for cross-language text categorization investigating the use of a bilingual lexicon to identify a novel model called domain alignment translation model. Their approach can achieve comparable performance with the machine translation approach using the Google translation tool, although their experiments only consider the word level but ignore the base phrase.

These last years, researches showed that using ontologies in monolingual text categorization is a promising track. J.Guyot proposed in [9] a new approach that consists in using a multilingual ontology for Information Retrieval, without using any translation. He tried only to prove the feasibility of the approach. Nevertheless, it still has some limits because the used ontology is incomplete and dirty. Intelligent methods for enabling concept-based hierarchical Multilingual Text Categorization using neural networks are proposed in [13]. These methods are based on encapsulating the semantic knowledge of the relationship between all multilingual terms and concepts in a universal concept space and on using a hierarchical clustering algorithm to generate a set of concept-based multilingual document categories, which acts as the hierarchical backbone of a browseable multilingual document directory. We have proposed in [10] a new approach for MTC based on spreading the use of WordNet in Text Categorization towards MTC in order to reduce noises introduced by machine translation.

3 Description of the two proposed approaches

As shown in figure 1, the two approaches are composed of three phases:

-
- Knowledge representation step;
 - Training step;
 - Predicting step.

For our experiments, the two approaches have the same training and prediction phases. The only difference is on the knowledge representation phase.

3.1 Knowledge representation

First approach The first approach consist on representing knowledge with the use of the Princeton WordNet. The labelled documents are mapped directly into the synsets of the princeton WordNet since they are expressed in English language. The unlabelled documents needs to be translated into the English language in order to be able to be mapped to the Princeton WordNet. The mapping into the princeton WordNet consists in replacing each term in a document by its most common meaning from the Princeton WordNet. We used a simple disambiguation strategy that consists of considering only the most common meaning of the term (first ranked element) as the most appropriate. Thus the synset frequency is calculated as indicated in the following equation:

$$sf(c_i, s) = tf(c_i, \{t \in T \mid first(Ref(t)) = s\}) \quad (1)$$

where:

- $tf(c_i, T')$: the sum of the frequencies of all terms $t \in T'$ in the train documents of category c_i .
- $Ref(t)$: the set of all synsets assigned to term t in WordNet.

Second approach The second approach excludes the direct use of machine translation techniques by incorporating the WordNet associated for document languages. Indeed, each term document will be firstly mapped to the WordNet synsets of the language in which the document is expressed. As result, the labelled documents and the unlabelled documents will be mapped on different taxonomies. The labelled documents will be mapped to the Princeton WordNet, and the unlabelled documents will be mapped to the WordNets associated to unlabelled documents languages. It is necessary to match the taxonomies of all the used WordNets to a common taxonomy in order to unify document representations. Since the Princeton WordNet is the richest taxonomy, we have chosen it to be the common taxonomy. This matching offers the following advantages:

- Avoiding the direct use of machine translation techniques which eliminate the problem of translation disambiguation.
- Interconnecting the different WordNets to the most rich WordNet (Princeton WordNet) which resolves the richness of some WordNets.

Formally, the synset frequency is calculated as indicated in the following equation:

$$sf(d, s) = tf(d, \{t \in T \mid match(first(Ref(t, L))) = s\}) \quad (2)$$

where:

- $tf(d, T')$: the sum of the frequencies of all terms $t \in T'$ in the unlabelled document d .
- L : The language of the unlabelled document d .
- $Ref(t, L)$: the set of all synsets assigned to term t in WordNet associated to language L .
- $match(s)$: the corresponding synset of the synset s on the Princeton WordNet.

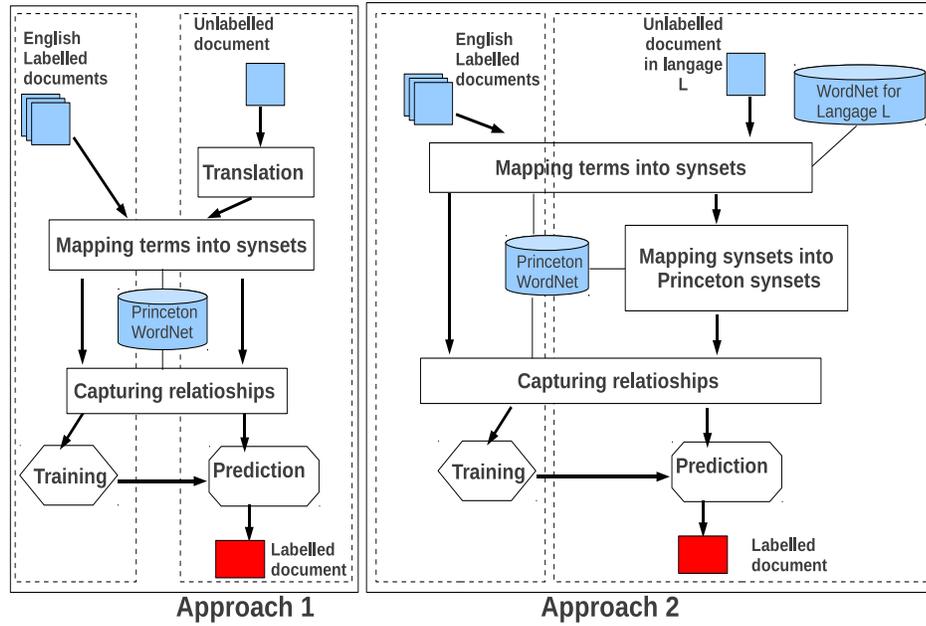


Fig. 1. The two compared approaches

Capturing relationships After mapping terms into Princeton WordNet synsets, this step consists in using the WordNet hierarchy to capture some useful relationships between synsets (hypernymy in our case). The synset frequencies will be updated as indicated in the following equation:

$$sf(c_i, s) = \sum_{b \in H(s)} sf(c_i, b) \quad (3)$$

Where:

- c_i : the i^{th} category.
- b and s are synsets.
- $H(s)$: the hyponyms set of synset s

3.2 Training

The training phase consists in using the labelled documents to create conceptual categories profiles. Formally, each category will be represented by a conceptual profile which contains the K better synsets (our features) characterizing best the category compared to the others. For this purpose we used the χ_2 multivariate statistic for feature selection. The χ_2 multivariate [24], noted $\chi_2^{multivariate}$ is a supervised method allowing the selection of features by taking into account not only their frequencies in each category but also interaction of features between them and interactions between features and categories. Given the matrix (synsets-categories) representing the total number of occurrences of the p synsets in the m categories. The contributions of these synsets in discriminating categories are calculated as indicated in the following equation, then sorted by descending order for each category.

$$C_{jk}^{\chi_2} = N \frac{(f_{jk} - f_{j \cdot} f_{\cdot k})^2}{f_{j \cdot} f_{\cdot k}} \times \text{sign}(f_{jk} - f_{j \cdot} f_{\cdot k}) \quad (4)$$

Where:

- $f_{jk} = \frac{N_{jk}}{N}$: the relative occurrence frequency.
- N : The total sum of the occurrences.
- N_{jk} : The frequency of the synset s_j in the category c_k .

Once the contributions of synsets are calculated and ordered for each category, the conceptual profile of each category contains the k first sorted synsets.

3.3 Prediction

The Prediction phase consists on using the conceptual categories profiles in classifying unlabelled documents. Our Prediction phase consists of:

- Weighting the conceptual categories profiles and the conceptual vector of the unlabelled document. In our experiments, we used the standard *tfidf* (term frequency - inverse document frequency) function [25], defined as:

$$w(s_k, c_i) = \text{tfidf}(s_k, c_i) = \text{tf}(s_k, c_i) \times \log\left(\frac{|C|}{df(s_k)}\right) \quad (5)$$

Where:

- $\text{tf}(s_k, c_i)$ denotes the number of times synset s_k occurs in category c_i .
- $df(s_k)$ denotes the number of categories in which synset s_k occurs.
- $|C|$ denotes the number of categories.
- Calculating distances between the conceptual vector of the document and all conceptual categories profiles and assigning the document to the category whose profile is the closest with the document vector. In our experiments, we used the dominant similarity measure in information retrieval and text classification which is the cosine similarity that can be calculated as the normalized dot product:

$$S_{i,j} = \frac{\sum_{s \in i \cap j} \text{tfidf}(s,i) \times \text{tfidf}(s,j)}{\sqrt{\sum_{s \in i} \text{tfidf}^2(s,i) \times \sum_{s \in j} \text{tfidf}^2(s,j)}} \quad (6)$$

With:
s: a synset,
i and *j*: the two vectors (profiles) to be compared.
tfidf(*s, i*): the weight of the synset *s* in *i*.
tfidf(*s, j*): the weight of the synset *s* in *j*.

4 Experimental results

4.1 Dataset for evaluation

For our experimentations, we extracted a bilingual dataset from Reuters Corpus Vol. 1 and 2 (RCV1, RCV2) using English training (RCV1) and Spanish test documents (RCV2). Our dataset is based on topic (category) codes with a rather varying number of documents per category as shown in Table1

Table 1. The 8 used Categories of the Multilingual Reuters corpus

Code category	Category Description	English labelled documents	Spanish unlabelled documents
C183	Privatisations	200	205
GSPO	Sport	401	84
GDIS	Disaster	278	116
GJOB	labour issues	401	197
GDEF	Defence	227	83
GCRIM	Crime, Law enforcement	401	157
GDIP	International relations	401	237
GVIO	War, Civil war	401	306

4.2 Results

For comparison, we have tested the two approaches on our multilingual dataset. Experimental results reported in this section are based on the so-called "F₁ measure", which is the harmonic mean of precision and recall.

$$F_1(i) = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (7)$$

The results of the experimentations are presented in Table2, Concerning the profiles size, the best performances are obtained with size profile *k* = 900 for the two approaches. Indeed, the performances improve more and more by increasing the size of profiles.

Comparing the results of the two approaches, the first approach largely outperform the second approach.

Table 2. Comparison of F-score results on the two approaches

Size of profiles	Approach1	Approach2
k=100	0.586	0.201
k=200	0.608	0.213
k=400	0.621	0.219
k=500	0.509	0.222
k=700	0.634	0.221
k=900	0.639	0.268

5 Conclusion

In this paper, we have compared two approaches for using WordNets for MTC. The first approach is based on using machine translation to use the Princeton WordNet while the second approach is based on replacing the use of machine translation by incorporating a WordNet for each language. The results of the experimentations show that the use of WordNets does not guarantee good results rather than those obtained by the Princeton WordNet. Future works will concern the experimentation of the second approach with different WordNets in order to be able to confirm the obtained results.

References

1. Jalam, R., Clesh, J., Rakotomalala, R.: Cadre pour la catégorisation de textes multilingues. 7^{èmes} Journées internationales d'Analyse statistique des Données Textuelles. Louvain-la-Neuve, Belgique (2004) 650–660
2. Rigutini, L., Maggini, M., and Liu, B.: An EM based training algorithm for Cross-Language Text Categorization. Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence. Compiegne, France. September 2005.
3. Wu, K., Lu, B.: A Refinement Framework for Cross Language Text Categorization. 4th Asia Information Retrieval Symposium, AIRS 2008, Harbin, China, (2008) 401–411
4. Gliozzo, A.M., Strapparava, C.: Cross Language Text Categorization by acquiring Multilingual Domain Models from Comparable Corpora. in Proceedings of the ACL Workshop on Building and Using Parallel Texts. Ann Arbor, Michigan, USA (2005) 9–16
5. Adeva, J. J., Calvo, R. A., and Ipiá, D.: Multilingual Approaches to Text Categorisation. The European Journal for the Informatics Professional, Vol 6, (2005) 43–51
6. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys, (2002) 1–47
7. Peters, C., Sheridan, P.: Accès multilingue aux systèmes d'information. In: In 67th IFLA Council and General Conference. (2001)
8. Nunberg, G.: Will the Internet speak english?. The American Prospect. (2000)
9. Guyot, J., Radhouani, S., Falquet, G.: Ontology-based multilingual information retrieval. In CLEF Workshop, Working Notes Multilingual Track, Vienna, Austria (2005) 21–23

-
10. Bentaallah, M.A., Malki, M.: WordNet based Multilingual Text Categorization. *Journal of Computer Science*, Vol 6 (2007)
 11. Hu, W., Jian, N., Qu, Y., Wang, Y.: Gmo: A graph matching for ontologies. In *Proceedings of the K-CAP 2005 Workshop on Integrating Ontologies*, (2005) 41–48
 12. Lacher, M.S., Groh, G.: Facilitating the exchange of explicit knowledge through ontology mappings. In *Proceedings of the 14th International Florida Artificial Intelligence Research Society Conference (FLAIRS01)*, AAAI Press, (2001) 305–309
 13. Chau, R., Yeh, C.H, Smith, K.: A Neural Network Model for Hierarchical Multilingual Text Categorization. In *proceeding of ISSN-05 Second International Symposium on Neural Networks*, Chongqing, China (2005) 238–245
 14. Chau, R., Yeh, C.: Multilingual Text Categorization for Global Knowledge Discovery Using Fuzzy Techniques. *Proceedings of the 2002 IEEE International Conference on Artificial Intelligence Systems (ICAIS)*, (2002) 82–86
 15. Ichise, R., Hamasaki, M., Takeda, H.: Discovering relationships among catalogs. In E. Suzuki and S. Arikawa, editors, *Proceedings of the 7th International Conference on Discovery Science (DS04)*, volume 3245 of LNCS, Springer, (2004) 371–379
 16. Nottelmann, H., Straccia, U.: A probabilistic, logic-based framework for automated web directory alignment. In Zongmin Ma, editor, *Soft Computing in Ontologies and the Semantic Web*, *Studies in Fuzziness and Soft Computing*, Springer Verlag, (2006) 47–77
 17. Miller, G.A.: WordNet: An On-Line Lexical Database. In *Special Issue of International Journal of Lexicography*, Vol 3, No. 4 (1990) 238–245
 18. Furst, F., Trichet, F.: Axiom-based ontology matching. In *Proceedings of the 3rd international conference on Knowledge capture (K-CAP 05)*, ACM Press, (2005) 195–196
 19. Do, H.H., Rahm, E.: Coma - a system for flexible combination of schema matching approaches. In *Proceedings of the 28th International Conference on Very Large Data Bases (VLDB 02)*, (2002) 610–621
 20. Gonzalo, J., Verdejo, F., Chugur, I., Cigarran, J.: Indexing with WordNet synsets can improve text retrieval. In: *Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*. (1998).
 21. Ide, N., Veronis, J.: Introduction to the special issue on word sense disambiguation: The state of the art. *Computational Linguistics*. (1998).
 22. Rahm, E., Bernstein, P.A.: A survey of approaches to automatic schema matching. *The VLDB Journal*, 10(4), (2001) 334–350
 23. Shvaiko, P., Euzenat, J.: A survey of schema-based matching approaches. *Journal on Data Semantics*, 4(LNCS 3730), (2005) 146–171
 24. Yang, Y., Pederson, J.O.: A comparative study on feature selection in text categorization. In *Proceedings of ICML-97, 14th International Conference on Machine Learning*. (1997) 412–420
 25. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. *Information Processing and Management*. (1988) 513–523
 26. Salton, G.: *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley (1989)

Enhanced Collaborative Filtering to Recommender Systems of Technology Enhanced Learning

Majda Maâtallah and Hassina Seridi-Bouchelaghem

LABGED Laboratory, University Badji Mokhtar Annaba, Po-Box 12, 23000, Algeria

majda.maataallah@univ-annaba.org, seridi@labged.net

Abstract. Recommender Systems (RSs) are largely used nowadays in many areas to generate items of interest to users. Recently, they are applied in the Technology Enhanced Learning (TEL) field to let recommending relevant learning resources to support teachers or learners' need. In this paper we propose a novel recommendation technique that combines a fuzzy collaborative filtering algorithm with content based one to make better recommendation, using learners' preferences and importance of knowledge to recommend items with different context in order to alleviate the Stability vs. Plasticity problem of TEL Recommender Systems. Empirical evaluations show that the proposed technique is feasible and effective.

Keywords: Technology-Enhanced Learning, Recommender Systems, Collaborative Filtering, Content Based Filtering, Learner Profile, Fuzzy C-means, Matrix Factorization.

1 Introduction

Web development has created a need for new techniques to help users find what they want and also to know that information exists, these techniques are called Recommender Systems (RSs). These systems are built generally based on two different types of methods that are Content Based Filtering (CBF) and Collaborative Filtering (CF). RSs suffer from several problems defined in [1], where one of them is the problem of the system's stability compared to the user's profile dynamicity (Dynamicity vs. Plasticity Problem) [1]. This problem comes from the system's incapability to track the user's behavioral evolution, because in RSs once a user's profile has been established, it is difficult to change it. RSs are widely used in many areas, especially in e-commerce [2],[3]and[4]. Recently, they are applied in the e-learning field, more specifically in Technology Enhanced Learning (TEL) [5], in order to personalize learning content and connect suitable learners with each other according to their individual needs, preferences, and learning goals.

TEL can be differentiated into formal and non-formal learning settings. In non-formal learning, the learners are acting more self-directed and they are responsible for their own learning. The learning process is not designed by an institution or responsible teachers like in formal learning, but it depends on individual learners' preferences or choices, which is similar to consumers looking for products on the internet. So, lifelong learners are need to have an overview of the available learning activities and

materials to decide which of them match better their personal needs, preferences, prior knowledge and current situation. Where the need to use Personalized Recommender Systems (PRS) that use efficiently the available resources in the network and propose learning resources and activities depending on individual needs, learning goals, context, and increase collaboration between learners. But the learner's need and preferences may change over period of time, also in the same time where he wants to learn from resources with different context. This creates the need of designing Adaptive RSs (ARS) capable of generating recommendations with different tastes depending on the learner's profile evolution. ARSs design is a great challenge because of the Stability vs. Plasticity problem of these systems.

Whereas recommendations in TEL field depend not only to learner's preferences but on the context as demonstrated in [6]; this makes more and more important the use of CBF in the recommendation process. To this end, we elaborate a hybrid technique that combines between a fuzzy-based CF algorithm and CBF using taxonomic information to generate multi-context recommendations with better performance.

The paper is organized as follows. Section two, presents the RSs field and the third section, contains some works deployed RSs in the TEL field. Then, we outline our proposed fuzzy hybrid technique to recommend learning resources with different tastes, in section four. Empirical results are presented in the fifth section. Finally, we give some conclusions and lines of future work.

2 Recommender Systems

RSs provide adequate information to people in need using a representation of the user called "User Profile". This profile is compared with different profiles available to determine those to which they correspond [7]. So, RSs intend to send from a large amount of information generated dynamically, the information judged relevant. Hence, filtering is interpreted as elimination of unwanted data on an incoming stream, rather than looking for specific data on this flow.

RSs are built generally based on two different types of methods that are Content Based Filtering (CBF) and Collaborative Filtering (CF). The CBF approach generates content recommendations based on the characteristics of users or items, while the CF method just use the evaluations made by users on the items to predict the unknown ratings of new user-item pair. Typical CF algorithms can be categorized into two classes: neighborhood methods and factorization methods. Generally factor-based algorithms are considered more effective than those based on neighborhood. But they are often complementary and the best performance is often obtained by blending them [8]. Hybridization between CF and CBF approaches has been the subject of interest in a lot of works on RSs, to enjoy their benefits.

One of the major problems of RSs is the Stability problem of these systems compared to the dynamic profile of the user (Stability vs. Plasticity Problem) [2]. To overcome this problem, we proposed a hybrid approach that combines between the CB approach that uses taxonomic information to represent the item's content and collaborative approach that uses preferences of similar learners (neighbors) to predict

the active learner's preferences, then, generating diversified recommendations that meet their needs according to his membership degrees to different clusters. These membership values can be obtained in the CF phase by applying the Fuzzy Logic [9], or by applying the Fuzzy C-means algorithm (FCM) [10].

In order to offer all needs of the active learner that fit their different tastes, we propose a fuzzy based clustering algorithm to regroup learners including the active learner, and that guarantees a multi-affectation of learners to nearest clusters allowing them to receive partial recommendations generated in each cluster according to their membership degrees. Due to the two major challenges for the CF based systems, which are the Scalability and Sparsity Problems, the application of traditional FCM algorithm can confront some difficulties. From this point, our goal was to design an efficient CF algorithm that guarantee a multiple assignment of a user to different clusters, by modifying the FCM objective function to a Matrix Factorization (MF) one [11].

3 Background

Many RSs have been deployed in TEL, as surveyed in Manouselis and al.[5], for recommending learning materials and resources to the learners in both formal and informal learning environment [12]. Concretely, Garcia and al.[13] uses association rule in the form of IF-THEN rules to discover information of interest through student performance data, then generating the recommendation based on those rules; Bobadilla and al.[14] had using a CF scheme where they incorporated learners' test score into the item prediction function; Soonthornphisaj and al.[15] applied CF to predict the most suitable learning objects to the learners; Ge and al.[16] have combined between CBF and CF to make personalized recommendation for a courseware selection module. Finally, Thay-Nghe and al.[17] applied the MF technique in the educational context, for predicting student performances. A critical study of recommender techniques regarding to their applicability and usefulness in TEL has presented in [12], providing an overview of advantages and disadvantages of each technique, and report the envisaged usefulness of each one for TEL recommenders. For more details on TEL Recommender Systems please refer to [5].

Generally RSs in e-learning deal with information about the learners and learning activities and would be able to track the evolution of the learner profile (behavior) during his different learning levels. For this aim, we propose a new hybrid technique that combines CF (using MF) with CBF to better predict the learner's need. The proposed technique allows generating learning resources recommendations to lifelong learners that correspond to their different interests, tracking their profiles evolution.

4 Contribution

To improve the recommendation quality, we are conducted toward hybridization between CF and a CBF to enhance the CF accuracy in TEL Recommender Systems in order to deal with the sparsity and scalability problems.

Our proposed approach can be divided into two main phases; the first one contains the description of the fuzzy-based CF algorithm and the CBF one, with their missing scores predictions. Then, it presents the hybrid scheme that blends the two predictions in order to obtain a full learner-course matrix. The second phase contains the recommendation algorithm adapted to TEL field by incorporating the learner's performances in order to generate effective recommendations.

4.1 Environment description

The universe of discourse considered in our system is based on pair-wise relationships between two types of entities u and t , which we call "user" and "item", or "learner" and "course", respectively. We envision a world with:

- A set of learners $U = \{u_1, u_2, \dots, u_N\}$; - A set of courses $C = \{c_1, c_2, \dots, c_M\}$.
- Each item is described by a set of descriptors $D(t) = \{d_1, d_2, \dots, d_n\}$ such that $|D(t)| \geq 1$. A taxonomic descriptor d is an ordered sequence of topics p denoted by $d = \{p_0, p_1, \dots, p_q\}$ where $d \subseteq D(c)$, $c \subseteq C$. The topics within a descriptor are sequenced so that the former topics are super topics of the latter topics, when the super topic covers the general term of the domain and sub-topic covers a more specific term.
- $r_{u,c}$ The evaluation of course c made by learner u . All evaluations made by the learner u form a vector r_u , that represents his profile. The evaluation matrix is R .
- $z_{u,k}$ The probability that learner u belongs to cluster k ; $Z = (z_{u,k})$ is the probability matrix $U \times K$, where U, K are number of learners and clusters, respectively.
- $c_{k,t}$ The average of evaluations made by members of cluster k to item t , and $C = (c_{k,t})$ is the matrix of centroids $K \times T$, where T is the number of items.

4.2 The Fuzzy-based Collaborative Filtering algorithm

As mentioned above, this part contains our novel CF algorithm description. From the literature survey on the CF algorithms, we have the main steps of our algorithm:

- First, the automatic construction of groups in the system from the evaluation matrix using the Non-Negative Matrix Factorization (NNMF) method. The reason behind this choice and use of this method, is the reduction of the scalability problem that occurs when adding a new user or a new item
- In addition, the resulting probability matrix can be used to process data to solve large-scale problems of CF more efficiently.
- Then, for the neighborhood selection, we propose to consider just the K -nearest neighbors belonging to the C -nearest clusters following the principle of [18], [8] but using the fuzzy extension of the algorithm.
- The prediction of learner's preferences.

4.2.1 Users clustering algorithm: Modified FCM to NNMF (MFCMtoNNMF).

In this step we will factorize the evaluation matrix R into two matrices Z and C . where Z is the probability matrix and C is the matrix of cluster centers.

We will use a modified version of *FCM* into *NNMF* following the same principle of WU and LI [11], with adding the non-negativity constraint on the elements of the matrix C . Since C is the matrix of cluster centers where each element is the evaluations' average made by members of a cluster c to a given course c , so its components must be ≥ 0 . The problem with new constrained to be solved is

$$H(Z, C) = \frac{1}{2} \sum_{(u,c) \in P} (r_{u,c} - \frac{1}{\sum_{k=1}^K e^{z_{u,k}}} \sum_{k=1}^K e^{z_{u,k}} c_{k,c})^2 + \lambda_c \|c_c\|_2^2 + \lambda_z \|z_u\|_2^2 \quad (1)$$

St. $Z_1 = 1$; $Z \geq 0$; $C \geq 0$.

To resolve this problem, we have used the ACLS algorithm (Alternating Constrained Least Squares Algorithm) proposed in [19]. And to initialize the ACLS algorithm, we proposed a modified version of the *random Acot initialization* method cited in [19] by initializing each row of the matrix C by averaging p random rows of the evaluation matrix R . we called this method *random Rows initiatization* method.

4.2.2 Neighbors Pre-selection and Selection

An important step in the CF algorithm is the search for neighbors of the current learner. Traditional methods generally need to search the entire database, which definitely suffer from the scalability problem. We proposed an adjusted version of the fuzzy neighborhood algorithm following the same principle as in [8], [18] as follows:

- Calculate similarity between the active learner and all clusters to select the Fuzzy C -Nearest Prototypes (*FCNP*) [20]. We have considered only the *FCNP* because it's uninteresting to assign the learner to dissimilar clusters.
- Calculate similarity between the active learner and members of the *FCNP* to select the Fuzzy K -Nearest Neighbors (*FKNN*) [20] using the learner membership degrees to clusters in order to minimize the calculations.

We proposed to use the difference between membership degrees to the same cluster as a similarity measure between the active learner and members of *FCNP*. Where, the similarity between two learners increases when the difference between their degrees of belonging tends to 0.

4.2.3 The CF-Based prediction of the learner preferences

Similar to the idea presented in [21], we propose a framework that can effectively improve the performance, by combining linearly the prediction results of user based and item based algorithms, respectively as a CF Based prediction.

$$CFB_pred(ua, c) = \delta \widehat{u}_{ua,c} + (1 - \delta) \widehat{i}_{ua,c} \quad (2)$$

Where $\widehat{u}_{ua,c}$ and $\widehat{i}_{ua,c}$ are user-based and item-based predictions.

After the application of CF-based prediction methods, values in the cells of learner-course matrix are recalculated and updated. So, the sparseness of the matrix is therefore reduced. However, there may still be some empty cells due to the inadequate number of nearest neighbors for that learner. For this reason, it is necessary to use content information to make prediction for each learner-course pair. Then, merging the two predictions types to make full evaluation matrix.

4.3 Content-Based Filtering

To predict missing values based on content, we must have a set of features to describe the items' content in order to correlate similar items. In our system, items are courses and features are topics (information used to describe the courses' content).

We propose to calculate the occurrence frequency of each topic in all evaluated items by the active learner ua . Then, we will give a score to each topic to promote courses according to the topics' appearance and evaluations made by learner ua to each course. The reason is that two topics p_1 and p_2 belong to two courses c_1 and c_2 , respectively, can have the same occurrence frequency in the set of items evaluated by ua , but the course c_1 had a better evaluation against the course c_2 . Hence, the learner's preferences should promote $p_1 \in c_1$ over $p_2 \in c_2$ through their scores in the preferences' vector. So, the score assigned to the topic p_n in the preferences' vector of the learner ua is computed as follows

$$score(p_n, \vec{V}_{ua}) = \frac{\sum_{p_n \in c(ua)} (rating(ua, c) \cdot Occur(p_n))}{|c(ua)|} \quad (3)$$

Such that $|c(ua)|$ is the number of items rated by ua . $rating(ua, c)$ is the evaluation made by ua to the course c containing topic p_n , and $Occur(p_n)$ represents the occurrence frequency of the topic p_n in the set of items evaluated by ua .

After have given a score for each topic, we calculate the similarity between the test course and the set of courses assessed by the active learner to select the T -nearest courses to the test course. We propose to use the cosine similarity measure to calculate the similarity between two course vectors.

4.3.1 Content-Based prediction

Finally, we make the content-based prediction of the missing values. The rating prediction for an unseen course is formulated as follows

$$B_pred(ua, c) = \frac{\sum_{m \in TNI(c)} rating(ua, m) \cdot sim(c, m)}{\sum_{m \in TNI(c)} sim(c, m)} \quad (4)$$

Where $rating(ua, m)$ represents the evaluation made by the learner ua to course $m \in TNI(c)$ and $sim(c, m)$ is the similarity calculated in the previous section.

This type of prediction use topics to predict missing ratings. So, it needs predictive features to achieve a good prediction which limit the effectiveness use of this prediction lonely. To address limitations of the CF-based and Content-based predictions, we are conducted toward hybridization between them.

4.4 Hybrid prediction

In this section, we will present our hybrid prediction scheme that combines between the CF-based prediction and the Content-based prediction in order to obtain a full user-item matrix. Our proposed hybrid prediction scheme is defined as follows

$$Final_pred = \alpha \times CFB_pred(ua, c) + (1 - \alpha) \times CB_pred(ua, c) \quad (5)$$

Where α is used to control the weight between the two predictions.

4.5 The Top-K Recommendation

After applying the hybrid algorithm cited above, we obtain predictions of the unviewed items by the active learner. Then, we apply the procedure for generating the recommendation. The first step is to calculate the scores of items based on clusters' preferences and the learner preferences' prediction, to select the Top-N items in each group. Then, generating a list of Top-K items selected from the Top-N items.

The preferred items (courses) will be determined by the number of nearest learners who evaluated the course (popularity) and their mean explicit evaluations by:

$$C_pref(uc, c) = \beta * moy(uc, c) + (1 - \beta) * pop(uc, c) \quad (6)$$

This formula is based only on the explicit evaluations. To apply this formula in the TEL field, we introduce the importance of knowledge proposed in [14]. So the average will be replaced by an evaluation estimation e_i of a course taking into consideration the importance of knowledge of learners who evaluated the course c ;

$$e(uc, c) = \frac{1}{\sum_{u=1}^{FKNN} \bar{s}_u} \sum_{u=1}^{FKNN} \bar{s}_u r_{u,c} \quad (7)$$

$$\text{Such as } \bar{s}_u \text{ is calculated as } \bar{s}_u = \frac{1}{t} \sum_{t=1}^T S_{u,t} \quad (8)$$

Where $c_{u,t}$ is the score obtained by the learner u in the test t . $r_{u,c}$ is the explicit evaluation of the learner u the course c . Thus, the C_pref formula becomes as follows;

$$C_pref(uc, c) = \beta * e(uc, c) + (1 - \beta) * pop(uc, c) \quad (8)$$

Then, a score (rank) is assigned to each item (course) in order to ranging items according to the cluster preferences and the predicted learner preferences. As

$$Rank_{u,c} = \alpha C_pref(uc, c) + (1 - \alpha) \hat{r}(u, c) \quad (9)$$

The list of recommendations to be generated in the cluster uc is chosen by selecting the $TOP-N$ items with the highest scores and the $TOP-K$ items will be set as follow

$$K = z_{u,uc} * N \quad (10)$$

Where N is the number of items selected from cluster uc and $z_{u,uc}$ is the membership degree of the learner u to cluster uc . The final recommendation is formally represented as

$$\sum_{uc \in C-FNP(u)} TOP - K(uc, c) \quad (11)$$

5 Application: Experiment and Results

5.1 Moodle Dataset

Moodle¹ is a free source e-learning software platform. Due to the lack of no data sets have been made publicly available for formal and non-formal learning, we used a database very known in RSs, BX-Book-Rating² and we consider that each book is a learning resource or a course. We restricted our validation to a subset of this base by selecting just 21 learners, 20 courses and we have added information about the knowledge level of the learner, which are his test scores. And we integrated it with our technique in the Moodle platform. As showed in Fig.1.

¹ www.moodle.org

² www.informatik.uni-freiburg.de/~cziegler/BX/

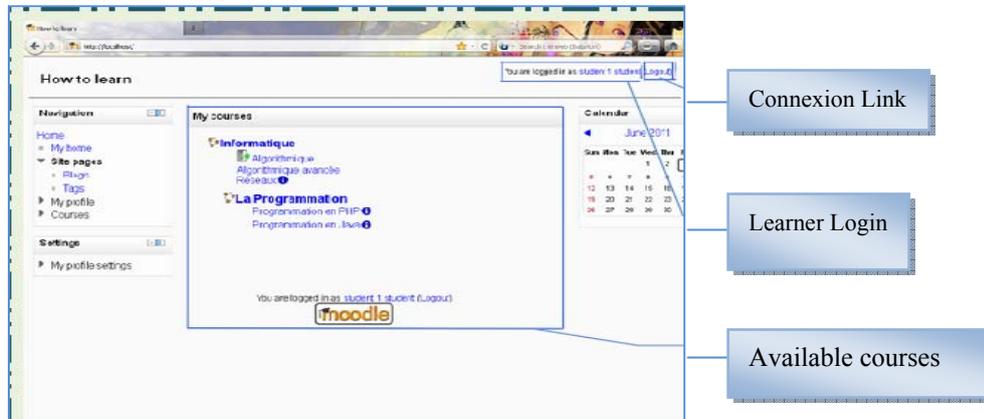


Fig. 1. Moodle Platform with our dataset

5.2 Mean Absolute Error (MAE)

We choose the Mean Absolute Error (MAE) as the evaluation metric to calculate the performance of our CF scheme.

$$MAE = \frac{\sum_{u,t} |x_{u,t} - \hat{x}_{u,t}|}{|N|} \quad (13)$$

N is the number of test evaluations. More MAE is lower, the performance is better.

As we are in the TEL field, we will apply the novel MAE metric proposed by [14], and adapted to the e-learning domain, in order to take in consideration the knowledge importance of the learner (his different test scores). The novel metric is as follow

$$MAE = \frac{1}{|k|} \sum_{k=1}^{|k|} r_{u,i} - [(1 - \alpha)]e_i + \alpha r_{u,i}; \quad 0 < \alpha < 1 \quad (14)$$

5.3 F1 metric

To evaluate the performance of *Top-K* recommendation, we used the F_1 metric,

$$F_1 = \frac{2PR}{P+R} \quad (15)$$

Where P and R are the precision and recall respectively. They are calculated as

$$P = \frac{N_t}{N} \quad , \quad R = \frac{N_t}{N_p} \quad (16)$$

N : The total number of items; N_t : Number of relevant items found and N_p : Total number of relevant items

5.4 Performance Evaluation of the CF technique

As the data sample on which we applied our algorithm is smaller the used by [14], therefore we cannot compare them. Such as [14] used four clusters of variant size between 15 and 90, we used only three groups with size between 5 and 10. We evaluate the performance using the novel MAE metric adapted to the e-learning field. Results are showed in Fig.2.

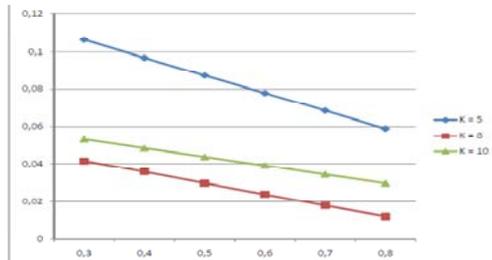


Fig. 2. MAE performance, less value means better performance

From Fig.2. We observe that the MAE has an inverse relationship with the clusters size K , and the different values of α (0.3+0.8), ie. Most K and α are large; most the value of MAE is small. We can notice that the new MAE in almost all cases is smaller than usual MAE, which is due to the subtraction of both products of the values on the y-axis and we know that the levels of RS accuracy are better when the new metric is applied, this is due to the favorable weighting of the users knowledge.

5.5 Performance Evaluation of the Top-K Recommendation

The figure below shows the evolution of the F1 metric with number of recommended courses. We observe from Fig.3 that the F1 metric increase until 15 courses evaluated. The F1 values are varied depending on the number of relevant items. It can be seen also that the recommendation performance of the system is good.

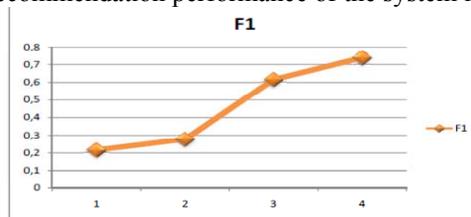


Fig. 3. F1 metric

6 Conclusion and future works

Recommender Systems are widely used recently in Technology Enhanced Learning which creates the need to adapt these systems to e-learning. For this and we proposed, in this paper, a novel approach which uses an adapted RS to TEL field. Especially when recommending learning objects that belong to different contexts. Experimental results show that the proposed approach can improve the recommendation accuracy. In the future work, we will elaborate this technique to generate multi-context recommendations taking in the account the temporal dynamics effect.

References

1. Burke R., “ Hybrid recommender systems: Survey and experiments “, User Modeling and User Adapted Interaction, vol. 12, n° 4, 2002, p. 331-370.

2. Li-Tung Weng, Yue Xu Yuefeng Li, R. Nayak. "Exploiting Item Taxonomy for Solving Cold-Start Problem in Recommendation Making". Tools with Artificial Intelligence, ICTAI'08, 20th IEEE International Conference, Volume: 2, Pages 113-120.
3. S. Rendle, C. Freudenthaler, L. Schmidt-Thieme, "Factorizing personalized Markov chains for next-basket recommendation", in Pro. of the 19th IC on WWW, ACM, 811-820, 2010.
4. Y. Koren, R. Bell, C. Volinsky, "Matrix Factorization Techniques for Recommender Systems", IEEE Computer Society Press 42 (8) (2009), 30-37, ISSN 0018-9162.
5. N. Manouselis, H. Drachsler, R. Vuorikari, H. Hummel, R. Koper, "Recommender Systems in Technology Enhanced Learning", in the 1st RSs Handbook, Springer-Berlin, 2010.
6. Drachsler, H., & Manouselis, N. "How Recommender Systems in Technology-Enhanced Learning depend on Context", the 1st workshop on Context-aware RSs for Learning, 2009.
7. Belkin N.J., Croft W.B., "Information filtering and information retrieval: two sides of the same coin? ", Communication of the ACM, vol. 35, n 12, Pages 29-38, Dec 1992.
8. Chen G., Wang F., Zhang C., "Collaborative Filtering Using Orthogonal Nonnegative Matrix Tri-factorization ", J. of Information Processing and Management, Volume 45, 2009.
9. Maatallah M. and Seridi H., "A Fuzzy Hybrid Recommender System", the 1st International Conference on Machine and Web Intelligence, ICMWI'10, Oct 2010.
10. J C. Bezdek., "Fuzzy mathematics in pattern classification" ph_D. dissertation. Cornell Univ.. Ithaca, NY, 1973.
11. Wu J. and Li T., "A modified fuzzy C-means algorithm for collaborative filtering" Proc. of the 2nd KDD Workshop on Large-Scale RSs and the NPC, ACM New York, 2008.
12. H. Drachsler, H. G. K. Hummel, R. Koper, "Identifying the Goal, User model and Conditions of Recommender Systems for Formal and Informal Learning", Journal of DI 10.
13. E.Garcia, C.Romero, S.Ventura, C.D.Castro, "An architecture for making recommendations to courseware authors using association rule mining and collaborative filtering", User Modeling and User-Adapted Interaction, Vol 19, issue (1-2), 2009.
14. J. Bobadilla, F. Serradilla, A. Hernando, "Collaborative filtering adapted to recommender systems of e-learning, Knowledge-Based Systems", Journal of KBS, vol 22, issu 4, 2009.
15. N. Soonthornphisaj, E. Rojsattarat, S. Yim-ngam, "Smart E-Learning Using Recommender System", in: International Conference on Intelligent Computing, 518-523, 2006.
16. L. Ge, W. Kong, J. Luo, "Courseware Recommendation in E-Learning System", in: International Conference on Web-based Learning (ICWL'06), 10-24, 2006.
17. Nguyen Thai-Nghe, Lucas Drumond, Artus Krohn-Grimberghe, Lars Schmidt-ThiemeThay, "Recommender System for Predicting Student Performance ", 1st Workshop on RSs for TEL, Procedia Computer Science, Elsevier,2010.
18. G.-R. Xue, C. Lin, Q. Yang, W. Xi, H.-J. Zeng, Y. Yu, and Z. Chen. "Scalable collaborative filtering using cluster-based smoothing". In Proc. of SIGIR, 2005.
19. Amy Langville, Carl Meyer, Russell Albright, James Cox and David Duling, "Algorithms, Initializations, and Convergence for the Nonnegative Matrix Factorization", KDD 2006.
20. James M.Keller, Michael R.Gray, James A.Givens, JR, "A Fuzzy Nearest Neighbor Algorithm", IEEE Transaction on Systems,Man and Cybernetics,VOL.SMC-15, 1985.
21. J. Wang, A. P. de Vries, and M. J. T. Reinders. "Unifying user-based and item-based collaborative filtering approaches by similarity fusion". In Proc. of SIGIR, 2006.

Meta-Learning for Escherichia Coli Bacteria Patterns Classification

Hafida Bouziane, Belhadri Messabih, and Abdallah Chouarfia

MB University, BP 1505 El M'Naouer
3100 Oran Algeria
e-mail: (h_bouziane,messabih,chouarfia)@univ-usto.dz

Abstract: In machine learning area, there has been a great interest during the past decade to the theory of combining machine learning algorithms. The approaches proposed and implemented become increasingly interesting at the moment when many challenging real-world problems remain difficult to solve, especially those characterized by imbalanced data. Learning with imbalanced datasets is problematic, since the uneven distribution of data influences the behavior of the majority of machine learning algorithms, which often lead to poor performance. It is within this type of data that our study is placed. In this paper, we investigate a meta-learning approach for classifying proteins into their various cellular locations based on their amino acid sequences. A meta-learner system based on k-Nearest Neighbors (k-NN) algorithm as base-classifier, since it has shown good performance in this context as individual classifier and DECORATE as meta-classifier using cross-validation tests for classifying Escherichia Coli bacteria proteins from the amino acid sequence information is evaluated. The paper reports also a comparison against a Decision Tree induction as base-classifier. The experimental results show that the k-NN-based meta-learning model is more efficient than the Decision Tree-based model and the individual k-NN classifier.

Keywords: Classification, Meta-Learning, Imbalanced Data, Subcellular Localization, E.coli.

1. Introduction

Most of the current research projects in bioinformatics deal with structural and functional aspects of genes and proteins. High-throughput genome sequencing techniques have led to an explosion of newly generated protein sequences. Nowadays, the function of a huge number among them is still not known. This challenge provides strong motivation for developing computational methods that can infer the protein's function from the amino acid sequence information. Thus, many automated methods have been developed for predicting protein structural and molecular properties such as domains, active sites, secondary structure, interactions, and localization from only the amino acid sequence information. One helpful step for understanding and therefore, elucidating the biochemical and cellular function of proteins is to identify their subcellular distributions within the cell. Most

of the existing predictors for protein localization sites are used with the assumption that each protein in the cell has one, and only one, subcellular location. In each cell compartment, specific proteins ensure specific roles that describe their cellular function which is critical to a cell's survival. This fact means that the knowledge of the compartment or site in which a protein resides allows to infer its function. So far, many methods and systems have been developed to predict protein subcellular locations and one of the most thoroughly studied single cell organism is *Escherichia coli* (*E.coli*) bacteria.

The first approach for predicting the localization sites of proteins from their amino acid sequences was a rule based expert system PSORT developed by Nakai and Kanehisa [1,2], then the use of a probabilistic model by Horton and Nakai [3], which could learn its parameters from a set of training data, improved significantly the prediction accuracy. It achieved an accuracy of 81% on *E.coli* dataset. Later, the use of standard classification algorithms achieved higher prediction accuracy. Among these algorithms, k-Nearest Neighbors (k-NN), binary Decision Tree and Naïve Bayesian classifier. The best accuracy has been achieved by k-NN classifier, that the classification of the *E.coli* proteins into 8 classes achieved an accuracy of 86% by cross-validation tests [4], The accuracy has been improved significantly compared to that obtained before. Since these works, many systems that support automated prediction of subcellular localization using variety of machine learning techniques have been proposed. With recent progress in this domain, various features of a protein are considered, like composition of amino acids [5], pseudo amino acids [6], and dipeptide and physico-chemical properties [7,8]. The performance of existing methods varies and different prediction accuracies are claimed. Most of them achieve high accuracy for the most populated locations, but are generally less accurate on the locations containing fewer specific proteins. Recently, there has been a great interest to the theory of combining classifiers to improve performance [9]. Several approaches known as ensembles of classifiers (committee approaches) have been proposed and investigated through a variety of artificial and real-world datasets. The main idea behind is that often the ensemble achieves higher performance than each of its individual classifier component. One can distinguish two groups of methods: methods that combine several heterogeneous learning algorithms as base-level classifiers over the same feature set [10], such as stacking, grading and voting, and methods which construct ensembles (homogeneous classifiers) generated by applying a single learning algorithm as base-classifier by sub-sampling the training sets, creating artificial data to construct several learning sets from the original feature set, such as boosting [11], bagging [12] and Random Forests [13]. In protein localization sites prediction problem, data distribution is often imbalanced. For the best of our knowledge, there are two major approaches that try to solve the class imbalance problems: the one which use resampling

methods and the one that modify the existing learning algorithms. Resampling strategy balances the classes by adding artificial data for improving the minority class prediction of some classifiers. Here, we focus on the resampling methods, since they are simplest methods to increase the size of the minority class. This article investigates the effectiveness of the meta-learning approach DECORATE [14] to create a meta-level dataset trained using a simple k-NN algorithm as base-classifier in classifying proteins in their subcellular locations in E.coli benchmark dataset using cross-validation and compares the results by using Decision Tree induction as base-classifier.

The rest of the paper is organized as follows. Section 2, presents the materials and the methodology adopted and presents a brief description of E.coli benchmark dataset as well as the evaluation measures used for performance evaluation. Then, section 3 summarizes and discusses the results obtained by the experiments, it also presents a comparison of Decision Tree induction against the k-NN algorithm as base-classifiers to the meta-classifier DECORATE. Finally, section 4 concludes this study.

2. Material and Methods

2.1 E.coli Dataset

The prokaryotic gram-negative bacterium Escherichia Coli is an important component of the biosphere, it colonises the lower gut of animals and humans. The Escherichia Coli benchmark dataset has been submitted to the UCI¹ Machine Learning Data Repository [15]. It is well described in [1,2,3]. The dataset patterns are characterized by attributes calculated from the amino acid sequences. Protein patterns in the E.coli dataset are classified to eight classes, it is a drastically imbalanced dataset of 336 patterns. One can find classes with more than 130 patterns and other ones with only 2 or 5 patterns. Each pattern with eight attributes (7 predictive and 1 name corresponding to the accession number for the SWISSPROT² database), where the predictive attributes correspond to the following features : (1) mcg: McGeoch's method for signal sequence recognition [16], the signal sequence is estimated by calculating discriminate score using length of N-terminal positively-charged region (H-region); (2) gvh: Von Heijne's method [17,18] for signal sequence recognition., the score estimating the cleavage signal is evaluated using weight-matrix and the cleavage sites consensus patterns to detect signal-anchor sequences; (3) lip: Von Heijne's Signal Peptidase II consensus sequence score; (4) chg: binary attribute indicating presence of charge on N-terminus of predicted lipoproteins; (5) aac: score of discriminate

¹ Web site: <http://archive.ics.uci.edu/ml>

² Web site: <http://www.uniprot.org/>

analysis of the amino acid content of outer membrane and periplasmic proteins; (6) alm1: score of the ALOM membrane spanning region prediction program, it determines whether a segment is transmembrane or peripheral; (7) alm2: score of ALOM program after excluding putative cleavable signal regions from the sequence.

Protein patterns in this dataset are organized as follows: 143 patterns of cytoplasm (cp), 77 of inner membrane without signal sequence (im), 52 of periplasm (pp), 35 of inner membrane without uncleavable signal sequence (imU), 20 of outer membrane without lipoprotein (omL), 5 of outer membrane with lipoprotein (omL), 2 of inner membrane without lipoprotein (imL) and 2 patterns of inner membrane with cleavage signal sequence (imS). The class distribution is extremely imbalanced, especially for imL and imS proteins.

2.2 Base-Classifiers

The problem considered here is multi-class, let us denote by Q the number of categories or classes, $Q \geq 3$. Each object is represented by its description $x \in X$, where X represents the feature set and its category $y \in Y$, where Y denotes a set of the Q categories and can be identified with the set of indices of the categories: $Y = \{1, \dots, Q\}$. The assignment of the descriptions to the categories is performed by means of a classifier, The chosen classifiers are then described in the following subsections.

2.2.1 k-Nearest Neighbors Classifier

The k-nearest neighbors (k-NN) rule [19] is considered as a lazy approach. It is one of the oldest and simplest supervised learning algorithm. Objects are assigned to the class having the majority of the k Nearest Neighbors in the training set. Usually, Euclidean distance is used as the distance metric. Given a test example x with unknown class, the algorithm assigns to the example x the class which is most frequent among the k training examples nearest to that query example, according to the distance metric. The classification accuracy of k-NN algorithm can be improved significantly if the distance metric is learned with specialized algorithms, many studies try to find the best way to improve the k-NN performance taking into account this factor. In practice, k is usually chosen to be odd. The best choice of this parameter depends on the data concerned with the problem at hand. This algorithm has shown good performance in biological and medical data classification problems.

2.2.2 Decision Tree Induction

A Decision Tree [20] is a powerful way of knowledge representation. The model produced by a decision tree classifier is represented in the form of tree structure. The principle, consists in building decision trees by recursively selecting attributes on which to split. The criterion used for selecting an attribute is information gain. A leaf node indicates the class of the examples.

The instances are classified by sorting them down the tree from the root node to some leaf nodes. Posterior probabilities are estimated by the class frequencies of the training set in each end node. In this study, we used a decision tree built by C4.5 [21].

2.3 Meta-Classifier

Meta-learners such as Boosting, Bagging and Random Forests provide diversity by sub-sampling or re-weighting the existing training examples [14]. Decorate (Diverse Ensemble Creation by Oppositional Relabeling of Artificial Training Examples) performs by adding randomly constructed examples to the training set when building new ensemble members (committee). It has been conceived basing on a diversity measure introduced by the authors. The measure defined expresses the ensemble member disagreement with the ensemble's prediction. If C_j is an ensemble member classifier, $C_j(x)$ the class label predicted by the classifier C_j for the example x and $C^*(x)$ the prediction of the ensemble, the diversity d_j of C_j on the example x is defined as follows :

$$d_j(x) = \begin{cases} 0 & \text{if } C_j(x) = C^*(x) \\ 1 & \text{otherwise} \end{cases} \quad (1)$$

The diversity of an ensemble of M members, on a training set of N examples is computed as follows :

$$D = \frac{1}{NM} \sum_{j=1}^M \sum_{i=1}^N d_j(x_i) \quad (2)$$

The approach consists in constructing an ensemble of classifiers which maximize the diversity measure D . Three parameters are needed: the artificial size which is a fraction of the original training set, the desired number of member classifiers and maximum number of iterations to perform. Initially, the ensemble contains the classifier (base-classifier) trained on the original data. The members added to the ensemble in the successive iteration are trained on the original training data combined with some artificial data. To generate the artificial training examples named as diversity data, the algorithm takes in account the specified fraction of the training set size. The class labels assigned to the diversity data differ maximally from the current predictions of the committee (completely opposite labels). The current classifier is added to the committee if it increases the ensemble diversity, otherwise it is rejected. The process is repeated until the desired committee size is reached or the number of iterations is equal to the maximum fixed. Each classifier C_j of the committee C^* provides probabilities for the class membership of each example to classify. If $P_{C_j,k}(x)$ represents the estimated probability of x to belong to the class labeled k according to the classifier C_j , to classify an example x , the algorithm considers the most probable class as the label for x as follows :

$$C^*(x) = \operatorname{argmax}_{k \in \{1, \dots, Q\}} P_k(x) \quad (3)$$

Where $P_k(x)$ represents the probability that x belongs to the class labeled k computed for the entire ensemble, it is expressed as :

$$P_k(x) = \frac{\sum_{C_j \in C^*} P_{C_j, k}(x)}{|C^*|} \quad (4)$$

In this paper, we performed two sets of experiments. In the first one, we used the k-NN classifier as base-classifier. In the second one, we used Decision Tree as base-classifier, which is used in the original DECORATE conception. Our goal was to empirically evaluate the two models on the E.coli dataset. For this purpose, we proceed for the two sets of experiments in two steps. In the first step, we evaluated both the two individual classifiers on Ecoli dataset applying cross-validation and in the second step we used the meta-learning system applying also cross-validation to prediction performance assessment. For all experiments, we made preliminary trials to select the appropriate parameters (model selection).

2.4 Evaluation Measures

Any results obtained by machine learning algorithms must be evaluated before one can have any confidence in their classifications, this aspect of machine learning theory is not only useful but fundamental. There are several standard methods for evaluation. In what follows, we present only the measures used in this study.

2.4.1 Cross Validation

In this study, we used Cross Validation tests to evaluate the classifier robustness, this methodology is most suitable to avoid biased results. Thus, the whole training set was divided into five mutually exclusive and approximately equal-sized subsets and for each subset used in test, the classifier was trained on the fusion of all the other subsets. So, cross validation was run five times for each classifier and the average value of the five-cross validations was calculated to estimate the overall classification accuracy.

2.4.2 Classification Accuracy Measurements

Some of the most relevant evaluation measures are precision, recall and F-measure. In this study, we adopted the three measures, for evaluating the effectiveness of the classification for each class and the classification accuracy for all the classes as performance measures. A confusion matrix

(contingency table of size $Q \times Q$ has been used, $M = (m_{kl})_{1 \leq k, l \leq Q}$, where m_{kl} denotes the number of examples observed in class k and classified in class l . The rows indicate different classes observed and the columns show the result of the classification method for each class. The number of correctly classified examples is the sum of diagonal elements in the matrix, all others are incorrectly classified. The F-measure has two components, which are: the Recall and the Precision. The Recall is the ratio of the number of positive examples (correctly classified) of class k and the number of all positive (observed) examples in class k . We can express this ratio using confusion matrix elements as follows:

$$Recall = 100 \times \frac{m_{kk}}{\sum_{l=1}^Q m_{kl}}, k \in \{1, \dots, Q\} \quad (5)$$

The Precision is the ratio of number of correctly classified examples of class k and the number of examples assigned to class k , it can be formulated as follows:

$$Precision = 100 \times \frac{m_{kk}}{\sum_{i=1}^Q m_{ik}}, k \in \{1, \dots, Q\} \quad (6)$$

The F-measure is then defined as :

$$F - measure = \frac{2 \times Recall \times Precision}{Recall + Precision} \quad (7)$$

The classification accuracy is the ratio of number of all correctly classified examples and the total number of examples (both positive and negative), it is given by :

$$Accuracy = 100 \times \frac{1}{N} \sum_{k=1}^Q m_{kk} \quad (8)$$

3. Experimental Results

In this section we report the results for each experiment by highlighting for each step the evaluation measure values. The most important evaluation values are shown with bold typeface. It is important to note, that adding training instance which is common characteristic of DECORATE implies increasing training time. This is visible when performing with a great number of needed classifiers for the ensemble and the desired number of artificial data to create for learning the meta-classifier.

The tables given above report the results of ensembles versus individual classifiers. In this experiment, we applied 5-fold cross-validations. The E.coli dataset is randomly partitioned approximately equally sized subsets. Table 1 and Table 3 summarize the performance in test of each individual classifier for each class. Table 2 and Table 4 give the number of patterns obtained for each class using DECORATE-based k-NN (Dk-NN) and DECORATE-based C4.5 (DC4.5). The best results for k-NN were obtained when setting k=9.

Table 1. Confusion matrix of k-NN as individual classifier on E.coli dataset using 5-CV

Observed	Predicted							
	cp	im	pp	imU	om	omL	imL	imS
cp (143)	141	0	2	0	0	0	0	0
im (77)	3	63	1	9	0	1	0	0
pp (52)	3	1	47	0	1	0	0	0
imU (35)	1	10	0	23	0	1	0	0
om (20)	0	0	4	0	15	1	0	0
omL (5)	0	0	0	0	0	5	0	0
imL (2)	0	1	0	0	0	1	0	0
imS (2)	0	1	1	0	0	0	0	0

Table 2. Confusion matrix of k-NN based-DECORATE (Dk-NN) on E.coli dataset using 5-CV

Observed	Predicted							
	cp	im	pp	imU	om	omL	imL	imS
cp (143)	141	0	2	0	0	0	0	0
im (77)	3	63	1	9	0	0	1	0
pp (52)	3	1	47	0	1	0	0	0
imU (35)	1	7	0	26	0	1	0	0
om (20)	0	0	2	0	17	1	0	0
omL (5)	0	0	0	0	0	5	0	0
imL (2)	0	1	0	0	0	1	0	0
imS (2)	0	1	1	0	0	0	0	0

The confusion matrix of Dk-NN in Table 2 shows a gain in classifying om and imU proteins. Whereas, no improvement has been observed for the two minority class proteins namely imL and imS, which are the most difficult to classify.

Table 3. Confusion matrix of C4.5 as individual classifier on E.coli dataset using 5-CV

Observed	Predicted							
	cp	im	pp	imU	om	omL	imL	imS
cp (143)	137	2	2	0	2	0	0	0
im (77)	2	59	1	13	2	0	0	0
pp (52)	4	2	45	0	1	0	0	0
imU (35)	1	12	1	20	1	0	0	0
om (20)	1	1	4	0	14	0	0	0
omL (5)	0	0	2	0	1	2	0	0
imL (2)	0	1	0	0	0	1	0	0
imS (2)	0	1	1	0	0	0	0	0

Table 4. Confusion matrix of C4.5 based-DECORATE (DC4.5) on E.coli dataset using 5-CV

Observed	Predicted							
	cp	im	pp	imU	om	omL	imL	imS
cp (143)	142	0	1	0	0	0	0	0
im (77)	2	66	0	8	0	0	0	1
pp (52)	4	2	46	0	0	0	0	0
imU (35)	1	13	0	21	0	0	0	0
om (20)	0	0	2	0	18	0	0	0
omL (5)	0	0	1	0	0	5	0	0
imL (2)	0	1	0	0	0	1	0	0
imS (2)	0	1	1	0	0	0	0	0

Table 3 and Table 5 show that Decision Tree used as individual classifier performs poorly than the individual k-NN. However, in Table 4 the improvement is well observed in both cp, im and om proteins. Not surprisingly, Dk-NN gives better results than DC4.5, which confirms once again its power in this context. What is important to notify is that even the ensembles Dk-NN and DC4.5 fail in classifying pp and imU with high confidence and fail completely for umL and imS. The influence of the number of ensembles (size) needed for the meta-classifier on the performance of the two ensembles Dk-NN and DC4.5 is shown in Fig.1.

Table 5. Test performance using 5- CV on E.coli dataset

Classifiers	Measures	Classes								Correctly classified	Accuracy
		cp	im	pp	imU	om	omL	imL	imS		
k-NN	Precision	95.3	82.9	85.5	71.9	93.8	55.6	0	0	294	87.5
	Recall	98.6	81.8	90.4	65.7	75.0	100	0	0		
	F-	96.9	82.4	87.9	68.7	83.3	71.4	0	0		
C4.5	Precision	94.5	75.6	80.4	60.6	66.7	66.7	0	0	277	82.4
	Recall	95.8	76.6	86.5	57.1	70.0	40.0	0	0		
	F-	95.1	76.1	83.3	58.8	68.3	50.0	0	0		
Dk-NN	Precision	95.3	86.3	88.7	74.3	94.4	55.6	0	0	299	88.9
	Recall	98.6	81.8	90.4	74.3	85.0	100	0	0		
	F-	96.9	84.0	89.5	74.3	89.5	71.4	0	0		
DC4.5	Precision	95.3	79.5	92.0	72.4	100	83.3	0	0	298	88.6
	Recall	99.3	85.7	88.5	60.0	90.0	100.0	0	0		
	F-	97.3	82.5	90.2	65.6	94.7	90.9	0	0		

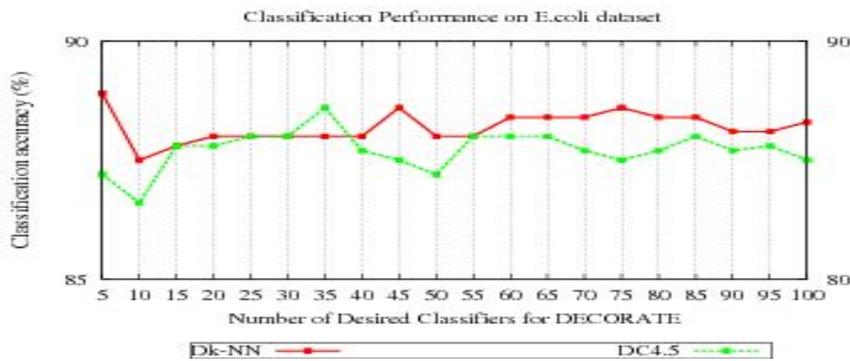


Fig. 1. Comparison of the classification performance (y axis), according to the desired size of classifiers (x axis) between the two ensembles Dk-NN and DC4.5 on E.coli dataset the individual classifiers (x axis)

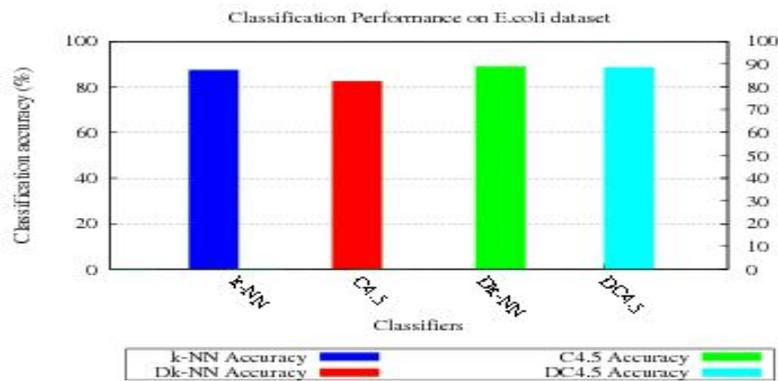


Fig. 2. Accuracy comparison between the four classifiers on E.coli dataset

The results reported for this study show that the classification attempts of inner membrane with lipoprotein (imL) and inner membrane with cleavable signal sequence (imS) proteins failed for each classifier and consequently also for Dk-NN and DC4.5. This situation is caused by the extremely low number of examples in these classes (one example used for training and one example for testing). On the other hand, outer membrane with lipoprotein (omL) proteins were classified with 100% success rate by kNN classifier and both Dk-NN and DC4.5. The cytoplasm (cp) proteins were relatively well classified by almost all classifiers. Fig.2 highlights the performance in test of each classifier and shows well the superiority of the ensembles Dk-NN and DC4.5 in classifying E.coli patterns. Finally; it should be emphasis that this results are better than those obtained by combining heterogeneous classifiers by majority voting rule, since an average classification success of 88.3% was

achieved [22]. Nevertheless, all these results prove that combining classifiers is indeed a fruitful strategy.

4. Conclusion

More recently, several ensemble learning algorithms have emerged that have different strengths regardless the type of data involved for the problem in question. One is often confused to make an effective choice among them. Protein cellular localization sites prediction is one among the most challenging problems in modern computational biology. Various approaches have been proposed and applied to solve this problem but the extremely imbalanced distribution of proteins over the cellular locations make the prediction much more difficult. In this study, we applied DECORATE ensemble learning, investigating two standard machine learning approaches to improve the performance in classifying E.coli proteins to their cellular locations, based on their amino acid sequences. The experiments show that the k-NN-based meta-learning model outperforms the individual k-NN classifier and achieves better classification accuracy than the Decision Tree-based model. Further investigations will be carried out to provide a much more improved ensemble model.

5. References

- [1] Nakai, K., Kanehisa, M.: Expert system for predicting protein localization sites in gram-negative bacteria. *Proteins: Structure, Function, and Genetics*. 11,95-110 (1991).
- [2] Nakai, K., Kanehisa, M.: A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics*. 14, 897-911 (1992).
- [3] Horton, P., Nakai, K.: A probabilistic classification system for predicting the cellular localization sites of proteins. In *Proceedings of Intelligent Systems in Molecular Biology*, pp 109-115. St. Louis, USA (1996).
- [4] Horton, P. , Nakai, K.: Better prediction of protein cellular localization sites with the k Nearest Neighbors classifier, pp. 147-152. *AAAI Press*. Halkidiki, Greece (1997).
- [5] Nakashima, H., Nikishawa, K.: Discrimination of intracellular and extracellular proteins using amino acid composition and residue pair frequencies. *J. Mol. Biol.* 238, 54–61 (1994).
- [6] Park, K. J., Kanehisa, M.: Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. *Bioinformatics*. 19, 1656–1663 (2003).

- [7] Sarda, D., Chua, G.H., Li, K. B., Krishnan, A. :pSLIP: SVM based protein subcellular localization prediction using multiple physicochemical properties. *BMC Bioinformatics*. 6 ,152 (2005).
- [8] Rashid, M., Saha, S., Raghava, G. P. S.: Support Vector Machine-based method for predicting subcellular localization of mycobacterial proteins using evolutionary information and motifs. *BMC Bioinformatics*. 8, 337 (2007).
- [9] Dietterich, T. G.: Ensemble methods in machine learning. In: Kittler, J., Roli, F. (eds.), *First International Workshop on Multiple Classifier Systems*, LNCS, pp. 1-15.. Springer-Verlag (2000).
- [10] Wolpert, D. H.: Stacked generalization. *Neural Networks*. 5, 241-259 (1992).
- [11] Freund, Y. , Schapire, R. E.: Experiments with a new boosting algorithm. In: Saitta, L. (Ed.), *Proceedings of the Thirteenth International Conference on Machine Learning (ICML96)*. pp. 148-156 (1996).
- [12] Breiman, L.: Bagging predictors. *Machine Learning*. 24 (2), 123-140 (1996).
- [13] Rodriguez, J. J., Kuncheva, L. I. : Rotation forest: A new classifier ensemble method. *IEEE Transaction in Pattern Analysis*. 28(10), 1619-1630 (2006).
- [14] Melville, P. , Mooney, R.: Constructing diverse classifier ensembles using artificial training examples. *The Eighteenth International Joint Conference on Artificial Intelligence*, pp. 505-510. Acapulco, Mexico, 2003.
- [15] Blake, C.L., Merz, C.J.: *UCI repository of machine learning databases* (1998).
- [16] McGeoch, D. J., Dolan, A., Donald, S. Rixon, F.J.: Sequence determination and genetic content of the short unique region in the genome of herpes simplex virus type 1. *J Mol. Biol*. 181, 113 (1997).
- [17] Heijne, G. V.: A new method for predicting signal sequence cleavage sites. *Nucleic Acids Research*. 14, 4683-4690 (1986).
- [18] Heijne, G. V. :The structure of signal peptides from bacterial lipoproteins. *Protein Engineering*. 2,531-534 (1989).
- [19] Cover, T. M. Hart, T, P. E.: Nearest neighbor pattern classification. *IEEE Transactions on information Theory*. 13 (1), 21–27 (1967).
- [20] Breiman, L., Friedman, J.H., Olshen, R. A., Stone, C. J.: *Classification and regression trees*. Monterey, Chapman & Hall (1984).
- [21] Quinlan, J.R. :C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo,CA (1993).
- [22] Bouziane, H., Messabih, B., Chouarfia, A.: A Voting-Based Combination System for Protein Cellular Localization Sites Prediction. In *IEEE International Conference on Information and Computer Applications (ICICA)*, pp. 166-173, Dubai (2011).

Ontology-based gene set enrichment analysis using an efficient semantic similarity measure and functional clustering.

Sidahmed Benabderrahmane¹ and Mekami Hayet²

(1): King Abdul Aziz University,
Faculty of Computing and Information Technology.
Jeddah 21589, Saudi Arabia.

`sidahmed.benabderrahmane@gmail.com`,

(2): Djillali Liabes University,
Laboratory of communication and
multimedia architecture,
Bel Abbes 22000, Algeria.

Abstract. Gene set enrichment analysis allows to extract specific biological functions relative to a group of genes. To this aim, we propose here a novel approach for mining biological data, using the Gene Ontology (GO) as main source of genes annotation terms. Firstly, we will use our new semantic similarity measure (*IntelliGO*) in a clustering process, for grouping genes sharing similar biological functions described in GO. Secondly, the clustering results are evaluated using the F-score method and public genes reference sets. After that, an overlap analysis is presented as a method for exploiting the matching between clusters and reference sets. This method is then applied to a list of genes found dys-regulated in cancer samples. In this case, the reference sets are replaced by gene expression profiles. Consequently, overlap analysis between these profiles and functional clusters obtained with the *IntelliGO*-based clustering, leads to characterize subsets of enriched biological functions of genes displaying consistent functions and similar expression profiles.

1 Introduction

In the last decade, DNA microarrays were used for measuring the expression levels of thousands of genes under various biological conditions [11, 8]. Thus, gene expression data analysis proceeds in two steps: Firstly, expression profiles are produced by grouping genes displaying similar expression levels under a given set of situations [13]. Secondly, a functional analysis, based on functional annotations, is applied on genes sharing the same expression profile, in order to identify their relevant biological functions [6, 16, 18, 19]. In fact, one important purpose of this functional analysis is to identify and characterize genes that can serve as diagnostic signatures or prognostic markers for different stages of a disease. One of the most interesting ontology in the biological domain is the Gene Ontology (GO), which is one of the most commonly used source of functional annotations

of genes [5, 1, 2].

This ontology of about 30,000 terms is organized as a controlled vocabulary describing the *biological process* (BP), *molecular function* (MF), and *cellular component* (CC) aspects of gene annotation, also called GO aspects [17]. The GO vocabulary is structured as a rooted Directed Acyclic Graph (rDAG) in which GO terms (concepts) are the nodes connected by different hierarchical relations (mostly *is_a* and *part_of* relations). The *is_a* relation describes the fact that a given child term is a specialization of a parent term, while the *part_of* relation denotes the fact that a child term is a component of a parent term. By definition, each rDAG has a unique root node, relationships between nodes are oriented, and there are no cycles, *i.e.* no path starts and ends at the same node. The GO Consortium regularly updates a GO Annotation (GOA) Database [2] in which appropriate GO terms are assigned to genes or gene products from public databases.

GO is widely used in several complex biological data mining problems. Authors in [20, 7] used GO for gene functional analysis in order to interpret DNA microarrays experiments, by exploiting the commonly accepted assumption that genes having similar expression profile should share similar biological functions. In such analysis, an enrichment study based on statistical P-value calculation are applied to genes sharing the same expression profile [10]. The results usually consist in sets of GO terms characterizing the biological function predominantly represented in a list of genes, thereby suggesting which function or process is affected when the behavior of this group of genes varies. However, the main limitation of these kinds of methods is that they consider the input list of genes in the enrichment analysis, as functionally homogeneous. Nonetheless in practice, genes present in the same expression profile could be involved in multiple biological processes. Thus the statistical tests for extracting specific GO terms could be biased. Moreover, the already proposed methods for gene functional enrichment analysis do not consider exclusively the three aspects of GO, that is not important for the biologists. To overcome this problem, we proposed here a new approach for analyzing gene expression data, by refining and creating subgroups of functionally homogeneous genes. The enrichment analysis could be then applied on each subgroup of genes, thus assuring the extraction of specific biological functions (GO terms) for those genes. The creation of subgroups of genes is performed using a clustering method based on our recently described semantic similarity measure called *IntelliGO* [4], that applies functional comparison between genes annotated by GO terms.

This paper is organized as follows. The next section, outlines the utilization of *IntelliGO* in a functional clustering approach and presents the evaluation results, using the F-score method and collections of reference sets. In second stage (Section III), we present and overlap analysis method that exploits the matching between functional clusters and reference sets. This method is then applied to a list of genes found dysregulated in cancer samples by replacing the reference sets by gene expression profiles. An enrichment analysis is then applied on overlapping genes, and leads to characterize subsets of enriched biological functions of

genes displaying consistent functions and similar expression profiles. Finally in the last section the relevance of the obtained results of the proposed algorithms are discussed.

2 The IntelliGO-based gene functional classification

2.1 Presentation of the datasets

Gene functional clustering aims to regroup genes sharing common biological functions. We used four datasets for evaluating functional classification of genes, already presented in our past study[4]. In each dataset we prepared a collection of reference sets. Each reference set represents a group of genes grouped by an expert due to their shared biological functions. We selected a total of 13 KEGG pathways from the KEGG database [15] for the Biological Process aspect of GO for human (total of 280 genes) and yeast (total of 185 genes) species. For the Molecular Function aspect of GO we chose 10 Pfam clans from the Sanger Pfam database [12] for both species(100 genes for human and 118 for yeast species).

2.2 Calculating similarity matrices and clustering

For performing a gene functional clustering for a given list of genes, the first step is to compute a matrix representing the semantic similarity values between all genes in the input list. This similarity matrix will be then used as parameter of a clustering algorithm. In our case we used both hierarchical and fuzzy clustering. The first method allows to have a global overview of the distribution of genes on different clusters, while the second method allows to a gene to belong to multiple clusters at one time. In fact, one gene could be involved in multiple biological process simultaneously. These two algorithms are available in R-Bioconductor package¹.

Pairwise similarity matrices were calculated for all genes present in the four datasets using our recently proposed similarity measure (*IntelliGO*) [4]. This measure is represented in an innovative vector space model (VSM), and takes into account both information content of annotation terms and their positions in the ontology rDAG [4]. With *IntelliGO VSM*, each gene is represented as a vector \mathbf{g} in a k -dimensional space where the basis vectors \mathbf{e}_i correspond to the k annotation terms. To measure the semantic relationships between terms, we defined a term similarity product as:

$$\mathbf{e}_i \cdot \mathbf{e}_j = \frac{2 * Depth(LCA)}{MinSPL(t_i, t_j) + 2 * Depth(LCA)}. \quad (1)$$

Moreover, we included in the *IntelliGO VSM* a novel weighting scheme in which a coefficient α_i is assigned to each \mathbf{e}_i so that the gene representation becomes: $\mathbf{g} = \sum_i \alpha_i \cdot \mathbf{e}_i$. The coefficients (α_i) combine a weight $w(g, t_i)$ which depends

¹ www.bioconductor.org

on the evidence code tracking the annotation of gene g with a GO term t_i and on the *Inverse Annotation Frequency* ($IAF(t_i)$) which is an estimation of the information content IC of the term t_i . Thus, the similarity between \mathbf{g}_1 and \mathbf{g}_2 is given by the following generalized cosine formula:

$$IntelliGO(\mathbf{g}_1, \mathbf{g}_2) = \frac{\mathbf{g}_1 \cdot \mathbf{g}_2}{\sqrt{\mathbf{g}_1 \cdot \mathbf{g}_1} \sqrt{\mathbf{g}_2 \cdot \mathbf{g}_2}}, \quad (2)$$

with: $\mathbf{g}_1 \cdot \mathbf{g}_2 = \sum_{i,j} \alpha_{1i} \alpha_{2j} e_i \cdot e_j$.

Remark that *IntelliGO* is a *pair-wise* measure involving both *node-based* and *edge-based* similarities. The measure, the clustering algorithms and the used datasets are available at <http://intelligo.loria.fr>.

2.3 Evaluation of the clustering using the F-score method

When reference sets are available, the best method for optimizing the number of classes produced by unsupervised classification approaches is the F-score method [22]. This method relies on pairing each reference set with the best-matched cluster and provides a quantitative estimation of the pairing efficiency (precision and recall). We decided to extend the F-score method in order to further investigate the pairing between reference sets and clusters in a so-called overlap analysis. Our approach is outlined in the following algorithms. Algorithm 1 describes unsupervised clustering optimization with reference sets and global F-score measure. We applied fuzzy C-means clustering on the gene-gene pairwise

Algorithm 1 Clustering optimization with reference sets and F-score measure.

Require: $\Sigma = \{R_1, R_2, \dots, R_p\}$: a collection of reference sets, (n_1, n_2) such that $n_1 < p < n_2$. The pairwise similarity matrix of all elements of Σ .

Ensure: The optimal number of generated clusters \hat{K} , $\widehat{Global\ F - score}(\hat{K})$.

```

1: for each  $K$  in  $[n_1, n_2]$  do
2:   Generate  $K$  clusters  $\Phi = \{C_1, C_2, \dots, C_K\}$ , using all elements in  $\bigcup R_i$ 
3:   for each reference set  $R_i \in \Sigma$  do
4:     for each cluster  $C_j \in \Phi$  do
5:       Precision( $R_i, C_j$ ) =  $|R_i \cap C_j| / |C_j|$ 
6:       Recall( $R_i, C_j$ ) =  $|R_i \cap C_j| / |R_i|$ 
7:        $F - score(R_i, C_j) = \frac{2 * Precision(R_i, C_j) * Recall(R_i, C_j)}{Precision(R_i, C_j) + Recall(R_i, C_j)}$ 
8:     end for
9:      $\widehat{F - score}(R_i) = Max_{C_j \in \Phi} (F - score(R_i, C_j))$ 
10:  end for
11:   $Global\ F - score(K) = \frac{\sum_{i=1}^p (|R_i| * \widehat{F - score}(R_i))}{\sum_{i=1}^p |R_i|}$ 
12: end for
13:  $\widehat{Global\ F - score}(\hat{K}) = Max_{K \in \Phi} Global\ F - score(K)$ 
14: return  $\hat{K}, \widehat{Global\ F - score}(\hat{K})$ .
```

similarity matrices calculated with *IntelliGO* for the four datasets. We used this

clustering algorithm since some genes can be involved in multiple biological processes or molecular functions. The same evaluation procedure was performed on a tool representing the state of the art for gene classification methods (DAVID: Database for Annotation, Visualization, and Integrated Discovery classification tool) [9, 14]. Each clustering result together with the corresponding collection of reference sets served as input to Algorithm 1 for determining global F-scores. Concerning the *IntelliGO*-based fuzzy clustering of Datasets 1 and 2, we varied the number of generated clusters, K , between 11 and 17 in steps of 1 since these datasets are composed of 13 pathways for human and yeast species. For Datasets 3 and 4, the values of K were taken between 8 and 14 with a step of 1, since these two datasets are composed of 10 Pfam clans for both species. For each K , the global F-score(K) value was calculated. Concerning the DAVID functional classification of the same datasets, we varied the *Kappa similarity threshold* between 0.3 to 0.7 with a step of 0.1 in order to obtain different numbers of clusters, since DAVID does not allow the number of clusters to be specified *a priori*. As in the previous case, the K clusters were matched with the input reference sets, and the *Global F - score(K)* value was calculated. The results are presented in Table 1.

Regarding the results obtained with Dataset 1 (13 human KEGG pathways) using our similarity measure, it can be seen that all global F-Score values are greater than 0.5, with a maximum value of 0.62 for $K = 14$. This means that the genes of the 13 human pathways considered in Dataset 1 are best grouped with our measure into 14 functional clusters. This result can reflect the fact that one pathway of the KEGG database encompasses two biological processes and/or that the clustering process has grouped together genes from various pathways sharing common BP annotations.

With DAVID (Table 1), the maximum global F-score (0.67) is reached when *Kappa* = 0.3, giving 10 functional clusters. At higher threshold, the number of genes excluded from the clustering increases, revealing one limit of the DAVID tool. Similar results are obtained with Datasets 2, 3 and 4 and are detailed in Table 1.

In summary these results indicate that *IntelliGO*-based clustering appears as a valuable alternative to DAVID classification tool. It is noteworthy that with DAVID classification tool all maximum values of global F-score are obtained for the minimal *Kappa* similarity threshold (0.3) which corresponds, according to DAVID, to the poorest quality of clustering. Moreover, the calculation of the global F-score is somewhat biased with DAVID as a certain number of genes are excluded from the classification results.

Dataset	A. IntelliGO			B. DAVID tool		
	K	Global F-score	Kappa Thr.	K	Global F-score	% excl.
1 (13 reference sets, total of genes =280)	11	0.59	0.3	10	0.67	20.7
	12	0.61	0.4	11	0.63	31.4
	13	0.61	0.5	14	0.66	38.2
	14	0.62	0.6	11	0.41	75.9
	15	0.56	0.7	8	0.31	68.2
	16	0.55				
	17	0.54				
2 (13 reference sets, total of genes =185)	11	0.59	0.3	9	0.68	17.8
	12	0.62	0.4	8	0.65	31.4
	13	0.64	0.5	9	0.55	43.2
	14	0.67	0.6	6	0.39	56.8
	15	0.66	0.7	7	0.20	69.2
	16	0.62				
	17	0.62				
3 (10 reference sets, total of genes =100)	8	0.70	0.3	11	0.64	27.0
	9	0.64	0.4	11	0.51	52.0
	10	0.68	0.5	8	0.31	66.0
	11	0.75	0.6	3	0.09	93.0
	12	0.66	0.7	2	0.01	96.0
	13	0.66				
	14	0.64				
4 (10 reference sets, total of genes =118)	8	0.79	0.3	10	0.70	40.7
	9	0.77	0.4	9	0.47	61.0
	10	0.78	0.5	9	0.39	69.5
	11	0.82	0.6	5	0.28	84.7
	12	0.78	0.7	3	0.21	91.5
	13	0.78				
	14	0.71				

Table 1. Variation of the global F-score values when (A) varying the number of generated fuzzy clusters K with the fuzzy C -Means algorithm using *IntelliGO* similarity measure and (B) varying the Kappa threshold (Kappa thr.) with DAVID functional classification tool. In B the percentage of genes that are excluded from the classification is indicated (% excl.). Results are shown for the four datasets used in this study (total number of genes between parentheses). The optimal K value and the corresponding maximal global F-score value are in bold.

3 Overlap analysis between functional clusters and reference sets

3.1 Overlap analysis algorithm

In order to refine our comparison, we decided to look at the matching between the reference sets and the clusters obtained with the optimal K value. We used Algorithm 2 to extract the top-ranked cluster from each list of clusters assigned to each reference set. This algorithm explains how clusters (C) are assigned to reference sets (R) according to the F-score values, allowing the identification of best-matching pairs ($R \cap C$).

The intersection $R \cap C$ is expected to display a highly homogeneous content composed of genes known as members of a reference set and found most similar by clustering. Alternatively, the two set-theoretic differences $C \setminus R$ and $R \setminus C$ can be considered in order to discover missing information. In our study, we are interested by genes present in $R \cap C$. Indeed, we apply an enrichment analysis on genes present in such intersection, in order to extract specific functions.

3.2 Application to cancer expression data

In this section, we present an application of the *IntelliGO*-based clustering and overlap analysis approach using a list composed of 128 genes relating to human

Algorithm 2 Assignment of clusters to reference sets according to the F -score values.

Require: $\Sigma = \{R_1, R_2, \dots, R_p\}$: a collection of reference sets, $\Phi_K = \{C_1, C_2, \dots, C_K\}$: a collection of clusters, $\forall (i, j) \mid 1 \leq i \leq p, 1 \leq j \leq K, F\text{-score}(R_i, C_j)$ (see Algorithm 1).

Ensure: A ranked list of clusters, ordered by decreasing F -score, assigned to each reference set.

- 1: **for** each reference set $R_i \in \Sigma$ **do**
 - 2: $List_i \leftarrow (C_j, F\text{-score}(R_i, C_j))$: A list of clusters C_j ordered by decreasing values of $F\text{-score}(R_i, C_j)$
 - 3: **print** $List_i$.
 - 4: **end for**
-

colorectal cancers. The idea here is to confront the *IntelliGO* functional clusters of the 128 genes, and to consider as reference sets the *fuzzy Differential Expression Profiles* (fuzzy DEP) obtained from the same list of genes [3]. Here, each DEP represents a group of genes having similar expression profile. We believe that overlap analysis may lead to discover hidden relationships between gene expression and biological function. Fuzzy DEPs are considered here as a collection of reference sets for overlap analysis. More precisely, 8 fuzzy DEPs containing genes with GO annotation are retained from our previous study [3]. The pair-wise similarity matrix was generated for the 128 genes, and the number of clusters, k , was optimized with the Algorithm 1 using the 8 fuzzy DEPs as reference sets ($\Sigma = \{DEP_i\}, i = 1..8$). The optimal number of cluster was obtained for $k = 3$ with and F-score value equals to 0.4.

After that, Algorithm 2 was used to extract lists of genes present in $C \cap DEP$, *i.e.* displaying both functional similarity (C) and present in one of the eight fuzzy DEPs (R). The enrichment analysis could be then applied on these signature genes, to discover among them statistically significant GO terms displaying low P-Value. In our case, the P-value is calculated for genes present in $C \cap DEP$ versus a background list (here all human genes) displaying GO annotation in the NCBI repository file², using the *hyper geometric test* [10].

Preliminary results have shown that very specific biological functions with inferior P-Values ($\leq 10E-04$) were extracted for genes present in $C \cap DEP$. For example, genes in $Cluster_1 \cap PED3$ have "regulation of transcription DNA-dependent" and "NADH oxidation", as very specific functions (non exhaustive). Genes of $Cluster_2 \cap PED2$ have the following functions: "cell differentiation", "multicellular organismal development", "insulin secretion". Genes of $Cluster_3 \cap PED14$ have the "Water transport" as specific function. The "Transport" processes are very important in the physiology of the digestive system. This function was found for the *AQP8* (Aquaporine 8) human gene, which is found in the literature under expressed in the tumoral tissues. This gene belongs to *PED14* which regroups genes under expressed in cancer versus normal tissue [3]. This observation could be considered as a positive witness of our strategy. Other similar results were obtained for the remaining PED, are not reported here.

² ftp://ftp.ncbi.nih.gov/gene/DATA/gene2go.gz

4 Conclusion and perspectives

In this paper, we have presented a gene set enrichment analysis based on functional clustering with the *IntelliGO* semantic similarity measure. In a first step, we proposed an algorithm for evaluation the clustering approach using reference sets and the F-score method. Very encouraging results were obtained with *IntelliGO* when compared with a well known classification method (DAVID tool). Beyond clustering *per se*, we have presented an overlap analysis method which leads to a pairing of clusters and reference sets and may be used for set-difference analysis. Applied to a list of genes from a transcriptomic cancer study, our method leads to identify subsets of genes displaying consistent expression and functional profiles. Promising results have been obtained using a simple GO term enrichment procedure. More sophisticated tools such as GSEA [21] could be used to improve the biological interpretation of these subsets of genes.

References

1. Michael Ashburner, Catherine Ball, Judith Blake, David Botstein, Heather Butler, Michael Cherry, Allan Davis, Kara Dolinski, Selina Dwight, Janan Eppig, Midori Harris, David Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel Richardson, Martin Ringwald, Gerald Rubin, and Gavin Sherlock. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.
2. Daniel Barrell, Emily Dimmer, Rachael P. Huntley, David Binns, Claire O’Donovan, and Rolf Apweiler. The GOA database in 2009—an integrated Gene Ontology Annotation resource. *Nucl. Acids Res.*, 37(suppl1):D396–403, 2009.
3. Sidahmed Benabderrahmane, Marie-Dominique Devignes, Malika Smail-Tabbone, Amedeo Napoli, Olivier Poch, Ngoc-H. Nguyen, and Wolfgang Raffelsberger. Analyse de données transcriptomiques: Modélisation floue de profils d’expression différentielle et analyse fonctionnelle. In *INFORSID*, pages 413–428, 2009.
4. Sidahmed Benabderrahmane, Malika Smail-Tabbone, Olivier Poch, Amedeo Napoli, and Marie-Dominique Devignes. Intelligo: a new vector-based semantic similarity measure including annotation origin. *BMC Bioinformatics*, 11(1):588, 2010.
5. Olivier Bodenreider. Special issue: Biomedical ontology in action. *Applied Ontology*, 4(1):1–4, 2009.
6. J-F Boulicaut and O. Gandrillon. *Informatique pour l’analyse du transcriptome*. Hermes Lavoisier, Paris, 2004.
7. Markus Brameier and Carsten Wiuf. Co-clustering and visualization of gene expression data and gene ontology terms for *saccharomyces cerevisiae* using self-organizing maps. *J. of Biomedical Informatics*, 40(2):160–173, 2007.
8. Alvis Brazma, Jaak Vilo, and Edited Gianni Cesareni. Gene expression data analysis. *FEBS Letters*, 480:17–24, 2000.
9. Glynn Dennis, Brad Sherman, Douglas Hosack, Jun Yang, Wei Gao, H Lane, and Richard Lempicki. David: Database for annotation, visualization, and integrated discovery. *Genome Biology*, 4(9):R60, 2003. A previous version of this manuscript was made available before peer review at <http://genomebiology.com/2003/4/5/P3>.
10. Eran Eden, Roy Navon, Israel Steinfeld, Doron Lipson, and Zohar Yakhini. Gorilla: a tool for discovery and visualization of enriched go terms in ranked gene lists. *BMC Bioinformatics*, 10(1):48, 2009.

-
11. M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 95(25):14863–14868, December 1998.
 12. Robert D. Finn, Jaina Mistry, Benjamin Schuster-Bockler, Sam Griffiths-Jones, Volker Hollich, Timo Lassmann, Simon Moxon, Mhairi Marshall, Ajay Khanna, Richard Durbin, Sean R. Eddy, Erik L. L. Sonnhammer, and Alex Bateman. Pfam: clans, web tools and services. *Nucl. Acids Res.*, 34.
 13. Audrey Gasch and Michael Eisen. Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biology*, 3(11):research0059.1–research0059.22, 2002.
 14. Da Huang, Brad Sherman, Qina Tan, Jack Collins, W Gregory Alvord, Jean Roayaei, Robert Stephens, Michael Baseler, H Clifford Lane, and Richard Lempicki. The david gene functional classification tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biology*, 8(9):R183, 2007.
 15. Minoru Kanehisa, Susumu Goto, Miho Furumichi, Mao Tanabe, and Mika Hirakawa. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic acids research*, 38(Database issue):D355–360, January 2010.
 16. Purvesh Khatri and Sorin Draghici. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, 21(18):3587–3595, 2005.
 17. P. W. Lord, R. D. Stevens, A. Brass, and C. A. Goble. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, 19(10):1275–1283, 2003.
 18. David Martin, Christine Brun, Elisabeth Remy, Pierre Mouren, Denis Thieffry, and Bernard Jacq. GOToolBox: functional analysis of gene datasets based on Gene Ontology. *Genome Biology*, 5(12), 2004.
 19. Brad Sherman, Da Huang, Qina Tan, Yongjian Guo, Stephan Bour, David Liu, Robert Stephens, Michael Baseler, H Clifford Lane, and Richard Lempicki. David knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis. *BMC Bioinformatics*, 8(1):426, 2007.
 20. Nora Speer, Christian Spieth, and Andreas Zell. A memetic co-clustering algorithm for gene expression profiles and biological annotation. 2004.
 21. Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, 2005.
 22. C. J. van Rijsbergen. *Information Retrieval*. Butterworth, 1979.

Theoretical Overview of Machine translation

Mohamed Amine Chéragui¹

¹ African University, Adrar, Algeria,
m_cheragui@univadrar.org

Abstract. The demand for language translation has greatly increased in recent times due to increasing cross-regional communication and the need for information exchange. Most material needs to be translated, including scientific and technical documentation, instruction manuals, legal documents, textbooks, publicity leaflets, newspaper reports etc. Some of this work is challenging and difficult but mostly it is tedious and repetitive and requires consistency and accuracy. It is becoming difficult for professional translators to meet the increasing demands of translation. In such a situation the machine translation can be used as a substitute.

This paper offers a brief but condensed overview of Machine Translation (MT). Through the following points: History of MT, Architectures of MT, Types of MT, and evaluation of M T.

Keywords: History of MT, Architecture of MT, Types of MT, evaluation of MT.

1 Introduction

After 65 years, this field is one of the oldest applications of computers. Over the years, Machine Translation has been a focus of investigations by linguists, psychologists, philosophers, computer scientists and engineers. It will not be an exaggeration to state that early work on MT contributed very significantly to the development of such fields as computational linguistics, artificial intelligence and application-oriented natural language processing.

Machine translation, commonly known as MT, can be defined as “translation from one natural language (source language (SL)) to another language (target language (TL)) using computerized systems and, with or without human assistance”[1] [2].

We try to give in this paper a coherent, if necessarily brief and incomplete, the development has been the field of machine translation through four points which are: first of all surveys the chronological development of machine translation, the different approaches developed (linguistic and computational), the types of machine translation and finally, we try to answer an important question which is how to evaluate a machine translation?

2 History of Machine Translation

Although we may trace the origins of machine translation (MT) back to seventeenth century ideas of universal (and philosophical) languages and of 'mechanical' dictionaries, it was not until the twentieth century that the first practical suggestions could be made. The history of machine translation can be divided into five (05) periods [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12] :

2.1 First period (1948-1960): The beginning.

- 1949 : Warren Weaver in his Memorandum of 1949 proposed the first ideas on the use of computers in translation, by adopting the term computer translation.
- 1952 : The first symposium of machine translation, entitled Conference on Machine Translation, held in July 1952 at MIT under leadership of Yehoshua Bar-Hillel.
- 1954 : The development of the first automatic translator (very basic) by a group of researchers from Georgetown University in collaboration with IBM, which translates into more than sixty (60) Russian sentences into English. The authors claimed that within three to five years, machine translation would not be a problem.
- 1954 : Victor Yngve published the first journal on MT, entitled « Mechanical translation devoted to the translation of languages by the aid of machines ».

2.2 Second Period (1960-1966) Parsing and disillusionment

- Early 1960s This parsing is put forward as the only possible avenue of research to advance the machine translation. Thus there are already many parsers developed from different types of grammars, such as grammar and dependency grammar Tesnière stratificationnelle Lamb
- 1961 : In February of this year that computational linguistics is born, thanks to weekly lectures organized by David G. Hays at the Rand Corporation in Los Angeles. These conferences will be included as papers at the First International Conference on Machine Translation of Languages and Applied Language Analysis of Teddington in September 1961 with the participation of linguists and computer scientists involved in the translation as: Paul Garvin, Sydney M. Lamb, Kenneth E. Harper, Charles Hockett, Martin Kay and Bernard Vauquois.
- 1964 : the creation of committee ALPAC(Automatic Language Processing Advisory Committee) with American government to studies the perspectives and the chances of machine translation
- 1966 : ALPAC published his famous rapport in which it concluded that its works on machine translation is just wasting of time and money ; the conclusion of this rapport is it had a negative impact on their search (MT) for a number of years

2.3 Third period (1966-1980): New birth and hope

- 1970 : Start of the project REVERSO by a group of Russian researchers.
- 1970 : Development of System SYSTRAN1 (Russian-English) by Peter Toma, who was at that time a member of a group search for Georgetown.
- 1976 : Creation of system WEATHER in the project TAUM (machine translation in the university of Montreal) under the direction of Alai Colmerauer for the machine translation weather forecasts for the general public, this system was created by group of researchers
- 1978 : Creation of system ATLAS2 by the Japanese firm FUJITSU, this translator was based on rules also he is able to translate from Korean to Japanese and vice versa

2.4 Fourth Period (1980-1990): Japanese invaders

- 1982 : The Japanese firm SHARP markets its Automatic translator DUET (English - Japanese), this translator was based on rules an approach to translation transfer
- 1983: as computer giant, NEC develops it's own system of translation based on algorithm called PIVOT. Marketed under the name of Honyaku Adaptor II, the version public the system of translation of NEC is also based on the method of pivot, by using Interlingua.
- 1986: Development of system PENSEE by OKI3, which is a translator (Japanese-English) based on rules.
- 1986: The group Hitachi developed his own translation system based on rules (which is an approach taken by transfer), christened on HICATS (Hitachi Computer Aided Translation System / Japanese- English).

2.5 Fifth Period (since 1990): the Web and the new vague of translators

- 1993: The project C-STAR (Consortium for Speech Translation Advanced Research) is an international cooperation. The theme of project is the machine translation of the parole in the field of tourism (dialogue client travel agent), by videoconference. these project birth the system C-STAR I which dealt three (03) languages (English, German et Japanese) and made the first demonstrations transatlantic trilingual in January 1993
- 1998: Marketing the translator REVERSO by the company Softissimo.
- 2000: the Development of system ALPH by Japanese laboratory ATR, this translator (Japanese-English and Chinese - English) takes an approach based on examples.

¹ The same translator was adopted by the European commission 1976 for the translation (Japanese-English)

² Currently we are in version 14 of the translator.

³ OKI : founded in 1881 Oki Electric Industry Co, is a Japanese manufacturer of telecommunications

- 2005: The appearance of the first web site for automatic translation ,like Google (<http://translate.google.fr/>).
- 2007: METIS-II is a hybrid machine translation system, in which insights from Statistical, Example based, and Rule-based Machine Translation (SMT, EBMT, and RBMT respectively) are used.
- 2008 : 23% of internet users, have used the machine translation and 40 % considering doing so
- 2009: 30% the professionals have used the machine translation and 18% perform a proofreading.
- 2010: 28% of internet users, have used the machine translation and 50% planning to do.

3 Architectures of machine translation systems

Different strategies have been adopted by different researchers at different times in the history of machine translation. The choice of strategy reflects one side of the depth and linguistic diversity but also the grandeur of ambition on the other side. There are generally two types of architecture for machine translation, which are:

3.1 Linguistic Architecture

In the linguistic architecture there are three basic approaches being used for developing MT systems that differ in their complexity and sophistication. These approaches are:

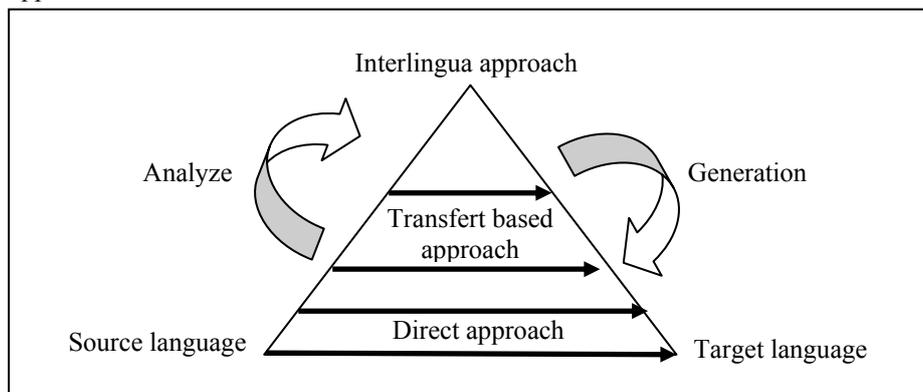


Fig1. The Vauquois triangle

- **Direct approach:** In direct translation, translation is direct from the source text to the target text. The vocabularies of SL texts are analyzed as needed for the resolution of SL ambiguities, for the correct identification of TL expressions as well as for the specification of word order in TL. This approach involves taking a string of words from the source language, removing the morphological

inflection from words to obtain the base forms, and looking them up in a bilingual dictionary between the source and the target languages. Components of this system are a large bilingual dictionary and a program for lexically and morphologically analyzing and generating texts [13].

- **Transfer-based approach:** In the Transfer approach, translation is completed through three stages: the first stage consists in converting SL texts into an intermediate representation, usually parse trees; the second stage converting these representations into equivalent ones in the target language; and the third one is the generation of the final target text [13].

In the transfer approach, the source text is analyzed into an abstract representation that still has many of the characteristics of the source, but not the target, language. This representation can range from purely syntactic to highly semantic. In the syntactic transfer, some type of tree manipulation into a target language tree converts the parse tree of the source input. This can be guided by associating feature structures with the tree. Whatever representation is used, transfer to the target language is done using rules that map the source language structures into their target language equivalents. Then in the generation stage, the mapped target structure is altered as required by the constraints of the target language and the final translation is produced.

- **Interlingua approach:** The Interlingua approach is the most suitable approach for multilingual systems. It has two stages: Analysis (from SL to the Interlingua) and Generation (from the Interlingua to the TL). In the analysis phase, a sentence in the source language is analyzed and then its semantic content is extracted and represented in the Interlingua form representation, where an Interlingua is an entirely new language that is independent of any source or target language and is designed to be used as an intermediary internal representation of the source text. The analysis phase is followed by the generation of the target sentences from the Interlingua representation. An analysis program for a specific SL can be used for more than one TL since it is SL-specific and not oriented to any particular TL. Furthermore, the generation program for a particular TL can be used again for translation from every SL to this particular TL since it is TL-specific and not designed for input from a particular SL [13].

3.2 Computational Architecture

- **Rule Based approach:** rule-based MT has two approaches: Interlingua and transfer. Rule-Based MT Systems rely on different levels of linguistic rules for translation. This MT research paradigm has been named rule-based MT due to the use of linguistic rules of diverse natures. For instance, rules are used for lexical transfer, morphology, syntactic analysis, syntactic generation, etc. In RBMT the translation process consists of:
 - Analyzing input text morphologically, syntactically and semantically.
 - Generating text via structural conversions based on internal structures.

The steps mentioned above make use of a dictionary and a grammar, which must be developed by linguists. This requirement is the main problem of RBMT as it is a time-consuming process to collect and spell out this knowledge, frequently referred as knowledge acquisition problem. It is not just very hard to develop and maintain the rules in this type of system, but one is not guaranteed to get the system to operate as well as before the addition of a new rule. RBMT systems are large-scale rule based systems; whereas their computational cost is high, since they must implement all aspects whether syntactic, semantic, structural transfer etc. as rules [14].

- **Corpus-based approach:** Corpus-Based Machine Translation, also referred as data driven machine translation, is an alternative approach for machine translation to overcome the knowledge acquisition problem of rule-based machine translation. There are two types of CBMT Statistical Machine Translation (SMT) and Example-Based Machine Translation (EBMT). Corpus-based MT automatically acquires the translation knowledge or models from bilingual corpora. Since this approach has been designed to work on large sizes of data, it has been named Corpus-Based MT ([17], [18], [16] and [15]).
- **Hybride approach:** Some recent work has focused on hybrid approaches that combine the transfer approach with one of the corpus-based approaches. This was designed to work with fewer amounts of resources and depend on the learning and training of transfer rules. The main idea in this approach is to automatically learn syntactic transfer rules from limited amounts of word-aligned data. This data contains all the needed information for parsing, transfer, and generation of the sentences ([19] and [20]). The following section covers part of the MT literature that gives details of specific systems for deriving the appropriate translation using different approaches.

4 Types of Machine Translation

4.1 Machine Translation for Watcher (MT-W)

This is intended for readers who wanted to gain access to some information written in foreign language who are also prepared to accept possible bad translation rather than nothing. This was the type of MT envisaged by the pioneers. This came in with the need to translate military technological documents. This was almost the dictionary-based translation far away from linguistic based machine translation [25].

4.2 Machine Translation for Revisers (MT-R)

This type aims at producing raw translation automatically with a quality comparable to that of the first drafts produced by human. The translation output can be considered only as brush-up so that the professional translator freed from that very boring and time consuming task can be promoted to revisers [25].

4.3 Machine Translation for Translators (MT-T)

This aims at helping human translators do their job by providing on-line dictionaries, thesaurus and translation memory. This type of machine translation system is usually incorporated into the translation work stations and the PC based translation tools. “Tools for individual translators have been available since the beginning of office automation.” And those systems running on standard platforms and integrated with several text processors are the ones that attained operational and commercial success [25].

4.4 Machine Translation for Authors (MT-A)

This aims at authors wanting to have their texts translated into one or several languages and accepting to write under control of the system or to help the system disambiguate the utterance so that satisfactory translation can be obtained without any revision. This is an “interactive MT, The interaction was however done both during analysis and during transfer, and not by authors, but by specialists of the system and language(s).” In short, there have been no operational successes yet in MT-A, but the designs are becoming increasingly user oriented and geared towards the right kind of potential users, people users, people needing to produce translations, preferably into several languages [25].

5 Evaluation of Machine Translation Systems

Evaluating Machine translation system is important not only for its potential users and buyers, also to researchers and developers. Various types of evaluation have been developed, such as :

5.1 BLEU (BiLingual Evaluation Understudy)

The BLEU metric, proposed by Papineni in 2001 was the first automatic measurement accepted as a reference for the evaluation of translations. The principle of this method is to calculate the degree of similarity between candidate (machine) translation and one or more reference translations based on the particular n-gram precision. The BLEU score is defined by the following formula [21]:

$$\text{BLEU} = \text{BP} \times e^{\left(\sum_{n=1}^N w_n \log p_n\right)} \quad (1)$$

Where:

- “ p_n ”: the number of n-grams of machine translation is also present in one or more reference translation, divided by the number of total n-grams of machine translation.
- “ w_n ”: positive weights.

- “BP”: Brevity Penalty, which penalizes translations for being “too short”. The brevity penalty is computed over the entire corpus and was chosen to be a decaying exponential in “r/c”, where “c” is the length of the candidate translation and “r” is the effective length of the reference translation.

$$BP = \begin{cases} 1 & \text{Si } c > r \\ e^{1-\frac{r}{c}} & \text{Si } c \leq r \end{cases} \quad (2)$$

5.2 WER (Word Error Rate)

The WER metric, Proposed by Popovic and Ney in 2007. Originally used in Automatic Speech Recognition, compares a sentence hypothesis refers to a sentence based on the Levenshtein distance. It is also used in machine translation to evaluate the quality of a translation hypothesis in relation to a reference translation. For this, the idea is to calculate the minimum number of edits (insertion, deletion or substitution of the word) to be performed on hypothesis translation to make it identical to the reference translation. The number of editss to be performed, noted “ $d_L(\text{ref}, \text{hyp})$ ” is then divided by the size of the reference translation, denoted “ N_{ref} ” as shown in the following formula [22]:

$$WER = \frac{1}{N_{ref}} \times d_L(\text{ref}, \text{hyp}). \quad (3)$$

Where:

- $d_L(\text{ref}, \text{hyp})$: is the Levenshtein distance between the reference translation “ref” and the hypothesis tanslation “hyp”.

A shortcoming of the WER is the fact that it does not allow reordering of words, whereas the word order of the hypothesis can be different from word order of the reference even though it is correct translation.

5.3 PER (Position-independent word Error Rate)

The PER metric, proposed by Tillman in 1997. compare the words of machine translation with those of the reference regardless of their sequence in the sentence. The PER score is defined by the following formula [23]:

$$PER = \frac{1}{N_{ref}} \times d_{per}(\text{ref}, \text{hyp}). \quad (4)$$

Where:

- d_{per} : calculates the difference between the occurrences of words in machine translation and the translation of reference.

A shortcoming of the PER is the fact that the word order can be important in some cases.

5.4 TER (Translation Error Rate)

The TER metric, proposed by Snover in 2006. Is defined as the minimum number of edits needed to change a hypothesis so that it exactly matches one of the references. The possible edits in TER include insertion, deletion, and substitution of single words, and an edit which moves sequences of contiguous words. Normalized by the average length of the references. Since we are concerned with the minimum number of edits needed to modify the hypothesis, we only measure the number of edits to the closest reference. The TER score is defined by the following formula [24]:

$$TER = \frac{Nb (op)}{Avreg N_{Ref}} \quad (5)$$

Where:

- Nb (op) : is the minimum number of edits;
- Avreg N_{ref}: the average size in words references.

6 Conclusion

In conclusion, we can say that the field of machine translation has been and remains a key focus of research on natural language processing and that led to the development of many positive results. However, perfection is still far away. If the translators have today reached a level of reliability and efficiency in a technical text, perfection is still a long way in the literary text, overwhelmed by the intricacies, the puns and colorful expressions. We think it must look to the construction of a translator hybrid (combining statistical and rules) at the end to increase the performance of the translation system.

References

1. Hutchins, W. J. and Somers, H. L., An introduction to machine translation, Academic Press, London. (1992)
2. Baumgartner-Bovier, “ La traduction automatique, quel avenir ? Un exemple basé sur les mots composés ”, Cahiers de Linguistique Française N°25, (2003).
3. J. Chandioux, “Histoire de la traduction automatique au Canada”, journal des traducteurs, vol. 22, n° 1, p. 54-56, (1977).
4. H. Kaji, “HICATS/JE : A Japanese-to-English Machine Translation System Based on Semantics ”, Machine Translation Summit, (1987).
5. Y. Lepage, E. Denoual, “ALEPH: an EBMT system based on the preservation of proportional analogies between sentences across languages ”, (2005).
6. Y. Fukumochi, “A Way of Using a Small MT System in Industry ”, the 5th Machine Translation Summit, July 10-13, (1995).

7. M. Cori et J. Léon , “ La constitution du TAL Étude historique des dénominations et des concepts ”, TAL. Volume 43 – n° 3/(2002).
8. C. Granell, “La Traduction automatique, Pour qui ? Pour Quoi ? ”, Support de cours, Novembre (2010).
9. P. P. Monty, “Traduction statistique par recherche locale”, , Mémoire de Maitre des sciences en informatique, Université de Montréal, (2010).
10. F. Yvon, “Une petite introduction au traitement Automatique du langage naturel, support de cours ”, Ecole Nationale Supérieure des télécommunications, Avril (2007).
11. C. Fuchs, B. Habert, “ Introduction le traitement automatique des langues : des modèles aux ressources ”, Article paru dans Le Français Moderne LXXII Volume1, (2004).
12. P. Bouillon, “Traitement automatique des langues naturelles ”, édition Duculot, (1998).
13. Hutchins J., Machine Translation: A Brief History, Concise History of the Language Sciences: From the Sumerians to the Cognitivists. Koerner E. F. K. and Asher R. E. (ed.). Oxford: Pergamon Press, pp. 431- 445, (1995).
14. Sumita E., Iida H., and Kohyama H., Translating with Examples: A New Approach to Machine Translation, the Third International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Language, pp. 203–212, (1990).
15. Lavie L., Vogel S., Peterson E., Probst K., Font-Llitjós A., Reynolds R., Carbonell J., and Cohen R., Experiments with a Hindi-to-English Transfer-Based MT System under a Miserly Data Scenario, ACM Transactions on Asian Language Information Processing TALIP, Papineni, K., Roukos, S., Ward, and T.: Maximum Likelihood and Discriminative, pp.143 – 163, (2004).
16. Imamura K., Okuma H., Watanabe T., and Sumita E., Example-based Machine Translation Based on Syntactic Transfer with Statistical Models, Proceedings of the 20th International Conference on Computational Linguistics, Vol. 1, University of Geneva, Switzerland, pp. 99-105, August (2004).
17. Imamura K., Doctor's Thesis Automatic Construction of Translation Knowledge for Corpus-based Machine Translation, May 10, (2004).
18. Lavie L., Vogel S., Peterson E., Probst K., Wintner S., and Eytani Y., Rapid Prototyping of A Transfer-Based Hebrew-to-English Machine Translation System, Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation TMI-04. Baltimore, MD USA, pp.1-10, October (2004).
19. Probst K., Peterson E., Carbonell J and Levin L., MT for Minority Language Using Elicitation-based Learning of Syntactic Transfer Rules. Machine Translation 17: 245-270, Kluwer Academic Publishers, pp. 245 – 270, (2002).
20. Zantout R., and Guessoum A., Arabic Machine Translation: A Strategic Choice for the Arab World, journal of King Saud University, Volume 12, pp. 299-335, (2000).
21. K. Papineni, S. Roukos, T. Ward, and W. Zhu, “Bleu: a Method for Automatic Evaluation of Machine Translation”, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, pp. 311-318, (2002).
22. M. Popovic and H. Ney.” Word error rates: Decomposition over POS classes and applications for error analysis”. In Proceedings of ACL Workshop on Machine Translation.
23. C. Tillman, , S. Vogel, H. Ney, H. Sawaf, and A. Zubiaga. .”Accelerated DP-based search for statistical translation”. In Proceedings of the 5th European Conference on Speech Communication and Technology, pp- 2667.2670. Rhodes, Greece. (1997).
24. M. Snover, B. Dorr, , R. Schwartz, L. Micciulla, J. Makhoul.: “A Study of Translation Edit Rate with Targeted Human Annotation”. In Proceedings of AMTA, Boston, (2006).
25. Abdullah H. Homiedan, “Machine translation”, Journal of King Saud University, Language & Translation Vol 10, pp.1.21, (1998).

Ontologies engineering and Applications



Effective Ontology Learning : Concepts' Hierarchy Building using Plain Text Wikipedia

Khalida Ben Sidi Ahmed, Adil Toumouh, and Mimoun Malki

Department of Computer Science, Djillali Liabes University,
Sidi Bel Abbes, Algeria
`send.to.khalida@gmail.com`

Abstract. Ontologies stand in the heart of the Semantic Web. Nevertheless, heavyweight or formal ontologies' engineering is being commonly judged to be a tough exercise which requires time and heavy costs. Ontology Learning is thus a solution for this exigency and an approach for the 'knowledge acquisition bottleneck'. Since texts are massively available everywhere, making up of experts' knowledge and their know-how, it is of great value to capture the knowledge existing within such texts. Our approach is thus an interesting research work which tries to answer the challenge of creating concepts' hierarchies from textual data. The significance of such a solution stems from the idea by which we take advantage of the Wikipedia encyclopedia to achieve some good quality results.

Keywords : domain ontologies, ontology learning from texts, concepts' hierarchy, Wikipedia.

1 Introduction : Ontology Learning

Ontologies are an extremely essential approach mainly used in order to represent acquired knowledge. The ontology of a certain domain is about all essential concepts of it, their specifications, their hierarchies, whatever relations they have, and the axioms that constraint their behaviour [1]. The greatest challenge to use ontologies is the Semantic Web. It should be noted that the success of this new Web generation is above all dependent on the proliferation of ontologies, which require speed and simplicity in engineering them [2].

However, ontology engineering is a tough exercise which can involve a great deal of time and considerable costs. The need for (semi) automatic domain ontologies' extraction has thus been rapidly felt by the research world. Ontology learning is then the research realm referred to. As a matter of fact, this field is the automatic or semi-automatic support for the ontology engineering. It has indeed the potential to reduce the time as well as the cost of creating an ontology. For this reason, a plethora of ontology learning techniques have been adopted and various frameworks have been integrated with standard ontology engineering tools [3]. Since the fully automation of these techniques remains in the distant

future, the process of ontology learning is argued to be semi-automatic with an insistent need for human intervention.

Most of the knowledge available on the Web represents natural language texts [4]. Semantic Web establishment depends a lot on developing ontologies for this category of input knowledge. This is the reason why this paper focuses especially on ontology learning from texts. One of the still thorny issues of domain ontology learning is concepts' hierarchy building. In this paper, we are primarily involved in creating domain concepts' hierarchies from texts. We plan to use Wikipedia in order to foster the quality of our results. From this optics, literature reviews few research works dealing with this issue and none is making use of Wikipedia on the same way that it is harnessed in our approach.

In fact, Wikipedia is recently showing a new potential as a lexical semantic resource [5]. When this collaboratively constructed resource is used to compute semantic relatedness [6, 7] using its categories' system, this same system is also used to derive large scale taxonomies [8] or even to achieve knowledge acquisition [9]. The idea of harnessing Wikipedia plain text articles in order to acquire knowledge is quite promising. Our approach capitalizes on the well organized Wikipedia articles to retrieve the most useful information at all, namely the definition of a concept.

First, we will describe in Section 2 the ontology learning layer cake. In Section 3, we move straightforward to the explanation of our approach which will be followed by a corresponding evaluation in Section 4. Finally, Section 5 sheds the lights on some conclusions and research perspectives.

2 Ontology Learning Layer Cake

The process of extracting a domain ontology can be decomposed into a set of steps, summarized by [10] and commonly known as "ontology learning layer cake". The following page contains the figure which illustrates these steps.

The first step of the ontology learning process is to extract the terms that are of great importance to describe a domain. A term is a basic semantic unit which can be simple or complex. Next, synonyms among the previous set of terms should be extracted. This allows associate different words with the same concept whether in one language or in different languages. These two layers are called the lexical layers of the ontology learning cake. The third step is to determine which of the existing terms, those who are concepts. According to [10], a term can represent a concept if we can define: its intention (giving the definition, formal or otherwise, that encompasses all objects the concept describes), its extension (all the objects or instances of the given concept) and to report its lexical realizations (a set of synonyms in different languages).

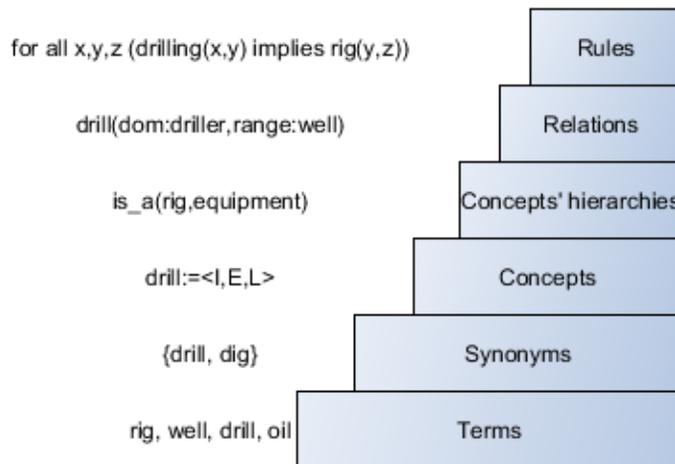


Fig. 1. Ontology learning layer cake (adapted from [10])

The extraction of concepts hierarchies, our key concern, is to find the relationship 'is-a', ie classes and subclasses or hyperonyms. This phase is followed by the non-taxonomic relations' extraction which consists on seeking for any relationship that does not fit in a previously described taxonomic framework. The extraction of axioms is the final level of the learning process and it is argued to be the most difficult one. To date, few projects have attacked the discovery of axioms and rules from text.

3 Concepts' Hierarchy Building Approach

Our approach tackles primarily the construction of concepts' hierarchies from text documents. We will make a terminology extraction using a dedicated tool for this task which is TermoStat [11]. The initial terms will be the subjects of a definitions' investigation within Wikipedia. Adapting the idea of the lexicosyntactic patterns defined by [12] to our case, the hyperonyms of our terms will be learned. This process is iterative which comes to its end when an in advance predefined maximum number of iterations is reached. Our algorithm generates in parallel a graph which unfortunately contains cycles and its nodes may have more then one hyperonym. The hierarchy we promise to build is the transformation result of the graph to a forest focusing on the hierarchic structure of a taxonomy. The figure on the following page gives the overall idea of the proposed approach.

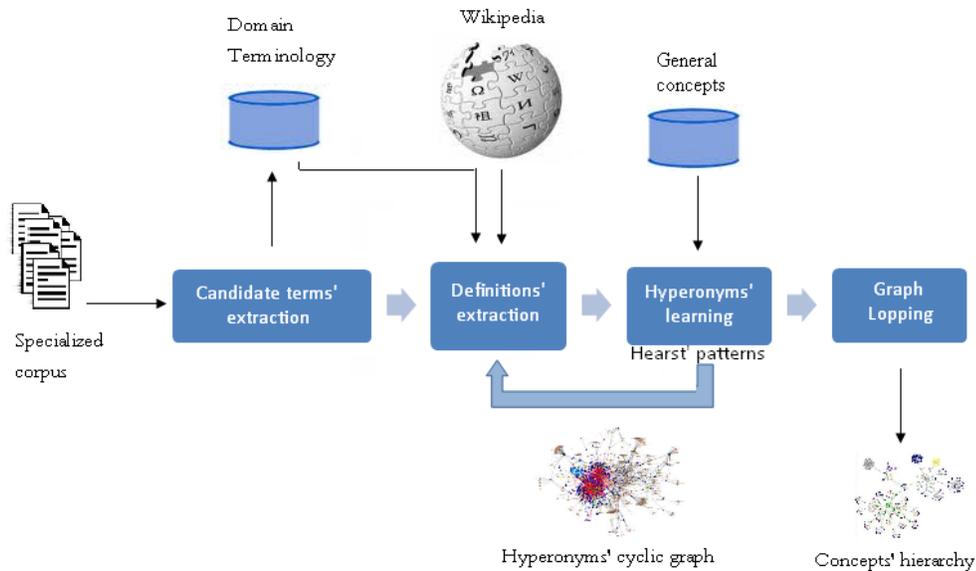


Fig. 2. Steps of the proposed approach

3.1 Preliminary Steps

In order to carry out our approach, we should first undergo the two lexical ontology learning's layers. The tool we used for the sake of retrieving the domain terminology is TermoStat. This web application was favored for determined reasons. In fact, TermoStat requires a corpus of textual data and, juxtaposing it to a generalized corpus such as BNC (British National Corpus), will give us a list of the domain terms that we need for the following step. Afterwards, we try to find out the synonyms among this list of candidate terms. The use of thesaurus.com as a tool in order to select synonyms was efficient. The third layer can be skipped in our context; concepts' hierarchies construction does not depend on the concepts' definitions. In other words, our algorithm needs mainly the candidate terms elected to be representative for the set of its synonyms (synset). The set of initial candidate terms is named \mathcal{C}_O .

3.2 Concepts' Hierarchy

The approach we are proposing belongs to two research paradigms, namely concepts' hierarchies construction for ontology learning and secondly the use of Wikipedia for knowledge extraction. The achievement of our solution relies heavily on concepts from graphs' theory.

a. Hyperonyms' Learning using Wikipedia

At the beginning of our algorithm, we have the following input data:

- $\mathcal{G} = (\mathcal{N}, \mathcal{A})$ is an oriented graph such as \mathcal{N} is the set of nodes and \mathcal{A} is the set of arcs, $\mathcal{N} = \mathcal{C}_O$. Our objective is to extend the initial graph with new nodes and arcs; the former are the hyperonyms and the later are the subsumption links. The extension of \mathcal{C}_i , i is the iteration index, is done by using the concepts' definitions extracted from Wikipedia.
- \mathcal{C}_{gen} is a set of general concepts for which we will not look for hyperonyms. These elements are defined by the domain experts including for example object, element, human being, etc.

S1 For each $c_j \in \mathcal{C}_i$, we check if $c_j \in \mathcal{C}_{gen}$. If it is the case, this concept will be skipped. Else, we look for its definition in Wikipedia. The definition of a given term is always the first sentence of the paragraph before the TOC of the corresponding article. Three cases may occur:

1. The term exists in Wikipedia and its article is accessible. Then we pass to the following step.
2. The concept is so ambiguous that our inquiry leads to the Wikipedia disambiguation page. In this situation, we ignore the word.
3. Finally, the word for which we seek a hyperonym does not exist in the database of Wikipedia. Here again, we skip the element.

S2 For the definition of the given concept, we apply the principle of Hearst's patterns. We attempt to collect exhaustive listing of the key expressions we need. For instance, the definition may contain: is a, refers to, is a form of, consists of, etc. This procedure permits us to retrieve the hyperonym of the concept c_j . The new set of concepts is the input data for the following iteration.

S3 Add into the graph \mathcal{G} the nodes corresponding to the hyperonyms and the arcs that link these nodes.

b. From Graph to Forest

The main idea which shapes the following stage shares a lot with [13]. In fact, the graph which results from the preceding step has two imperfections. The first one is that many concepts are connected to more than one hyperonym. In addition, The structure of the resulting graph is patently cyclic which does not concord with the definition of a hierarchy. An adequate treatment is paramount in order to clean up the graph from circuits as well as multiple subsumption links. Thus, we will obtain, at the end, a forest respecting the structure of a hierarchy.

The following illustrative graph is a piece taken from the whole graph that we obtained during the evaluation of our approach. It represents a part of drilling wells' HSE namely the PPE (Personal Protective Equipment). The green rectangles are the initial candidate concepts.

The resolution of the first raised imperfection implies obviously the resolution of the second one. Therefore, we will use the following solution:

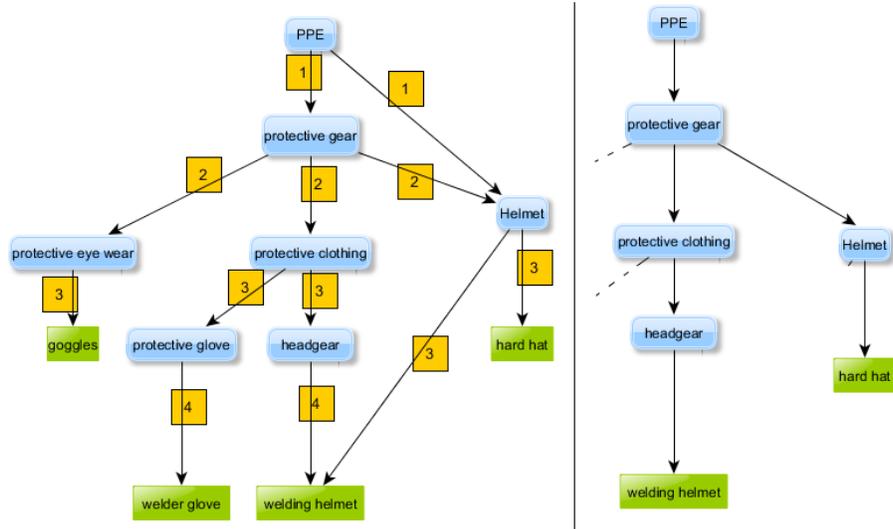


Fig. 3. From wells' drilling HSE graph to forest

1. Weigh the arcs as such as to foster long roads within the graph. We will increment the value assigned to the arc the more we go in depth (it is already done in fig.3).
2. We apply the Kruskal's algorithm[1956] which creates a maximal covering forest from a graph (fig.3).

Finally we have reached the aim we have planned.

4 Our Approach's Evaluation

Our evaluation corpus is a set of texts that are collected in the Algerian/British/Norwegian joint venture Sonatrach / British Petroleum / Statoil. This specialized corpus deals with the field of wells' drilling HSE . Throughout our approach, interventions from the experts are inevitable.

Tex2Tax is the prototype we have developed using Java. Jsoup is the API which allows us to access online Wikipedia. The same result is reached if using JWPL with the encyclopedia's dump. JUNG is the API we have used for the management of our graphs. The following page's figure is the GUI of our prototype.

The terminology extraction phase and the synonyms retrieving have given a collection of 259 domain concepts. The final graph is formed by 516 nodes and 893 arcs. After having done the cleaning, the concepts' forest holds 323 nodes, among them 211 are initial candidate terms. The amount of remaining arcs is of 322. In order to study the taxonomy structure we calculate the compression

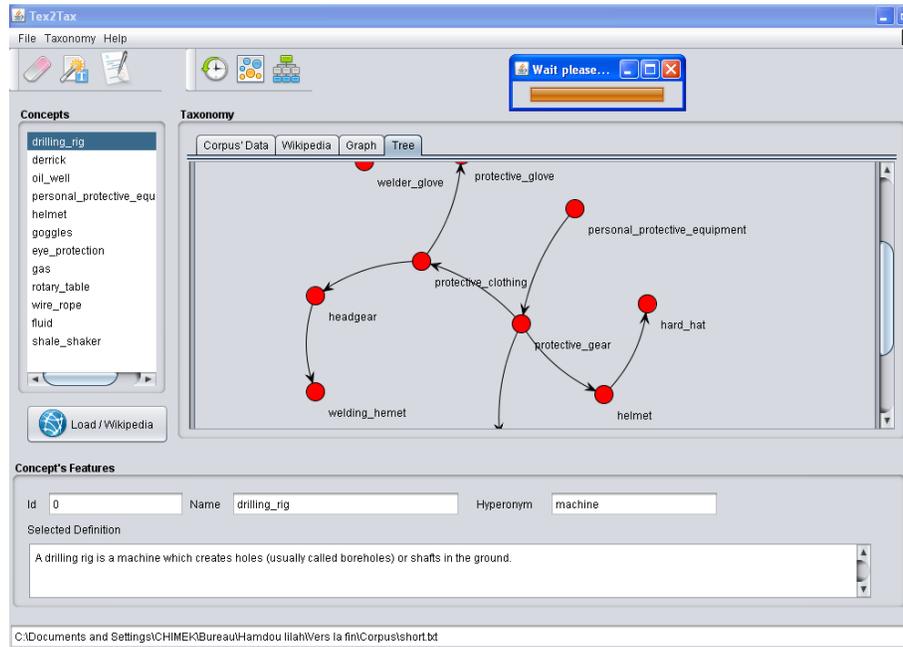


Fig. 4. Tex2Tax prototype's GUI

ratio for the nodes which is $0.63(323 = 516)$ and the one of the arcs which equals to $0.36(322 = 893)$.

$$LP = 0.63(323/516).$$

$$LR = 0.36(322/893).$$

The precision of our taxonomy is relatively low. This phenomenon is mainly due to the terms that do not exist in the database of Wikipedia. The graph's lopping is also responsible of some loss of nodes containing appropriate domain vocabulary.

5 Conclusion

Despite all the work which is done in the field of ontology learning, a lot of cooperation, many contributions and resources are needed to be able to really automate this process. Our approach is one of those few works that harness the collaboratively constructed resource namely Wikipedia. The results achieved and which are based on the exploitation of the idea of Hearst's lexico-syntactic patterns and the graphs' pruning is seen to be very promising. We intend to improve our work by addressing other issues such as enriching the research base

by the Web, exploiting the categories' system of Wikipedia in order to attack higher levels of the ontology learning process such as non-taxonomic relations. Dealing with disambiguation pages of Wikipedia is of great value and multilingual ontology learning is, in addition, an alive research area which is just timidly evoked.

Acknowledgement We are thankful to the Sonatrach / British Petroleum / Statoil joint venture's President and its Business Support Manager for giving us the approval to access the wells' drilling HSE corpus.

References

- [1] Cimiano,P., Mädche, A., Staab, S., and Völker, J. Ontology Learning. In: S. Staab and R. Studer. Handbook on Ontologies. 2nd revised edition. Springer, 2009.
- [2] IJCAI'2001 Workshop on Ontology Learning, Proceedings of the Second Workshop on Ontology Learning OL'2001, Seattle, USA, August 4, 2001. CEUR Workshop Proceedings, 2001.
- [3] Mädche, A. Ontology Learning for the Semantic Web. Kluwer Academic Publishing, 2002.
- [4] Zouaq, A. and Nkambou, R. A Survey of Domain Ontology Engineering: Methods and Tools, In Nkambou, Bourdeau and Mizoguchi (Eds): 'Advances in Intelligent Tutoring Systems', Springer, 2010.
- [5] Zesch, Z., Müller, C., and Gurevych, I. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary . In Proceedings of the Conference on Language Resources and Evaluation (LREC). European Language Resources Association, 2008.
- [6] Ponzetto,S.P.and M.Strube.Knowledge Derived from Wikipedia for Computing Semantic Relatedness. Journal of Artificial Intelligence Research 30, 2007.
- [7] Strube M. et Paolo Ponzetto S. Wikirelate ! computing semantic relatedness using wikipedia. Proceedings of the National Conference on Artificial Intelligence (AAAI), 2006.
- [8]Ponzetto S. P. et StrubeM. Deriving a Large Scale Taxonomy from Wikipedia. AAAI '07, 2007.
- [9] Nastase V. et Strube M.. Decoding Wikipedia Categories for Knowledge Acquisition. AAAI '08, 2008.

[10] Buitelaar, P., Cimiano, P., Magnini, B. Ontology learning from text: An overview. *ontology learning from text: Methods, evaluation and applications. Frontiers in Artificial Intelligence and Applications Series 123*, 2005.

[11] Drouin P., Acquisition automatique des termes : l'utilisation des pivots lexicaux specialises, thse de doctorat, Montral : Universit de Montral, 2002.

[12] Hearst M. A. et Schutze H. Customizing a lexicon to better suit a computational task. *Proceedings of the ACL SIGLEX Workshop on Acquisition of Lexical Knowledge from Text*, 1993.

[13] R. Navigli, P. Velardi, S. Faralli. A Graph-based Algorithm for Inducing Lexical Taxonomies from Scratch. *Proc. of the 22nd International Joint Conference on Artificial Intelligence*, 2011.

Security Ontology for Semantic SCADA

Sahli Nabil, Benmohamed Mohamed

(LIRE) Distributed Computer Science Laboratory
Mentouri Constantine University & SONELGAZ Group
Po.Box 325, Route Ain El Bey 25017 Constantine Alegria
n.sahli@sonelgaz.dz, ben_moh123@yahoo.com

Abstract. Web services have become a significant part of embedded systems as SCADA and internet applications embedded in RTU, because (WS) was XML/SOAP support, independent to platform and very simple to use, these advantages make (WS) vulnerable to many new and old security attacks. Now, it becomes easier to attack (WS) because their semantic data is publicly accessible in UDDI registry and (WS) use http protocol and the 80 TCP port as an open tunneling as a very big vulnerability. We work for the development of better distributed defensive mechanisms for (WS) using semantic distributed (I/F/AV) bloc, security ontology's and WS-Security framework accelerated by ECC mixed coordinates cryptography integrated in our global security solution.

Keywords: SCADA; Web Services (WS); IDS/Firewall/Antivirus (I/F/AV) bloc; ECC Cryptography; Security Ontology.

1 Introduction

The XML Web services open 70% of root for the hackers that firewall and IDS can't detect [2]. Hackers can transport all data with the 80 port, and firewall can't detect this attack [2]. With HTTP protocol Web services can destroy the security strategy the 80 port is always open because it is used by the HTTP protocol used by the web navigators, to create a tunneling, became a very big vulnerability. One of the key challenges to successful of the integration Web services technologies in the embedded system and the SCADA RTU (Remote Terminal Unit) is how to address crosscutting architectural concerns such as policy management and security, governance, authentication, a hacker's attacks, semantic attack and traditional attack.

To address this challenge, this article introduce the notion of semantic attacks in SCADA RTU using the semantic information in the UDDI registry and security concerns lead to the enhancement of SOAP messages via WS-Security framework. In our research, we work to secure the semantic and intelligent Web services embedded in the SCADA RTU, as presented in the figure 1.

We present in this article our approach of accelerating and optimizing security ontology with mixed coordinates ECC cryptography. We begin our article with

presenting SCADA platform used in our research, after that we present security of semantic web services embedded in SCADA RTU, then we present a modified semantic Mitnick attack, after that we present our ontology based semantic distributed (I/F/AV) bloc for SCADA, also we present our solution for optimizing WS-Security framework with mixed coordinates ECC for complex embedded system as SCADA, finally we conclude with a conclusion and our future work and perspectives in our research.

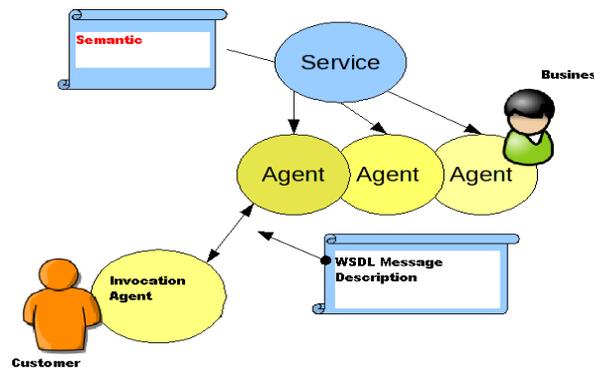


Fig1. Intelligent and semantic Web services embedded in SCADA RTU

2 SCADA Platform Used In Our Research

We use the first IP-based RTU solutions that enable complete integration of SCADA, control, and communications functionality in one rugged package. Our simple yet powerful products leverage easy-to-use Web technologies and inexpensive public networks. They are easy to configure and offer dramatically reduced costs versus traditional SCADA/PLC systems as presented in the figure 2.



Fig 2. Web services and XML technologies embedded in the SCADA RTU [25]

The SCADA RTU integrate, internet compatibility, E-mail messaging, SMS text messaging, Web pages served via the internet or intranets, using FTP file transfer as (CSV, JPEG, etc.), Embedded internet and Web server text messaging, SCADA compatibility with protocols (MODBUS, DNP3, ...etc), SCADA protocol messaging to host computer system, multi communications include (Ethernet, RS-232, RS-485, Fiber optics, GSM/GPRS, PSTN modem, private line modem, and radio) each port operates independently of each other, programmable control, alarm management, data logging and intelligent end device compatibility as (sensors, actuators, digital and intelligent camera, electronic metering devices and process inputs/outputs (fixed and mobile assets as filters, generators, motors, pumps, valves)), as presented in the figure 3.

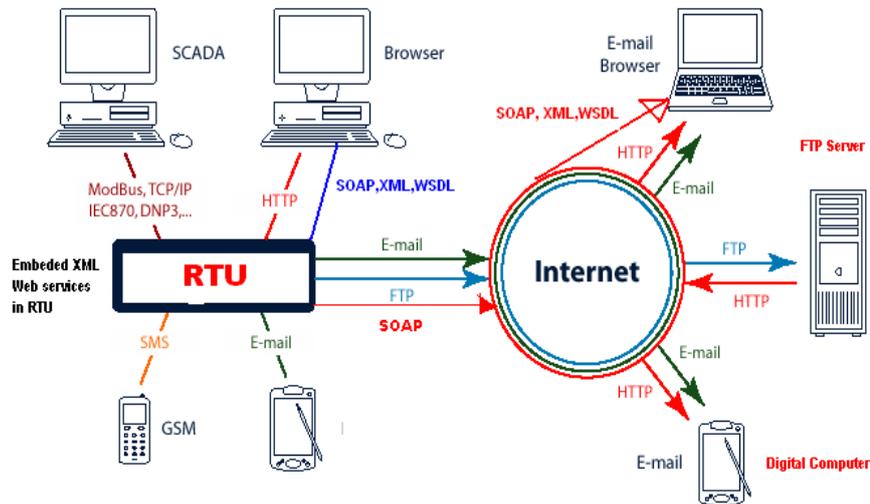


Fig3. SCADA platform and protocols used in our research

For critical applications as SCADA in energy networks security and monitoring, communications redundancy is supported. The RTU SCADA used in Algerian Ministry of Energy and Mining offer an ultra-compact OEM solution, it can be rapidly adapted to many embedded applications and can be connected to the internet for worldwide monitoring, can be served to internet portals regularly or upon events.

3 Security of Semantic Web Services Embedded In SCADA RTU

Semantic (WS) have raised many new unexplored security issues as new ways of exploiting inherit old security threats, semantic (WS), which can publish the information about their functional and non-functional properties, add additional security threats. The hackers do not need to scan the Web and SCADA network to find targets. They just go to UDDI Business Registry in the SCADA control room and get all the information's they need to attack semantic Web services. Now, the whole semantic (WS) attack consists of several stages during which a hacker discovers weakness, then penetrates the semantic (WS) layer and gets access to SCADA critical applications and infrastructures.

For example, the XML Injection attack [7] occurs when user input is passed to the XML stream, it can be stopped by scanning the XML stream. Another type of attacks on (WS) is Denial of Service (DoS) attack when attackers can send extremely

complicated but legal XML documents, it forces the system to create huge objects in a memory and deplete system's free memory. Distributed and multi-phased attacks such as the Mitnick attack [8] are more dangerous for semantic (WS) embedded in the SCADA RTU because IDS [9, 18] can detect them only by acting as a coalition with firewall as a semantic bloc. We need antivirus in the coalition bloc for other kind of vulnerability as distributed and mobile virus. Semantic (WS) embedded in the SCADA RTU are vulnerable at a lot of attacks as: (Application Attacks, Discovery attacks, Semantic Attacks, SOAP Attacks, XML Attacksetc.), as presented in the figure 4, in the following subsections.

The attacker begin by finding Web services using UDDI registry, after that he discover points of weakness in WSDL documents which can be used as a vulnerability guide book for getting access to SCADA RTU critical applications and infrastructures, and create a lot of damages as different kind of semantic Web services attacks: Discovery Attacks [12], WS DoS Attacks [7], CDATA Field Attacks [7], SOAP Attacks [12], Application Attacks [7] [9] [10] [11], XML Attacks [7], Semantic WS Attacks [7]

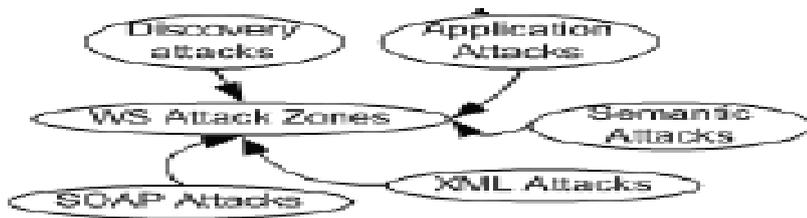


Fig. 4. Attack Zones [12]

4 Modified Semantic Mitnick Attack

The Mitnick attack step is presented in the figure 5 below.

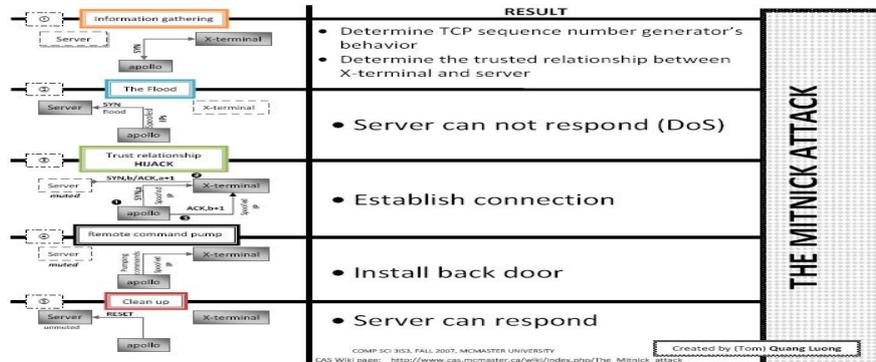


Fig .5. The Mitnick attack Steps [22]

The Mitnick attack can be modified for using in conjunction with the XML Injection attack, semantic (WS) Mitnick attack is organized as follows:

1. An **Attacker** navigates to UDDI registry and asks for a service (Gas temperature) for example.
2. The **Attacker** attaches to UDDI and asks for WSDL files.
3. For blocking communications between **Host1** and **Host2**, **Attacker** starts a Syn/Flood attack against **Host1**.
4. **Attacker** sends multiple TCP packets to **Host2** in order to predict a TCP sequence number generated by **Host2**.
5. **Attacker** pretends to be **Host1** by spoofing **Host1**'s IP address and tries to establish a TCP session between **Host1** and **Host2** by sending a Syn packet to **Host2** (the **Step 1** of a three way handshake).
6. **Host2** responds to **Host1** with a Syn/Ack packet (**Step2** of a three way handshake), however, **Host1** cannot send a RST packet to terminate a connection because of a Syn/Flood (Dos) attack from Step3.
7. **Attacker** cannot see a Syn/Ack packet from Step 6, however, **Attacker** can apply a TCP sequence number from Step4 and **Host1**'s IP address and send a Syn/Ack packet with a predicted number in response to a Syn/Ack packet sent to **Host1** (**Step 3** of a three way handshake).
8. Now, a **Host2** thinks that a TCP session is established with a trusted **Host1**. **Attacker** can attack **Host2** semantic Web services that believe that has a session with **Host2**.
9. **Attacker** inspects **Host2** WSDL files in order to find dangerous methods.
10. **Attacker** tests these methods in order to find possibilities for the XML Injection attack.
11. An **attacker** applies XML Injection for changing Attacker's ID and getting more privileges.
12. If the XML Injection attack is not successful **Attacker** can try the SQL Injection attack or any other injection attacks as XPATH attack or others, against semantic Web services because **Host2** still believes that it is connected to **Host1**.

Our OWL class for the modified Mitnick attack is shown as follows:

```
<owl:Class rdf:ID= '&WSAttacks ;WSMitnick'>
  <owl:intersectionOf rdf:parseType="Collection">
    <owl:Class rdf:about="#Probing"/>
    <owl:Class rdf:about="#WSProbing"/>
    <owl:Class rdf:about="#SynFlood"/>
    <owl:Class rdf:about="#XMLInjection"/>
  </owl:intersectionOf>
</owl:Class>
```

To detect the modified Mitnick attack, the distributed bloc (I/F/AV) installed in the network between Host1 and Host2 should operate as a coalition using the security attack ontology based on distributed (I/F/AV) bloc cooperation, in SCADA systems Host1 must be client and Host2 the RTU.

5 Our Ontology Based Semantic Distributed (I/F/AV) Bloc for SCADA

Using Ontology for creating distributed defenses using IDS [17] is introduced in [8], but, it takes into account only application attacks. A lot of security ontology's of Web services are described in [19], describes types of security information including security mechanisms, objectives, algorithms, credentials and protocols using security ontology's as SWSL[3], WSMO [4], KAoS[5], METOR-S[6], OWL-S [20]. It's applied to SOA to show how Web services can publish their security requirements and capabilities. Security properties and security policies of Web services must be expressed in SCL [14, 15, 16], as automatic reasoning. Our security threats of embedded semantic Web services in SCADA RTU and our proposed defense techniques based distributed semantic (I/F/AV) bloc presented in the figure 6 bellow using VPN Tunneling security technique (VPN1 for ERP and information system and VPN2 for SCADA system), Packet Filtering and Port Filtering.

As shown in the table 1, Web services are generally modeled as resting on top of TCP/IP application protocols such as HTTP. For securing embedded Web services in SCADA RTU we use protocols as (HTTPS, IPSEC, SSL) and other techniques as content filtering and a mixed coordinates ECC encryption with (affines, Montgomery and jacobian) coordinates.

Our Security solution for embedded semantic (WS) uses standards as (OWL/OWL-S) [21, 20], for more detail read [11, 14]. We use WS-Security framework (XML Signature, XML Encryption, WS-Security, WS-SecureConversation equivalent as SSL in SOAP level, WS-Trust, WS-Federation, WS-Policy and WS-SecurityPolicy,

WS-Privacy for management of confidentiality politic with the use of jetton and WS-Authorization) as specified in the figure 7.

Our security solution uses WS-Security framework as presented in the figure 8, our solution include all XML security techniques as transforming, caching, ECC encryption and decryption, auditing, logging, screening and filtering, verification, validation, authentication, authorization, and accounting.

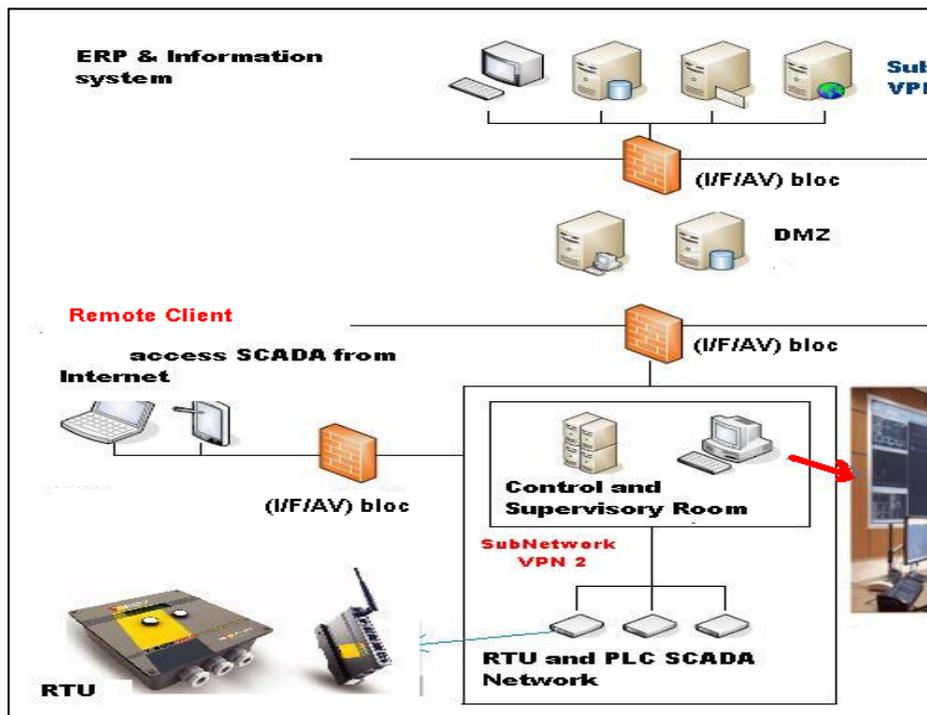


Fig. 6. Our security solution Platform for SCADA

Network Layer	Protocols	Security Technique
Application	HTTP, HTTPS	Content Filtering & ECC Encryption (a mixed of affine, Montgomery and Jacobian coordinates) & SSL Protocol
Transport	TCP, UDP	Port Filtering
Inter network	IP, ICMP	Packet Filtering & IPV6
Data Link	PPTP, L2TP	VPN Tunneling (VPN1 & VPN2)

Table 1. Security techniques proposed for SCADA systems

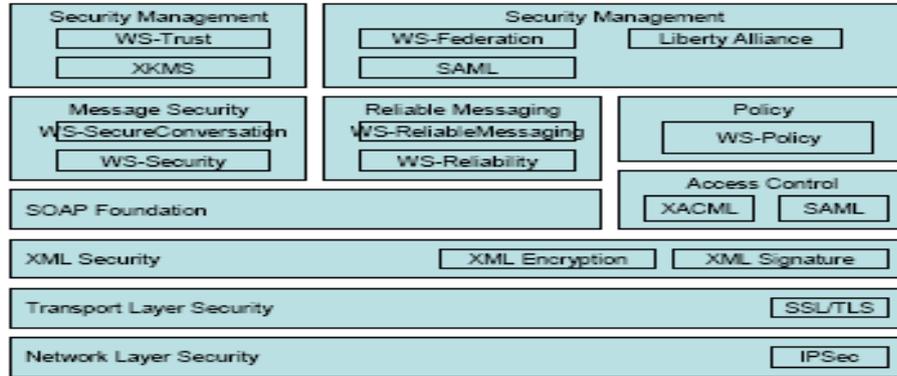


Fig.7. WS-Security framework stack [1] [13]

Our solution use ten (10) steps as : message signature operation, message crypt operation, associating a jetton in the SOAP message (steps : 1,5) and the SOAP message preparation (step 4) in distance customer, and SOAP message transmission (step 7), validation operations , decrypting SOAP messages and to certificate them (steps: 8,9,10) in SCADA RTU, also the Service Registry, Policy Store and Identity Provider (steps :2,3,6) , as presented in the figure 9.

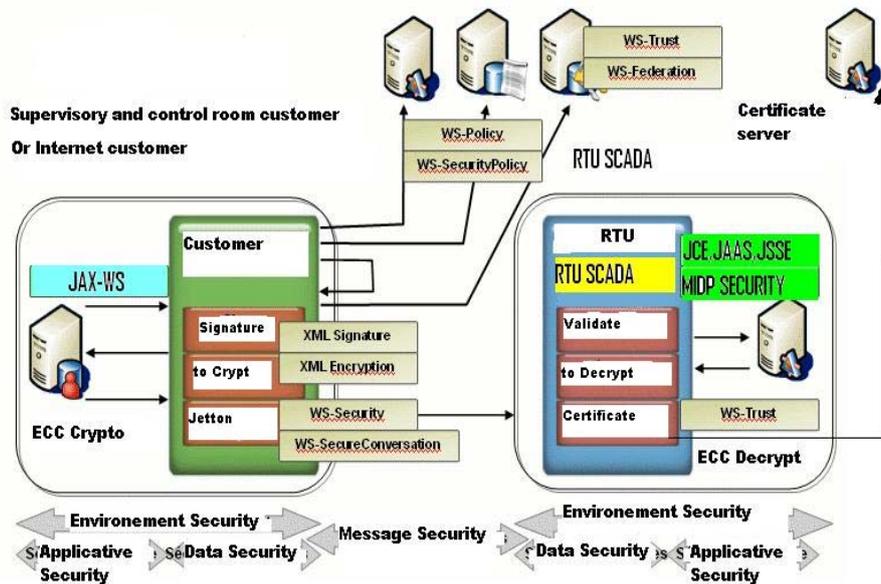


Fig. 8. Our Security solution for SCADA

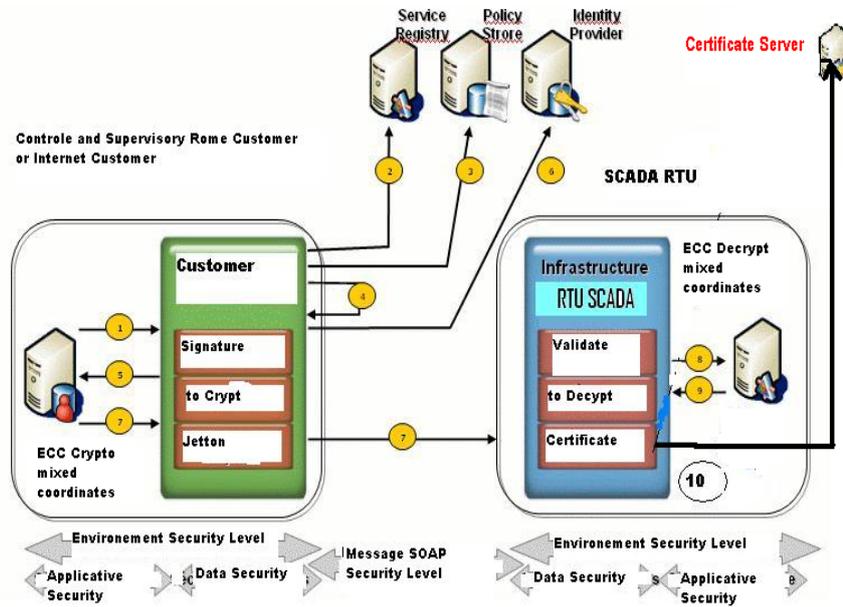


Fig.9. The Ten (10) steps of our security solution for SCADA

Our solution includes a lot of security levels as (applicative security, data security, environment security and message SOAP security). We present in the figure 10 and 11 our SOAP message security proposed solution.

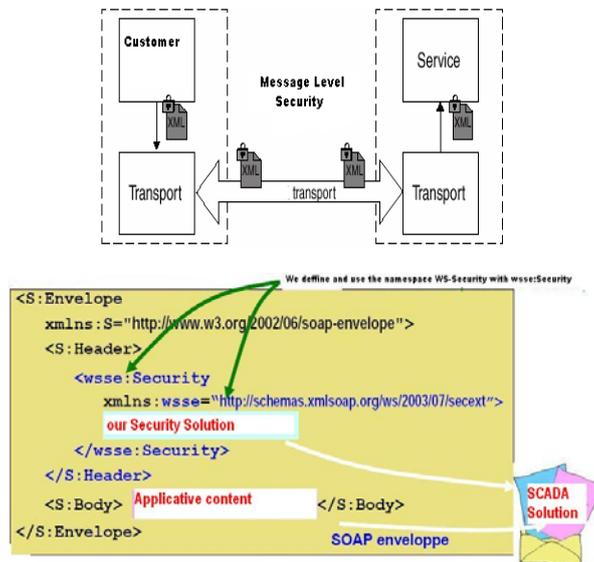


Fig.10. A SOAP security message solution Fig.11. SOAP message security implemented in RTU

6 Optimizing WS-Security Framework with Mixed Coordinates ECC

Elliptic curve cryptography (ECC), independently introduced by Koblitz and Miller in the 80's [27], has attracted increasing attention in recent years due to its shorter key length requirement in comparison with other public-key cryptosystems such as RSA. Shorter key length means reduced power consumption and computing effort, and less storage requirement, factors that are fundamental for SCADA systems as presented in the figure 12. Comparing (WS) secured by WS-Security framework to unsecured Web services, the WS-Security is by factor 100 slower than Web services. WS-Security should be used only where security has the highest priority over performance, but it is not the case of the embedded complex system as SCADA system and embedded Web services in the SCADA RTU. Our approach is to optimize WS-Security framework by using our solution based mixed coordinates ECC [24] for the operations (to crypt, to decrypt, to sign and to verify signature) SOAP messages as presented in our solution figures 8 and 9.

Algorithmne	Size of Key (bits)	Signature		Vérification	
		Time (s)	Energy consumption (mW)	(s)	Energy consumption (mW)
RSA	1024	22.03	726.99	0.86	28.38
RSA	2048	166.85	5506.05	3.89	128.37
ECC	160	1.65	54.45	3.27	107.91
ECC	224	4.46	174.18	8.84	291.72

Fig. 12. ECC and RSA comparative [26]

Our analyze in the database « **Explicit-Formulas Database** » [23] determine the result shown below in Figure 13 and 14.

Coordinats	Addition	Doubling	Mixed Addition
Modifiees	13M+6S	4M+4S	-
Brier & Joye	9M+2S	6M+3S	-
Montgomery	4M+2S	3M+2S	-
Affines	I+2M+S	I+2M+2S	-
Projectives	12M+2S	7M+5S	9M+2S
Jacobiennes	12M+4S	4M+6S	7M+4S
Chudnovsky	11M+3S	5M+6S	-

Fig. 13. Cost of ECC coordinates in the field F_p

Coordinats	Addition	Doubling	Mixed Addition
Affines	I+M	I+M	-
Projectif (c=1, d=1)	13M	7M	12M
Jacobien	14M	5M	10M
Lopez-Dahab	14M	4M	8M

Fig.14. Cost of ECC coordinates in the field $F2^m$ (**M**: Multiplication, **S**: square, **I**: Inversion)

Our ECC optimized algorithm « **Mixed-Coordinates-ECC-Algo** » is presented below:

1. We compute the doubling operations with « Montgomery » coordinates for preparing the addition operation in the field Fp and with « Affines » coordinates for the field $F2^m$.
2. We compute the addition of the last point computed and another point in the curve, with « Affines » coordinates, for the two fields Fp and $F2^m$.
3. All addition operation will be computed with « Affines » coordinates for the two fields Fp and $F2^m$.
4. All mixed addition operation will be computed by « Lopez-Dahab » coordinate for the field $F2^m$ and « Jacobiennes » coordinates for the field Fp .

7 Conclusion

The SCADA RTU including embedded Web services and embedded XML creates a new big challenge in security for SCADA, because network security is maturing and semantic embedded Web services security not mature. Specific procedures for securing embedded XML SCADA network applications are not yet widely known. We present our security solution introduced in this paper for embedded SOA security design, with a distributed implementation of a distributed semantic bloc (I/F/AV) between client and RTU. Our approaches is composed with ten (10) steps using ECC mixed coordinates cryptography solution and WS-Security framework, adapted and optimized for SCADA systems. We use a bloc of products such XML semantic firewalls, proxies, IDS, gateways, VPN technologies, security protocols (HTTPS, IPSEC, and SSL), a security framework as WS-Security and ECC mixed coordinates cryptography integrated in our solution. We propose a security solution of semantic Web services embedded in RTU using security ontology's as OWL/OWL-S. We work to do more optimization and to implement our solution with a real material used in Algerian ministry of energy and mining as TBOX RTU manufactured by CSE-Semaphore Group Company [25] and TOSSIM (PowerTossim & TinyViz) simulator [28].

References

1. Aymen BOUGHATTAS & Med Aymen BAOUAB, Web service security (WS-Security), a master degree memory 2008/2009 Nancy Franch University, 2009.
2. Soaj2ee.blogspot.com/files/whitepaper/soaj2ee-security-transport.pdf
3. SWSL, <http://www.daml.org/services/swsl/>
4. WSMO, <http://www.wsmo.org/>
5. KAoS, <http://www.ihmc.us/research/projects/KAoS/>
6. METEOR-S, <http://lsdis.cs.uga.edu/projects/meteor-s/>
7. A.Stamos and S.Stender, "Attacking Web Services: The Next Generation of Vulnerable Enterprise Apps, BlackHat2005, USA, 2005.
8. J.Undercoffer, A.Joshi, T.Finin, and J.Pinkston, A target-centric ontology for intrusion detection, Int. Joint Conference on Artificial Intelligence, Mexico, 2004.
9. J.Mirkovic, "D-WARD: Source-End Defence Against Distributed Denial-of-Services Attacks", The Phd thesis, University of California, 2003.
10. P.Lindstrom, "Attacking and Defending Web Services", A Spire Research Report, January 2004.
11. W.Negm, "Anatomy of a Web Services Attack: A Guide to Threats and Preventive Countermeasures", 2004.
12. S.Faut, "SOAP Web Services Attacks: Are you web applications vulnerable", SPI Dynamics, 2003.
13. T.Erl, "WS-* Specifications, An Overview of the WS-Security Framework", 2004.
14. K.Khan and J.Han, "A Security Characterisation Framework for Trustworthy Component Based Software Systems", COMPSAC2003, USA, 2003.
15. A.Vorobiev and J.Han, "Specifying Dynamic Security Properties of Web Service Based Systems", SKG2006, Guilin, China, 2006.
16. K.Khan, "Security Characterisation and Compositional Analysis for Component-based Software Systems", PHD thesis, Monash University, April 2005.
17. S. Axelsson, "Research in Intrusion-Detection Systems: A Survey, Technical report 98-17, Chalmers University of Technology, 1998
18. G. Denker, S.Nguyen, and A.Ton, OWL-S Semantics of Security Web Services: a Case Study, SRI Internayional, Menlo Park, California, USA, 2004
19. A.Kim, J.Luo, and M.Kang, Security Ontology for Annotating Ressources, ODBASE 2005, Cyprus, 2005.
20. OWL-S: Semantic Markup for Web Services, November 2004, <http://www.w3.org/Submission/OWL-S/>
21. OWL, <http://w3.org/TR/owl-features/>
22. http://wiki.cas.mcmaster.ca/index.php/The_Mitnick_attack
23. <http://www.hyperelliptic.org/EFD/> Explicit-Formulas Database
24. H. Cohen, A. Miyaji, and T. Ono. Efficient elliptic curve exponentiation using mixed coordinates. In ASIACRYPT, LNCS. Springer, 1998.
25. CSE-Global group company, europe Belgium www.CSE-Semaphore.com
26. A.Patel, Arvinderpal Wander, Hans Ebele, Sheulling C hang Shantz, comparing elliptic curve cryptography and RSA on 8-Bit CPUs, 2004.

27. N. Koblitz. A Family of Jacobians Suitable for Discrete Log Cryptosystems. In Shafi Goldwasser, editor, *Advances in Cryptology - Crypto '88*, volume 403 of *Lecture Notes in Computer Science*, pages 94 – 99, Berlin, 1988.
28. Victor Shnayder, Mark Hempstead, Borrong Chen, Geoff Werner Allen, and Matt Welsh, *Simulating the Power Consumption of LargeScale Sensor Network Applications*, Harvard University, Baltimore, Maryland, USA, *SenSys'04*, November 3–5, 2004.

Automatic construction of ontology from Arabic texts

Ahmed Cherif Mazari¹, Hassina Aliane², and Zaia Alimazighi³

¹Electrical Engineering and Computer science Department, University of Médéa
mazari.ac@gmail.com

²CERIST, Research Center on Scientific and technical Information, Algiers.
haliane@mail.cerist.dz

³Computer Science Department, USTHB, Algiers.
alimazighi@wissal.dz

Abstract. The work which will be presented in this paper is related to the building of an ontology of domain for the Arabic linguistics. We propose an approach of automatic construction that is using statistical techniques to extract elements of ontology from Arabic texts. Among these techniques we use two; the first is the “repeated segment” to identify the relevant terms that denote the concepts associated with the domain and the second is the “co-occurrence” to link these new concepts extracted to the ontology by hierarchical or non-hierarchical relations. The processing is done on a corpus of Arabic texts formed and prepared in advance.

Keywords: Ontology, Information Extraction (IE), Arabic Natural Language Processing (Arabic-NLP), Statistical methods for text processing.

1 Introduction

Existing methods of ontologies construction differ mainly according to the information that they treat (concepts, relations, properties ...) and techniques for extracting these elements from texts. These techniques are carried out either by methods that require linguistic corpus annotated or by statistical methods that do not need the annotation text. In our approach, we are oriented toward the use statistical methods, since these methods do not require these types of annotated corpora and NLP¹ analyzers (such as the lexical analyzer and parser). These methods are based on two criteria: the relevance of a term from a domain that is defined by the number of occurrences of the word in the corpus and the co-occurrence of two terms at a frequency more high.

2 Overview of the Approach

In our approach, we started the initialization of the ontology manually, by the general (generic) concepts retrieved from the ontology of GOLD (General Ontology for Linguistic Description) [Far03], it is a general ontology for descriptive linguistics and is applicable to most human languages. It was created on the base of the general

¹ NLP: Natural Processing Language.

ontology of SUMO² (the Standard Upper Merged Ontology). Then, we adopted the process of extraction from the domain text which can be summarized in three main steps; the first is the formation of the domain corpus, this step is fundamental since the quality of the corpus will depend on the quality of processing and the corpus must fully cover the domain treated. The second step is the extraction of candidate terms (*these terms may be among the elements that make up the ontology: a concept, a relation or an individual*). Finally, we make the junction of these new elements to the ontology.

2.1 Constitution and preparation of the corpus

In a project of construction ontologies from texts, the corpus, its status and its collection are of paramount importance both as a source of knowledge to build the model and also a source of reference throughout the process development [BoA03]. So the questions addressed in the constitution of the corpus include: the type of corpus (a corpus "specialized" is a corpus containing texts on a topic related to a domain of knowledge as our case Arabic linguistics), and the suitability for the project referred (the quality of the results of a corpus largely is depending on the quality of the corpus, this means, that the domain texts are well defined and delimited, they are fairly representative). However, size is often limited by the availability of texts and issues of copyright). Representativeness (variety of texts, authors, sources, etc) and using full-texts or samples. [Mar03]

Preparation of corpus. After the formation of crude corpus, it must be prepared for processing. This phase is performed by a set of preprocessing steps to remove some ambiguity, reduce the number of transactions and adapt the corpus following the final objective "extraction of candidate terms".

Normalization. In the corpus, we will encounter elements that do not carry information and increase the processing time. This is mostly special characters, numbers, non-Arabic words, abbreviations and single letters. These should be deleted:

- Special characters: include any special sequence of characters delimited by letters or spaces.
- Numbers: We regroup all the character sequences located between two spaces containing numbers in a single occurrence. This method also has the advantage to combine the dates, the actual numbers and percentages.
- Words in Latin characters: The non-Arabic words, mainly in Latin characters are simply detected by their graphic.
- Abbreviations and isolated letters: The list of words to a single letter in the Arabic texts reveals the presence of a significant number of these words. These letters are often used in abbreviations. It may designate a variable, for example « ب الفنة », « category B », numbering ; « الفقرة أ » section A, ت for « تاريخ » date, م for « ميلادي », « category B », numbering ; « الفقرة أ » section A, ت for « تاريخ » date, م for « ميلادي »

² <http://suo.ieee.org> developed in the project *IEEE SUO* Working Group.

ص for «page». We can find also letters that form a grammatical category for example [AbD08]. حروف العلة: (ا، و، ي)

- Character 'ـ': The typographers make frequent use of the character 'ـ', allowing the extension of the line in the middle of words, for better readability, to limit the white space on a line justified, even for purely aesthetic reasons. This character is not part of the Arabic alphabet. It is therefore necessary to eliminate it.
- To remove the vowel signs, which are written in the form of diacritics placed above or below letters.
- Because of graphs variations that may exist when writing the same word and that they can be sources of ambiguity. We will make some substitutions as follows: Substituting letters ا, آ, and أ by ا. Substituting of end letters ي, ة, by ي. [Dou05]

Deletion of Stop-Words. These are grammatical or lexical words; they are so often grouped together in a "stop-list." It is generally accepted that these words very common (about half of the occurrences of a text) are not indexed because they are not informative [Ver04]. It is a list with all the words of tools, connection and articulation (pronouns, articles, conjunctions, prepositions, etc.). (Example: في، ان، على، التي، عن، .. الذي، مع، في، بعد، بين، هذه، هذا، انه، منذ، ما، لم).

Light stemming. Using words as linguistic unity is possible, but also raises a number of problems of ambiguity in the morphological analysis, the fact that Arabic (unlike the Latin languages) is an inflected language, and strongly differentiable agglutinative, articles, prepositions and pronouns stick to adjectives, nouns, verbs. To resolve the ambiguity [Bou05] showed that stemming is a very useful preprocessing, which involves finding the root of each word. It makes a deletion of prefix and suffix to identify the root word. These suffixes and prefixes are grouped in a dictionary. Since most of the Arabic words have a root with three or four letters, keeping the word at least three letters will allow us to preserve the integrity of sense. So we conducted light stemming by identifying prefixes and suffixes that were added to the word. We use the list of prefixes and suffixes proposed by [Dar03], it was determined by a frequency calculation on a corpus of Arabic articles. This list includes prefixes and suffixes commonly used in the Arabic language such as conjunctions, verbal prefixes, possessive pronouns, pronouns name or verbal suffixes expressing the plural and so on.

Table 1. Prefixes and suffixes list.

Prefixes						
والـ	بـ	وتـ	بـ	كـ	لـ	فـ
فـ	يـ	سـ	لـ	فـ	لـ	وا
بـ	مـ	نـ	ومـ	الـ	ويـ	فا
Suffixes						
اتـ	وهـ	تهـ	همـ	نا	ينـ	هـ
وا	انـ	تمـ	هنـ	تـكـ	يهـ	سيـ
تيـ	كمـ	ها				ونـ

2.2 Automatic extraction of “candidate terms”

After preparing the corpus, we move to the extraction step of ontology elements. The processing is done in two passages. In the first; we will extract all the terms (one or more words) used to denote concepts in the domain, using the method of “repeated segments” based on the following prepositions: *A significant term is used several times in a specialized text.*

- Terms can be complex, that are composed of several words used individually (ex. جملة اسمية).
- Complex terms are constructed using a finite number of sequences of words.

In the second passage; we will seek the pairs of terms that co-occur more frequently in the corpus. The result of this processing provides us with a list of pairs of terms that will be used to update the ontology. Therefore, the objective of the first pass is to identify the terms that denote the concepts related to the domain, however the second pass is to identify among these terms, couples who have links with elements of the ontology.

Applying the method of “repeated segments”. It is a statistical technique for extracting information from texts unlabelled. The repetition of these segments indicates that these can be used to denote concepts of domain of the corpus. A text segment consists of one or more words and delimiters are punctuation marks or spaces. The method performs an index of all words in the text by assigning a code corresponding to their positions in the corpus. Then it identifies of all repeated segments in a window of four words (number of four is chosen on the principle that a term denoting a concept contains a maximum of four words) in limiting itself to the same sentence. During this phase, redundancies are eliminated by removing the segments included in others with the same number of occurrences. At this step a large number of segments are extracted, some of which are incorrect. All of these segments are then filtered to remove unwanted segments and retain only those who are selected as candidate terms. In our approach, we use two filters; filter of weights [Her06] and a cutting filter³. The weighting filter is used to select terms with enough weight with respect to this weighting; it is a global threshold and fixed indicating the relevance (a relevant term is used several times in a specialized text). The weight is measured by the total frequency of a term; it is the total number of occurrences of the word in the corpus. If this *frequency* exceeds a *global threshold*, then the term is part of the domain.

The “cut filter” removes the segments containing certain words such as verbs, named entities, numbers into letters or other. The words of "cut filter" may be present at the beginning, the end and within the segment. The list of words of the filter can be easily adapted and expanded by the user depending on the specifics of the corpus treated. The words of the "cut filter" cannot be present in a segment after application of this filter.

³ Used in the MANTEX (it is a system of terminology extraction from texts unlabelled. [RoF02])

Applying the method of “co-occurrence”. The technique is based on the extraction of binary cooccurents or pairs of terms that meet one of the other more frequently than by chance and these two terms were included in the list found in the previous phase (phase detection of repeated segments). The method starts by identifying cooccurents of a given term in a window of fixed size (example ten words) and in the same sentence, examining the cooccurents relative to the target term. The method measures the attraction in pairs (the terms in some order) and not in pairs. Pair {جملة, اسم} corresponds to two pairs < اسم , جملة > (جملة is the first term and اسم appears to the left in the text) et < جملة , اسم > (This time it is جملة than appears in the left). Finally, we will select the cooccurents with a frequency exceeding a statistically significant frequency due to chance. A numerical threshold of 80%⁴ is defined a priori to estimate a relation between two terms is significant.

2.3 Update of the ontology

The principle of the approach is to compare the pair of candidate terms extracted (<t1,t2>) with the labels of the ontology concepts, we find four possible cases; t1 (t2) belongs to the labels of ontology and t2 (t1) is not, t1 and t2 are in the same time labels of the ontology, t1 and t2 or not belong to the labels of the ontology.

Relation by linguistic marker. To identify relations between terms, we will study the context surrounding these terms in a small window (eg, four words) [Koo03]. From this context the method will look for lexico-syntactic elements for identifying a relation between them. These elements are called linguistic markers⁵.

Example « T1 is-a T2 », « T1 part-of T2 », ...

But as the same relation can be expressed by different markers so they are organized into categories or separate lists depending on the type of relation to be extracted, which will be incremented progressively.

Thus we have in each list (or category), a kind of paradigm of linguistic units which are sometimes heterogeneous categories (nouns, verbs, function words or grammatical, etc.). But always it fulfills the same functions for the relation type.

- Hyponymy or Generalization relation « is-a » : list = { ... هم ، هي }
- Meronymy relation «part-of » : list= { تتألف-من ، تنقسم-الى ، تتكون-من }

Accordingly to the specific morphology of Arabic at the vocalization and agglutination, the list of markers should be clustered all forms and other morphological variants likely to be encountered in the texts. We can add new relations and to update the lists of pre-existing relations. The process of updating the ontology is as follows:

- If one term of the pair is found among the labels of the ontology concepts, the second term of the pair will be proposed for a new concept in the ontology and will be linked to the first concept for a relation defined by linguistic marker.

⁴ The numerical threshold used in the "Xtract" extractor is 80%. [Sma93]

⁵ CAMELEON is a software research of lexical relations from linguistic markers. [Ség01]

- If both terms are among the labels of the ontology concepts and there was no relation between these two concepts, a new relation will be proposed from marker linguistic.
- In case where neither the first nor the second term do not belong to the ontology labels. The process does nothing and let these cases for future running.

Hierarchical relation. If the linguistic markers are absent in the context of words, the approach based on a parent-child relation where the parent term is more general than the child term. This relation between terms is extracted from the asymmetric co-occurrence of terms. The relation is characterized by the following two rules: $P(x/y) \geq 0.8$ and $P(y/x) < P(x/y)$; $P(x/y)$ is the probability of term 'x' occurrence then the term 'y', inversely for $P(y/x)$ [HeM06]. First rule ensures that both terms appear together enough (ie 80% of cases). According to the second rule, x subsumes y where the probability of occurrence of x before y is upper than the reverse. Using the transitive property of the relation we can eliminate some relations, e.g. if the relation "a" subsumes "b", "a" subsumes "c" and "b" subsumes "c" are extracted, the relation "a" subsumes "c" can be deleted because it is deductible from the other two [Her06]. However, the process of updating the ontology is as follows:

- If the first term (or second) is found among the labels of the ontology concepts and the second (or first) term of the couple is not, then it will be proposed a new *son-concept* (*father-concept*) related to the first (second) concept by subsumption relation "is-a".
- In the case where both terms are among the labels of the ontology concepts and there was no relation between these two concepts, a new relation of subsumption "is-a" will be proposed.
- In case where neither the first nor the second term do not belong to the ontology labels. The process does nothing and let these cases for future running.

3. Experimentation and results

We were able to test the approach using the Python programming language, due to its power and through its NLTK⁶ (Natural Language Toolkit) library.

3.1 Constitution of corpus

We selected a sample of texts from documents written in Arabic sought in the following resources: books on Arabic linguistics, and journal articles (N°7 and N°8 of AL-LISANIYYAT) published by the CRSTDLA⁷ in Arabic language and through

⁶ http://nltk.sourceforge.net/index.php/Main_Page

⁷ Center for Scientific Research and Technical Development of Arabic Language (Algiers)

Deletion of stop words (2). We need to eliminate stop words again, since in the results of light stemming we found these words again after deleting some of the prefixes and suffixes: Example (following cases are present: بعده-بعد ، اخرى-ال اخرى)

Result. 261 715 words are found and 39 207 words are removed (13%).

3.3 Processing

Extraction of “repeated-segments”. We set the following parameters:

- Segment size = 4 words. It indicates the maximum size of a complex term, usually a complex term in Arabic is made up of 4 words.
- Weighting threshold: The weight of a term is calculated by the total frequency, is the total number of occurrences in the corpus. Threshold weight of a simple word is = 100. Threshold weight of a compound term is = 20. The number 100 and 20 are randomly selected relatively to the corpus size.

Result. The program extracts 281 200 different segments, but it only selects a list of 445 segments in accordance with the thresholds defined above. In analyzing this list, we have identified the following comments:

1. Words appear that are outside domain (personal names, object names ...). We can update the list of stop words by these words and to redo processing.
2. Two morphological forms of same word are identified as two different segments. Example (عناصر ، عناصر) (حرف ، حروف) (لغه ، لغوى، لغات ، للغه). We can regroup the different morphological forms in the same form then replace them in the corpus and repeat the processing.

The following table shows a sample of selected segments:

Table 3. Sample of selected segments.

Segment	Frequency	Segment	Frequency	Segment	Frequency
لغه	5071	فاعل	592	مفعول مطلق	84
فعل	2449	ظاهر	579	جمل اسم	83
اسم	1938	ضمير	575	علام رفع ضمه	78
...

Extraction of “co-occurents”. We set the following parameters:

- Window size of co-occurrence = 10 words.
- Co-occurrence threshold = 80% (percentage of appearance two terms together).
- Co-frequency threshold = 100 (number of appearance two terms together).

The program gives the result in a marked file where each line contains the co-occurring, their frequency and their co-frequency. As the following example:

```
< t1="نصب" t2="فتح" Ft1="672" Ft2="129" CF="211"/>
< t1="اسم" t2="فعل" Ft1="1938" Ft2="2449" CF="210"/>
```

Suggestion. This result file must be validated by an expert (a linguist).

4. Conclusion

In this paper, we have shown an approach for the automatic construction of ontology from a corpus of domain "Arabic linguistics". We reused information extraction techniques for extracting new terms that will denote elements of the ontology (concept, relation). To analyze the texts of the corpus, two statistical methods were used, the "repeated segments" to identify the candidate terms and "co-occurrence" to the updating of ontology. So, we have formed a domain corpus by the recovery of text from articles of journals and books of the domain and also the collection of documents over the Web. This corpus was preprocessed to remove some ambiguity, reduce the number of transactions and adapt the corpus according to our aim.

Many perspectives are offered based on our work, among them; we proposed an ontology that represents the fundamentals notions of Arabic linguistics, this ontology can be useful for developing NLP tools that analyze Arabic texts. A second perspective would be to use our techniques and statistical methods for information extraction on Arabic texts for other works (e.g. terminology extraction, creation of electronic dictionaries and thesaurus ...).

References

- [AbD08] Ramzi Abbès, Joseph Dichy « Extraction automatique de fréquences lexicales en arabe » JADT 2008 :« 9^{ème} Journées internationales d'Analyse statistique des Données Textuelles » Université Lumière Lyon 2, ICAR-CNRS.
- [BoA03] Didier Bourigault et Nathalie Aussenac-Gilles. «Construction d'ontologies à partir de textes ». conférence sur le traitement automatique des langues (TALN), France, Juin 2003.
- [Dar03] Darwish K « *Probabilistic methods for searching OCR-Degraded Arabic Text* » Thèse de Doctorat Université de Maryland 2003.
- [Dou05]. F. S. Douzidia, G. Lapalme « *Un système de résumé de texte en arabe* » université de Montréal exposé en deuxième conférence International de "l'Ingénierie de la Langue et Ingénierie de l'Arabe " Alger 2005.
- [Far03] : Farrar, William D. Lewis, and D. Terence « *An Ontology for Linguistic Annotation* » Department of Linguistics, University of Arizona 2003.
- [HeM06] N. Hernandez, J. Mothe « *TtoO: une méthodologie de construction d'ontologie de domaine à partir d'un thésaurus et d'un corpus de référence* » IRIT, Toulouse, 2006.
- [Her06] Nathalie HERNANDEZ « *Ontologies de domaine pour la modélisation du contexte en recherche d'information* » Thèse de Doctorat à l'Université Paul Sabatier France 2006.
- [Koo03] S. Koo, S.Y. Lim, S.J. Lee, « *Building an Ontology based on Hub Words for Informational Retrieval* », the IEEE/WIC International Conference on Web Intelligence, 2003.

- [Mar03] Elizabeth Marshman «Construction et gestion des corpus : Résumé et essai d'uniformisation du processus pour la terminologie » Janvier 2003, "Observatoire de linguistique Sens-Texte" (OLST) de l'Université de Montréal.
- [RoF02] F. Rousselot et P. Frath, « Terminologie et Intelligence Artificielle » (12^{èmes} rencontres linguistiques), Presses Universitaires de Caen, 2002.
- [Ség01] Patrick Séguéla « Construction de modèles de connaissances par analyse linguistique de relations lexicales dans les documents techniques » thèse TOULOUSE III. 2001.
- [Sma93] Frank. Smadja, « Retrieving collocations from text: Xtract, Computational Linguistics », université de Columbia 1993.
- [Ver04] Jacques Vergne « Découverte locale des mots vides dans des corpus bruts de langues inconnues, sans aucune ressource » JADT 2004 :« 7^{ème} Journées internationales d'Analyse statistique des Données Textuelles » GREYC – Université de Caen.

Model driven approach for specifying WSMO ontology

Djamel Amar Bensaber¹, Mimoun Malki¹

¹EEDIS laboratory, University of Sidi Bel Abbes, Algeria
{amarbensaber@yahoo.com, mymalki@gmail.com}

Abstract. The semantic web promises to bring automation to the areas of web service discovery, composition and invocation. In order to realize these benefits, rich semantic descriptions of web services must be created by the software developer. A steep learning curve and lack of tool support for developing such descriptions thus far have created significant adoption barriers for semantic web service technologies. In this paper, we present a model-driven architecture approach for specifying semantic web service through the use of a UML profile that extends class diagrams. In this paper we describe our efforts to develop a transformation approach based MDA to translate XMI specifications (e.g., XML encodings of UML) into equivalent WSMO specifications via the use of ATL transformations.

Keywords: Model driven Architecture (MDA), WSMO, ATL, Metamodel.

1 Introduction

The potential to achieve dynamic, scalable and cost-effective infrastructure for electronic transactions in business and public administration has driven recent research efforts towards so-called Semantic Web services, that is enriching Web services with machine-processable semantics. As a matter of fact, describing Web services through aforementioned submissions are not easy for service developers to write. Although, several tools and editors such as OWL-S Editors, WSMO studio [1], and WSMOViz [2] have been proposed to facilitate writing Semantic description, developers still need to know the concepts and syntaxes of the Semantic Web service languages. This lack of knowledge and also the complexity of these languages cause the adoption of Semantic Web services slow down [3].

In order to tackle this problem, several approaches have been proposed based on Model driven Architecture (MDA) [4] for automatically generating semantic web service descriptions from a set of graphical models. MDA is an approach presented by OMG for developing application system in the way of creating model rather than code. The portability, interoperability, and reusability are primary goals of MDA, which are acquired via separation of concerns between the implementation and specification. In most of MDA-based approaches, Unified Modeling Language (UML) [5] is used as modeling language due to its widespread adaption among software developers [6].

In this context, we are developing an approach that allows a developer to focus on creation of semantic web services and associated WSMO [7] specifications via the

development of a standard UML model. We describe our efforts to develop a transformation model for translating UML specifications into equivalent WSMO specifications. The approach relies upon the use of MDA concepts by developing two meta-models (source and target one) and a transformation model to translate XMI specifications (e.g., XML encodings of UML) into WSMO via the use of ATL transformations [8]. By using transformations from equivalent UML constructs, difficulties caused by a steep learning curve for WSMO can be mitigated with a language that has a wide user base, thus facilitating adoption of semantic web approaches.

The remainder of this paper is organized as follows. Section 2 describes the related works for semantic web services approaches, a WSMO overview is presented in section 3. The specifics of our approach, details and the main parts of our solution are presented in Section 4. Sections 5 and 6 discuss implementation and conclusions, respectively.

2 Related works

In this section we present briefly various approaches which allow to use UML for the creation of ontologies. Gasevic [9] suggests using an UML profile for ontology as well as the standards of the OMG concerning the approach MDA. By this method he wishes to insure the generation automatic of complete ontologies (in OWL [10]) by using transformations of models. The approach of Gasevic and his colleagues relays on the principles of MDA and transformation of models. For it they defined an UML profile named OUP (Ontology UML Profile) which takes back the concepts of ontologies such as they are defined in OWL. Their second contribution is to supply bidirectional transformations between it profile UML and the ODM (Ontology Definition Metamodel) metamodel, proposed by the OMG. Their last contribution holds in transformations between ODM and the languages of ontology such OWL.

Brambila et al in [11] present a model-driven methodology to design and develop WSMO-based Semantic Web services using Business Process Model and Notation (BPMN) [12] in conjunction with Web Modeling Language (WebML) [13].

MIDAS-S [14] is based on the expansion of MIDAS [15] which is a model driven methodology to develop Web Information System (WIS). This approach present a methodology to develop semantic Web service based on WSMO. The four top-level elements such as ontologies, goals, mediators, and Web services are formed in the PSM level.

3 WSMO Overview

The WSMO initiative aims at providing an overarching framework for handling Semantic Web services (SWSs). WSMO identifies four main top-level elements:

1. Ontologies that provide the terminology used by other elements;
2. Goals that state the intentions that should be solved by Web Services;
3. Web Services descriptions which describe various aspects of a service;

4. Mediators: to resolve interoperability problems.

Each of these WSMO Top Level Elements can be described with non-functional properties like creator, creation date, format, language, owner, rights, source, type; etc. WSMO comprises the WSMO conceptual model, as an upper level ontology for SWS, the WSML[16] language and the WSMX [17] execution environment.

The Web Service Modeling Language (WSML) is a formalization of the WSMO ontology, providing a language within which the properties of Semantic Web Services can be described.

WSMX provides an architecture including discovery, mediation, selection, and invocation and has been designed including all required supporting components enabling an exchange of messages between requesters and the providers of services.

4 Our approach

The approach relies upon the use of MDA concepts by developing two metamodels (source and target one) and a transformation model to translate XMI specifications (e.g., XML encodings of UML) into WSMO.

Figure 1 shows the overview of our approach. The model transformation is based on ATL language, and it relates two metamodels (source:UML and target: WSMO). A transformation engine takes a source model as input, and it executes the transformation program to transform this source model into the target model. The business model is created by any UML tool, consistent with the UML metamodel (UML profile). The obtained WSML document will be exported and validated by WSMO studio Tool.

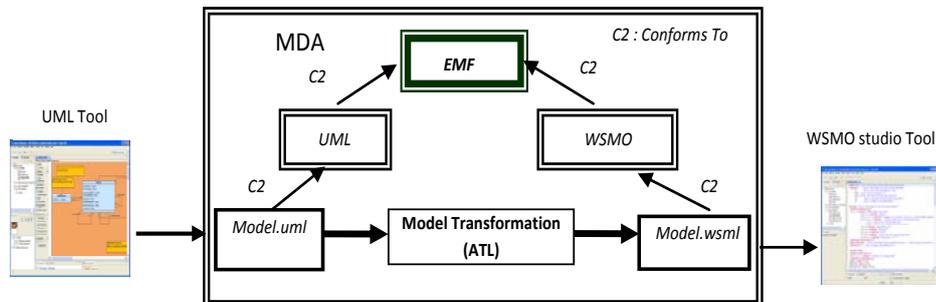


Fig.1. Architecture of our approach

4.1 The source metamodel

A UML profile [18] is a collection of stereotypes, tagged values and custom data types used to extend the capabilities of the UML modeling language. We use a UML profile to model various WSMO constructs in conjunction with the UML static structure diagram. In terms of MDA, the stereotypes, tagged values, and data types serve to mark-up the platform-independent model, or PIM, in order to facilitate transforma-

tion to WSMO specification. Stereotypes work well to distinguish different types of classes and create a meta-language on top of the standard UML class modelling constructs. Tagged values allow the developer to attach a set of name/value pairs to the model. Figure 2 shows a metamodel of our profile UML, where a group of extensions UML is introduced. The source metamodel consists of:

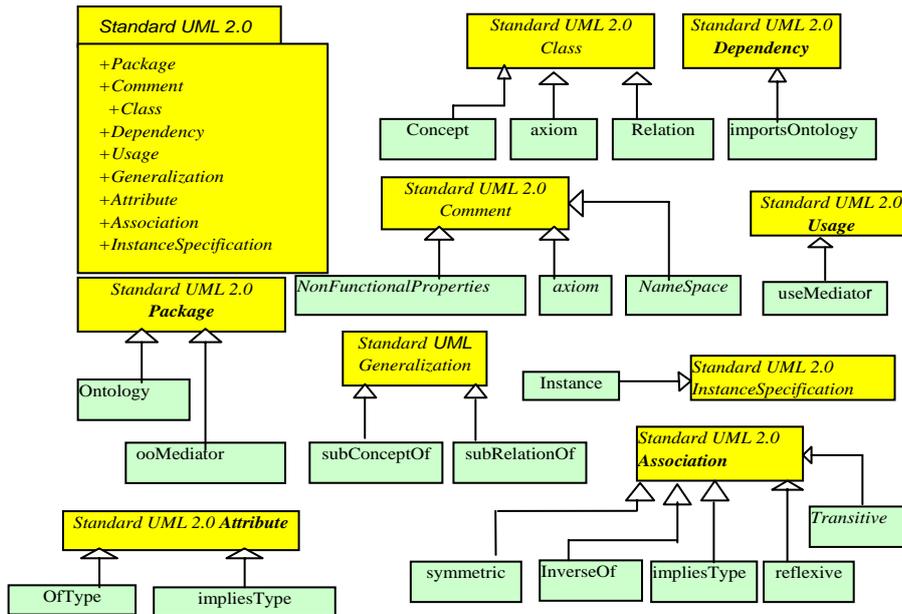


Fig.2. The source Metamodel : UML profile for WSMO

- The standard elements of UML: which are represented in the figure 2 by yellow color, we used: package, comment, Class, Dependency, Usage, Generalization, Attribute, Association, and InstanceSpecification. All these elements can be used in the class diagram for modeling the business model.
- Stereotypes: represented in the figure by the green color, they are introduced to allow the modeling of the diverse WSMO's constructs. The WSMO's constructs that we used are:
 - "Concept", " axiom ", "relation" which extend the " Class " element.
 - "Ontology", "ooMediator " which extend the "Package" element.
 - "NonFunctionalProperties","axiom" and "NameSpace" which extend the "comment" element.
 - " ImportsOntology " which extends the " Dependency " element.
 - "Instance" which extends the " InstanceSpecification " element.
 - « subConceptOf» and « subRelationOf » which extend « Generalization » element.
 - « OfType » and « impliesType » which extend « Attribute » element.

- « symmetric », « InverseOf », « impliesType », « reflexive » and « Transitive » which extend « Association » element.

4.2 The WSMO Ontology Target metamodel

This metamodel [19] is used by our transformation to generate the WSMO ontology. It consists of Ontology composed of : Concept, Relation, Axiom, Instance.

The figure 3 shows a fragment of WSMO metamodel.

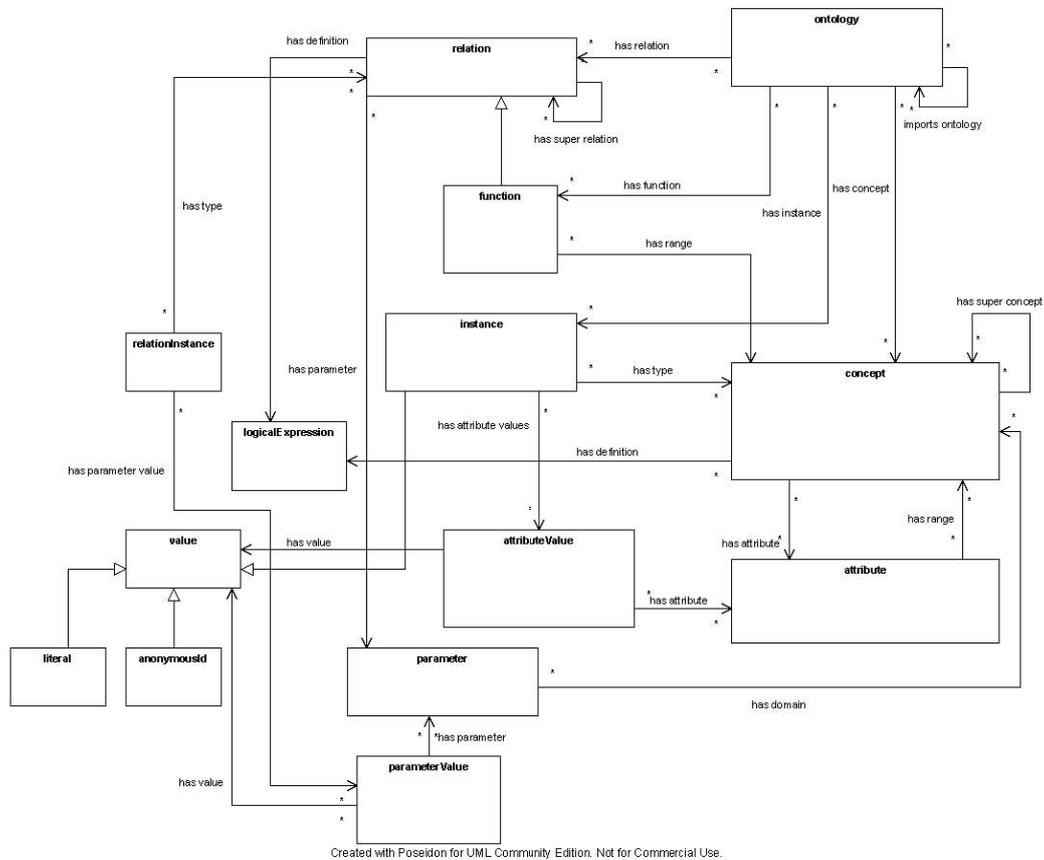


Fig.3. Fragment of WSMO Ontology metamodel

4.3 UML to WSMO transformations

4.3.1 Principle

The overview of the transformation is detailed in Figure 4. The set is split between two areas of modeling: MDE for space engineering models in which is defined the different metamodels described above and the transformation between UML and

WSMO, and WSMO space that defines the WSMO ontologies. In M3 layer we find ECORE language of metamodelisation. The both metamodels (UML and WSMO) and ATL metamodel located in M2 layer are based on ECORE.

At M1 layer we found the source model expressed in UML 2.0 conforms to our metamodel WSMO UML Profile, the model transformation UML2WSMO implemented in ATL language, WSMO ontology model resulting from the transformation process which is conforms to WSMO target metamodel in M2 layer, and a WSMO / WSML projector. This projector is a particular transformation that allows to switch from one model space to another. In our case it is used to transform the WSMO ontology in WSML document. Now we will explain in detail the transformation rules between our UML profile and WSMO ontology.

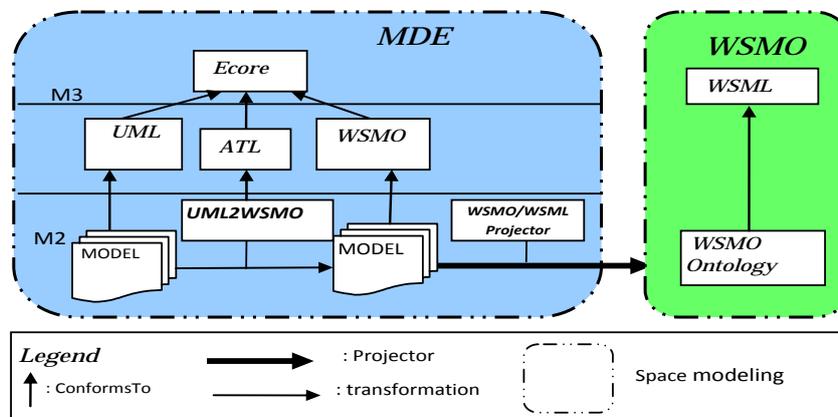


Fig. 4. UML to WSMO transformations

4.3.2 Transformation UML into WSMO

In our approach, a transformation definition is implemented in ATL language based on a mapping specification; we use the term mapping as a synonym for correspondance between the elements of two metamodels, while a transformation is the activity of transforming a source model into a target model in conformity with the transformation definition.

The source metamodel UML profile includes stereotypes, tagged values and constraints, each of which map to a particular construct in target metamodel WSMO description, as shown in table1. The left hand column provides the abstract type represented by the constructs. The middle column shows the UML constructs used to specify semantic services. Finally, the right hand two columns name the corresponding target construct in WSMO specification and present the target elements which are defined in this transformation.

Once mappings are specified between the two metamodels (e.g. UML and WSMO), transformation definitions are implemented using transformation languages such as Atlas Transformation Language (ATL). An extract of these rules is illustrated below.

Table 1. UML to WSMO Mapping

<i>UML TYPE</i>	<i>UML Construct</i>	<i>WSMO Construct</i>	<i>Elements defined in target WSMO construct</i>
Package	« ontology » stereotype	Wsmo Ontology	URI,nonfonctionnal properties,imports ontology,usesmediators,concept,relation,axiom,instances
Class	« Concept » stereotype	WSMO Concept	Concept name,subconceptof, nonfonctionnal properties, attribute
Class	« Relation » stereotype	Wsmo relation	Relationname, nonfonctionnalproperties, parameters,subrelationof
Class	« axiom » stereotype	Wsmo axiom	Name, Type,className,attributeName
Attribute	« oftype » attribute	Wsmo attribute	Name, Type,className,attributeName
Association	« implies type » stereotype	Wsmo attribute	Name, Type,className,attributeName
Association	« transitive_impliesType » stere.	WSMO attribute	Name, Type,className,attributeName
Association	« symmetric_impliesType »stere.	Wsmo attribute	Name, Type,className,attributeName
Dependency	« importsOntology» stereotype	WSMO import ontology	supplierName, clientName
Depenency	« usemediator» stereotype	Wsmo usemediator	supplierName, clientName
Comment	« nonfonctionalProperties» stereo.	Wsmo nonfonctional Properties	Name, body
Comment	« namespace » stereotype	Wsmo namespace	Name, body
Comment	« axiom » stereotype	Wsmo axiom	Name, expression_definition
Instance	« instance Speciiation» stere.	Wsmo Instance	Instance Name, memberOf, attributeValue
Specification			
Instance	« instanceproperty » stereotype	Wsmo attribute value	Instance Name,Instance, attributeName
Specification			

• **Rule : UMLClass2WSMOConcept**

This rule allows to create a concept WSMO from a class UML stereotyped "Concept". Any class UML stereotyped "Concept" is transformed into WSMO concept. This one is defined by the concepts from which it inherits "subConceptOf" , by a Name, NonFunctionalProperties and by these attributes.

```

rule UMLClass2WSMOconcept {
  from s : UML!"uml::Class" (s.hasStereotype('WSMO_Profil::Concept'))
  to t : b_Ontology_WSMO!Concept (
    Nom_concept <- s.name->debug('Cette Class Est Un Concept '),
    SubConceptOf <- if s.general.oclIsUndefined()then "
      else s.general->collect(a|a.name).first() endif,
    NonFunctionalProperties<-
    b_Ontology_WSMO!NonFunctionalProperties.allInstances() ->select(b|b.Nom_De = s.name)-
    >collect(b1|b1.Corps).first(),
    Attribute <- b_Ontologie_WSMO!Attribute.allInstances() ->
    select(b|b.Nom_De_Class = s.name)->collect(b1|b1.Nom_Attribute) ) }

```

The helper " hasStereotype " receives a string and returns a boolean. It is used to know if the current UML element is stereotyped as the string taken in parameter.

helper context UML!"uml::Element" **def:** hasStereotype(name : String) : Boolean = self.getAppliedStereotype(name)->oclIsKindOf(UML!Stereotype);

- **Rule : Property2Attribute**

This rule allows to create attributes WSMO from UML properties stereotyped "OfType ". Any property UML stereotyped "OfType" is transformed into WSMO attribute. This one is defined by a Name, Type, class names and Attribute name.

rule Property2Attribut

```
{ from P : UML!Property ( P.hasStereotype('WSMO_Profil::ofType'))
  to A : b_Ontology_WSMO!Attribute (
    Nom_Attribute <- P.name + 'ofType' + ' ' +
    P.type.toString().substring(4,P.type.toString().size()),
    Type_Attribute <- P.type.toString().substring(4,P.type.toString().size()),
    Nom_De_Class <- P.class.name ),
  At : b_Ontologie_WSMO!AttRelation (
    Nom_Attribute <- 'ofType' + P.type.toString().substring(4,P.type.toString().size()),
    --Type_Attribute <- P.type,
    Nom_De_Class <- P.class.name ) }
```

- **Rule : InstanceSpecification2WSMOInstance**

This rule transforms UML instance into WSMO instance. This one is defined by the classes that are "MemberOf", InstanceName and AttributeValues.

rule UMLInstance2WSMOInstance

```
{ from I : UML!InstanceSpecification
  to Ins : b_Ontology_WSMO!Instance ( Nom_Instance <- I.name,
    MemberOf <- I.classifier->collect(a|a.name),
    AttributeValues <- b_Ontology_WSMO!AttributeValue.allInstances( )
    ->select(b|b.name = I.name)->collect(a|a.Nom_Att_Ins ) }
```

4.3.3 WSMO2WSML Projector

A projector consists of one or several transformations allowing realizing the projection of an artefact belonging to a technological space towards another. In our case, the artefact is a model in compliance with the WSMO metamodel, belonging to the technical space of MDE. We aim to project this artefact towards the WSMO space in WSML syntax. A model in the MDE space is serialized in the XMI format, whereas a document in the WSMO space is in WSML format. The projector implemented here includes a single ATL query file allowing the transformation of WSMO model into a WSML document. Among the main rules which compose the transformation file named Ecore2wsmo, we quote the first rule:

- **Rule 1** : This rule allows to translate the "concept" element of WSMO modeled in ECORE Format into WSML syntax.

helper context UML!Concept **def:** toString_b() : String =
 '\n'+ 'concept' + self.Nom_concept + **if** self.SubConceptOf->iterate(e; acc : String =
 "/acc + e.toString()+ ") = 'OclUndefined' **then** '\n'

```

else
'subConceptOf'+self.SubConceptOf->iterate(e; acc : String = '' |acc + e.toString() + ') + '\n'
endif
+ if self.NonFunctionalProperties.oclsIsUndefined() then "
else ' nonFunctionalProperties '+ '\n' + self.NonFunctionalProperties + '
endNonFunctionalProperties '+ '\n' endif
+ if self.Attribute.size() <> 0
then self.Attribute->iterate(e; acc : String = '' |acc + e.toString() + '\n') + '\n'
else " endif;

```

5 Implementation

To show the feasibility of our approach, we have developed a tool on the Eclipse environment (Eclipse Ganymede 3.4.2), an ATL project is created, the UML metamodel profile is modeled by UML diagrams 2.1 Integrated in EMF eclipse, and the WSMO metamodel is modeled by the Ecore language (Ecore diagram), then two ATL files containing the transformation rules (UML2WSMO and WSMO2WSML) are written in ATL. For each execution, we introduce a source model conforms to our source metamodel and we obtain a WSMO document.

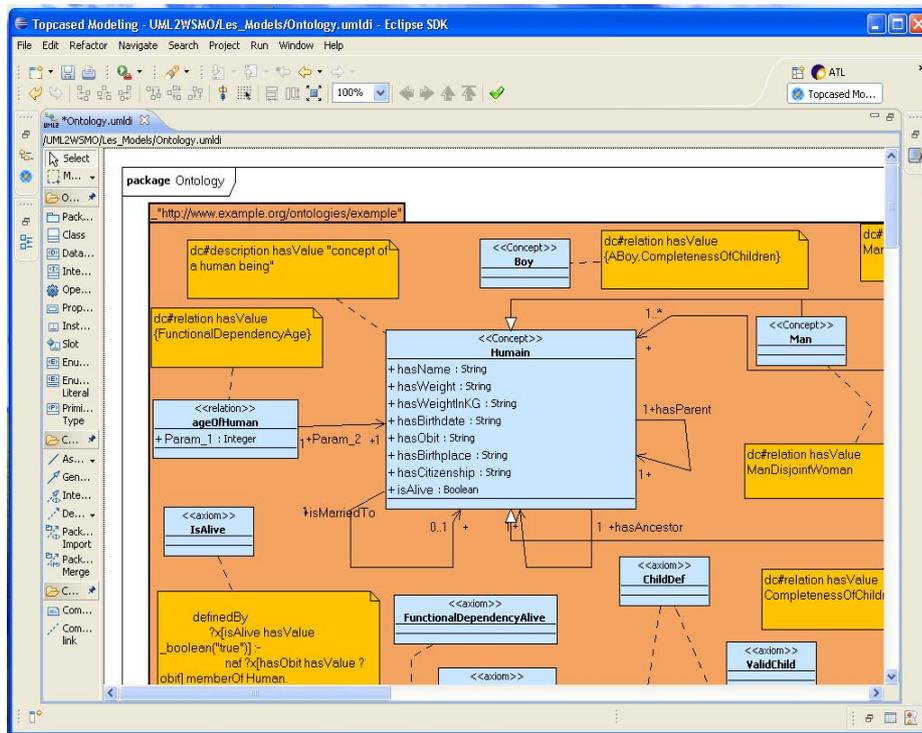


Fig.5. Class diagram of source model

To illustrate our approach, we take the human ontology as an example, the source model is expressed in a UML2.0 class diagram in conformity with our source metamodel (see Fig.5). The transformation engine takes the source model serialized in XMI [20] format as input and executes the transformation rules contained in UML2WSMO ATL file that generates a WSMO ontology in Ecore format, then the projector tool executes the ATL query file WSMO2WSML to transform the resulting WSMO ontology into WSML format. The obtained WSML document of the human ontology is depicted in Fig. 6. At the end, the output of our system is imported by the WSMO STUDIO tool to validate the correctness of our transformations.

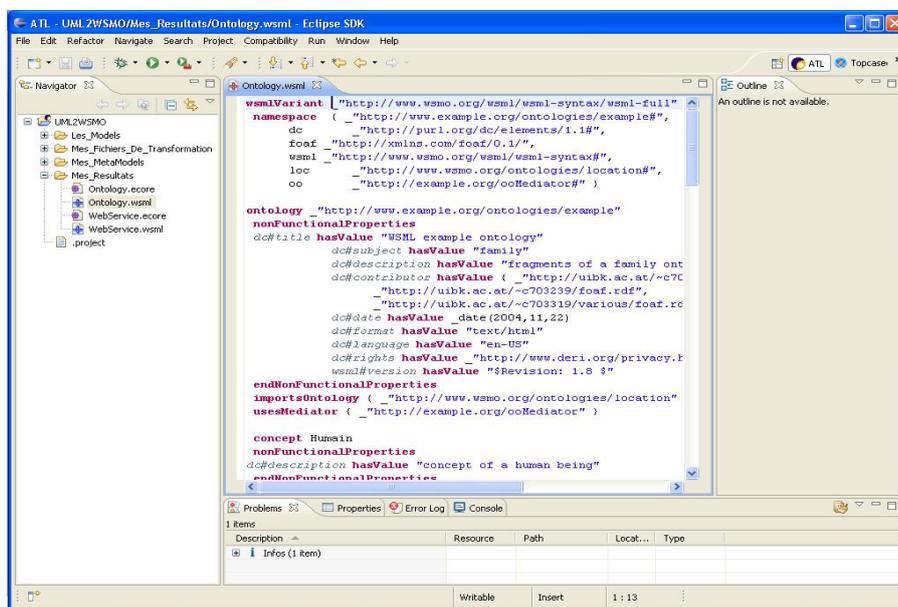


Fig.6. The resulting Human ontology file in WSML format

6 Conclusions and Future Work

Semantic Web Services can potentially change the way software is both developed and used. In order to realize that great promise, the software development community must embrace the technology. The barriers to adoption must be bridged in a manner that leverages the capabilities of developers.

We have been developing an MDA-based approach for facilitating such adoption by using Metamodels at PIM levels in such way that it more adequately hides the details of semantic web service technologies and allows the developer to focus on creating models of semantics web services, Then a transformation model consisting of a defined set of rules that specify the correspondence between the elements of both source and target metamodels, is applied to generate a WSMO ontology encoded in WSML.

Our current investigations involve developing the framework in such a way that it improves the transformation model by handling the logical expressions to cover "axiom" and covers the other parts of WSMO like capability, mediators and goals.

7 References

1. "WSMO Studio." vol. 2010, 2009: <http://www.wsmostudio.org>.
2. Kerrigan M.: "WSMOViz: An ontology visualization approach for WSMO," London, United kingdom, 2006, pp. 411-416.
3. Timm., J.T.E.: "A model-driven framework for the specification, grounding, and execution of semantic Web services." vol. PH.D: Arizona State University, 2008, p. 170.
4. Mukerji J.M.: "MDA Guide Version 1.0.1." vol. 2010: OMG Group, 2003.
5. O. M. Group, "OMG Unified Modeling Language (UML)." vol. 2010: Object
6. Wang Q., Garousi V., Madachy R., Pfahl D., Bendraou R., J.-M. Jezéquel J.M. and F. Fleurey F. : "Combining Aspect and Model-Driven Engineering Approaches for Software Process Modeling and Execution," in Trustworthy Software Development Processes. vol. 5543: Springer Berlin / Heidelberg, 2009, pp. 148-160.
7. Roman D., Lausen H., and Keller U.. Web service modeling ontology (WSMO). Final Draft D2v1.3, WSMO, 2006. Available from:<http://www.wsmo.org/TR/d2/v1.3/>.
8. Jouault, F., Kurtev, I. : Transforming Models with ATL, Proceedings of the Model Transformations in Practice Workshop at MoDELS'05, Montego Bay, Jamaica, 2005.
9. Gasevic, D., Djuric, D., Devedzic, V. : MDA-based automatic OWL ontology development, International Journal on Software Tools for Technology Transfer, June 2006.
10. Hori, M., Euzenat, J., Patel-Schneider, P., F. : OWL (Web Ontology Language) XML Presentation Syntax W3C Note 11 June 2003, <http://www.w3.org/TR/owl-xmlsyntax/>
11. Brambilla M., Ceri S., Facca F.M. , Celino I., Cerizza D., and Valle E.D.: "Model-driven design and development of semantic Web service applications," ACM Transactions on Internet Technology, vol. 8, 2007.
12. "Business Process Model and Notation (BPMN)," in Version 1.2. vol. 2010: OMG, 2009.
13. Stefano C., Piero F., Aldo B., Marco B., Sara C., and Maristella M. :Designing Data-Intensive Web Applications: Morgan Kaufmann Publishers Inc., 2002.
14. Sanchez D.M. , Acuna C.J., Cavero J.M. , and Marcos E. : "Toward UML-Compliant semantic web services development," International Journal of Enterprise Information Systems, vol. 6, pp. 44-56.
15. Cáceres P., Marcos, E., Vela, B. A. : "MDA-Based Approach for Web Information System Development," in Workshop in Software Model Engineering, 2003.
16. Lausen H., Bruijn J., Polleres A., and Fensel D. : WSML - A Language Framework for Semantic Web Services. In Proc. of the W3C Workshop on Rule Languages for Interoperability, Washington DC, USA, 2005.
17. Haller A., Cimpian E., Mocan A., Oren E., and Bussler C. WSMX - A Semantic ServiceOriented Architecture. In Proceedings of the International Conference on Web Service (ICWS 2005), 2005.
18. Atkinson, C. and Kühne, T.: Profiles in a strict metamodeling framework, Science of Comp.Prog., Vol. 44, No. 1 (2002) 5-22
19. <http://www.w3.org/Submission/2005/SUBM-WSMO-20050603/>.
20. OMG. XML Metadata Interchange (XMI) specification, version 2.0, formal/03/05/02, may 2003.

Foundations on Multi-Viewpoints Ontology Alignment

Lynda Djakhdjakha^{1,3}, Mounir Hemam², and Zizette Boufaida³

¹Department of Computer Science, University 08 May 1945 of Guelma,
Guelma 24000, Algeria
ldjakhdjakha@yahoo.fr

²Department of Computer Science, University of Khenchela,
Khenchela 40000, Algeria
Mounir.hemam@gmail.com

³LIRE Laboratory, Department of Computer Science,
Mentouri University of Constantine, Constantine 25000, Algeria
zboufaida@gmail.com

Abstract. In the last few years, a lot of effort has been paid to support both consensus and heterogeneity in the same ontology. As a result, multi-viewpoints ontologies have become essential for heterogeneous organizations and for diverse user communities that need to share and exchange information in an application domain. The development of multi-viewpoints ontologies by different communities makes distributed and heterogeneous knowledge resources not accessible. So, to solve this problem and in order to support users in sharing and reusing vocabularies and knowledge, we need for techniques for solving heterogeneity problems between different multi-viewpoints ontologies. In this case, multi-viewpoints ontology alignment is required to provide a consensual understanding of that domain represented by these ontologies. In the literature, there are much alignment systems, but existing systems are not capable to support multiple viewpoints. So, our challenge is to introduce the notion of multiple viewpoints in the alignment process. Therefore, in order to align multi-viewpoints ontologies, we present in this paper first, the definition of multi-viewpoints ontology in description logics extended by a stamping mechanism, and then we deal with their alignment problems and definitions.

Keywords: multi-viewpoints ontology, ontologies alignment, description logics, stamping mechanism.

1 Introduction

Ontologies can apprehend, capitalize, represent, operate and share semantically knowledge and information. However in reality, there are usually several ways to capture the knowledge of a given domain, that is to say different viewpoints (or perceptions) by which this knowledge can be represented. Thus, the same domain may have more than one ontology, where each one is described according to a viewpoint

or a particular perception. Indeed, in a large organization, there may be several communities or groups of individuals with their own viewpoints. These viewpoints depend on the type of person (occupation, age, educational level, experience ...) or use (the same person may have different views depending on the task which he is trying to accomplish) [1].

The two concepts ontology and viewpoint are complementary [2], indeed the ontology represents the knowledge shared by multiple users and the viewpoint represents the domain of knowledge that is relevant at a given viewpoint. With the coupling of these two notions we are talking about multi-viewpoints ontology. The latter gives the same universe of discourse several partial descriptions such that each one is on a particular viewpoint.

The exploration of multi-viewpoints ontologies [1], [3-4] could be an efficient way for heterogeneous organizations to share knowledge. The development of these ontologies by different communities makes distributed and heterogeneous knowledge resources not accessible. So, to solve this problem and in order to support users in sharing and reusing vocabularies and knowledge, we need for techniques for solving heterogeneity problems between different multi-viewpoints ontologies. In this case, multi-viewpoints ontology alignment is required to provide a consensual understanding of that domain represented by these ontologies.

An ontology alignment is defined as a set of correspondences between ontological entities, i.e. classes, properties, and individuals, of two ontologies [5]. In the literature, there are much alignment systems presented in [6] and [7]. However, existing systems are not capable to support the notion of multiple viewpoints. So, our challenge is to introduce this notion in alignment process to align multi-viewpoints ontologies.

Therefore, the objective of this work is to find an alignment definition of the multi-viewpoints ontologies described in description logics extended by a stamping mechanism, and to find a definition of the semantics of these alignments. The aim of these definitions is to be able to align multi-viewpoints ontologies.

The rest of the paper is organized as follows. Section 2 presents a multi-viewpoints ontology. In section 3 we detail syntax and semantics of multi-viewpoints ontology described in description logics extended by stamping mechanism. Section 4 looks at the problems of the multi-viewpoints ontology alignment. In section 5 we present a set of definitions for the multi-viewpoints ontology alignment, and Section 6 concludes with the future directions of work.

2 Definition of Multi-Viewpoints Ontology (MVp Ontology)

2.1 Viewpoint Approach

For a given domain of knowledge, several criteria can be used to observe an object. These different perceptions of the world are called viewpoints or perspectives. In computer science, most of data modeling systems don't deal with the variety of perceptions related to the same universe of discourse, and develop tools to create a single model for a single vision of the observed world. The viewpoint approach is opposed

to this monolithic approach and makes it possible to model the same reality according to different points of view [8].

The viewpoint approach is constructed on the conjunction actor/information. Therefore, it is necessary to include the actor in the action. In [9] viewpoint is defined as “*a conceptual manner binding, on the one hand an actor who observes and, on the other hand, a universe of discourse which is observed*”.

2.2 Multi-Viewpoints Ontology

A multi-viewpoints ontology is defined as an ontology in which a concept can have multiple definitions, each definition corresponding to a particular viewpoint on the concept [1], [3].

In [4], a multi-representation ontology is seen as an ontology that characterizes an ontological concept by a variable set of properties (static and dynamic) or attributes in several contexts and granularities.

The accepted definition in our work is that mentioned in [2] and [8], where a multi-viewpoints ontology is defined as an ontology that gives the same universe of discourse, several partial descriptions, in a way that, each one is on one viewpoint. These partial descriptions share on a global level the ontological elements (concepts and global roles) and semantic links constitute a consensus between the different viewpoints.

3 A Multi-Viewpoints Ontology in Description Logics Extended by a Stamping Mechanism

Description logics (DL) [10-11] are a family of knowledge representation languages that can be used to represent the knowledge of an application domain in a structured and formally well-understood way. A stamping mechanism allows multiple representations of concepts. In [2] and [8], description logics extended by a stamping mechanism have a signature, which is based on the following types: global concepts, local concepts, global roles, local roles, individuals and bridge rules. This language allows the use of constructors to create complex elements.

Definition 1 (Multi-viewpoints Ontology in Description Logics Extended by a Stamping Mechanism). A multi-viewpoints ontology described in description logics extended by a stamping mechanism is a multiple descriptions of the same universe of discourse according to different viewpoints. It is defined as the quadruple $O_{mpv} = \langle C^G, R^G, V_p, M \rangle$ where: C^G : is a set of global concepts, R^G : is a set of global roles, M : is a set of bridge rules and V_p : is a set of viewpoints, where a viewpoint is a partial description of a universe of discourse in a particular perception. It is defined by the triplet $\langle C^L, R^L, A^L \rangle$, where: C^L : is a set of local concepts, R^L : is a set of local roles and A^L : is a set of local individuals.

Definition 2 (Syntax of a Global Concept). Given $V_p = \{vp_i, \dots, vp_k, \dots, vp_m\}$ a set of viewpoints. A global concept C^G can be formed using the Boolean manufacturers (conjunction, disjunction) and the following global restrictions manufacturers:

- $\forall_{vp_1, \dots, vp_k} R.C, \exists_{vp_1, \dots, vp_k} R.C$: defines a new concept that all their instances are connected by the role R ,
- $\leq_{vp_1, \dots, vp_k} R.C, \geq_{vp_1, \dots, vp_k} R.C$: specifies the minimum or maximum cardinality of the role R in the viewpoints vp_i to vp_k .

Definition 3 (Syntax of a Local Concept). Given $vp_i \in V_p$. A local concept $vp_i: C^L$ is either a primitive local concept or a defined local concept:
 $vp_i: C^G \rightarrow (Gocal\ Concept) \mid (\neg C^L) \mid (C^L \sqcap D^L) \mid (\exists R^L.C^L) \mid (\forall R^L.C^L) \mid$
 $(\geq n R^L.C^L) \mid (\leq n R^L.C^L) \mid (a_1, a_2, \dots)$, where C^L and D^L are local concepts, R^L is a local Role, a_1, a_2, \dots are individuals, and n is a natural number.

Definition 4 (Syntax of a Local Role). A given local role $vp_i: R^L$ is defined as: $vp_i: R^L(C^L, D^L)$, where R^L is the name of the local role defined in the viewpoint vp_i , C^L and D^L are two local concepts defined in this viewpoint vp_i . As well, a local role R^L can be a primitive local role or a defined local role: $(R^L \sqcap S^L), R^L \sqcup S^L, \neg R^L, R^L \circ S^L$ and R^{L-}, R^{L+} , where R^L and S^L are given local roles.

Definition 5 (Syntax of a Global Role). A global role R^G is defined as: $R^G(vp_i: C^L, vp_j: D^L)$, where R^G is the name of the global role, C^L and D^L are two local concepts defined in two different viewpoints. As well as a local role, a global role can be a primitive or a defined global role.

Definition 6 (Syntax of a Subsumption Relation). Below a viewpoint vp_i , a local hierarchy vp_i/H , is defined as a triplet $(C^L, \theta, \sqsubseteq)$ where: C^L is a set of local concepts, θ is a function of C^L in C^G , which associated for all root concept denoted C_S^L of C^L a global concept of C^G , and \sqsubseteq is the subsumption relation used to express explicitly a direct order relationships as follows:

- $vp_i: C^L \sqsubseteq vp_i: D^L$, where C^L and D^L are two local concepts defined in the same viewpoint vp_i .
- $vp_i: C_S^L \sqsubseteq C^G$, where C_S^L is the most general concept defined in the viewpoint vp_i , and C^G is a global concept.

Definition 7 (syntax of a Bridge Rule). We distinguish:

- *Inclusion.* $vp_i: X \xrightarrow{\sqsubseteq} vp_j: Y$. It expresses the set inclusion between a local concept extension of a viewpoint and another concept of another viewpoint.

- *Inclusion with Multiple Sources.* $vp_i: X \sqcap \dots \sqcap vp_k \xrightarrow{\sqsubseteq} vp_j: Y$. It expresses an inclusion relation between a list of local concepts belonging to several viewpoints and another destination concept belonging to another viewpoint.
- *Bidirectional Inclusion.* $vp_i: X \xleftrightarrow{=} vp_j: Y$. It expresses equality between two local concepts belonging to two different viewpoints.
- *Bidirectional Exclusion.* $vp_i: X \xleftrightarrow{\perp} vp_j: Y$. It expresses a relationship between two local concepts belonging to two different viewpoints.

3.1 Semantics of Multi-Viewpoints Ontology

In [2] and [8], the semantics of description logics extended by the stamping mechanism is defined by a global interpretation, a set of local interpretations, and a set of domain relations:

Definition 8 (Local Interpretation). A local interpretation $(\Delta^{I_k}, \cdot^{I_k})$, is associated for each local element, where Δ^{I_k} is a domain of local interpretation and \cdot^{I_k} is a local interpretation function such that for all local concepts of \mathcal{C}^L , $\mathcal{C}^{L^{I_k}} \subseteq \Delta^{L^{I_k}}$, for all local roles R^L , $R^{L^{I_k}} \subseteq \Delta^{L^{I_k}} \times \Delta^{L^{I_k}}$, and for all individuals a , $a^{I_k} \subseteq \Delta^{I_k}$.

Definition 9 (Global Interpretation). A global interpretation (Δ^I, \cdot^I) is associated for each global element, where $\Delta^I = \Delta^{I_1} \cup \dots \cup \Delta^{I_k} \dots \cup \Delta^{I_m}$ is a domain of global interpretation and \cdot^I is a global interpretation function such that for all global concepts \mathcal{C}^G , $\mathcal{C}^{G^I} \subseteq \Delta^I$, for all global role R^G , $R^{G^I} \subseteq \Delta^I \times \Delta^I$.

Definition 10 (Domain Relation). A relation domain defines how two different viewpoints interact.

4 The problems of Multi-Viewpoints Ontology Alignment

In this work, we will take into consideration the notion of viewpoint in the alignment process. Thus, the multi-viewpoints ontologies alignment is the task to find the relationships that hold between the entities belonging to these ontologies.

Multi-viewpoints ontologies cover different domains that can be modeled differently and can represent several viewpoints. They can support both heterogeneity (at a local level) and consensus (at a global level). In the multi-viewpoints ontologies alignment, there is a great heterogeneity resides in the variations present in the semantic coverage of comparable concepts, especially local concepts that are defined according to different viewpoints and that can be semantically very similar. Indeed, the mapping task or alignment task between multi-viewpoints ontologies is more difficult than that between classical ontologies, because there is much specificity for this process. Among the specificities of multi-viewpoints ontologies alignment, we may find:

4.1 Elements to Align

In the alignment process between classical ontologies, the task is to discover correspondences between concepts, properties and individuals. In our context of multi-viewpoints ontologies described in description logic extended by stamping mechanism, it is necessary to take into account the different types of concepts (global and local), the different types of roles (global and local), individuals.

4.2 Localization of Local Elements According to the Different Viewpoints

In our context, there are two description types: a global description and other partial descriptions defined according to different viewpoints. So, taking into account the localization of the local concept (local role) in the alignment process influenced on the remaining correspondences that can be discovered.

4.3 Bridge Rules

The consideration of bridges rules between elements of different viewpoints impact on the set of correspondences found.

5 Multi-Viewpoints Ontologies Alignment

We adopt the definitions of classical ontology alignment presented in [12] and [13], to define the MVP ontology alignment.

Definition 11 (MVP Ontology Alignment). Given \mathcal{O}_{mvp} , \mathcal{O}'_{mvp} two multi-viewpoints ontologies in description logics extended by a stamping mechanism. An alignment between \mathcal{O}_{mvp} , \mathcal{O}'_{mvp} is defined as the task to find the *best subset of the correspondences* between multi-viewpoints ontology elements belonging to \mathcal{O}_{mvp} and \mathcal{O}'_{mvp} .

Definition 12 (MVP Ontology Element). A multi-viewpoints ontology element is a term of the multi-viewpoints ontology (e.g., global concept, local concept, global role, local role or individual).

Definition 13 (Correspondences). Given \mathcal{O}_{mvp} , \mathcal{O}'_{mvp} two multi-viewpoints ontologies in description logics extended by a stamping mechanism, a correspondence between \mathcal{O}_{mvp} , \mathcal{O}'_{mvp} is defined as a triple $\langle e_{mvp}, e'_{mvp}, A_{mvp} \rangle$ where: e_{mvp} and e'_{mvp} are multi-viewpoints ontology elements from the two multi-viewpoints ontologies to align, and A_{mvp} is a relation that is asserted to hold between e_{mvp} and e'_{mvp} .

Multi-viewpoints ontologies alignment contains correspondences between global concepts, local concepts, global roles, local roles or between individuals belonging to these ontologies. These correspondences are similar to the bridge rules in a same mul-

ti-viewpoints ontology. We can identify the different types of correspondences between two multi-viewpoints ontologies i and j as follows: $i: C^G \stackrel{\equiv}{\leftrightarrow} j: D^G$, $i: C^L \stackrel{\equiv}{\leftrightarrow} j: D^L$, $i: C^L \stackrel{\equiv}{\leftrightarrow} j: D^G$, $i: R^G \stackrel{\equiv}{\leftrightarrow} j: S^G$, $i: R^L \stackrel{\equiv}{\leftrightarrow} j: S^L$, $i: R^L \stackrel{\equiv}{\leftrightarrow} j: S^G$, $i: C^G \stackrel{\perp}{\leftrightarrow} j: D^G$, $i: C^L \stackrel{\perp}{\leftrightarrow} j: D^L$, $i: C^L \stackrel{\perp}{\leftrightarrow} j: D^G$, $i: R^G \stackrel{=}{\leftrightarrow} j: S^G$, $i: R^L \stackrel{=}{\leftrightarrow} j: S^L$, $i: R^L \stackrel{=}{\leftrightarrow} j: S^G$, $i: a \stackrel{\in}{\leftrightarrow} j: C^G$, $i: a \stackrel{\in}{\leftrightarrow} j: C^L$, $i: a \stackrel{=}{\leftrightarrow} j: b$, where C^G and D^G are global concepts, C^L and D^L are local concepts, R^G and S^G are global roles, R^L and S^L are local roles, a and b are individuals, $\stackrel{\equiv}{\leftrightarrow}$ is a subsumption relation, $\stackrel{\perp}{\leftrightarrow}$ is a disjunction relation, $\stackrel{\in}{\leftrightarrow}$ is a membership relation, and $\stackrel{=}{\leftrightarrow}$ is an identity relation.

Consequently, a multi-viewpoints ontology alignment involves a set of entities connected by symbols of relations. It is seen as a pair (E_{mvp}, A_{mvp}) , where E_{mvp} is a set of entities described in description logics extended by a stamping mechanism. And A_{mvp} is a set of symbol of relationships between these entities. So, it is necessary to interpret the pair (E_{mvp}, A_{mvp}) before interpreting a multi-viewpoints ontology alignment.

Definition 14 (Entities Interpretation). Interpretation of entities is a pair (Δ^I, \cdot^I) , where Δ^I is a domain of interpretation and \cdot^I is an interpretation function such that for all local concepts C^L , $C^{L^I} \subseteq \Delta^{L^I}$, for all global concepts C^G , $C^{G^I} \subseteq \Delta^I$, for all local roles R^L , $R^{L^I} \subseteq \Delta^{L^I} \times \Delta^{L^I}$, for all global roles R^G , $R^{G^I} \subseteq \Delta^I \times \Delta^I$, and for all individuals a , $a^I \subseteq \Delta^I$, where Δ^{L^I} is a local domain of interpretation and it is a subset of Δ^I .

Definition 15 (Multi-Viewpoints Ontologies Alignment Interpretation). In our context, all entities are coming from the same description language. An interpretation of an alignment relation between two multi-viewpoints ontologies i and j is a pair (Δ^I, \cdot^I) , where Δ^I is a global domain of interpretation and \cdot^I is a binary relation of interpretation such that for all relations $A_{mvp}, A_{mvp}^I \subseteq \Delta^{I^i} \times \Delta^{I^j}$, where Δ^{I^i} and Δ^{I^j} are domain of interpretation of the multi-viewpoints ontologies i and j respectively, and are subsets from Δ^I . Each multi-viewpoints ontology has its own domain of interpretation.

6 Conclusion

In this paper, we specified the basic concepts that constitute a starting point for multi-viewpoints ontology alignment.

Several research directions are considered to carry out this work. We will use these definitions to propose a method of multi-viewpoints ontology alignment. We, also, plan to include bridge rules between multi-viewpoints ontologies elements for the reasoning on the multi-viewpoints ontologies and alignments in order to obtain the best set of alignments between these ontologies.

References

1. Falquet, G., Mottaz, J.C.L. : Navigation Hypertexte dans une Ontologie Multi-Points de Vue. In NimesTIC'2001, Nîmes, France (2001)
2. Hemam, M., Boufaïda, Z.: MVP-OWL: a multi-viewpoints ontology language for the Semantic Web. In International journal of Reasoning-based Intelligent Systems (IJRIS). Inderscience Publishers, Vol. 3, No. 3/ 4, pp. 147-155 (2011)
3. Falquet, G., Mottaz, J.C.L. : A Model for the Collaborative Design of Multi-Point-of-View Terminological Knowledge Bases. In R. Dieng and N. Matta (Eds) Knowledge Management and Organizational Memories, Kluwer, 2002. Preliminary version published in Proceedings of the Knowledge Management and Organizational Memory workshop of the International Joint Conference on Artificial Intelligence, Stockholm (1999)
4. Benslimane, D., Arara, A., Falquet, G., Maamar, Z., Thiran, F., Gargouri, F.: Contextual Ontologies : Motivations, Challenges, and Solutions, T. Yakhno and E. Neuhold (Eds.): ADVIS 2006, LNCS 4243, pp. 168–176, c_Springer-Verlag Berlin Heidelberg (2006)
5. Euzenat, J. , Shvaiko, P.: Ontology matching. In Springer, Heidelberg (DE) (2007)
6. Rahm, E. , Bernstein, P.: A survey of approaches to automatic schema matching. The LDB Journal, 10(4):334–350 (2001)
7. Shvaiko, P., Euzenat, J.: A survey of schema-based matching approaches. Journal on Data Semantics, IV:146– 171(2005)
8. Hemam., M. et Boufaïda., Z. : Représentation d'ontologies multi-points de vue: une approche basée sur la logique de descriptions. In 20^{es} Journées Francophones d'Ingénierie des Connaissances (IC'09) (2009)
9. Benchikha, F., Boufaïda, M. : The Viewpoint Mechanism for Object-oriented Databases Modelling, Distribution and Evolution” In Journal of Computing and Information Technology. Vol 15, p. 95-110 (2007)
10. Baader, F., Calvanese, D., McGuinness, D., Nardi, D., Patel-Schneider, Peter F., editors, The Description Logic Handbook: Theory, Implementation and Applications, chapter 2, pages 43-95. Cambridge University Press (2003)
11. Baader, F., Horrocks, I., Sattler, U.: Chapter 3 Description Logics. In Frank van Harmelen, Vladimir Lifschitz, and Bruce Porter, editors, Handbook of Knowledge Representation. Elsevier (2007)
12. Zimmermann, A., Euzenat, J.: Three semantics for distributed systems and their relations with alignment composition. In : Proc. 5th conference on International semantic web conference (ISWC), Athens (GA US), Lecture notes in computer science 4273:16-29 (2006)
13. Zimmermann, A., Duc, C. : Reasoning with a Network of Aligned Ontologies. In RR '08 Proceedings of the 2nd International Conference on Web Reasoning and Rule Systems (2008)

Networks and Security



A Flexible Integration of Security Concern in Rule based Business Process modeling

Khadhir Bekki¹ , Hafida Belbachir²

¹ Department of Computer science, Ibn khaldoune University, Tiaret, Algeria

Bekki_kh@Yahoo.fr

²Department of Computer science, Mohamed Boudiaf University, Oran, Algeria

H_Belbach@Yahoo.fr

Abstract Today, to stay competitive, organizations are in the quest to execute their business processes correctly and continuously. This need require to apply risk, security and business process management in a more integrated way. At the same time, business processes need to be more flexible and adaptable. Habitually, The business rules represent main driving force for adaptability and competitiveness in organizations. The ECA (Event-condition-action) is a popular way to incorporate flexibility into a process design. As well, separation of concerns becomes one of the cornerstone principle in software engineering, and it supports adaptation in several ways. In this paper, we propose a flexible way to integrate security concern into rule based business process modeling. First, we govern any business activity through our ECATE formalism (Event-Condition-Action-Temporal condition- trigger Event) based on business rules. Then, we integrate the security requirements in a separate concern as EUCATE rules (a variant of ECATE rule). The rules based process will verified before being deployed in the runtime environment

Keywords: Business processes modeling, business rules, flexible modeling, separation of concerns, security.

1 Introduction

Actually, companies are more to more in the quest to execute their business processes correctly and continuously. Within the last years, the private sector has noticed a growing need to improve security to meet tighter regulative and legal requirements[1]. This need forced organizations to integrate the capture of security requirement in the business process modeling.

The early design of security requirements have some benefits [2] (1) use the security knowledge of security business process analysts at high level in modeling step.

(2) reduce potential costs avoiding the additional implementation of business process security after the implementation of business process. (3) simplify the capturing of the security requirements. As well, flexibility, adaptability and correctness, besides knowledge-intensiveness belong to the most challenging issues of business process [3].

The BPEL language does not provide any support for the specification of either authorization policies or authorization constraints on the execution of activities composing a business process [4]. It is important that such an authorization model be high-level and expressed in terms of entities that are relevant from the organizational perspective [4]. The regulations and policies in organizations are often expressed in terms of business rules that are sometimes defined as high-level structured statements that constrain, control, and influence the business logic [5]. Business rules are defined as [5]: "the set of policies for regulating the whole business within and out-side an organization". They represent main driving force for adaptability and competitiveness. The ECA pattern has been widely adopted for business rules [6]. They are an interest way to incorporate flexibility into a process design. And, they are a popular approach to catch unanticipated events and adapt to exceptions [7].

As well, separation of concerns provides a way to separate development of the functionality and the crosscutting concerns (e.g., quality of service, security). This principle has become one of the cornerstone principle in software engineering, and has lead to a wide spread of aspect-oriented programming (AOP) approach [8].

The advantages in addressing each concern separately are transparency, evolution, understandability and scalability. More, it is necessary to bring them together to understand which global system properties emerge at any given activity [9].

In order to incorporate flexibility and adaptability into a business process design, and benefit of the advantages of separation of two concerns: security and functional in business process modeling, we propose, in this paper, a new rule based model that wants to improving the flexibility, adaptability of business process.

First, for the functional concern, we govern any business activity through our ECATE formalism (Event-Condition-Action-Temporal condition- trigger Event) based on business rules. Then, we integrate the security requirements in a separate concern as EUCATE rules (a variant of ECATE rule).

The rest of this paper is organized as follows. In the second section, we present rule based business process modeling as set of ECATE rules. The third section explain how to integrate flexibly the security requirement in the ECATE rules based process. The section 4 gives a related works. Finally, wrapped up by some concluding remarks and further required extensions of this work.

2 A Rule based business process modeling

2.1 Definition

The process modeling aims to provide high-level specification independent from implementation of such a specification. To support verification, validation, simulation of the automated process, the process modeling language provides the appropriate syntax

and semantics to specify the precise requirements of business processes and reflect the logic of the underlying process

As given in [10], two formalisms on which the most predominant process modeling languages are developed, are graph-based formalism and rule based formalism.

Rule-based approach proposes to model the logic of the process with a set of business rules. Each rule specifies properties of one or more business activity, such as the pre and post conditions of execution. In comparison with graph based approaches, the rule based approaches are more expressive and flexible [10]. They are able to express the temporal requirements. They take advantage in adaptation to ad hoc modification at runtime and exception.

Business rules are considered as policies, laws and know-how for doing business in any cross-organizations. The ECA pattern has been widely adopted for business rules [6]. It is an interest way to incorporate flexibility into a process design. The E-C-A paradigm has been the foundation for many rule-based processes modeling approaches. A survey of rule based approaches is given in [10].

To cope with flexibility, adaptability and temporal requirements of business process, we propose an ECA based formalism ECATE to govern business rules as follows:

<i>ON</i>	Event
<i>IF</i>	Condition
<i>DO</i>	Action
<i>TIME</i>	Constraint of execution Time
<i>Trigger</i>	Post Event

Its semantics is: for each concern (C) when the event (E) occurs, the activated rule evaluates the condition (C). The condition is either a Boolean expression or a SQL query on the database. If the condition is satisfied, the action (A) is executed. The Time (T) is a condition on the execution time. It captures the constraints of time. This condition is of type “before t”, “after t”, “during t” or a combination of three types. before t means that the action A should be performed before the time t, “after t” means that the action A should be performed after the time t. “during t” means that the execution time of the action A should not exceed the time t. If the time constraint is violated then the process will be interrupted and a compensating action will be launched. The event triggered E design the set of events raised after the execution of the action.

2.2 Example

In order to give an intuitive idea about our formalism, let us consider the following scenario, inspired from [11]. Upon receipt of customer order, the calculation of the

initial price of the order and shipper selection is done simultaneously. When both tasks are complete, a purchase order is sent to the costumer. In case of acceptance, a bill is sent back to the customer. Finally, the bill is registered. A Functional constraint exists in this scenario: the bill payments must be made 15 days before the delivery date. The security constraints in this scenario are: 1) the client must be authenticated in the company system to control purchases. 2) The client must be authenticated in bank system to do banking. 3) If the amount of the bill exceeds some value m, the client must have an authorization between 08h00 and 19h00 to pay bill. The figure 1 shows the modeling of the functional concern of this example.

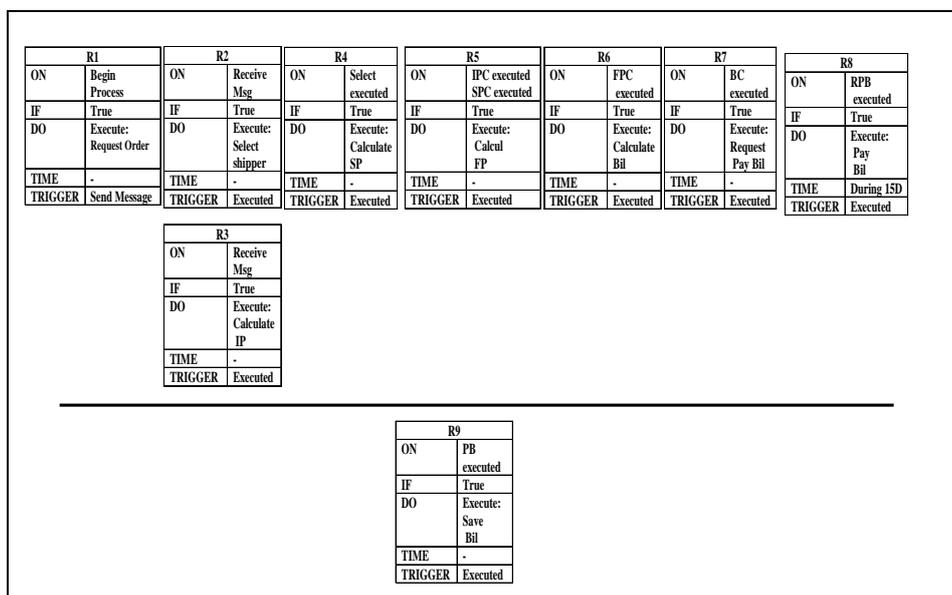


Fig. 1. ECATE rules based Business process

This model represents the business process of the purchase order process as set of ECATE Rules. So, The business rules are governed as ECATE rules. The event "begin process" activates the business process. It represents customer order (it may be, for example, clicking on the button "Place an order"). The two rules R2 (policy of initial price calculation), R3 (policy of shipper selection) have the same event to be activated. They represent two Parts of business process which will be executed in parallel. The constraint "the bill payments must be made 15 days before the delivery date" is specified in the time condition of the rule R8. The attribute time contains the value "during 15D" which means that if the execution of the action pay bill exceed 15 day after the activating event "request pay bill executed", so the order will be rejected, and a compensation action, to compensate the executed action part effects, will

Be launched. The successful execution of the rules R2 ,R3 actions will activate the rule R4. In turn, the execution of this rule action activates another rules. And so on, until the end of process rules set.

So, the business process of the purchase order, in this example, is governed in a flexible way as a set of ECATE rules. A flexibility way mean that we can implement changes in some rules (parts of a business process) without affecting the rest of rules (other parts).

But, this ECATE rule based model take only the functional concern of the process.

3 Flexible integration of security concern

Separation of concerns provides a way to separate development of the functionality and the crosscutting concerns (e.g., quality of service, security). This principle has become one of the cornerstone principle in software engineering, and has lead to a wide spread of aspect-oriented programming(AOP) approach [8]. The advantages in addressing each concern separately are transparency, evolution, understandability and scalability. More, it is necessary to bring them together to understand which global system properties emerge at any given activity [9]. Some scientific research efforts have interested to integrate the capture of security requirements in business process modeling. A survey of these works is given in [3]. But, they haven't used an ECA based formalism to capture the security requirement. Governing the business rules as ECA rules with separation of concerns have many benefits including[9] (1) the inherent ability of adapting any concern rules before imposing them on running services or components; (2) the promotion of understandability of each concern in isolation and then the study of the coherent composition.

In order to integrate the security concern flexibly into a business process design, and benefit of the advantages of separation of two concerns: functional and security in business process modeling, we use, in this section, EUCATE rule, which is a variant of ECATE, to govern the security requirement.

Our formalism EUCATE is defined as follow:

<i>ON</i>	Event
<i>USER</i>	Activity User
<i>IF</i>	Condition
<i>DO</i>	Action
<i>TIME</i>	Constraint of execution Time
<i>Trigger</i>	Post Event

It have the same semantic of ECATE. The added attribute user specifies the activity user. The figure 2 shows the integration of security requirements in the previous model, using EUCATE rule in separate concern.

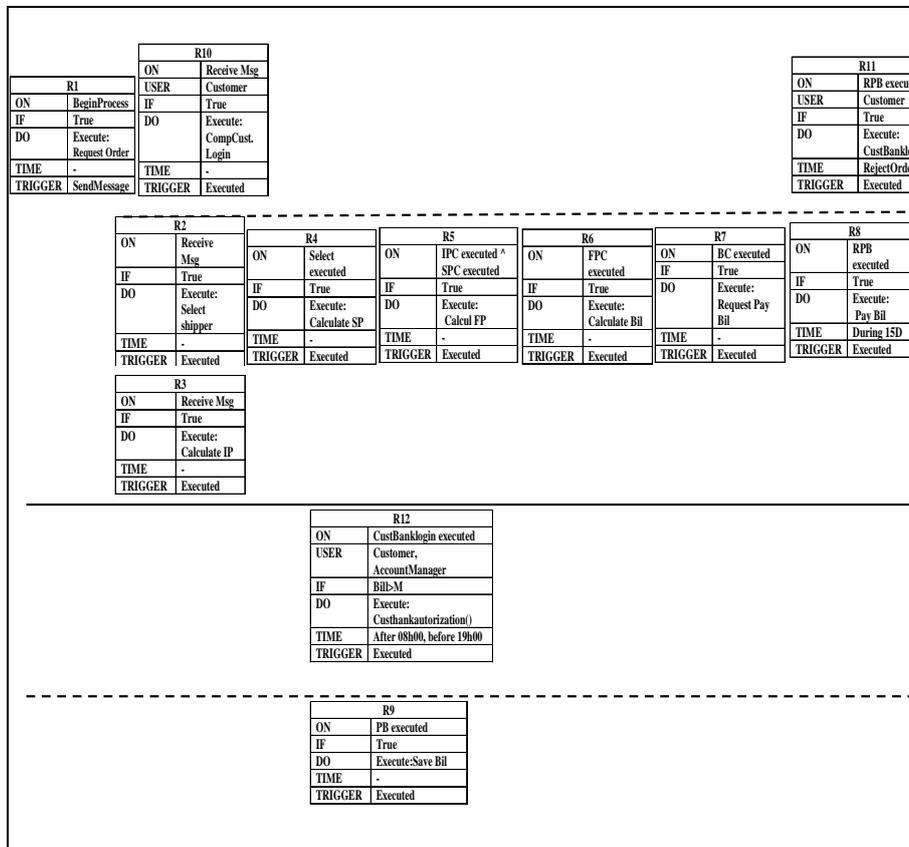


Fig. 2. Integration of security concern

The security requirements are modeled separately as set of EUCATE Rules. The separation of concerns promotes the understandability of each concern in isolation. For example, The rules R10, R11, R12 are of security concern that govern a security constraints. These rules may be modeled and handled by a security expert designer, independently of other concerns. The three rules R10 (policy of Company customer login) R2 (policy of initial price calculation), R3 (policy of shipper selection) have the same event to be activated. It is “begin process” event that represents customer order (it may be, for example, clicking on the button "Place an order"). However, they can't be activated at the same time, because they are of two different concerns. To avoid conflict between concerns, the security concern has more priority. In result, the rule R10 is activated before the rules R2 and R3. More, the rules R2 and R3 can not be activated if the R10 is not activated successfully. In other words, the condition and the time condition of R10 must be satisfied. If not, the order will be rejected. So, it will be

useless to activate the rules R2 and R3. In a positive case, R2 and R3 will be activated in the same time, because they are of the same concern. In turn, the execution of these rules actions activates another rules. And so on, until the end of process rules set.

So, the business process of the purchase order is governed now in a flexible way as a set of rules divided on two concerns: security concern and functional concern. A flexibility way mean that we can implement changes in some rules (parts of a business process) without affecting the rest of rules (other parts).

4 Verification of rules based process

It is important that a process model is correctly defined, analyzed, refined and verified before being deployed in the runtime environment[10].

The exceptions healing of the business process means that detecting the functional errors on the process and the risks on changing rules. These risks may be exceptions raised at run time like infinite loop and process non-termination, services deny.

The verify of functioning of the business process by analyzing the graph of rules based process is not scope of this paper . We are interested here by the formal verification of the rules based process. Our verification consists of two steps : the transformation of ECATE/EUCATE rules into a Petri net, and verification of such Petri net

The steps of such verification are summarized in the following diagram:

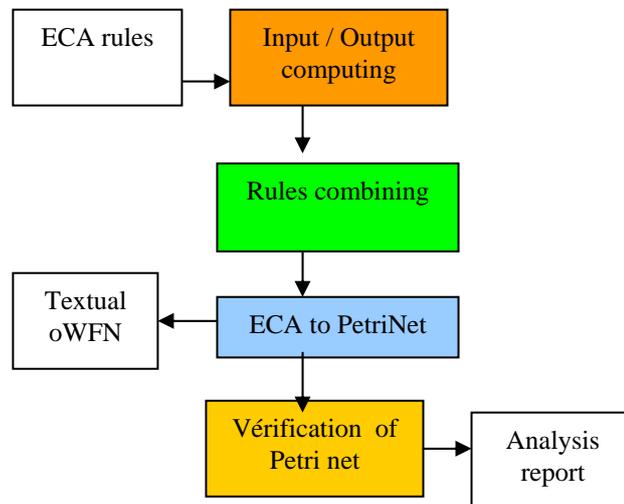


Fig. 3. Verification Environment

The oWFN (open WorkFlow) is a kind of Petri nets in order to verify the controllability property. The transformation of ECA rules to Petri Net allows to verify rules based business process and to exploit technical verification of Petri nets in the framework of business processes.

The transformation steps are as follows:

- a. Structuring the used ECA rules
In our case, the used rules must be simple: the two sides must contain only one variable, in order to have reduction during the following steps. The complex rules can be represented by simple sub-rules.
- b. Research inputs and outputs
The inputs are variables with non beginning and not having predecessors, the outputs are variables non final and have no successors.
- c. Combining rules
This step consists to reduce simple rules number applying the following principles:
Each left side of the rule must contain one input variable and one variable or one input and several variables.
Each right side of the rule must contain one output and one variable or one output and several variables.
A rule doesn't contain an input and output in the same time. All the rules must respect the previous principles and are able to be combined.
- d. Rules from ECA to Petri Net
Each rule becomes transition. The event and action becomes places.
- e. Verification of Petri net
We verify the properties of Deadlock, Live lock, Boundness and controllability on the produced Petri net using tools of Petri net verification as Lola[12] and Fiona[13]. The detail of the verification is not given in this paper.

5 Related work

The authors in [6] believe that it is important to couple WS-BPEL with a model for expressing authorization policies and constraints, and a mechanism to enforce them. They see that it is important that such an authorization model be high-level and expressed in terms of entities that are relevant from the organizational perspective. They propose an extension of WS-BPEL syntax with an authorization model that also supports the specification of a large number of different types of constraints. But, BPEL is not flexible.

[14] propose a flexible access control policies through the use of three classes of restraint rules in active cooperation: authorization rules, assignment rules and activation rules. A restraint rule consists of prerequisite conditions and a consequence. Each condition is in form of one or more weighted atomic conditions combined through logic operation connectors.

To enable a dynamic business process management, the authorization policies in [15] are expressed in an SQL-like language which can be rewritten into query sentences for execution. The framework proposed supports dynamic integration and execution of multiple access control policies from disparate enterprise resources.

In order to support the authorization policy development, [16] introduce a simple and readable authorization rules language implemented in a Ruby on Rails [17] authorization plug-in that is employed in workflow application. Ruby on Rails is a Web development framework that supports agile development and draws from the meta-programming features of the programming language Ruby.

Authors in [18] propose active role-based access control model to assign permissions to users in real time and automatically. They combine the role-based access control model with the active database. They exploit the characteristics of the active database to assign roles to users based on the event trigger, user and environmental conditions, and to assign permissions to roles using the RBAC model.

6 Conclusions and future work

In this paper, we present a flexible integration of security concern in a rules based business process modeling. We are proposed a new ECA based rules to govern the functional and security business rules in multi-concerns view. The approach is thoroughly illustrated using an order purchase example.

How to manage this flexibility? What are the relationships between the rules of different concerns? How to recognize and heal the functional exceptions in rules based process? How to verify this rules based business process? Some answers for these questions will be subjects of future works.

7 REFERENCES

1. Jakoubi S., Tjoa S., Goluch G., Quirchmayr G., A Survey of Scientific Approaches Considering the Integration of Security and Risk Aspects into Business Process Management, in DEXA '09 Proceedings of the 20th International Workshop on Database and Expert Systems Application, IEEE Computer Society Washington, DC, USA ©2009.
2. Rodríguez A., Fernández-Medina E., Piattini M., Towards a UML 2.0 Extension for the Modeling of Security Requirements in Business Processes, In: Proceedings of Trust and Privacy in Digital Business (TrustBus 2006), Springer, 2006
3. Papazoglou M.P., Traverso P., Dustdar S., and Leymann F, Service-Oriented Computing: a Research Roadmap. *Int. J. Cooperativ Inf. Syst.*, 17(2):223–255, 2008
4. Bertino E., Crampton J., Paci F., Access Control and Authorization Constraints for WS-BPEL, *icws*, pp.275-284, IEEE International Conference on Web Services (ICWS'06),2006.
5. Business Rules Group. Defining Business Rules - What Are They Really? www.businessrulesgroup.org, 2005.
6. Wan-Kadir W.M.N. and Loucopoulos P., Relating Evolving Business Rules to Software Design. *Journal of Systems Architecture*, 2003.

7. Ahn G.-J., Sandhu, R., Kang M., and Park J., , Injecting RBAC to secure a web-based workflow system. In Proceedings of the 5th ACM Workshop on Role-Based Access Control, pages 1–10, 2000.
8. Kazhamiakin R., Benbernou S., Baresi L., Plebani P., Uhlig M. and Barai O., Adaptation of Service-Based Systems Service Research Challenges and Solutions for the Future Internet, Lecture Notes in Computer Science, Springer-Verlag, 2010.
9. Aoumeur N., Barkaoui K., Saake G., , A multi-dimensional architectural approach to behavior-intensive adaptive pervasive applications, in ISWPC'09 Proceedings of the 4th international conference on Wireless pervasive computing, IEEE Press Piscataway, NJ, USA, 2009.
10. Ruopeng L., Sadiq S., a Survey of Comparative Business Process Modeling Approaches, in BIS'07 Proceedings of the 10th international conference on Business information systems Springer-Verlag Berlin, Heidelberg ©2007.
11. Boukhebouze M., Amghar Y., Benharkat A., Maamar Z., Rule-based Approach to Model and Verify Flexible Business Processes, International Journal of Business Process Integration and Management: IJBPIIM, 2011.
12. Massuthe P., Weinberg D., Fiona: A Tool to Analyze Interacting Open Nets. AWPN 2008: 99-104
13. Schmidt K, LoLA: A Low Level Analyser, Application and Theory of Petri Nets 2000: 21st International Conference, ICATPN 2000, Aarhus, Denmark, June 2000. Proceedings, volume 1825 of Lecture Notes in Computer Science, pages 465–474, June 2000. Springer-Verlag
14. Yuqing Sun , Bin Gong , Xiangxu Meng , Zongkai Lin , Bertino E., , Specification and enforcement of flexible security policy for active cooperation, Information Sciences: an International Journal, July , v.179 n.15, p.2629-2642,2009.
15. Cao J., Chen J., Zhao H., Minglu Li, A policy-based authorization model for workflow-enabled dynamic process management, Journal of Network and Computer Applications, March, v.32 n.2, p.412-422, 2009.
16. Bartsch S., Sohr K., Bormann C., , Supporting Agile Development of Authorization Rules for SME Applications, Collaborative Computing: Networking, Applications and Worksharing, 4th International Conference, CollaborateCom November 13-16, Orlando, FL, USA, 2008.
17. Ruby on rails, website: <http://rubyonrails.org/>
18. Mei-Yu Wu, Chih-Kun Ke, Jung-Shin Liu, "Active Role-based Access Control Model with Event-Condition-Action Rule and Case-Based Reasoning", JCIT: Journal of Convergence Information Technology, Vol. 6, No. 4, pp. 328 - 339, 2011

Security Requirements Analysis of Web Applications using UML

Salim Chehida ¹, Mustapha kamel Rahmouni ²

¹ Department of Informatics, University of Mostaganem, Algeria
salimchehida@yahoo.fr

² Department of Informatics, University of Oran Es-Senia, Algeria
kamel_rahmouni@yahoo.fr

Abstract— The security problems of the Web applications (processes and data) take a great importance nowadays. The transactions made through the network can be intercepted, more especially since adequate legislation has not yet been fully enforced on the Internet. The functional specification of the Web applications is not sufficient, the design and the realization of these systems must take into account the various security requirements. Taking into account the various security constraints (Availability, Authentication, Integrity, Secrecy, Non-Repudiation, etc.) in the modeling process constitutes one of the principal challenges for the designer of these systems. UML is the standard language for the modeling of the multiple views of systems by using the various mechanisms of extension. In this paper we describe our return on experiment concerning the modeling of the Web applications in order to analyze the security requirements of these systems by proposing new extensions of UML and a case study as illustration.

Keywords: *Web applications, Computer Security, Modeling, and UML.*

1. Introduction

If the generalization of the Internet connections offers new and promising possibilities, it also introduces a certain number of risks which we should be aware of, weigh their possible consequences, and take adequate measures. A company communicates today with its subsidiaries, its partners, and that induces a massive opening to information. The Web applications are thus increasingly likely to be the subject of various disturbances such as congestions, malicious accesses and attacks. The number of security problems has recently drastically increased and, unfortunately, this ascending curve certainly would not dip. In 2003, according to a study published [4], the damage caused by security incidents can amount, in Europe, between 0,2 and 0,5% of the sales turnover.

Security aspects of systems should be analysed and modeled during the entire system development process, so that the violated security requirements can be identified in the early stages of the development process. [17] UML is a standard language that is

used to visualize, specify, build and document a software system. This language is not adapted to all the system views: it uses extension mechanisms to model various aspects of the system.

This study proposes new extensions of *UML language* for the modeling of security requirements of *Web applications*, these extensions relate to the various phases of development (Specification, Logical analyze and technical structure). Firstly, we present the new vision of the computer security which makes it possible to treat the security constraints in the level of the development process, we explain after the *UMLsec* version; a whole of UML profiles proposed by Jürjens (Munich University of Technology) for security on the level of the conceptual models, and finally, we present the new extensions proposed and a case study of the *COMEX* system, an Information System of Commercial Management for a Harbor Company.

2. Security at the development process

Security of Information System consists in identifying the vulnerabilities, evaluating the threats and determining the risk which vulnerability allows threat given to be carried out, it uses a methods, techniques and tools to protect the resources of information system in order to ensure the availability of the services, the confidentiality and the integrity of information.

- The availability of the services: the services and information must be accessible to the authorized entity when they need some.
- Confidentiality of information: information does not belong to everyone; only can reach it those which have the right of it.
- Integrity of information: information (files, messages...) can be modified only by the authorized entity.

Adding security solutions to a system that has already been functionally realized is very difficult, and can make the system instable. The security requirements should then be *integrated at the design stage*, so that they can be identified with the first parts of development process. The posteriori security of critical systems (Firewall, Antivirus, etc.) does not constitute the best security policy. We think that the development of a security policy must be done at the same time than the functional design stage, and that the final model must integrate, at the same time, the functional and security specifications. The security of the critical systems must start with the development of a “model” which would represent: what are the threats? What do we have to protect? Why? This new approach makes the transformation of the security concept from a posteriori vision to a priori vision (at the development process level). “Security concern must inform every phase of software development, from requirements engineering to design, implementation, testing and deployment”. [11] This central activity consists in foreseeing the threats and the vulnerabilities induced by the use of the system.

3. UMLsec Profiles

UMLsec is an extension of UML proposed by J.Jürjens that includes profiles for secure systems development. Stereotypes¹ are used to formulate the security requirements. The tables below show some *UMLsec* stereotypes with their labels². [9]

TABLE I. UMLSEC STEREOTYPE

Stereotype	Description	Label
Secure dependency	Package to identify the secure dependency relations in the static models	
Secure links	Package to identify the secure dependency relations between the system's components	
Data security	Package to specify the critical objects and the various properties of security on the data	secrecy, integrity, high, fresh
Fair exchange	Package to represent the fair exchange scenarios in the electronic transactions	start, stop
No down – flow	Package to secure the information flow	high
Provable	Package to express non- repudiation in the electronic transactions	action, cert
Guarded access	Package to control the access to the objects	
Internet	Internet connection	
Encrypted	Encrypted connection	
LAN	Local area network connection	
Secrecy	Confidentiality of dependence	
Integrity	Integrity of dependence	
Guarded	Guarded object	guard
LAN	Local area network node	
Smart card	Smart card node	

TABLE II. UMLSEC LABELS

Label	Description
Secrecy	Data which should be secret
Integrity	Data which should not be modified
Fresh	Data which should not be re-used
Start	Initial state
Stop	Final state
Action	Provable action
Cert	Certificate activity
Guard	To guard an object

4. UML Security Extensions

The extensions which we have just proposed concern different views of the system; the *secure context model*, *security cases* and *critical scenarios* for the specification of the security requirements, the *secure interactions of objects* and the *security*

¹ The stereotypes make it possible to extend the semantics of the modeling elements and to define new UML elements classes.

² A label or marked value is a pair (name, value) which adds a new property to UML modeling element.

constraints on the data for the logical view and finally the *protected hardware configuration* for the technical view of the system. As an illustration, examples of the *COMEX* system will be presented.

4.1. Secure Context Model

Many authors, like G.Booch in [8] or more recently P. Roques and F.Vallee in [13], recommended the use of collaboration diagrams to represent, in a synthetic manner, the various functional requirements of a system. After the definition of security conditions, we can present the various security requirements of Web applications on a diagram, which can be called *secure context model*. This model consists in defining the various expected security services of the system considered as a black box. The collaboration diagram is used in the following way:

- The system is represented by a central object; this object is surrounded by other objects symbolizing the various actors.
- The objects are connected by bonds; on each bond are shown output messages which represent the various security services provided by the system.

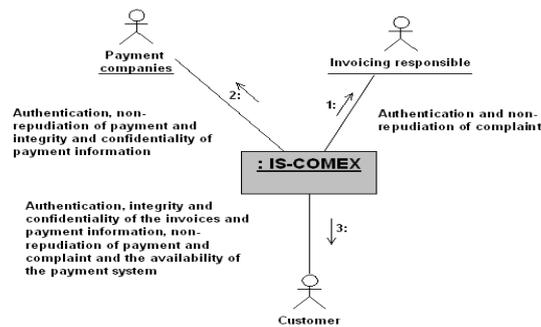


Figure 1. Example of a secure context model

4.2. Security Cases Model

In this model, we are interested in the specification of the Web applications requirements in terms of security. To do that, we use the *use cases* in a different manner by introducing the concepts of *security cases* and *security cases model*. The *security cases model* is used to structure the security services provided by the system (always considered as a black box) for the various actors as a set of *security cases*. A *security case* represents a security service returned by the system for one or more actors. For example: to verify the identity of user, to ensure the integrity and the secrecy of the exchanged information, to ensure the non-repudiation of transactions, etc. A *security case* specifies an awaited system behavior to meet security needs without imposing the realization mode of this behavior. It makes it possible to describe what the future system will have to do in terms of computer security without defining how to do it. *Security cases* are distinct from *use cases*; they do not produce a functional added value but they indeed cover all security services that a user needs.

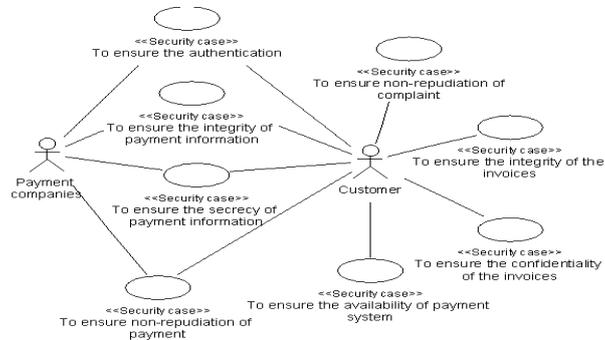


Figure 2. Example of a security cases model

4.3. Critical Scenarios

The critical scenarios consist in describing and representing the critical interactions or actions using the various services of security specified by the *security cases*. A critical scenario represents a particular succession of sequences (interactions between the actors and the system) which involves a risk in terms of computer security. To underline this risk, we will associate the various constraints of security on the interactions between the system considered as a black box and the various actors. For example: the scenarios which ensure the non-repudiation in the electronic transactions, the scenarios which specify the interactions with exchange of critical information, etc. We used the sequence diagram which makes it possible to better visualize the interactions.

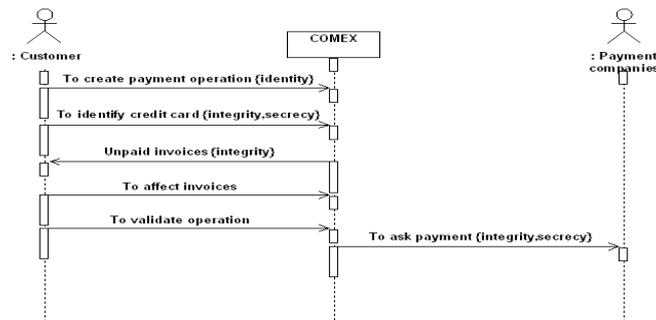


Figure 3. Example of critical scenarios model

We used three constraints³ for the interactions between system and actors:

- The {secrecy} constraint to ensure the secrecy of the interactions.
- The {integrity} constraint to ensure the integrity of the interactions.
- The {identity} constraint to ensure the identity of the parties during the execution of interaction action between an actor and the system.

³A constraint is a semantic relation between UML modeling elements. Each constraint is indicated between braces and is placed close to the element (stereotyped or not).

4.4. Secure Interactions of Objects

After the identification of the classes and objects of the system (the Static Model), we now replace the system by a collaboration of objects. A scenario of secure interactions of objects represents an ordered set of messages exchanged between objects (instances of classes and actors) with the specification of the security constraints on these messages. A message represents the specification of a one-way communication between objects which transports information and whose goal is to generate a reaction from the receiver. It can include parameters which transfer values from the transmitter to the receiver. [15] For the representation of secure interactions of objects, we used the sequence and the collaboration diagrams of the UML.

- The {secrecy} constraint to ensure the secrecy of the messages;
- The {integrity} constraint to ensure the integrity of the messages;
- The {identity} constraint to ensure the identity of the transaction parties.

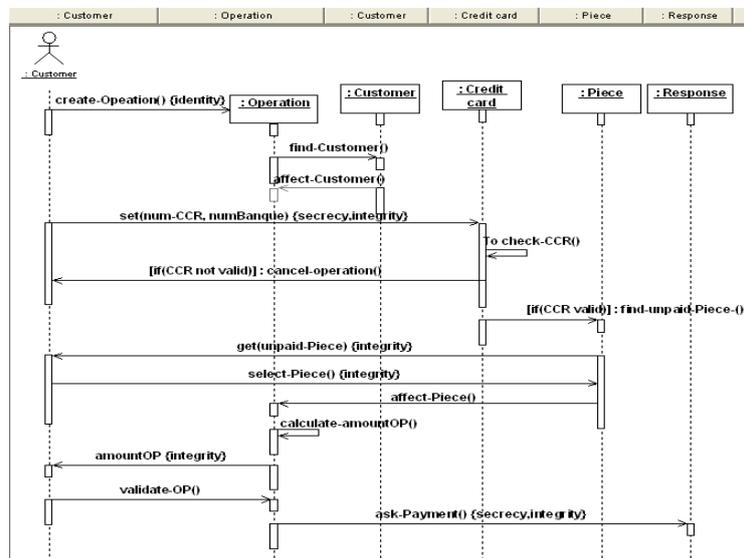


Figure 4. Example of the secure interactions of objects model

4.5. Data Security

The set of security cases discovered through the specification of security constraints guides all the dynamic views, by representing the critical scenarios, the collaborations and the interactions of objects with the sequence diagrams. In order to benefit from the security analysis phase, it is necessary to update the class diagram by adding security constraints on the data. The class diagram is viewed as the most important diagram in the object methods. After having developed the class diagram, we will define security constraints on the attributes and the operations starting from the critical scenarios represented on message flows between objects. The {secrecy} constraint specifies the data being confidential, the {integrity} constraint is used to ensure the integrity of the data and the {identity} constraint indicates that only the authorized parts can reach the data.

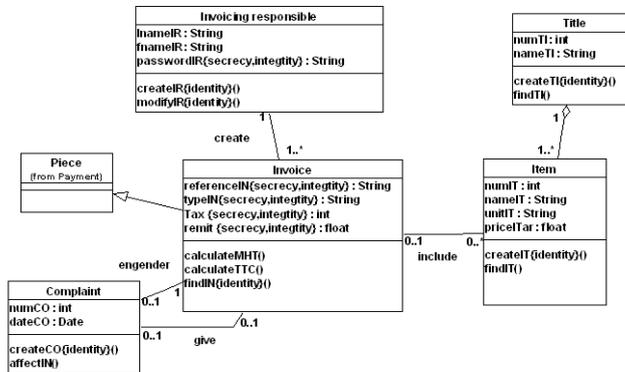


Figure 5. Example of security constraints on the data

4.6. Protected Hardware Configuration

The protected hardware configuration model consists in expressing the implementation constraints at the physical level represented by nodes and connections, which are the various types of machine connected by various means with the integration of the prevention tools (Firewall, IDS, etc) to implement the security constraints. This model also allows representing the types of connections (LAN, VPN, etc) between the various nodes. The deployment models and hardware configuration models are both expressed by using a deployment diagram. However, they do not quite express the same description level. The hardware configuration model is used to express the constraints of implementation at the physical level; it consists of the nodes and the physical connections of the system. On the other hand, the deployment model expresses the physical distribution of the system's functions and permits to justify the localization of the data bases and working environments.

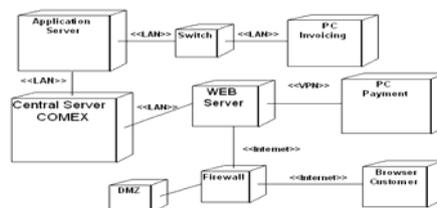


Figure 6. Example of a protected hardware configuration model

5. Conclusion

In this paper, we have tackled the highly vast subject of computer *security*, while concentrating on security of Web applications at the model level. It is a transverse approach, where the security concept is being included in the modeling of Web applications, and where the UML extensions are able to help master the control of security. The security model is a representation of security derived from a “vision of the world”. The model defines what must be defended (information flow), against what (threats)

and why (sensitivity of information). It can be more or less complete, but in all cases it emphasizes just about the risks from where we can deduct a security policy. [6]

UML is not a closed notation: it is generic, extensible and configurable by the user. Where necessary, we can use extension mechanisms. This paper presents new profiles of UML for the modeling of security aspects. The *secure context model* and the *security cases model* for the specification of the security needs, the *critical scenarios model* consist in describing the interactions or the actions which involve a risk and the *secure interactions of objects model* for the specification of the security constraints on the messages exchanged by objects. In the analysis model, we defined security properties on the data. At last, for the modeling architecture, the *protected hardware configuration model* allows to express the implementation constraints at the physical level with the integration of the prevention tools in order to fulfill the security requirements. The important points which remain to be developed are: the realization of attack simulations on protected UML models in order to validate these models and to correct the security weaknesses found, and the integration of these extensions in a development process.

References

- [1] M. LOULOU ALOULOU, « Approche Formelle pour la Spécification, la Vérification et le Déploiement des Politiques de Sécurité Dynamiques dans les Systèmes à base d'Agents Mobiles », Thesis of doctorate, University of Bordeaux I, (2010).
- [2] A.Abou el Kalam, « Modèles et politiques de Sécurité pour les Domaines de la Santé et des Affaires Sociales », Doctorate Thesis, Institut National Polytechnique of Toulouse ,(December 2003).
- [3] Réda KADRI, « Une approche pour la Modélisation d'Applications Web à base de Composants Logiciels », Thèse de Doctorat, Université de Bretagne Sud, (2009).
- [4] <http://www.cert.org/stats>.
- [5] C.Larman, « UML et les Design Patterns », Campus Press, (2002).
- [6] CNRS, “Computer security: number 31,... , 35”, Site: <http://www.cnrs.fr/Infosecu>, (2001).
- [7] E.Maiwald, « Sécurité des réseaux », Campus Press, (2001).
- [8] G.Booch,“Object Solutions:Managing the Object-Oriented Project”,Addison Wesley, (1996).
- [9] J. Jurjens, « Secure Systems Development with UML: a Foundation », Thesis of doctorate, Munich University of Technology, (2003).
- [10]N. Mayer and J Humbert. «La gestion des risques pour les systèmes d'information ». Magazine MISC n°24. ISSN: 1631-9036. (April-May 2006).
- [11]P. Devanbu, «Software Engineering for Security: A Roadmap », (2000).
- [12]P. Mell, K. Scarfone, S. Romanosky, « A complete guide to the Common Vulnerability Scoring System, version 2.0 », Forum of Incident Response and Security Team. (2007).
- [13]P. Roques et F. Vallee, « UML en action », Eyrolles, (2002).
- [14]P. Roques, « Modéliser un site e-commerce», Eyrolles, (2002).
- [15]P.Roques, « UML par la Pratique », Eyrolles, second edition (2003).
- [16]Robert Longeon et Jean-Luc Archimbaud,«Guide de la sécurité des systèmes d'information a l'usage des directeurs », CNRS, Site : <http://www.cnrs.fr/Infosecu>.
- [17]S. Meng , « Security Requirements Analysis and Modeling of Distributed Systems », Thesis of Master, Munich University of Technology, (2004).
- [18]BLOCH, Laurent, WOLFHUGEL, Christophe. « Sécurité informatique : principes et méthodes »,Eyrolles, (2007).

Development of RSA with random permutation and inversion algorithm to secure speech in GSM networks

Khaled Merit¹ and Abdelazziz Ouamri²

¹ National Institute of Telecommunications and Information and communication technologies,
INT&TIC Oran, Algeria
merit1984@gmail.com

² Signals and Images Laboratory, University of Sciences and Technology of Oran
USTO Oran, Algeria
ouamri@univ-usto.dz

ABSTRACT

Global System for Mobile Communications (GSM) is one of the most commonly used cellular technologies in the world. One of the objectives in mobile communication systems is the security of the exchanged data. GSM employs many cryptographic algorithms for security like A5/1, A5/2 and A5/3. Even so, these algorithms do not provide sufficient level of security for protecting the confidentiality of GSM. Therefore, it is desirable to increase security by additional encryption methods. This paper presents a voice encryption method called: "RSA with Random permutation and Inversion", based on current voice channel, which overcomes data channel's insufficiencies and solves the problem of penetrating the RPE-LTP vocoder by the encrypted voice. The proposed method fulfils an end-to-end secured communication in the GSM; insure a good compatibility to all GSM networks, and easy implementation without any modification in these systems.

KEYWORDS : SECURITY, GSM, SPEECH CHANNEL, RSA, SPEECH CODEC

1. INTRODUCTION

Security presents a very important axis in wireless communication systems. This is obviously because of the ubiquitous wireless medium's nature that makes it more susceptible to attacks. Any eavesdropper can get over to whatever is being sent over the network through the wireless medium. In addition, the presence of communication does not uniquely identify the originator. Besides this, any eavesdropping or tapping cannot even be detected in a medium as ubiquitous as the wireless medium which makes the latter situation even worse. Hence, security plays a fundamental task for the successful operation of a mobile communication system.

To secure data in GSM, encryptions and mechanisms to grant it are obligatory. In this paper, a new approach has been proposed which includes extra encryption RSA with random permutation and inversion algorithm. GSM employs stream ciphers for encryption which requires the data to be in its binary form [1]. Our encryption technique processes directly on symbols without passing to the bit level. In addition, this technique does not need any hardware; it is totally based on software. This technique is much simpler than existing techniques, thus a more robust and efficient system is achieved. The following sections discuss the proposed scheme: Section 2 enumerates the security requirements of mobile networks. Section 3 gives a quick overview of existing GSM encryption algorithms and a variety of attacks on these

algorithms. Section 4 illustrates the proposed End-To-End encryption method. Section 5 the simulation results, and Finally, concludes this paper by summarizing the key points and proposing related suggestions.

2. SECURITY REQUIREMENTS OF MOBILE NETWORKS

Security has become an essential topic in current mobile and wireless networks. As the security procedures for such networks elevates, the tools and techniques used to attack such networks also increases. Wireless communications security is the measures or methods used to protect the communication between certain entities. To protect the entity from any third party attacks, such as revealing a particular identity, data modification or data-hijacking, eavesdropping, impersonating an identity, Protection mechanisms are used. Devoted technologies for securing data and communication are mandatory in wireless networks, which vary according to the category of wireless technology deployed. Security in mobile networks handles a diversity of issues, from authenticating a user accessing a network, to data integrity and data encryption. GSM, like a lot of other systems with huge users' numbers, contains numerous precious resources that need protection against misuse and deliberate attacks. This section highlights the GSM Network precious resources, which are important to protect for the best of the system's shareholders.

The facilities listed below are provided to insure security to the users of the communication networks [3]-[7]:

Confidentiality: This means that the transmitted information is only disclosed to the authorized parties. Sensitive information disclosed to an adversary could have severe consequences.

Integrity: This assumes that a message is not altered in transit between sender and receiver. Messages could be corrupted due to network malfunctioning or malicious attacks.

Authentication: Authentication guarantees the identity of the entity with which communications are established, before granting it the access to the resources of the network. In the absence of authentication mechanisms, an attacker could masquerade as a legitimate entity and attempt to violate the security of the network.

Nonrepudiation: This means that the source of a message cannot deny having sent the message. An attacker could generate a wrong message that appears to be initiated from an authorized party, with the aim of making that party the guilty one. If non-repudiation is guaranteed, the receiver of a wrong message can prove that the originator has transmitted it, and that, therefore, the originator misbehaved.

Access control: Access control means that only authorized parties can be allowed to access a service on the network, use a resource, or participate in the communications;

any other entity is denied access. The access control assumes the authentication of the entity trying to get access to the network.

Network availability: Availability ensures that all resources of the communications network are always utilizable by authorized parties. An attacker may launch a Denial of Service (DoS) attack by saturating the medium, jamming the communications, or keeping the system resources busy in any other way or by any other means. The aim here is just to slow down or stop authorized parties from having access to the resources, thereby making the network unusable.

3. GSM encryption and attacks

In GSM, A5 stream cipher is used [5]. Versions A5/1 and A5/2 were kept secret for a long period of time. Briceno et al reverse-engineered A5/1 and A5/2 from a GSM handset and published them. After which, attacks were rapidly found for these algorithms. The principal problem is the small key length of the session key Kc. The actual length of Kc is 64 bits. However, only 54 bits are effective. Even though this key size is sufficiently big to protect against real-time attacks, the hardware state available today makes it possible to record the packets between the mobile subscriber and the BTS and then decrypt them afterward [6].

Biryukov et al. found a known-key stream attack on A5/1 that needed about two seconds of the key stream and recovers Kc in a few minutes on a PC after a large pre-processing stage. Barkan et al. [5] have proposed a ciphertext-only attack on A5/1 that also recovers Kc using only four frames; the problem was its complexity. A5/2 was also cracked and proved to be totally vulnerable. The attack needed very few pseudo random hits and only 216 steps [5].

A new security algorithm, known as A5/3 provides users of GSM mobile phones with an even higher level of protection against eavesdropping than they have already. A5/3 is based on the Kasumi algorithm, specified by 3GPP for use in 3G mobile systems. The A5/3 encryption algorithm particularly provides signaling protection to protect important information such as telephone numbers as well as user data protection to secure voice calls and other user generated data. This algorithm were so far assumed to be stronger than A5/1 and A5/2, but the Biham et al attack shows that the key can be obtained quickly without applying exhaustive key search.

4. END-TO-END encryption method

4.1. Review of GSM Voice Transmission

The process of GSM voice channel transmission illustrated in Figure 1, includes five components: A/D module, RPE-LTP vocoder, channel coding/decoding module, wireless encryption /decryption module, and GMSK modulation/demodulation module. The wireless encryption/decryption module only works on wireless channel. So it cannot provide end-to-end secure communication in GSM system.

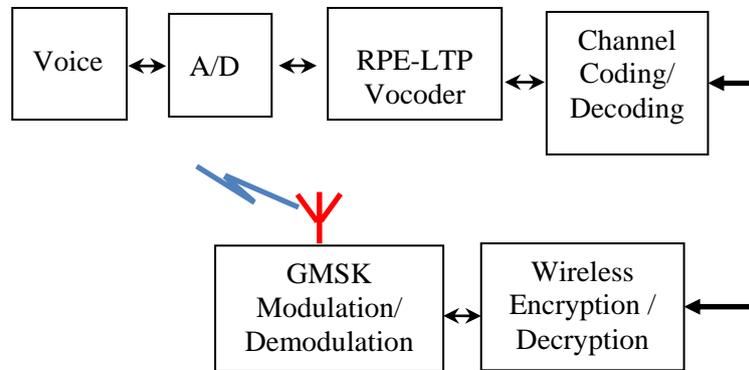


Figure 1. GSM voice transmission process.

4.2. Review of GSM Voice Transmission

The RPE-LTP [4]-[2] is an important algorithm in the field of voice encoding. It is not only used in GSM, but also used in Internet.

At the transmitter, the processing in the RPE-LTP Encoder includes pre-processing, LPC analysis, short-term analysis filtering, Long-Term Prediction and Regular Pulse Excitation sequence coding. The details is described as following: first Encoder samples original digital voice signal at 8kHz sampling rate, and removes the direct current component, then it can make use of FIR filter to pre-emphasis the high frequency. Secondly, LPC analysis takes every 160 sample points (20ms) as one frame and figures out 8 logarithm acreage ratio parameter for each frame. Short-time analysis filter produces LPC residual signal. It removes redundancy farther coding with RPE-LTP, and outputs 260 bits coding every frame at last. At the receiver, it practices a reverse processing and rebuilds the original speech signal.

4.3. Voice Encryption Method

In a general way, encryption/decryption module is put before the RPE-LTP vocoder, which is easy to implement in MT (Mobile Terminal). But it cannot accomplish the end-to-end secured communication, and need to be modified in BS (Base Station). So a novel voice encryption/decryption method is proposed based on voice channel, which can fulfill the end-to-end secured communication without any modification in BS. The novel voice encryption/decryption scheme is depicted in Figure 2.

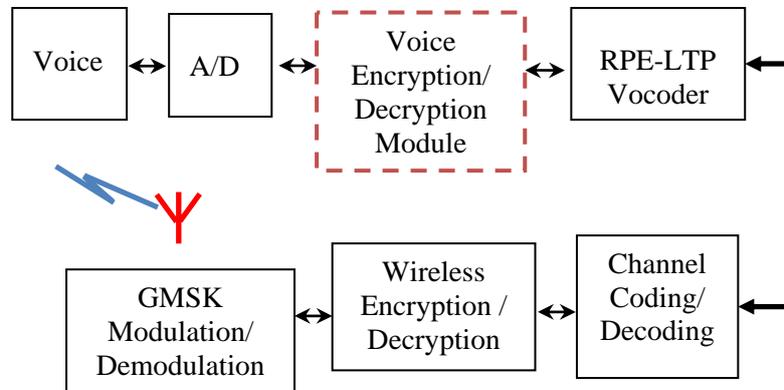


Figure 2. Voice encryption module access point in MT.

In Figure 2, the new voice encryption/decryption module is inserted between A/D and RPE-LTP vocoder in MT. The coming voice signal from the A/D module would firstly arrive to the newly-added Voice Encryption/Decryption module and finishes the encryption. After that, it is sent to the RPE-LTP vocoder. Hence, this encryption method must penetrate the RPE-LTP vocoder and have ideal encryption intensity. Simultaneously this encrypted signal can be recovered to get the original understandable speech at the receiver. This new voice encryption method is a kind of signal source encryption technology, so it could achieve the end-to-end secured communication.

4.4. Encryption algorithm

Principle of the encryption algorithm:

For implementing encryption algorithm, we follow the following steps:

- Decomposition of a speech signal in sub-frame, each frame is represented by an index.
- Encrypting data with inversion and random permutation algorithm, which gives the permutation indexes.
- Encrypting these indexes with RSA algorithm.
- Used these indexes to decrypt the signal.

In this paper, we propose an algorithm that combines between permutation and inversion of the voice signal samples, giving as a result the permuted indexes. These indexes are processed in an encryption/decryption module by RSA algorithm, and finally, these encrypted permuted indexes are added to the compressed encrypted voice signal samples after the RPE-LTP module. So it has a good recovery character to RPE-LTP vocoder, and its encryption intensity also can meet the special requirement. The algorithm is mainly intended to make the encrypted voice signal to be similar to the natural human voice signal, and can penetrate the RPE-LTP vocoder, and then it can execute all the encryption and decryption process. (See Figure 3).

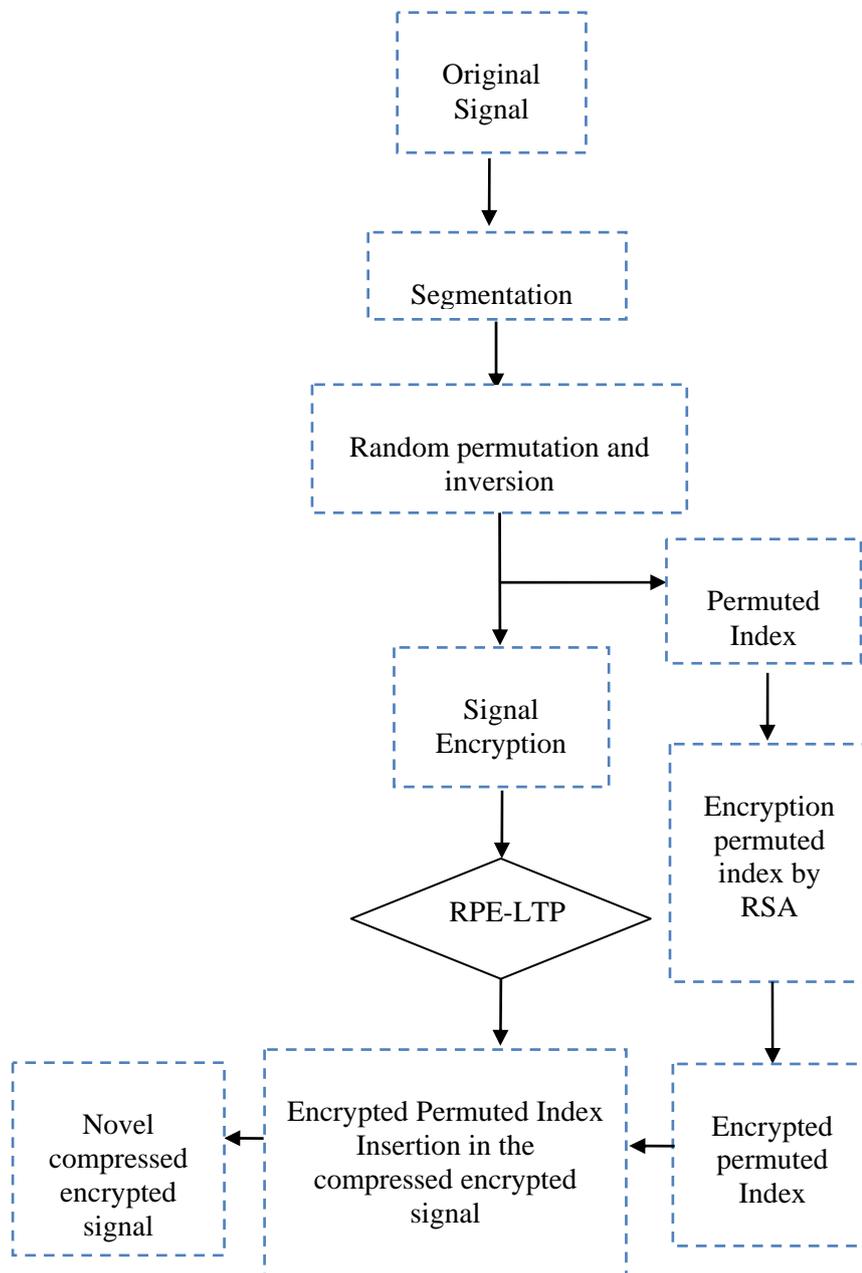
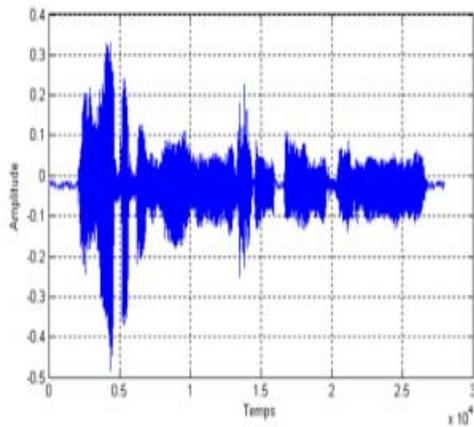


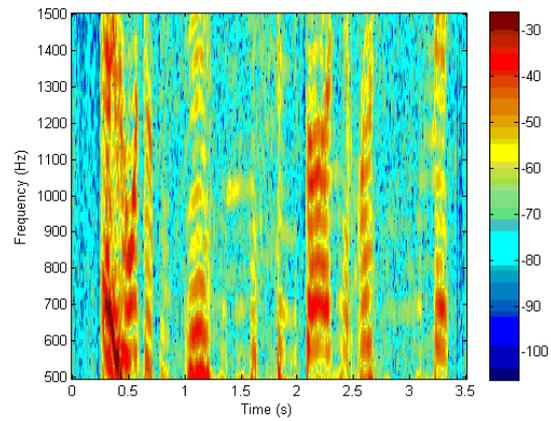
Figure 3. Encryption chart

5. Simulation results

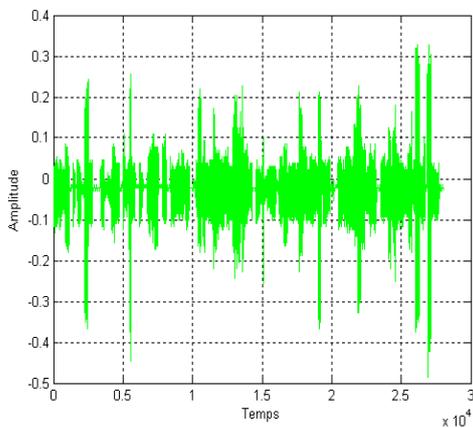
This section presents the results for the proposed method adopted in Section 3. This section also discusses the obtained results from implementing the system. In order to implement such a system, one must go through several steps which were described in details in the preceding sections. The implementation for this simulated project is written by MATLAB.



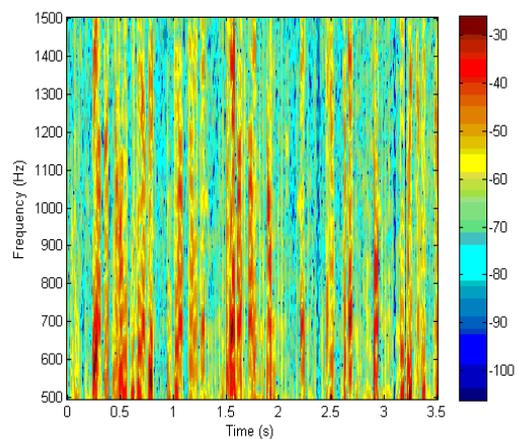
a.1 Original temporal signal



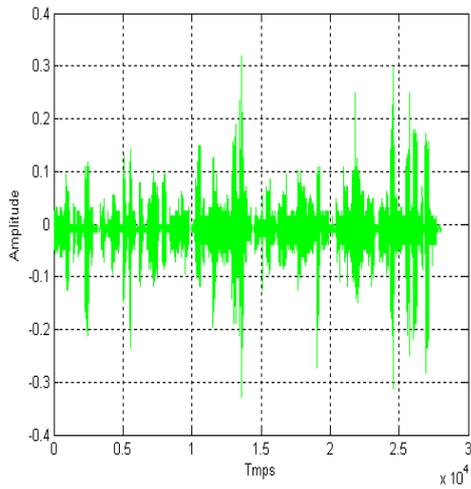
b.1 Spectrogram of original Signal



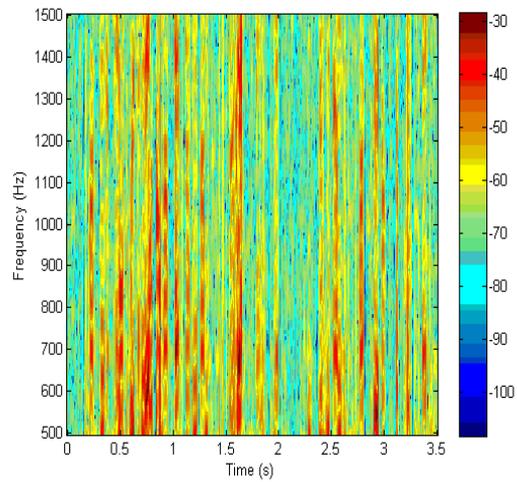
a.2 Encrypted temporal signal
before LPC



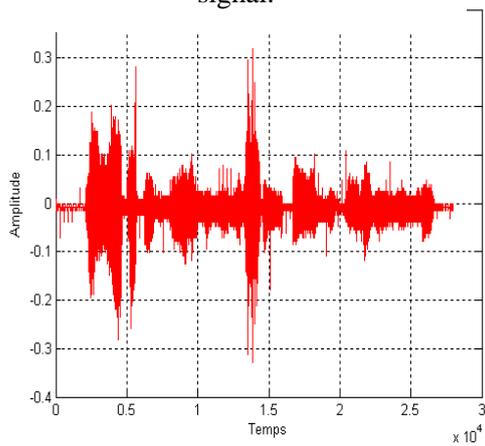
b.2 Spectrogram of encrypted
signal before LPC



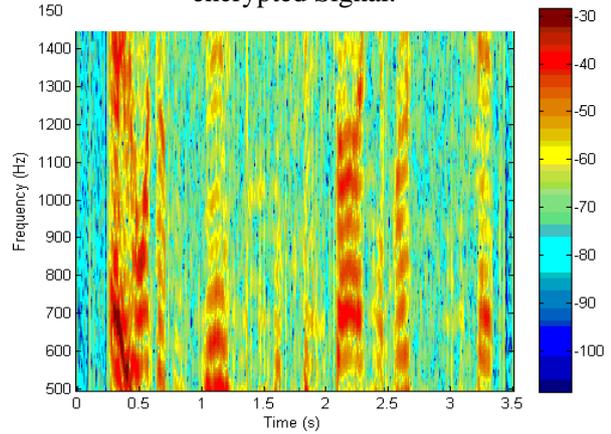
a.3 Synthesized encrypted temporal signal.



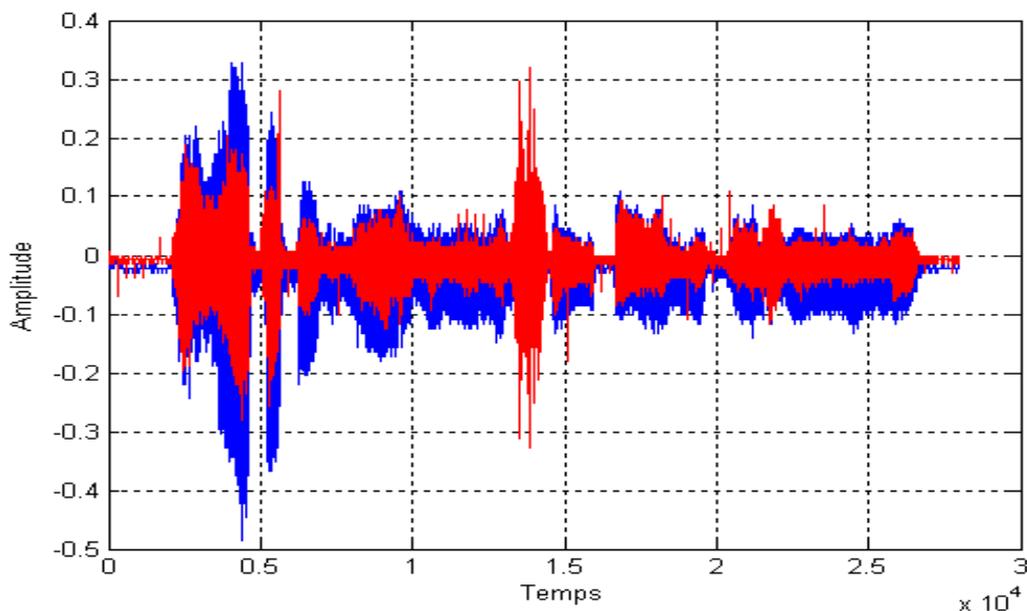
b.3 Spectrogram of Synthesized encrypted Signal.



a.4 Synthesized deciphered temporal signal.



b.4 Spectrogram of Synthesized deciphered Signal.



C. Comparison between the original signal and the synthesized signal.

In figure C, the margin between the original signal (blue) and the synthesized one (red) is due to the reduction of the bit rate imposed by the RPE-LTP module.

6. Conclusion

In this paper, a novel kind of encryption method is proposed to fulfill the end-to-end secured communication in the GSM voice channel. The new encryption method solves the problem that traditional encryption algorithms cannot be used in voice channel directly because of RPE-LTP vocoder requirements in GSM system. In addition, this encryption method has the advantages of suiting the RPE-LTP compression module requirements, good compatibility to GSM networks, and suitable implementation without any adjustment in current GSM signalling system. The algorithm presented in this paper is made by the RSA algorithm, but it can be done also by other encryption methods such as: DES, RC4 and AES.

REFERENCES

- [1] David G. W. Birch and Ian J. Shaw, "Mobile communications security private or public", IEEE, June 1994.
- [2] ETSI Speech processing functions General Description, GSM06.01-1999, Version 8.0.1 pp22-53.
- [3] H.Imai, M.G. Rahman, K. Kobara "Wireless Communications Security" ARTECH HOUSE 2006.
- [4] K.Hellwig, P.Vary, D. Massaloux, et al. Speech Codec for the European Mobile Radio System. IEEE Global Commu Conf, 1989: pp1065-1069.

- [5] Ross Anderson, Mike Roe "A5 - The GSM Encryption Algorithm", 1994.
- [6] Dr. S. Muhammad Siddique and Muhammad Amir "GSM Security Issues and Challenges" (SNPD'06) 2006.
- [7] Nouredine Boudriga "Wireless Communications Security",CRC 2010 by Taylor and Francis Group, LLC

Spam Detection System Combining Cellular Automata and Naive Bayes Classifier

F. Barigou*, N. Barigou**, B. Atmani***

Computer Science Department, Faculty of Sciences, University of Oran
BP 1524, El M'Naouer, ES-SENIA, 31 000 Oran, Algeria
{*fatbarigou, **barigounaouel, ***atmani.baghdad}@gmail.com

Abstract. In this study, we focus on the problem of spam detection. Based on a cellular automaton approach and naïve Bayes technique which are built as individual classifiers we evaluate a novel method combining multiple classifiers diversified both by feature selection and different classifiers to determine whether we can more accurately detect Spam. This approach combines decisions from three cellular automata diversified by feature selection with that of naïve Bayes classifier. Experimental results show that the proposed combination increases the classification performance as measured on LingSpam dataset.

1 Introduction

Spam is rapidly becoming a major problem on the Internet. Some recent studies shows that about 80% of the e-mails sent daily are Spam [8]. The major problem concerning spam is that it is the receiver who is paying in terms of its time, bandwidth and disk space. To address this growing problem of spam, many solutions have emerged. Some of them are based on the header of the email such as black list, white list and DNS checking. Other solutions are based on the text content of the message such as filtering based on machine learning. Many techniques have been developed to classify e-mails –for good review the reader can look, e.g., [9]. In a previous study [4], we proposed CASD (a Cellular Automaton for Spam Detection) a new approach to spam detection, based on symbolic induction by cellular automata [3]. Experiments show a very high quality of prediction when using stemming and Information gain as a features selection function [5]. A performance improvement is also observed over NB and KNN proposed in [2] on Ling Spam corpora. In this paper, our aim is to further improve the spam detection by adopting a combination strategy of classifiers. One technique to create an ensemble of classifiers is to use different feature subsets for each individual classifier. We believe that by varying the feature subsets to train the classifiers we can improve the performance of filtering, since it is possible to incorporate diversity and produce classifiers that tend to have high variety in their predictions. In a set of experiment to prove this, the same learning algorithm of CASD is trained over three different subsets of features and combined by voting, with a naïve Bayes algorithm.

The remainder of this paper is organized as follows; in section 2, we give an overview of the different types of strategies for classifier combination and we follow with the related work in combining multiple classifiers for spam detection. Section 3, first introduces the Naïve Bayes classifier and the CASD based cellular automaton

and then moves to the proposed combination approach. Experimental results are presented in section 4. Conclusions are finally drawn in section 5.

2 Background

A general overview of classifier combination is given in section 2.1. Some background on the spam detection using classifier combination is given in section 2.2.

2.1 Combining Classifiers

An ensemble of classifiers combines the decisions of several classifiers in some way in an attempt to obtain better results than the individual members. Such systems are also known under the names multiple classifiers, committees or classifier fusion. Numerous studies have shown that combining classifiers yields better results than achievable with an individual classifier. A good overview of different ways of constructing ensembles as well as an explanation about why ensemble is able to outperform its single members is pointed in [11].

An ensemble of classifiers must be both diverse and accurate in order to improve accuracy, compared to a single classifier. Diversity guarantees that all the individual classifiers do not make the same errors. If the classifiers make identical errors, these errors will propagate to the whole ensemble and so no accuracy gain can be achieved in combining classifiers. In addition to diversity, accuracy of individual classifiers is important, since too many poor classifiers can overwhelm correct predictions of good classifiers [7, 15].

In order to make individual classifiers diverse, many ensemble methods use feature selection so that each classifier works with a specific feature set. To contribute to this research, we propose to employ multiple classifiers, each making predictions based on subsets of features.

2.2 Spam detection using multiple classifiers

In the context of spam filtering, a number of ensemble classification methods have been studied. Sakkis et al. [13] combined a Naïve Bayes (NB) and k-nearest neighbor (k-NN) classifiers by stacking method and found that the ensemble achieved better performance. Carreras and Marquez [6] used boosting decision trees with the AdaBoost algorithm. Compared with two learning algorithms, the induction decision trees (DT) and Naïve Bayes, Adaboost clearly outperformed the above two learning algorithms in terms of the F1 measure. Rios and Zha [12] applied random forests, an ensemble of decision trees, using a combination of text and meta data features. For low false positive spam rates, RF was shown to be overall comparable with support vector machines (SVM) in classification accuracy. Also, Koprincha et al. [10] studied the application of random forests to Spam filtering. The LingSpam and PU1 corpora with 10-fold cross-validation were used, selecting 256 features based on either information gain or the proposed term-frequency variance. Random forests produced the best overall results. Shih et al. [14] proposed an architecture for collaborative agents, in which algorithms running in different clients can interact for the

classification of messages. The individual methods considered include NB, Fisher's probability combination method, DT and neural networks. In the framework developed, the classification given by each method is linearly combined, with the weights of the classifiers that agree (disagree) with the overall result being increased (decreased). The authors argued that the proposed framework has important advantages, such as robustness to failure of single methods and easy implementation in a network.

3 Proposed Framework

In this research, we propose an ensemble of classifiers diversified by both manipulating input data and using two different classifiers Cellular automaton CASD [4] and Naïve bayes approach. These two classifiers are given in section 3.1 and 3.2 while the design of the proposed combination is discussed in section 3.3.

3.1 Naive Bayes Classifier

Naïve Bayes (NB) which has been widely used for spam filtering [1,2, 13] is a simple but highly effective classifier. It uses the training data to estimate the probability that an instance belongs to a particular class. NB requires little storage space during both the training and classification stages; the strict minimum is the memory needed to store the prior and conditional probabilities. In our experiments, each message is represented as a binary vector (x_1, \dots, x_m) , where $x_i=1$ if a particular token X_i of the vocabulary is present, otherwise $x_i=0$.

From Bayes' theorem, the probability that a message with vector $\vec{x} = (x_1, \dots, x_m)$ belongs in category c (= spam or leftitimate) is: $P(c|\vec{x}) = \frac{P(c) \times P(\vec{x}|c)}{P(\vec{x})}$. NB classifies each e-mail in the category that maximizes the product $P(c) \times P(\vec{x}|c)$. The a priori probabilities $p(c)$ are typically estimated by dividing the number of training e-mails of category c by the total number of training e-mails. And the probabilities $P(\vec{x}|c)$ are calculated as follows: $P(\vec{x}|c) = \prod_{i=1}^m P(x_i|c) = \frac{Xc+1}{Nc+|vocabulary|}$ where Xc is the number of occurrences of token X in e-mails with label c , Nc is the total number of token occurrences in e-mails labeled c and $|vocabulary|$ is the number of unique tokens across all e-mails.

3.2 CASD : a Cellular Automaton for Spam Detection

CASD is a classifier which is built on the cellular automaton CASI [3]. Besides its high classification accuracy, CASD also has advantages in terms of simplicity, classification speed, and storage space [5].

Cellular automaton CASI (Cellular Automaton for Symbolic Induction) is a cellular method of generation, representation and optimization of induction graphs generated from a set of learning examples. It produces conjunctive rules from a Boolean induction graph representation that can power a cellular inference engine. This Cellular-symbolic system is organized into cells where each cell is connected

only with its neighbors (subset of cells). All cells obey in parallel to the same rule called local transition function, which results in an overall transformation of the system. CASI uses a knowledge base in the form of two layers of finite automata. The first one, called CelFact, represents the facts base and the second one, called CelRule, represents the rule base. In each layer, the content of a cell determines whether and how it participates in each inference step; at every step, a cell can be active or passive, can take part in the inference or not. The states of cells are composed of three parts; EF, IF and SF, and ER, IR and SR which are the input, internal state and output parts of the CelFact cells, and of the CelRule cells, respectively. The neighborhood of cells is defined by two incidence matrices called R_E and R_S respectively. They represent the input respectively output relation of the facts and are used in forward chaining.

- The input relation, noted iR_{Ej} , is : *if (fact $i \in$ Premise of rule j) then $iR_{Ej} = 1$ else $iR_{Ej} = 0$.*
- The output relation, noted iR_{Sj} , is : *if (fact $i \in$ Conclusion of rule j) then $iR_{Sj} = 1$ else $iR_{Sj} = 0$.*

The cellular automaton dynamics is implemented as a cycle of an inference engine made up of two local transitions functions δ_{fact} and δ_{rule} .

The transition function δ_{fact} which corresponds to the evaluation, selection and filtering phases is defined as: $(EF, IF, SF, ER, IR, SR) \xrightarrow{\delta_{fact}} (EF, IF, EF, ER + (R_E^T \times EF), IR, SR)$

The transition function δ_{rule} which corresponds to the execution phase is defined as: $(EF, IF, SF, ER, IR, SR) \xrightarrow{\delta_{rule}} (EF + (R_S \times ER), IF, SF, ER, IR, \overline{ER})$

3.2.1 Learning classifier system

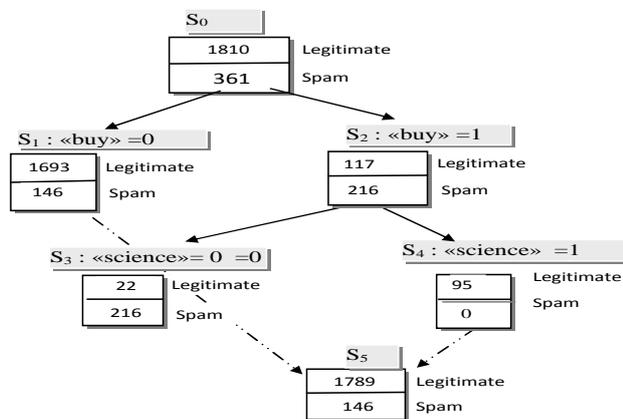


Fig.1. Example of an induction graph with only two terms.

During the learning phase, the Sipina method produces a graph. From this graph, a set of rules is inferred. They are in the form of "if condition1 and condition 2 and ...condition n then conclusion". For example, in the graph of Figure 1, if we look to partition number 2 at node number 1 (S1) we have the rule "if the term 'buy' is not

present then the email is legitimate", because the majority of emails (1693) which do not contain this term are legitimate.

The set of rules generated from induction graph are modeled by the CASI automaton as follows:

- The set of all conditions and conclusions are represented by a Boolean facts base called *CelFact*.
- The set of rules is represented by a Boolean Rule-based called *CelRule*.
- An input matrix R_E which memorizes conditions of the rules.
- and finally, an output matrix R_S which memorizes conclusions of the rules.

Forward chaining will allow the model to move from initial configuration to the next configurations G_0, G_1, \dots, G_n . The inference stops after stabilization with a final configuration. At this step the construction of cellular model is complete.

Table 1 presents the final configuration corresponding to the example of Figure 1. Three rules, represented by *CelRule* layer are deduced from the graph. The conditions and conclusions of these rules are stored in *CelFact* layer. The premises are the terms used in classification and the last two facts present the two classes. Note that no facts are established: $EF = 0$.

In the input matrix R_E (respectively output matrix R_S) are stored the premises (respectively the conclusions) of each rule. For example, the rule R2, has premises "buy = 1", "science=0" and a conclusion "class = spam".

Interaction between these two layers (*CelFact* and *CelRule*) is done by *Fact* and *Rule*.

Table 1. Final Configuration: *CelRule*, *CelFact*, R_E , and R_S .

Rules	ER	IR	SR
R1	0	1	0
R2	0	1	0
R3	0	1	0
CelRule			

Facts	EF	IF	SF
buy=0	0	1	0
buy=1	0	1	0
science=0	0	1	0
science=1	0	1	0
S3:class=spam	0	1	0
S5:class=legitimate	0	1	0
CelFact			

R_E	R1	R2	R3
buy = 0	1	0	0
buy = 1	0	1	1
science=0	0	1	0
science=1	0	0	1
S3:class=spam	0	0	0
S5:class=legitimate	0	0	0

R_S	R1	R2	R3
buy = 0	0	0	0
buy=1	0	0	0
science=0	0	0	0
science=1	0	0	0
S3: class=spam	0	1	0
S5:class=legitimate	1	0	1

3.2.2 Classification

We can use the model composed of *CelFact*, *CelRule*, R_E and R_S to classify new e-mails. The classification process is illustrated in Figure 2.

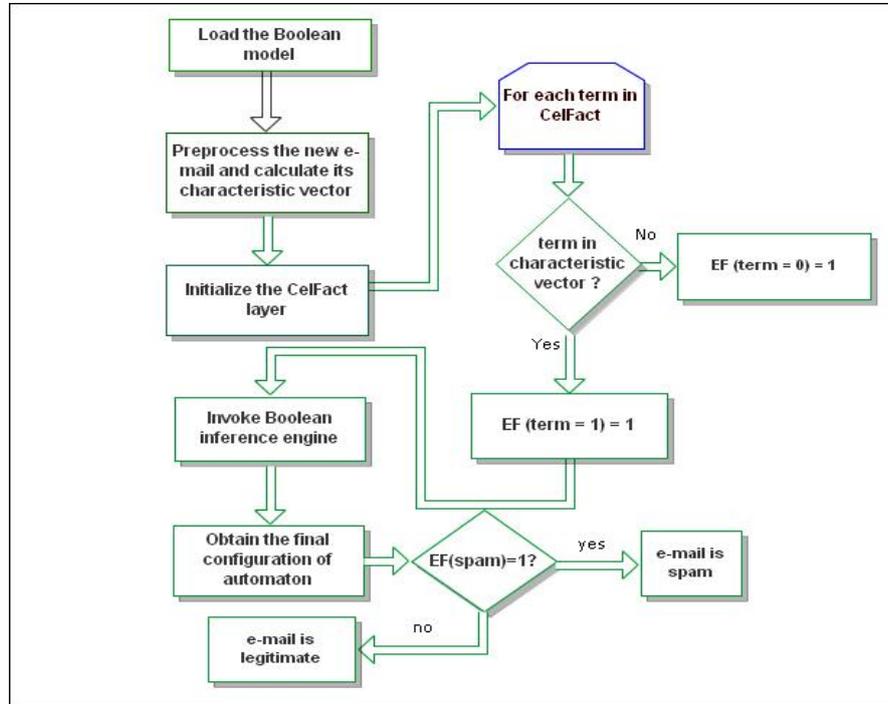


Fig.2. Classification Process

3.3 Proposed classifier combination

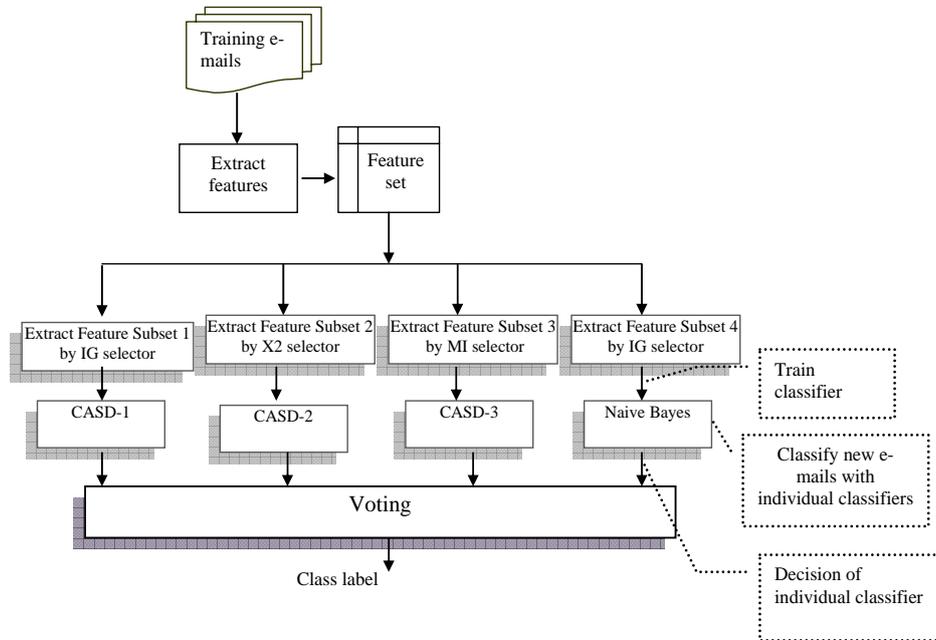


Fig.3. Architecture of the proposed ensemble classifiers for spam detection: 3CA-1NB (Three Cellular Automata combined with one Naïve Bayes).

Methods for creating ensembles [7, 15] focus on producing diversified base classifiers. Indeed, combination can be done by manipulating the training data, manipulating the input features, using different learning techniques to the same data. In this paper, we have chosen to consider combination by manipulating both features and using two different classifiers (CASD and NB).

The proposed approach termed 3CA-1NB (See Figure 3) combines three cellular automata classifiers (CASD), where each one is trained only with a feature subset. These subsets are generated with three different feature selection functions [17]: Information gain (IG), mutual information, and Chi-2 statistic respectively. We combine the decisions of these classifiers with that of Naïve bayes decision using voting¹ strategy. Our motivation for using this combining technique by varying feature selectors and using two different classifiers emerged from our preliminary results [4, 5] which indicate:

- The set of features selected by CASD during the learning phase depends on the selection function used to select features. For example, we observe that the features subset used by CASD after a selection based on information gain is generally

¹ E-mail is classified spam when at least two classifiers decide spam

different from that which was selected by the Chi-2 statistic or MI function. Therefore, we are guaranteed of having high feature set diversity.

- When using CASD, The quality of detection is better when features selection is done by MI or χ^2 (we have precision=100%), while the coverage is very low in the case of a selection with χ^2 (recall is very low) but very good with IG selector (see table 2 below). We want a classifier with high quality of detection and high coverage.
- Besides their simplicity, classification speed, CASD and NB also have advantages in terms of high classification accuracy.

4 Experimental study and results

We used the publicly available LingSpam corpora [2]. It comprises 2893 different e-mails, of which 2412 are legitimate e-mails obtained by downloading digests from the list and 481 are spam e-mails retrieved from one of the authors of the corpus [1, 13].

4.1 Linguistic preprocessing and feature selection

The first step in the process of constructing a classifier is the transformation of the e-mails into a format appropriate for the classification algorithms. We use an indexing module to:

- (a) Tokenize texts and establish an initial list of terms;
- (b) Eliminate stop words using a pre-defined stop list and;
- (c) Perform stemming with a variant of the Porter² algorithm.

Prior experiments [5] have shown that stemming improves classification performance. In this paper we report results on stemmed data. Since the number of terms after this preprocessing phase is very high, and to reduce the computational cost and improves the classification performance, we must select those that best represent the emails and remove less informative and noisy ones. Based on a study of [17] indicating the most used feature selectors in text categorization, we have implemented three feature selectors: Information gain (IG), mutual information (MI) and χ^2 -statistic (CHI). The system calculates the chosen measure for all the terms, and then takes the first k terms corresponding to larger scores. In our experiments the threshold's parameter is set to $k= 500$. After feature selection process, each e-mail is represented by a vector that contains a weighting for every selected term. This weighting represents the importance of that term in that e-mail. In this paper, we deal with a binary weighting. The k^{th} document is represented by the characteristic vector $X_k=(a_{1k}, a_{2k}, \dots, a_{Mk})$. $(a_{ik})=1$ if the term "i" is present in document "k", 0 otherwise and M is the index size.

² <http://tartarus.org/~martin/PorterStemmer/>

4.2 Performance measures

To evaluate performance we calculated spam precision (SP), spam recall (SR), spam F1 measure (F1) and accuracy. (Shown in equations 1 to 4). Let TN: the number of legitimate e-mails classified as legitimate (true negatives), TP: the number of spam emails classified as spam (true positives), FP: the number of legitimate e-mails classified as Spam (False Positives) and FN: the number of spam e-mails classified as legitimate (false negatives), then we have:

$$SP = \frac{TP}{TP+FP} \quad (1) \quad SR = \frac{TP}{TP+FN} \quad (2)$$

$$F1 = \frac{2 \times SP \times SR}{SP+SR} \quad (3) \quad A = \frac{TP+TN}{TP+FP+TN+FN} \quad (4)$$

Weighted accuracy (WA) was also calculated. More formally, WA is defined as follows:

$$WA = \frac{\lambda TN + TP}{\lambda(TN+FP) + TP + FN} \quad (5).$$

Three scenarios are evaluated and compared with previous work:

- (a) $\lambda=1$; no cost considered;
- (b) $\lambda=9$; semi-automatic scenario for moderately accurate filter, and
- (c) $\lambda=999$ completely automatic scenario for a highly accurate filter.

The experiments were performed with a k-fold cross validation with $k = 10$. In this way, our dataset was split 10 times into 10 different sets of learning sets (90% of the total dataset) and testing sets (10% of the total data). We conduct the training-test procedure ten times and use the average of the ten performances as final result.

4.3 Results and discussion

To evaluate 3CA-1NB and to show improvement over our previous work, we include the results of experiments on the LingSpam corpus with the CASD classifier using three subsets of features and NB classifier. In Table 2, we reproduce the best performing configuration. These configurations were used as members of the ensemble.

Table 2. Best configurations of NB, CASD and the corresponding performance.

Classifier	Feature Selector	Feature Size	SP (%)	SR (%)
NB	IG	500	99,00	82,10
CASD-1	IG	500	98,10	99,02
CASD-2	χ^2	500	100	2,5
CASD-3	MI	500	100	44,30

Figure 4 illustrates the ensemble results obtained using the 3CA-1NB classifier alongside those cited above. The results indicate improved performance when classifying with 3CA-1NB. It is clear that the former outperforms individual classifiers in accuracy and F1-measure. We conclude that the proposed ensemble

approach gives better performance than the four base classifiers used separately. The ensemble approach exploits the differences in misclassification by individual classifier and improves the overall performance. We also compare 3CA-1NB with the ensemble approaches developed by [13]. Table 3 reports the best results that we have achieved with 3CA-1NB and which are actually better than the results of [13].

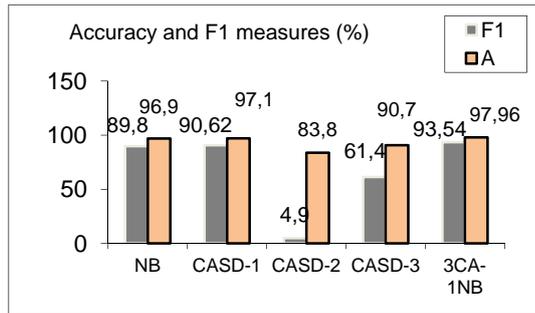


Fig.4. Performance of individual classifiers and 3CA-1NB on Spam Filtering

Table 3. Performance of stacking and 3CA-1NB on spam filtering.

Classifier	Performance Measures (%)				$\lambda=9$	$\lambda=999$
	SP	SR	SFI	A	WA	WA
Stacking[13]	90,80	91,90	91,30	97,10	98,00	98,10
3CA-1NB	98,20	89,36	93,54	97,96	99,37	99,58

5 Conclusion

In this paper a new approach for creating a diversity ensemble of classifiers is proposed. This method uses feature subset selection to train and construct a diversified set of base classifiers. We combine the predictions from the different classifiers by a voting technique in order to increase the performance of spam detection.

The results of experiencing on LingSpam datasets show better performance of the proposed method. As a future perspective, we will investigate the effect of combining more types of classifiers, and also, exploring other combination techniques [11] to further increase accuracy.

References

1. Androutsopoulos, I., Koutsias, J (2000a), "An Evaluation of Naive Bayesian Networks.", In: Machine Learning in the New Information Age. Barcelona Spain (2000) 9-17
2. Androutsopoulos, I., Paliouras, G., Karkaletsis, V., Sakkis, G., Spyropoulos, C.D., and Stamatoopoulos, P. (2000b). "Learning to filter spam e-mail: a comparison of a naive Bayes-

- ian and a memory based approach”. In Proc. of the Workshop on ML and Textual Information Access, PKDD 2000, France.
3. Atmani B., Beldjilali B. (2007). Knowledge Discovery in Database: Induction Graph and Cellular Automaton, Computing and Informatics Journal, 26, 171-197.
 4. Barigou F., Atmani B., Beldjilali B.: Utilisation de la machine cellulaire pour la détection des courriels indésirables. EGC 2011: 321-322, Revue des Nouvelles Technologies de l'Information, RNTI-E-20.
 5. Barigou N, Barigou F, Atmani B., “A Boolean model for spam detection”, In: Proceedings of the International Conference on Communication, Computing and Control Applications, Tunisia (2011).
 6. Carreras X., Marquez L., (2001), “Boosting Trees for Anti-Spam Email Filtering” in Proc. of RANLP-01, 4th International Conference on Recent Advances in Natural Language Processing.
 7. Dietrich T.G., Ensemble methods in machine learning. In: Kittler J., Roli F. (eds), Proc. of 1st Int. Workshop on Multiple Classifier Systems, Springer Verlag LNCS 1857, 2000, 1-15
 8. Flavio D. Garcia , Jaap-henk H. , Jeroen van N., “spam filter analysis” in ‘Proc. of 19th IFIP International Information Security Conference, 2004
 9. Guzella T. S., Caminhas W. M. 2009, “A review of machine learning approaches to spam filtering”, Expert Systems with Applications, 36(7), 10206-10222.
 10. Koprinska I., Poon J., Clarck J., Chan J. Learning to classify e-mail. Info. S. 177: 2167-2187, 2007.
 11. Kuncheva L., Combining Pattern Classifiers, Methods and Algorithms, Wiley Inter Science, 2005.
 12. Rios G., Zha H. Exploring support vector machines and random forests for spam detection, in: Proc. First International Conference on Email and Anti Spam (CEAS), 2004.
 13. Sakkis, I. Androutsopoulos, G. Paliouras, V. Karkaletsis, C. D. Spy-ropoulos, and P. Stamatopoulos. Stacking classifiers for anti-spam filtering of e-mail. Proceedings of 6th Conference on Empirical Methods in Natural Language Processing, 1:44-50, 2001.
 14. Shih D. H., Chiang S., Lin I. B. Collaborative spam filtering with heterogeneous agents. Expert systems with applications, 34(4), 1555-1566, 2008.
 15. Valentini G., Masuli F., Ensembles of Learning Machines. In: R.Tagliaferri, M. Marinaro (eds), Neural Nets WIRN Vietri-2002, Springer-Verlag LNCS, vol. 2486, 2002 , 3-19.
 16. Zighed. “Graphe d’induction: Apprentissage et data mining”. HERMES, 2000.
 17. Yang Y., Pedersen J. O., “A comparative study on feature selection in text categorization”, FISHER D. H., Ed., Proceedings of ICML-97, 14th International Conference on Machine Learning, Nashville, US, Morgan Kaufmann Publishers, 412–420,1997.

Clustering-based data in ad-hoc networks

Bakhta Meroufel¹, Ghalem Belalem²
[bakterasba@gmail.com](mailto:bakhtasba@gmail.com), ghalem1dz@univ-oran.dz
Department of computer science
Faculty of Sciences
Oran University (Es Senia)
-Algeria-

Abstract: Clustering is an important mechanism in large wireless sensor networks for obtaining scalability, reducing energy consumption and achieving better network performance. Most of the research in this area has focused on clustering based on physical parameters such as: energy, mobility, connectivity, density..., without taking in the count the data stored in each nodes. The main objective of this paper is to provide a useful fully-distributed algorithm for clustering that maximize the intra-cluster access, so we used a new heuristic parameter that combine between energy, mobility and number of data to select the Cluster-Heads. Our clustering strategy was compared with Lowest-ID Cluster Algorithm (*LID*) and the results show that our algorithm improves system performance and increases its life.

Keywords: Clustering, data, ad-hoc networks, availability, mobility.

1. Introduction

A mobile ad-hoc network (MANET) is a collection of mobile nodes that form a wireless network without the existence of a fixed infrastructure or centralized administration. This type of network can survive without any infrastructure and can work independently. Hosts forming an ad hoc network can take equal responsibility in maintaining the network. Each host provides routing services to other hosts to deliver messages to remote destinations. As such a network requires no fixed infrastructure; it makes them better for deployment in a volatile environment such as battlefield and disaster relief situations [6]. Some of the constraints in MANETs are: limited bandwidth, low battery nodes and link frequent breaks due to node mobility. These constraints must be taken into account, while maintaining the connectivity between nodes. Clustering plays an important role in solving such problems [11]. It hides the dynamic structure of the system by forming a hierarchical topology. There are many researches that offer different strategies for clustering based on different metrics.

Unfortunately, the majority of this works do not take into account the management of requests in the system [1]. In a network where energy is critical, the management of user requests and the difference between intra-and inter-cluster access can consume a lot of energy, which degrades system performances and reduces its life. In this paper, we propose a new strategy of clustering that takes into account the data of the system in addition to mobility and energy of the nodes. Our goals are minimizing the energy used by managing the user's requests and at the same time improving the response time and the stability of the system. The rest of the paper is structured as follows: in the second section, we present some related works on clustering. The third section presents the environment of our work. In

section four we introduce our clustering algorithm based on the data, energy and mobility of each node in the system. We validate the proposed approach by a set of experimental results shown in the fifth section, and we finish this work with a conclusion and perspectives.

2. Related works

The algorithms differ on the criterion of selection of cluster-head. Among these algorithms we have: the Lowest-ID Cluster Algorithm (*LID*) [6], in this algorithm, each mobile host in the network must have a unique identifier *id*. The node that has the smallest *id* among all its neighbors is elected as the cluster-head. The cluster is formed by the cluster-head and all its neighbors. *LID* has the advantage of being simple and rapid but also generates a large number of clusters and can be adjustable to changes in the topology. The Last cluster-head Change algorithm (*LLC*) [4], is designed to minimize the change of cluster-head and provides better stability in the composition of system. In High-Connectivity Clustering (*HCC*) [9], [11], the cluster-head election is based on the degree of connectivity (number of neighbors of the node) instead of the identities of the nodes. A node is elected as a cluster-head if it has the highest connectivity among all its neighbors. This algorithm suffers from frequent changes of cluster-head. In [3], two clustering algorithms are proposed, providing a new approach. The first, Distributed Clustering Algorithm (*DCA*) which is targeted to "quasi-static" in which the movement of nodes must be "slow". A weight is defined by the speed of each node. The criterion for the election of the cluster-head is the maximum weight in its neighborhood. The node whose weight is the greatest among all its neighbors is elected cluster-head. The second algorithm is designed for networks and mobility called Mobility-Adaptive Clustering algorithm (*DMAC*). Each node reacts locally to all changes depending on its status: member node or cluster-head. In the two algorithms, it is assigned different weights to the nodes and it is assumed that each node has knowledge of its weight. A node is selected as cluster-head if its weight is greater than among all its neighbors. There are other works that take into account data management in ad hoc networks, such as work [13] which proposes a strategy with two steps: first create the cluster and then replicate the data requested in each cluster. The works [12] and [7] also proposed methods of replication to improve availability of data or to facilitate the update and allocation of data. All works cited precisely separate between the clustering and data management and use the replication to improve their approaches. In this paper we propose a strategy of clustering that includes data management and the creation of clusters in one step by taking advantage of existing data without creating other replicas in the system.

3. Contributions

Proper management of queries in an ah-hoc network improves the reliability and usefulness of the system but increases energy and reduces the lifetime of the nodes. The majority of the works cited above focus on the physical characteristics of networks such as: energy, connectivity, mobility, density, without taking into account the usefulness of the network itself, that is to say the goal of building the networks, the type of services provided and the types of treatments that can be achieved. In this paper, we propose a new algorithm for non-overlapping clustering (see definition 2) to minimize power consumption and

response time in the system. Minimizing the access time of queries depend on the location and distribution of data in the system [8]. The main idea in this work is to maximize intra-cluster access by maximizing the number of different data in the same cluster and remove the replica of same data. In the example of Figure 1, the system contains two clusters, each cluster has two different data (two colors) and in this case, the cluster can satisfy only queries that research these two data. As against the second system in Figure 2, each cluster contains three data which increases the number of requests satisfied at the intra-cluster.

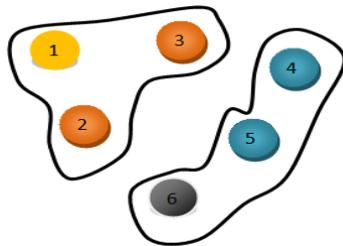


Fig1: First system: two colors in each cluster

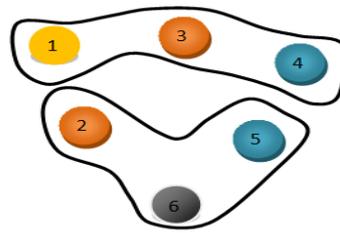


Fig2: Second system: three colors in each cluster

Model

The system is modeled by an undirected graph $G = (V, E)$ where V is the set of network nodes and E models all the connections between these nodes. An edge $(u, v) \in E$ if and only if nodes u and v can mutually receive transmissions of each other. This means that all links between nodes are bidirectional. In this case, we say that u and v are neighbors. The set of neighbors of a node $v \in V$ is denoted $Neigh_v$. Each node u of the network is associated to a unique *id* and can communicate with its neighbors $Neigh \subseteq V$. Each node is characterized by its mobility Mob and energy $Energ$ remaining in the battery, it also can store zero or more data D_i . Users can initiate read requests to access data in the system.

For the sake of clarity, we will use later in this paper the following notations:

- $Neigh_u$: The neighbors of node u .
- $Energ_u$: The remaining energy of the battery of node u .
- Mob_u : The mobility of the node u .
- $ListL_u$: List local data stored in the node u .
- $ListT_u$: Total list of data stored in the node u and its neighbors (see formula 2).

Before describing our clustering algorithm in detail, we make the following assumptions, which are common in the design of clustering algorithms for MANETs [1], [3], [6], [11], [13]:

- The network topology is static during the execution of the clustering algorithm.
- A packet transmitted by a node can be received correctly by all its neighbors in a finite time.
- Each node has a unique *id* and knows its neighbors and vice versa.
- The inter-cluster access is expensive compared to intra-cluster access in terms of bandwidth and energy.

One more of these assumptions, we also assume that all requests made by users are of the reading type [10].

4. Clustering

The clustering algorithm consists of two steps: the selection of cluster-head and then the construction of clusters.

4.1 Metric of clustering

Given the interest in the concept of clustering and its undeniable contributions to improve the performance of an ad hoc network, the choice of the clustering mechanism is important. Thus, a clustering algorithm must first be able to select the appropriate nodes to ensure the functionality of the cluster-head. In our algorithm, the cluster-head selection is based on a new metric γ :

$$\gamma_n = \frac{Energy_u}{Mob_u} * ||ListT_u|| \quad (1)$$

- $\tilde{\alpha}_u$: Metric of clustering.
- $||ListT_u||$: The cardinality of the data list of the node u and its neighbors.

$$ListT_u = \bigcup_{i=1}^m ListL_m \cup ListL_u \quad (2)$$

- m : The number of neighbors of node u .

The node with the maximum γ in the neighborhood can be selected as a cluster-head. The rapport between energy and mobility helps to select a cluster-head which has a relatively high energy capacity and low mobility witch increase the stability of the system. Held that the parameter $||ListT_u||$ can elect a cluster-head which has a great opportunity to satisfy read requests in a single degree at the neighborhood.

Maximize the number of different data per cluster can be achieved by choosing as cluster-head node, which in its first neighborhood has many different data including its own list of data (Maximum $||ListT_u||$). For example in Figure 3, assuming that the report $\frac{Energy_i}{Mob_i}$

$/i \in [0,3]$ is the same for all nodes i (just to explain) then:

- For node 0, the list $ListT_0 = \{A\} \cup \{A, Z\} \cup \{A, H\} = \{A, H\}$.
- For node 1, the list $ListT_1 = \{A, H\} \cup \{C, B, F\} \cup \{A\} \cup \{A, Z\} = \{A, B, H, F, C, Z\}$.
- For node 2, the list $ListT_2 = \{A, Z\} \cup \{A, H\} \cup \{A\} = \{A, E, H\}$.
- For node 3, the list $ListT_3 = \{C, B, F\} \cup \{A, H\} = \{A, H, C, B, F\}$

We note that the node1 in its neighborhood contains a lot of data compared to its neighbors: $|ListT_1| > |ListT_3| > |ListT_2| > |ListT_0|$, so node 1 is capable of meet a lot of requests (read requests) by at most one jump (zero jump if data is stored in the node itself). In this case, the node 1 will be selected as a cluster-head and broadcasts its status to its neighbors.

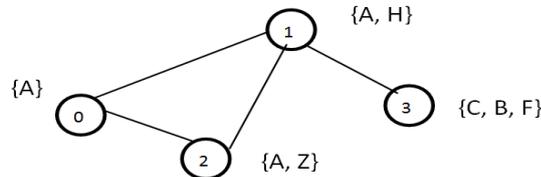


Fig3. A system with 4 nodes

4.2 Clustering algorithm

The construction of the clusters is through periodic exchange of messages which we will call *hello* messages. Each network node exchanges with its neighbors the *hello* messages.

Each *hello* message sent by a node u contains four values are: id_u , $Statut_u$, $(Energ_u / Mob_u)$ and $ListL_u$.

- id_u : the identifier of the node u .
- $Statut_u$: The role of the node u in the system (See definition 3).

This message is also used for each node to verify the presence of its neighbors. Thus, if a node no longer receives *hello* message from a neighbor at the end of a period, it considers that this neighbor has disappeared. So each node waits a specified period in advance and it is assumed that during this period, all nodes have sent their *hello* message. In our algorithm we use the following definitions:

Definition 1 (Cluster): We define a cluster C_i by a connected sub-graph of the network, with a diameter less than or equal to 2. The cluster has an identifier corresponding to the identity of the node with the higher γ in the cluster, that is to say that if the cluster is the cluster C_i then the cluster-head of this cluster has the identifier $id = i$.

Definition 2 (Non-overlapping): a non-overlapping clustering ensures that there is no node that belongs to two different clusters at the same time witch minimizes the conflict.

Definition 3 (node status): Each node u has a status $Statut_u$, the $Statut_u$ can be one of the following: *CH*: cluster-head node, *MN*: member node, or *GN*: gateway node. These three roles are defined in definitions 3.1, 3.2 and 3.3:

Definition 3.1 (Cluster-head): A node u is called cluster-head of cluster C_i iff: $id_u = i \wedge \forall v \in Neigh_u, \gamma_u > \gamma_v$.

Definition 3.2 (member node): A node u is said member node if it is not a cluster-head and it does not have in its neighborhood a node associated to a different cluster. $u \in C_i$ Then $id_u \neq i \wedge [\nexists v \in (C_j / C_j \neq C_i) \wedge v \in Neigh_u]$

Definition 3.3 (Gateway node) A node u is a Gateway node iff: $u \in C_i$ Then $\exists v \in Neigh_u / v \in C_j \wedge C_j \neq C_i$. Gateway node has a special role. It provides access to one or more neighboring clusters.

Definition 4 (Degree of no-similarity): the degree of non-similarity $\bar{\beta}$ between two nodes u and v is the size of the list of data that exists in a node and does not exist in the other (formula 2). Two nodes with a high $\bar{\beta}$ need each other more then two nodes with a lower $\bar{\beta}$ (disjointness of contents).

$$\bar{\beta}(u,v) = \|(ListL_u \cup ListL_v) - (ListL_u \cap ListL_v)\| \quad (3)$$

Where:

- u, v : Nodes.
- $ListL_u$: List of local data in the node u .

For example, if local lists of data in nodes u, v, h are $ListL_u = \{A, B, C\}$, $ListL_v = \{A, E, F, G, H\}$, $ListL_h = \{A, B, C, D\}$ respectively then:

- $\bar{\beta}(u, v) = \|\{B, C, E, F, G, H\}\| = 6$.
- $\bar{\beta}(u, h) = \|\{D\}\| = 1$.

Clustering steps are:

1. Each node calculates the parameter γ and disseminates information on its neighbors by the *Hello* packet.

2. If the node has the maximum γ among its neighbors then it becomes the cluster-head (Status = CH) and sends requests to join the cluster to the other. If two nodes u and v have $\gamma_u = \gamma_v$, then the cluster-head is the node that has the lowest id .
3. The node that receives a request to join a cluster, then it became with a Status = MN.
4. If the node receives several requests from different cluster-head to join their clusters, then it becomes a gateway node: Status = GN. But as a cluster selects the one with the maximum γ . In case of a tie, the node selects the cluster-head that has the maximum non-similarity degree \bar{a} with it.
5. If two neighboring nodes have both a Status = CH, then the first node that detects the conflict through *hello* messages received compared γ to that of its cluster-head neighbor. If the γ is smaller, it abandoned his status as cluster-head and becomes a member node and sends a hello message to its neighbors announcing its new status. Otherwise, it retains the role of cluster-head.
6. The algorithm terminates when each node in the system specifies its status.

For example in Figure 4: $\gamma_0 = 20$, means the node will be a cluster-head. The node 2 will be a node member in the cluster of node 0. Node 7 is a gateway node. The result of clustering is shown in Figure 4.

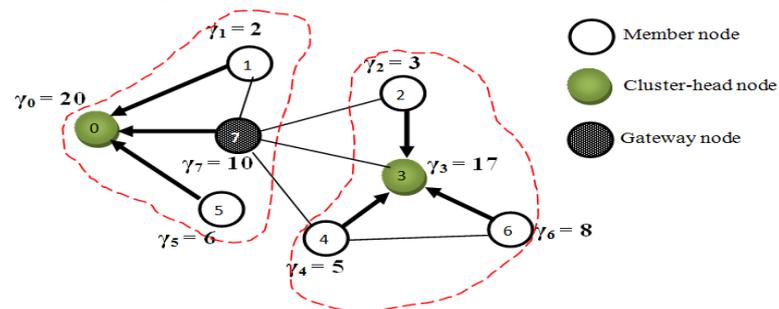


Fig 4. Example of clustering of 8 nodes

4.3 Maintenance of the topology

In ad hoc networks, topology changes frequently due to node mobility and energy. So we have to manage: 1) nodes that move, 2) nodes that disappear and nodes that appear. Our algorithm can also manage these cases.

- The appearance of a new node. When a new node enters the network, it broadcasts at a regular interval of time a message of type *hello*. At the end of the timeout, if it has not received any type *hello* message from a cluster-head, the node becomes cluster-head. If the node received a *hello* message from a cluster-head, the node becomes a member node. In the event that it receives from more than one cluster-head, the node becomes a gateway node (Execute step 4 of the clustering algorithm)
- Disappearance of a node. In the same way as for the appearance by exchanging *hello* messages the neighbors are aware of the disappearance.
- Moving a node. Case equivalent to the emergence of a new node. In this case the *hello* messages will be received by the other new neighbors.

5 Implementation and validation

To validate our approach, we used our simulator that allows us to run and measure the performance of our strategy of clustering *CD* (*Data based Clustering*) and compare it with the *LID* strategy [6]. As it is already mentioned in the section of related works, The *LID* is a one hop clustering that selects as cluster-head the node with the minimum *id* among its neighbors without taking into account other characteristics of system such as energy, mobility. The parameters of simulations are presented in Table 1.

Table 1: Parameters of simulations

Variable	Value
Total number of nodes	1-1000
Node mobility (%)	1-100
Node energy (%)	10-100
bandwidth (Mb/s)	20-100
range (m)	10-100
Total number of data	10-300
Total number of data per node	0-10
Data size (Mo)	1-10

In the first experiment, we studied the impact of the range on the number of clusters in both clustering approaches *LID* and *CD*. The range specifies the number of possible neighbors for each node. We studied the impact of this parameter on the number of clusters in the system. The results are shown in Figure 5. Increasing the range minimizes the number of clusters in the both approaches because it increases the number of neighbors per node. But our approach optimizes the number of clusters with a gain of 25% compared to the *LID* approach.

The number of nodes also affects the number of clusters. According to the results obtained in the second experiment (see Figure 6). Increasing the number of nodes increases the number of clusters, but not on the same frequency for both approaches. Our approach *CD* is better than *LID* approach with a gain estimated by 22.8%.

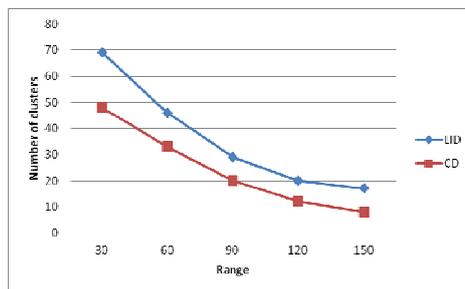


Fig5: Range vs Number of clusters.

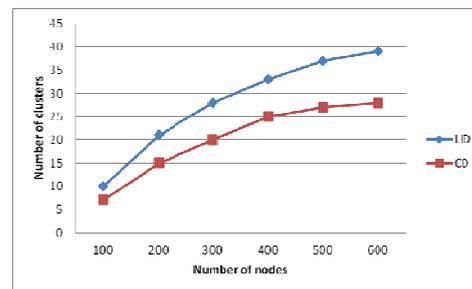


Fig6: Number of nodes vs Number of clusters.

The number of transitions is a very important parameter; it reflects the stabilization of the topology constructed by the clustering. The transition is the passage from configuration *i* to

configuration $i + 1$. The configuration is the result of the execution of at least one step of the clustering algorithm [2]. The results of our third experiment (see Figure 7) show that despite the increased number of nodes in random graphs we used, the number of transitions for stabilization in our algorithm *CD* varies very slightly, while still better with 30 % than the clustering approach *LID*, because *CD* is purely local and does not require that information obtained through the neighborhood *hello* messages. This ensures the scaling.

Minimizing response time increases system reliability. The results of the fourth experiment shown in Figure 8 prove that in the *LID* approach, the response time increases with increasing number of nodes and the time is greater compared to our approach because the data in cluster formed by *LID* are located randomly. The *CD* approach works better because the CH maximizes the number of data in its neighborhood which increases intra cluster access from access inter-cluster. The *CD* minimizes 43% of response time compared to *LID* approach.

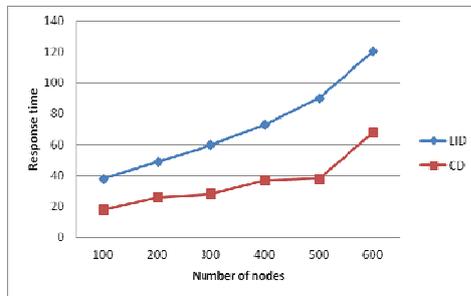
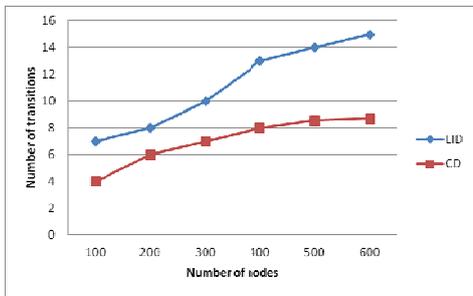


Fig7 : Number of nodes vs Number of transitions **Fig8**: Number of nodes vs response time

In the last experiment, we studied the impact of number of read requests to the energy consumed in the system. We note that the energy increases with increasing number of applications for both approaches because each query must be redirected to other nodes in different clusters to reach the answer. But our approach *CD* decreases remarkably the energy consumption (a gain of 25%) because it maximizes the number of different data in each cluster taking in the account the mobility and the energy of each node, which improves the intra-cluster access and stabilize the topology (see Figure 9).

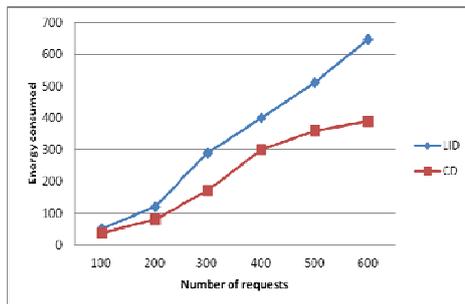


Fig9: Number of requests vs Consumption of energy.

6 Conclusion

In this report, we proposed a clustering strategy that maximizes intra-cluster access and minimizes energy consumption. This strategy uses energy, mobility and the types of data stored in the neighborhood to elect the cluster-head. The experimental results demonstrate the effectiveness of our proposal. Our strategy can be used to create clusters in unstructured P2P system as *FastTrack* and *Gnutella* [5], where the number of message and consumption of bandwidth are large. To avoid thus problems and increase the quality of services in requests management, we use our clustering algorithm where for each node u : $\frac{Energy_u}{Mob_u} = 1$, so $\gamma = |ListT_u|$. Perspective, we offer extension work by taking into account the availability of links between the different neighboring data to improve system reliability.

References

1. A. Abbasi, M. Younis, *A survey on clustering algorithms for wireless sensor networks*, *Journal Computer Communications*. 30(14): 2826–2841, 2007.
2. D.J. Baker, A. Ephremides and J. A. Flynn. *The Design and Simulation of a Mobile Radio Network with Distributed Control*. IEEE Journal on Selected Areas in Communications, 2(1): 226–237, 1984.
3. S. Basagni, *Distributed Clustering Ad Hoc Networks*, In Proceedings of the IEEE International Symposium on Parallel Architectures, Algorithms, and Networks (I-SPAN), pages 310-315, august 1999.
4. C-C. Chiang, H-K. Wu, W. Liu, M. Gerla, *Routing in Clustered Multi hop, Mobile Wireless Networks With Fading Channel*. IEEE Singapore international conference on networks, pages 197-211, 1997.
5. K. Eger, T. Ho_feld, A. Binzenhofer, and G. Kunzmann, *Efficient simulation of Large-Scale P2P Networks: Packet-level vs. Flow-level Simulations*, in Proc. UPGRADE-CN'07, HPDC Workshops, pages 1-16, USA, 2007.
6. J. W. A. Ephremides and D. J. Baker, *A design concept for reliable mobile radio networks with frequency hopping signaling*, Proceeding of IEEE, pages 53-76, January 1987.
7. T. Hara, *Replica Allocation Methods in Ad Hoc Networks with Data Update*, *Journal Mobile Networks and Applications*, 8(4): 342-354, 2003.
8. B-J. Ko and D. Rubenstein. *Distributed self-stabilizing placement of replicated resources in emerging networks*. IEEE/ACM Trans. Netw., 13(3): 476–487, 2005.
9. S-J. Lee, W. Su, J. Hsu, M. Gerla and R. Bagrodia, *A Performance Comparison Study of Ad hoc Wireless Multicast Protocols*. Proceedings of the IEEE International Conference on Computers and Communications, pages 565 – 574, March 2000.
10. B. Meroufel, G. Belalem: *Availability Management in Data Grid*, Lecture Notes in Electrical Engineering, 1, Volume 107, IT Convergence and Services, Part 1, Pages 43-53.
11. H. Taniguchi, M. Inoue, T. Masuzawa, H. fujiwara, *Clustering Algorithms in Ad Hoc Networks*, *Electronics and Communications in Japan*, Part2, 88(1): 51-59, 2005.
12. L. Yin and G. Cao, *Balancing the tradeoffs between data accessibility and query delay in ad hoc networks*, Proceedings of the 23rd IEEE International Symposium on Reliable Distributed Systems, pages 289 – 298, 2004.
13. J. Zheng, J. Su, and X. Lu. *A Clustering based Data Replication Algorithm in Mobile Ad hoc Networks for Improving Data Availability*. In Proc 2nd International Symposium on Parallel and Distributed Processing and Applications (ISPA 2004), Pages 399–409, 2004.

Posters



A Recommendation-based Approach for Communities of Practice of E-learning

Lamia Berkani^{1,2}, Omar Nouali³ and Azeddine Chikh⁴

¹ Department of Computer Science, USTHB University, Bab-Ezzouar, Algiers, Algeria

² Higher National School of Computer Science, ESI, Oued Smar, Algiers, Algeria

³ Department of Research Computing, CERIST, Algiers, Algeria

⁴ Department of Information Systems, KSU University, Riyadh, Saudi Arabia

l_berkani@hotmail.com, onouali@mail.cerist.dz, az_chikh@ksu.edu.sa

Abstract. The paper presents a recommendation-based approach for knowledge resources in Communities of Practice of E-learning (CoPEs). The proposed approach is based on the hybrid semantic information filtering (IF), integrating the content-based filtering, the collaborative filtering and the ontology-based filtering approaches. The main idea is to apply a multi-level filtering, where three dimensions have been proposed for the profile: collaborative, social and semantic.

Keywords: CoP of e-learning, knowledge resource, recommendation, information filtering, ontology-based filtering, profile.

1 Introduction

According to Wenger [1], Communities of Practice (CoPs) are “*groups of people who share a concern, a set of problems, or a passion about a topic, and who deepen their knowledge and expertise in this area by interacting on an ongoing basis*”. CoPs allow members to share their practices, to develop their knowledge and skills. They are embedded within all areas and domains including education, engineering, management, health, etc. They are seen as a new organizational structure offering innovative means for creating and sharing knowledge.

The authors in [2, 3] extended the application of this concept to the domain of e-learning. They considered CoPs of e-learning (CoPEs) as a virtual framework for exchanging and sharing techno-pedagogic knowledge and know-how between actors of e-learning. CoPEs give the possibility for professionals in e-learning to gather, collaborate, and organize themselves in order to: (i) share information and experiences related to e-learning development and use; (ii) collaborate in order to solve together e-learning problems and to build techno-pedagogic knowledge and best practices; (iii) learn from each other and develop their competences and skills in their domain of expertise.

In order to participate effectively to the knowledge management and learning processes in a CoPE, members need guidance to find and synthesize information.

They need to find the adequate resources for their activities within the CoPE or to be used for example to design their courses within the e-learning platform.

This paper will focus on the recommendation of knowledge resources using Information Filtering (IF) approach that will attempt to present to the member information items, according to his interests.

The rest of this paper is organized as follows: Section 2 presents the background and related work about IF approaches. Section 3 discusses the application of IF in CoPEs and proposes a hybrid semantic IF approach for the recommendation of knowledge resources in CoPEs. Finally the conclusion highlights the main results of this work and presents some perspectives.

2 Information Filtering

We present in this section the different IF techniques and some related works close to our context of study.

2.1 Background

Information filtering (IF) is the process allowing, starting from an incoming volume of dynamic information, to extract and present the only information interesting either a user or a group of users having relatively similar interests. The filtering system makes a "prediction" about the usefulness of the information to the user. This prediction is based on the "profile" of the user and leads to a decision-making: "recommend" or "not recommend" information [4]. The problem of IF can be expressed as follows [5]: C is a set of users, S a set of documents to be recommended, and u a function which measures the importance that represents a document s to a user c . The objective is to search about documents s so as to maximize the utility function u , as described formally:

$$U: C \times S \rightarrow R \\ \forall c \in C, s'_c = \arg_{s \in S} \max u(c, s)$$

The IF systems are classified into three categories: the content-based filtering systems, the collaborative filtering systems, and the hybrid ones.

- The content-based filtering systems recommend the similar documents to those the user has already liked. This is calculated by comparing the interests of users introduced explicitly (e.g. through a questionnaire) or implicitly (through a behavior supervision) with the characteristics of the documents [6].
- Collaborative filtering or social recommender systems recommend data items to a user by taking into account the opinions of other users [7]. Instead of recommending data items because they are similar to items the user preferred in the past (content-based recommendation), collaborative approaches generate recommendations about data items that users with similar interests liked in the

past. In order to estimate user's preference for an item, collaborative filtering systems collect ratings through explicit means (e.g. the user is asked to rate the item), implicit means (e.g. the system infers user's preference by observing user's actions) or both. More formally, the utility of a document s to a user c , $u(c, s)$ will be calculated based on the $u_j(c_j, s)$ that are similar. The prediction function F uses the vote matrix $C \times S$ and proceeds in two steps [8]: (1) calculate the similarity between the users and infer communities, (2) predict notes for a few documents and select only those with a high score.

There are two major collaborative approaches, an approach based memory (the note given by a potential user to a document is calculated based on ratings given by other users for the same document) and another based model (learn a descriptive model linking users, documents and votes). With the growth of e-commerce, collaborative filtering techniques have become well known through their use in commercial web sites such as Amazone.com.

- The hybrid systems, combine in different ways the two previous approaches and try to overcome their shortcomings: the "cold start" problem when there are not enough ratings, the inability to recommend non-textual documents that do not have information about their content, quality criteria and reliability of the source are not considered in the content-based systems, etc.

Recently, with the emergence of the semantic Web, a new generation of recommender systems has emerged [9]: *(1) the ontology-based IF systems* (conversion from a description of the documents by key words to a semantic description based on concepts); *(2) the collaborative annotations systems* (assigning to resources a set of words called tags or annotations to describe their content or provide a more contextual and semantic information); *(3) the social networks-based IF systems* (managing the friends lists and expressing their interests such as in Facebook, and LinkedIn, encouraged the reuse of this social data in the IF systems).

2.2 Related work

The state of the art shows an important number of proposed recommender systems. We present some works related to our context of study.

QSIA (Questions Sharing and Interactive Assignments) for learning resources sharing, assessing and recommendation has been developed by Rafaeli et al. [10]. This system is used in the context of online communities, in order to harness the social perspective in learning and to promote collaboration, online recommendation, and further formation of learner communities.

ReMashed is a recommender system that addresses learners in informal learning networks [11; 12]. The authors created an environment that combines sources of users from different Web2.0 services and applied a hybrid recommender system that takes advantage of the tag and rating data of the combined Web2.0 sources.

3 Contribution

We propose a recommendation system based on the hybrid semantic IF (see Fig.1). In the CoPE, one member or a group of members need a recommendation of knowledge resources in the following situations: (1) information retrieval; (2) when a new resource has been added to the memory and that can be interesting for the member; (3) during an activity (e.g. design of a learning scenario); and (4) for a new member who integrate the community.

Accordingly, we propose a recommendation system based on the hybrid semantic IF. The main idea is to apply a multi-level filtering approach and to consider a multi-level profile according to the need, context, and conditions (availability of information), so as to make an effective recommendation. As illustrated in Fig. 1, resources are represented semantically using OntoCoPE ontology [13] and three dimensions are considered for the profile: collaborative (implicit/explicit evaluations), social (a set of personal information: name, specialty, email, a set of contacts...), and semantic (members' interests represented in the form of concepts with weight corresponding to their degrees of importance). Each dimension produces a set of recommendations that can be classified, using for example an adaptive classification:

$$u(c,s) = \alpha \cdot u \text{ Coll} (c, s) + \beta \cdot u \text{ Social} (c, s) + \gamma \cdot u \text{ Sem} (c, s); \text{ where } : \alpha + \beta + \gamma = 1$$

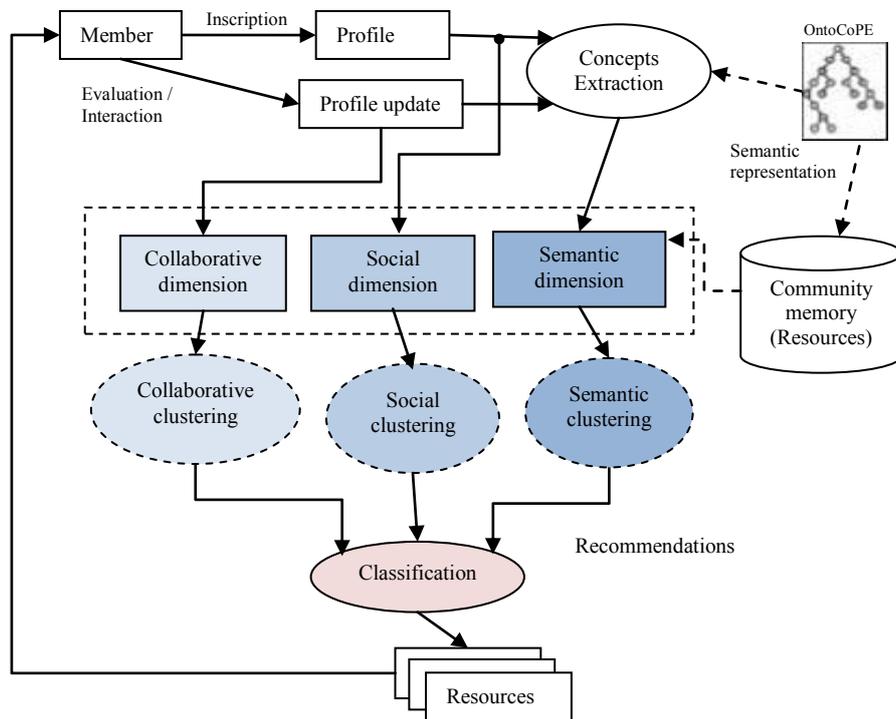


Fig. 1. The Hybrid semantic filtering system adapted from [9].

The social recommendation has the priority, if there are no or not enough evaluations or if the semantic dimension is not yet well defined. The collaborative recommendation has the priority, if we want to discover new interests to a member. Otherwise, the semantic recommendations will have the priority as they more correspond to the members' interests.

4 Conclusion

The paper presents proposes a recommendation-based approach for knowledge resources in CoPEs, using the hybrid semantic IF. The main idea is to apply a multi-level filtering, where three dimensions has been proposed for the profile: collaborative, social and semantic. However, the proposed approach needs to be evaluated in a real situation. We envisage in a near future to develop the recommendation system and to evaluate its performance using a learning community of students within the USTHB University in Algeria.

References

1. Wenger, E., McDermott, R., Snyder, W.M.: *Cultivating Communities of Practice: A guide in Managing Knowledge*. Harvard Business School Press (2002)
2. Berkani, L.: *Communities of Practice of E-learning – « CoPE »: Definition of Concepts and Proposition of a Learning Scenario Specification Language*. Master Thesis, Higher National School of Computer Science (ESI), Algiers, Algeria (2007)
3. Chikh, A., Berkani, L., Sarirete, A.: *Modeling the communities of practice of e learning (CoPEs)*. In 4th Annual Conference proceedings of Learning International Networks Consortium, pp. 428--441, Jordan (2007)
4. Lopez M.: *Information Access through a controlled collaborative filtering system*, PhD Thesis, Joseph Fourier University, (2005).
5. Adomavicius G., Tuzhilin A.: *Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions*, IEEE Transactions on knowledge and data engineering, vol. 17, n° 6, (2005).
6. Peis E., Morales-del-Castillo J. M., Delgado-López J. A.: *Semantic Recommender Systems. Analysis of the state of the topic*, *Hipertext.net*, n° 6, (2008).
7. Shafer, J.B., Frankowski, D., Herlocker, J., Sen, S.: *Collaborative filtering recommender systems*. In: Brusilovsky, P., Kobsa, A., Neidl, W. (eds.) *The Adaptive Web: Methods and strategies of Web Personalization*. LNCS, Vol. 4321, pp. 291--324, (2007).
8. Woerndl W., Eigner R.: *Utilizing Physical and Social Context to Improve Recommender Systems*, Workshop on Web Personalization and Recommender Systems (WPRS07), USA, (2007).
9. Oufaida H., Nouali O.: *The Collaborative Filtering and the web 2.0. Literature Review*, Document numérique 2008/1, Vol. 11, pp. 13--35, (2008)
10. Rafaeli, S., Barak, M., Dan-Gur, Y., & Toch, E. *QSIA: a web-based environment for learning, assessing and knowledge sharing in communities*. *Computers & Education*, 43(3), 273-289, (2004).
11. Drachsler, H., Pecceu, D., Arts, T., Hutten, E., Rutledge, L., Van Rosmalen, P., Hummel, H.G.K., Koper, R.: *ReMashed - Recommendations for Mash-Up Personal Learning Environments*. In: Cress, U., Dimitrova, V., Specht, M. (eds.): *Learning in the Synergy of*

- Multiple Disciplines, EC-TEL 2009, LNCS 5794, Berlin; Heidelberg; New York: Springer, pp 788-793, (2009)
12. Drachsler, H., Pecceu, D., Arts, T., Hutten, E., Rutledge, L., Van Rosmalen, P., Hummel, H.G.K., Koper, R.: ReMashed - An Usability Study of a Recommender System for Mash-Ups for Learning. 1st Workshop on Mashups for Learning at the International Conference on Interactive Computer Aided Learning, Villach, Austria. (2009)
 13. Berkani, L., Chikh, A.: Towards an Ontology for Supporting Communities of Practice of E-learning "CoPEs": A Conceptual Model. In: Cress, U., Dimitrova, V., and Specht, M. (eds.) EC-TEL 2009. LNCS 5794, pp. 664–669. Springer, Heidelberg (2009)

AMSI: An Automatic Model-Driven Service Identification from Business Process Models

Mokhtar Soltani¹, Sidi Mohammed Benslimane²

¹ Sciences and Technology Department, Ibn Khaldoun University, Tiaret, Algeria

² Computer Sciences Department, Djillali Liabes University, Sidi Bel Abbes, Algeria

m_soltani@mail.univ-tiaret.dz

benslimane@Univ-sba.dz

Abstract. The evolution of software engineering has passed through various paradigms; including structured programming, object oriented programming, component-based approaches and in recent years service-oriented computing. One of the key activities needed to develop a quality service oriented solution is the specification of service model. The majority of existing methods for service model specification are developed manually because they are based on the competence of the developers. The integration of Business Process Modeling (BPM) and Model-Driven Development (MDD) allows the automation of the SOA (Service-Oriented Architecture) services development. Three steps are used for developing an SOA solution: service identification, service specification and finally service realization. In this paper we propose a method called AMSI (Automatic Model-Driven Service Identification) that automatically identifies the architecturally significant elements from a high level business process model to specifying service model artifacts. The main goal of this work is to support the automation of the development process of service-oriented enterprise information system.

Keywords: Business Process Modeling, Model-Driven Development, Service-Oriented Architecture, Service Identification.

1 Introduction

Bridging the gap between Enterprise Modeling methods and Semantic Web services is an important yet challenging task. For organizations with business goals, the automation of business processes as Web services is increasingly important, especially with many business transactions taking place within the Web today [7].

Nowadays, the enterprises are organized in networks, in which various actors can be interacting. The competitiveness of these companies is deeply related to the capacity to structure, share and exchange knowledge with the participants in the collaborative network. This need to exchange knowledge obliges the companies to evolve their information systems and their applications in order to return them interoperable. The interoperability of enterprise applications allows ensuring the exchange of the

functionalities and the services in a transparent way. Each functionality, service, or data have a specific model. Several transformations of these models are essential to ensure interoperability between the various heterogeneous entities of the enterprise. So that these model transformations become an effective solution for establishing interoperability in a purely heterogeneous environment; it is necessary that they must be guided by a standard modeling framework. The MDA approach (Model-Driven Architecture) provides the bases to support the model-driven interoperability.

The development of an enterprise application to large scale always starts with the highest level abstraction where they are the specification and the representation of the business in the form of business process models. These models must be projected gradually on an adapted architecture to the need for interoperability. Currently, the more adapted paradigm to the realization of the interoperable applications is the service-oriented paradigm because it brings a certain simplification and facilitates the establishment of a reconfigurable collaboration through the dynamic construction of software services. We try, in this paper, to answer the following question: How to ensure a permanent and flexible evolution of enterprise information system when business requirements change? For answer to this question, we proposed a method for deriving automatically a Service Oriented Architecture starting from a high level collaborative business process.

The remainder of this paper is organized as follows. In section 2 we presented a basic concepts needed to understand our approach including Business Process Modeling and Service-Oriented Architecture. In section 3 we presented our approach called AMSI (Automatic Model-driven Service Identification). Section 4 concludes this paper.

2 Business Process Modeling and Service-Oriented Architecture Handshake

Service identification is one of the core elements of the BPM and SOA handshake that reinforces the current mantra that “BPM should be service oriented, SOA should be business process focused, and SOA takes over where BPM leaves the enterprise in a path towards agility” [3].

2.1 Business Process Modeling

The process vision plays a significant role in the theories of the organizations as in the information system field where the process modeling is regarded as a key element of the representation of dynamics. The business process modeling is a prerequisite necessary to design an organizational information system. The business process definition reflects the functional needs implicitly. However, it is not sufficient to just conceive the business activities connected by control flows of the process. To represent the complete whole of the requirements, a process definition must explicitly indicate all the entities which take part in the process. These requirements should be

transformed, without loss of information, in semantic specifications of which different software components can be derived.

2.2 Service-Oriented Architecture (SOA)

Service-Oriented Computing (SOC) is a new paradigm for distributed computing that uses services to support the development of interoperable, evolvable, and distributed applications. Services are autonomous, platform-independent entities that can be described, published, discovered, and loosely coupled by using standard protocols. Service-Oriented Architecture is the main architectural style for SOC.

The main idea of Service-Oriented Architecture is the restructuring of enterprise information systems into loosely coupled, independent services. These services should allow the reuse of existing implemented functionality in order to minimize the time between design and implementation when business requirements change. The key challenges in developing the service oriented systems are the mapping of business processes models into service models. Service models play an important role during service-oriented analysis and design phases. According to the IBM SOMA [1], service-oriented modeling lifecycle has three main phases:

Service identification. This phase is about identifying the architecturally significant elements of the target solution. The output artifact of this phase is analysis-level service model.

Service specification. This phase is about describing a service: what it offers, what it requests and how it is exposed. It also describes dependencies with other services, service composition, and service messages. The main model related to this phase is the design-level service model.

Service realization. This phase is about providing a solution for a particular service. We represent here, how a service is realized. The model related with this phase is the design model. This model has to be traced back to the service model, because it represents its realization.

3 Automatic model-driven service identification

Service identification is one of the main activities in the modeling of a service-oriented solution, and therefore errors made during identification can flow down through detailed design and implementation activities that may necessitate multiple iterations, especially in building composite applications. According to [1], the initial activity in the development of a new SOA solution is the service identification. The result of the service identification is a set of candidate services.

The first stage in the service identification process is the modeling of a high level business process that is represented as a CIM model. Metadata are added to the CIM model in the second stage. This operation is based on a whole of generic concepts stored in the process model ontology. The third stage allows the transformation, after the interrogation of the process model ontology, of the input business process model into an executable process model expressed as a PIM model. In the fourth stage, the service identification engine generates a whole of candidate's services for

implementing the input process. These services are grouped in a set of clusters (composed services) by a multi-objective clustering algorithm (cf. **Fig. 1.**)

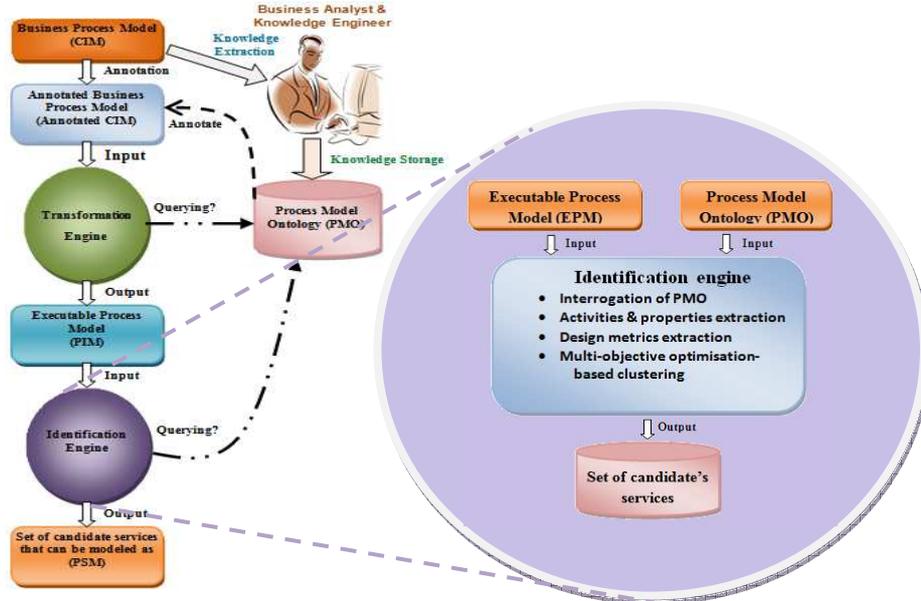


Fig. 1. Automatic Model-Driven Service Identifier

3.1 Business Process annotation

In this stage, complementary information is added to the business model elements such as the nature of the activities (manual, semi-automatic, automatic), the composition of the activities (decomposable, atomic activity), the goal of each activity etc. All knowledge's about the initial business process are extracted and stored in the Process Model Ontology by business analyst and knowledge engineer.

The PMO is based on two principles: to unify the different existing Business Process Metamodels, and to provide the necessary properties for deriving software services from a height level business processes. The PMO captures generic concepts associated with business processes and the relationships among them. To facilitate the extraction of the multiple views on a process model the PMO allows a business analyst and a knowledge engineer to mark the visibility of activities to different collaboration roles. Thus various views on the business process model can be extracted. The PMO will capture the notions of process models at the business level. This ontology defines concepts like *Process*, *Activity*, *Event*, *Control Node*, *Message Flow*, *Sequence Flow*, etc. and the relationships between them. It is regarded as a generic metamodel of business process that is used as a knowledge base for deriving software services from a height level business process model.

Our proposition is that constructing the PMO through a careful analysis of existing reference metamodels, guarantees the representational width of the ontology, i.e. that

all existing business process models can be represented and all software services can be extracted from it.

3.2 Business Process transformation

The first step for identifying SOA services is the business process modeling, and the business functionalities are understood, then the service identification step starts. We cannot directly transform a high level business model into SOA solution because it is independent of any computation specification and it comprises manual, semi-automatic and automatic activities. As well as the high level activities have a great granularity. The same business activity can be transformed into several SOA services. Thus it is necessary firstly to transform the high level business process into an intermediate process called executable process in order to identify the candidate services. The PMO is queried for transform the annotated business process model into an executable process. During this stage, the transformation engine executes the following operations:

- **Rename a business activity:** In the simplest case, a small business process activity is mapped to exactly one SOA service
- **Split a business activity into several:** Service-oriented system requires a strict distinction between activities with user interface (Human Activities) and Service Calls. One business activity can be mapped into several activities in the executable process (Human Activities that need a user interface and automatic services)
- **Merge two business activities in only one:** If small activities of a continuous sequence are always realized by the same person, it makes sense to merge them into one SOA service
- **Insert a new activity in the executable process:** Additional activities, such as authentication service, are necessary for the normal execution of executable process
- **Remove an activity from the business process:** A business process model often contains activities whose execution should not be controlled and monitored by a business process management system. Consequently these activities shall not be to be implemented as SOA services, but are important at business model level for calculating processing times and simulating process costs.

3.3 Service identification

The identification engine queries the ontology and takes the executable process as input in order to generate automatically the candidate services. In this phase, a set of design metrics which satisfy business goals should be calculated from the PMO such as cohesion, coupling, granularity, maintainability, and reusability of activities, etc. that are considered as input parameters for classifying the candidate services in a groups (composite services). The identification of the services corresponding to the executables activities is possible via their names. The identification engine searches the name of the activity in the concepts taxonomy of the ontology and extracts the properties of the activity and its relations with the other concepts of ontology for

calculate design metrics. After this research phase, the identification engine generates a set of candidate services equivalents (implementing) to the activity in question. Thus their initial descriptions (name of the service, names of the interfaces, etc.).

The service identification engine use relationships between individuals of the concepts *Activity*, *Resource*, and *Participant* for calculate different design metrics such as *Service Cohesion*, *Service Coupling*, *Service Granularity*, *Service Maintainability*, and *Service Reusability*. These design metrics are used as input parameters in the clustering algorithm that generates as output a set of optimal service clusters. We can formulate our identification algorithm as a multi-objective optimization problem that classifies candidate's services according to optimal values of design metrics.

4 Conclusion and outlook

In this paper we outlined some of our initial work in the development of AMSI, a method for identifying automatically candidate SOA services from a high level business process. The method defines how a high level business process should be transformed into an executable process in order to identifying SOA services. Our automatic service identifier uses a Process Model Ontology (PMO) to annotate the business process model. The annotated business process model is used as input of a transformation engine which transforms it, after the interrogation of ontology, into an executable process. Finally an identification engine querying the ontology and take the executable process as input in order to generate the candidates services automatically. Currently, our work is at formalization stage. Future work is mainly related to the implementation and evaluation of our approach.

5 References

1. Arsanjani: (SOMA) Service-Oriented Modeling and Architecture: How to identify, Specify, and Realize services for your SOA. IBM developer Works (2004)
2. Fareghzadeh N.: Service Identification Approach to SOA Development. In: Proceedings of World Academy of Science, Engineering and Technology, vol. 35, pp. 258–266 (2008)
3. Inaganti S. and Behara G. K.: Service Identification: BPM and SOA Handshake. BPTrends, March (2007)
4. Jamshidi P, Mansour S, Sedighiani K, Jamshidi S, Shams F.: An Automated Service Identification Method. Technical Report, Department of Electrical and Computer Engineering, Shahid Beheshti University, (2012)
5. Jamshidi P., Sharifi M., and Mansour S.: To Establish Enterprise Service Model from Enterprise Business Model. IEEE International Conference on Services Computing (2008)
6. Klose K., Knackstedt R., Beverungen D.: Identification of Services - A Stakeholder based Approach to SOA Development and its Application in the Area of Production Planning. In: ECIS 2007, pp. 1802–1814 (2007)
7. Nadarajan, G., and Chen-Burger, Y.-H.: Translating a Typical Business Process Modeling Language to Web Services Ontology through Lightweight Mapping. IET Software In: Formerly IEE Proceedings Software, Vol. 1, Issue 1, p.1-17, Feb (2007)

Relations extraction on patterns lacking of Resulting Context

Asma HACHEMI¹, Mohamed AHMED-NACER¹

¹Computer science department, USTHB, Algiers, Algeria

ashachemi@usthb.dz, anacer@mail.cerist.dz

Abstract. Many software patterns are available nowadays. They allow the reuse of proved solutions in various areas of software engineering, but are expressed in different formalisms. This diversity is detrimental for patterns reuse, since it is difficult to compare and compose heterogeneous patterns (patterns expressed in different formalisms). Moreover, patterns composition is based on inter-patterns relationships, that are difficult to discern if they are not explicit. Thus, an automatic method that extract non explicit relations between patterns even if those latter are heterogeneous, becomes a necessity. In this context, we improve an existing method of automatic inter-patterns relations analysis. As many patterns lack of resulting context, our aim is to enable that method to extract relations on this kind of patterns.

Keywords : Patters formalisms, inter-patterns relationships, automatic relations extraction.

1 Introduction

Nowadays, the WWW supplies an increasing amount of knowledge covering diverse domains. Among others, it supplies a large number of software patterns that are a formidable tool allowing the reuse of proved solutions, in various areas of software engineering. A software pattern presents an issue to a recurrent problem, by offering a proven solution. Software patterns need to be composed, in order to solve complex problems that are not dealt by a single pattern. Inter-patterns relationships are on the basis of the patterns composition. However, it is difficult to discern these relations if they are not explicit in each pattern. Moreover, even if those relations are explicit, they are limited to intra-catalog relationships.

Indeed, a pattern is expressed through a pattern formalism, which is a syntactic structure of the pattern content. The majority of pattern formalisms in the literature differ in the number and degree of detail of their items. So, it is difficult to interpret and compare heterogeneous patterns. It is also difficult to compose them into larger solutions; a fact which is detrimental for patterns reuse.

This paper presents our approach that improves an existing method of automatic inter-patterns relations analysis. This method is the first automatic approach which handles relations between heterogeneous patterns, cross different catalogs. The aim of our improvement is to enable that method to handle more patterns formalisms; specially, those formalisms lacking of resulting context. So, works related to inter-patterns relationships extraction are described in section 2. Our approach for automatic relations analysis on patterns lacking of resulting context is presented in section 3. Finally, we conclude this paper and give some research perspectives in section 4.

2 Related works

Many researcher were interested in defining inter-patterns relationships, like [1], [2], [3], [4], ... but very few works treat relationships extraction. In this area, Prabhakar *et al.* [15] propose a graphical model called Design Decision Topology Model, in order to represent design patterns and extract relationships between them; unfortunately, this work is limited to the analysis of relations on design patterns only. The method of Kubo *et al.* [10] is the first automatic approach able to extract relations on heterogeneous patterns, belonging to different catalogs, and is the only approach able to deal with different kinds of software patterns. Kubo *et al.* method is an interesting approach based on its own pattern model (consisting of *Starting Context*, *Forces*, *Resulting Context*), and on several text processing techniques (stop word removal [11], stemming [12], the TFIDF term weighting [13], vector space model [11] and the cosine similarity). However, Kubo *et al.* method is not able to treat patterns which lack of Resulting Context, so we propose to improve it to extract relations on this kind of patterns.

4 Our approach

Our approach towards an automatic way to analyze relations between patterns is based on Kubo *et al.* method, that we propose to improve in terms of the pattern forms handled. As many patterns do not express the Resulting Context explicitly (like those in [5], [7], [8], [9], [14], ...), they cannot be represented and analyzed by the model of Kubo *et al.*. Therefore, a value-added of our approach is the proposition of a solution to represent this kind of patterns and analyze relations on them.

Resulting Context is the result or product generated by the pattern application. So, each pattern has its resulting context either explicit in a dedicated section, or implicitly given in the Solution section. Here is an example of a resulting context expressed within the Solution : The pattern *Declare Before First Use* [8] aims to ensure that the declaration of an element is positioned before the reference to it (in an XML document). The resulting context of this pattern is the increase of the probability of treating the document in a single pass. Actually, this resulting context is expressed within the Solution section : "*This gives the processing software a better chance of doing a single pass traversal of the document*".

Our idea is to overcome the absence of a dedicated section for Resulting Context by using the Solution section, so that one be able to represent the third element of the pattern model (Resulting Context). A such use of the Solution does not alter the significance of the relationships that we are interested in (*Starting-Starting* [10], *Same*[10], *Resulting-Starting*[10], *Uses*[6] and *Refines*[6]). The reasons are :

- The relation *Starting-Starting* : The analysis of this relation is based on the Starting Context [10]. The Resulting Context is used only to represent the pattern in the Kubo *et al.* model. Thus, the use of the Solution section instead of the Resulting Context does not affect the meaning of this relation.

- The relations *Refines* : The analysis of this relation is based on the Starting Context and Forces [6]. The Resulting Context is used only to represent the pattern in the Kubo *et al.* model. So, the use of the Solution instead of the Resulting Context does not alter the meaning of this relationship.
- The relation *Same* : When two patterns share the same Starting Context and the same Solution, this means that these two patterns deal with the same problem and provide the same result. Thus, we can use the Solution section instead of the Resulting Context to analyze this relation.

For example, let's consider the pattern Navigation Tabs [9] (called P1) which is Same as the pattern Navigation Tabs [7] (called P2). This relation is given by the author of [9], so we consider it as correct and process the analysis using our method. This latter starts by eliminating stop words [11] and applying the Stemmer [12] on the elements of P1 and P2. After that, the terms of these elements are weighted using the TFIDF method [13], and the cosine similarity is calculated as explained in [10]. Also, our method checks the inclusion [6], either it is true or false between each couple of elements. We obtain the results shown in Table 1.

Table 1. Results of comparisons between patterns elements

Compared Elements	Results
SC of P1 and SC of P2	Similarity = 0,934 SC of P1 includes SC of P2 = False SC of P2 includes SC of P1 = False
RC of P1 and RC of P2	Similarity = 0,956 RC of P1 includes RC of P2 = True RC of P2 includes RC of P1 = False
Forces of P1 and Forces of P2	Similarity = 1 Forces of P1 includes Forces of P2 = False Forces of P2 includes Forces of P1 = False
RC of P1 and SC of P2	Similarity = 0,136
RC of P2 and SC of P1	Similarity = 0,135

Since we obtain the similarity and inclusion results, we calculate the value of each relation between P1 and P2, using the definition of each relation (Uses [6] Refines [6], Same [10], Resulting-Starting [10], Starting-Starting [10]). For instance, the relation Starting-Starting between the patterns P1 and P2 is represented by the similarity value of the Starting Contexts of these patterns. The results of the relations analysis are shown in Table 2.

Table 2. Results of relations analysis

Relationship	Its value
P1 Uses P2	0
P1 Refines P2	0
P2 Uses P1	0
P2 Refines P1	0
Same	0,945

Starting-Starting	0,934
Resulting-Starting (P1 then P2)	0,136
Resulting-Starting (P2 then P1)	0,135

Finally, as in Kubo *et al.* method, the strongest relation of the eight types (P1 Uses P2, P1 Refines P2, P2 Uses P1, P2 Refines P1, Same, Resulting-Starting (P1 then P2), Resulting-Starting (P2 then P1), Starting-Starting) is assumed as the representative relationship. So we conclude in this case that the patterns P1 and P2 are *Same*.

- The relation *Resulting-Starting* : When the Solution of a pattern and the Starting Context of another are similar, this means that the second pattern (that we are interested in its Starting Context) can be applied after the first pattern (that we are interested in its Solution), because the solution of the first one provides the preconditions necessary to apply the second pattern. So, we can use the Solution instead of the Resulting Context to analyze the *Resulting-Starting* relation.

For example, let's consider the patterns Titled Sections [14] (called P1) and Closable Panels [14] (called P2) related by the Resulting-Starting relation. This relation is given by the author of these patterns, so we consider it as correct and process the analysis using our method. We compare the elements of P1 and P2 and obtain the results shown in Table 3.

Table 3. Results of comparisons between patterns elements

Compared Elements	Results
SC of P1 and SC of P2	Similarity = 0,156
	SC of P1 includes SC of P2 = False
	SC of P2 includes SC of P1 = True
RC of P1 and RC of P2	Similarity = 0,119
	RC of P1 includes RC of P2 = True
	RC of P2 includes RC of P1 = False
Forces of P1 and Forces of P2	Similarity = 0,097
	Forces of P1 includes Forces of P2 = False
	Forces of P2 includes Forces of P1 = True
RC of P1 and SC of P2	Similarity = 0,248
RC of P2 and SC of P1	Similarity = 0,060

After that, we calculate the value of each relation between those patterns. The results are shown in Table 4.

Table 4. Results of relations analysis

Relationship	Its value
P1 Uses P2	0
P1 Refines P2	0
P2 Uses P1	0
P2 Refines P1	0,127
Same	0,137

Starting-Starting	0,156
Resulting-Starting (P1 then P2)	0,248
Resulting-Starting (P2 then P1)	0

Finally, considering the strongest relationship, we conclude that P1 and P2 are related via the relationship *Resulting-Starting*.

- The relation *Uses* : When the Starting Context and the Solution of a pattern are respectively included in the Starting Context and the Solution of another pattern, then this means that the second pattern *Uses* the first one. So, we can utilize the Solution instead of the Resulting Context to analyze this relation.

For example, let's consider the pattern Extras On Demand [14] (called P1) which Uses the pattern Closable Panels [14] (called P2) according to the author of these patterns. So, we consider this relation as correct and process the analysis via our method. We compare the different elements of P1 and P2 and obtain the results shown in Table 5.

Table 5. Results of comparisons between patterns elements

Compared Elements	Results
SC of P1 and SC of P2	Similarity = 0,216
	SC of P1 includes SC of P2 = True SC of P2 includes SC of P1 = False
RC of P1 and RC of P2	Similarity = 0,223
	RC of P1 includes RC of P2 = True RC of P2 includes RC of P1 = False
Forces of P1 and Forces of P2	Similarity = 0,130
	Forces of P1 includes Forces of P2 = True Forces of P2 includes Forces of P1 = False
RC of P1 and SC of P2	Similarity = 0,073
RC of P2 and SC of P1	Similarity = 0,114

After that, we calculate the value of each relation between those patterns. The results are shown in Table 6.

Table 6. Results of relations analysis

Relationship	Its value
P1 Uses P2	0,220
P1 Refines P2	0,173
P2 Uses P1	0
P2 Refines P1	0
Same	0,220
Starting-Starting	0,216
Resulting-Starting (P1 then P2)	0,073
Resulting-Starting (P2 then P1)	0,114

Finally, considering the strongest relationship, we conclude that the pattern P1 *Uses* the pattern P2.

5 Conclusion and perspectives

Our way of looking at the analysis of relations between patterns is based on Kubo *et al* method. We improved this method to enable it dealing with patterns which do not give their *Resulting Contexts* in an explicit manner. Our idea consisted of using the *Solution* section. As we have explained earlier, a such use does not alter the significance of the different relations treated. Some other improvements can be addressed to face the drawbacks inherent to Kubo *et al.* method, and to offer more benefits for patterns composition. Such as :

- The block HTML Analysis of the method is limited to the treatment of patterns expressed in HTML. This block can be extended to deal with patterns expressed in other ways.
- The method can be improved to treat patterns lacking of Starting Context and/or Forces, which are two necessary elements to represent patterns in the model of Kubo *et al.*
- The method can be extended to offer the functionality of Patterns Retrieval, which provides to a user having a particular problem, all available patterns that treat this problem.

References

1. Zimmer, W.: Relationships between design patterns. In: PLoP (1994)
2. Conte, A., Fredj, M., Giraudin, J.P., Rieu, D.: P-Sigma : un formalisme pour une représentation unifiée de patrons. In: Inforsid, Genève (2001)
3. Gnatz, M., Marschall, F., Popp, G., Rausch, A., Schwerin, W.: The Living Software Development Process. Journal Software Quality Professional, Volume 5, Issue 3 (2003)
4. Henney, K.: Patterns Inside Out. Talk presented at Application Development, London (1999)
5. Crumlish, C., Malone, E.: Designing social interfaces. available at http://www.designingsocialinterfaces.com/patterns/Main_Page (2009)
6. Hachemi, A., Ahmed-Nacer, M.: Primary inter-patterns relationships analysis. In: CARI2012, Algiers (2012) unpublished.
7. Yahoo design pattern library. <http://developer.yahoo.com/ypatterns/>.
8. Develop effective XML documents using structural design patterns. <http://www.xmlpatterns.com/> (2000)
9. Lammi, J., Varjokallio, M., Hocksell, J.: A user interface design pattern library. <http://www.patternry.com>.
10. Washizaki, H., Kubo, A., Takasu, A., Fukazawa, Y.: Analyzing Relations among Software Patterns based on Document Similarity. In: IEEE Internationale Conference on Information Technology : Coding and Computing (2005)
11. Salton, G., McGill, M.: Introduction to Modern Information Retrieval. MG-Hill Inc., NY (1983)
12. Paice, C.: Another stemmer. SIGIR Forum, Vol. 24, No. 3, pp. 56–61 (1990)
13. Salton, G., Yang, C.: On the specification of term values in automatic indexing. Journal of Documentation, Vol. 29, pp. 351–372 (1973)
14. Tidwell, J.: Designing interfaces - Patterns for effective interaction design. available at <http://designinginterfaces.com/firstedition/> (2005)
15. Prabhakar, T.V., Kumar, K.: Design Decision Topology Model for Pattern Relationship Analysis. In: Asian PLOP 2010 Tokyo (2010)

Reverse Engineering Process for Extracting Views from Domain Ontology

Soraya Setti Ahmed¹ and Sidi Mohamed Benslimane²

¹ Mascara University, Computer Science Department, Algeria
{settoraya@yahoo.fr}

² Djillali Liabes University, Research Laboratory, Computer Science Department
Sidi Bel Abbes, Algeria
benslimane@univ-sba.dz

Abstract. Ontology Modularization is one of the techniques that bear good promises of effective help towards scalability in ontology design, use, and management. The development of proper ontological modules should provide a mechanism for packaging coherent sets of concepts, relationships, axioms, and instances, and a means for reusing these sets in new environments, possibly heterogeneous with respect to the environment the modules were first built. The main contribution of this paper is to describe an approach for extracting views from domain ontology using existential dependency (ED) by reverse engineering process. The extraction process based on ED could provide a coherent fragment of ontology parts together with transitive closure of dependant parts. The goal of reverse engineering process is to output a possible conceptual model, which is more readable to extracting the views, on the basis of the code in which the ontology is implemented. Thus, a set of translation rules is used to convert owl ontology in a UML class diagram.

Keywords: Modularization, Reverse Engineering, Existential Dependency, Ontology Views, Guizzardi Metamodel, UML profiles, OWL.

1 Introduction

Ontology Modularization techniques identify coherent and often reusable regions within an ontology. The ability to identify such modules, thus potentially reducing the size or complexity of an ontology for a given task or set of concepts is increasingly important in the Semantic Web as domain ontologies increase in terms of size, complexity and expressivity[1].

In conceptual modelling, the Foundational Ontology is needed as domain independent theoretical basis to guide and validate models of particular domains, as using of right modelling concepts and rules is making a great influence on the quality of Information Systems [2]. For such purpose, the transformations between conceptual models (expressed, for example, in UML) and ontological models, expressed in ontological languages (for example, OWL) are needed. The extraction process using lightweight ontologies like UML and OWL generates strictly unnecessary classes and individuals, for this reason the first step of our approach is based on the reverse engineering process whose goal is to output a possible conceptual model, which is more readable to extracting the views, on the basis of the code in which the ontology

is implemented [3] and [4]. Thus, a set of translation rules is used to convert owl ontology in a UML class diagram.

The rest of the paper is organized as follows: In section 2 we describe the architecture and the main steps of our approach. Section 3 introduces implementation of our system. Finally, we conclude this paper and outline our future work in section 4.

2 Our approach

In this section, the global architecture of our system is presented. Figure 1 illustrates the main steps of the proposed approach.

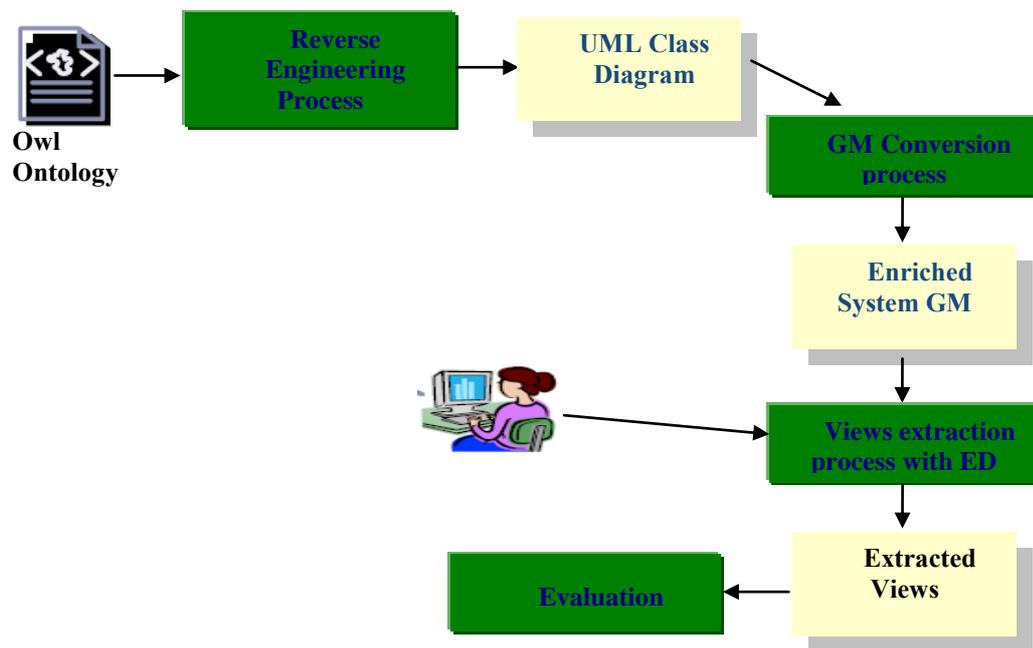


Fig.1. Main steps of our approach

2.1 Reverse engineering process

The designer initiates transformation of domain ontology described in OWL file into UML class diagram by reverse engineering transformation. At first, system transforms ontology classes, then object and data type properties, and finally constraints.

Algorithm of mapping OWL Ontology to UML class diagram

```

Input: OWL file ontology
Output: UML class diagram
Begin
For all OWL class (concept) defined into ontology do
Create UML class with same name.
    If the ontological class is sub class of restriction then
For all restriction do
    If type of this restriction is : cardinality, minCardinality or maxCardinality then
        transforme these in multiplicities for propriety specified on Property of restriction
    Else
Define the name of role of toClasse classe with object property name specified in onProperty.
    Endif
Endfor
Endif
    If this class is sub class of other class then
        Define UML generalisation element

    Endif
Endfor
For all DataTypeProperty Do
    Create an attribute whose domain is class and whose range is the type of property
Endfor
For all ObjectProperty Do
    Create UML association whose domain is class and whose range is class
Endfor
End

```

Table 1 summarised the important rules of mapping Owl2Uml

OWL constructor	UML constructor
DatatypeProperty	Property ownedAttribute
ObjectProperty	Property memberEnd
InverseOf subClassOf, subPropertyOf	Binary Association superClass,Genearlization
Cardinality, MinCardinality, MaxCardinality	Multiplicities
Ontology Union, Intersection	Package Ontology Generalization isDisjoint, isCovering
one of	Enumeration
Individual	Instance

Table 1. Rules of mapping Owl2Uml

2.2 GM conversion process

A Conversion tool implements a transformation from UML class diagram obtained in step 2 to Guizzardi Metamodel (GM) [5]. We introduce a formal ontology, the GM to resolve some highlighted anomalies. We adopt GM to enrich our diagram with several existential dependencies and to define some extraction rules under tree main structural relationships in GM such as association, subtype and part whole.

Guizzardi's concepts *kind*, *subkind*, *phase*, *role* and *relator* are all represented as stereotypes of the UML metaclass *Class*, for example, and all inherit the semantics of *Class* in UML. Any UML metaclass can be stereotyped.

Some examples of transformation rules:

Rule1: In UML Class Diagram, a collection of instances of classes are, respectively, instances of UML G-M profiles including concrete classes (<<kind>>, <<subkind>>, <<quantity>>, <<collective>>, <<phase>> and <<role>>).

Rule 2: In UML Class Diagram, concrete classes (and their instances) are related via UML G-M profiles including properties (<<mediation>>, <<derivation>>, <<characterisation>>, <<material>> and <<formal>>) as well as complex objects or part-whole (*subQuantityOf*, *subCollectionOf*, *memberOf*, *componentOf*).

Rule3: In UML CD, concrete classes (and instances) can be categorised accordingly by UML G-M profiles via abstract classes (<<category>>, <<roleMixin>> and <<mixin>>) and other rules.

2.3 Views extraction process with ED

This step present extraction cases and rules for how these views can be extracted using existential dependency, especially where the ontology is constructed using the GM formal ontology. We note that user in this case should specify certain individuals and classes. The extraction process produces a more focused and smaller portion and reduces the costs to the user. There are several systems under the 3G-M like systems of (kind, phase, role, mixin, quality, formal, relator, material, mode, Q-parthood, C-parthood, M-parthood and system of CF-parthood). All these systems contribute to the ED.

Some examples of extraction cases and rules:

System of Kind: Super kind is Mandatory (+M), subkind is mandatory (+M), siblings are optional (-M): This case applies general rules "requires all superclasses" and "siblings optional".

System of Relator: A relator is mandatory (+M) and mediated classes are mandatory (+M)

A mediated class is mandatory (+M), a relator is mandatory (+M) and a pair of mediated classes is mandatory (+M): Every instance of mediated class does not make sense without every instance of another (pair) mediated class with the relator mediates to.

System of Role:

Superkind is an ultimate substance sortal that supplies a principle of identity. Superkind does not make sense without the roles and vice versa. Supermixin (role mixin) is **optional** (-M) since it does not supply a principle of identity.

An application **may not** (-M) need sibling roles since they carry an incompatible principle of identity supplied by its superkind respectively. An individual must be not a member of its siblings. This case applies general rules: **“some superclasses optional”** and **“siblings optional”**.

Superkind is mandatory (+M), role is mandatory (+M), supermixin is optional (-M) and sibling roles are optional (-M).

2.4 Evaluation

Correctness of the extracted views translates the fact that no information is lost in the process.

Information preservation may be defined as the fact that the result of a query addressed to the collection is functionally (i.e., not from a performance viewpoint) the same as the result of the same query addressed to the original ontology.

3 Implementation

The architecture of our system has been conceived to follow a Model-Driven Approach. In particular, we have adopted the OMG MOF (Meta-Object Facility) metamodeling architecture [6]. In order to describe constraints in UML/MOF (meta) models, the OMG also proposes the declarative formal language OCL (Object Constraint Language) [7]. On the formalization of the UML profile we have used OCL expressions mainly to: define how derived attributes/associations get their values; define default values of attributes/associations, *i.e.*, define their initial values; specify query operations and specify invariants, *i.e.*, integrity constraints that determine a condition that must be true in all consistent system states.

The full set of OCL expressions including: OCL expressions to specify derivation rules; OCL expressions to define default values; OCL expressions to specify operations created to support some OCL derivation rules and invariants, and invariants to model the constraints stated on the UML profile. An example of an OCL invariant representing the essential parthood axioms is shown in the code below. One can notice that in this expression the modal existential dependence constraint of essential parthood from UFO (Unified Foundational Ontology) is emulated via the existence condition (lower cardinality ≥ 1) plus the immutability constraint (`isReadOnly = true`).

```
Inv: if (self.isEssential = true) then self.target-> forAll(x | if x.ocIsKindOf(Property)
then ((x.ocAsType(Property).isReadOnly = true) and ((x.ocAsType(Property).lower
>= 1)) else false endif) else true endif
```

4 Conclusion and future work

In This paper we describe our approach for extracting views from domain ontology by reverse engineering process witch consists of transforming the OWL file ontology of E-Tourism into UML class diagram. there is an implementation of the metamodel proposed by Guizzardi [8] by using MDA (Model-Driven Architecture) technologies, in particular, the OMG MOF (Meta-Object Facility) and OCL (Object Constraint Language).

Future work will concern the implementation of process of extracting views with rules proposed here to confirm the useful of our approach.

References

1. Doran,P.,Tamma, V.,Payne,T,R Pal misano,I. : An entropy inspired measure for evaluating ontology modularization. in :5th International conference on knowledge capture(KCAP'09).(2009)
2. Rajugan,R.,Tharan,S.,T.S.Dillon.: Modeling views in the layered view model for XML using UML, journal of Web information System 2 (2006) 95-117.
3. Chikofsky,E.J.,Cross II, J. H., 1990 Reverse engineering and design recovery: a taxonomy. Software Magazine 7 (1990) 13-17.
4. Fernandez-Lopez,M.,Gomez Pérez,A.: Overview and analysis of methodologies for building ontologies. *The Knowledge Engineering Review*, Vol. 17:2, 129– 156. © 2002, Cambridge University Press
5. Guizzardi, G., “On Ontology, ontologies Conceptualizations, Modeling Languages, and (Meta)Models”, *Frontiers in Artificial Intelligence and Applications, Databases and Information Systems IV*, Olegas Vasilecas, Johan Edler, Albertas Caplinskas (Editors), ISBN 978-1-58603-640-8, IOS Press, Amsterdam, (2007).
6. Object Management Group (OMG):Meta Object Facility MOF core Specification, v2.0,Doc # ptc/06-01-01 (2006)
7. Object Management Group (OMG): Object Constraint Language, v2.0, Doc.# ptc/06-05-01 (2006)
8. Guizzardi,G.: *Ontological Foundations for Structural Conceptual Models*, Ph.D. Thesis, University of Twente, The Netherlands (2005)

Multi-Agents Model for Web-based Collaborative Decision Support Systems

Abdelkader Adla, Bakhta Nachet and Abdelkader Ould-Mahraz

Department of Computer Science, University of Oran
Oran, Algeria
{adla abdelkader, nachet.bakhta}@univ-oran.dz

Abstract. In this paper, we propose a Multi-agent model for web-based collaborative decision support system in which a facilitator and group decision makers are supported by agents. The integrated agents into web-based collaborative decision support system constitute a collection of autonomous collaborative problem solving intelligent agents, goal-directed, proactive and self-starting behaviour; interact with other agents and humans in order to solve problems. Specifically, agents were used to collect information and generate alternatives that would allow the user to focus on solutions that were found to be significant. The decision making process, applied to the boilers defects in an oil plant, relies on a cycle that includes recognition of the causes of a defect (diagnosis), plan actions to solve the incidences and, execution of the selected actions.

Keywords: Collaborative decision making, Web-based decision support systems, Multi-agent systems, Decision support

1 Introduction

As organizations seek to adapt in a world of rapid change, decision making becomes increasingly dynamic and complex. Collaborative decision support systems provide a means by which a larger number of organizational stakeholders can efficiently and effectively participate in the decision making process. A greater number of organizational members participating in the decision making process logically leads to a better decision. The resulting decision should benefit by the richness of knowledge provided by the greater representation of organizational members. A success factor critical to this involvement is the successful organization of massive amounts of information generated by such a group.

On the other hand, the Distributed Artificial Intelligence (DAI), which is commonly implemented in the form of intelligent agents, offers considerable potential for the development of information systems and in particular Decision Support Systems (DSS). Widely range applications domains, in which agent solution is suggested, are being applied or investigated [Cheung, 2005]. This is because of the reason that intelligent agents have a high degree of self-determination capabilities, and they can decide for themselves when, where, and under what condition their action should be performed. Intelligent agents have the promise to provide timely assistance in various areas of such environments as information gathering, information dissemination, monitoring of team progress and alerting the team to various unexpected events.

This article takes a multi-agent view of the web-based collaborative decision making process and examines the potential integration of agent technology into a distributed group decision support systems. It considers group participants as multiple agents concerned with the quality of the collaborative decision. We define a facilitator agent as that agent responsible for the overall decision making process. This includes managing the complex negotiation processes that are required among those participants collaborating on decision making.

We take first a literature survey of some related work in section 2 and 3. Then we propose a multi-agent architecture for web-based collaborative decision support systems in section 4. We also present some implementations issues in section 5. Finally, we conclude with future research direction in section 6.

2 Collaborative Decision Support Systems

Decision aid and decision making have greatly changed with the emergence of information and communication technology (ICT). Decision makers are now far less statically located; on the contrary they play the role in a distributed way. This fundamental methodological change creates a new set of requirements: web-based collaborative decisions are necessarily based on incomplete data. “web-based collaborative decision” means that several entities (humans and machines) cooperate to reach an acceptable decision, and that these entities are distributed and possibly mobile along networks. Distributed decision making must be possible at any moment. It might be necessary to interrupt a decision process and to provide another, more viable decision.

Collaborative or Group Decision Support Systems (GDSS), which are closely related to DSS, facilitate the solution of unstructured and semi-structured problems by a group of decision makers working together as a team [Ribeiro, 2006; DeSanctis, and Gallup, 1997; Nunamaker, 1997]. Group Decision Support Systems (GDSS) are interactive computer-based environments which support concerted and coordinated team effort towards completion of joint tasks. DeSanctis and Gallup [1997] defined GDSS as a combination of computers, communications and decision technologies working in tandem to provide support for problem identification, formulation and solution generation during group meetings.

Research that studied group decision support systems in the existing literature used mainly face-to-face facilitated collaborative decision support systems. Some of its results may not apply to distributed teams that, it is difficult for distributed teams to arrange face-to-face meetings or to meet at the same time virtually.

3 Multi-Agent Systems

In recent years, there has been considerable growth of interest in the design of a distributed, intelligent society of agents capable of dealing with complex problems and vast amounts of information collaboratively. Various researches have been conducted into applying intelligent agent-based technology toward real-world problems. Furthermore, there has been a rapid growth in developing and deploying intelligent agent-based systems to deal with real-world problems by taking advantage of the intelligent, autonomous, and active nature of this technology. The main benefits of an agent-based approach come from its flexibility, adaptability, and decentralization.

The definition of multi-agent systems (MAS) is well known and accepted as a loosely coupled network of agents that work together to find answers to problems that are beyond the individual capabilities or knowledge of each agent and there is no global control system.

An agent's architecture is a particular design or methodology for constructing an agent. Wooldridge and Jennings refer to an agent's architecture as a software engineering model of an agent [Jennings, 1996]. Using these guidelines, agent architecture is a collection of software modules that implement the desired features of an agent in accordance with a theory of agency. This collection of software modules enable the agent to reason about or select actions and react to changes in its environment.

MAS are software systems composed of several autonomous software agents running in a distributed environment. Beside the local goals of each agent, global objectives are established committing all or some group of agents to their completion. Some advantages of this approach are: 1) it is a natural way for controlling the complexity of large and highly distributed systems; 2) it allows the construction of scalable systems since the addition of more agents become an easy task; 3) MAS are potentially more robust and fault-tolerant than centralised systems.

As is typical with an emerging technology, there has been much experimentation with the use of agents in DSS, but to date, there has been little discussion of a framework or methodological approach for using agents in DSS, and while DSS researchers are discussing agents as a means for integrating various capabilities in DSS and for coordinating the effective use of information [Whinston, 1997], there has been little discussion about why these entities are fit for such tasks.

4 A Multi-Agent Architecture for Web-based Collaborative Decision Support Systems

We started our framework with the following fundamentals:

1. The first fundamental, in keeping with [Adla et al., 2007], was to segment web-based collaborative decision support systems into two components: Facilitator and participants (decision-makers)
2. The second fundamental we adopted was to include in each collaborative decision support system component an agent to oversee or manage the other agents within the component;

4.1 The Web-based Collaborative Decision Making Framework

In [Adla et al., 2007] we consider the paradigm of web-based collaborative decision-support systems, in which several decision-makers geographically dispersed who must reach a common decision. The networked decision-makers can evaluate and rank alternatives, determine the implications of offers, maintain negotiation records, and concentrate on issues instead of personalities.

In our proposed framework [Adla et al., 2007], the group is constituted of two or several decision-makers (participants) and a facilitator. Each participant interacts with individual DSS integrating local expertise and allowing him to generate one or several alternatives of the problem submitted by the facilitator. The group (facilitator and participants) use the

group toolkit for alternative generation, organization, and evaluation as well as for alternative choice which constitutes the collective decision. Therefore, we view the individual DSS as a set of computer based tools integrating expert knowledge and using collaboration technologies that provide decision-maker with interactive capabilities to enhance his understanding and information base about options through use of models and data processing, and collaborate with him.

Agents were integrated into the DSS for the purpose of automating more tasks for the user, enabling more indirect management, and requiring less direct manipulation of the collaborative decision support system. Specifically, agents were used to collect information outside of the organisation and to generate decision-making alternatives that would allow the user to focus on solutions that were found to be significant. A set of agents is integrated to the system and placed in the collaborative decision support system components, according to our framework [Adla et al. 2007].

4.2 The Multi-Agent Architecture

The goal of Distributed Group Decision making is to create a group of coarse-grained cooperating agents that act together to come to a collective decision. Participants in a collaborative decision making meeting are considered as a set of agents involved in creating a collective decision. These participant agents are involved with the content knowledge of the particular group problem at hand. The responsibility of managing any decision making process is typically put upon a supervisory agent. We call this agent the facilitator. We view the participants as multiple agents responsible for creating the *content* of the decision, and the facilitator as an outside agent responsible for managing the decision process that the participant agents use to come to common decision

For each participant (decision's maker), the following agents are defined:

- DA (Decision-maker Assistant): it's the interface between the participant and the system. During idea (solution) generation stage, a decision maker can use its proper DSS (Decision Support System) through the DA.

- CA (Collaborator Assistant): The role of this agent is devoted exclusively to the collaboration of the decision maker in the process of decision making support. The only interaction it manages is with CRA of the facilitator and does not communicate directly with agents of other decision makers.

For the facilitator side, the following agents are defined:

- FA (Facilitator Assistant): it manages the interface between the system and the facilitator. It provides a private workspace for the facilitator and a public space for the group. It also allows the facilitator to communicate at any time with group members outside the decision making process, helps to establish communications with other system users through their assistants (DA).

- CRA (CooRdinator Agent for the decision making process): It's the central agent of the decision making process. It is supervised by the facilitator via the FA. Its role is to ensure the rules checking and application during the various phases of the decision making process. FA starts the decision making session. The CRA takes in charge the following tasks of this activity. It guides the group through the activity phases.

- MA (Mediator Agent): is requested by the CRA during the alternatives organisation phase. Its role is to refine the alternatives (deletes or merges synonymous, redundant or inconsistent alternatives) and to classify the alternatives as well.

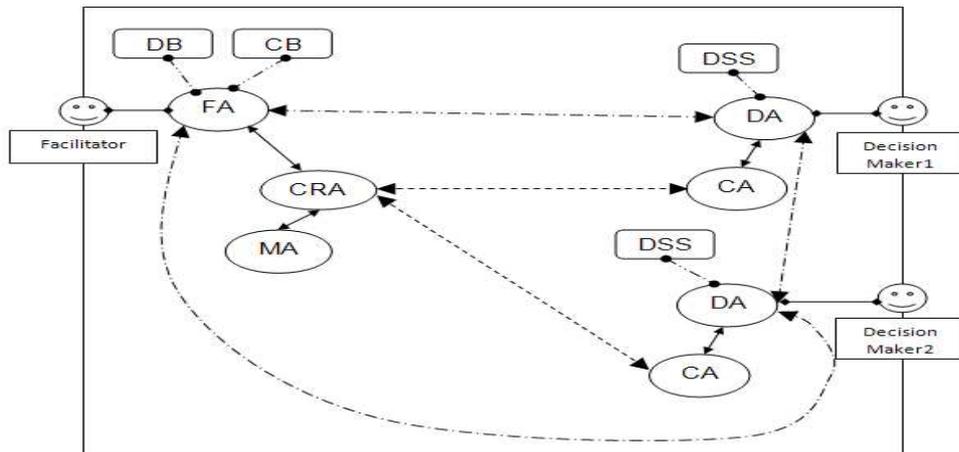


Figure 2: Distributed DSS-MAS logical Architecture

5 Implementation Issues

A prototype of the multi-agent architecture for distributed group decision support system is being implemented in order to generate results that can be analyzed and validate our work. To this end, we have used the FIPA compliant JADE platform to implement our system. Some implementation details are given in the next section.

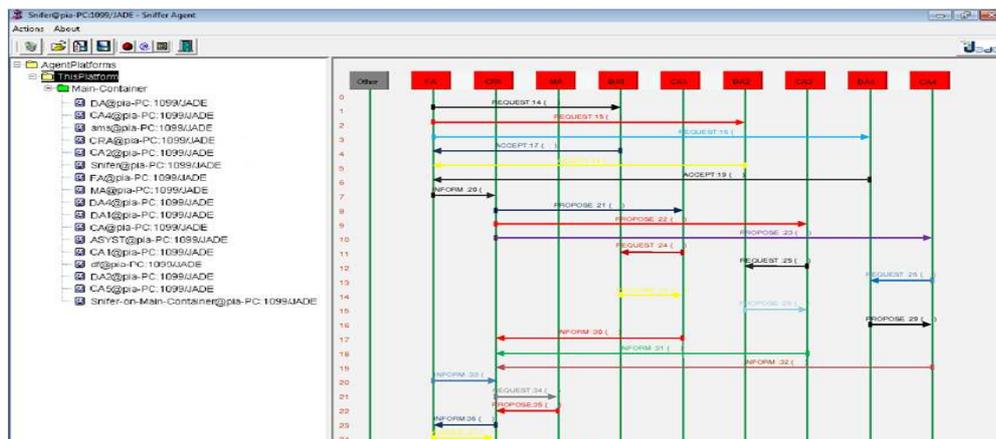


Figure 4: Partial result (sniffer screen)

As depicted in figure 4, a decision group composed of a facilitator and four decision makers collaborate and interact to solve a problem; the decision maker number three doesn't appear on the figure as it's disconnected and does not participate to the decision making session. A partial result of the interactions between agents (JADE's sniffer screen) is given Figure 4.

6 Conclusion

In this paper we presented a web-based collaborative decision support system based on a multi-agent architecture. We have integrated agents into a cooperative intelligent decision system for the purpose of automating more tasks for the decision maker, enabling more indirect management, and requiring less direct manipulation of the DSS. In particular, agents were used to collect information and generate alternatives that would allow the user to focus on solutions found to be significant. Agents are normally used to observe the current situation and knowledge base, and then make a decision on an action consistent with the domain they are in, and finally perform that action on the environment.

The use and the integration of software agents in the decision support systems provide an automated, cost-effective means for making decisions. The agents in the system autonomously plan and pursue their actions and sub-goals to cooperate, coordinate, and negotiate with others, and to respond flexibly and intelligently to dynamic and unpredictable situations.

References

- Adla, A., J-L, Soubie, and P. Zarate, "A Co-operative Intelligent Decision Support System for Boilers Combustion Management based on a Distributed Architecture", *Journal of decision Systems*, Lavoisier, 2007, Vol. 16, pp. 241-263. Systems, Lavoisier.
- Cheung, W. (2005): "An Intelligent decision support system for service network planning", *Decision Support Systems*, Lavoisier, 2005, Vol. 39, pp. 415- 428.
- G. DeSanctis, and B. Gallup, "A foundation for the study of group decision support systems", *Management Science*, 1997, Vol. 13, pp. 1589-1609.
- E. Jennings, "Using intelligent agents to manage business processes", In B. Crabtree and N. R. Jennings editors, In *Proceedings of the 1st international conference on practical applications of intelligent agents and multi-agent technology (PAAM96)*, 1996, pp. 345-360.
- J. Nunamaker, "Lessons from a dozen years of group support systems research", *Journal of MIS*, 1997, Vol. 13, pp. 163-207.
- R. Ribeiro, "Intelligent Decision Support Tool for Prioritizing Equipment Repairs in Critical/Disaster Situations", *In Proceedings of Workshop on Decision Support Systems*, 2006
- Whinston, A. (1997). *Intelligent Agents as a Basis for Decision Support Systems*. *Decision Support Systems*, 20(1).

Agent-based Approach for Mobile Learning using Jade-LEAP

Khamsa Chouchane¹, Okba Kazar², and Ahmed Aloui¹

¹Computer Science Department, Faculty of Sciences
University Hadj Lakhdar 05000 Batna, Algeria

² Computer Science Departement, Faculty of Science And Engineering
Science, University Mohamed Khider 07000 Biskra, Algeria
khamsa.info@yahoo.fr, kazarokba@yahoo.fr, ahmed0725@gmail.com

Abstract. The rapid evolution of mobile and wireless technologies has created a new dimension of modern people's lifestyles; it facilitates their daily activities and summaries distances between them, and allowed them to do several tasks whenever they want and wherever they go. When these technologies started to be used in conjunction with learning a new paradigm has been emerged, it's about mobile learning. Since its emergence it has been raised a lot of attention by researchers whose attempt to propose approaches that address limitations of mobile learning environment. A promising technology which can reduce most of these limits is used in this paper which is mobile agent technology. This paper seeks to provide an agent-based approach for mobile learning systems using jade-LEAP platform.

Keywords: mobile learning, mobile agent, jade-LEAP.

1 Introduction

Mobile learning has emerged as an "anytime anywhere learning". Therefore, learning content and services must be always available and delivered to the learner whenever he wants and wherever he goes. However, mobile learning environment has a number of constraints which may hinder mobile learning applications designers to reach this potential. These constraints are related to the limitations of the mobile devices themselves which have reduced processing power, low memory capability, limited battery life and display capability. However, these limitations are reduced at present, since the exponential growth of mobile devices and adoption of the computer capabilities in those devices. Other limitations are related to the wireless networks which have high latency and transmission delays, and low bandwidth especially with considerable number of users, as a result the size of data exchanged should be optimized. Moreover, wireless link may not be available in permanent way, in addition to the expensive and fragile network connections which creates problems for services designed to operate with fast and reliable and continuously open connection.

The other side, mobile agents are a promising solution that can reduce problems

mentioned above; furthermore they facilitate introducing automatic and dynamically adaptive learning methods. Thus, we propose an agent based approach for an effective mobile learning systems using jade-LEAP platform. The remainder of this paper is organized as follows. First, we present an overview of jade-LEAP platform. Second, we describe in detail our proposal. Finally, our conclusion and future work is given.

2 Jade-LEAP in mobile devices

JADE-LEAP (Lightweight and Extensible Agent Platform) is an extension of JADE platform that can be deployed not only on PCs and servers, but also on lightweight resource devices such as Java enabled mobile phones. In order to achieve this, JADE-LEAP can be shaped in different ways corresponding to the two configurations of the Java Micro Edition and the Android Dalvik Java Virtual Machine: [1]

- **Pjava:** to execute JADE-LEAP on handheld devices supporting J2ME CDC or PersonalJava such as PDAs.
- **Midp:** to execute JADE-LEAP on handheld devices supporting MIDP1.0 (or later) only, such as the Java enabled cell phones.
- **Android:** to execute JADE-LEAP on devices supporting Android 2.1 (or later).
- **Dotnet:** to execute JADE-LEAP on PC and servers in the fixed network running Microsoft .NET Framework version 1.1 or later.

These versions provide the same APIs to developers thus offering a homogeneous layer over a diversity of devices and types of network, except the midp's version which have some unsupported features compared with the other versions of jade-LEAP. [1] Jade-LEAP provides two execution modes to adapt to the device's

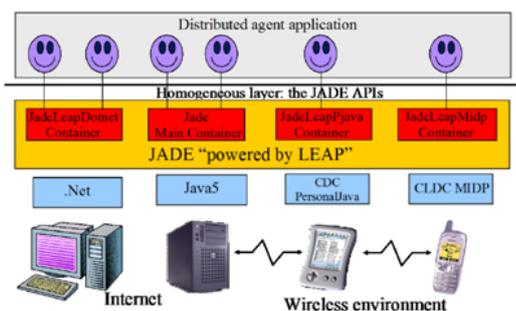


Fig. 1. The JADE-LEAP runtime environment [1]

constraints; the normal "Stand-alone" execution mode suggested in .net environment and supported in Pjava and Android. In this execution mode a complete

container is executed on the device/host where the JADE runtime is activated. The "Split" execution mode is mandatory in Midp and strongly suggested in Pjava. In this execution mode the container is split into a FrontEnd (actually running on the device/host where the JADE runtime is activated) and a Back-End (running on a remote server) linked together by means of a permanent connection.

This execution mode is very useful for our work because it use less memory and need less processing power on the mobile device, since the Front-End is definitely more lightweight than a complete container. Furthermore, it allows us to let the intensive processing tasks to the remote server and let the mobile device. It has the advantage of minimizing the bandwidth and optimizes wireless connection to the main container, since all communications with the Main container required to join the platform are performer by the Back End and therefore they are not carried out over the wireless link. Thus, the bootstrap phase is much faster.

In our work we attempt to implement the Jade-LEAP in mobile learning environment and benefit with the advantages of the split execution mode mentioned above, which addresses some limits of the mobile learning environment such as: low bandwidth.

There are several multi-agent platforms for mobile devices such as The MobiAgent [2], AgentLight [3], MicroFIPA-OS Agent Platform [4], and jade-LEAP [1]. We choose the jade-LEAP platform for many reasons such as: [5]

- Extension to JADE which written in java, and have features such as the possibility of executing multiple concurrent tasks (behaviours) in a single Java thread, matched well the constraints imposed by devices with limited resources. [1]
- Supports large variety of devices such as Java MIDP-capable phones, PDA devices,
- Smallest available platform in terms of footprint size,
- Proprietary device-initiated and socket based communication channel with main container,
- Developed within LEAP project,
- Open-source.

3 The proposed Architecture

We are proposing a multi-agent architecture for implementing mobile learning system which supports context-awareness and adaptive learning content using jade-LEAP platform. In our proposal we used agents to benefit of their advantages such as autonomous, reactive, proactive and social. The other side, we need to reduce wireless network problem by the use of mobile agents through the wireless connections to the mobile devices. The detailed description of these agents is articulated below:

1. **Interface Agent:** it is a stationary agent which have several tasks:

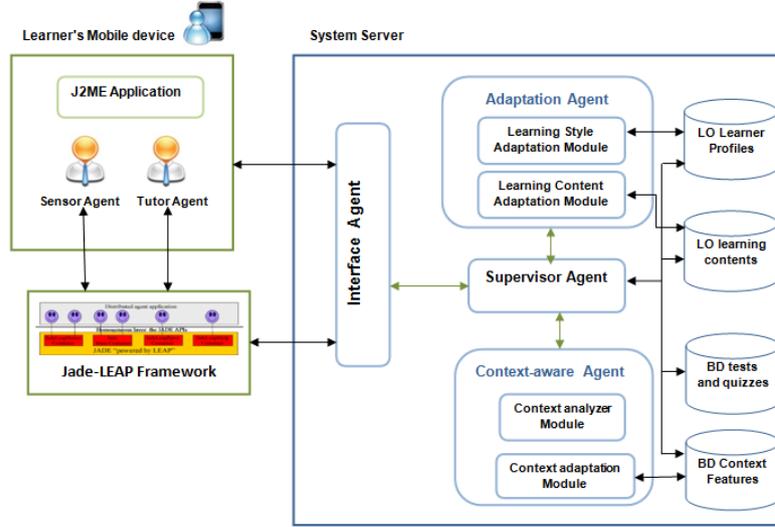


Fig. 2. Proposed Architecture

- It performs the authentication of the new learner, and checks user authorization by verifying the password.
 - It acts as a communication point between learners' devices and the system.
 - Send requests to Jade-LEAP platform to create and send mobile agents to the learner device.
 - It informs the Supervisor Agent to update or store information concerning the learner profile.
2. **Sensor Agent:** we called it sensor because it sense the learning environment and react accordingly to changes. This mobile agent has a role of monitoring and tracking the learner in his learning process and save his behavior and relevant data about it.
 - Send information about the device's features (memory size, processing power, available connectivity, communication costs, bandwidth, and battery level) to be saved in the context device features database.
 - Send observation about the learner; the duration of learning a course, concentration level (how often he interrupted by an external event such as a call or a message, navigation behavior, etc), how often he check the help page, duration between two connection to the system, and then send a report to the system when the student is disconnected.
 - Save the current learner location and send a request to the system contains the current learner location when the learner changes it to update context data base and to adapt course content to the user location.
 3. **Tutor Agent:** A mobile agent that manages the course delivery to the remote learner. The main tasks of the tutor agent are:

- Carry and manage the adaptive course material based on the learning style of the student.
 - It saves the pause point of the learner when he logout, and start from this point when learner login.
 - It insures the display of services and learning content according to the user preferences and device capabilities, in collaboration with the sensor agent.
 - Bring the test content to the learner and retrieve his answers to the adaptation module which calculate and send him his note.
4. **Context-aware Agent:** Context-aware Adaptation Agent consists of Context analyzer module and context adaptation module. Context analyzer module charges of analyze the information sent by the sensor agent and filter it to extract data related to the context, it receives periodically data from sensor agent, then it models this data and classify it according to its priority to be treated effectively by the context adaptation module, it send user profile information and context information to the supervisor agent who associates it to the context features and to the learner profile.
- Context adaptation module use the information retrieved by context analyzer module and apply it. For example, if the user has a limited bandwidth connection, then we must reduce multimedia content, and in the worse case we can replace it with text. On the basis of the present context, context adaptation agent predicts the future context and performs appropriate activity. For the previous example, it will transmit only data with small size. Finally, context adaptation module transmits context into adaptation module via the supervisor agent, which in turn save the learning context and incorporates it with adaptable learning content.
5. **Supervisor Agent:** It is a supervisor agent which has the role of monitoring the functionality of the system. It considered as a mediator between the system modules and it coordinate between them. It is the only agent who has the ability to change and update data in the learning object repositories (context features, learner profiles), with the help of interface agent which request it to create a new learner profile and informs it about data changed in the learner context.
6. **Adaptation Agent:** Since learners have different learning styles and devices have different characteristics, it has been necessary of personalized learning content. This task is realized by the adaptation agent, which consists of two modules; learning styles adaptation module and learning content adaptation module. These two modules coordinate between them, that is, learning style adaptation module matches the appropriate learning objects according to the learner style to be chosen later by the learning content adaptation module who manages the knowledge about courses and teaching strategies, and packaging the course material and tests according to the user profile and device profile.
7. **J2ME Application:** The Java 2 Micro Edition was, at the time, quickly becoming a de facto standard to develop mobile client-based applications [1]. This application is deployed and runs in learner's mobile device such as

java-enabled mobile phone, PDA, Smart phones, etc. after the learner download the jar file, he could install the application on his device. It displays a usable and appropriate interface which suit to the screen display capabilities. Via this interface user access to the learning material, and benefit to services offered by the system. So it act as a mediator between leaner and mobile learning system.

4 Conclusion and future work

In this paper we have described our proposed context-aware and adaptive learning system for Mobile Learning using mobile agent technology, which considered as promising solution in mobile learning systems, it may facilitate introducing automatic and dynamically adaptive learning for effective mobile learning systems. We are currently designing the system prototype which will be implemented using JADE-LEAP platform.

References

1. Bellifemine, F. & Caire, G. & Greenwood D.: Developing Multi-Agent Systems with JADE, John Wiley & Sons Ltd, England, 145–161 (2007)
2. Mahmoud, Q.H.: MobiAgent: An Agent-based Approach to Wireless Information Systems. In Proc. of the 3rd Int. Bi-Conference Workshop on Agent-Oriented Information Systems, Montreal, Canada. May 28 - June 1, (2001)
3. AgentLight - Platform for Lightweight Agents. <http://www.agentlight.org>
4. microFIPA-OS Agent Platform. <http://www.cs.helsinki.fi/group/crumpet/mfos>.
5. Mikko Laukkanen, Agents on Mobile Devices, Sonera Corporation, (2002)

New Web tool to create educational and adaptive courses in an E-Learning platform based fusion of Web resources

Mohammed Chaoui¹, Mohamed Tayeb Laskri²

^{1,2}Badji Mokhtar University – Annaba, Algeria

¹chaoui.mohamed@yahoo.fr, ²laskri@univ-annaba.org

Abstract. The evolution of new communication and information technologies led to a very high rate of innovation in online education. This opens doors for several major research projects at universities, institutes and research centers, all over the world. The content of training courses and quality are two key points presented in each E-learning platform system. Our working interest registers in these two points, or the need for powerful tool to create automatic creation of course content and the source is of course the Web, which has a huge space of information available requires good and over filtering. Our new tool increase the quality of being given the wealth of Web resources, direct adaptation based fusion of Web resources to learner profiles give high performance of our new tool, and enrichment of courses directly from the Web with backup of extracted resources, ensures the reusability of E-learning platform resources. This is also important that teachers receiving full benefits, time and effort will be reduced, and they just control over resources created in databases of system.

Keywords: Web Resources, E-learning Platform, Reusability, Adaptation, Fusion, Learner Profiles.

1 Introduction

The amount of learning material on Internet has grown rapidly in recent decades. Therefore, the information consumers are challenged to choose the right things. In systems of e-learning, most approaches have led to confusion for learners. Inevitably, adaptive learning has gained much attention in this area [1], [2].

We aim through our new tool reduce the huge space of the Web, containing billions of Web pages, to a personal space and direct adaptive to learners, to increase their satisfaction and provide good training scalable to any change or update [3], but with reliable and academic resources [4]. We must find good research and precise filtering to extract the most relevant information, because we are facing a very large mass of information available on the WEB, and editors spend an indefinite time to create courses and more specifically, having a content database that will be adapted to learner profiles. And before the learner needs to cultivate, to deepen more on such field or theme of learning [5], we are obliged to produce system that uses the Web as a documentary medium, and provides techniques to custom navigation for learners.

The rest of paper is organized as following: the second part related works and learner needs to construct adaptive and personalized learning domain. In the third part, we present our new tool and approach to create educational and adaptive courses in an E-learning platform based fusion from Web resources. And finally, we terminate with discussion and conclusion.

2 Related works

To create a practical learning environment for e-users, and to a broad audience (different objectives, knowledge levels, funds or learning abilities), it is necessarily that the designers of e-learning systems thinking on adaptive learning environments and flexible with this potential need, so they must improve the performance to the learners [6], [7]. Recent works dealing with the problem of adaptation have very powerful difficulty, because such learner profile can change a lot of time in learning [7], [3].

Some researchers are in making extensions for learning content standards to improve the quality of learning process. These researchers argue that current standards do not support an adaptive system so that they must be changed to have good adaptation to learner model. Much effort has been made in the field of adaptive systems to offer a user model. In learning systems, most of these works are about learning styles of learners to gain more [8], [9], [10], [11]. Learning style is an acceptable factor of adaptation, as it reflects the characteristics of learner preferences and needs.

There are two different general approaches of learning content adaptation [12], [13], [14]. The first approach seeks to adapt learning content with special needs, and the second focuses on the provision of the most appropriate learning content to needs of learners. The first is called adaptation of content level and the second is called the link-level adaptation. Neither approach has been preferred to another in the literature. Several research projects have been targeted to lead to propose new methodology for appropriate content. Some of these studies are underway on the extension of learning content standards to improve the quality of learning process. One group argues that current standards do not support an adaptive system so they must be modified in some respects [15], [16]. In response to fact that metadata standards of learning content are somehow inadequate for some applications, group of researchers tried to replace these standards with ontologies 'Semantic Web' [17], [18], [19], [20], [21]. Ontologies modeling course and give interaction between learners and systems, such as [17], [18], [20], [21], [22]. There are some studies that have used agents in adaptive learning [3], [4].

Current generation of E-learning platform is not yet ready for commercialization [23]. In other words, current studies are so focused on quality of adaptation [2] that result in special systems designated for learning purposes and does not work with other systems. In addition, no work to date has begun the next content before the adaptation, that is, to adapt content unorganized or non-existent [24]. The new in our research (addition to last work [25]), is the fusion of several fragments of Web resources, to increase the quality of training content via adaptive, reliable, very rich and dynamic learning domain in the sense enrichment and update.

3 Proposed approach

We must first searching in the Web by Google API; we can with this API finding Web resources to be filtering in another processing step. In second time, we consider implementation and the use of ontology in our system. A simple idea is the extraction of concepts, slots and instances. To do this step, we need an API called Jena. This API allows reading and writing of ontology (RDF or OWL type) in Java Platform. Our domain ontology is OWL-type, which has facilitated its implementation. We keep the hierarchy of ontology after extraction of concepts to give hierarchy that preparing our Learning Domain to saving in next time all extracted segments in correspondent parts of course in new segments database 'NSDB' after fusion of sub segments. 'NSDB' database use Excel model (as in Table 1.), to do this, we used a Java Excel API; this API allows reading and writing an Excel document in Java Platform. For each part of course, we define some semantic rules 'SR' to calculate degree of relevance 'DR' and distance based semantic rules 'DBSR' of each sub segment 'SS' of one Web resource part. The semantic rules of each course part defining in table are organized vertically and for each one, we define their correspondent sub segments, these later are extracted from Web resources. After this, we start fusion process (as in Fig. 1.) for each course part in table, for example for Part 1, we choose the content stored in sub segment 1 to N and save new segment or course part in correspondent column, in the same part of course Part 1, we save result in FSS1 'Fusion of Sub Segments'.

Table 1. Portion of Excel Model to save Filtering Results

			DR	DBSR			DR	DBSR
Part 1	SR1	SS1	0	0	...	SSN	0	0
FSS1	SR2	SS1	0	0	...	SSN	0	0
	0	0	0	0
	SRN	...	0	0	0	0
...
...
Part N	SR1	SS1	0	0	...	SSN	0	0
FSSN	SR2	SS1	0	0	...	SSN	0	0
	0	0	0	0
	SRN	...	0	0	0	0

We obtained a comprehensive approach that meets our needs:

- The hierarchy of course is mined from ontology of domain.
- Annotations and keywords of each concept in ontology are extracted and assured calculation of degree of relevance 'DR' (1) of each segment extracted from Web resources to finding the most relevant portions.
- We calculate Distance Based Semantic Rules 'DBSR' (2) for all relevant portions to extract the most relevant sub segments.
- Finally, we order the most relevant sub segments in Excel Model to create our New Segments Database 'NSDB'.

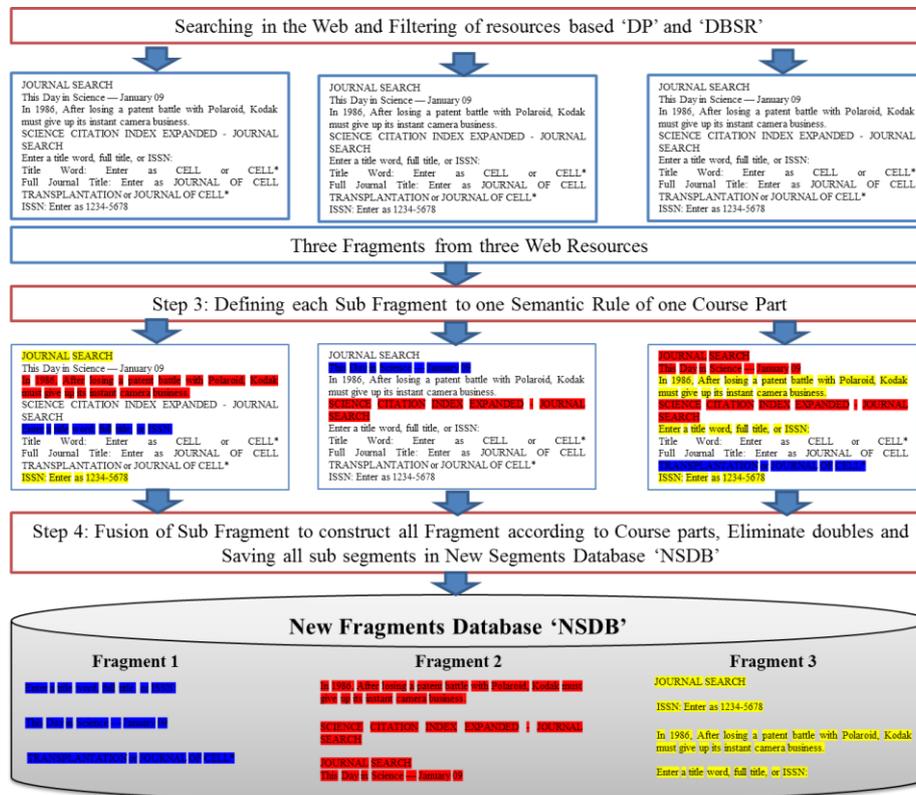


Fig. 1. Fusion process of three Web resources

3.1 Degree of Relevance 'DR'

It is a statistical result (1), based on the frequency of ontology concept (which presents a component of the course) in Web resource segment in the first part and the existence of keywords and their frequencies in the same segment in the second part. The frequency 'F' of a word in one segment is the number of times that word appears in this segment. Degree of relevance equal frequency of ontology concept in fragment, adding sum of keyword (k=0...n) frequencies of one ontology concept in one Web resource segment, multiple by correspondent keyword weight 'W'. The all is divided by total number of words in one Web resource segment.

3.2 Distance Based Semantic Rules 'DBSR'

It is a semantic result (2), based on the distance between terms in sub fragments, we must firstly extract terms from one sub fragment, and we calculate distance only between terms that defined in semantic rules. DBSR present a projection of semantic rule on sub fragment of Web resource to extract the most relevant sub segment appropriate to one sub part of course.

3.3 Fusion & Adaptation process

When processing of one document is finished, same steps were doing to other documents, but provided to relevant parts in order in Excel file for each component (column) of the course. If processing is completed, 'NSDB' database is full accomplished. After that, our 'NSDB' database present mean of adaptation based fusion process. We can adapt courses to level in learner profiles. Each level has number of course parts, and number of semantic rules. If level augments, course parts and semantic rules augment.

4 Discussion and Conclusion

Through this study developed, we succeed in building new Web tool with new adaptation approach in E-learning platform, based research and filtering of Web resources, after that, creating areas of learning with possibility of fusion of extracted resources, and the most important, adaptation of Web content to learner profiles. The world in the last years saw very rich side resources available on the Web; our method is to reduce this informational space in an adaptive educational space, personalized and mostly reusable for entire community of learners.

The study improves the quality of segments after fusion of several Web resources, and reusability of segments stored in our database gives performance in E-learning platform, and finally the augmentation of construction courses quality with enrichment by Web resources and the good methods of research and filtering implicated in our tool.

5 References

1. Chaoui, M., & Laskri, M-T.: Towards the Creation of Adaptive Content from Web Resources in an E-Learning Platform to Learners Profiles. *International Journal of World Academy of Science, Engineering and Technology WASET*. 77 (27), 157--162 (2011)
2. Caravantes, A., & Galn, R.: Generic Educational Knowledge Representation for Adaptive and Cognitive Systems. *Educational Technology & Society*. 14 (3), 252--266 (2011)
3. Chen, C.: Intelligent Web-based learning system with personalized learning path guidance. *Computers & Education*. 51(2), 787--814 (2008)
4. Canales, A., Pena, A., Peredo, R., Sossa, H., & Gutierrez, A.: Adaptive and intelligent Web based education system: Towards and integral architecture and framework. *Expert Systems with Applications*. 33(4), 1076--1089 (2007)
5. Papanikolaou, Mabbott, A., Bull, S., & Grigoriadou, M.: Designing learner-controlled educational interactions based on learning/cognitive style and learner behavior. *Interacting with Computers*. 18, 356--384 (2006)
6. Wang, M., Ran, W., Liao, J., and Yang, S.J.H.: A Performance-Oriented Approach to E-Learning in the Workplace. *Educational Technology & Society*. 13(4), 167--179 (2010)
7. Chen, C.: Personalized E-Learning System with Self-Regulated Learning Assisted Mechanisms for Promoting Learning Performance. *Expert Systems with Applications*. 36, 8816--8829 (2009)
8. Yang, Y., & Wu, C.: An attribute-based ant colony system for adaptive learning objects recommendation. *Expert Systems with Applications*. 36(2), 3034--3047 (2009)

9. Liegle, J., & Janicki, T.: The effect of learning styles on the navigation needs of Web-based learners. *Computers in Human Behavior*. 22(5), 885--898 (2006)
10. Magoulas, G., Papanikolaou, K., & Grigoriadou, M.: Adaptive Web-based learning: Accommodating individual differences through systems adaptation. *British Journal of Educational Technology*. 34(4), 511--527 (2003)
11. Stach, N., Cristea, A., & De Bra, P.: Authoring of learning styles in adaptive hypermedia. In *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*. pp. 114--123 (2004)
12. Olfman, L., & Mandviwalla, M.: Conceptual versus procedural software training for graphical user interfaces: A longitudinal field experiment. *MIS Quarterly*. 18(4), 405--426 (1994)
13. Papanikolaou, K., Grigoriadou, M., Magoulas, G., & Kornilakis, H.: Towards new forms of knowledge communication: The adaptive dimension of a Web based learning environment. *Computers & Education*. 39(4), 333--360 (2002)
14. Samuelis, L.: Notes on the components for intelligent tutoring systems. *Acta Polytechnica Hungarica*. 4(2), 77--85 (2007)
15. Lu, E., & Hsieh, C.: A relation metadata extension for SCORM content aggregation model. *Computer Standards & Interfaces*. 31(5), 1028--1035 (2008)
16. Rey-Lopez, M., Diaz-Redondo, R., Fernandez-Vilas, A., Pazos-Arias, J., Garcia-Duque, J., Gil-Solla, A., et al.: An extension to the ADL SCORM standard to support adaptivity: The T-learning case study. *Computer Standards and Interfaces*. 31(2), 309--318 (2009)
17. Chi, Y.: Ontology-based curriculum content sequencing system with semantic rules, Expert Systems with Applications. 36(4), 7838--7847 (2009)
18. Jovanovic, J., Gasevic, D., Knight, C., & Richards, G.: Ontologies for effective use of context in e-learning settings. *Educational Technology & Society*. 10(3). 47--59 (2007)
19. Lee, M., Tsai, K., & Wang, T.: A practical ontology query expansion algorithm for semantic-aware learning objects retrieval. *Computers & Education*. 50(4), 1240--1257 (2008)
20. Shih, W., Yang, C., & Tseng, S.: Ontology-based content organization and retrieval for SCORM-compliant teaching materials in data grids. *Future Generation Computer Systems*. 25(6), 687--694 (2009)
21. Zitko, B., Stankov, S., Rosic, M., & Grubisic, A.: Dynamic test generation over ontology-based knowledge representation in authoring shell. *Expert Systems with Application*. 36(4), 8185--8196 (2009)
22. Lee, Y.-H., Hsieh, Y.-C., & Hsu, C.-N.: Adding Innovation Diffusion Theory to the Technology Acceptance Model: Supporting Employees' Intentions to use E-Learning Systems. *Educational Technology & Society*. 14 (4), 124--137 (2011)
23. Whurle 2.0: An adaptive Web service e-learning environment for the Web http://www.cs.nott.ac.uk/mzm/Saudi_Conference_mmecaw13.pdf
24. Chaoui, M., & Laskri, M.T.: New method of finding information on the Web in unstructured information resources for educational use by learners, *International Journal of Research and Reviews in Computer Science*, Science Academy Publisher, United Kingdom. 2 (1), 33--39 (2011)
25. Chaoui, M., & Laskri, M.T.: Automatic construction of an on-line learning domain. In *IEEE Proceedings of International Conference on Machine and Web Intelligence*. pp. 439--443, November 2010.

Complete and incomplete approaches for graph mining^{*}

Amina Kemmar ¹, Yahia Lebbah ¹, Mohammed Ouali ¹, and Samir Loudni ²

University of Oran, Es-Senia, Lab. LITIO,
B.P. 1524 EL-M'Naouar, Oran, Algeria

² University of Caen - Campus II, Department of Computer Science, France
{kemmami,ylebbah}@yahoo.fr, mohammed.ouali@gmail.com, samir.loudni@unicaen.fr

Abstract. In this paper, we revisit approaches for graph mining where a set of simple encodings is proposed. Complete approaches are those using an encoding allowing to get all the frequent subgraphs. Whereas incomplete approaches do not guarantee to find all the frequent subgraphs. Our objective is also to highlight the critical points in the process of extracting the frequent subgraphs with complete and incomplete approaches. Current canonical encodings have a complexity which is of exponential nature, motivating this paper to propose a relaxation of canonicity of the encoding leading to complete and incomplete encodings with a linear complexity. These techniques are implemented within our graph miner GGM (Generic Graph Miner) and then evaluated on a set of graph databases, showing the behavior of both complete and incomplete approaches.

Keywords: Graph mining, frequent subgraph, pattern discovery, graph isomorphism.

1 Introduction

Graph-mining represents the set of techniques used to extract interesting and previously unknown information about the relational aspect from data sets that are represented with graphs.

We revisit some approaches for graph mining where a set of simple encodings is proposed. Complete approaches are those using an encoding enabling to get all the frequent subgraphs. Whereas incomplete approaches do not guarantee to find all the frequent subgraphs. Our objective is also to highlight the critical points in the process of extracting the frequent subgraphs. The introduced techniques are implemented within GGM (generic graph miner). We provide an experimentation with GGM showing the behavior of complete and incomplete approaches. It is not proven if the canonical encoding of graphs is in the class of NP-complete problems, nor in polynomial class. This is also verified in practice,

^{*} This work is supported by TASSILI research program 11MDU839 (France, Algeria).

since that all the current canonical encodings have complexities which are of exponential nature. This motivates deeply our work on proposing a relaxation of the canonicity of the encoding, leading us to what we qualified in this paper *complete* and *incomplete* encodings with low polynomial complexities.

The following section 2 introduces preliminaries on graph mining and the current approaches to solve frequent subgraph discovery problem. Section 3 explains our graph mining algorithm GGM. Experimental results of GGM are given in section 4. Section 5 concludes the paper and addresses some perspectives.

2 Frequent subgraph discovery problem

An undirected graph $G = (V, E)$ is made of the set of vertices V and the set of edges $E \subseteq V \times V$. Each edge (v_1, v_2) is an unordered pair of vertices. We will assume that the graph is labeled with vertex labels L_V and edge labels L_E ; the same label can be assigned to many vertices (or edges) in the same graph. The size of a graph $G = (V, E)$ is defined to be equal to $|E|$.

Definition 1 (Frequent Subgraph discovery). *Given a database \mathcal{G} which contains a collection of graphs. The frequency of a graph G in \mathcal{G} is defined by $freq(G, \mathcal{G}) = \#\{G' \in \mathcal{G} | G \subseteq G'\}$. The support of a graph is defined by*

$$support(G, \mathcal{G}) = freq(G, \mathcal{G})/|\mathcal{G}|.$$

The frequent subgraph discovery problem consists to find all connected undirected graphs F that are subgraphs of at least $minsup|\mathcal{G}|$ graphs of \mathcal{G} :

$$F = \{G \in \mathcal{G} | support(G, \mathcal{G}) \geq minsup\},$$

for some predefined minimum support threshold $minsup$ that is specified by the user.

Generally, we can distinguish between the methods of discovering frequent subgraphs according to the way the three following problems are handled:

Candidates generation problem This is the first step in the frequent subgraph discovery process which depends on the search strategy. It can be done with breadth first or depth first strategies. With breadth first strategy, all k -candidates (i.e., having k edges) are generated together, then $(k + 1)$ -candidates and so on; making the memory consumption huge [4][3]. But with a depth approach, the k -candidates are iteratively generated, one by one.

Subgraph encoding problem When some new candidate is produced, we should verify that it has been already generated. This can be resolved by testing if this new candidate is isomorphic to one of the already generated subgraphs. The canonical DFS code [6] is usually used to encode the generated frequent subgraphs. By this way, verifying that the new candidate is isomorphic to one of the already generated candidates is equivalent to testing if its encoding is equal to the encoding of some already generated candidate.

Frequency computation problem If some new candidate is declared to be not isomorphic to any of the already produced candidates, we should compute its frequency. It could be done by finding all the graphs of the database which contain this new candidate.

In the following section, we present a new algorithm GGM - Generic Graph Miner - for finding connected frequent subgraphs in a graphs database. We propose also some simple encodings to handle efficiently the frequency counting problem.

3 GGM, a generic graph miner

GGM finds frequent subgraphs, parameterized with some encoding strategies detailed in section 3.1. It is generic, because we aim to make the key steps of GGM easily parameterized.

Algorithm 1 $\text{GGM}(\mathcal{G}, f_{min})$

Require: \mathcal{G} represents the graph dataset and f_{min} the minimum frequency threshold.

Ensure: \mathcal{F} is the set of frequent subgraphs in \mathcal{G} .

- 1: $\mathcal{F} \leftarrow \emptyset$
 - 2: $\mathcal{E} \leftarrow$ all frequent edge labels in \mathcal{G}
 - 3: $\mathcal{N} \leftarrow$ all frequent node labels in \mathcal{G}
 - 4: $\mathcal{P} \leftarrow \text{Generate-Paths}(\mathcal{N}, \mathcal{E}, \mathcal{G}, f_{min})$
 - 5: $\mathcal{T} \leftarrow \text{Generate-Trees}(\mathcal{P}, \mathcal{E}, \mathcal{G}, f_{min})$
 - 6: $\mathcal{C} \leftarrow \text{Generate-Cyclic-Graphs}(\mathcal{T}, \mathcal{E}, \mathcal{G}, f_{min})$
 - 7: $\mathcal{F} \leftarrow \mathcal{P} \cup \mathcal{T} \cup \mathcal{C}$
 - 8: RETURN \mathcal{F}
-

The general structure of the algorithm is illustrated in algorithm 1. The algorithm initializes the frequent subgraphs with all frequent edges and nodes within the graph database \mathcal{G} . Then, the algorithm proceeds with three separated steps:

1. enumerating frequent paths from the frequent nodes,
2. generating the frequent trees from the frequent paths by keeping the same extremities of each initial path,
3. extending the frequent paths and trees by adding an edge between two existing nodes to obtain cyclic graphs.

This approach is inspired from GASTON [5] in which these three steps are repeated for each discovered subgraph. In other words, GASTON loops on the above three steps, whereas in our approach, they are executed one time only.

3.1 Graph encoding

The canonical labeling is used to check whether a particular candidate subgraph has already been generated or not. However, developing algorithms that can efficiently compute the canonical labeling is critical to ensure that the mining algorithm can scale to very large graph datasets. There exists different ways to assign a code to a given graph, but it must uniquely identify the graph such

that if two graphs are isomorphic, they will be assigned the same code. Such encoding is called a *canonical encoding*. It is not proven if the canonical encoding of graphs is in the class of NP-complete problems, nor in polynomial class. This is also verified in practice, since that all the current canonical encodings have complexities which are of exponential nature.

The idea of our encoding is to use a non-canonical encoding, resulting in two kinds of encodings : complete and incomplete.

Definition 2 (Complete and incomplete encodings). *Let f be an encoding function. For any two distinct non-isomorphic graphs G_1 and G_2 , f is complete if $f(G_1) \neq f(G_2)$. Otherwise, f is said to be incomplete.*

DFS based complete encoding This encoding is a relaxation of that defined in [6]. Such encoding is processed by taking only one walk through a depth first search (and not the minimum as in [6]). It is straightforward that this encoding is complete, and the same graph can be generated several times as illustrated in Figure 1 which shows that two isomorphic graphs can have different codes. For the graph (a) in Figure 1, there exists several DFS codes. Two of them, which are based on the DFS trees in Figure 1(b)-(c) are listed in Table 1.

edge	0	1	2	3	4	5
Fig 1.(b)	(1, 2, X, s, Y)	(2, 3, Y, t, X)	(3, 1, X, s, X)	(3, 4, X, q, Z)	(4, 2, Z, t, Y)	(2, 5, Y, r, Z)
Fig 1.(c)	(1, 2, Y, s, X)	(2, 3, X, s, X)	(3, 1, X, t, Y)	(3, 4, X, q, Z)	(4, 1, Z, t, Y)	(1, 5, Y, r, Z)

Table 1. DFS codes for Figure 1 (b)-(c)

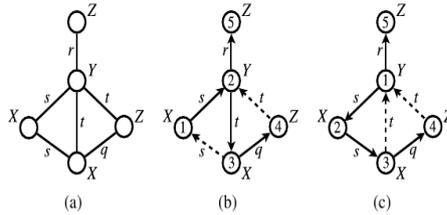


Fig. 1. Different DFS trees associated to the labeled graph (a)

Since this encoding visits only once the edges, it is then straightforward that the worst-case complexity is $O(m)$, where m is the number of edges.

Edge sequence based incomplete encoding Given a graph G and an edge $e_{ij} = (v_i, v_j) \in G$, where $deg(v_i) \leq deg(v_j)$, the edge e_{ij} is represented by the 5-tuple: $(deg(v_i), deg(v_j), l_v(v_i), l_e(v_i, v_j), l_j(v_j))$, where $deg(v)$ is the degree of v in the graph G , $l_v(v)$ and $l_e(e)$ are the labels of the vertex v and the edge e respectively.

Given a graph G , we denote $SEQ-DEG(G)$ the sequence of its edge codes :

$$SEQ-DEG(G) = code(e_1)code(e_2)...code(e_{|E|})$$

where $e_i <_l e_{i+1}$, the relation $<_l$ defines a lexicographic order between edges (e.g. $(2, 3, X, s, Y) < (2, 3, Z, s, Y)$). The code associated to the graph (a) of Figure 1 is: $(1, 4, Z, r, Y)(2, 3, X, s, X)(2, 3, Z, q, X)(2, 4, X, s, Y)(2, 4, Z, t, Y)(3, 4, X, t, Y)$.

Enumerating the edges is done with $O(m)$, where m is the number of edges, but sorting lexicographically the edges requires $O(m \log(m))$ which is the worst case complexity of sorting algorithms. Thus, in final, the complexity of this encoding is $O(m \log(m))$. This encoding is not complete because we can find examples of non-isomorphic graphs having the same code.

4 Experimental Results

We performed a set of experiments to evaluate the performance of our algorithm GGM on two kinds of graph databases. The first databases of large graphs contain some molecular structures of chemical compounds (PTE ¹). The second databases of small graphs are extracted from the database PTE (PTE1,PTE2,PTE3). The characteristics of these datasets are illustrated in Table 2. All experiments were done on 2.4Ghz Intel Core 2 Duo T8300 machines with 2GB main memory, running the Linux operating system.

Name	#graphs	#nodes	#edges	average #nodes	average #edges	#node labels	# edge labels
PTE1	1	8	7	8	7	2	1
PTE2	5	98	102	19	20	10	3
PTE3	20	519	530	25	26	10	4
PTE	340	9189	9317	27	27	66	4

Table 2. Characteristics of graph datasets used in the experiments.

Dataset	PTE1	PTE2	PTE3
MinFreq	1	3	8
Gaston	0,00	0,00	0,00
GGM SEQ-DEG	0,02	0,13	0,17
GGM DFS	0,22	3,42	3,94

Table 3. Runtimes in second of Gaston and GGM on simple graph datasets.

MinSup % = MinFreq	20% = 68			50% = 170			60% = 204		
Algorithm	GASTON	SEQ-DEG	DFS	GASTON	SEQ-DEG	DFS	GASTON	SEQ-DEG	DFS
#freq. paths	53	53	-	17	17	17	9	9	9
#freq. trees	124	97	-	15	14	66	2	2	6
#freq. cyclic graphs	13	12	-	2	2	10	0	0	0
# Total	190	162	-	34	33	93	11	11	15
runtime (s)	0,02	2,29	>2000	0,01	0,60	4,29	0,01	0,26	0,68

Table 4. Results of Gaston and GGM (with the DFS encoding and SEQ-DEG) on the graph dataset PTE. MinSup represents the minimum support threshold and MinFreq the minimum frequency.

We have done a comparison between our algorithm and the Gaston tool. Table 3 (resp. Table 4) shows the results of our algorithm with the first database (resp. second database). The minimum frequency was set to 1 for the first results. So, this table presents the runtimes to generate all the subgraphs of each database. While for the second case, we choose different MinFreq expressed also by MinSup (i.e. minimum support). For instance, for the PTE database which contains 340 graphs, the number of graphs that are subgraphs of at least

¹ The Predictive Toxicology dataset (PTE) can be downloaded from <http://web.comlab.ox.ac.uk/oucl/research/areas/machlearn/PTE>.

60% = $\frac{204}{340}$ graphs of PTE, is given by $\#Total = \#freq.paths + \#freq.trees + \#freq.cyclic\ graphs$. From this frequency, we see that there is no frequent cyclic graphs (i.e. $\#freq.\ cyclic\ graphs = 0$).

Concerning the results on both databases, the complete encoding is usually less performant than the incomplete one. This is explained by the fact that the complete encoding handles larger set of candidates than the incomplete one. For the incomplete encoding, the number of frequent subgraphs discovered by GGM is not too far from that of Gaston. We notice also that from the frequency of 170 graphs, the result is the same.

5 Conclusion

In this paper, we have presented algorithm GGM for the frequent subgraph discovery problem in a graph datasets. We pointed out the key points in the graph mining process. We combined several strategies inspired from existing algorithms to implement *GGM*. The two important points in the process of discovery are the generation of new candidates and the frequency counting. Our experimentations show the effectiveness of the incomplete approach compared to the complete one. It shows the importance of handling a reasonable amount of candidates. The main perspective is to improve our incomplete encoding. We have to experiment our incomplete approach on huge graph databases such as those coming from chemistry. Actually, we have implemented the whole mining algorithm in GGM and confront its performance to Gaston.

References

1. Bettina Berendt, Andreas Hotho, and Stum Gerd. Towards semantic web mining. In *Proceedings of the First International Semantic Web Conference on The Semantic Web*, ISWC '02, pages 264–278, London, UK, UK, 2002. Springer-Verlag.
2. Mukund Deshpande, Michihiro Kuramochi, Nikil Wale, and George Karypis. Frequent substructure-based approaches for classifying chemical compounds. *IEEE Trans. on Knowl. and Data Eng.*, 17:1036–1050, August 2005.
3. Akihiro Inokuchi, Takashi Washio, and Hiroshi Motoda. An apriori-based algorithm for mining frequent substructures from graph data. In *Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery*, PKDD '00, pages 13–23, London, UK, 2000. Springer-Verlag.
4. Michihiro Kuramochi and George Karypis. Frequent subgraph discovery. In *Proceedings of the 2001 IEEE International Conference on Data Mining*, ICDM '01, pages 313–320, Washington, DC, USA, 2001. IEEE Computer Society.
5. Siegfried Nijssen and Joost N. Kok. The gaston tool for frequent subgraph mining. *Electron. Notes Theor. Comput. Sci.*, 127:77–87, March 2005.
6. Xifeng Yan and Jiawei Han. gspan: Graph-based substructure pattern mining. *Order A Journal On The Theory Of Ordered Sets And Its Applications*, 02:721–724, 2002.
7. Ken'ichi Yoshida and Hiroshi Motoda. Clip: concept learning from inference patterns. *Artif. Intell.*, 75:63–92, May 1995.

Alignment between versions of the same ontology

Ahmed ZAHAF

Evolutionary Engineering and Distributed Information Systems (EEDIS)
University of Djillali Liabès , Sidi BEL-ABBES, Algeria
zhmida@gmail.com

Abstract. Ontological change constitutes a knowledge source on the similarity and difference between versions. The usual algorithms of ontology matching do not take this knowledge into account. An ontology is a logical theory consisting of a pair of signature and axioms. The persistent signature is the focus of alignment between versions. We reformulate the alignment problem between versions as a problem of choosing among the elements of the signature persistent those who can form an ontology signature isomorphism. To ensure alignment coherence, we introduce a constraint on alignment semantics which we call changed meaning conservation. This constraint allows extending the computed alignment with correspondences for the remaining elements in a coherent manner. Regardless of the change, our approach identifies the persistent signature and provides an initial alignment between the elements of the two persistent signatures. Then it calculates the difference between versions to form the ontological change. The constraint of changed meaning conservation helps us on one side to revise the initial alignment to form an ontology signature isomorphism and on the other side to adjust the eliminated correspondences in the revision step to form a coherent alignment. Finally, we discuss the prototype implementation of our approach.

Keywords: ontological change, versions difference, alignment between versions, ontology signature morphism, changed meaning conservation, alignment revision, alignment Coherence.

1. Introduction

The alignment between versions of ontology facilitates the evolution of ontology based semantic systems by allowing them to continue to interoperate with each other.

The ontological change constitutes a knowledge source on the similarity and difference between versions. Usual algorithms of ontology matching [1] do not take this knowledge into account.

An ontology is a logical theory consisting of a pair of signature and axioms [2]. The signature is the vocabulary used to name the ontological entities (concept, property and individual). The axioms express intentional interpretation of this vocabulary.

The ontological change affects both the signature and axioms. The signature change is the set of added or deleted signature elements. The remaining elements form the persistent signature. The axiomatic change is the set of added or deleted axioms. The remaining axioms form the persistent axioms.

The persistent signature is the focus of alignment between versions. There are two types of such signature: the elements whose intentional interpretation, as specified by the axioms is not affected by the axiomatic change and those whose intentional interpretation is affected. We express the alignment between the elements of the first type as ontology signature isomorphism between the versions. Then, the alignment problem consists in finding the maximal ontology signature isomorphism. We introduce a constraint on alignment semantics which we call changed meaning conservation. This constraint ensures that the intentional interpretation of an element of the signature in a version is maintained vis-à-vis the knowledge propagation through alignment semantics. We then say, an alignment between versions that conserves changed meaning is a coherent alignment. The problem of alignment between the elements of the second type is therefore to extend the maximal ontology signature isomorphism with correspondences while respecting the constraint of changed meaning preservation.

Our approach to solve the problem of alignment as posed above spread over four steps. Regardless of the change, our approach identifies the persistent signature and provides an initial alignment between the elements of the two persistent signatures. Then it calculates the difference between the versions to form the ontological change in the second step. In the third step, the initial proposed alignment must be revised if a violation of changed meaning conservation constraint is detected. To support revision, we introduce a relevance relation on the elements of the signature of the axiom. This relation compares the degrees of intentional persistence of these elements. The last step is user driven to extend the revised alignment.

Section 2 is the foundation of our work. We present the problem statement of the alignment in Section 3, and then we present our approach in Section 4. Section 5 is reserved to describe the platform of the prototype implementation of our approach. We compare our approach with related works in Section 6 and we summarize our work in Section 7.

2. Preliminaries and notations

The concept of Ontology can be seen as a logical theory [2]. So it is a pair $(\mathcal{S}, \mathcal{A})$, where \mathcal{S} is the signature - describing the vocabulary - and \mathcal{A} is a set of axioms - specifying the intended interpretation of the vocabulary in a domain of discourse. The signature is the set $\mathcal{S} = \mathcal{C} \sqcup \mathcal{P} \sqcup \mathcal{I}$. \mathcal{C} represents the set of vocabulary to designate concepts. \mathcal{P} is the set of vocabulary to designate properties and \mathcal{I} is the set of vocabulary to designate individuals. We distinguish between the origins axioms \mathcal{A} and their logical consequences \mathcal{A}^* (also called closure). Theory $(\mathcal{S}, \mathcal{A})$ is called the presentation of $(\mathcal{S}, \mathcal{A}^*)$. In this work, we limit ourselves only to $\mathcal{S} = \mathcal{C} \sqcup \mathcal{P}$ and we designate by ontological entity a concept or a property.

The ontological change affects both the signature and axioms. The signature change is the set of added or deleted signature elements. The remaining elements form the persistent signature. The axiomatic change is the set of added or deleted axioms. The remaining axioms form the persistent axioms.

We use the following notation: \mathcal{S}_i^p is persistent signature of a version i . \mathcal{S}^- is the removed signature. \mathcal{S}^+ is the added signature. Similarly, \mathcal{A}_i^p is the set of persistent axioms of a version i . \mathcal{A}^- is the set of deleted axioms. \mathcal{A}^+ is the set of added axioms.

We express the ontological change as the difference between versions.

Definition 1 (difference between versions). Given two versions of an ontology $\mathbf{v}_1 = (\mathcal{S}_1, \mathcal{A}_1)$ and $\mathbf{v}_2 = (\mathcal{S}_2, \mathcal{A}_2)$, the ontological change is the difference given by the set:

$$diff = \{(\mathcal{S}^-, \mathcal{A}^-), (\mathcal{S}^+, \mathcal{A}^+)\}.$$

Ontology alignment is the task to detect links between elements from two ontologies. These links are referred to as correspondences and express a semantic relation. According to Euzenat and Shvaiko [1] we define a correspondence as follows and introduce an alignment as set of correspondences.

Definition 2 (Correspondence and Alignment). given ontologies \mathbf{o}_1 and \mathbf{o}_2 , let Q be a function that defines sets of matchable elements $Q(\mathbf{o}_1)$ and $Q(\mathbf{o}_2)$. A correspondence between \mathbf{o}_1 and \mathbf{o}_2 is a 4-tuple (e, e', r, n) such that $e \in Q(\mathbf{o}_1)$, $e' \in Q(\mathbf{o}_2)$, r is a semantic relation, and $n \in [0; 1]$ is a confidence value. An alignment M between \mathbf{o}_1 and \mathbf{o}_2 is a set of correspondences between \mathbf{o}_1 and \mathbf{o}_2 . We restrict r to be one of the semantic relations from the set $\{\sqsubseteq, \supseteq, \equiv, \perp\}$.

In order to reason about alignment, two classes of approaches have been introduced. The first class is based on model theory. IDDL [7] and DDL [6] are two examples of approaches of this class. Based on an axiomatic approach, the second class called reductionist semantics [8] is to interpret correspondences of the alignment as axioms in some merged ontology. In this paper, we use an example of this semantic called natural semantic. It involves building a merged ontology through the union of the two ontologies to align and axioms obtained by translating relations of the alignment. We introduce this semantic through its merged ontology.

Definition 3 (Merged Ontology). given an alignment M between two ontologies \mathbf{o}_1 and \mathbf{o}_2 and $trans: M \rightarrow \mathcal{A}$, a function that transforms a correspondence to an axiom. The merged ontology is defined by

$$\mathbf{o}_1 \cup_M \mathbf{o}_2 = \mathbf{o}_1 \cup \mathbf{o}_2 \cup trans(M).$$

Since an ontology is a logical theory, an ontology signature morphism (an important notion used as a foundation to our work) is a theory morphism. Theory morphism is a signature morphism which preserve the axioms [18].

Definition 4 (ontology signature morphism). given two ontologies $\mathbf{o}_1 = (\mathcal{S}_1, \mathcal{A}_1)$ and $\mathbf{o}_2 = (\mathcal{S}_2, \mathcal{A}_2)$, an ontology signature morphism is a function $f: \mathcal{S}_1 \rightarrow \mathcal{S}_2$ such that $\mathcal{A}_2 \models f(\mathcal{A}_1)$, i.e., all models of \mathcal{A}_2 are models of the image of \mathcal{A}_1 by f . The image of an axiom is obtained by systematically replacing signature elements of this axiom by their correspondents, according to the signature morphism f . When f is bijective, we say f is an ontology signature isomorphism.

3. Problem Statement

The persistent signature is the focus of alignment between versions. The objective is to establish semantic relations between elements of the two persistent signatures. The persistent signature includes two types of elements: element whose meaning as specified by the axioms is not affected by the axiomatic change and those whose meaning is affected. Thus, alignment must establish relations between elements of the first type so that their meanings are completely preserved. It therefore defines an ontology signature isomorphism (See Definition 4) between elements of this type. We call such condition, meaning preservation and we define it formally as follows,

Definition 5 (meaning preservation). given two versions of an ontology $\mathbf{v}_1 = (\mathcal{S}_1, \mathcal{A}_1)$ and $\mathbf{v}_2 = (\mathcal{S}_2, \mathcal{A}_2)$, an alignment M between \mathbf{v}_1 and \mathbf{v}_2 preserve meaning if and only if it define an ontology signature isomorphism $M: \mathcal{S}_1 \rightarrow \mathcal{S}_2$ such that:

$$\mathcal{A}_2 \models M(\mathcal{A}_1) \text{ and } \mathcal{A}_1 \models M^{-1}(\mathcal{A}_2);$$

We can establish a variety of ontology signature isomorphism between versions depending on the number of correspondences established. The goal is to find the maximal one (M_{max}).

The alignment is known to propagate knowledge from one version to another. If this propagation is not controlled, it can affect the meaning of elements of the second type. The control of knowledge propagation amounts to establish correspondences between the signature elements such that the changed meaning in one version is preserved. We call such condition, changed meaning conservation and we define it formally as follows,

Definition 6 (changed meaning conservation). an alignment M between two versions $\mathbf{v}_1 = (\mathcal{S}_1, \mathcal{A}_1)$ and $\mathbf{v}_2 = (\mathcal{S}_2, \mathcal{A}_2)$ conserve the changed meaning if and only if M verifies the following two properties:

$$\forall \delta \in A^- , \mathbf{v}_1 \cup_M \mathbf{v}_2 \not\models M(\delta);$$

$$\forall \delta \in A^+ , \mathbf{v}_1 \cup_M \mathbf{v}_2 \not\models M^{-1}(\delta).$$

We then say, an alignment between versions that conserves the changed meaning is a coherent alignment and it is incoherent otherwise.

The first property ensures the coherence of alignment with regard to the propagation by its natural semantics (see Definition 3) of deleted axioms. The second property ensures the coherence of alignment with regard to the propagation of the added axioms.

The problem of alignment between the elements of the two persistent signatures which their intentional interpretation is altered is therefore to extend M_{max} with correspondences between them so that the alignment is coherent.

4. Alignment Method

The objective of our alignment method is to compute an alignment between the elements of persistent signatures of different versions of the same ontology. This alignment must satisfy meaning preservation condition for the signature elements whose meaning is not altered by the ontological change and the conservation of the changed meaning vis-à-vis the propagation of knowledge by the semantics of the alignment for the other elements. Our method satisfies meaning preservation condition by establishing equivalence relations between signature elements of the first type. We can establish a variety of alignments of this type depending on the number of correspondences established. Our method tends to generate the maximal one in three steps: version matching, version difference and

alignment revision. Version matching step identifies persistent signature based on the comparison between the terminology elements of both signatures. Assuming no changed meaning had occurred for persistent elements, our method generates an initial alignment by establishment of equivalence relations between the elements of the two persistent signatures. The persistent signature serves as a guide to determine the ontological change as the difference between versions in the second step. The revision step of the initial alignment eliminates just the correspondences that are responsible for the incoherence of this alignment. The alignment result is the desired maximal alignment.

The alignment method is completed by the extension step. In this step, the eliminated correspondences must be reviewed by the user to establish the appropriate relations while respecting changed meaning conservation condition. This step is semi automatic. We describe in what follows only the first three steps.

4.1 Version Matching

The objective of this step is the identification of the persistent signature in both versions and expresses the correspondences between the elements of the two persistent signatures with equivalence relations to form the initial alignment. The identification of the persistent signature is based on the existence of a terminological matcher. The terminological matcher can be based on the syntax of terms to be compared or a relationship of synonymy from a thesaurus in the field of ontology versions. Formally defined,

$$s_1 \in S_1^p \text{ and } s_2 \in S_2^p \text{ if and only if there exists a matcher } M \text{ such that } s_2 = M(s_1);$$

4.2 version difference

The objective of this step is to compute the ontological change in the form of semantic difference between versions. First, our method computes the set of persistent axioms then use this set to deduce the sets of deleted and added axioms. An axiom in a version is considered as persistent if the other version contains its image. The image of an axiom is obtained by systematically replacing signature elements of this axiom by their correspondents, according to a matcher M . Formally defined,

$$\delta_1 \in A_1^p \text{ and } \delta_2 \in A_2^p \text{ if and only if there exists a matcher } M \text{ such that } \delta_2 = M(\delta_1);$$

The following rules express the semantic difference:

$$A^- = A_1 - A_1^p; (\text{deleted axioms})$$

$$A^+ = A_2 - A_2^p; (\text{added axioms})$$

However, there may be exceptions to this, especially when considered as added or removed axioms can still be deduced. Therefore, we must refine the difference as follows,

$$\delta \in A^- \text{ and } v_2 \models M(\delta) \text{ then } A^- = A^- - \{\delta\}; (\text{refined deleted axioms})$$

$$\delta \in A^+ \text{ and } v_1 \models M^-(\delta) \text{ then } A^+ = A^+ - \{\delta\}; (\text{refined added axioms})$$

4.3 Alignment Revision

In general, initial alignment cannot be coherent. Because, some correspondences propagate axioms from one version to another that violate the constraint of changed meaning conservation. The objective of this step is to identify these correspondences and provide a means to choose among them which must be eliminated. The identification of these correspondences is simply obtained by identifying the signature of the axiom propagated. To choose among correspondences, we introduce an order relation which we call relevance relation on the signature elements of the propagated axiom. The relevance relation (noted $<_{rel}$) compares the degrees of intentional persistence of these elements. The intentional persistence of an element signature s denoted ($intPersistence(s)$) is expressed as the ratio of the

number of occurrences of this element in the persistent axioms set (denoted $nboccurrence(s, A_i^p)$ for a version i) on the total number of persistent axioms. Formally defined,

$s_1 <_{rel} s_2$ if and only if $intPersistence(s_1) < intPersistence(s_2)$ and

$$intPersistence(s) = nboccurrence(s, A_i^p) / |A_i^p|.$$

The signature element that has the less intentional persistent with respect to the relevance relation allows to choose the correspondence to eliminate from the initial alignment. When two of the signature elements have the same degree of persistence intentional, the choice is left to the user.

5. Implementation

The implementation of our method is at the time of this writing in an advanced stage. The first three phases of our approach is fully implemented. It remains to design and implement the extension phase. This phase requires a user-friendly interface to help the user to handle correspondences of the alignment with a flexible manner. Currently, the platform of our prototype is for OWL ontologies. We hope to extend it to other ontology languages in the near future. The platform is based on OWL API [14] and Align API [15]. The platform integrates pellet [16] as the main reasoning engine on OWL ontologies.

6. Related works

The problem of ontology matching has known the emergence of several approaches in recent years [1]. The main distinction between them is due to the nature of the knowledge encoded in the ontology, and how it is used in the identification of correspondences [9]. Terminological methods compare the lexicon used to designate ontological entities, while the semantic methods are based on model theory to determine the existence of a correspondence between two entities. Some approaches consider the internal structure of the ontology. Other approaches consider the external structure of the ontology. The ontology extension can also be used. The majority of the existing matching systems combine these techniques to cover different aspects of the ontology. With the exception of a few systems, such ASMOV [10] and S-Match [11], the alignment result is subject to logical contradictions. Other approaches [12] and [13] propose an additional component to revise the alignment. The revision is intended to ensure alignment coherence. Alignment coherence requires satisfiability preservation of ontological entities by alignment. None of these approaches considers the ontological change as a source of information about the similarity and difference between the versions. Meaning preservation and changed meaning conservation by alignment ensures alignment coherence between versions. In the case of alignment between versions, these two conditions are more general than satisfiability preservation of ontological entities.

The comparison between versions has been the subject of several approaches ([3], [4], [5], [17]). The purpose of the comparison is to calculate the difference between versions. Each approach is influenced by the underlying representation of the ontology. For example, PromptDiff [3] consider ontology as a graph. Ontoview [4], SemVersion [5] and [17] consider ontology as a set of RDF triples. None of these approaches match our vision of ontology as a logical theory.

7. Conclusion and future works

We presented the problem of alignment between versions as a problem of establishing an ontology signature isomorphism between the persistent signature elements of versions. We introduced changed meaning conservation constraint both in building this isomorphism and its extension in a coherent manner. We also proposed an approach based on the concepts of meaning preservation and changed meaning conservation to build a coherent alignment between versions of the same ontology. We discussed the platform of the prototype of our approach and we hope to automate the extension step of our method and to evaluate the prototype in the near future.

References

1. Euzenat, J., Shvaiko, P.: *Ontology Matching*. Springer Verlag, 2007.
2. Kalfoglou, Y., Schorlemmer, M.: Ontology mapping: the state of the art. *The Knowledge Engineering Review* 18(1), 1–31 (2003).
3. Noy, N.F., Musen, M.A.: Promptdiff: A fixed-point algorithm for comparing ontology versions. In *National Conference on Artificial Intelligence and Fourteenth Conference on Innovative Applications of Artificial Intelligence*, pages 744–750, Edmonton, Alberta, Canada, July 2002.
4. Michel, C., Klein, A.: *Change Management for Distributed Ontologies*. PhD thesis, Vrije Universiteit, Amsterdam, 2004.
5. Vålolkel, M.: D2.3.3.v2 SemVersion Versioning RDF and Ontologies. Technical report, University of Karlsruhe, January 2006.
6. Borgida, A., Serafini, L.: Distributed description logics: Assimilating information from peer sources. *Journal on Data Semantics*, 2003.
7. Zimmermann, A., Le Duc, C.: Reasoning with a network of aligned ontologies. *Proceeding of the 2nd International Conference on Web Reasoning and Rule systems (RR2008)*, 2008.
8. Meilicke, C., Stuckenschmidt, H.: An Efficient Method for Computing Alignment Diagnoses. *Proceedings of the Third International Conference on Web Reasoning and Rule Systems (RR-09)*, Chantilly, Virginia, USA, 2009.
9. Euzenat, J., Meilicke, C., Stuckenschmidt, H., Shvaiko, P., Trojahn, C.: Ontology Alignment Evaluation Initiative: six years of experience. *Journal on Data Semantics (JoDS)*, XV, pp. 158-192, 2011.
10. Jean-Mary, Y.R., Shironoshita, E.P., Kabuka, M.R.: Ontology matching with semantic verification. *Journal of Web Semantics* 7(3), 235–251 (2009).
11. Giunchiglia, F., Shvaiko, P., Yatskevich, M.: Semantic schema matching. In: *Proceedings of the 13rd International Conference on Cooperative Information Systems (CoopIS)*, Agia Napa (CY). LNCS, vol. 3761, pp. 347–365 (November 2005).
12. Meilicke, C., Stuckenschmidt, H.: Incoherence as a basis for measuring the quality of ontology mappings. In: *Proceedings of the ISWC 2008 Workshop on Ontology Matching*, Karlsruhe, DE pp. 1–12 (October 2008).
13. Qi, G., Haase, P. and Ji, Q. A Conflict-based Operator for Mapping Revision--Theory and Implementation, In *Proceedings of the 8th International Semantic Web Conference (ISWC'09)*, 2009.
14. Horridge, M., Bechhofer, S.: The OWL API: A Java API for Working with OWL 2 Ontologies. *OWLED 2009, 6th OWL Experienced and Directions Workshop*, Chantilly, Virginia, October 2009.
15. Euzenat, J.: An API for ontology alignment. In: McIlraith, S.A., Plexousakis, D., van Harmelen, F. (eds.) *ISWC 2004*. LNCS, vol. 3298, pp. 698–712. Springer, Heidelberg (2004).
16. Sirin, E., Parsia, B., Grau, B.C., Kalyanpur, A., Katz, Y.: Pellet: A practical OWL-DL reasoner. *Journal of Web Semantics*, 5(2), 2007.
17. Papavassiliou, V., Flouris, G., Fundulaki, I., Kotzinos, D., Christophides, V.: On Detecting High-Level Changes in RDF/S KBs. *Proc. ISWC*, 2009.
18. Meseguer, J.: General logics. In *Logic Colloquium '87*, pages 275–329. North Holland, 1989.

Discovery of similar blocks from very large-scale ontologies

Aicha Boubekour¹, Abdellah Chouarfia²

¹ Computer Science Department, University of Tiaret, Algeria,
boubakeur@univ-tiaret.dz

² Computer Science Department, University of UST-Oran, Algeria,
chouarfia@univ-usto.dz

Abstract. Large scale ontology matching is a labour-intensive and time-consuming process. To alleviate the problem, many automated solutions have been proposed. In order to avoid the drawbacks of the existing solutions, this paper proposes to cut down the number of concept pairs for which a similarity measure must be computed during ontology matching. More important, the main contribution is to deal subsets of concepts pair: on the one hand, if two concepts are highly similar, we leverage the concept hierarchy to skip subsequent matching between sub-concepts of one concept and super-concepts of the other concept. On the other hand, if two concepts are weakly similar, we leverage the locality phenomenon of matching to skip subsequent matching between one concept and the neighbours of the other concept.

Keywords. Large ontology, neighbour, similarity, link, search space.

1 Introduction and Motivation

In recent years, many large ontologies are created and maintained in the areas including machine translation, information retrieval, e-commerce, digital library, medicine, and life science. These ontologies have more than thousands to millions of concepts and properties, and some of them contain more than billions of instances such as Cyc¹, WordNet², SUMO³, Gene Ontology⁴ and UMLS⁵.

It has been argued that the difficulties of the operations of constructing, matching, reusing, maintaining, and reasoning on large ontologies would be extremely simplified by splitting large ontologies into smaller modules which cover specific subjects [6, 8]. Ontology modularization is the collective name of two approaches for frag-

¹ <http://www.cyc.com/>

² <http://wordnet.princeton.edu/>

³ <http://www.ontologyportal.org/>

⁴ <http://www.geneontology.org/>

⁵ <http://www.nlm.nih.gov/research/umls/>

menting ontologies into smaller, coherent components (modules), which are themselves ontologies [5]:

Ontology Partitioning Approaches. The ontology is partitioned into a number of modules $\{M_1, \dots, M_n\}$ such that the union of all the modules is semantically equivalent to the original ontology $\{M_1 \cup M_2 \cup \dots \cup M_n\} = O$; i.e. the M_i modules are not necessarily disjoint. Thus, a function partition (O) can be formally defined as follows:

$$\text{Partition (O)} \rightarrow M = \{\{M_1, M_2, \dots, M_n\} \mid \{M_1 \cup M_2 \cup \dots \cup M_n\} = O\} \quad (1)$$

The partitioning method reduces the search space and thus leads to better efficiency. The space complexity of the matching process is also reduced. Four partition based methods COMA++ [2], FalconAO [7,3], Taxomap [3], anchor flood [7] will be discussed below.

Module Extraction techniques. Concepts that form a coherent fragment of an ontology O are extracted to form a module M, such that it covers a given vocabulary (based on an initial module signature) $\text{Sig}(M)$, Such that $\text{Sig}(M) \subseteq \text{Sig}(O)$ [9]. In fact this task consists in reducing an ontology to the sub-part, the module, that covers a particular sub-vocabulary of O, as such $M \subseteq O$ [9]. Note that M is now ontology itself. A function extract (O, Sig (M)) can be defined as follows:

$$\text{Extract (O, Sig (M))} \rightarrow \{M \mid M \subseteq O\} \quad (2)$$

There are numerous techniques [4, 5] for module extraction, more than ontology partitioning approaches that have been developed for different purposes. The main usage of these approaches concerns partial reusing, when an application or a system needs only a part of ontology. Broadly speaking, modularization approaches aim to identify the minimal set of necessary concepts and definitions for different parts of the original ontology.

However, ontology partitioning approaches present several drawbacks. They cannot control the size of blocks, which may be too small or too large for matching [3, 4, 5, 6,9]. They can also cause another problem, namely, the partitioning can make the elements on the boundaries of blocks lose some semantic information, which in turn affects the quality of final matching results. This paper proposes a generic solution to assess preliminary 1:n mappings between any two concepts from two given ontologies based on their descriptive (semantic) information. On the one hand, if two concepts are highly similar, we leverage the concept hierarchy to skip subsequent matching between sub-concepts of one concept and super-concepts of the other concept. On the other hand, if two concepts are weakly similar, we leverage the locality phenomenon of matching to skip subsequent matching between one concept and the neighbors of the other concept.

The paper is structured as follows: Section 2 discusses large scale matching techniques. Section 3 presents definitions and basic concepts used throughout the paper. Section 4 describes our structure-based matching approach. Finally, Section 5 provides some concluding remarks.

2 Related work

According to Shvaiko P and Euzenat J [1] one of the toughest challenges for matching system is handling large scale schemas or ontology. Large-scale ontologies are a kind of ontologies created to describe complex real world domains. So, various large scale matching techniques are categorized in [2]:

- The early pruning strategy is to reduce the search space for matching; one matcher can prune entity pairs whose semantic correspondence value is very low, thus re-

ducing search space for the subsequent matcher (Quick Ontology Matching algorithm (QOM), Eric peukert et al. schema and ontology matching algorithm).

- The partition strategy is performed in such a way that each partition of first ontology is matched with only small subset of the partitions of the second ontology. This method reduces the search space and thus better efficiency (Coma++, Falcon-AO, Taxomap and Anchor flood).

- The parallel matching technique has two kinds' inter-matcher and intra-matcher parallelization. Inter-matcher parallelization deals with parallel execution of independently executable matchers while intra-matcher parallelization deals with internal decomposition of individual matchers or matcher parts into several match tasks that can be executed in parallel (Gross & al. ontology matching algorithm).

- Other matching tool: RiMOM and ASMOV ontology matching tools, Agreementmaker schema and ontology matching tool.

3 Preliminaries

The following definitions and basic concepts are used throughout the paper:

Definition 1 (schema graph): A schema graph (directed acyclic graph) of an ontology is given by (V, E, Lab_v) , where: $V = \{r, v_2, \dots, v_n\}$ is a finite set of nodes, each of them is uniquely identified by an object identifier (OID), where r is the schema graph root node. $E = \{(v_i, v_j) | v_i, v_j \in V\}$ is a finite set of edges. Lab_v is a finite set of node labels. These labels are strings for describing the properties of the element and attribute nodes, such as name and data type.

Definition 2 (neighbor): A neighbor concept c can be defined as follows: $Neighbors(c) = \{Sub(c) \cup Sup(c)\}$ avec $Sub(c) = \{c' | c' \text{ sub-concept } c\}$ and $Sup(c) = \{c' | c' \text{ sup-concept } c\}$

Definition 3 (strong-Links): Given two schema graph $G=(O1, E, Lab_v)$ and $G'=(O2, E', Lab_v)$ of ontologies $O1$ and $O2$, the similarity values between $a_i \in S$ and concepts b_1, b_2, \dots, b_n in ontology $O2$ are $Sim(a_i) = \{sim(a_i, b_j) \in G \times G' | j=1..n\}$, and the strong-Links of a node $a_i \in O1$ is given by $S_N(a_i) = \{b_j | sim(a_i, b_j) \geq \text{thresold}\}$, thresold is a high value in $[0..1]$.

Definition 4 (low-Links): Given two schema graph $G=(O1, E, Lab_v)$ and $G'=(O2, E', Lab_v)$ of ontologies, the similarity values between $a_i \in O1$ and concepts b_1, b_2, \dots, b_n in ontology $O2$ are $Sim(a_i) = \{sim(a_i, b_j) \in G \times G' | j=1..n\}$, and the low-Links of a node $a_i \in O1$ is given by $L_N(a_i) = \{b_j | sim(a_i, b_j) < \text{thresold}\}$, thresold is usually a small value in $[0..1]$.

Through these two last definitions, the matching process can reduce maximum times of similarity computation and thus reduce the time complexity significantly.

4 Structure connected Links

Our structure-based matching approach is realized by:

Step1. This phase is concerned with the representation of heterogeneous ontologies as sequence representations. First, each ontology is parsed and represented internally as a rooted ordered labeled graph, wherein each graph component (element and/or attribute) is represented as a node, while edges are used to represent relationships between components. Each node in the schema graph carries the associated element properties.

Step2. Compute preliminary similarities between any two entities for two given ontologies based on their descriptive information i.e. generate set of concepts pairs or links. It utilizes both structural and linguistic information for initial alignment and then applied subsequent similarity propagation strategy to produce more alignments if necessary. Its main function is to match the heterogeneous ontologies.

Step3. The first issue is to extract two kinds of virtual sub-graph for highly / weakly similar concepts (links) across ontologies. The second issue is to reduce the search space (i.e. Space and time complexity of the matching process), concerning wide-scale semantic heterogeneity in matching: this phase specifies all the similarity to be computed, and among these calculations, several links can be skipped in matching process.

Step4. During matching process, if credible alignments are computed, the corresponding high similarity links are isolated. Such links are to predict the ignorable similarity calculations in the remaining matching process. Also if the incredible matching results are found, the corresponding 'negative reduction' Links according to the locality of matching are also constructed, and such links to predict the ignorable similarity calculations are utilized. The similarity measure between entities from the two ontologies is computed by analyzing the literal and structural information in semantic subgraph extract in previous part.

Step 5. Repeat the two last steps for more alignment.

To this end, this process aims at providing high quality alignments between concept pairs with a time processing limit reasonable and it not needs to modularize or partition the large ontologies.

Therefore, considering structural information is a natural way for enhancing ontology mapping as illustrated by: Given two entities ai from O1 and bj from O2, we first apply and compute the similarities between entities based on the similarities of words e.g. the string-based and WordNet-based methods:

String-based method. the similarity measure between words w_i and w_j is defined as:

$$simStr(w_i, w_j) = comm(w_i, w_j) - diff(w_i, w_j) + winkler(w_i, w_j) \quad (3)$$

where $comm(w_i, w_j)$ stands for the commonality between w_i and w_j , $diff(w_i, w_j)$ for the difference between w_i and w_j , and $winkler(w_i, w_j)$ for the improvement of the result using the method introduced by Winkler in [7].

WordNet-based method. We use an electronic lexicon, WordNet, for calculating the similarity values between words. The similarity between two words w_i and w_j is measured by using the inverse of the sum length of the shortest paths [6]:

$$sim_{WN}(w_i, w_j) = 1 / (llength + rlength) \quad (4)$$

Where $llength$ is the shortest path from word node w_i to its common hypernym with word node w_j and $rlength$ denotes the shortest path from w_j to its common hypernym with w_i .

Instead of matching to all concepts by traversing taxonomies completely, the goal is to find Links between ontologies, at this step, it only considers on finding Links from the Cartesian product (X) of the two ontologies. These Links, are very important matching concepts, are used to reduce the time complexity in matching without exploring other commonalities between neighbors from the corresponding Links (initial Links generation). The algorithm proposed here generates a set of matching concepts as the initial links (see Algorithm1). The function *Sim* is an aggregated similarity function incorporating name and structural similarities (step 2):

Algorithm 1:

Input: Two ontologies O1,O2
Output: Neighbor-set
For each pair (ai,bj) ∈ O1xO2 **do**
 Compute sim(ai,aj)
 If (sim (ai,bj) > 0)
 then Links ← U {(ai,bj)}
 End
Return (Neighbor-set)
End

Algorithm 2:

Input: Ontology O1, Ontology O2, Links
Output: Set of Strong-Links
Links are generated by algorithm1
Getstrong-Links ← ai
SN ← ∅
For each bj ∈ O2 **do**
 Compute sim(ai,bj)
 If sim(ai,bj) > threshold
 then SN ← U {bj}
 End
End
Return Getstrong-Links

For finding efficient results, two possibly solutions are provided:

- If concept A matches concept B, it needs not to calculate the similarity between sub-concepts (/super-concepts) of A and super-concepts (/sub-concepts) of B, thus we can reduce the total times of similarity calculations.
- If A does not match B, it is very possible that their neighbors also do not match each other that imply we can ignore many similarity calculations.

Obviously, it needs to discover the high-Links and the low-Links dynamically in matching, and then uses these Links to optimize similarity calculations. For $S_N(a_i) = \{b_1, b_2, \dots, b_n\}$, the strong-Links set $RS_N(a_i)$ is calculated by:

$$RS_N(a_i) = \bigcap_{j=1}^k RS_N(a_i|b_j) = [\text{sub}(a_i) \times \text{sup}(\text{lub}(b_1, \dots, b_k))] \cup [\text{sup}(a_i) \times \text{sub}(\text{glb}(b_1, \dots, b_k))]$$

With $\text{lub}(b_1, \dots, b_k)$ and $\text{glb}(b_1, \dots, b_k)$ are the least upper bound and the greatest lower bound for (b_1, \dots, b_k) . Apparently, the total strong-Links sets during the matching process is $RS_N = \cup RS_N(a_i)_{i=1,n}$ (see Algorithm2 & 3):

Algorithm 3:

Input: Ontology O1, Ontology O2, Strong-Links
Output: total strong-Links sets
StrongLinks are generated by algorithm2
Matchedset ← strong-Links (ai)
Generates the neighbors of ai {sub(ai) | sup(aj)}
For each bj ∈ SN
 Generates the neighbors of bj {sub(bj) | sup(bj)}

$RS_N \leftarrow U \{ [\text{sub}(a_i) \times \text{sup}(\text{lub}(b_1, \dots, b_k))] \cup [\text{sup}(a_i) \times \text{sub}(\text{glb}(b_1, \dots, b_k))] \}$
End
Return Matchedset

5 Conclusion

First of all, the analysis in the existing matching systems depicts that there is always a tradeoff between effectiveness and efficiency. The main goal of this paper is to deal with wide-scale semantic heterogeneity in large scale ontology matching. For this purpose, we focus on reducing complexity, concerning wide-scale semantic heterogeneity in space matching. To accomplish this, we propose to skip subsequent matching between sub-concepts of one concept and super-concepts of the other concept (of shortcuts) of ontologies as input. However, it may be asked if this solution is quite adapted to find the most correct mappings between two concepts and the off-line discovering mappings from different ontologies. As a future work, we aim at answering these questions.

6 References

1. Shvaiko P, Euzenat J, "Ten challenges for ontology matching," *Confederated International Conference on the Move to Meaningful Internet Systems*, pp. 1164–1182, 2008.
2. Rahm E, "Towards Large-Scale Schema and Ontology Matching," *Schema matching and mapping*, Bellahsene Z, Bonifati A Rahm E, eds. New York: Springer Heidelberg, pp. 3–27, 2011.
3. F. Hamdi, B. Safar, C. Reynaud, and H. Zargayouna. Alignment-based partitioning of large-scale ontologies. In *Advances in Knowledge Discovery and Management*, volume 292, pages 251–269. Springer, 2010.
4. J. Seidenberg and A.L. Rector, "Web ontology segmentation: Analysis, classification and use", In *Modular Ontologies*, H. Stuckenschmidt, C. Parent and S. Spaccapietra, LNCS 5445, Springer, 2009, pp. 211–243.
5. Doran, Paul *Ontology modularization: principles and practice*. Doctoral thesis, University of Liverpool, octobre (2009).
6. P. Bouquet, L. Serafini, S. Zanobini: Semantic coordination: A new approach and an application. In *Proceedings of the 2nd Int. Semantic Web Conf. (ISWC'03)*. (2003) 130-145
7. G. Stoilos, G.B. Stamou, S.D. Kollias: A string metric for ontology alignment. In *Proceedings of the 4th Int. Semantic Web Conference (ISWC'05) (ISWC'05)*. (2005) 624-637
8. I. Palmisano, V. Tamma, T. Payne and P. Doran, "Task oriented evaluation of module extraction techniques", In *ISWC*, LNCS 5823, Springer, 2009, pp. 130–145.
9. M. d'Aquin, A. Schlicht, H. Stuckenschmidt and M. Sabou, "Criteria and evaluation for ontology modularization techniques", In *Modular Ontologies*, H. Stuckenschmidt, C. Parent and S. Spaccapietra, LNCS 5445, Springer, 2009, pp. 67–89.

From UML class diagrams to OWL ontologies: a Graph transformation based Approach

Aissam BELGHIAT, Mustapha BOURAHLA

Department of Computer Science, University of Md Boudiaf, Msila, 28000, Algeria

Belghiatissam@gmail.com

mbourahla@hotmail.com

Abstract. Models are placed by modeling paradigm at the center of development process. These models are represented by languages, like UML the language standardized by the OMG which became necessary for development. Moreover the ontology engineering paradigm places ontologies at the center of development process, in this paradigm we find OWL (the description language adopted by a great community of users) the principal language for knowledge representation. The bridging between UML and OWL appeared on several regards such as the classes and associations. In this paper, we propose an approach based graph transformation and registered in the MDA architecture for the automatic generation of OWL ontologies from UML class diagrams. The transformation is based on transformation rules; the level of abstraction in these rules is close to the application in order to have usable ontologies.

Keywords: UML, Ontology, OWL, ATOM3, MDA.

1 Introduction

UML is the unified object oriented modeling language which became an important standard. In the other side, the ontologies became the backbone of the semantic web which described formally using a standard language called OWL (Ontology Web Language). In this work we propose a set of rules for transforming classes diagrams into OWL ontologies in the order to profit from the power of ontologies so that the information described by those diagrams can be shared and linked with other information and we could start dealing with the overlaps, gaps, and integration barriers between modeling languages and get greater value out of the information capture. These rules will be implemented within ATOM3 to automate this transformation.

The rest of the paper is organized as follows: In Section 2, we present some related works. In Section 3, we present some basic notions about UML, OWL. In Section 4, we present concepts about model and graph transformation. In Section 5, we describe our approach. Finally concluding remarks drawn from the work and perspectives for further research are presented in Section 6.

2 Related Works

The idea of our work is not innovating, indeed several works exist in the literature tackle this subject. In [6] the OMG notices the interest of such subject and proposed in its turn the ODM which provides a profile for writing RDF and OWL within UML, it also includes partial mappings between UML and OWL. In [9], the author presented an implementation of the ODM using ATL language. In [5], the author used a style sheet “OWLfromUML.xsl” applied to an XMI file to generate an ontology OWL DL represented as RDF/XML format. In the other side Atom3 has been proven to be a very powerful tool allowing the meta-modeling and the transformations between formalisms, in [1] and other works we can found treatment of class diagrams, activity, and other UML diagrams. In these works the Meta modeling allows visual modeling and graph grammar allows the transformation.

Obviously, the heart of our work is articulated on transformation rules and their implementation. In preceding works, the transformation rules are more specific and reflect a general opinion of the author often related to a specific field which he works on (specific transformation). In this paper we propose that transformation rules are in a level of abstraction close to the application in order to obtain usable ontologies.

3 Bridging UML and OWL

UML (Unified Modeling Language) is a language to visualize, specify, build and document all the aspects and artifacts of a software system [7].

OWL (Ontology Web Language), was recommended by the W3C in 2004, and its version 2 in 2009, is designed for use by applications that need to process the content of information instead of just presenting information to humans [10].

UML and OWL have different goals and approaches; however they have some overlaps and similarities, especially for representation of structure (class diagrams). UML and OWL comprise some components which are similar in several regards, like: classes, associations, properties, packages, types, generalization and instances [6]. UML is a notation for modeling the artifacts of objects oriented software [2], whereas OWL is a notation for knowledge representation, but both are modeling languages.

4 Graph Transformation

Model transformation play an essential role in the MDA. MDA recommends the massive use of models in order to allow a flexible and iterative development.

A model transformation is a set of rules that allows passing from a meta-model to another, by defining for each one of elements of the source their equivalents among the elements of the target. These rules are carried out by a transformation engine; this last reads the source model which must be conform to the source meta-model, and applies the rules defined in the model transformation to lead to the target model which will be itself conform to the target meta-model (see fig. 1).

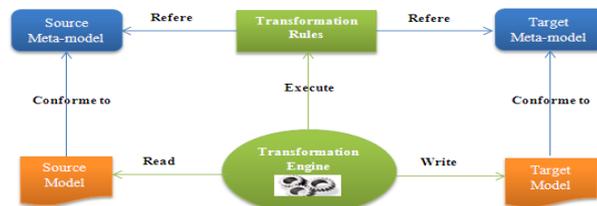


Fig. 1. Model transformation principle.

Graph transformation was largely used for the expression of model transformation [4]. Particularly transformations of visual models can be naturally formulated by graph transformation, since the graphs are well adapted to describe the fundamental structures of models. The set of graph transformation rules constitutes what is called the model of graph grammar, each rule of a graph grammar is composed of a left hand side (LHS) pattern and of a right-hand sided (RHS) pattern.

AToM3 [1] “A Tool for Multi-formalism and Meta-Modeling” is a visual tool for model transformation, written in Python [8] and is carried out on various platforms. It provides visual models those are conform to a specific formalism, and uses the graph grammar to go from a model to another.

5 Our approach

Our solution is implemented in AToM3. Our choice is quickly related to AToM3 because of the advantages which it presents like its simplicity, and its availability.

For the realization of this application we have to propose and to develop a meta-model of class diagram (fig.2), this meta-model allows us to edit visually and with simplicity class diagrams on AToM3 canvas. In addition to meta-model proposed we develop a graph grammar made up of several rules which allows transforming progressively all what is modeled on the canvas towards an OWL ontology stored in a disk file (fig.2). The graph grammar is based on transformation rules; those rules try to transform the class diagram in the implementation level, always in order to obtain at the end a usable description of ontology. For ontology, the choice among OWL profiles is made on OWL DL because it places certain constraints on the use of the structures of OWL [10][11].

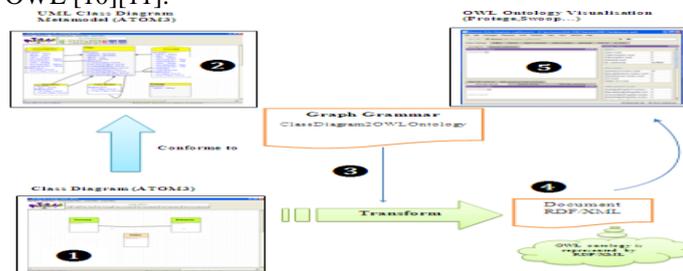


Fig. 2. Transformation sequence.

5.1 Transformation rules

Our approach is realized according to suggested transformation rules (Table 1). We propose a set of rules for all elements of a class diagram. The level of abstraction of rules is close to the application. For lack of space, we have presented one rule.

Table 1. UML to OWL Transformation rules.

Class
An UML class is transformed to an OWL class; the name of the class is preserved.
<code><owl:Class rdf:ID="ClassName"/></code>

5.2 Meta-model of UML Class diagram

To build UML class diagram models in AToM3, we have to define a meta-model for them. Our meta-model is composed of two classes and four associations developed by the meta-formalism (CD_classDiagramsV3), and the constraints are expressed in Python [8] code (fig.3):

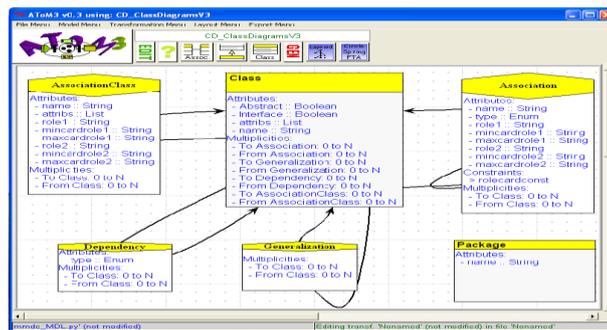


Fig. 3. Class diagram meta-model.

After we built our meta-model, it remains only its generation. The generated meta-model comprises the set of classes modeled in the form of buttons which are ready to be employed for a possible modeling of a class diagram.

5.3 The Proposed Graph grammar

To perform the transformation between class diagrams and OWL ontologies, we have proposed a graph grammar composed of an initial action, ten rules, and a final action. For lack of space, we have not presented all the rules.

Initial Action: Ontology header

Role: In the initial action of the graph grammar, we created a file with sequential access in order to store generated OWL code. Then we begin by writing the ontology header which is fixed for all our generated ontologies (fig. 4).

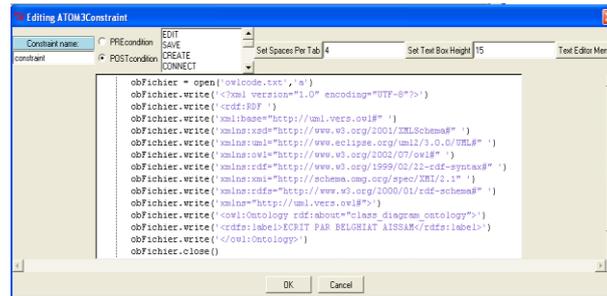


Fig. 4. Ontology header definition.

Rule 1: Class transformation

Name: class2class

Priority: 2

Role: This rule transforms an UML class towards an OWL class (cf. Table 3). In the condition of the rule we test if the class is already transformed, if not, in the action of the rule we reopen the OWL file to add the OWL code of this class.

Table 2. Class transformation.

Condition	
<pre>node = self.getMatched(graphID, self.LHS.nodeWithLabel(1)) return not hasattr(node, "rule executed")</pre>	
LHS	RHS
<p style="text-align: center;">::=</p>	
Action	
<pre>node = self.getMatched(graphID, self.LHS.nodeWithLabel(1)) classname = node.name.getValue() node.rule_executed = True abst = node.Abstract.getValue()[1] interf = node.Interface.getValue()[1] if abst == 1: self.getMatched(graphID, self.LHS.nodeWithLabel(1)).name.setValue('Abstract-'+classname) elif interf == 1: self.getMatched(graphID, self.LHS.nodeWithLabel(1)).name.setValue('Interface-'+classname) obFichier = open('owlcode.txt', 'a') node = self.getMatched(graphID, self.LHS.nodeWithLabel(1)) classname = node.name.getValue() obFichier.write('<owl:Class rdf:ID="'+classname+'"/>') obFichier.close()</pre>	

Final Action: Definition of the end of ontology

Role: In the final action of the graph grammar, we end our ontology, we will have to open our file and to add '</rdf:RDF>' (cf. fig. 5).

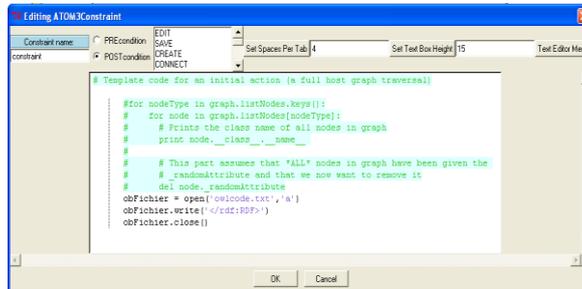


Fig. 5. End of ontology.

6 Conclusion

We saw in this paper how to implement an application which makes a transformation from a UML class diagram to an OWL ontology based on graph transformation and by using the tool AToM3.

For the realization of this application we developed a meta-model for UML class diagrams, and a graph grammar composed of several rules which enables us to transform all what is modeled in our AToM3 generated environment to an OWL ontology stored in a hard disk file.

In future work, we plan to extend the transformation of semantic rules models towards the language of rules SWRL (Semantic Web Rule Language).

7 References

1. AToM3. Home page: <http://atom3.cs.mcgill.ca>.2002.
2. Laurent AUDIBERT, "UML2", <http://www.lipn.univparis13.fr/audibert/pages/enseignement/cours.htm>, 2007.
3. Fowler, Martin, "UML Distilled - Third Edition - A Brief Guide to the Standard Object Modeling Language", 2003.
4. G. Karsai, A. Agrawal, "Graph Transformations in OMG's Model-Driven Architecture", Lecture Notes in Computer Science, Vol 3062, Springer, juillet 2004.
5. Sebastian Leinhos, <http://diplom.ooyoo.de>, 2006.
6. OMG, "Ontology Definition Metamodel", <http://www.omg.org/spec/ODM/1.0>, May 2009.
7. OMG, "OMG Unified Modeling Language, Infrastructure, v2.3", <http://www.omg.org/spec/UML/2.1.2/Infrastructure/PDF>, May 2010.
8. Python. Home page: <http://www.python.org>.
9. SIDo Group, "ATL Use Case - ODM Implementation (Bridging UML and OWL)", <http://www.eclipse.org/m2m/atl/usecases/ODMImplementation/>, 2007.
10. W3C OWL Working Group, "OWL Web Ontology Language-Overview", <http://www.w3.org/TR/2004/rec-owl-features-20040210/>. W3C Recommendation 10 February 2004.
11. W3C OWL Working Group, "OWL Web Ontology Language-Guide", <http://www.w3.org/TR/2004/REC-owl-guide-20040210/>. W3C Recommendation 10 February 2004.

Automatic composition of semantic Web services-based alignment of OWL-S

Adel BOUKHADRA¹, Karima BENATCHBA¹, Amar BALLA¹,

¹ National School of Computer Science, BP 68M, 16270, Oued-Smar, Algeria
{a_boukhadra, k_benachtba, a_ballla}@esi.dz

Abstract. Web services transform the Web into a platform for distributed components, heterogeneous, loosely coupled and integrated automatically. This technology is now widely used as a support for interoperability between distributed applications, which operate independently of the design features and technical specifications in order to achieve a feature previously established. The creation of a complex distributed application can be obtained by the composition of Web services. To build our platform-based semantic Web services, we strive to establish an architecture in which the semantic Web services interact with each other only, so they allow compositions of Web services to meet the up a different user requirements. The aim of our work is to achieve semantic interoperability in a heterogeneous, distributed architecture, based on the automatic dialing services Semantic Web. The special feature of this architecture is to place the alignment of OWL-S in the heart of this process, depending on the quality of services (QoS).

KEY WORD: automatic composition, Semantic Web services, semantic interoperability, ontology alignment OWL-S, QoS.

1 Introduction

Web services are as stateless software entities, betting provided by suppliers on the Internet and invoked by clients (users or other Web services). The architecture and Web services technology define a set of specifications for the description (WSDL), publishing (UDDI) and communication (SOAP) Web services between to promote interaction in an open, heterogeneous, and is versatile Web [2] [12].

The composition of Semantic Web services is the process of building new Web services to add value from two or more Web services are already present and published on the Web. The study of the composition of Semantic Web services is handled by several scientific communities [17] [18].

The ontology alignment is a very promising to enable semantic interoperability. It is the heart of this interoperability. The purpose of ontology alignment is to establish links, or semantic correspondences between entities belonging to different heterogeneous ontologies, to enable their semantic interoperability in a distributed and

heterogeneous. The ontology alignment based on the calculation of similarity measures.

The evaluation of the similarity between concepts in an ontology is a known problem in many areas. There are different measures of similarity, categorized according to the techniques used (terminology, Structural, linguistic, extensional semantics,...) [6] [7].

In this paper, we focus on the use of technical alignment of OWL-S, for the automatic composition of semantic Web services in distributed and heterogeneous. In cases where multiple Web services can meet the needs of users at the same time, we take into consideration the service that has a better quality of service parameter.

The rest of the paper is organized as follows. In Section 2, we present the problem with the objectives. Subsequently, in Section 3, described in detail our approach and presents our main contributions. In Section 4 illustrates the application of our approach through an implementation, and we end with a conclusion and give some perspectives in Section 5.

2 Problem and Objective

WSDL specifies the interface of a Web service: the operations performed, the types of messages sent and received, the formats of inputs and outputs. However, these specifications were insufficient for an automatic use of Web services (discovery, composition, ... etc.). The WSDL specification is too low-level operation of a Web service [10] [11].

Really, it is not always easy to find Web services that pair up with user requests. Therefore, the composition of Web services satisfying the query is a growing need today. To resolve this problem, the idea is to enrich the descriptions of Web services with other information understandable by machines. The description of the interface of a Web service can be completed with the OWL-S.

The current trend for the automatic composition of Web services is to enable semantic interoperability between Web services. There are other ways to automatically dial Web services, such as workflows, the calculation of situations, but planning is currently one of the most suitable and most studied by the community of this area [18] [19].

The objective of our work is to develop a system approach that aims to automatically dial the Semantic Web Services. For this, we propose to use the techniques of alignment of ontologies in the context of automatic scheduling to meet the problems described above. Indeed, support for the alignment during the composition, will minimize false responses, and significantly improves the overall quality of results.

3 Presentation of the proposed architecture

The architecture we propose is divided into the following modules (see Figure 1):

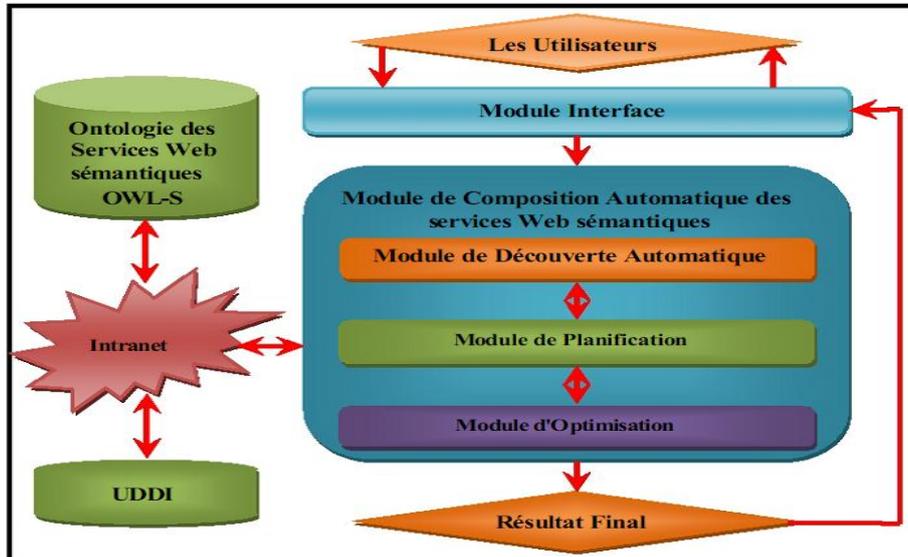


Fig. 1. Architecture for the Automatic Composition of Web Services.

3.1 Interface Module

The module interface is considered as the first window system to the world of users, it is the visible part of the architecture. The user has at its user-friendly interface, simple and allowing it to make its application according to their preferences in terms of quality of Web service (in our case, we are interested only in the following parameters: response time, the execution time and cost). Similarly, the user formulates his needs through the parameters Input, Output, Precondition, Result and TextDescription.

For this, we present in detail the architecture of automated dialing module, which is based on the following modules (see Figure 1):

3.2 Interface Automatic discovery of semantic Web services

We propose an algorithm for automatic discovery of Web services that is based entirely on two technical alignment of OWL-S [1].

3.3 Planning Module

Arrange Web services with inputs and outputs is very similar to a planning problem automatically, and this to find the correct order in the Web services automatic composition of Semantic Web services. Indeed, the role of planning is to find a sequence of actions, or plan to, from an initial state, reaching a goal state, expressed by the user.

In this context, we propose an algorithm for automatic scheduling of semantic Web services, which is based on two technical alignments of ontologies: the technical terminology and technique extrinsic [1].

```

Paramètre := { Input, Output, Precondition, Result et TextDescription } ;
Algorithme Alignement_Ontologie () ;
I := 1 ; Similarité := 0 ; N := Nombre des ontologies OWL-S découverts ;
M := Nombre des concepts de (Paramètre) d'une ontologie ;
L := Nombre des concepts de (Paramètre) d'une autre ontologie ;
Début
Tant Que (I ≤ N) faire
J := 1 ;
Tant Que (J ≤ M) faire
Terme1 := Paramètre_Requete [J] ; K := 1 ; MaxTerm := 0 ; SimTerm := 0 ;
SimilaritéTerm := 0 ; MaxExtr := 0 ; SimExtr := 0 ; SimilaritéExtr := 0 ;
Tant Que (K ≤ L) faire
Terme2 := Paramètre_Ontologie [K] ;
SimTerm := Similarité_Terminologique (Terme1, Terme2) ;
SimExtr := Similarité_Extrinsèque (Terme1, Terme2) ;
Si (SimTerm > MaxTerm) alors MaxTerm := SimTerm ; Fin Si
Si (SimExtr > MaxExtr) alors MaxExtr := SimExtr ; Fin Si
K := K + 1 ;
Fin Tant Que
Si (MaxTerm > SimilaritéTerm) alors SimilaritéTerm := MaxTerm ; Fin Si
Si (MaxExtr > SimilaritéExtr) alors SimilaritéExtr := MaxExtr ; Fin Si
J := J + 1 ;
Fin Tant Que
Similarité := SimilaritéTerm × PoidsTerm + SimilaritéExtr × PoidsExtr ;
Tab_Paramètre [I] := Similarité ;
I := I + 1 ;
Fin Tant Que
Retourner (Tab_Paramètre) ;
Fin

```

Algorithm. 2. Ontology Alignment.

This algorithm is based on the function Similarity_Terminology (word1; word2): It compute a similarity measure for the concepts of input parameters and output, between two ontologies OWL-S. The measure used is the metric Jaro-Winkler [15].

In this algorithm, the parameters (Precondition, Result and TextDescription) are often in the form of a long text including phrases, sentences or even paragraphs. For this reason, the similarity measures designed to deal with short strings, such as Jaccard, Hamming, and Jaro are no longer appropriate. Instead, we propose a measure that is based on a hybrid method to compare the length [8].

Similarly, this algorithm is based on the function Similarity_Extrinsic (word1; word2). It is used to compute a similarity measure for the concepts of the above parameters between OWL-S ontology concepts with two ontologies OWL-S to describe Web service semantics. There are several methods to calculate semantic relations in WordNet, among these methods; we chose to use the Jiang-Conrath measure [9].

The construction of a plan is based an Algorithm for automatic discovery of semantic Web services; we have shown previously to find the similarity between two different ontologies OWL-S. That is to say, the plan starts from the first Web service in which its parameters (Input, Precondition, Result and TextDescription) are semantically similar to the same parameters of the user query. Then Output parameter of the first Web service is semantically similar to the Input of another semantic Web service. Then, the parameters (Input, Precondition, Result and TextDescription) of the second Web service are semantically similar with the same parameters of another semantic Web service.

This process continues until the last Web service, such as its Output parameter is semantically similar to the output parameter of the user request. At the end of this

algorithm, we finally get a plan or several plans of automatic composition of semantic Web services.

3.4 Optimization Module

In fact, if the automatic discovery process is complete, a large number of semantic Web services can be found. As a result, the number of Web services increases and thus candidates for automatic composition process of Web services can take a long time. Under these conditions, the following criteria: Input, Output, Precondition, Result and TextDescription are not sufficient to allow a selection of Web services. We must use other criteria and parameters such as Quality of Web service (QoS) to distinguish between these Web services.

It is necessary to add an optimization phase whose goal is to provide the user with the best semantic web services according to certain criteria. This step takes into account user preferences in terms of quality Web service it wants, since each user has different needs and preferences, so it would be interesting to customize the dialer to provide improved results to users' needs. For example, a user prefers a web service with response time less than 12 ms execution time greater than 30 ms, and a cost of 13 cents per call, in this case, we select only the Web services that have these properties.

4 Implementation

We have developed a web application using JSF technology Eclipse Galileo, to show the proper functioning of our architecture in a distributed and heterogeneous. With regard to the different similarity measures that are implemented in our architecture, we used the Java API SIMPACK. We use two APIs to query WordNet 2.0, the API's functionality JWNL extracted for each lemma the list of its corresponding synsets in WordNet ontology and the API JWordNetSim to measure the similarity between synsets in WordNet. And to manipulate OWL-S, we used the OWL-S API provides a Java API to access programs, in addition, the Jena API is a Java framework for building Semantic Web applications.

5 Conclusion and Future Work

It is important that our proposed architecture for the composition can be made in a clear manner. From the perspective of the user, once the request is set, the platform began to compose semantic Web services automatically required existing and propose at the end of the compositions found.

We intend in the near future enrich our approach using optimization techniques such as heuristics and meta heuristics to select the best candidate Semantic Web Services in terms of quality of service after the stage of automatic discovery. This

work can be completed by the introduction of a formal semantics for verification of a composition.

References

1. Hakim Amrouche, Adel Boukhadra, Karima Benatchba, Walid Khald Hidouci, Amar Balla. Une approche sémantique pour la découverte automatique des services Web sémantiques. Workshop sur les services Web, WWS'10, CERIST, Algérie, (2010)
2. Fabien Baligand. Une Approche Déclarative pour la Gestion de la Qualité de Service dans les Compositions de Services. Thèse de Doctorat, Université de Nantes, France, (2008)
3. A. Budanitsky et G. Hirst. Evaluating wordnet-based measures of semantic distance. *Computational Linguistics*, 32(1), pages 13-47, (2006)
4. W. Cohen, P. Ravikumar, et S. Fienberg. A Comparison of String Distance Metrics for Name-Matching Tasks. In *Proceedings of KDD 2003 Workshop on Data Cleaning and Object Consolidation*, (2003)
5. Rémi Emonet. Semantic Description of Services and Service Factories for Ambient Intelligence. Thèse de Doctorat, Université de Grenoble INP, France, (2009)
6. J. Euzenat, et P. Valtchev. Similarity-based ontology alignment in OWL-Lite. In *Proceedings of 15th ECAI, Valencia, Espagne*, (2004)
7. Euzenat, J., Bach, T.L., Barrasa, J., Bouquet, P., Bo, J.D., Dieng-Kuntz, R., Ehrig, M., Hauswirth, M., Jarrar, M., Lara, R., Maynard, D., Napoli, A., Stamou, G., Stuckenschmidt, H., Shvaiko, P., Acker, S.V. et Zaihrayeu, I. Stat of the art on ontology alignment, *IST Knowledge Web NoE, Knowledge Web NoE*, (2004)
8. J. Euzenat, et P. Shvaiko. *Ontology Matching*. Edition Springer, Berlin Heidelberg, (2007)
9. J. Jiang, et D. Conrath. Semantic similarity based on corpus statistics and lexicalterminology. In *Proceedings of the International Conference on Computational Linguistics, RoclingX*, (1997)
10. Heather Kreger. *Web Service Conceptual Architecture*, IBM Software Group, (2001)
11. H. Lausen, D. Innsbruck. *Semantic Annotations for WSDL and XML Schema*, Édition Springer, (2007)
12. C. Lopez-Velasco. Sélection et composition de services Web pour la génération d'applications adaptées au contexte d'utilisation. Thèse de Doctorat, Université Joseph Fourier, France, (2008)
13. Julien Ponge. Model-based Analysis of Time-aware Web Services Interactions. Thèse de Doctorat, Université de Blaise Pascal - Clermont-Ferrand II, France, (2008)
14. N. Seco, T. Veale, et J. Hayes. An intrinsic information content metric for semantic similarity in Wordnet. In *Proceedings of ECAI'2004, the 16th European Conference on Artificial Intelligence, Valence, Espagne*, (2004)
15. W. E. Winkler, The state of record linkage and current research problems. *Statistics of Income Division, Internal Revenue Service Publication*, (2004)
16. Ustun Yildiz, Décentralisation des procédés métiers qualité de services et confidentialité. Thèse de Doctorat, Université de Henri Poincaré - Nancy 1, France, (2008)
17. Elie Abi Lahoud, Composition dynamique de services application à la conception et au développement de systèmes d'information dans un environnement distribué. Thèse de Doctorat, Université de Bourgogne, France, (2010)
18. H. Reza Motahari-Nezhad, R. Saint-Paul, F. Casati, B. Benatallah, *Event correlation for process discovery from Web service interaction logs*, Springer-Verlag New York, (2011)
19. C. Ba, M. Halfeld Ferrari, Martin A. Musicante, PEWS platform: a Web services composition environment. *WEWST '11: Proceedings of the 6th International Workshop on Enhanced Web Service Technologies*, (2011)

Author Index

A

Adel, Boukhadra	336
Adla, Abdelkader	294
Ahmed-Nacer, Mohamed	282
Aicha, Boubekeur	324
Aissam, Belghiat	330
Aliane, Hassina	193
Alimazhighi, Zaia	83, 193
Aloui, Ahmed	300
Amar Bensaber, Djamel	203
Amar, Balla	336
Amarouche, Idir Amine	4
Amel, Boussis	102
Atef, Chorfi	30
Atmani, Baghdad	22, 250

B

Barigou, Fatiha	250
Barigou, Naouel	250
Beghdadi, Hadj Ali	93
Bekakria, Hychem	12
Belalem, Ghalem	261
Belayachi, Naima	93
Ben Sidi Ahmed, Khalida	170
Benabderrahmane, Sidahmed	151
Benatallah, Boualem	1
Benslimane, Djamal	4
Benslimane, Sidi Mohammed	40, 276, 288
Bentaallah, Mohamed Amine	121
Berkani, Lamia	273
Bouamrane, Karim	93
Bouchiha, Djelloul	60
Boukhalfa, Kamel	83
Bouziane, Hafida	139

C	
Chaoui, Mohammed	139, 306
Cehida, Salim	232
Chemakhi, Imed	12
Chikh, Azeddine	273
Chkiwa, Mounira	70
Chouarfia, Abdallah	139, 324
Chouchane, Khamsa	300
D	
Derbal, Khalissa	83
Djeddai, Ala	50
H	
Hachemi, Asma	282
Hafida, Belbachir	222
Hayet, Djellali	112
Henni, Fouad	22
J	
Jedidi, Anis	70
K	
Karima, Benachtba	336
Kazar, Okba	300
Kemmar, Amina	312
Khadhir, Bekki	222
Khadir, Tarek	50
Khebizi, Ali	12
L	
Laskri, Mohamed Tayeb	112, 306
Lebbah, Yahia	312
Lezzar, Fouzi	30
Loudni, Samir	312
Lynda, Djakhdjakha	214
M	
Maatallah, Majda	129
Malki, Abdelhamid	40
Malki, Mimoun	60, 121, 203
Mazari, Ahmed Cherif	193
Mekami, Hayet	151

Merit, Khaled	240
Meroufel, Bakhta	161
Messabih, Belhadri	139
Mohamed Amine, Cheragui	160
Mohamed, Benmohammed	179
Mounir, Hemam	214
Mustapha, Bourahla	330
N	
Nabil, Sahli	179
Nachet, Bakhta	294
Nader, Fahima	102
Nouali, Omar	270
O	
Ouali, Mohammed	312
Ouamri, Abdelazziz	240
Ouksel, Aris M.	2
Ouzzani, Mourad	3
R	
Rahmouni, Mustapha Kamel	232
S	
Seridi-Bouchelaghem, Hassina	12, 50,129
Setti Ahmed, Soraya	288
Soltani, Mokhtar	276
T	
Toumouh, Adil	170
Z	
Zahaf, Ahmed	318
Zidani, Abdelmadjid	30
Zizette, Boufaïda	214

**Proceedings of the 4th International Conference on Web and Information Technologies
ICWIT 2012
Sidi Bel-Abbes, Algeria, April 29-30 2012**



ICWIT 2012
April 29-30 2012, Sidi Bel-Abbes
Algeria

