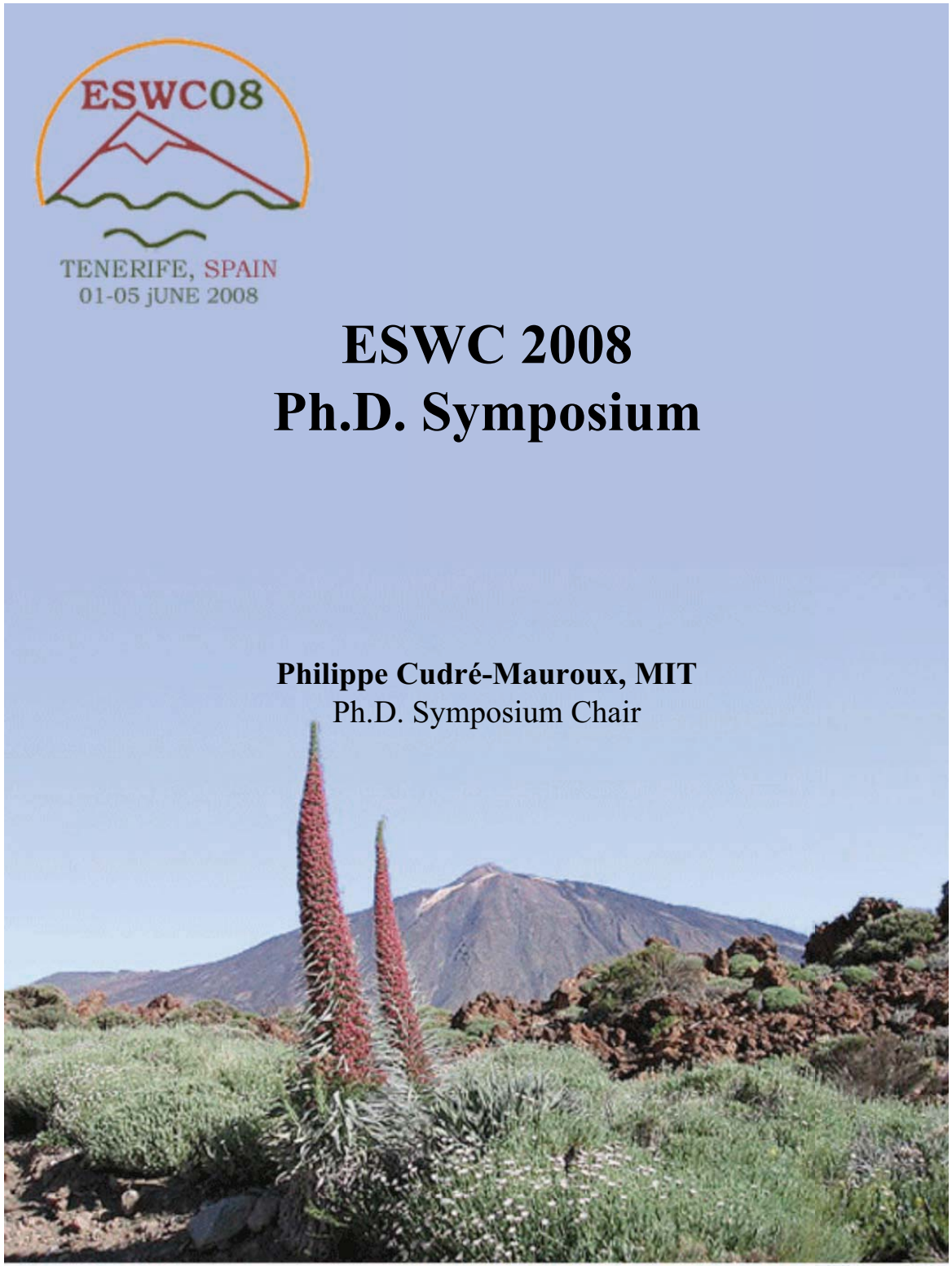

ESWC 2008
Ph.D. Symposium

Philippe Cudré-Mauroux, MIT
Ph.D. Symposium Chair

Preface

ESWC is the leading European conference on theory, practice and application of Semantic Web technologies, constantly attracting a large number of high quality submissions from both academia and industry. In addition to its plenary scientific sessions, the conference includes a Ph.D. Symposium to allow doctoral students to present their work and obtain guidance from leading scientists in the field.

The ESWC 2008 Ph.D. Symposium took place in Tenerife – Spain, on June 2, 2008. We received this year a total of 34 contributions, proving once again the sustained attention raised by Semantic Web technologies. The contributions originated from all over the world, with a majority coming quite naturally from Europe. Asia, the United States, South America, and Northern Africa were all represented as well.

Each of the submissions we received was thoroughly reviewed by at least two reviewers from the Program Committee (see below). The contributions were evaluated in terms of innovation, scientific soundness, clarity of presentation, and potential impact. We decided in the end to accept nine papers for regular presentation, and seven papers as posters. The present proceedings collet the five-page abstracts of all accepted contributions.

The success of the symposium is heavily dependent on the benevolent contribution of the members of the Program Committee, both for the reviewing phase and during the event itself. We would like to take the opportunity to heartily thank once again the Program Committee members and the additional reviewers for their outstanding work. At this point, we would also like to express our gratitude to the local organizing committee of ESWC 2008, whose help was instrumental in setting up the event. Finally, we would like to thank the Okkam project (http://fp7.okkam.org/) for sponsoring the event and STI International (http://www.sti2.org/) for its help in organizing the symposium.

June 2008
Philippe Cudré-Mauroux
ESWC 2008 Ph.D. Symposium Chair

The ESWC 2008 Ph.D. Symposium is sponsored by the Okkam project
(http://fp7.okkam.org/)

Program Committee

Daniel Abadi, Yale University (U.S.A.)
Karl Aberer, EPFL (Switzerland)
Harith Alani, University of Southampton (U.K.)
Abraham Bernstein, University of Zurich (Switzerland)
Paolo Bouquet, University of Trento (Italy)
Jeremy Carroll, HP Labs (U.K.)
Isabel Cruz, University of Illinois at Chicago (U.S.A.)
Jerome Euzenat, INRIA Rhone-Alpes (France)
Boi Faltings, EPFL (Switzerland)
Enrico Franconi, Free University of Bozen-Bolzano (Italy)
Carole Goble, University of Manchester (U.K.)
Peter Haase, University of Karlsruhe (Germany)
Siegfried Handschuh, NUI Galway (Ireland)
Frank van Harmelen, Vrije Universiteit Amsterdam (The Netherlands)
Matthias Klusch, DFKI (Germany)
Diana Maynard, University of Sheffield (U.K.)
Riichiro Mizoguchi, Osaka University (Japan)
Enrico Motta, The Open University (U.K.)
Natasha Noy, Stanford University (U.S.A.)
Daniel Olmedilla, L3S Research Center (Germany)
Jeff Z. Pan, University of Aberdeen (U.K.)
Axel Polleres, NUI Galway (Ireland)
Marta Sabou, Open University (U.K.)
Sebastian Schaffert, Salzburg Research (Austria)
Guus Schreiber, Free University Amsterdam (The Netherlands)
Luciano Serafini, Istituto Trentino di Cultura (Italy)
Amit Sheth, Wright State University (U.S.A.)
Pavel Shvaiko, University of Trento (Italy)
Elena Simperl, University of Innsbruck (Austria)
Heiner Stuckenschmidt, University of Mannheim (Germany)
Hideaki Takeda, University of Tokyo (Japan)
Tomas Vitvar, NUI Galway (Ireland)
Chris Welty, IBM Watson Research Center (U.S.A.)
Ilya Zaihrayeu, University of Trento (Italy)
Michal Zaremba, University of Innsbruck (Austria)

Additional Reviewers

Liliana Cabral, Gianluca Correndo, Armin Haller, Birgitta Koenig-Ries, Meena Nagarajan, Cartic Ramakrishnan, and Alexander Schutz.

Table of Contents

Towards Semantic Web-based Adaptive Hypermedia Model 1
Martin Balik

Techniques to Produce Good Web Service Compositions in The Semantic Grid 6
Eduardo Blanco

Metadata goes where Metadata is: contextual networks in the photographic domain 11
Rodrigo Fontenele Carvalho

Process Mediation for Semantic Web Services 16
Emilia Cimpian

Towards Ontology Mapping for Ordinary People 21
Colm Conroy

Methodology for Searching Entities on the Web 26
Renaud Delbru

Event and Sentiment Detection in Financial Markets 31
Uta Hellinger

Access rights and collaborative ontology integration for reuse across security domains 36
Martin Knechtel

Acquisition and Management of Semantic Web Service Descriptions 41
Maria Maleshkova

Ontological Description of Image Content Using Regions Relationships ... 46
Zurina Muda

Semantic Web-based Group Formation for E-learning 51
Asma Ounnas

Identifying Individuals using Identity Features and Social Information ... 56
Matthew Rowe

An Approach to Evaluate and Enhance the Retrieval of Web Services Based on Semantic Information 61
Stefan Schulte

OntoGame: Games with a Purpose for the Semantic Web 66
Katharina Siorpaes

Trend Mining with Semantic-based Learning 71
Olga Streibel

Semantics-aware Software Project Repositories 78
Jonas Tappolet

Towards Semantic Web-based Adaptive Hypermedia Model

Martin Balík, Ivan Jelínek

Department of Computer Science and Engineering, Faculty of Electrical Engineering
Czech Technical University
Karlovo náměstí 13, 121 35 Prague, Czech Republic
balikm1@fel.cvut.cz

Abstract. At present, most hypermedia systems display the same content for all users. To allow users working effectively, we need adaptive personalization. We are developing a general model for adaptive hypermedia that should provide a formal description and allow simple development of such systems. We use an innovative approach of utilizing Semantic Web technologies to enable data reuse and system interoperability. In this work we give the description of the research problem, introduce our General Ontological Model for Adaptive Web Environments (GOMAWE), demonstrate experiments used for evaluations and indicate the steps leading to completion of the work.

Keywords: adaptive hypermedia, personalization, user modeling, general model, formal description, e-learning, Semantic Web, ontologies, interoperability.

1 Introduction

Humans constantly monitor the world around them. Computers, however, are built to do what they are told to do, nothing more, and nothing less. Generally, computers do not exhibit such modeling behavior as humans do [1]. Adaptive systems should remedy this by storing a user model containing each user's preferences. User adaptive systems perform incremental behavior analysis to model the user and using the stored information the adaptation is performed.

Many researchers are working on the development of solutions for content and navigation adaptation of hypermedia spaces. Adaptive systems are mostly used in the fields of e-commerce or e-learning. Several models have been proposed for the description of adaptive hypermedia architecture. However, most of the current systems are developed using ad-hoc approaches and as a result are unable to cooperate and reuse their data.

Our work is motivated by the state of the art in the area of adaptive hypermedia systems. There are still shortages which need to be addressed. One obvious problem is that authoring of adaptive system is a difficult task. Therefore we need automatic authoring techniques. The next problem is related to the reuse and exchange of data.

Current systems do not store the data in a machine-understandable way, which makes data reuse and collaboration impossible. This is where the latest research started to utilize semantic web technologies in the context of adaptive personalization.

We would like to be able to create good adaptive systems with the ability to reuse data and cooperate with each other. This includes more individual problems. We can define these problems as research questions which will be answered through the results of our work.

- *How to create a user adaptive system?*
- *How to store the data within the system to enable their reuse and interchange?*
- *How to evaluate the user adaptive system?*

A strong formal theory of adaptive hypermedia is needed to correct the aforementioned shortcomings. Such theory is still missing, although the first attempts have been made [2], [3]. We want to use Semantic Web to define a formal description of adaptive systems and throughout our work we want to answer the research questions based on such theory.

2 Related Work

Personalization is the activity where a system is changed to conform better to the user. This is typically performed by explicit user actions (e.g. preference screen). Opposed to this, *adaptive personalization* means that the user's actions are observed by the system and used to base a user model. Data in the user model is used to personalize the presented information.

Several approaches can be used to personalize the information presented to the user. In adaptive hypermedia that our research is focusing on, the content of regular pages can be adapted (*content-level adaptation*) as well as the links from regular page, index pages, and maps (*link-level adaptation*). The overview of the adaptation techniques can be found in [4], [5].

The user model is a representation of information about an individual user that is essential for an adaptive system to provide the adaptation effect. We can classify user models along three layers: what is being modeled, how this information is represented and how different kinds of models are maintained [4]. Important user features to be modeled are the *user's knowledge, user's interests, user's goals, cognitive styles* and *context of the user's work*. For the development of personalized web there have been designed several models. In our work we analyzed the most known of them and we compared their features [5].

Semantic Web technologies have begun to creep into use in the hypermedia applications. These technologies are becoming popular in the field of adaptive hypermedia, because they provide means to overcome the interoperability problems connected with current adaptive systems. Improved solutions will be based on ontologies and also take into account existing standards.

3 Contributions

The aim of our work is to develop a formal model for adaptive web systems and a methodology for the development of adaptive systems based on this model. In our previous work we have analyzed the most desirable requirements of good adaptive systems which are missing in most of the current systems. Based on these requirements we have extended the modeling loop of adaptive systems [5]. We based our work on the GAM [1] foundations and we intend to make it a powerful tool for designing adaptive systems, by extending this model with new functions, new approaches and Semantic Web technologies.

We will introduce the General Ontological Model for Adaptive Web Environments (GOMAWE[1]). The GOMAWE is a model based on the semantic data representation. Such representation can be utilized by machines in process automation, data integration and reuse of knowledge across applications. The machine understandable contents are called metadata and their semantics can be specified using ontologies. Ontologies play an important role in the Semantic Web as they provide a common shared model to represent a domain and to reason about the objects in the domain.

The architecture of GOMAWE can be divided into several layers as depicted in Fig. 1. Important part of the model is the storage layer. The data structure is represented by an ontology, which enables the storage of metadata together with the data. Furthermore, the ontology is not a monolithic detailed ontology, but the data structure consists of multiple lighter weight ontologies, which can be used together. These ontologies should be independent, modular and layered. The user model is in fact an overlay model consisting of instances of objects described by the ontologies.

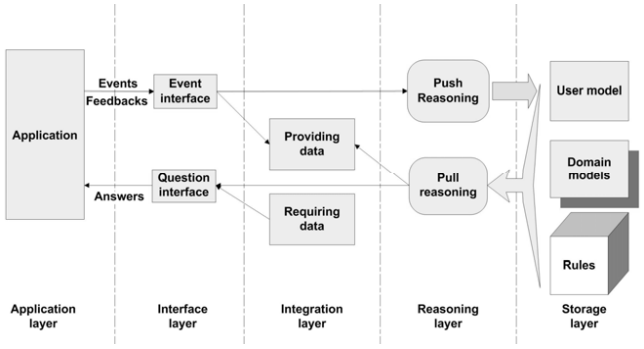

Fig. 1. Overview of the General Ontological Model for Adaptive Web Environments

[1] Gomawe is a New Caledonian god who created humans. Similarly, our model will be used to create adaptive web systems for humans.

To query for information from ontologies we have reasoning mechanisms derived from the description logic. However, we need rules to make further inferences and to express further relations not provided by the ontological reasoning. A rule is formed by an event, a condition and an action to be performed. Selected ontologies appear as layers on the dimensions of a multidimensional matrix. This was inspired by the semantic framework proposed by Italian researchers in [6]. Multidimensional matrix structure is used to select corresponding rules and thereby infer further information not included in the user model.

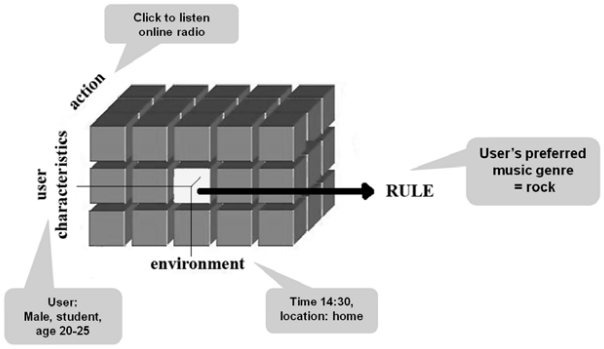

Fig. 2. Multidimensional matrix example (three dimensions)

The basic idea of the matrix is that the rules can be defined on the points of intersection between planes. The rule specifies classes of properties of classes that contribute to define the value of the inferred feature. For demonstration, let's assume situation on adaptive music web portal (Fig. 2). We will consider the following ontologies: *user characteristics*, *environment description* and *user's actions*. Then we can provide a rule that defines the property *user's preferred music genre* as an intersection of the planes.

4 Evaluation

To verify the theory, we need to perform experiments and implement a system based on this theory. This is an important step for making the model exercisable in practice. In the first phase of our research we have focused on the field of e-learning and developed ontologies of course materials and student progress in the course. We have also experimented with the tools for accessing the ontology data such as RDFReactor, RDF2GO and Jena semantic framework. Based on these experiments we are now developing a fully functional adaptive web portal. This web application is written in

java language, is based on MVC architectural design paradigm, utilizes the Spring framework and the above mentioned tools to access ontological data in the database. With a functional adaptive system based on our model, we will have the possibility to perform further experiments.

5 Work Plan

Our work should be heading towards the development of a general model for adaptive hypermedia systems. We have identified six important steps leading to completion of our work. 1. We have analyzed requirements of adaptive systems. 2. We have proposed a general model for adaptive systems in correspondence with the requirements. 3. We have defined fundamentals of the formal description, which will be extended in our future work. Now is the appropriate time for the next step. 4. We are implementing a prototype hypermedia system based on the theoretical model. It will serve for further experiments and model verification. 5. During the experimental implementation we will define a methodology for the development of adaptive systems. 6. The last step of our work will be the evaluation of the experimental system to verify our theoretical proposals.

Acknowledgments. This research has been supported by MSMT under research program No. 6840770014. This research has been supported by the grant of the Czech Grant Agency No. 201/06/0648.

The results of our research are part of the work of a special research group WEBING (http://webing.felk.cvut.cz).

References

1. de Vrieze, P.T., van Bommel, P., van der Weide, T.P.: GAM: A Generic Model for Adaptive Personalisation. Technical report: ICIS-R06022, Radboud University Nijmegen (2006)
2. de Vrieze, P.T.: Fundaments of Adaptive Personalization. PhD Thesis, Radboud University Nijmegen (2006)
3. Bureš M.: Formal Description of Adaptive Hypermedia System. PhD. Thesis. Faculty of Electrical Engineering, Czech Technical University in Prague. Czech Republic (2006)
4. Brusilovsky P., Millán E.: User Models for Adaptive Hypermedia and Adaptive Educational Systems. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.) The Adaptive Web, LNCS, vol. 4321, pp. 3--53. Springer, Heidelberg (2007)
5. Balík, M., Jelínek, I.: General Architecture of Adaptive and Adaptable Information Systems. ICEIS Doctoral Consortium, Funchal, Madeira – Portugal, pp. 29--34 (2007)
6. Carmagnola F., Cena F., Gena C., Torre I.: A Semantic Framework for Adaptive web-based Systems. In Proceedings of SWAP 2005, the 2nd Italian Semantic Web Workshop, Trento, Italy, CEUR Workshop Proceedings, ISSN 1613-0073 (2005)

Techniques to Produce Good Web Service Compositions in The Semantic Grid

Eduardo Blanco

Universidad Simón Bolívar,
Departamento de Computación y Tecnología de la Información,
Apartado 89000, Caracas 1080-A, Venezuela
{eduardo}@ldc.usb.ve

Advisors: Yudith Cardinale and María Esther Vidal

Abstract. The Grid has emerged as a new distributed computing infrastructure for advanced science and engineering. It aims at enabling resource sharing as means to facilitate a problem-solving approach for dynamic environments that is based on the integration of available resources. Wide agreement that a flexible Grid is not possible without the support of semantic technologies, has lead to the term "Semantic Grid". Transparent service composition has a great potential to facilitate the integration of Web Services deployed in the Grid. However, discovering, dynamically composing and optimizing large-scale Web Services (e.g., in the rank of 1,000 to 100,000 services) is a challenging problem. We approached the problem of Web Service compositions in the Grid as an optimization problem, in which for a given request a good plan has to be generated. In this work, we propose the definition of several optimization strategies to identify good orderings of Web Service compositions.

1 Research Problem

The Grid provides the protocols, services, and software development kits needed to enable flexible, controlled resource sharing on a large scale. Automatic discovery and coordination of services in the Grid is not only desirable, but necessary. Users need tools that assist them on taking full advantage of the possible large number of available services. Any form of automated composition needs an infrastructure where all resources are adequately described in a machine-processable form. Semantic Web technologies provide the means to incorporate such descriptions to the Grid infrastructure leading to the Semantic Grid. Thus, it is possible to apply Semantic Web technologies in grid-computing developments.

Numerous efforts have focused on the task of service discovery and Web Service composition (the WSC problem). However, few have addressed the problem of identifying efficient Web Service compositions in a large-scale scenario, as the one presented in the Grid vision [1]. In such scenarios, the association among services may be extremely complex, and the search space of possible Web Service compositions could be very large. In consequence, the task of

finding an optimal plan could be not feasible in real time; thus, our objective is to find sub-optimal plans in a reasonable time.

In GRID platforms requests are usually scheduled as batch processes, and real time constraints are relaxed. Thus, it is possible to do a greater exploration of the space of possible WEB SERVICE compositions, while the batch process is waiting to be evaluated.

On the other hand, some GRID resources can only be accessed at certain times. It may be required to incorporate into the space of alternatives, the possibility of executions where intermediate results are stored and retrieved according to time restrictions. In other words, all services used by a plan do not need to be available at the same time.

In this work, we propose the definition of optimization strategies to identify good orderings of WEB SERVICE compositions in large-scale WEB SERVICEenvironments, e.g., in the rank of 1,000 to 100,000 available services. These strategies will follow different meta-heuristics to explore the space of possibilities, while minimizing the estimated evaluation cost. We have already defined two strategies that prune the space of possibilities. Our initial results show that the compositions identified by our algorithms, are close to the optimal solution.

2 Related Work

The problem of identifying a good WSC in the GRID is related to problems that have been studied in query optimization and WSC. In this section we consider the main related approaches in each of these areas.

In the context of the Web, several strategies have been presented to identify good evaluation plans where sources have limited query processing capabilities; then, the optimizer task is to identify a good ordering of the sub-goals of a query where limited processing capabilities of each considered source, are satisfied [2, 3].

Typically, existing approaches achieve the challenge of identifying good plans by representing query capabilities as binding patterns and using these patterns and meta-heuristics to traverse the space of source plans. Each sub-goal in the query is rewritten in the plan by using sources that define the sub-goal; limited processing capabilities of the sources are satisfied in the plan with query bindings or attributes projected out by previous sources in the plan.

The WSC problem has been extensively treated in the literature, and diverse solutions that take advantage of AI techniques and Search Meta-Heuristics, have been proposed. First, in the context of AI, the WSC problem has been represented as a planning problem where actions to be taken by the planner are defined in terms of service preconditions and effects.

The description of a planning domain includes a set of planning operators and methods that establish the way a task can be decomposed into smaller subtasks. The description of a planning problem contains an initial state as in classical planning. Instead of a goal formula, there is a partially-ordered set of tasks to accomplish. Planning proceeds by decomposing tasks recursively into smaller and

smaller subtasks, until primitive tasks, which can be performed directly by using one planning operator, are reached. For non-primitive tasks, the planner chooses an applicable method, instantiates it to decompose the task into subtasks, and then chooses and instantiates other methods to decompose the subtasks even further. If the constraints on the subtasks or the interactions among them prevent the plan from being feasible, the planning system backtracks and tries other methods.

As any planning problem, the approach presented in [4], requires the formalization of the domain-dependent control knowledge in the planner. Thus, a domain expert is needed in order to achieve good performance in real-world domains.

An approach that uses *Answer Set Programming* is presented in [5]. It shows that service descriptions can be expressed in a rule based language that allows to search a repository efficiently and to build solutions that solve a goal with respect to soft and hard constrain. The author reports that the solution performs very well in this rather simple domain. It defines a simple cost function instead of a utility function. Its strength is that it provides means to gain all solutions for a given problem despite the cost of computation time and space. He states that dedicated software employing fast heuristic algorithms could rapidly find a good solution for user requests in much reasonable times.

In the context of Search Meta-Heuristics, the SAM (Service Aggregation Matchmaking) algorithm is defined [6]. It makes use of an OWL-S ontology, and explicitly returns a sequence of atomic processes that need to be executed in order to achieve the desired result.

SAM follows a greedy approach in which only one sub-plan is generated in each iteration. In each sub-plan, a sub-set of the output attributes is produced considering some of the bindings given in the query. The algorithm ends when all the output attributes are produced. In terms of time, SAM is able to scale up in environments with a moderate number of services (e.g., in the rank of 100 to 200 services). However, since SAM does not consider any cost metric or optimization criteria to compose the services, plans produced by SAM may be costly. To exacerbate this problem, SAM may add processes to the plan that are not needed to produce the output required by the user. Thus, the quality of generated plans may be far from optimal.

Semantic Grid [7] is an infrastructure where it is possible to apply Semantic Web technologies in grid-computing developments [8, 9]. In the context of Semantic Grid, the project ARGUGRID [10] provides a new model for programming the Grid; it uses argumentative agent technology and semantic descriptions to facilitate the dynamic composition of services.

Finally, in [11], the `WSC` problem is defined as a planning problem where a workflow is generated automatically. It takes as inputs the desired data products, and the planner uses heuristic control rules to identified a high-quality solution. Plan quality is measured with respect to a global utility function that does not represent the dynamic properties of Grid environments.

3 Contributions

We hypothesize that if there is a large number of Web Services published in different sites then, services' performance may vary. Thus, it is imperative that solutions to the WSC problem consider an estimate of the service evaluation cost. This cost is used to prune the space of possibly costly plans of services, and to identify the service composition with the lowest estimated cost.

We propose solutions to the WSC problem for large-scale platforms, e.g., Grids. These solutions will be adapted to consider the dynamic characteristic presented in such platforms. In particular, we address the problem of selecting and coordinating services that satisfy some of the constraints presented in the Grid, while the evaluation time is minimized.

Plans will be generated considering the time constraints of Grid resources. Thus, we will incorporate into the space of alternatives, the possibility of executions where intermediate results are stored and retrieved according to these constraints.

4 Evaluation

To validate the quality of our techniques, we plan to conduct experimental studies that compare our approach against some of the existing projects. We propose to use at least two available test sets - EEE05 and ICEBE05[1].

We will also analyze the performance of our solutions in real-world scenarios. For this, we will generate a large number of ontological Web Service descriptions for a given scientist community. WSBen[12] is a tool able of automatically generate sets of WSDL service descriptions. WSBen is inspired by extensive studies on real Web Services and therefore, it is designed to support various network topologies and distributions. We will extend this tool to generate ontological service descriptions.

5 Work Plan

We consider that the project can be achieved within the following stages:

1. We propose approaches that adapt some optimization techniques to the WSC problem. We have already defined two extensions to the algorithm SAM. Initial results show that the extensions are effective. In terms of plan quality and optimization time, our generated plans, outperform results produced by SAM [13].
2. We will extend WSBen to generate large sets of ontological Web Service descriptions.
3. We will propose cost models and test them with the generated sets; we will study the behavior of our approaches using different distributions and Web Service inter-relations.

[1] They are available at *http : //www.comp.hkbu.edu.hk/ ctr/wschallenge/news.html#dataset*

4. We will evaluate existing strategies proposed to discover and compose services in The Semantic Grid.
5. We will enrich the service composition algorithm and the cost models with the experiences on Semantic Grid. In this sense, it is mainly important to consider the dynamic characteristic of grid platforms.

References

1. Foster, I., Kesselman, C., Tuecke, S.: The Anatomy of the Grid: Enabling Scalable Virtual Organizations. International Journal of High Performance Computing Applications **15**(3) (2001)
2. Levy, A., Rajaraman, A., Ordille, J.: Querying heterogeneous information sources using source descriptions. In: Proceedings of the VLDB Conference. (1996)
3. Papakonstantinou, Y., Gupta, A., Haas, L.: Capabilities-based query rewriting in mediator systems. In: Conf. on Parallel and Distributed Inform. Systems. (1996)
4. Kuter, U., Sirin, E., Nau, D.S., Parsia, B., Hendler, J.A.: Information gathering during planning for web service composition. In: Intl. Semantic Web Conf. (2004) 335–349
5. Rainer, A.: Web service composition using answer set programming. In: PuK 2005. (2006)
6. Brogi, A., Corfini, S., Popescu, R.: Composition-oriented service discovery. In: Proc. of Software Composition'05, LNCS. Volume 3628. (2005) 15–30
7. De Roure, D., Jennings, N., Shadbolt, N.: The semantic grid: A future e-science infrastructure. In Berman, F., Fox, G., Hey, A.J.G., eds.: Grid Computing - Making the Global Infrastructure a Reality. John Wiley and Sons Ltd. (2003) 437–470
8. Goble, C., Roure, D.D.: The grid: an application of the semantic web. SIGMOD Rec. **31**(4) (2002) 65–70
9. Pahl, C.: An ontology-based framework for semantic grid service composition. Grid Services Engineering and Management **3270** (2004) 63–77
10. Programme, E.C.S.F.: ARGUGRID (2006) http://www.argugrid.org/.
11. Deelman, E., Blythe, J., Gil, Y., Kesselman, C., Mehta, G., Vahi, K., Lazzarini, A., Arbree, A., Cavanaugh, R., Koranda, S.: Mapping abstract complex workflows onto grid environments. Journal of Grid Computing, Vol. 1, No. 1, (2003)
12. Oh, S.C., Kil, H., Lee, D., Kumara, S.R.T.: Wsben: A web services discovery and composition benchmark. In: ICWS '06: Proceedings of the IEEE International Conference on Web Services (ICWS'06), Washington, DC, USA, IEEE Computer Society (2006) 239–248
13. Blanco, E., Cardinale, Y., Vidal, M.E., Graterol, J.: Techniques to produce optimal Web Service Compositions. In: 2008 IEEE Congress on Services 2008 - Part I (SERVICES-1 2008), Honolulu, Hawaii, USA, IEEE Computer Society (2008) To appear.

Metadata goes where Metadata is: contextual networks in the photographic domain

Rodrigo Fontenele Carvalho

The University of Sheffield, Web Intelligence Technologies lab,
211 Portobello, Sheffield, S1 4DP
R.Carvalho@dcs.shef.ac.uk

Abstract. Realizing the potential of digital media within the Semantic Web, relies on the existence of detailed, semantically unambiguous metadata. However, in the domain of personal photography within family centered communities, the generation of such metadata remains an unsolved issue. This paper introduces a solution to this problem based on exploiting the existence of strong semantic connections among photographs and entities specific to the user's context. In particular, three scenarios where this approach would be useful are discussed such as: (1) the reuse of existing metadata in order to generate more semantic metadata for unannotated photographs, (2) taking advantage of the connections among photographs to enhance search over a user's collection and (3) the support for communities by an external framework allows the use of image-based communications (e.g. sharing) to aid in metadata generation by reinforcing the connections across multiple graphs.

Keywords: photograph, metadata, semantic web, social web, context, graph

1 Introduction

In order to realize the potential of digital media within the Semantic Web, detailed and semantically unambiguous metadata (i.e. data about data) is needed, as it will provide the handles that will enable automated agents to interact with the external world to perform the more tedious tasks of finding, sharing and combining information on the web [1].

In the domain of personal photography, and more specifically within family centered communities, there is a growing need to exploit the existence of semantic metadata for facilitating the management of photographic collections. However the issue of generating such metadata for photographs remains unsolved.

The development of an approach that tackles this problem in the domain of family photography presents one main challenge: the use of appropriate data capture and reuse strategies. This challenge has not only been previously referred to as a crucial step to the wider adoption of the Semantic Web [2] but is also key in personal photography as previous studies by Frohlich et al [3], and Miller and Edward [4], have found that users in this domain are unlikely to make extensive contributions to generating semantic metadata about their ever expanding photographic collections.

Furthermore, offering great affordances for image-based communications (e.g. sharing) is an important requirement for this domain. The importance of this has only become clearer over the past years with the observation of an ever increasing number

of users that take advantage of web applications that enable the sharing of photographs and associated metadata via direct or indirect social interactions[1].

To address the issue of generating semantic metadata for personal photographs in the family domain an approach is proposed that assumes that (1) although users are unlikely to generate semantic metadata for their entire photographic collection, it is likely that some such metadata be generated for key images within a collection [5], and (2) photographs and entities specific to the user's context share strong semantic connections. By taking an approach based on directed graphs to representing and integrating user generated content with their social and physical environment, a **Context Network** can be created for modeling among other things: the content of photographs as well as user input (direct[2] or indirect[3]); user acquaintances (e.g. friends and genealogy tree), familiar locations and social interactions; automatic input from capture devices such as GPS coordinates, Exif and accelerometer data[4].

By taking advantage of the data structure formalization offered by ontologies to represent such context networks, the existing link structure between photographs and other context specific entities can be explored. More specifically, this can be targeted at facilitating the inference of detailed and semantically unambiguous metadata from the reuse of metadata that has been previously associated with other nodes in the graph (i.e. photographs). Another advantage of representing this data as a directed graph include improved searching capabilities over the photographic collection.

While the requirement for image-based communications is not directly addressed by the approach proposed, it can be implemented externally by a supporting framework. The social interactions between users that are triggered by the sharing of photographs can then be explored for the generation of semantic metadata.

The following sections introduce related work, details how the approach addresses the issue of semantic metadata generation as well as how it enhances search and capitalizes on the sharing of photographs for the generation of semantic metadata.

3 Related Work

Recent work in this area demonstrates different approaches for tackling the problem of metadata generation. Systems such as Photocopain [6] and MediAssist [7] rely on cheap contextual information as well as active human input for automatically or semi-automatically inferring textual metadata about the image contents. The use of collective input about a particular resource in order to make similar metadata suggestions for previously unseen resources has also been explored in AktiveMedia [8]. The application of human computation for annotation tasks has also been well researched, and can be exemplified by the game-like platform seen in [9].

The appearance of the Web 2.0 (i.e. support for social interaction and user generated content) has also motivated the creation of many photography applications

1 Flickr (http://www.flickr.com), Zooomr (http://www.zooomr.com), Riya (http://www.riya.com), MyHeritage (http://www.myheritage.com)

2 Direct user input includes any user interaction that has the explicit intent of generating photographic metadata such as adding captions, placing photographs on a map, etc.

3 Indirect user input includes any user interaction whose intent is not directed towards the generation of metadata, such as sharing photographs with other family members.

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

(e.g. Flickr, Facebook[5]) that make use of active human input about photographs in the form of tags, comments or other annotation types (e.g. geographic, face, object, audio, etc). Others use input from the capture device or from processing the medium to infer data such as the geographic location (e.g. ZoneTag[6]) or face detection and recognition (e.g. Riya, MyHeritage.com[7]) for creating folksonomies of image metadata.

These approaches however miss out on the integration of the context of images as well as their owner's social environment[8] in generating semantic metadata for representing photographs. This network of data gives essential clues as to what kind of metadata is likely to be relevant and not using it in an integrated manner can be very limiting. For instance, in processing semantic captures such as Ben Nevis, contextual information about the photograph and the user can help determine whether the entity refers to the Scottish mountain or a person (e.g. does GPS data show the photograph was taken in Scotland? Does the user know anyone called Ben Nevis?).

4 Proposed Approach

The approach proposed for facilitating the generation of photographic metadata is founded on the notion that data mining requires taking into account not only the features of the entities of interest, but also the underlying structure of the data [11,10]. Its relation with the Semantic Web comes from a clear parallel that exists between such data structures and those imposed by the latter. Coupling graph based algorithms with the highly interlinked nature of RDF data is therefore not only natural but a sensible approach to developing novel techniques within this field of research.

Metadata Generation. For addressing the problem of metadata generation, the approach relies on a single core principle: ***"a little metadata goes a long way."*** This principle is based on the assumption that users often provide metadata for key photographs in their collections and addresses the challenge of making appropriate reuse of existing semantic metadata for the generation of more semantic metadata for unannotated photographs.

Fig. 1: Only the first image in this series contains textual metadata – *Everest trek*.

Fig. 1[9] exemplifies a situation commonly found in a person's collection of photographs: higher proportion of unannotated images than annotated ones. By taking advantage of strong semantic connections between the seed image and other images

5 http://www.facebook.com/
6 http://zonetag.research.yahoo.com/
7 http://www.myheritage.com/
8 This includes with whom a photograph has been shared, the user's closer acquaintances, familiar locations or even the similarity with other photographs in the same set.
9 Images contributed by Flickr user *Fernweh*.

in the same collection, as well as information in the context network, and applying algorithms that explore not only the links between these images but also the similarities between them [10] (e.g. visual similarity, closeness in time and GPS data), the information that is known about the seed image could be propagated to other neighbouring nodes (i.e. the other two images).

Searching. The existence of a graph with strong semantic connections between entities can also be explored for use in the retrieval of photographs. Queries based on keywords or visual similarities could be used for retrieving base images from a collection, and the context network commonly shared between these could be used to discover other resources that are also likely to be relevant but that may not share the exact same metadata as the base images. This would enable keyword queries to return images that have no textual metadata or content based[10] queries to return images that are conceptually similar but not necessarily visually similar. For instance, a seed image of a predominantly orange sea at sunset would not prevent images from the sea at sunrise to be retrieved in the same result set. Similar approaches have shown promising results in the domain of video retrieval [12].

Sharing. The requirement that any technique or tool developed for the family photography domain offers image-based communications can be turned into a feature of this approach. Intuitively, in a directed graph, data points (i.e. photographs) would be represented as both annotated and unannotated nodes and the edges between them are weighted so that the closer the nodes are according to a distance metric, the larger the weight of their connection. The social aspect of image sharing would play a vital role in controlling the weights between nodes from different graphs by modelling a direct relationship between people's social clique and each edge weight. That is, the stronger the social clique between people, the more likely it is that their images will influence the semantic metadata in each other's collections. The benefits of this would then be twofold: (1) for predicting who is likely to be interested in a specific image; (2) for using the information in other people's RDF graph to increment a user's graph.

Making predictions about who may be interested in any one specific photograph within a user's collection would involve an analysis on the link strength between nodes in the context network of photographs in two collections. The images with stronger links given by their contextual similarity as well as the user's social clique can be proposed for sharing.

An extension to the metadata generation approach, where nodes that have stronger links should be attributed similar semantic metadata, can be proposed for allowing the graph from two separate users to influence the semantic metadata of each other's photographs. By using the social clique between people as a control variable on the weight of node edges, an approach similar to semi-supervised learning, such as label propagation [10], can be used in the context of the social interaction between people. For instance, if a user tends to share a lot of photographs with his family, the weights between similar nodes across graphs in this community should reflect this fact. The contrary would happen with, for example, work colleagues who you don't usually share anything with.

---

10 CBIR

5 Evaluation and Work Plan

The evaluation of the results obtained for this work will depend largely on the collection of a corpus of photographs that is representative of the domain together with any associated metadata. It is envisaged that the use of a mixture of sources from the web (e.g. Flickr, Facebook, etc.) as well as directly from human users will be enough to provide a good sized heterogeneous corpus. This will not only aid the development of an approach that is valid in the real world, but also the testing of any algorithm that is derived from this work in terms of its accuracy.

It is also possible that some user based testing may be necessary in order to determine the level of satisfaction that is achieved by using this approach.

Some of the work that has been achieved so far involved an experiment related to capturing semantics from photographic descriptions contained within a large corpus of photographs from Flickr. For more details see [5].

Future plans include developing a preliminary model of the context network as well as experimenting with the graph based approach detailed previously for propagating metadata from annotated photographs to unannotated ones according to a user's context network.

References

[1] T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web," *Scientific American Magazine*, May. 2001.

[2] M.C. Schraefel, "What is an analogue for the semantic web and why is having one important?," *HT '07: Proceedings of the 18th conference on Hypertext and hypermedia*, Manchester, UK: ACM Press, 2007, pp. 123--132.

[3] D. Frohlich et al., "Requirements for photoware," *2002 ACM conference on Computer supported cooperative work*, New Orleans, Louisiana, USA: ACM Press, 2002, pp. 166-175.

[4] A. Miller and K.W. Edwards, "Give and Take: A Study of Consumer Photo-Sharing Culture and Practice," *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*, San Jose, California, USA: ACM Press, 2007, pp. 347--356.

[5] R.F. Carvalho, S. Chapman, and F. Ciravegna, "Extracting semantic meaning from photographic annotations using a hybrid approach.," *MMIU'08: Proceedings of the 1st International Workshop on Metadata Mining for Image Understanding*, Funchal, Madeira - Portugal: 2008.

[6] M.M. Tuffield et al., "Image Annotation with Photocopain," *Proceedings of the Fifteenth World Wide Web Conference (WWW2006)*, 2006.

[7] N. O'Hare et al., "MediAssist: Using Content-Based Analysis and Context to Manage Personal Photo Collections," *Image and Video Retrieval, 5th International Conference*, Tempe, AZ, USA: Springer, 2006, pp. 529-532.

[8] A. Chakravarthy, F. Ciravegna, and V. Lanfranchi, "Cross-media document annotation and enrichment," *1st Semantic Authoring and Annotation Workshop (SAAW2006)*, 2006.

[9] L. von Ahn and L. Dabbish, "Labeling images with a computer game," *ACM Conference on Human Factors in Computing Systems (CHI)*, 2004, pp. 319--326.

[10] X. Zhu, "Semi-supervised learning with graphs," May. 2005.

[11] E. Minkov and W.W. Cohen, "An Email and Meeting Assistant using Graph Walks," *Third Conference on Email and Anti Spam*, Mountain View, California USA: 2006.

[12] W.H. Hsu, L.S. Kennedy, and S. Chang, "Video search reranking through random walk over document-level context graph," *Proceedings of the 15th international conference on Multimedia*, Augsburg, Germany: ACM Press, 2007, pp. 971-980.

Process Mediation for Semantic Web Services

Emilia Cimpian

Semantic Technology Institute Innsbruck
emilia.cimpian@sti2.at

1 Research Problem

It is a common scenario for the business environment that one process needs to *communicate* with another process, in order to fulfill its goal. For example, the simple action of paying a bill to a service provider can be seen as two processes that are communicating: one process is defined by the client, its own steps to be taken for paying the bill; the other one belongs to the service, the sequence of activities it performs in order to obtain the payment. If a bank is also involved (which is currently the case in most of this type of situations), we can even talk of three different processes performed by three different entities in order to obtain the final result: the bill had been paid.

The problem addressed by this thesis is how two (or more) processes can successfully interact in order to accomplish a common goal. The processes considered are semantically defined and any inputs and outputs of a process needs to also be represented using an ontology.

This thesis addresses the problem of solving heterogeneity mismatches between previously defined processes. The assumption made is that the processes should not adjust in order to match the processes they want to interact with, from various reasons. Either they are involved in more than one interaction, and this adjustment will damage those, or the business partner owning the processes simply does not want to change anything. In this case, the communication can be hampered even if all the data is available [2].

This thesis makes the distinction between two different heterogeneity problems: process model heterogeneity and communication heterogeneity. In the first case, the processes are incompatible, that is no automatic solution can be developed for overcoming the heterogeneity problem. In this situation the inputs of a business expert are needed, and the process mediator will have to provide semi-automatic support for the domain expert. In the second case the processes are compatible, the mismatch existing only on the message exchange level. In this case the process mediator can provide a completely automatic mediation solution. An example of mismatch that can not be automatically solved is when one process expects a message that the other one will never send. In this case the domain expert can select a third process that will generate that message, or manually create it.

The first step for process mediation is to determine the nature of the problem, if it can be solved automatically or not. The heterogeneity problems that can be automatically solved are called *solvable* (or *communication*) *mismatches*, while the ones that require domain expert interactions are called *unsolvable* (or *process model*) *mismatches* [1].

2 Related Work

Process mediation is still a poorly explored research field, in the context of Semantic Web Services. The existing work represents only visions of mediator systems able to resolve in a (semi-) automatic manner the processes heterogeneity problems, without presenting sufficient details about their architectural elements. Still, these visions represent the starting points and valuable references for the future concrete implementations.

Two integration tools, Contivo[1] and CrossWorlds[2] seemed to be the most advanced ones in this field.

Contivo is an integration framework which uses metadata representing messages organized by semantically defined relationships. One of its functionalities is that it is able to generate transform code based on the semantic of the relationships between data elements, and to use this code for transforming the exchange messages. However, Contivo is limited by the use of a purpose-built vocabulary and of pre-configured data models and formats.

CrossWorlds is an IBM integration tool, meant to facilitate the B2B collaboration through business processes integration. It may be used to implement various e-business models, including enhanced intranets (improving operational efficiency within a business enterprize), extranets (facilitating electronic trading between a business and its suppliers) and virtual enterprizes (allowing enterprizes to link to outsourced parts). The draw-backs of this approach is that different applications need to implement different collaboration and connection modules, in order to interact. As a consequence, the integration of a new application can be done only with additional effort.

3 Contributions

The main contribution of this thesis is *the development of a semantic process mediation solution*. This overall accomplishment consists of a number of smaller contributions:

1. Identification and formalization of a set of atomic problems that can be automatically solved by a mediator (solvable or communication mismatches), as well as identification of a set of problems that can not be automatically overcome (unsolvable or process model mismatches).

2. Development of a run-time process mediator able to address the solvable mismatches.

3. Development of a design-time process mediation for allowing the domain expert to accommodate for the unsolvable mismatches.

4. Development of a comprehensive architecture for process mediation.

Because of space constraints this extended summary of the thesis contains only details of how the formalization and of the algebra developed in the thesis.

3.1 Notations and Definition

The service mediator performs an automatic analyze of the two processes involved in a communication. The internal decisions taken inside any of the processes are not relevant

[1] http://www.contivo.com/

[2] http://www.sars.ws/hl4/ibm-crossworlds.html

in this case, the mediator operating on the level of messages sent and received during the actual communication. In this sense it can be considered that the mediator operates on one particular branch of each process involved in the communications. That is, if depending on one condition one of the processes can perform one activity or another, the run-time mediator sees only the result of evaluating that condition, only the activity that is performed.

Furthermore, in a semantic environment the messages are important only from the point of view of the semantic information they carry. This information consists of instances of concepts defined in an ontology used in the description of the process (in the process model). If the process description specifies that *message* M_1 *contains instance* I_1 *of concept* C_1, the mediator understands this as M_1 *consists of an instance of* C_1, or in other words *an instance of* C_1 *is being sent or received*. The previous two formulations are further simplified to *Message* C_1.

If a message M_1 consists of multiple instances of multiple concepts (C_1, C_2,... C_n) it will be referred to as: *message* C_1 *and* C_2 *and ... and* C_n. This definition still holds if multiple instances of the same concept are part of the same process, in which case the message will refer to every one of these instances.

The notation used for denoting that message C_1 is to be sent by a process is ***S(C)***, while a message that should be received by a process is represented by ***R(C)***. For denoting that a process should be either sent or received the notation used is ***A(C)*** (an action for handling the message C). If the message carries more then one instance, of types C_1, C_2,...C_n, this is denoted by ***A(***C_1**+**C_2**+...+**C_n***)***.

The order of messages is represented by using the symbol $\twoheadrightarrow$.

The message sequence of a process P is represented as $MS(P)$. If P exchanges n messages during a communication, then: $MS(P) = A(C_1) \twoheadrightarrow A(C_2) \twoheadrightarrow ... \twoheadrightarrow A(C_n)$

For representing the communication between two processes P_1 and P_2 the notation $\frac{MS(P_1)}{MS(P_2)}$ is used.

The fractions for representing a communication can be decomposed in multiple fractions, respecting the messages sequences of the processes involved in the communication.

$$If: \ MS(P_1) = MS_1(P_1) \twoheadrightarrow MS_2(P_1) \ and \ MS(P_2) = MS_1(P_2) \twoheadrightarrow MS_2(P_2)$$
$$then: \ \frac{MS(P_1)}{MS(P_2)} = \frac{MS_1(P_1)}{MS_1(P_2)} \twoheadrightarrow \frac{MS_2(P_1)}{MS_2(P_2)}$$

Furthermore, the following terms are defined:

Definition 1. *An* ***Atomic Send/Receive*** *(****Atomic S/R****) is considered to be that particular fragment of a communication consisting of one process sending a message and the other process receiving it.*

$$Atomic \ S/R(C) = \frac{S(C)}{R(C)} \ or \ Atomic \ S/R(C) = \frac{R(C)}{S(C)}$$

Definition 2. *A* ***Projection*** *of a process, denoted by* $\pi(P)$*, is a derived process obtained from P as the result transformations performed by the run-time Process Mediator.*

The communication between two processes is equivalent with the communication between one process and the projection of the other process, which is denoted by the symbol $\approx$.

$$\frac{MS(P_1)}{MS(P_2)} \approx \frac{MS(\pi(P_1))}{MS(P_2)}$$

Definition 3. *There is a **Match** between two given processes if the communication between them can be represented as a sequence of Atomic S/R.*

The notation used for denoting that two processes P_1 and P_2 match is:

$$Match(P_1, P_2)$$

Definition 4. *Two processes are considered to be **Compatible** if there is a Match between them or if every mismatch is at the message sequence level.*

The notation used for denoting that two processes P_1 and P_2 are compatible is:

$$Compatible(P_1, P_2)$$

Both *Match* and *Compatible* relationships are symmetric.

3.2 Process Mediation - Lemmas and Theorems

A set of lemmas can be defined for obtaining the projection of a process, given its message sequence. An example of such lemma is:

Lemma 1. *For a given process P where $MS(P) = MS_1(P) \twoheadrightarrow S(C) \twoheadrightarrow MS_2(P)$, a process P' such as $MS(P') = MS_1(P) \twoheadrightarrow MS_2(P)$ is a projection of P (i.e., $P' = \pi(P)$).*

A total of 8 lemmas are defined for governing the creation of the projections, based on the message exchange sequence of all the processes involved in the communication. All of them define the conditions under which the messages can be interchanged in order to create projections.

Furthermore, the thesis defines and proves several theorems for the process interoperability, given the relationships between their projections. The most general one is:

Theorem 1 *Any two processes P_1 and P_2 are compatible if and only if exist two projections P_1^n and P_2^m that match, where $P_1^i = \pi(P_1^{i-1})$ and $P_2^j = \pi(P_2^{j-1})$ where $1 \leq i \leq n$ and $1 \leq j \leq m$.*

As part of this thesis, a run-time process mediator able to apply the projections described above for each process in respect with the process it communicates with was developed. The appropriate projections are determined based on the exchange patterns of the processes involved in the communication, involving a detailed analyze of the processes and the evaluation of the rules that govern the message ordering. For dealing with the heterogeneity problems that cannot be automatically solved, a design-time process mediator which provides support to the domain expert was also developed.

4 Evaluation

The approach and prototypes developed in this thesis will be evaluated based on a two-fold methodology. Firstly, the thesis will consider a real use-case scenario developed as part of the SUPER project; this type of evaluation will prove that the approach is applicable in a real scenario. Secondly, in order to prove the correctness and completeness of the formal modeled developed in this theses, it will be evaluated against the existing workflow data patterns, based on the data *visibility*, *interaction*, *transfer* and *routing* [3].

5 Work Plan

The most important steps in accomplishing the objectives of this theses were already performed:

1. Identification of the types of mismatches that can be automatically solved.
2. Formalization of the operations that can be automatically performed by a mediator without breaking the communication.
3. Development of proof of concepts design-time and run-time prototypes needed for the process mediation.
4. Identification of a real-use case scenario, detailed analyze of the problems raised by the scenario.

However, important phases needed for the completion of the theses are still ongoing, such as:

1. Development of a comprehensive architecture for process mediation which will allow the integration of the two prototypes previously developed, providing complete solutions for process mediation;
2. Evaluation of the prototypes based on the available scenario;
3. Evaluation of the completness and correctness of the approach based on the existing workflow data patterns

References

1. C. Bussler. *B2B Integration: Concepts and Architecture*. Springer, 2003.
2. E. Cimpian, A. Mocan, and M. Stollberg. Mediation enabled semantic web services usage. *Proceedings of the First Asian Semantic Web Conference*, 09 2006.
3. N. Russell, A. H. ter Hofstede, D. Edmond, and W. M. van der Aalst. Workflow data patterns. Technical report, Workflow Patterns Initiative, 2005.

Towards Ontology Mapping for Ordinary People

Colm Conroy[1],
(Co)Supervisors: Declan O'Sullivan[1], David Lewis[1]

[1] Knowledge and Data Engineering Group,
Trinity College Dublin
{coconroy,declan.osullivan,dave.lewis}@cs.tcd.ie

1 Introduction

The reasons for the lack of uptake of the semantic web amongst ordinary users can be attributed to technology perception, comprehensibility and ease of use. It is perceived that the creation of ontologies is a top-down and complex process, whereas in reality ontologies can emerge bottom-up and be simple. Ontology technology is based on formal logics that are not understandable for ordinary people. Finally there is significant overhead for a user in the creation of metadata for information resources in accordance with ontologies. To address these three problems, we believe that the interfaces to ontology tools will need to be engineered in such a way as the tools disappear into the background from the ordinary person's perspective.

There is a common diversity between the semantic models of interest between people. If users of the semantic web choose to model their interest with a personal ontology, their ontology will need to be mapped to the models used in the various diverse communities by the person himself or herself. The automatic and efficient matching between the personal ontology and the models used by others (collaborative tags and/or community ontologies) can be achieved through the application of a variety of matching techniques [1]. Fully automatic derivation of mappings is considered impossible as yet [2], and the majority of state of the art tools in the ontology mapping area [3] and the community ontology creation area [4] rely on a classic presentation of the class hierarchy of two ontologies side by side and some means for the user to express the mappings. These approaches predominately assume that the mapping is being undertaken by an expert: who does not require a personalized interface; whose explicit task is to generate a "one size fits all" full mapping (to be used in common by several applications); and who undertakes the task during a small number of long sessions. The number of user trials that have taken place have also been small [5] and those that have, have focused purely on the effectiveness and do not address usability issues (an exception recently being that of [6]). In contrast, we propose that the user who will benefit from mappings (through usage by their applications), will undertake themselves partial targeted mappings, gradually and over time, using techniques that address usability issues, support personalization and enable control of the mapping interactions.

2 State of Art

There is an emergence of focusing on better support for users within the ontology mapping area with cognitive support for user [7], a community-driven matching [8], explanation of matches to users [9], and developing a formal model for ontology mapping [10]. One of the key problems which have seen little research from a cognitive perspective is how to display the match information in a manner that is natural for the user. The visualization of ontology (schema) mapping can be categorised into four different categories: *tree-type*, *object-type*, *instance based spreadsheet*, and *hyperbolic*. *Tree-type interfaces* are the most common for mapping tools and represent the semantic models side by side in a tree form, e.g. COMA++ [3], and mappings can be represented in two different ways with lines drawn between matching terms or via a mapping table that contains all the mappings. *Object type interfaces* represent the ontology in an object type or UML structure, with mappings between the two schemas being drawn using lines (and symbols to represent the type of mapping), e.g. SMART [11] '=' is used for equals, ')' for subset and so on. *Spreadsheet type interfaces*, e.g. Webscripter [12], uses spreadsheet functionality to display global mapping tables containing instance data from different users. Finally a *hyperbolic interface* is one where the source and target models are represented by hyperbolic graphs in different frames, e.g. Schema Mapper [13]. AlViz [14] is a tab plug-in for Protege which provides visualization techniques to facilitate user understanding of alignment results. There are other types of ontology tools which use different types of interfaces; GINO [15] is a guided input natural language ontology editor that allows users to edit and query ontologies in a language akin to English. OntoViz [16] is a protégé plugin which represents ontologies via a direct graph with concepts in boxes and relations defined with lines.

3 Research Question/Contribution

The research question we are addressing is *what kind of interactions will be acceptable, efficient and effective for an ordinary user to achieve semantic mappings gradually and over time between information models of interest to the user*. In particular our work will provide:

- *Design of a mapping framework in support of ordinary people*: There is a need to make the ontology mapping process as unintrusive and as natural as possible, as it is important not to interrupt ordinary users during their daily life, so that they do not see mapping as inconvenient work but more as something that will be beneficial to them, and where they can clearly see the benefits.
- *Determine the most appropriate user interaction in the process of constructing mappings*: There has been little or no research by the current state of the art completed regarding user interaction within mapping systems. While visualization is a key factor in this problem another key issue is determining the most appropriate (different) way(s) for (different) people to construct mappings, e.g. draw lines graphically, 'yes/no' answer on questionnaire, etc…

- *Determine the appropriate ways to engage the user over time*: There is a need to engage the user in the mapping process over time rather than one long session. This need comes from the realization that by reducing the mapping process into piece wise comparison will make the mapping process easier to comprehend and also will be able to give feedback to the user of the mapping choices they make outside of the mapping sessions. Another issue will be does collaborative knowledge sharing (via groups) assist ordinary people in the mapping task

An outcome of this research will be a process (Figure 1), methodology and tools.

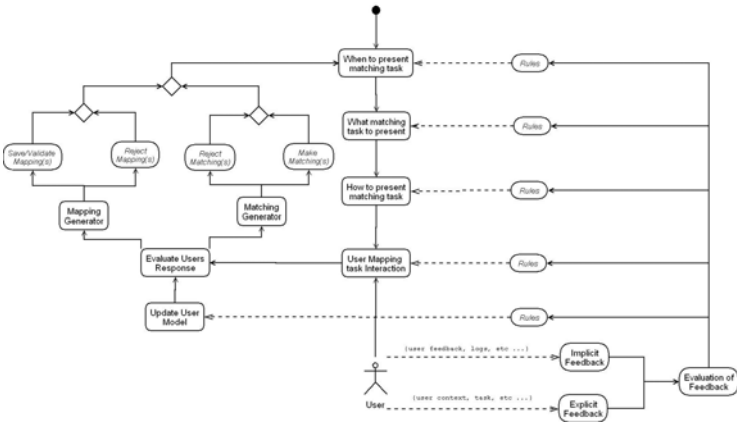

Figure 1: Mapping Process

4 Evaluation

In our initial experiment undertaken early in 2007 we aimed to determine the most practical way of visually displaying the mapping information for different groups of users. Our hypothesis was *using a Question & Answer natural language interface to visually display ontological information helps in making mapping more familiar and accessible and also reduces the complexity of the mapping process for users*. The intention of our experimentation was to investigate the effect and usability of a natural language prototype tool (NL) on three groups of users and to determine whether it made the mapping task more user friendly for one group over another. In addition we wanted to contrast our tool against a current state of the art mapping tool. We chose COMA++ as our State of the Art tree type graph mapping tool. The three different groups of users were: Ontologically aware, Technology aware and Non-Technology Aware (Casual Web User). The paper [17] goes into detail about the experiment, some key conclusions drawn were:

- On the positive side, results suggested casual web users can map effectively and efficiently even compared to ontology aware users. Using Natural Language seemed to help people read and understand the information and the Q&A approach helped in navigating through the mapping task.
- On the negative side, casual web users found it very restrictive to be limited to a narrow range of mapping terminology, e.g. "corresponds" and "similar to" when answering mapping questions. In addition, some users were unclear about the benefit in engaging in the mapping task.

In our current experiment (finishing May 2008) we are focusing on whether it is valuable to embed the mapping process within the user environment, designing a user-centric mapping process, and addressing the negative concerns garnered from the previous experiment by allowing the user to be more expressive by allowing them to 'tag' the mapping relation. Our hypothesis is *the mapping task can be simplified and become unintrusive by embedding the mapping process within the user environment and by using a 'tagging' approach paradigm.* By using the power of "Web 2.0" through a Firefox extension within our new 'tagging' prototype, we aim to engage the user and display matching collections at appropriate times within their own work environment, see [18] for more details. We use online questionnaires, interviews, and a log of each user's actions to evaluate the impact of the 'tagging' prototype. In particular through the use of our implementation over the coming months we aim to investigate whether casual web users will be able to use tagging to turn matches into expressive mappings in a straightforward, practical and natural manner. We will also investigate whether embedding the mapping interface inside a browser extension will allow the mapping process to take place over time within a casual web persons work environment in an unobtrusive, sensible, and normal way.

Our next experiment is going to be investigating the effects of collaborative knowledge sharing via groups (from June 2008 to October 2008). A key factor of this experiment will be to investigate whether the knowledge shared by ontological aware users can be beneficial to casual web users. Another issue is which users are the most beneficial for validating each matching pair question, i.e. will users with musical background be better than ontological aware user in relation to music based matching's. Another aspect is whether it is better to divide the users into different groups and what type of characteristics determine which group(s) each user is put into. A final feature of investigation will be if categorising the 'user-defined' tags collaboratively, whether globally or within groups, is of any help to users.

In our first experiment we looked at displaying ontological information with natural language and to analysis the difference between a graph types interfaces. The natural language used was far from ideal and although the results showed the benefits of using this type of interface we intend to revisit this objective to make tests with different visual types of interfaces (from November 2008 to April 2009). We intend to investigate different visual types of natural language, such as representing the concepts with shallow text generation as discussed in [19], and other different visual types. Although finding a one-size-fits-all interface for everybody may be impossible, we believe there can be a basic interface which can be suitable to most

casual web users while also allowing these users to change the visualization to their needs, i.e. some may rather graph based to natural language and vice versa, etc...

A key problem is reducing mapping tasks to being *unintrusive* to casual web user. We intend to explore in what context and at what time should a mapping task be performed by the user (from April 2009 to June 2009). Finally the thesis write up will occur from July 2009 till November 2009.

References

1. Shvaiko P. and Euzenat J., "A Survey of Schema-based Matching Approaches", Journal of Data Semantics, Long version in DIT Technical Report DIT-04-87, 2004.
2. Noy N., "Semantic Integration: A Survey of Ontology-Based Approaches", Special Issue on Semantic Integration, SIGMOD Record, Volume 33, Issue 4, pages 65-70, 2004.
3. Aumüller D., Do H., Massmann S., Rahm E., "Schema and Ontology Matching with COMA++", Proc. SIGMOD 2005 (Software Demonstration), Baltimore, June 2005.
4. Zhdanova A., "Towards Community-Driven Ontology Matching", Proceedings of K-CAP, Banff, Canada, 2005.
5. Jameson A., "Usability and the Semantic Web", 3rd European Semantic Web Conference, ESWC 2006, Budva, Montenegro, June 11-14, 2006, Springer LNCS Vol. 4011
6. Falconer S. M., Noy N. F., Storey M., "Towards understanding the needs of cognitive support for ontology mapping", ISWC 2006, 2006, Springer LCNS Vol. 4273
7. Falconer S. M., Storey M., "A Cognitive Support Framework for Ontology Mapping", In Proceedings of ISWC+ASWC, pages 114-127, 2007
8. Zhdanova A., Shvaiko P., "Community-driven ontology matching", In Proceedings of ESWC, pages 34-49, 2006
9. Shvaiko P., Giunchiglia F., Pinheiro da Silva P., McGuinness D., "Web Explanations for Semantic Heterogeneity Discovery", In Proceedings of ESWC, pages 303-317, 2005
10. Mocan A., Cimpian E., Kerrigan M., "Formal Model for Ontology Mapping Creation", In Proceedings of ISWC, pages 459-472, 2006
11. Morishima A., Okawara T., "SMART: a tool for semantic-driven creation of complex XML mappings", SIGMOD '05, Baltimore, MD, USA, ACM Press, NY, 2005
12. Yan, B., Frank, M., Szekely, P., "WebScripter: Grass-roots Ontology Alignment via End-User Report Creation.", ISWC 2003, LNCS, vol. 2870, Springer, Heidelberg, 2003
13. Raghavan, A., Rangarajan, D., Shen, R., Goncalves, "Schema Mapper: A Visualization Tool for DL Integration.", In Proceedings JCDL2005, Denver, p. 414, 2005.
14. Lanzenberger M., Sampson J., "Alviz – a tool for visual ontology alignment", In Proceedings of the 10th International Conference Information Visualization, 2006
15. Bernstein A., Kaufmann E., "GINO – A Guided Input Natural Language Ontology Editor", ISWC, 2006
16. OntoViz, http://protege.stanford.edu/plugins/ontoviz/
17. Conroy C., O'Sullivan D., Lewis D., "A Tagging Approach to Ontology Mapping", The 2nd International Workshop on Ontology Mapping, ISWC, Korea, 2007 (poster)
18. Conroy C., O'Sullivan D., Lewis D., "Ontology Mapping Through Tagging", The International Workshop on Ontology Alignment and Visualization, CISIS, Barcelona, 2008
19. Kalyanpur A., Hashmi N., Golbeck J., Parsia B., "Lifecycle of a Casual Web Ontology Development Process", Proceedings of the WWW2004, May 18, 2004

Methodology for Searching Entities on the Web*

Renaud Delbru

Digital Enterprise Research Institute
National University of Ireland, Galway
firstname.lastname@deri.org

1 From a Web of Documents to a Web of Entities

The Semantic Web is driven by the idea of moving from a Web of documents, designed for human consumption, to a Web of data in order to "create a universal medium for the exchange of data where data can be shared and processed by automated tools as well as by people"[1].

Nowadays, more and more machine-readable annotations and meta-data are available on the Web. This data, typically codified using the Resource Description Framework (RDF) or Microformats, is accessible directly via HTTP. Microformat enables the annotation of an entity in a web page, whereas RDF enables the description of anything that can be named using a Uniform Resource Identifier (URI). By describing the relationships between resources, the Web moves from a Web of documents to a semantically interconnected Web of entities.

Although data are available, data consumers face a challenge due to the decentralised publishing infrastructure of the Web: they need to locate information about an entity and handle multiple, possibly discording, views of the entity.

Search engines are the primary method for accessing information on the Web, i.e. finding relevant documents given a keyword-based query. By leveraging the Web of entities, we can imagine an entity-centric search engine which, given a query, would support the user in obtaining an aggregated and balanced view of the data available on the Semantic Web. Given that the Semantic Web data are machine processable, the most interesting use of such an engine could be made by machines themselves: any application could use one such engine directly to find, interconnect and enrich information.

Searching information about a particular entity on the Web raises new challenges: (i) how to efficiently locate and retrieve Semantic Web data and (ii) how to integrate data on a decentralised and heterogeneous information space. We aims to propose a comprehensive methodology for searching entities on the Web along these requirements.

2 Research Problem

We plan to tackle such complex problems by exploring how existing and proven robust technologies can be advanced and specialized specifically to address the needs of an

* This material is based upon works supported by the European FP7 project *Okkam - Enabling a Web of Entities* (contract no. ICT-215032), and by Science Foundation Ireland under Grant No. SFI/02/CE1/I131.

[1] Semantic Web Activity Statement: http://www.w3.org/2001/sw/Activity.html

entity-centric search engine. In particular, our work will focus on the topics illustrated in the following sections.

2.1 Adapting Information Retrieval engines for Semantic Web Data

Standard Web search engines are intensively using Information Retrieval (IR) techniques for locating relevant information on the Web. Information Retrieval is a well studied field [1] and many optimisations have been developed for efficiently storing and querying large amount of information. Techniques such as inverted indexes [2] have proved to scale to the size of the Web (e.g. Google). The shortcoming of such systems is that they can only answer simple queries, e.g. a boolean combination of words, but are not really meant to query relationships between entities, e.g. a graph pattern.

On the contrary, entity-centric search engines such as SWSE [3] are built on a data structure which is more similar to relational databases than to IR engines. SWSE relies on YARS [4], a distributed RDF store, for storing and querying large amounts of graph-structured data. Such systems can typically answer complex conjunctive queries involving large joins, but they are in turn difficult to scale since they need clever indexes for query efficiency which are however computationally expensive to update.

Our intuition is that it is possible to construct a fast and scalable entity centric search engine based on a two-tier architecture: a modified IR engine to efficiently perform a preliminary semantic document selection, and an optimised triple level post-processing to answer complex queries. Our research will therefore focus on how to employ existing IR engines to perform useful queries over semantically structured documents.

Information Retrieval engines, however, are primarily designed for unstructured text information, and not for graph-structured information such as RDF. Information Retrieval engines for Semantic Web data have notable previous works with Semplore [5] and ESTER [6] which however were developed with different goals than those we consider. The developers of Swoogle [7] have also discussed the problem of introducing a new search paradigm for Semantic Web resources and emphasized the importance of combining knowledge inference with information retrieval methods.

2.2 Optimising Inference at Web Scale

Reasoning over semantically structured documents enables to make explicit what would otherwise be implicit knowledge: it adds value to the information and enables an entity-centric search engine to ultimately be much more competitive in terms of precision and recall [7]. The drawback is that inference can be computationally expensive, and therefore prevent efficient indexing.

The novel aspect that our work covers is how to reason over semantically structured documents that have been harvested from the Web. To reason on documents, we assume that ontologies, which are referenced explicitly with OWL:IMPORTS or implicitly by using properties and classes of a certain namespace, are also part of the Semantic Web as dereferenciable data, in accord with the W3C Best Practice[2]. As ontologies might

[2] Best Practice Recipes for Publishing RDF Vocabularies: http://www.w3.org/TR/swbp-vocab-pub/

refer to other ontologies, the web fetching process is recursive and should, in theory, be repeated for each harvested documents independently.

The proposed research will focus on how to maximally reuse the results of such "web closure reasoning", i.e finding and exploiting the referenced ontologies, that has been performed over previously indexed semantically structured documents in order to minimise the computational cost of indexing. We will also considers how to "keep in quarantine" reasoning tasks and inference results in order to prevent maliciously crafted web ontologies to alter the semantics of agreed ontologies published by third parties on a global level. For example, if an ontology states that FOAF:NAME is an inverse functional property, an inferencing agent should not consider this axiom outside the scope of the document that references this particular ontology.

The coordinate use of the features offered by the IR and inference engines will be demonstrated in the applications described in the following sections.

2.3 Identification, Coreference Resolution and Information Merging

Due to its decentralised publishing infrastructure, information about an entity are generally spread across the Web. The identification of an entity is fundamental for discovering complementary data sources. The use of URI makes easier the identification of an entity, but the Unique Name Assumption (UNA) does not hold. In theory, a single URI uniquely identifies a resource, but it is unrealistic to assume that data publishers can universally agree on a single identifier for each resource. Therefore, the identification of an entity among the Semantic Web becomes uncertain since two identifiers, apparently distinct, can refer to a unique entity.

The coreference problem is well known across various research communities with a variety of different names, such as record linkage [8], entity resolution [9], reference reconciliation [10] or object consolidation [11]. A wide variety of algorithms has been developed for resolving the coreference problem, but these are generally not designed for Web scale and semi-structured data. Recent initiatives amongst the Semantic Web community addressed the problem of resource identification: [12] described the phenomenon of the proliferation and coreference of URIs and the OKKAM project[3] proposed to research an infrastructure for assigning global identifiers at Web scale.

The problem of identification and coreference resolution will be a natural testbed for the IR and inference engines that we described previously. The IR engine will enable to perform a blocking pass [13] before executing complex coreference resolution and, coupled with the inference engine, will permit more advanced reasoning than what was possible in the Semantic Web object consolidation work described in [11].

Clearly, coreference resolution is an important enabler for information merging. More factors, however, have to be taken into consideration before aggregating diverse information sources. Entity descriptions are generally produced under a certain context (provenance, time, etc.). The descriptive information is usually a subjective view of the entity with a certain level of reliability. Merging these descriptions can result in inconsistent and contradictory information. In order to enable a proper data integration, we

[3] http://www.okkam.org/

have to keep information in its context which is naturally supported by the document-centric storage we adopt.

3 Methodology

We will evaluate the methodology for searching entity by implementing a solution for each identified problem, by integrating them into a single software platform and by performing a qualitative evaluation of the resulting platform. In addition, we will perform an evaluation of each solution with a dedicated corpus, as described below.

Information Retrieval Engine for Semantic Web Data A benchmark, including index size and query response time, against other systems is planned.

Semantic Web Inference Engine The evaluation of the inference engine will include an analysis of its complexity in term of size and response time.

Entity Identification and Coreference Resolution The evaluation of the coreference resolution system requires a gold standard dataset for analysing the precision of the different algorithms.

4 Achievements and Work Plan [4]

Information Retrieval Engine for Semantic Web Data We achieved, as part of the Sindice project [14], a first prototype of the Information Retrieval engine. The current system has currently indexed more than 2 billions of triples. The system enables fast lookup of URIs, keywords and Inverse Functional Properties (IFP) through a human interface or a HTTP API for machine access. We are currently finishing a second prototype that enables queries of increased complexity and semantic meaning, i.e combining URIs and keywords and adding triple-structure. We foresee a third and final prototype capable of answering more complex queries involving simple joins.

Semantic Web Inference Engine We have developed a prototype of an optimised inference engine that enables inference of a subset of OWL at indexing time. Preliminary results of this work has been published in [14]. We have formalised an advanced inference engine that avoid malicious users from "infecting" cached data on a global scale. Its development is in progress.

Entity Identification and Coreference Resolution As a next step we will tackle a coreference resolution system, based on the IR and inference engines. The first task will be to implement a prototype for identifying entities with the help of OWL:SAMEAS statement and IFPs, e.g. the e-mail of a person. The prototype will be able to return an aggregated view of the entity information available on the Web. The second task will be to improve the system with "pair-wise" matching algorithms.

[4] The work on the thesis has formally started in February, 2007.

References

1. Baeza-Yates, R.A., Ribeiro-Neto, B.A.: Modern Information Retrieval. ACM Press / Addison-Wesley (1999)
2. Zobel, J., Moffat, A.: Inverted files for text search engines. ACM Computing Surveys **38** (2006) 6
3. Harth, A., Hogan, A., Delbru, R., Umbrich, J., Ó'Riain, S., Decker, S.: SWSE: Answers before links! In: Proceedings of the Semantic Web Challenge, 6th International Semantic Web Conference. (2007)
4. Harth, A., Umbrich, J., Hogan, A., Decker, S.: YARS2: A federated repository for querying graph structured data from the web. In: Proceedings of the 6th International Semantic Web Conference. (2007) 211–224
5. Zhang, L., Liu, Q., Zhang, J., Wang, H., Pan, Y., Yu, Y.: Semplore: An IR approach to scalable hybrid query of semantic web data. In: Proceedings of the 6th International Semantic Web Conference. (2007) 652–665
6. Bast, H., Chitea, A., Suchanek, F., Weber, I.: ESTER: efficient search on text, entities, and relations. In: SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, ACM (2007) 671–678
7. Mayfield, J., Finin, T.: Information retrieval on the Semantic Web: Integrating inference and retrieval. In: Proceedings of the SIGIR Workshop on the Semantic Web. (2003)
8. Fellegi, I.P., Sunter, A.B.: A theory for record linkage. Journal of the American Statistical Association **64** (1969) 1183–1210
9. Benjelloun, O., Garcia-Molina, H., Jonas, J., Su, Q., Widom, J.: Swoosh: A generic approach to entity resolution. Technical report, Stanford University (2006)
10. Dong, X., Halevy, A.Y., Madhavan, J.: Reference reconciliation in complex information spaces. In Özcan, F., ed.: SIGMOD Conference, ACM (2005) 85–96
11. Hogan, A., Harth, A., Decker, S.: Performing object consolidation on the semantic web data graph. In: Proceedings of the WWW2007 Workshop I3: Identity, Identifiers, Identification, Entity-Centric Approaches to Information and Knowledge Management on the Web. (2007)
12. Jaffri, A., Glaser, H., Millard, I.: URI identity management for semantic web data integration and linkage. In: 3rd International Workshop On Scalable Semantic Web Knowledge Base Systems, Springer (2007)
13. Elmagarmid, A.K., Ipeirotis, P.G., Verykios, V.S.: Duplicate record detection: A survey. IEEE Transactions on Knowledge and Data Engineering **19** (2007) 1–16
14. Oren, E., Delbru, R., Catasta, M., Cyganiak, R., Stenzhorn, H., Tummarello, G.: Sindice.com: A document-oriented lookup index for open linked data. International Journal of Metadata, Semantics and Ontologies **3** (2008)

Event and Sentiment Detection in Financial Markets

Uta Hellinger

AIFB, Universität Karlsruhe, Germany
hellinger@aifb.uni-karlsruhe.de

Abstract. Today, traders in financial markets are confronted with the problem that information is distributed over diverse sources and that there is too much information available. In our work we develop methods and tools to help traders to overcome this information overload by enabling the integrated view on news from various sources, by filtering relevant news and by providing decision support for traders. Another goal of our work is to propose a formal model of the impact of news on asset prices and thus enable better predictions of stock prices than possible with purely text mining based approaches.

1 Research Problem

Traders in financial markets are confronted with the problem that too much information is available from various, heterogenuous sources like newswires, forums, blogs and collaborative tools. In order to make accurate trading decisions, traders have to filter the relevant information efficiently so that they are able to react to new information in a timely manner.

The focus of our work is the development of methods and tools to support traders in this process. The goal is to provide an integrated view on news from different sources, to filter those news that have significant market impact and to help the users to decide how to react to newly published information. The development of these methods and tools raises the following questions:

- How can information from various, heterogeneous sources be integrated? News that are found in various sources differ in their content and in available annotations: news published by newswires are annotated with standardized metadata (which differ between newswires), blog posts are at the most tagged with some keywords. The information published on different web sites has to be collected and metadata has to be mapped to a single format such that all news can be processed using the same algorithms.
- How can the important news be filtered? Users can not monitor all relevant news services and process all the information that is published by these services. Therefore, methods are needed to filter news that have significant market impact. These methods have to detect important events, sentiments and expectations concerning the market.

– How can the users' trading decisions be supported? Price changes are caused by changes in the expectations concerning a company. Therefore, expectations and precise information (like the amount of the quarterly result) should be used for the prediction of price changes. This requires mechanisms that extract necessary information from texts, formalize it and make predictions based on it.

Although our work focuses on a specific application domain, its results will be relevant for other applications, as we show how information from different sources can be integrated and used to provide decision support.

2 Related Work

Our work is related to research in the domains of text mining (especially event and sentiment detection), information extraction, semantic web and finance.

A variety of systems for the prediction of asset price developments based on recently published news have been developed (see [1] for an overview). These systems are based on text classification, where the target categories are derived from financial data. Although they are closely related to our intended application, they have two important weaknesses: (i) expectations, which heavily influence the development of asset prices, and (ii) quantified information (like the value of paid dividends or the amount of the annual profit), which enables the quantification of the expected price change, are not considered in these systems.

Online event detection methods have for example been developed by [2], and [3]. These methods only attempt to identify new events mostly using clustering techniques without trying to formalize them semantically, which is required for matching them against our expectation models.

Work on sentiment detection in a finance context include [4] and [5]. While Das and Chen [4] use linguistic features to classify messages in negative and positive ones and then examine the correlation with stock price changes, Koppel et al. [5] use stock price changes to identify positive and negative news from which then describing features can be extracted.

While to the best of our knowledge no method for modelling expectations exists, Halaschek-Wiener and Hendler [6] have proposed an OWL-based news syndication framework to match publications and information needs. The subscribers' information needs which are described by conjunctive ABox queries are matched against publications which are formalized as ABox assertions.

Relation extraction is applied to populate ontologies from text - the problem we have to solve for extracting information on events and expectations from news. Bootstrapping usually is applied in these methods, where the web [7] or Wikipedia [8] are used as corpora to find patterns for describing relations. These methods have very low precision and recall in some applications which is problematic for our application.

Event studies study the impact of certain events on a firm's value (see [9] for an overview of the methodology). These will be useful to find events and aspects that should be taken into account in our system.

3 News Analysis Tool

Our solution to the problems discussed above is a tool which offers support in filtering important information and in decision-making. The planned architecture can be seen in Fig. 1. The components of this framework and requirements regarding each of them will be discussed in the following.

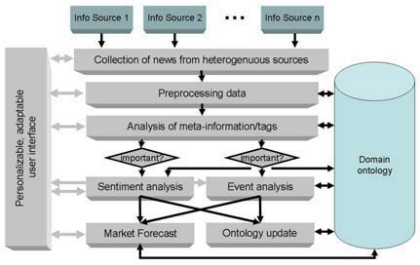

Fig. 1. Preliminary version of news analysis framework

The first component of our tool is responsible for the *collection of news from various, heterogenuous sources*. This component will monitor a huge number of relevant sources for new information and will make it available to the *Preprocessing data* component. The latter maps the available metadata to a single representation such that all data is processable in the same way in later steps. We will define an ontology for each news source's metadata and map these ontologies to one general one which will be used for further processing. If no metadata is available, the extraction of some annotations of the news' content like named entities becomes necessary to enable filtering of these news. However, only very efficient methods can be applied here as information extraction is quite expensive in general.

The *Analysis of meta-information/tags* component will examine the metadata to filter relevant news. It will decide whether a news item contains expectations concerning future events and will thus be processed by the *Sentiment analysis* component and whether it contains information on an actual event and will thus be processed by the *Event analysis* component. It is possible that a news item is processed by both components or that a news item is not important and thus will not be processed further.

Both the *Event analysis* and the *Sentiment analysis* component will apply information extraction methods to extract formal descriptions of the news' content. These descriptions will be used by the *Market forecast* component to predict the impact of a news item on the market by quantifying the difference of the actually published information from the expectations and the current status of the

domain as described in the ontology. The *Ontology update* component is responsible for the integration of the changes that occur due to published expectations and events into the ontology.

The *Domain ontology* is the backbone of the three previously described components. It describes the current status of the market and expectations concerning future events. We currently try to identify the most predictive features of news items. We develop a linear regression model that predicts market responses based on text features. This is the technique of choice as the predicted impact can be quantified and as the influence of each feature on the result can easily be seen. The developed model will help us in identifying the information that should be modelled in the domain ontology. It will also provide some prediction facility that can serve as a base line for the evaluation of more elaborate methods.

The tool will be personalisable and adaptable in the sense that users can specify their preferences, e.g. companies that they are especially interested in. This will be possible through the *Personalizable and adaptable user interface*.

An important requirement for the whole process is that it has to be extremely fast as significant price changes (in the short-term trade that we consider) can only be observed within one minute after the publication of a news item by a newswire.

4 Evaluation

The framework presented in the previous section requires a set of different evaluations. Firstly, an evaluation of the methods employed in each component has to be done. This especially means that an evaluation of the classification and information extraction methods in terms of precision and recall is required.

As our goal is the prediction of price changes based on expectation changes, the quality of the predictions serve as an evaluation of the domain ontology.

Finally, a user study is necessary to see how well users are supported by this kind of system and whether it helps them to make better trading decisions.

5 Work plan

So far, we have acquired news published by Reuters and information on intraday trades and quotes for over 240 markets in 2003 and we have developed aggregation functions for the financial data such that it can be used as training data for the methods we will develop. Given the huge amount of data we focus our experiences on the German market.

As mentioned in section 3 we currently work on the identification of the most predictive features. The next steps will be to compare the features we find to results of event studies available in the finance literature. In parallel to these steps we will develop the metadata ontologies and mappings between them. We will then build a classifier that filters the relevant news based on metadata. Once these methods are available we will build our domain ontology that models events

and expectations, develop the necessary information extraction methods and define how discrepancies between expectations and newly published information can be quantified for predicting the associated asset price changes.

The last component that we develop will be the user interface before we finish our project with user experiments that will hopefully show the benefit of the proposed tool.

6 Conclusion

The goal of our work is the development of a news analysis tool that supports traders in financial markets by filtering news and making predictions on the impact of news on the market. The contributions of our work will be:

- refined information extraction methods for the analysis of financial news
- ontologies of the financial domain that allow the formalization of news and their annotations as well as of expectations and events
- a method to quantify the distance of an event from expectations

7 Acknowledgements

This work was funded by the German Research Foundation (DFG) in scope of Graduate School Information Management and Market Engineering.

References

1. Mittermayer, M.A., Knolmayer, G.: Text mining systems for market response to news: A survey. Working paper (2006)
2. Zhang, J., Ghahramani, Z., Yang, Y.: A probabilistic model for online document clustering with application to novelty detection. In: Neural Information Processing Systems Conf. (2004)
3. Makkonen, J., Ahonen-Myka, H., Salmenkivi, M.: Applying semantic classes in event detection and tracking. In Sangal, R., Bendre, S.M., eds.: Proc. of Int. Conf. on Natural Language Process. (2002) 175–183
4. Das, S., Chen, M.: Yahoo! for amazon: Sentiment extraction from small talk on the web. Management Science **53**(9) (2007) 1375–1388
5. Koppel, M., Shtrimberg, I.: Good news or bad news? let the market decide. Computing Attitude and Affect in Text: Theory and Applications (2006) 297–301
6. Halaschek-Wiener, C., Hendler, J.: Toward expressive syndication on the web. In: WWW '07: Proc. of the 16th Int. Conf. on World Wide Web, ACM (2007) 727–736
7. Pantel, P., Pennacchiotti, M.: Espresso: leveraging generic patterns for automatically harvesting semantic relations. In: ACL '06: Proceedings of the 21st Int. Conf. on Computational Linguistics and the 44th annual meeting of the ACL. (2006) 113–120
8. Blohm, S., Cimiano, P.: Using the web to reduce data sparseness in pattern-based information extraction. In: Proc. of the 11th European Conf. on Principles and Practice of Knowledge Discovery in Databases, Springer (2007) pp. 18–29
9. MacKinlay, A.C.: Event studies in economics and finance. J. of Economic Literature **35**(1) (March 1997) pp. 13–39

Access rights and collaborative ontology integration for reuse across security domains *

Martin Knechtel

SAP AG, SAP Research CEC Dresden
Chemnitzer Str. 48, 01187 Dresden, Germany
martin.knechtel@sap.com

1 Research Problem

This section gives a description of the overall research problem tackled in context of the Ph.D. and its relevance to the Semantic Web area.

The problem domain for this extended abstract is a collaborative marketplace in the Semantic Web. In the planned pilot 2 of the application scenario *PROCESSUS* of the research program *THESEUS* [1], described products to be sold are Web services. They are traded like goods and described in documents.

Ontologies can be used to define a shared vocabulary with concepts, properties and axioms. By referencing this shared vocabulary in product descriptions, a conceptual navigation over heterogeneous resources is possible.

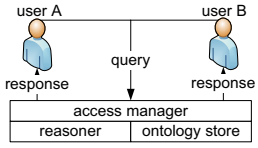

Fig. 1. Different responses for the same query, dependent on access rights

This induces that the ontology alone already contains insights about resources' contents. Different user roles are involved when accessing a semantic marketplace, e.g. visitor, customer, high volume customer, provider. Since all of them get different conditions and information detail about products, they get different answers for ontology queries when posing the same question (cf. Fig. 1). Access Control inside ontologies is one focus of the thesis.

A second focus is collaborative ontology integration. Given the functionality to have different views on a ontology for different user roles, one might also

* The project was funded by means of the German Federal Ministry of Economy and Technology under the promotional reference "01MQ07012". The author takes the responsibility for the contents.

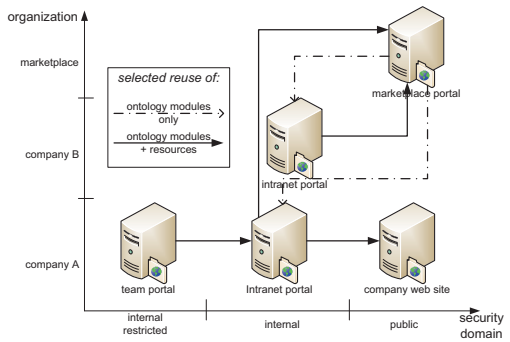

Fig. 2. Reuse of ontology subsets across security domains and different organizations

define a public view on the ontology which can be distributed for from a company internal server to a Web marketplace (cf. Fig. 2). In opposite direction, a company can also import the marketplace ontology, which might be based on a product standard like eCl@ss [2].

2 Related Work

This section discusses the state of the art in the fields affected by the given scenario.

Semantic content management is studied e.g. for semantic portals [3]. Also wikis can be used for semantic content management [4]. The contributions describe the motivation and implementation of content management with ontology support.

Collaborative Ontology Engineering seems well investigated. Examples are Ontolingua Server [5] and Collaborative Protégé [6]. Since the marketplace is a Web application, it is desirable to edit the ontology directly in the browser [7]. This makes no tool change for contribution and consumption necessary and changes can be tested directly in the application. This thesis' focus is how ontology contextualization can support collaboration.

To reuse parts of an ontology from a company internal context, ontology modularization is involved to decide if a module is complete [8]. In the other direction conservative extensions are extensions of an ontology without changing existing subsumption relations [9]. An interesting question for the thesis is how ontology modularization is influenced by assigned access rights. This is not investigated so far.

Fine grained access control inside ontologies is not well investigated in the research community yet. The contribution [10] presents basic access control meth-

ods and brings them in relation to ontologies. Although this work does not provide technical details, recommendation for authority based access control (ABAC) is given and justified. They propose that hierarchies can be used to inherit rights. According to [11], information about axioms of an ontology can be represented as context. This might be a starting point to represent access rights.

Authorization in other fields like file systems, content management systems, database management systems etc. is modeled by access control lists or by capabilities. Approaches often use hierarchies to inherit access rights. Due to the nature of ontologies, having no tree but a graph structure, access rights inheritance is of limited use. In the subsumption hierarchy a concept can be subconcept of several others, which leads to multiple inheritance. Object relations between concepts may form cycles. And it may be desired that a user can only see the superconcepts but not the subconcepts or the other way round. There is a simliar behaviour commonly used for FTP servers called *chroot jail*.

There are approaches for access rights inside ontologies. While [12] is based on a three-valued semantics and assumes an RDF tree without cyclic references, we want to use Description Logics and not restrict ontology structure to a tree. In [13] the focus is to restrict access on syntactically heterogeneous resources with help of a harmonizing ontology. A security policy is stored separately from the ontology, while we want to integrate it. An own ontology definition is used which is not conform to OWL-DL [14] since e.g. axioms and individuals are missing, while we want to use OWL-DL.

3 Contributions

This section describes how the proposed project will advance state of the art and summarizes expected contributions.

From the related work section it seems that context can store information about ontology axioms. The thesis will investigate if this context is suitable to store access rights and collaboration information to support ontology reuse. The following research questions will be subject of the thesis:

1. What is the right granularity for access control within an ontology: axiom, module, whole ontology, others?
2. How are axiom rights propagated to resources?
3. Can ontology axiom rights be derived from resource rights, to improve usability?
4. What effect has access control on reasoning and modularization?

The contribution of the thesis will be a framework to answer the conceptual questions, and an implementation to demonstrate the results. Therefore a conception and a syntactical representation of access rights will be developed. One candidate is to save context within an ontology with annotation properties according to [11]. The OWL1.1 standard will allow annotation properties for axioms and reference by axiom URI. This allows fine grained access control

similar to XML query languages. In the following example the URI is printed in brackets following the axiom.

$$DesignDocument \sqsubseteq Document\ [axiom1]$$
$$access(axiom1, companyInternal)$$
$$access(DesignDocument, companyInternal)$$

Argumentations for axioms and other ontology elements can be recorded analogously. In further processing steps the ontology can be stripped down to a version which only contains elements for public use and is therefore contextualized. But this naive syntactic process will not be enough since the remaining axioms may not make sense alone. The implications of access rights assignment concerning rights inheritance and ontology modularization will be investigated.

4 Evaluation

This section describes the methodology used to evaluate and validate results of the project.

In the above mentioned application scenario *PROCESSUS*, different user roles will get access to different parts and granularity level of the ontology. This offers an evaluation opportunity for the thesis' results.

Also collaborative ontology integration might be evaluated in the application scenario, since product descriptions on the marketplace have to be imported from somewhere. They might be interpreted as a subset of the company internal resources and ontology. It is a subset because, whitepapers and other marketing documents are intended to be made publicly available whereas design documents and test protocols which reference the same product are not.

5 Work Plan

This section sketches the different stages of the project and differentiates between current status, work in progress and planned future work.

Results achieved. The overall thesis work time is planned to be three years. Six months have passed so far. Currently the idea outline exists as presented in this abstract.

Current work. Current work is to investigate the two considered aspects of ontology reuse on behalf of an example case. Next planned step is to finish a paper in 2008-07 to present a first concept and a deeper related work analysis than given in this extended abstract.

Planned work. Further coming steps are the following. Until 2008-08 a first draft of the exposé is planned. Up to 2008-10 the structure of the manuscript and potential diploma thesis topics are formulated. Until 2009-10 the conceptual part of the thesis shall be finished, to have time for implementation until 2010-05. The thesis manuscript is planned to be finished in 2010-09.

References

1. THESEUS research program, "PROCESSUS - optimisation of business processes." available at `http://theseus-programm.de/scenarios/en/processus`, retrieved March 7, 2008.
2. M. Hepp and J. de Bruijn, "GenTax: A generic methodology for deriving OWL and RDF-S ontologies from hierarchical classifications, thesauri, and inconsistent taxonomies," in *ESWC' 07: Proceedings of the 4th European Semantic Web Conference*, pp. 129–144, 2007.
3. J. Hartmann and Y. Sure, "An infrastructure for scalable, reliable semantic portals," *IEEE Intelligent Systems*, vol. 19, pp. 58–65, 5 2004.
4. M. Krötzsch, D. Vrandecic, and M. Völkel, "Semantic MediaWiki," in *ISWC '06: Proceedings of the 5th International Semantic Web Conference*, (Athens, GA, USA), pp. 935–942, Springer, 11 2006.
5. A. Farquhar, R. Fikes, and J. Rice, "The Ontolingua server: a tool for collaborative ontology construction," *International Journal of Human-Computer Studies*, vol. 46, no. 6, 1997.
6. Stanford University, "Protégé 3.3.1 ontology editor." available at `http://protege.stanford.edu`, retrieved January 3, 2008.
7. A. V. Zhdanova, R. Krummenacher, J. Henke, and D. Fensel, "Community-driven ontology management: DERI case study," in *WI '05: Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, (Washington, DC, USA), pp. 73–79, IEEE Computer Society, 2005.
8. B. C. Grau, I. Horrocks, Y. Kazakov, and U. Sattler, "Just the right amount: extracting modules from ontologies," in *WWW '07: Proceedings of the 16th international conference on World Wide Web*, (New York, NY, USA), pp. 717–726, ACM, 2007.
9. S. Ghilardi, C. Lutz, and F. Wolter, "Did I damage my ontology? a case for conservative extensions in description logics," in *Proceedings of the Tenth International Conference on Principles of Knowledge Representation and Reasoning (KR'06)* (P. Doherty, J. Mylopoulos, and C. Welty, eds.), pp. 187–197, AAAI Press, 2006.
10. M. Dzbor, A. Kubias, L. Gridinoc, A. Lopez-Cima, and C. B. Aranda, "The role of access rights in ontology customization," Deliverable 4.4.1, NeOn Project, 2007.
11. G. Qi, P. Haase, and S. Pinto, "Context representation formalism," Deliverable 3.1.2, NeOn Project, 2007.
12. S. Kaushik, D. Wijesekera, and P. Ammann, "Policy-based dissemination of partial web-ontologies," in *SWS '05: Proceedings of the 2005 workshop on Secure web services*, (New York, NY, USA), pp. 43–52, ACM, 2005.
13. C. Farkas, A. Jain, D. Wijesekera, A. Singhal, and B. Thuraisingham, "Semantic-aware data protection in web services," in *IEEE Web Services Security Symposium (WSSS) 2006*, (Berkeley, California, USA), 5 2006.
14. S. Bechhofer, F. van Harmelen, J. Hendler, I. Horrocks, D. L. McGuinness, P. F. Patel-Schneider, and L. A. Stein, *OWL Web Ontology Language Reference*. World Wide Web Consortium, 2 2004. W3C Recommendation, available at `http://www.w3.org/TR/owl-ref/`, retrieved January 3, 2008.

Acquisition and Management of Semantic Web Service Descriptions

Maria Maleshkova

Institute of Applied Informatics and Formal Description Methods (AIFB),
University of Karlsruhe (TH), Germany
maleshkova@aifb.uni-karlsruhe.de

Abstract. The increasing importance and use of Web services have resulted in a number of efforts targeted at automating Web service discovery and composition based on semantic descriptions of their properties. However, the progress in the automation of Web service discovery is still held back by the fact that the description of Web services in terms of semantic metadata is still mainly manually. This Ph.D. thesis addresses this problem by developing an approach for the acquisition and management of semantic Web service descriptions in order to facilitate efficient service discovery and composition. Specifically, this involves the collection of information about a Web service, the acquisition of semantic descriptions based on the collected information, and the structured storage of the generated semantic descriptions.

1 Introduction

A clearly developing trend in the recent years is towards exposing the functionalities provided by existing software components in the form of services and facilitating in this way software reuse and added value through service composition. This development is strongly supported by the Web, which enables ubiquitous access via a set of Web standards and protocols to software components residing on different platforms. As a result, Web services are seen increasingly as the basic construct for the development of rapid, low-cost and easy-to-compose distributed applications in heterogeneous environments.

When it comes to finding appropriate services and composing distributed applications, current technologies still require a large amount of human interaction. This results in the restricted number of use cases for service integration and the limited scalability of solutions involving manual activities in the process of service discovery and compositions. In order to address these problems, the idea of supplementing Web services with a semantic description of their functionality, that could facilitate their discovery and integration, has been developed. However, the information, on which these semantic descriptions are based, is rarely discussed. In addition, there is still no established method for acquiring semantic Web service descriptions, which not only describe the functionality but also specify functional and non-functional properties in a way that supports automatic

matchmaking processes. Therefore, there is the strong need of developing an automatic approach, which enables the acquisition and management of semantic Web Service descriptions with the goal of supporting efficient service discovery and composition.

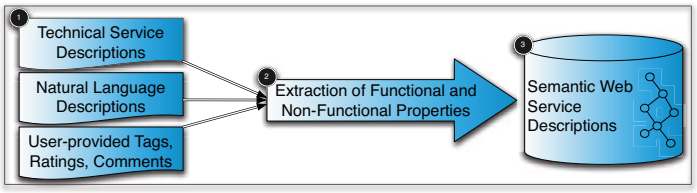

Fig. 1. Acquisition and Management of Semantic Web Service Descriptions

This Ph.D. thesis targets to explore the problem of semi-automatically acquiring and publishing semantic descriptions for Web services and aims to use Semantic Web techniques to develop a methodology to enhance the situation. Figure 1 shows the three main challenges connected with the acquisition and management of semantic Web Service descriptions.

1. **How can information about Web services be collected?** To answer this question, the relevant sources of information have to be identified, including not only technical and text descriptions, but also information inspired by Web 2.0 such as user tags, ratings and comments. After this, automatic mechanisms for data collection from the Web have to be developed.
2. **How can semantic Web service descriptions be automatically acquired?** This question is in the focus of the Ph.D. thesis and is addressed by exploring existing methods for metadata acquisition and identifying their advantages and disadvantages. The main goal is to develop a method, which semi-automatically acquires metadata by extracting functional and non-functional service properties in order to facilitate discovery and composition.
3. **How can semantic Web service descriptions be stored?** The acquired semantic descriptions have to be stored in a way that facilitates the manual as well as the automatic matchmaking processes. It is important to employ the metadata's syntactical or semantical interconnections in order to enable indexing and efficient service search.

2 Related Work

There is already some work done related to the automatic creation of semantic Web service descriptions. In particular, there are two main areas of research: the acquisition of a suitable Web service domain ontology and the actual process of

annotating Web services. Sabou et al. [1] present two ontology building process in the context of two concrete research projects, revealing some of the major aspects that make Web service ontology building difficult.

Focusing on the Web service annotation task, Patil et al. [2] apply graph similarity techniques to select a relevant domain ontology for a given WSDL file from a collection of ontologies. Hess and Kushmerick [3] employ Naive Bayes and SVM machine learning methods to classify WSDL files in manually defined task hierarchies. However, none of the developed approaches focuses on facilitating Web service discovery by specifying functional and non-functionally properties and in the same time taking into consideration temporal conditions on effects, trust and access control policies.

There are a number of state of the art Web service repositories including UDDI, Bindingpoint, .NET XML Web Services Repertory, WebserviceX.NET, Web Service List, Xmethods and SalCental. An overview of these repositories presented in [4] shows that in contrast to traditional software libraries, Web service repositories rely on little metadata to support service discovery, mainly because of the difficulty of automatically deriving metadata describing Web service collections. In order to overcome the lack of metadata, there are a number of approaches which aim to enhance existing Web service repositories, in particular UDDI, with complex semantic markup [5]. Still, the existing repositories do not target the structured storage of semantic Web service descriptions in order to facilitate the effective service discovery but rather use semantic information as an extension to stored descriptions.

3 Contributions

A number of Web service tasks can be automated by using semantic descriptions. In particular, service offers and requests can be matched automatically. However, semantic descriptions still have to be manually generated. The major goal of this Ph.D. thesis is to develop an semi-automatic method for the acquisition of semantic descriptions of Web services, based on metadata and tags extracted from the Web. The resulting semantic descriptions will be stored in an online semantic Web service descriptions repository, which can be used for both manual and automatic service discovery. The acquisition of semantic descriptions facilitates matchmaking approaches and can reduce the required human interaction to a minimum. Especially in use cases, where Web services are used as building blocks in distributed applications or where Web services provide functionality integrated in business processes, the automated deriving of semantic description plays a key role.

There are three main contributions, which are to be archived. First, the automatic collection of information describing a Web service is facilitated. Second, semantic descriptions of Web services are semi-automatically acquired from the collected information, in the form of semantic annotations. Finally, Web service semantic annotations are stored in an online repository, which provides semantic

indexing to allow for efficient queries and support matchmaking processes. Each of these contributions is described in detail in the following sections.

Web Service Crawlers are developed to support the automatic collection of information about Web services. Web service descriptions are usually given in a standardized form such as WSDL. Unfortunately, WSDL provides mainly technical description of services and this information is insufficient for automatic service discovery because it presumes increased human interaction. Therefore, it is necessary to automatically collect additional information. A Web service crawler can effectively be used to perform this task. In particular, this involves the identification of relevant sources and types of Web service information. Sources include natural language text descriptions, for example, documentation on provider Web-sites, and technical Web service descriptions such as WSDL, or WSMO variants. In addition, sources of information inspired by Web 2.0 such as user tags, ratings, and community inputs are also considered. The definition of the Web service crawlers is based on semantic attributes for service description, based on information needed for service discovery and composition. This involves mainly the specification of functional and non-functional Web service properties.

Web Service Semantic Annotations comprise the acquired semantic Web services descriptions, which are based on the data collected by the Web service crawlers. The main focus of this Ph.D. thesis is on developing a mechanism for the acquisition of semantic descriptions from the collected data, based on a determined formalism for the specification of semantic information. This mechanism is developed by using existing techniques for data mining and ontology learning and thereby, automatically constructing service annotations. In addition, the formalism, chosen as the basis for the semantic annotation, annotates functional and non-functional properties of Web services in a unifying way. The approach described in [6] will be used as a basis where semantic, temporal and security constraints are described by using a combination of the π-calculus and description logics and will be adjusted based on requirements resulting from the specification of the developed mechanism for acquisition of semantic descriptions.

Semantic Indexing is necessary in order to facilitate efficient queries on top of the semantic annotations. The structured storage of the acquired semantic annotations is targeted to improve performance of automatic matchmaking processes and speed up manual search. During semantic indexing, the collection of Web service semantic annotations will be analyzed and organized in a way that allows the efficient answering of expressive queries.

4 Evaluation

The planned evaluation consists of three main steps. First, the developed mechanism for data collection is evaluated by determining a fixed domain of source data and assessing how much data was correctly retrieved. This evaluation can easily be performed by using the well-established recall/precision metrics. Second, semantic annotations evaluation assesses the performance of extracting Web

service properties relevant for matchmaking from the corpus of collected data. Naturally, the quality of semantic annotation extraction has a direct influence on the quality of the service discovery and composition, which are based on these semantic annotations. This evaluation can be done by comparing the Web service functional and non-functional properties present in the collected data to the ones present in the acquired semantic descriptions. Third, the chosen structure for storage of Web service semantic annotations is evaluated by comparing the performance of matchmaking algorithms ran on the repository to the performance of these matchmaking algorithms ran on the same semantic annotations but in unstructured list collection. This evaluation will point out what improvements the developed repository structure brings in means of search performance. Finally, the best evaluation for the developed approach would be the making public of the semantic annotations repository so that other researchers can use it to test their discovery and composition algorithms.

5 Work Plan

The work plan is divided into three stages, each stage marking one year of the Ph.D. studies. **Initialization** State of the art report on formalisms for Web service semantic descriptions, in the context of service discovery (M6). State of the art report on data-mining techniques and ontology learning (M6). State of the art report on possible structures for semantic annotations storage (M6). Identified sources of Web service information (M12). Identified formalism for semantic annotations (M12). **Development** Definition of semantic attributes for the Web service crawlers. First Web service crawlers prototype (M18). Identified structure for annotations storage (M18). Identified data-mining and ontology learning approach (M24). First prototype implementation of acquisition of semantic annotations (M24). Evaluation of the first prototypes (M24). **Refinement** First prototype of semantic annotations repository (M30). Refined methodology based on the first prototypes (M30). Refined implementation (M36). Evaluation (M36).

References

1. Sabou, M., Wroe, C., Goble, C., Stuckenschmidt, H.: Learning Domain Ontologies for Semantic Web Service Descriptions. Journal of Web Semantics 3(4) (2005)
2. Patil, P., Oundhakar, S., Sheth, A., Verma, K.: METEOR-S Web service Annotation Framework. In Proceedings of the 13th World Wide Web Conference (2004)
3. Hess, A., N. Kushmerick, N.: Machine Learning for Annotating Semantic Web Services. In AAAI Spring Symposium on Semantic Web Services (March 2004)
4. Sabou M., Pan, J.: Toward Improving Web Service Repositories through Semantic Web Techniques. Workshop on Semantic Web Enabled Software Engineering (SWESE) at ISWC (November 2005)
5. Akkiraju, R., Goodwin, R., Doshi, P., Roeder, S.: A Method for Semantically Enhancing the Service Discovery Capabilities of UDDI. IIWeb 87-92 (2003)
6. Agarwal, S., Studer, R.: Automatic Matchmaking of Web Services. International Conference on Web Services (ICWS'06). IEEE Computer Society (September 2006)

Ontological Description of Image Content Using Regions Relationships

Zurina Muda

School of Electronics and Computer Science
University of Southampton, United Kingdom
{zm06r@ecs.soton.ac.uk}

Extended Abstract

Keywords: Spatial relationships, image annotation and ontology.

1 Research Problem And Aim

Rapid growth in the volume of multimedia information creates new challenges for information retrieval and sharing, and thus anticipates the emergence of the Semantic Web [2, 3]. The principal component in most of multimedia applications is the use of visual information and new approaches are essential to improve the inferring of semantic relationships from low-level features for semantic image annotation and retrieval. Much initial research on image annotation represents images in terms of colours, texture, blobs and regions, but pays little attention to the spatial relationships between regions or objects. Annotations are most frequently assigned at the global level [17] and even when assigned locally the extraction of relational descriptors is often neglected. However, current annotation system might recognise and identify a beach and an ocean in an image but fail to represent the fact that they are next to each other. Therefore, to enrich the semantic description of the visual information, it is important to capture such relations.

The aim of this research is an attempt to develop a new approach or technique for enhancing annotation systems, either through automatic or semi-automatic means, by capturing the spatial relationships between labelled regions or objects in images and incorporating such knowledge in a knowledge base such as an ontology. By this means, human users and software agents alike will be able to search, retrieve and analyze visual information in more powerful ways.

2 Related Work

Ontologies play an important role for knowledge intensive applications to enable content-based access, interoperability and communication across the Web. These ontologies become the backbone for enabling the Semantic Web [20]. The number of multimedia ontologies available is still rather small, and well-designed ontologies that

fulfill the requirements [5] of reusability, MPEG-7 compliance, extensibility, modularity and interoperability are rare [18]. The COMM ontology which is under development elsewhere and is based on DOLCE ontology as a foundational ontology is of particular relevance.

A pure combination of traditional text-based and content-based approaches is not sufficient for dealing with the problem of image retrieval on the Web, mostly because of the problem of its text based orientation. Some Web images have irrelevant, few or even no surrounding texts. Thus, the problem of limited collateral text for the annotation of images needs to be solved. Besides, manual image annotation is a tedious task and often it is difficult to make accurate annotations on images. There are many annotation tools available but human input is still needed to supervise the process. So, there should be a way to minimize the human input by making the annotation process semi or fully automatic. In the latter case, although there is much research on automatic image annotation, the results often do not really satisfy the retrieval requirements because of the flexibility and variety of user needs.

To date, many contend-based image retrieval research systems, frameworks and approaches have been reported. Li et. al[14] presented Integrated Region Matching, a similarity measure for region based image similarity comparison. Ko & Byun[8] used Hausdorff Distance to estimate spatial relationships between regions in their Integrated Finding Region In the Pictures (IFRIP) as extension to their previous FRIP [9]. Laaksonen et. al[10] proposed a context-adaptive analysis of image content, by using automatic image segmentation. Lee at. al[12] proposed a new domain-independent spatial similarity and annotation-based image retrieval system. Zhou et. al [21] proposed an approach for computing the orientation spatial similarity between two symbolic objects in an image. Wang [18] proposed a new spatial-relationship representation model called two dimension begin-end boundary string (2D Be-string), based on previous research in 2D String [11]. Ahmad & Grosky [2] proposed a symbolic image representation and indexing scheme to support retrieval of domain independent spatially similar images.

However, all the research in spatial relationships has been pursued independently without taking into consideration the problems of integrating them with an ontology. Such integration would be valuable in producing high level semantics by making semantic annotation systematically easier and more meaningful. In doing so, existing ontologies such as DOLCE and COMM will be evaluated to identify both their relevance and effectiveness in achieving the research aim.

3 Contributions And Evaluation

As part of a preliminary experiment, a comparative analysis of three existing annotation tools has been carried out: Caliph & Emir [15], AKTive Media [6], and M-OntoMat-Annotizer [16]. Each of these tools has been explored individually by using a group of images and a comparative study based on an evaluation framework adapted from Lewis[13] and Duineveld[7] has been performed and results obtained. The comparative study investigated image description features (including annotation) and

user interface components to find out the capabilities of existing image descriptions tools and to establish whether the spatial relationships are included and, if so, what the relationships might be. For image description components, follow-up with the developer of the tools has been established to ensure the reliability of the result.

The study shows that, each of the tools offered some special features compared to others and all tools were involved with manual annotations of the whole image. In addition M-OntoMat-Annotizer and AKTive Media allowed segmentation and annotation of the selected regions in images. Caliph & Emir and AKTive Media support some relations but not spatial relationships. Neither of these tools considered the specific locations of objects nor regions in the image for annotation or retrieval.

Based on the study and the previous research, currently, several existing annotation or description tools enable automatic segmentation by grouping multiple regions together and use manual annotation to annotate those regions. By adding the locator description where spatial relationships are considered, the knowledge of the image content becomes more specific and retrieval could be more efficient and performed in an explicit way.

This research will use existing automatic segmentation algorithm when available and manual combining of regions into composite regions for recognised objects. These will be manually annotated in the first instance together with spatial relationships between the objects. From there, an automatic annotation of spatial relationships among the objects in the image plane could be developed based on various available approaches by integrating directional and topological representation of spatial relationships. The process is simplified as illustrated in Fig.2.

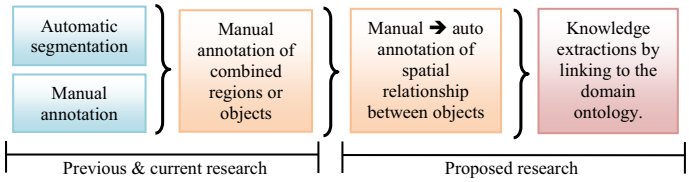

Fig.2. Research outline.

Therefore the expected contribution will be a new approach or technique to automate spatial relationships extraction between the composite regions or objects in images and linking the knowledge to an extended multimedia ontology. The approach or technique should be reliable in order to counter the uncertainty of matching images with the real world cases. For example, this is how it would works when given an image of a beach:

1. Existing tool would provide the annotation of regions of the image corresponding to: the beach, the ocean, the sky and the coconut tree objects are recognised.
2. Our approach then identify that: a. The coconut tree is within the beach; b. The beach is next to the ocean; c. The ocean is below the sky.

3. By reasoning over appropriate domain ontology, and exploiting the entailed spatial relationships, we would be able to infer that if the beach is in Hawaii, then the ocean must be the Pacific Ocean.

For the time being, the domain of the research would be a subset of everyday scenes such as city scenes or places of interest, but later other domains such as medical domain, may be considered to test the generality of the approach. Evaluation on ground truth with spatial relationships in term of precision and recall test will be made to see how well the automated extraction of spatial relationships has been achieved. The evaluation will use sufficient images such as Corel dataset to ensure statistical significance of the result obtained.

4 Work Plan

In order to accomplish the aim, the research plan is assigned into two levels – a macro plan using a Gantt chart for general activities and corresponding timelines, and micro plan using a K-chart [1] for the specific planning and execution of research. The research framework is illustrated in Fig. 3 and consists of:

1. Annotation component – automatically extracts and identifiers spatial relationships between multiple segmented regions or objects.
2. Ontological component – logics and reasoning of the extended existing multimedia ontology specifically in terms of spatial descriptors and locators.
3. Retrieval component – image retrieval mechanisms based on spatial relationships to evaluate the functionality and effectiveness of the approach.

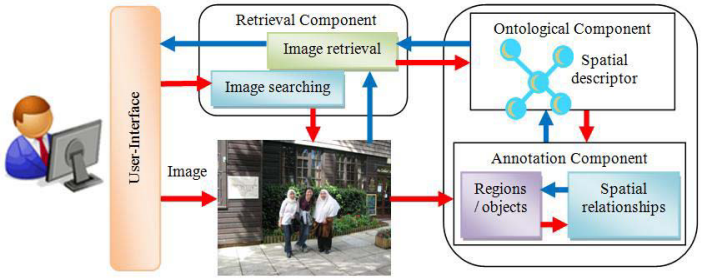

Fig.3. Research framework.

So far, the literature reviews and some preliminary experiment have been performed. However further practical works in the research and development phase is now being carried out. As a conclusion, it is hoped that this research will generate a constructive semantics approach in enabling the Semantic Web as well as bridging the Semantic Gap in image retrieval, while at the same time contributing new finding to human knowledge as a whole.

References

1. Abdullah, M. K., Mohd Suradi, N. R., Jamaluddin, N, Mokhtar, A.S., Abu Talib, A. R. & Zainuddin, M. F.: K-Chart: A Tool for Research Planning and Monitoring. J. of Quality Management And Analysis, vol 2(1), 123-130 (2006.)
2. Ahmad, I. & Grosky, W. I.: Indexing and Retrieval of Images by Spatial Constraints. J. of Visual Communication and Image Representation, vol. 14(3), Elsevier, 291-320, (2003)
3. Berners-Lee, T., Hendler, J. & Lassila, O.: The semantic web. Scientific American, (2001)
4. Berners-Lee, T., Fischetti, M. & Francisco, H.: Weaving the web: The original design and ultimate destiny of the World Wide Web by its Inventor (1999)
5. Bloehdorn, S. Petridis, K. Saathoff, C. Simou, N. Tzouvaras, V. Avrithis, Y. Handschuh, S. Kompatsiaris, I. Staab, S. & Strintzis, M.G.: Semantic Annotation of Images and Videos for Multimedia Analysis. In Proc. of the 2nd ESWC2005, (2005)
6. Chakravarthy, A., Ciravegna, F. & Lanfranchi, V.: Cross-media document annotation and enrichment. In: Proc of the 1st SAAW2006, (2006)
7. Duineveld, A. J., Stoter, R., Weiden, M. R., Kenepa, B. & Benjamins, V.R.: WonderTools? A comparative study of ontological engineering tools. In: Proc. of the 12th Workshop on Knowledge Acquisition, Modeling and Management. Alberta, Canada, October (1999)
8. Ko, B. & Byun, H.: Multiple Regions and Their Spatial Relationship-Based Image Retrieval. In: Proc.of the international CIVR2002. Lew, et al. (Eds). LNCS, Vol. 2383. Springer-Verlag, London, 81-90(2002)
9. Ko, B. C., Lee, H. S. & Byun, H.: Region-based Image Retrieval System Using Efficient Feature Description. In: Proc. of the 15th ICPR2000, vol. 4, 283-286. Spain, Sept., (2000)
10. Laaksonen, J., Koskela, M & Oja, E.: PicSOM-Self-organizing image retrieval with MPEG-7 content descriptions. IEEE Trans. on Neural Networks, 13(4), 841–853 (2002)
11. Lee, S.C., Hwang E.J & Lee, Y.K.: Using 3D Spatial Relationships for Image Retrieval by XML Annotation. ICCSA2004, LNCS 3046, 838–848 (2004)
12. Lee, S. & Hwang, E.: Spatial Similarity and Annotation-Based Image Retrieval System. IEEE 4th Inter. Sym. on Multimedia Software Engineering, Newport Beach, CA (2002)
13. Lewis, J.R.: IBM Computer Usability Satisfaction Questionnaires: Psychometric Evaluation and Instruction for Use. Inter. J. of HCI, 7(1), 57-78 (1995)
14. Li, J., Wang, J. Z., Wiederhold G.: IRM: Integrated Region Matching for Image Retrieval. ACM Multimedia, pp. 147-156, (2002)
15. Lux, M. Becker, J. & Krottmaier, H. Calph & Emir: Semantic Annotation and Retrieval in Personal Digital Photo Libraried. In: Proc. of 15th CAiSE'03. pp. 85-89, Austria (2003)
16. Saathoff, C., Petridis, K., Anastasopoulos, D., Timmermann, N., Kompatsiaris, I. & Staab, S.: M-OntoMat-Annotizer: Linking Ontologies with Multimedia Low-Level Features for Automatic Image Annotation. In: Posters of the 3rd ESWC 2006, Montenegro, (2006)
17. Srikanth, M., Varner, J., Bowden, M. & Moldovan, D.: Exploiting Ontologies for Automatic Image Annotation. In: Proc. of the 28th Annual Inter. ACM SIGIR Conference on Research and Development in Information Retrieval, Salvador, Brazil, 552-558 (2005)
18. Staab, S.: Multimedia Ontology. Summer School in Multimedia Semantics (SSMS2007), Glasgow, (2007)
19. Wang, Y. H.: Image Indexing and Similarity Retrieval Based on Spatial Relationship Model. Inf. Sci. Comput. Sci. 154, 1-2, Elsevier, New York, USA, pp. 39-58, Aug. (2003)
20. Ying D.: Ontology: The enabler for the Semantic Web, http://citeseer.ist.psu.edu/601004.html (2002)
21. Zhou X.M., Ang C. H. & Ling T. W.: Image Retrieval based on object's orientation spatial relationship. Pattern Recognition Letters 22. Elsevier Science, 469-477 (2001)

Semantic Web-based Group Formation for E-learning

Asma Ounnas

School of Electronics and Computer Science
University of Southampton, UK
ao05r@ecs.soton.ac.uk

1. The Research Problem

Collaboration has long been considered an effective approach to learning. However, forming optimal groups can be a time consuming and complex task. Teachers often need to set some constraints for the grouping based on the aim of the collaborative task. To achieve optimal grouping, the formation needs to satisfy these constraints, even when the list of students is unknown. In this research, we investigate the use of Semantic Web technologies to assist teachers in overcoming this problem. In particular we investigate the following problems with forming groups:

- Describing the students: how do we describe the students in a way that is meaningful to the group formation.
- Specifying and satisfying the constraints: how do we model the constraints for the different collaboration goals, and how can we form optimal groups by satisfying the given constraints.
- Avoiding *"Orphan Students"*: when assigning students to groups, some students remain unassigned to any groups, often because some constraint has been violated. This problem is known as the orphan students.
- Handling the formation with incomplete data: If the students do not provide the relevant data, how do we ensure that the formation is still efficient.

We know that the Semantic Web aims at providing a promising foundation for enriching resources with well defined meanings and inferring new data from existing one. Therefore, we study the use of Semantic Web technologies and mainly ontologies and deduction rules in describing students and handling incomplete data in constraint-based group formation. The challenges of the research reside in applying the potential of the Semantic Web is real life applications such as forming groups for learning; especially with the growing use of collaboration applications over the Web.

2. Related Work

Group formation is a well-known problem in various disciplines including Psychology, Philosophy, social studies, Economics, and Education. In learning, different applications have been developed to automate the process of allocating participants to groups as tool for Computer Supported Collaborative Work. Most of the existing applications follow a self-selecting formation approach, where the learner

selects the potential learners that can assist him or her in achieving the learning goal, and the selected learners get to accept or reject joining the group [1], [2]. These systems usually model the learners' context, experience, and previous performance in the subject of the collaboration. Similar systems employ agents to negotiate the participation in the groups [3], which facilitates the dynamic formation of the groups (coalition formation) through agents' communication and decision-making [4]. In terms of constraint-based formation, we argue that:

- Existing systems only model a fixed set of constraints.
- Most of these systems are based on self-selecting group formation [1], [3] which is not the most efficient approach in forming teams, as it does not ensure balanced grouping, and usually end with some students being unassigned to any group (orphan students).
- Most systems use *Opportunistic Group Formation* concepts where the system initiates the collaboration and sets up a learning goal for the learner [2], [1], [4], [3]. OGF ensures the satisfaction of the participants in the group through negotiation, but does not discuss the efficiency of the negotiation if all students are grouped simultaneously. Furthermore, OGF is usually more beneficial in short-term groups.
- In [5] and [6], the authors introduced tools that assign all the students in the class to groups simultaneously. However, although these applications only model a limited number of constraints, their evaluation showed that manual corrections to the results were needed due to the appearance of orphan students in the generated formation.

3. Expected Contribution

The theoretical contribution of this research is a study of the feasibility and usefulness of employing Semantic Web technologies for group formation within a complex domain such as learning, and particularly in handling incomplete data. To overcome the complexity of allocating students to groups, we provide a framework to assist the teacher in forming groups based on their chosen set of constraints. The framework handles the group formation process based on the following concepts:

Modeling the students' features: We model a large range of features that can be considered for different group formations using a number of domain ontologies, which can form a reliable dynamic learner profile [7]. In this context, semantic modeling provides meaningful descriptions of the students and the relationships between them. Examples of the modeled features are: personal details, course details, interests, team roles, preferences, friends, collaborators, trust ranking, and so on [7]. The ontologies we use are based on *Friend Of A Friend* (FOAF), an existing ontology that describes people for building social groupings. This allows us to identify the relations between the participants in order to form groups from social networks. This feature allows the teacher to be aware of the social connections between the students and therefore controlling the group dynamics, or detecting plagiarism.

Negotiating the group formation: We model the students' allocation problem as a *Constraint Satisfaction Problem (CSP)* [8] with strong and weak constraints. The negotiation process can be then handled by a constraint satisfaction solver. We emphasize that, in this research, we are not concerned with proving that any particular

set of constraints leads to better results in terms of the performance of the groups; neither do we claim that any particular algorithm leads to best grouping.

Handling Incomplete Data: The semantic representation of students' data, to which the instructor constraints can be mapped to, allows inferences to generate more data. We use domain ontologies and deductions rules for substituting missing data with data mined from the Web. For example, if the information about whether student John is a leader or not is missing, and we know from John's web page that he is a captain of the football team, then we can infer that John is a leader; or if we are grouping student by skills, and we don't know Sarah's skills, but we know that Sarah has a high grade in discrete mathematics and Sarah has a high grade in Logic then we can infer that Sarah will perform well in formal methods.

Calculating the group formation quality: To evaluate the generated group formation, we provide a metrics framework for calculating its quality in terms of the satisfied constraints [9], and hence the collaboration goals set by the teacher. Using these metrics, the teacher can check the confidence of the group formation framework in generating the groups.

In general, although it is applied to learning, this research can be employed in other domains as a solution to any type of constrained group formation. When completed, the system will form a standard semantic technology that allows groups of users to be generated based on a set of constraints and a range of information about themselves.

Research Methodology

In the early stage of the research, we run an observational study to analyze the different constraints teachers consider when forming groups. We studied the possible students' features that can be relevant to forming different types of groups by investigating the available literature on collaborative learning theories [7], and asking teachers what constraints they employ for different educational goals. We then gave a class of 66 undergraduate students some questionnaires to monitor the data they provide for the grouping and their satisfaction with the groups at the end of course. The study enabled us to realize the depth of the problem and the pedagogical issues that accompany it. We then modeled the group formation problem as a constraint satisfaction problem to be implemented as the semantic (group formation) framework [9], [10]. We also reviewed the different techniques for evaluating group formation, and provided a model for the formation quality metrics framework in terms of the formation goals and hence the constraints satisfaction [9]. For the next step, we started implementing the group formation framework [10] based on the following:

The Student Interface: The student can enter their data through a web-based form composed of the student's personal data, a list of their friends, their interests and preferences, and information about their course such as the modules they are taking.

The Ontology: We created an ontology called *Semantic Learner Profile* (SLP) that extends FOAF with a description of a large range of student's personal, social, and academic data such as learning styles and collaborators. We also use the trust ontology (http://trust.mindswap.org/ont/trust.owl) to allow the students to rank their trust towards each other in specific topics. Following the vision of James Hendler in

reusing and sharing small ontological components instead of large complex ontologies [11], we intend to enrich our learner profile with more features by employing other domain ontologies (competency and interest topics ontologies). Once the student submits the profile data through the interface, an RDF file is created (FOAF+SLP), and processed using Jena, a Semantic Web inference engine.

The Instructor Interface: Through this web-based interface, the instructor is given a degree of freedom in selecting the constraints they care about for the formation they are initiating. They are provided with an option that enables them to set a priority value for each constraint. Ranking the importance of the constraints enables the application to manage compromises based on these priorities. The constraints are then written as a constraint satisfaction problem in the group generator.

The Group Generator: The generator is based on a DLV solver, an implementation of disjunctive logic programming, used for knowledge representation and reasoning. DLV's native language is *Disjunctive Datalog* extended with constraints, true negation and queries [12]. DLV performs a simple forward checking algorithm [8] on the data provided to process the groups. The use of strong and prioritized weak constraints in DLV enables the framework to always generate a solution with all students allocated even if some of the weak constraints are violated [13]. This avoids the orphan students' problem. The solver returns the optimal group formation that minimizes the number of violated constraints and returns the list of the violated constraints, which can be used in calculating the group formation quality.

For our future work, we plan to add a module to the architecture of the framework that mines data from web pages and connect it to the ontology and a set of deduction rules to infer the missing data from the knowledge base. In this case, if the needed data is incomplete, the system will substitute the necessary data and subsequently feed it to the solver. So far, we evaluated the framework with two classes of undergraduate students. However, since the teachers had a maximum of three constraints, the framework returned a best model in both cases with violation of one constraint for one group in both courses. Future evaluation of the framework will include running it with different scenarios on simulated classes of students. The simulated data will be based on the population statistics collected from our observational study. The framework will be tested with various constraints, different in content and number. Since groups can also be generated from social networks, a range of constraints will be based on the social connections between the learners. We intend to use the metrics framework we introduced in [9] to record the formation quality for the evaluation. Once the framework is refined with deduction rules, the evaluation of its performance with incomplete data will be compared to its performance with complete data (and no deduction rules), and its performance with incomplete data (and no deduction rules). For handling incomplete data, due to privacy issues, we aim at using students' web pages from the University of Southampton as a base for our mining.

Conclusion and Future Work

In this research, we propose an approach to learner group formation, based upon satisfying the constraints of the person forming the groups by reasoning over possibly

incomplete semantic data about the potential participants. We are currently evaluating the semantic (group formation) framework with complete data. Within the next few months, we intend to implement extensions to allow for handling incomplete data, and for forming groups from social networks. The research can then be fully evaluated and results published to the community with more results in more depth.

We believe that by reasoning on learners' profiles and the teacher's constraints, we can achieve a powerful foundation for automated group formation. The use of Semantic Web technology demonstrates the powerful characteristics of the Semantic Web that can be put in use to facilitate daily life tasks such as allocating students to groups for collaborative learning. The use of the Semantic Web in this domain can be extended to other areas of group formation such as forming teams within organizations, sports, or even military. The research can also be extended to other constraint satisfaction problems where data is key to the solution of the problem. The interoperability of the Semantic Web facilitates the use of such an application in different platforms and systems, even when the participants are geographically distributed. For this, the only challenge to applying this research in other areas is the development of domain ontologies and deduction rules.

References

1. Hoppe, H.U. Use of multiple student modeling to parametrize group learning, In proc of AI-ED 95, AACE, Charlottesville, VA, USA, 1995, pp. 234-241.
2. Wessner, M. and H-R. Pfister, Group formation in computer-supported collaborative learning, Proc. of ACM SIGGROUP, Boulder, Colorado, USA, 2001, pp. 24 - 31.
3. Inaba, A., T. Supnithi, M. Ikeda, R. Mizoguchi, and J.I. Toyoda. How Can We Form Effective Collaborative Learning Groups? In ITS, Montréal, Canada, 2000, pp. 282-291.
4. Soh, L.K., N. Khandaker, X. Liu, and H. Jiang. A Computer-Supported Cooperative Learning System with Multiagent Intelligence. In AAMAS'06, Japan, 2006, pp. 1556-1563.
5. Redmond, M.A., A computer program to aid assignment of student project groups, In proc of ACM SIGCSE, Charlotte, NC, USA, 2001, pp. 134-138.
6. Tobar, C.M., de Freitas, R.L. A support tool for student group definition. In Proceedings of the 37th ASEE/IEEE Frontiers in Education Conference. 2007.
7. Ounnas, A., Davis, H. C. and Millard, D. E. Semantic Modeling for Group Formation. In: PING Workshop at the 11th UM2007 conference, Corfu, Greece, 2007
8. Kumar, V. Algorithms for Constraint Satisfaction Problems: A Survey. In AI Magazine, Vol 13. Issue 1, 1992, pp. 32-44.
9. Ounnas, A., Millard, D. and Davis, H. A Metrics Framework for Evaluating Group Formation. In Proc of ACM Group'07, Sanibel Island, Florida, USA, 2007, pp 221-224.
10. Ounnas, A., Davis, H. C. and Millard, D. E. A Framework for Semantic Group Formation. To appear in the 8th IEEE ICALT'08, Santander, Cantabria, Spain. 2008.
11. Hendler, J., Agents and the Semantic Web. IEEE Intelligent Systems, 2001 pp.30-37.
12. Leone, N., Pfeifer, G., Faber, W., eiter, T., Gottlob, G., Perr, S., Scarcello, F. The DLV system for knowledge representation and reasoning. In ACM Transactions on Computational Logic (TOCL), Vol 7, Issue 3, 2006, pp. 499 - 562.
13. Buccafurri, F., Leone, N., Rullo, P. Strong and weak constraints in disjunctive Datalog. In Proc of the 4th International Conference on Logic Programming and Nonmonotonic Reasoning. In LNCS, Vol. 1265. 1997, pp. 2 -17.

Identifying Individuals using Identity Features and Social Information

Matthew Rowe
Web Intelligence Technologies Lab
Department of Computer Science
University of Sheffield, UK
m.rowe@dcs.shef.ac.uk

Abstract: This paper presents an approach for the disambiguation of individuals using the semantics of identity and social circles. Identity information is extracted and integrated to provide a presentation of existing identity information currently on the web relating to a given individual. Communities are used to discover identity resources, share them socially, and critique the resources based on the accuracy and volume of their content. As motivation for this research issues concerning identity theft, online fraud and cyber stalking are considered, where the growth of the social web has contributed to the rise in such practices. Monitoring identity on the web would go some way to addressing these issues.

Keywords: community, disambiguation, identity, integration, semantic web, social web

1. Research Problem Overview

The motivations behind my research have been the growth of the social web over the past 2 years, and the rise in online identity theft and cyber stalking [1]. The work presented in this extended abstract provides an approach that identifies, extracts and integrates occurrences of identity information from the web. The approach is split into three parts: Extracting identity information and social network mining, integrating identity information, and resource discovery.

The semantics of identity are used to perform the extraction process by recognising identity features within a web resource and extracting the content relating to these features. Social networks are mined and pruned to derive the social circle that an individual belongs to using social content such as socially tagged images and conversation data. The social circle is then used to recognise identity information on the web, by parsing text content from web pages to derive the names of individuals, and comparing them against the social circle.

Disambiguating between information describing different individuals using features of identity and the pruned social circles is used to aid with the integration of identity information. Social circles are used to provide a useful technique to disambiguate individuals by their acquaintances. Resource discovery is supported using a community of users by sharing resources containing identity information; the

community is responsible for rating, and critiquing the resources, and discovering more identity resources upon which they are shared with the community.

Semantic technologies provide useful techniques and methods to identify individuals. An individual's identity will be formalised to encapsulate uniquely identifiable properties, reasoning is performed to derive additional information relating to the individual. The extraction of information will be carried out using a populated ontology containing personal details belonging to an individual. Identity information from various resources will be integrated together, and disambiguated according to their semantics. Involving a community to aid with extraction, by sharing vital resources, will also use social technologies. Social feedback methods will also be used for feature selection by allowing an individual to select the properties of their identity they believe to be the most prevalent.

The work presented within this abstract employs a both a combination of existing methodologies such as social networking mining, and original techniques for disambiguation of individuals. The motivation of this work places emphasis on monitoring online information, and providing risk assessments to concerned users, those who wish to discover what information exists relating to them. This extended abstract is structured as follows: Section 2 discusses the state of the art divided into the three previously mentioned areas. Section 3 is similarly divided into three areas outlining the work plan by explaining the various investigations being conducted. Section 4 explains the evaluation methods to be used, and section 5 presents concluding remarks.

2. Related Work and Contributions

Extracting Identity Information and Social Network Mining

The state of art on information extraction distributed throughout various textual sources includes standard information extraction mechanisms and approaches that can be applied to identity data such as classic wrapper induction [6] for information extraction from structure sources, and more up to date approaches such as support vector machines [4] for the extraction of information from free text. My work has focused on the semantics of identity, what properties are more prevalent than others, and how the community can influence the prevalence of identity features when extracting identity information.

The state of the art within the area of social network mining commonly uses techniques such as entity co-occurrence [8], [3] for extracting the strengths and ties among individuals. A seed set of entities is produced that models the names of individuals that commonly co-occur together in the same context. State of the art work presented in [9] also demonstrates how social networks can be mined from Semantic description files via FOAFnet. Advancement on previous work is demonstrated in [10] and [5] where relations between individuals within the same social network are not only identified but are also given labels denoting the tie between them. My research will investigate the inference of relationship strengths that bind relationships. Using such methods I am investigating how social cliques and circles play an important role in identifying individuals, similar to real life identification through acquaintances.

Integrating Identity Information
Social networks from two separate sources are integrated together in [11], enabling the integration of identity information. Disambiguation is performed using a context sensitive algorithm, considering the properties and relations surrounding the entities in question. Utilising both community selected prevalent identity features and social circles, my work will contribute to the state of the art by offering a social approach to the feature selection problem and disambiguating objects using social bonds. By incorporating a user within the disambiguation process bootstrapping is performed by allowing the user to select the features of their identity that they believe provide their most unique features.

Resource Discovery
Work in [7] describes a framework to allow users to share information within a community portal by adding and removing metadata from already existing information. My work contributes to the state of the art by sharing resources containing identity information, and supervising the process of information extraction by allowing individuals to select their prevalent identity features. User based feedback is used enabling individuals to rate resources based on their usability and accuracy of retrieved content.

3. Work Plan

Extracting Identity Information and Social Network Mining
Regarding the discovery of identity information I have focussed on the semantics of identity. I am currently defining a manually created ontology encapsulating the properties of identity, and able to capture an individual's identity properties. Future work includes the designing of a methodology to efficiently discover identity information from the wider web. In order to extract identity information I have investigated the use of support vector models for community supported identity extraction from semi-structured web resources. Following work will investigate focussed crawling using specialised web queries, indexing a subset of the web, and community supported blocking mechanisms.

To mine social networks I have created several mechanisms to extract social network data from social networking sites that will be used to seed sets for a wider mining process. A working prototype of this approach is available for use[1]. The next stage of work will investigate the pruning of social network data to derive social cliques and circles, and the investigation of the effects and application of identity discovery through the use of social cliques and clusters.

Integrating Identity information
The work I have done to date has investigated the integration of object data from heterogeneous web resources, and the disambiguation of objects. The disambiguation of objects can then be applied to identity information. Further work will investigate

[1] http://apps.facebook.com/socialcircular

the use of pairwise decision models, and community supervision of integration where I anticipate that the use of feature selection using decision models will be an important aspect of disambiguating identity information.

Resource Discovery
To date I have researched approaches to share resources through social bookmarking tools and similar web applications. Future work will investigate the adoption of collective intelligence approaches when sharing identity resources, and the use of feedback mechanisms to rate and prioritise identity resources.

4. Evaluation

Extracting Identity Information and Social Network Mining
Identity extraction will be evaluated for precision, recall and error rate. Evaluators of the approach will be required to find all occurrences of their identity manually to create a gold standard, detailing what identity details are present. Extracted identity information will then be evaluated against the gold standard. Precision and error rate will be used to evaluate extracted social networks through comparison against real life social networks for each individual performing the evaluation. Evaluators will also be required to validate relationship strengths within their social circle, and the members of their social circle.

Integrating Identity Information
Evaluation will be performed using exhaustive user testing to derive the precision, recall and error rate. Each individual will verify the all information items relating or not relating to them, and the integrated information to identify incorrectly integrated information and incorrectly excluded data.

Resource Discovery
The evaluation of the sharing mechanism will be performed using social studies of users when using the approach, testing for user satisfaction through questionnaires. The approach should provide a useful means for sharing identity resources through an easy to use, yet effective methodology.

5. Conclusions

This paper presents an overview of the research that I am currently conducting. The research when broken down into the three areas can be summarised further to include information extraction, information integration and sharing mechanisms. The first areas being largely concerned with existing semantic web technologies and their adaptation to these areas of work. The third area is largely centred around the social web, and current sharing mechanisms being employed by social web sites and services. A combination of both semantic web and social web technologies would incorporate the user at a more intrinsic level by supervising the information extraction and integration stages.

The state of the art will be contributed to mainly in the area of social network mining, and the use of the derived social circles to disambiguate individuals. The work that I have carried out so far has been largely concerning research within each separate area of work, and I have reached a position to begin implementation of an approach to disambiguate individuals using social circles. Information retrieval metrics have been chosen to evaluate the extraction and integration of information because of their widespread usage in similar applications. Evaluating for user satisfaction was selected when evaluating the discovery of resources in order analyse the effectiveness of the sharing mechanism.

In relation to the addressing of issues such as identity theft, online fraud, and cyber stalking, the presented approach provides a methodology to monitor the occurrence of identity information, and using semantic technologies reasoning can be performed to asses the risk of an individual being a victim of such practices. The approach must have sufficient flexibility to allow assessments to be made based on alternative requirements, such as different identity features. Disambiguating identity information performs a crucial role when assessing the risk of identity theft, any information that is wrongly classified could contribute to providing a false analysis.

References

1. Atkinson. S., Jagodzinski. P., Johnson. C., Phippen. A. D.: Personal Privacy: Exploitation or Control through Technology. Proceedings of the Sixth International Network Conference (INC2006), Plymouth, UK. 11-14 July, pp. 269-276 (2006).
3. Hamasaki. M., Matsuo. Y., Ishida. K., Nakamura. Y., Nishimura. T., Takeda. H.: Community Focused Social Network Extraction. Proceedings of 2006 Asian Semantic Web Conference (2006).
4. Huang. T-M., Kecman. V., Kopriva. I.: Kernel Based Algorithms for Mining Huge Data Sets, Supervised, Semi-supervised, and Unsupervised Learning, pp. 260. Springer-Verlag, Berlin, Heidelberg (2006).
5. Jin. Y., Matsuo. Y., Ishizuka. M.: Extracting Social Networks among Various Entities on Web. The Semantic Web. pp. 487-500. International Semantic Web Conference (2006).
6. Kushmerick. N., Weld. D., Doorenbos. R.: Wrapper induction for information extraction, IJCAI-97 (1997).
7. Maneewatthana. T., Wills. G., Hall. W.: Adaptive Personal Information Environment based on the Semantic Web. In: HT 2005 - ACM Workshop on Hypertext and Hypermedia, 6-9 September. Salzburg, Austria (2005).
8. Matsuo. Y., Hamasaki. M., Nakamura. Y.: Spinning Multiple Social Networks for the Semantic Web. Proceedings of the 2006 Asian Artificial Intelligence Conference. (2006).
9. Mika. P.: Bootstrapping the FOAF-Web: An Experiment in Social Network Mining. 1st Workshop on Friend of a Friend, Social Networking and the Semantic Web, Galway, Ireland (2004).
10. Mori. J., Tsujishita. T., Matsuo. Y., Ishizuka. M.: Extracting Relations in Social Networks from the Web Using Similarity Between Collective Contexts. Proceedings of ISWC 2006 (2006).
11. Aleman-Meza. B., Nagarajan. M., Ding. L., Sheth. A., Arpinar. B., Joshi. A., Finin. T.: Scalable semantic analytics on social networks for addressing the problem of conflict of interest detection. ACM Transactions on the Web Journal. Vol. 2. (2008).

An Approach to Evaluate and Enhance the Retrieval of Web Services Based on Semantic Information

Stefan Schulte

Multimedia Communications Lab (KOM)
Technische Universität Darmstadt, Germany
schulte@kom.tu-darmstadt.de
Phone: +49-6151-166187
Fax: +49-6151-166152

1 Research Problem

Web services have the potential to be composed to cross-organizational workflows. Due to their loose coupling, Web services provided by internal and external parties can be integrated into workflows at runtime. This vision aims for dynamic ad hoc collaborations between different business partners and entities.

In order to achieve the (semi-) automatic composition of Web services to business processes and workflows, it is necessary to identify the appropriate services. Unfortunately, a syntactic description of a Web service's capabilities is sufficient only if all potential parties (i.e., service providers, service brokers, and service requesters) use the exact same vocabulary. However, this is quite unlikely even in a corporate environment. Therefore, it is necessary to enrich Web service descriptions with semantic annotations and use them in the discovery process.

Even though the retrieval of Web services based on semantic information has already been investigated in several approaches, differences in Web service standards and the repositories used for the evaluation of these approaches has led to both a lack of in-depth evaluations and comparability of the proposals. Until now, surprisingly little effort has been put into the measurement of semantic Web service (SWS) retrieval performance.

Nevertheless, in order to identify the "best of breed"-approach to SWS retrieval it is necessary to have the means to compare the different approaches. Subsequently, it is possible to enhance or combine current techniques in order to improve retrieval results. The different methods for SWS retrieval should be evaluated at least regarding computation time and measures to identify the quality of results, i.e., precision and recall.

2 Related Work

SWS retrieval is based on a matchmaking engine, i.e., an algorithm that finds the best fitting Web services for a precise service request. There are no limitation

regarding how this algorithm is actually implemented, the form of the request, the number and the sequence of "best fitting" services, or which service feature is retrieved. Several authors have proposed different kinds of matchmaking based on the degree of conformity between the requests and Web service descriptions. In most cases, service requests are expressed as Web service descriptions that perfectly meet the requests; a query in terms of keywords or the ability to browse a service repository are not provided. Hence, it is necessary to identify the inputs and outputs of the Web service which fits perfectly, thereby making it more difficult for uninformed users to find appropriate services.

An obvious approach to SWS matchmaking has been proposed, i.e., by [8] and [5] with the matching of capabilities: A service is deemed to be of use for a requester if all outputs requested are matched by the outputs advertised and if all inputs needed by the service advertised can be covered by the inputs provided by the requester. Matches between inputs/outputs requested and advertised are categorized into *exact*, *plug in*, *subsumes*, and *fail* matches [5]. Thus, it is possible to arrange services by the degree to which they match the inputs/outputs requested.

The four categories mentioned may also be employed to measure the degree to which an advertised Web service can meet a request. A detailed implementation of this approach is presented in [4]. The authors enhance the four mentioned categories by *intersection*. However, it is not possible to assess, for example, which of two *plug-ins* better meets a request. Xu et al. propose the use of semantic distances between concepts in an ontology which extends this categorization and introduces a feasibility to rank Web service [9]. Klusch et al. enhance the frequently applied logic-based approaches with content-based information retrieval [1].

Regarding the evaluation of approaches to SWS matchmaking, most researchers fall back on their own Web service data sets. Consequently, this constrains the ability to compare evaluation results. To overcome this issue, the research community has come up with different contests in which researchers can bring in their approaches for such evaluations.

The *S3 Matchmaker Contest* adopts the OWLS-TC2 test data set [2]. As its name implies, this constrains the deployment of Non-OWL-S algorithms. Even though OWLS-TC2 can be regarded as a state-of-the-art test data set at the moment, it lacks real world examples and the semantic richness of Web services [3]. It is planned to include WSDL-S/SAWSDL and WSML test data sets and approaches in the next executions of this contest, but even then it would "only" be possible to compare one SAWSDL-based approach to another SAWSDL-based approach etc.

The *Semantic Web Service Challenge* is currently the most established contest and has been carried out several times since 2006. Its aim is to develop a test bed for different matchmaking frameworks. Hence, this contest is independent of Web service standards. All services are only specified by natural language descriptions and hence must be adopted to the matchmaking approach at hand [6].

3 Expected Contributions

The contributions of my thesis include both an evaluation workbench that covers the issues regarding evaluation approaches currently used and the evaluation of a heuristic-based algorithm for SWS matchmaking.

The approaches presented to SWS matchmaking evaluation lack at least *one* of the following issues (a detailed discussion of the pros and cons of these contests is presented in [3]):

- Lack of real world example Web services
- A too small set of Web services
- Some meaningful evaluation criteria are not examined
- Limitation to one Web service standard
- Results are often not published in detail, i.e., the actual retrieval results per query etc. are missing in the concerned publications
- Degree of matching is only of subordinated importance

While it is very difficult to address the first two issues without contributions from a large community, it is possible to counter the remaining problems. Hence, the implementation of the workbench in my thesis is based upon the following principles:

- Provision of a Web-based workbench for SWS matchmaking algorithms which can be used by the research community.
- Requests may be expressed in different Web service languages.
- Answer sets are not constrained by the language of existing Web services.

It must be noted that it should not and cannot be the aim of the proposed workbench to replace the well-established contests mentioned above. Quite on the contrary, the goal is to provide researchers with another possibility to evaluate SWS matchmaking algorithms, especially the approaches to SWS matchmaking in our working group.

Current approaches to SWS matchmaking which do not take their performance into account are hardly feasible in dynamic real-time scenarios due to the large number of potential Web services involved. This is especially complicated if a workflow has to be replanned at runtime. In such a case the computation time of a composition becomes crucial. Hence, it is necessary to find the appropriate Web services in a very short period of time. Replanning at runtime becomes necessary, if the Web services chosen at design time are not available anymore. Obviously, a service consumer is not willing to wait for the transaction of a predefined functionality or workflow. Thus, it is necessary to identify and make use of other possibilities to minimize the retrieval time for Web services.

This leads to an optimization problem based on the objective function and constrains which have to identified. Instead of using time-consuming linear integer programming, I propose the usage of heuristics in order to minimize computation time.

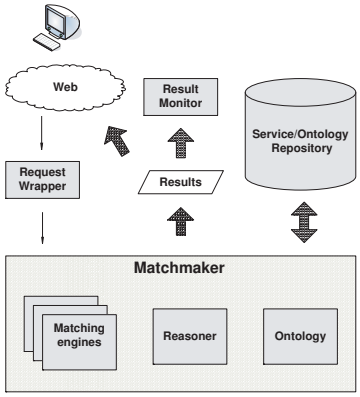

Fig. 1. Overview of Evaluation Workbench SEM.KOM

4 Next Steps

A prototypical implementation of the evaluation workbench has already been carried out in SEM.KOM (cp. Fig. 1 and [7]). Although the capabilities of this workbench are permanently enhanced, we have not yet implemented all possible features of SEM.KOM. In particular, it is necessary to include more retrieval approaches and Web service standards.

Currently, the components illustrated in Fig. 1 have been realized as follows:

- The *request wrapper* uses a RDF/XML format to convert the service request into a comparable format. It is possible to use a keyword-based search or to post the request in terms of a complete OWL-S description. The *Jena Semantic Web Framework* (version 2.5.4) is used to read and write RDF- and OWL-statements. In order to parse OWL-S, we use OWL-S API 1.1.0 beta (`http://www.mindswap.org`).
- It is possible to choose from two approaches of SWS retrieval (*matching engines*): the implementation of either logic-based reasoning as presented in [8] or keyword-based search. A combination of these approaches is also provided.
- The *service repository* is available in the form of files in a directory and can be accessed via an interface which wraps all the services advertised. At the moment, we are deploying OWLS-TC (version 2.2 [2]) as the dataset for testing.
- The *result monitor* provides the quality metrics precision, recall, F1 score and average query response time and stores them and all corresponding metadata.

In the near future, the following steps will be performed:

The Workbench will be enhanced by more service wrappers in order to support SAWSDL-based service requests. Furthermore, it is planned to provide the workbench via a website or as a Web service. One further long-term objective could be the provision of an ontology-based GUI which allows for the browsing of services.

The Heuristics have to be (mathematically) derived, implemented and evaluated both within SEM.KOM and in the contests mentioned in Sect. 2. Furthermore, the heuristics should be able to deal with incomplete semantic information.

Acknowledgements

This work is supported in part by the E-Finance Lab e. V., Frankfurt am Main, Germany (`http://www.efinancelab.com`).

References

1. Klusch, M., Fries, B., Khalid, M., Sycara, K.: OWLS-MX: Hybrid Semantic Web Service Retrieval. In: Proceedings Proceedings of the 1st International AAAI Fall Symposium on Agents and the Semantic Web, Arlington VA, USA (2005)
2. Klusch, M.: OWLS-TC-v2.2. `http://projects.semwebcentral.org/projects/owls-tc` (2007)
3. Küster, U., Lausen, H., König-Ries, B.: Evaluation of Semantic Service Discovery – A Survey and Directions for Future Research. In: Preliminary Proceedings of the 2nd Workshop on Emerging Web Services Technology (WEWST), Halle, Germany, pp. 37-53 (2007)
4. Li, L., Horrocks, I.: A Software Framework For Matchmaking Based on Semantic Web Technology. In: Proceedings of the 12th International World Wide Web Conference (WWW 2003), Budapest, Hungary, pp. 331-339 (2003)
5. Paolucci, M., Kawamura, T., Payne, T. R., Sycara, K. P.: Semantic Matching of Web Services Capabilities. In: Proceedings of the First International Semantic Web Conference (ISWC'02), Sardinia, Italy, pp. 333-347 (2002)
6. Petrie, C., Margaria, T., Küster, U., Lausen, H., Zaremba, M.: SWS Challenge: Status, Perspectives, and Lessons Learned so far. In: Proceedings of the Ninth International Conference on Enterprise Information Systems (ICEIS 2007), Funchal, Madeira, Portugal, pp. 447-452 (2007)
7. Schulte, S., Eckert, J., Repp, N., Steinmetz, R.: An Approach to Evaluate and Enhance the Retrieval of Semantic Web Services. Forthcoming in: Proceedings of the 5^{th} International Conference on Service Systems and Service Management (ICSSSM'08), Melbourne, Australia (2008)
8. Sycara, K. P., Paolucci, M., Ankolekar, A., Srinivasan, N.: Automated discovery, interaction and composition of Semantic Web services. Journal of Web Semantics **1** (1) (2003) 27-46
9. Xu, B., Zhang, P., Li, J., Yang, W.: A Semantic Matchmaker for Ranking Web Services. Journal of Computer Science Technology **21** (4) (2006), 574-581

OntoGame: Games with a Purpose for the Semantic Web

Extended PhD Thesis Abstract

Katharina Siorpaes
STI, University of Innsbruck, Austria
katharina.siorpaes@sti2.at

1. Research Problem

A pre-requisite for the Semantic Web to become a reality is the availability of ontologies [1] and meta-data. In many cases, it might be necessary to align between different ontologies in order to ensure interoperability. The research in the area of semantic content authoring has brought up an inventory of mature techniques and tools for semantic content creation. However, there is a severe lack of semantic data available on the Web: one can only find few well-maintained ontologies, respective alignments and very little semantic annotation. For instance, a search on Watson[1] or Swoogle[2] for a tourism ontology does not deliver a proper tourism ontology even though travel and tourism ontologies have been created in many academic projects in the last couple years. Furthermore, one can observe very little involvement of Web users in the process of semantic content creation. However, this involvement is urgently needed: there are tasks that are trivial for a human user but still difficult for a computer [2, 3]. Conceptual modeling and semantic annotation are tasks that depend on human intelligence: even though approaches for automating these activities exist, the problem has not been solved completely yet and human input is required at some stage. Therefore, we are now confronted with the situation that even though the technology is available, there is very little semantic content which can be traced back to only little user involvement. We believe that this is caused by missing incentive structures: the effort of building ontologies currently outweighs the benefit.

2. Motivation and Contribution

This is in sharp contrast to the Web 2.0 movement, which has proper incentive structures in place [4-6]. In my thesis, I investigate intrinsic motivations of users for contributing to Web 2.0 applications and propose to define possible incentive models for the Semantic Web. More precisely, I propose to masquerade core tasks of weaving the Semantic Web behind on-line, multi-player game scenarios, in order to create proper incentives for humans to contribute. Doing so, I adopt the findings from the already famous "games with a purpose" by von Ahn [2], who has shown that presenting a useful task, which requires human intelligence, in the form of an on-line game can motivate a large amount of people to work heavily on this task, and this for free.

[1] http://watson.kmi.open.ac.uk/WatsonWUI/
[2] http://swoogle.umbc.edu/

The **contribution of my thesis** is (1) an overview of incentives for users to contribute to Web 2.0 applications, (2) a survey on serious games and games with a purpose, (3) a conceptual framework that aims at (a) defining incentives (more precisely, intrinsic motivations) for the Semantic Web and (b) describing how to hide semantic content creation and maintenance tasks behind online games. Furthermore, I will provide (3) a proof-of-concept implementation with four cool games scenarios that will be available to the general public. Finally, I will (4) evaluate the fun factor of the games and (5) analyze the output of the games checking the correctness and the usefulness of the resulting data. OntoGame is an approach to the massive generation of lightweight knowledge structures that can serve as a starting point for further axiomatization, as training sets for semi-automatic approaches, and that can be useful for machine learning techniques.

3. Related Work

Several "games with a purpose" have been described by **Luis von Ahn** and colleagues; they also coined the term *"human computation"*: The ESP game [7] aims at labeling images on the Web - two players, who do not know each other, have to come up with identical tags describing an image. Peekaboom [8] works similar and has the objective of locating objects within images. Verbosity [9] is a game for collecting common sense facts. Phetch [10] is a computer game that collects explanatory descriptions of images in order to improve accessibility of the Web for the visually impaired. Law, von Ahn, and colleagues [11] came up with a game called Tagatune for music and sound annotation based on tags. However, their current prototypes remain mostly at the **level of lexical resources only**, i.e. terms and tags and are not directly connected with Semantic Web research.

Liebermann and colleagues describe the game Common Consensus [12], which aims at collecting human goals in order to recognize goals from user actions and conclude a sequence of actions from these goals.

Another approach to collecting common sense knowledge is the FACTory Game[3] published by **Cycorp**[4]: FACTory is a single-player online game that randomly chooses facts from the Cyc knowledge base [13] and presents them to the players. The player has to say whether the statement is true, false, doesn't make sense, or whether the user does not know. The answers are scored depending on accordance with the majority of answers.

A different type of games are so called passively multiplayer online games[5], a term coined by **Justin Hall**. The idea of the PMOGs[6] is to create avatars and game moves in multiplayer online games from user behavior on the Web. In other words, PMOGs translate e-mail content, chat logs, pictures, etc. into hunting parties, teams, puzzles, and so on.

4. Approach

In my PhD thesis, I propose to hide relevant tasks of semantic content authoring behind online games. In this section, I outline the challenges of my approach and the design

[3] http://game.cyc.com
[4] http://www.cyc.com
[5] http://passivelymultiplayer.com/PMOGPaper.html
[6] http://www.passivelymultiplayer.com

principles. Finally, I describe four cool OntoGame scenarios and explain how they address tasks in the Semantic Web lifecycle.

Challenges

The design of games for building the Semantic Web involves several challenges:
(1) Conceptual model of the games: it is crucial to make sure that the games are interesting and deliver useful output at the same time. This involves not only nice user interface design but also methods to keep up interest. One example for this would be revealing information about the partner (gender, age, nationality, etc.).
(2) Input data: For most game scenarios, a large corpus of knowledge, such as Wikipedia or YouTube, is required.
(3) Deriving formal semantics: from the games, formal semantics must be extracted, i.e. exports in common languages such as OWL.
(4) Cheating: ways to avoid cheating must be described and implemented.
(5) Re-use and analysis of generated data: in order to increase the amount of diverging data gathered as well as further deepening the degree of detail of the data, one must find algorithms and mechanisms to re-use gained data.
(6) Typical Mistakes: from first experiments, it is obvious that there are some cases where users tend to make mistakes, i.e. classifying something as a sub-class of a concept that is not correct. One has to find ways to avoid the impact of these "false friends".

Design Principles

I build my work on OntoGame on the following design principles:
I. Fun and Intellectual Challenge
Fun and the game challenge are the predominant user experience. The actual tasks are very well hidden such that their serious and useful nature does not decrease the "fun factor". Additionally, the games should comprise an intellectual challenge being fun and interesting at the same time.
II. Consensus
In our games, I adopt the "Wisdom of crowds" [14] paradigm. Groups only perform well under certain conditions: the group must be diverse, geographically dispersed, and members must be unable to influence each other. The settings of our games fulfill these requirements in order to tap the "wisdom of crowds".
III. Massive Content Generation
The assumptions about the intelligence of groups are only true given mass participation. Our games aim at the massive generation of semantic content, and thus mass user participation.

Four Cool Scenarios

In order to evaluate the set of abstract game scenarios, four games were implemented[7] that address the whole Semantic Web lifecycle (Fig. 1): certain tasks involved in ontology construction, alignment, and semantic annotation can be hidden behind online games. In a

[7] OntoPronto is released, OntoTube and SpotTheLink are close to release, OntoBay implementation is starting now. All four scenarios can be expected by summer 2008 latest.

nutshell, the OntoGame[8] series includes the following games: **OntoPronto** is a game for annotating Wikipedia and for creating a huge general interest ontology (the English Wikipedia currently contains more than 2 Million articles). **SpotTheLink** aims at aligning the product and service classifications eCl@ss and UNSPSC, respectively their OWL counterparts eClassOWL and unspscOWL. **OntoTube** (Fig. 2) produces annotations for YouTube videos. A fourth upcoming scenario is called **OntoBay** and is a game for annotating eBay auctions by expressing the type of goods being offered using eClassOWL.[9]

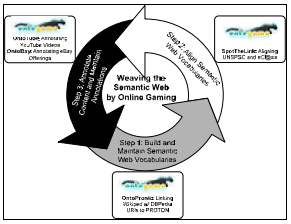

Fig. 1. Games for the Semantic Web Lifecycle

Fig. 2. Annotating YouTube

5. Evaluation and Preliminary Evidence

The objective of the evaluation is twofold: (1) to evaluate whether our first prototype creates an entertaining **gaming experience** and (2) whether the consensual conceptual choices of players in the **games are correct**. My hypothesis is that the majority of the players' decisions are ontologically correct and players will enjoy the games. We plan to release the four prototypes to the general public on several game platforms and make many users play the games. I will then analyze the resulting data: absolute number of games, absolute number of resources (Wikipedia articles, YouTube videos, etc.), ratio of single-player games, time invested by users, figures about the degree of consensus, and most importantly, the quality of conceptual salutation, i.e. mistakes that were made. For this purpose I will take representative samples and will ask experts to judge the correctness of the data. Furthermore, I will conduct surveys among players evaluating the fun factor, similar to the survey described in [15].

Preliminary evidence [15] indicates that this hypothesis is correct: OntoPronto, the first game of the OntoGame series, was released to the general public in Dec. 2007. Within the first two days, more than 200 players registered and played the game. The results of the analysis are promising: players make few mistakes and manage to find consensus in the majority of cases.

6. Expected Impact and Roadmap

Designers of semantic applications should start to think about incentives for users to invest time in those applications: in my thesis, I will provide helpful guidelines for adopting those

[8] http://www.ontogame.org
[9] A more detailed description of the games can be found in [16].

from Web 2.0 to Semantic Web. More precisely, the thesis will focus on games, implementing the motivation fun and competition. I believe that the games described in my PhD thesis have the potential to generate a huge amount of lightweight knowledge structures that are useful in several aspects: (1) use of the resulting data with very little or no changes as lightweight ontologies and annotations, (2) use of the resulting knowledge structures as a basis for domain ontologies for further axiomatization, (3) use as training data for semi-automatic approaches, and (4) for machine learning.
So far, OntoPronto has been released to the general public. OntoTube and SpotTheLink are currently being tested. All four scenarios are expected to be online and broadly published by summer 2008. So far, my work was published in [15-17].

References

1. Gruber, T.R., *Toward principles for the design of ontologies used for knowledge sharing.* International Journal of Human-Computer Studies, 1995. **43**: p. 907-928.
2. Von Ahn, L., *Games with a Purpose.* IEEE Computer, 2006. **29**(6): p. 92-94.
3. Von Ahn, L., et al. *CAPTCHA: Using Hard AI Problems for Security*. *Eurocrypt 2003*.
4. Marlow, C., et al., *Position Paper, Tagging, Taxonomy, Flickr, Article, ToRead*, *Proceedings of the World Wide Web Conference (WWW2006)*. 2006, ACM: Edinburgh, Scotland.
5. Hemetsberger, A., *When Consumers Produce on the Internet: The Relationship between Cognitive-affective, Socially-based, and Behavioral Involvement of Prosumers.* The Journal of Social Psychology, 2003.
6. Kuznetsov, S., *Motivations of Contributors to Wikipedia.* ACM SIGCAS Computers and Society, 2006. **36**(2).
7. Von Ahn, L. and L. Dabbish, *Labeling Images with a Computer Game*, *CHI 2004*. ACM.
8. Von Ahn, L., *Peekaboom: A Game for Locating Objects in Images*, *Proceedings of the SIGCHI conference on Human Factors in computing systems*. 2006, ACM: Montréal, Québec, Canada.
9. Von Ahn, L., M. Kedia, and M. Blum, *Verbosity: a game for collecting common-sense facts*, *Proceedings of the SIGCHI conference on Human Factors in computing systems*. 2006, ACM: Montréal, Québec, Canada.
10. Von Ahn, L., et al., *Improving Accessibility of the Web with a Computer Game*, *Proceedings of the SIGCHI conference on Human Factors in computing systems CHI '06*. 2006, ACM.
11. Law, E., et al. *Tagatune*. in *ISMIR 2007*. 2007. Vienna, Austria: OCG.
12. Lieberman, H., D. Smith, and A. Teeters, *Common Consensus: A Web-based Game for Collecting Commonsense Goals*, in *Workshop on Common Sense for Intelligent Interfaces, ACM International Conference on Intelligent User Interfaces (IUI-07)*. 2007: Honolulu.
13. Lenat, D.B. and R.V. Guha, *Building Large Knowledge-based Systems: Representation and Inference in the Cyc Project*. 1990, Boston, Masschusetts: Addison-Wesley.
14. Surowiecki, J., *The Wisdom of Crowds*. 2003, New York: Anchor Books Random House.
15. Siorpaes, K. and M. Hepp, *Games with a Purpose for the Semantic Web.* IEEE Intelligent Systems, Special Issue on Semantic Web, Summer 2008. (Forthcoming)
16. Siorpaes, K. and M. Hepp, *OntoGame: Weaving the Semantic Web by Online Games*, in *European Semantic Web Conference (ESWC 2008)*. 2008, Springer LNCS: Teneriffe, Spain.
17. Siorpaes, K. and M. Hepp, *OntoGame: Towards Overcoming the Incentive Bottleneck in Ontology Building*, in *Proceedings of the 3rd International IFIP Workshop On Semantic Web & Web Semantics (SWWS '07) co-located with OTM Federated Conferences.* 2007, Springer LNCS: Vilamoura, Portugal.

Trend Mining with Semantic-Based Learning

Olga Streibel

Networked Information Systems, Free University Berlin,
Königin-Luise-Str.24-26 , 14195 Berlin, Germany
streibel@inf.fu-berlin.de
http://www.ag-nbi.de

Abstract. Mining trends by analyzing text streams could enhance the standard trend analysis based on numeric data. The use of qualitative information in the process of trend recognition, in addition to that of quantitative data, requires new analysis techniques. Since Semantic Web enables the appropriate and advantageous formalization of knowledge, we propose to include formalized expert knowledge in the process of trend recognition. In this preliminary work, we introduce our approach based on Semantic Web technologies combined with Data Mining methods for mining trends in a given domain.

Key words: trend mining, trend recognition, semantic technologies, pattern recognition, trend patterns, learning methods, trend pattern ontology

1 Introduction

"Stock market news has gone from hard to find (in the 1970s and early 1980s), then easy to find (in the late 1980s), then hard to get away from."[1]
A huge amount of textual information like business news is freely available on the Internet[2]. This abundance of information makes the access of new information far easier, as is also true of previously hidden knowledge. On the other hand, in order to retrieve required information and discover the potential knowledge, we need to utilize appropriate search and analysis techniques. Regarding business news and the stock market, a "human" specialist can deduce information and knowledge she needs for the prediction of market movements. However, this recognition and comprehension process is very complex and requires experience as well as the initial context knowledge.
In our work, we concentrate on the *trend mining* process based on numeric data and on textual information. Research projects like GIDA and TREMA have shown that there is a huge demand for the research on and development of

[1] Peter Lynch,2000 "One Up On Wall Street: How To Use What You Already Know To Make Money In The Market"
[2] i.e. http://news.bbc.co.uk/2/hi/business, http://www.tagesschau.de/wirtschaft/index.html, http://faz.net

useful trend mining methods that are able to include analyses of textual information in the process of trend recognition. In our work, we define repositories consisting of quantitative data and qualitative data as simple hybrid information systems. Regarding specific application fields, i.e. financial markets, the qualitative data is represented by financial news whereas the quantitative data means the numeric values of different trade instruments. Consequently, we aim to use text corpus consisting of financial news in German language[3] and correlate this corpus with the trading values of a chosen financial instrument. In particular, we concentrate on the analysis of the business news filtered over a period of 12 months due to the trend segments deduced from the market values of a trading instrument. The focus of our research is on developing a solution relevant to the trend mining problem in simple hybrid information systems which combines a Data Mining approach and adequate Semantic Web technologies. There are many other examples of simple hybrid information systems in application areas like weather forecasting, traffic analysis, customer opinion mining, etc. We will work on a solution that will be applicable in those different systems.
In the following, we outline briefly the idea of our novel approach for trend mining. Section 2 gives an insight into research relevant to our work. In section 3, we specify our definition of a "trend" and outline the issues of our research. Describing briefly the different methods from Computer Linguistic which can be partially applied to the trend mining difficulty, we introduce *Extreme Tagging System* (ETS) in 3.2. We close the section with a short paragraph about learning methods that we aim to apply in the future.

2 Related Work

The research project GIDA[4][6][1] and its follower, TREMA[5], concentrated on the fusion of multimodal market data in order to mine trends on financial markets (GIDA, TREMA) and in market research (TREMA). These projects provide us with our research direction. Since we aim to focus only on a fraction of the whole trend mining process, in particular, on the search for the trend indicating language patterns in news, we are not going to concern ourselfs with the conception of a complex trend mining framework as the project TREMA does. Similar to TREMA, we are using the Semantic Web technologies in order to support the textual trend recognition. The difference lies in our idea of applying an ETS, as described in section 3.2, instead of applying classic and hierarchical ontologies.
In [3] the concept of velocity density estimation is discussed for the trend mining in supermarket customers' data.This work "provides the user generic tool to understand, visualize and diagnose the summary changes in data characteristics". The aspects of dynamics and evolving data included in this research, could also

[3] The corpus is available due to the cooperation with the German company, neofonie GmbH
[4] Description online: www.computing.surrey.ac.uk/ai/gida
[5] Project website: www.trema-projekt.de

be important for our work. The authors of [16] introduce a simple and interesting knowledge-based approach for the kidney function monitoring in medical diagnosis systems. In particular, the trends appear in the form of trend reports which are counted on the numeric data and explained using a knowledge-base. The use of a semantic knowledge-base will also be a part of our work. We are going to use the knowledge base not only to explain the emerging trends but also to learn from them. Trends based on keyword search statistics are well visualized by the Google-Trends [24] feature. Here, the trend mining of searches actually shows anomalies appearing in the historic patterns of Google search on the Web. Search for certain text patterns in the text corpus is also a part of our work. The difference is that we aim to search for trend indicating keywords that have been learned from historic data using semantic, not only statistic methods. Another interesting tool is the BlogPulse [25] that identifies topics and subjects that people are talking about in their blogs. BlogPulse shows the complex trend concept. A trend is a phenomenon that consists of trend setters (blogs' authors), detected topics, "buzz" words, etc. In our work, we are assuming a simplified, data and text oriented, trend definition that can be treated as a fraction of the complex trend mining process.
As last, the work described in [10] could be very useful for us. In particular, the definitions of *theme*, *theme life cycle*, and *theme snapshot* could be important for our approach.

3 Mining Trends

In order to analyze trends, we have to define what is a *trend*. Since we aim firstly to originate our trend recognition process in the numeric data, we will treat the given text stream in a similar way as we might a data stream. With regard to the trend analysis based on time series, the analysis process consists of four major *components* or *movements* for characterizing time-series data [8]. We refer to the *long-term movements* that can be visualized by a *trend curve*. Based on the trend curve generated over quantitative data, we identify *time segments* for those long-term movements that can have positive or negative trend values ("ups" and "downs" on the market). Correlating this segments to the news stream, we identify a priori three trend classes: positive, negative and neutral class and divide the news stream in the 3-category text corpus. Analyzing text corpus, we will search for specific, so called *trend-indicating* keywords and statements. Trend-indicating keywords from the financial market domain are i.e. *cut*, *concern*, *recession*, etc. These simple keywords are subject to what we call trend indicating *language patterns*.
When analyzing text corpus, we are concentrating on trend indicating language structure and on the characterization of this structure. Firstly, we propose to divide the identification of trend indicating language patterns (in the following also called simple *trend patterns*) in the non-semantic feature extraction and in semantic feature annotation (more in sections 3.1 and 3.2).

In the following, we briefly describe stages in our proposed approach for the trend recognition method.

3.1 Non-semantic trend patterns

Since we analyze a given text corpus that is divided in trend classes, the "simplest" method for identifying trend patterns is the counting of the most frequent keywords or the TFIDF-method[15]. Different methods from *text mining* can be successfully applied in order to identify keywords or simple statements from the text corpus. However, we assume that not every keyword or statement extracted from the given trend class in text corpus is the trend-indicating one. The interesting point is how to recognize whether given keywords or statements are trend-indicating or not.
In particular we rely on the observation that there are characteristic words used in different domains describing the customer's opinion and/or her sentiment[2][9][19]. Following from this, since most sentiment indicating words are *adjectives* whereas the *nouns* build the sentiment concepts, then a possible and very simple trend pattern in the text could consist of an adjective-noun word pair. Using WordNet[6] or a Part-Of-Speech analysis, we could identify these pairs in the text corpus. Regardless, we assume that the search for trend patterns requires more complex text analysis then the POS. We assume, that we should investigate taxonomic and non-taxonomic relations between identified keywords or simple statements. Additionally, we should consider the semantic orientation as described in [7] and [19].

3.2 Trend pattern ontology

The non-semantic trend feature extraction provides a basis for a trend pattern structure. This can be useful for both, analyzing trend patterns on the non-semantic level and creating a trend knowledge base that provides insight into the general characteristic of the trend patterns. A knowledge base can be realized as a classic ontology. We propose the application of an adapted Extreme Tagging System (ETS) as a knwoledge base for trend recognition. An ETS as introduced in [18], is an extension of collaborative tagging systems which allow for the collaborative construction of knowledge bases. An ETS offers a superset of the possibilities of collaborative tagging systems in that it allows us to collaboratively tag the tags themselves, as well as the relations between tags. ETS are not destined to exclusively produce hierarchical ontologies but strive to allow the expression and retrieval of multiple nuances of meaning, or semantic associations. Our propose in this research is to use these novel knowledge acquisition techniques, which are based on lightweight annotations in social environments, in order to generate a semantic description for the analyzed application field. We will apply an adapted ETS in order to gain expert knowledge of trend recognition in the business field. We expect that the use of an ETS will bring an easy

[6] http://wordnet.princeton.edu/

retrieval and extraction of the expert knowledge in the form of a RDF triple set. An initial set of tags (which should be tagged by experts in a given domain) will be generated from the selected trend features that are extracted in a non-semantic way from the text corpus (described in 3.1). Experts using the ETS will play the "association game" on the initial tag set. Created association sets will be automatically converted to RDF-data. Produced RDF triple set will be then used to generate a trend scheme. Furthermore, we will use the data from ETS as the input for another feature extraction from the texts.
Combining the non-semantic search for trend patterns with the association sets based on expert knowledge, we aim to create an appropriate semantic trend pattern scheme- a trend pattern ontology- that will be applied to a learning algorithm.

3.3 Learning Trends

Regarding different possibilities of learning methods from machine learning [11][14][21], we firstly propose to use the supervised learning approach. Hence we work with strictly separable text classes- the texts with positive trend indicating patterns cannot belong to the neutral or negative trend category at the same time- standard classification seems to be an appropriate learning form for the trend recognition problem, particularly where the trend classes' ranges are well separable. With regard to the evaluation of the advantages achieved through applied semantics to the learning process, we propose to use firstly decision trees (i.e. C4.5) or decision rules [21] which both allow the vizualization of the learned model. Learning trends with decision trees means here learning trend indicating language patterns that are expressed in RDF-triples.
However, once the feature space has been created from the text corpus (as described in 3.1 and 3.2), we can use the features in order to validate the assumptions about the positive, negative and neutral trend indicating patterns. Therefore, we can use clustering as the alternative learning method for automatically assigning the trend classes' ranges. In our research we are considering also different alternative learning algorithms like rough sets, fuzzy case reasoning, neural networks or inductive learning approaches [14][21][13][8] in order to find the most appropriate one for the semantic-based trend recognition.

4 Future work

Given the directions for research outlined in section 3, we have chosen to continue our work on the theoretical and the practical solutions in order to create a prototype of here described semantic-based learning method for trend recognition in simple hybrid information systems.

5 Acknowledgments

This work has been partially supported by the "InnoProfile-Corporate Semantic Web" project funded by the German Federal Ministry of Education and Research (BMBF) and the BMBF Innovation Initiative for the New German Länder - Entrepreneurial Regions. The author would like to thank their supervisor, Prof.Robert Tolksdorf and the TREMA-project partners for the support of this work.

References

1. Ahmad, K.: Events and the Causes of Events, In Conference on Terminology and Knowledge Engineering 2002, online: http://www.computing.surrey.ac.uk/ai/TKE
2. Archak, K., Ghose, A., Ipeirotis, P. G.: Show me the Money! Deriving the Pricing Power of Product Features by Mining Consumer Reviews
3. Charu, C. Aggarwal: A framework for diagnosing changes in evolving data streams, SIGMOD 2003: Proceedings of the 2003 ACM SIGMOD international conference on Management of data, 575-586,(2003)
4. Hevner, A. R., March, S.T., Park, J., Ram, S.: Design Science in Information System Research,MIS Quarterly 2004
5. Esuli, A. and Sebastiani, F.: SentiWordNet: Publicly Available Lexical Resource for Opinion Mining
6. Gillam, L., Ahmad, K., Ahmad, S., Casey, M., Cheng, D., Taskaya, T., Oliveira, P.C.F. and Manomaisupat, P.: Economic News and Stock Market Correlation: A Study of the UK Market. In Conference on Terminology and Knowledge Engineering 2002, online: http://www.computing.surrey.ac.uk/ai/TKE
7. Hatzivassiloglou, V. and McKeown, K. R.: Predicting the semantic orientation of adjectives. In Proceedings of the 35th Annual Meeting of the ACL
8. Han, J., Kamber, M.: Data Mining Concepts and Techniques, 2.Ed. Morgan Kaufmann 2006
9. Hu, M., and Liu, B.: Mining and summarizing customers reviews. In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004) (2004), pp. 168-177
10. Mei, Q., Liu, C., Su, H., and Zhai, C.: A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In Proceedings of the 15th International Conference on World Wide Web (Edinburgh, Scotland) WWW'06 ACM Press, New York, NY, 533-542.
11. Mitchell, T.M.: Machine Learning, Mc-Graw-Hill, 1997
12. Morinaga, S., Yamanishi, K..: Tracking Dynamics of Topic Trends, KDD'04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge Discovery and Data Mining, 811-816, ACM NY
13. Pal, S.K. and Mitra, P.: Pattern Recognition Algorithms for Data Mining, CRC Press LLC 2004
14. Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach, Prentice Hall, 2.Ed.2003
15. Salton, G., Buckley Ch.: Term Weighting Approaches in Automatic Text Retrieval, 1988 Information Processing and Management: an International Journal archive Volume 24 , Issue 5 (1988) Pages: 513 - 523

16. Schleutermann, S. and Heidl, B. and Finsterer, U.: Trenderkennung beim Nierenfunktionsmonitoring auf der Intensivstation, GMDS 139-142, 1996
17. Simon, H.A.: The Science of the Artificial, Ch.4: Remembering and Learning, MIT Press, Third Edition (1996)
18. Tanasescu, V., Streibel, O.: Extreme Tagging: Emergent Semantics Through the Tagging of Tags. In International Workshop on Emergent Semantics and Ontology Evolution, ISWC2007
19. Turney, P.D., and Littman, M.L.: Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems 21, 4 (2003), 315-346
20. Vejlgaard, H.: Anatomy of a Trend Mc-Graw-Hill, 1.Ed. 2007
21. Witten, I.h., Frank, E.: Data Mining Practical Machine Learning Tools and Techniques, 2.Ed.Morgan Kaufmann 2005
22. Witten, I.H., Gori, M., Numerico, T.: Web Dragons: Inside The Myths of Search Engine Technology, Morgan Kaufmann 2007
23. Wong, W.-K., Moore, A., Cooper, G., Wagner, M. What is Strange About Recent Events (WSARE) in Journal of Machine Learning Research 2005
24. `www.google.com/trends`
25. `www.blogpulse.com`
26. `www.projekt-trema.de`

Semantics-aware Software Project Repositories

Jonas Tappolet

Department of Informatics, University of Zurich, Switzerland
tappolet@ifi.uzh.ch

Abstract. This proposal explores a general framework to solve software analysis tasks using ontologies. Our aim is to build semantically annotated, flexible, and extensible software repositories to overcome data representation, intra- and inter-project integration difficulties as well as to make the tedious and error-prone extraction and preparation of meta-data obsolete. We also outline a number of practical evaluation approaches for our propositions.

1 Research Problem

In *Software Engineering*, many tools have been used for years to support the collaboration of development teams. Among others, these tools are typically a version control system holding the project's files in a versioned manner and a bug tracking system in which defects and enhancement requests are stored. Research has shown that these repositories contain a huge amount of additional information that can be exploited to enhance the quality of software systems (such as the detection of error-prone software patterns or the prediction of the number of defects).An in-depth analysis reveals, however, that there are three main obstacles to seamless software analysis: (1) *data representation* — the natural structure of the contents of these repositories is a typed graph (*e.g.*, source code syntax trees). However, most of current software analysis tools use relational database management systems requiring a transformation of the data to the relational format. More importantly, they rely on propositional (i.e., one table) representations for analysis, as most data mining algorithms use this format. Moreover, the data is often extracted in a form that is suitable for only a particular task. These transformations are tedious, error-prone, and, usually, lossful. (2) *Intra-project repository integration* — each of these repositories is designed as a stand-alone system covering only a specific part of the software development process. In order to generate uniform views on software projects, methods are needed to overcome the boundaries of each isolated repository. (3) *Inter-project repository integration* — fewest software projects use solely their own code. Developers make pervasive use of components and frameworks hosted in remote repositories, weaving a world-wide call graph. Hence, repositories of different projects need to be accessed in a integrated manner.

In this paper we present our EvoOnt approach to address these three problem areas. By integrating different repositories (data sources) using Semantic Web technologies, we end-up in a graph-based approach that is capable of handling distributed and heterogeneous software project data.

2 Related Work

In this section, we summarize a small selection of the most closely related studies addressing the identified problem areas.

Graph based software analysis. Collberg *et al.* [4] present Gevol, a graph-based visualization tool for CVS and Java. The aim is to support developers understanding the software by providing visual representations of the source code and its history. Sager *et al.* [13] present *Coogle* (Code Google) that implements a set of tree similarity measures to detect similarities between Java classes of different releases of software projects. Coogle's approach is to first transform the abstract syntax tree (AST) representations of Java classes into intermediary FAMIX tree representations [6], and second, to measure their similarity by applying tree similarity algorithms. FAMIX is a software source code meta model designed to serve as an exchange format for object-oriented programming languages using flat text streams. Dietrich [7] proposed an OWL ontology to model the domain of software design patterns to automatically generate documentation about the patterns used in a software system. With the help of this ontology, the presented pattern scanner inspects the ASTs of source code fragments to identify the patterns used in the code.

baetle is an open-source project that aims to add semantics to software repositories with a strong emphasis on bug tracker data.[1] We tightly work together with the *baetle* developers to combine our ontologies with theirs.

Intra-project repository integration. Fischer *et al.* [8] present their *Release History Database* (RHDB) that integrates versioning and bug tracking systems. The data is extracted and stored in a traditional relational database.

Antoniol *et al.* [1] combined the relational RHDB with FAMIX to integrate source code with bug tracking and versioning system information.

Inter-project repository integration. Chang and Mockus [3] present a method to detect file copies among different versioning systems to build a unified version history.

None of the approaches above combines the strengths of an integrated software ontology to address all three obstacles identified in the introduction.

3 Contributions

Data representation. Most of the studies presented in Section 2 use flat representations of source code. This forces analyses to be on a textual level. Although, valuable information can be extracted using text mining techniques, it is generally hard to detect different types of source code changes (*e.g.*, structural vs. nominal changes). Consider a similarity measure which is able to find changes on the textual source code level (i.e., treating software code as a string of characters). Although the measure can, for instance, say that two software artifacts are different if their copyright notes have changed, it can, however, not say anything about the impact of this change on the software's functionality. Therefore, we use

[1] `http://code.google.com/p/baetle/`

a graph-based approach to model the repositories using RDF/OWL ontologies, which allows both textual and structural analyses.

Intra-project repository integration. There are two different relation types among repositories. *Implicit* connections are defined by the data itself or given by the nature of a repository. The relation between a file's meta-data (e.g., create date, file name, version information) and its content implicitly connects different versions of the source code. *Explicit* connections need to be manually defined. The connection between a bug report (*i.e.*, a bug fix typically reported in a bug reporting tool such as Bugzilla) and a specific file version (typically stored in a version control systems such as CVS) needs to be explicitly defined by a developer. This is usually done by mentioning a bug number in the comment of a file's new version (commit message), which can be extracted using simple text mining techniques [8] to link the respective bug report with a version of a file. However, the extractability of such explicit connections relies on disciplined and uniform reporting practices of a development team. Another method of linking a bug with a version is to compare the closing date of a bug report with the creation date of a new version. Having matching or near-matching dates is a strong indicator for a connection.

With the integration of repositories we can access the history of a file with all changes made during a file's life cycle. We can differentiate between evolutionary changes (extension of functionality) and maintenance changes (fixes of bugs).

Inter-project repository integration. In a next step we can integrate a software project's model with the models of used external components. Whenever a program makes a call to an entity that is not located inside the same project, this can be considered an external function call. Our aim is to map these calls to their representing source code model in a remote repository. One convenient method is to relate external calls in the same way as internal ones differentiating them only using their differing namespaces (which need to have a uniform transformation to the source-code-namespace/package-declaration). Having this integration, we can explore the dependencies between a software and its components analyzing, for example, the impact of a component's replacement with another (e.g., How does the code need to get adapted?) or the relation between bugs and the usage of external components.

4 Research Plan

4.1 Current State of our Research

In a first step, we implemented a set of tools to extract and interconnect data from software repositories (*i.e.*, CVS, Bugzilla, and Java), and store it as instances according to EvoOnt's model. So far, we conducted several experiments using query techniques, reasoning, similarity measures, and machine learning to evaluate EvoOnt's ability to serve as a general software analysis framework. We briefly summarize our conducted experiments. In our previous work [11, 10], the experiments are described in detail with example data from the Eclipse project.

Software metrics. We used plain SPARQL queries to compute object-oriented software metrics [12]. These metrics are, *e.g.*, the number of bugs per file, the relative number of bugs or the fan-in fan-out of a class.

Software pattern detection. Using ontological reasoning, we are able to detect software patterns as well as anti-patterns, and code smells [12]. We achieved this by defining a pattern (anti-pattern) using the concepts from our EvoOnt. We build up an own pattern ontology which can, when conducting pattern detection, be combined with the existing ontologies an a reasoner will then match the defined patterns in the data.
Similarity measures / Software evolution. Having the data in a graph-based format, we are able to calculate structural similarities between two versions of a source code file using iSPARQL [9] running similar analyses as Coogle by executing a single iSPARQL query.
Machine learning. Using SPARQL-ML, a SPARQL extension with machine learning algorithms, we were able to simply reproduce tedious bug prediction analyses [2].

4.2 Next Steps

Intra-project integration. So far, we used Bugzilla-, CVS-, and Java-repositories as data sources to extract software information from. However, there are various other products, we plan to integrate: Jira (bug tracker), Subversion (verisoning system), and C# (programming language) are our next candidates to write RDF/OWL interfaces for. On the other hand, there are other repository types involved in the software development process such as mailing-lists and forums which we plan to integrate as well into our unified framework. These types reflect the social network structure around the development process.
Inter-project integration. Inspired from Data-Warehousing, where heterogeneous data is accessed through a sole interface, we plan to implement such behavior in EvoOnt as well. With the implementation of web-based, RDF/OWL enabled interfaces exposing SPARQL-endpoints (*e.g.*, for versioning systems), it would be able to execute and answer SPARQL queries over the web. This enables a repository to be linked from any other software project using this component's functionality.
Integration with software project repositories. We intend to integrate the semantic capabilities of EvoOnt into commonly used software project repository tools (such as Subversion and Jira) making the tedious extraction and preparation of meta-data obsolete.
Evaluation. In our first set of experiments, we evaluated EvoOnt against a wide variety of software analysis tasks. In a next step, we plan to deepen certain experiments. A first selection is:
The identification of the location of a bug. Usually, a developer links a version of a source code file with a bug report whenever she fixed a specific bug. Derived from this information, we can use graph algorithms to compute deltas between the fixed and the pre-fixed source code version resulting in a subgraph exactly representing the change made for fixing a bug. Having this portion of change, we can try to identify the point in the history of the software when this changed code fragment was inserted or modified. Our hypothesis would be that this is the point in the software history where the bug was introduced.
Analyze the code co-evolution of projects and their components. This important

task has been very difficult so far, as the inter-project connections has been largely missing. The inter-project integration allows to uncover the relation between bugs of different projects. A bug may be misleadingly reported in a project due to a bug in the referenced project. Moreover, the impact of updating to a new version of a component that includes several bug fixes, and may have changed its behavior can be made visible.
Analyze the connection between code elements and people with respect to their relationship to Conway's Law [5] and perform other in-depth social network analyses.

5 Conclusions

This proposal outlined the advantages of applying Semantic Web technologies to the field of Software Analysis. Specifically, we discussed how Semantic Web technologies allow to overcome the data representation, intra-, and inter-project repository integration problems. We, furthermore, succinctly outlined how we intend to evaluate our approach by showing its usefulness for a variety of software analysis tasks and publish the findings in the software engineering literature. We also indicated our plans to develop semantically annotated software repositories, which will make the extraction and preparation of meta-data obsolete.

References

1. G. Antoniol, M. D. Penta, H. C. Gall, and M. Pinzger. Towards the integration of versioning systems, bug reports and source code meta-models. *ENTCS*, 2005.
2. A. Bernstein, J. Ekanayake, and M. Pinzger. Improving defect prediction using temporal features and non linear models. *IWPSE*, 2007.
3. H.-F. Chang and A. Mockus. Constructing universal version history. *MSR*, 2006.
4. C. Collberg, S. Kobourov, J. Nagra, J. Pitts, and K. Wampler. A system for graph-based visualization of the evolution of software. *SoftVis*, 2003.
5. M. Conway. How do committees invent? *Datamation*, 1968.
6. S. Demeyer, S. Tichelaar, and P. Steyaert. FAMIX 2.0 - the FAMOOS information exchange model. 1999.
7. J. Dietrich and C. Elgar. A formal description of design patterns using owl. *ASWEC*, 2005.
8. M. Fischer, M. Pinzger, and H. Gall. Populating a release history database from version control and bug tracking systems. *ICSM*, 2003.
9. C. Kiefer, A. Bernstein, and M. Stocker. The fundamentals of iSPARQL - a virtual triple approach for similarity-based semantic web tasks. *ISWC*, 2007.
10. C. Kiefer, A. Bernstein, and J. Tappolet. Analyzing software with iSPARQL. *SWESE*, 2007.
11. C. Kiefer, A. Bernstein, and J. Tappolet. Mining software repositories with iSPARQL and a software evolution ontology. *MSR*, 2007.
12. M. Lanza and R. Marinescu. *Object-Oriented Metrics in Practice*. Springer, 2006.
13. T. Sager, A. Bernstein, M. Pinzger, and C. Kiefer. Detecting similar java classes using tree algorithms. *MSR*, 2006.