SEMAPRO 2013

The Seventh International Conference on Advances in Semantic Processing

ISBN: 978-1-61208-293-6

September 29 - October 3, 2013

Porto, Portugal

SEMAPRO 2013 Editors

Alexey Cheptsov, High Performance Computing Center Stuttgart (HLRS), Germany

SEMAPRO 2013

Foreword

The Seventh International Conference on Advances in Semantic Processing (SEMAPRO 2013), held between September 29 and October 3, 2013 in Porto, Portugal, continued a series of events highlighting the most recent advances in ontology, web services, semantic social media, semantic web, deep semantic web, semantic networking and semantic reasoning.

The inaugural International Conference on Advances in Semantic Processing, SEMAPRO 2007, was initiated considering the complexity of understanding and processing information. Semantic processing considers contextual dependencies and adds to the individually acquired knowledge emergent properties and understanding. Hardware and software support and platforms were developed for semantically enhanced information retrieval and interpretation. Searching for video, voice and speech [VVS] raises additional problems to specialized engines with respect to text search. Contextual searching and special patterns-based techniques are current solutions.

We take here the opportunity to warmly thank all the members of the SEMAPRO 2013 Technical Program Committee, as well as the numerous reviewers. The creation of such a broad and high quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and efforts to contribute to SEMAPRO 2013. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

Also, this event could not have been a reality without the support of many individuals, organizations, and sponsors. We are grateful to the members of the SEMAPRO 2013 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that SEMAPRO 2013 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in the field of semantic processing.

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

SEMAPRO 2013 Chairs:

SEMAPRO Advisory Chairs

Wladyslaw Homenda, Warsaw University of Technology, Poland
Bich-Lien Doan, SUPELEC, France
Alexey Cheptsov, High Performance Computing Center Stuttgart (HLRS), Germany

SEMAPRO 2013 Industry Liaison Chairs

Peter Haase, Fluid Operations, Germany
Thorsten Liebig, derivo GmbH - Ulm, Germany

SEMAPRO 2013 Research Chair

Nima Dokoohaki, Royal Institute of Technology (KTH)-Kista, Sweden

SEMAPRO 2013

Committee

SEMAPRO Advisory Chairs

Wladyslaw Homenda, Warsaw University of Technology, Poland
Bich-Lien Doan, SUPELEC, France
Alexey Cheptsov, High Performance Computing Center Stuttgart (HLRS), Germany

SEMAPRO 2013 Industry Liaison Chairs

Peter Haase, Fluid Operations, Germany
Thorsten Liebig, derivo GmbH - Ulm, Germany

SEMAPRO 2013 Research Chair

Nima Dokoohaki, Royal Institute of Technology (KTH)-Kista, Sweden

SEMAPRO 2013 Technical Program Committee

Nasser Alalwan, King Saud University - Riyadh, Saudi Arabia
Riccardo Albertoni, IMATI-CNR-Genova, Italy
José F. Aldana Montes, University of Málaga, Spain
Panos Alexopoulos, iSOCO S.A., Spain
Eckhard Ammann, Reutlingen University, Germany
Mario Arrigoni Neri, University of Bergamo, Italy
Sofia J. Athenikos, Amazon, USA
Isabel Azevedo, ISEP-IPP, Portugal
Ebrahim Bagheri, Athabasca University & University of British Columbia, Canada
Khalid Beljhajjame, University of Manchester, UK
Helmi Ben Hmida, FH MAINZ, Germany
Jorge Bernardino, ISEC - Polytechnic Institute of Coimbra, Portugal
Christopher Brewster, Aston University - Birmingham, UK
Volha Bryl, University of Mannheim, Germany
Dilletta Romana Cacciagrano, University of Camerino, Italy
Ozgu Can, Ege University, Turkey
Tru Hoang Cao, Vietnam National University - HCM & Ho Chi Minh City University of Technology, Vietnam
Nicoletta Calzolari, CNR-ILC (Istituto di Linguistica Computazionale del CNR), Italy
Delroy Cameron, Wright State University, USA
Sana Châabane, ISG - Sousse, Tunisia
Sam Chapman, Knowledge Now Limited, UK
Shu-Ching Chen, Florida International University, U.S.A.
Alexey Cheptsov, High Performance Computing Center Stuttgart (HLRS), Germany

Dickson Chiu, Dickson Computer Systems, Hong Kong
Smitashree Choudhury, UK Open University - Milton Keynes, UK
Sunil Choenni, Ministry of Security and Justice, Netherlands
Daniele Christen, Parsit Company, Italy
Soon Ae Chun, City University of New York, USA
Paolo Ciancarini, Università di Bologna, Italy
Ruben Costa, UNINOVA - Instituto de Desenvolvimento de Novas Tecnologias, Portugal
Frithjof Dau, SAP Research Dresden, Germany
Juri Luca De Coi, Université Jean Monnet - Saint-Etienne, France
Geeth Ranmal De Mel, University of Aberdeen - Scotland, UK
Cláudio de Souza Baptista, Computer Science Department, University of Campina Grande, Brazil
Thierry Declerck, DFKI GmbH, Germany
Jan Dedek, Charles University in Prague, Czech Republic
Chiara Di Francescomarino, Fondazione Bruno Kessler - Trento, Italy
Alexiei Dingli, The University of Malta, Malta
Bich Lien Doan, SUPELEC, France
Milan Dojčinovski, Czech Technical University in Prague, Czech Republic
Nima Dokoohaki, Royal Institute of Technology (KTH) - Stockholm, Sweden
Laura Dragan, National University of Ireland, Ireland
Raimund K. Ege, Northern Illinois University, USA
Enrico Francesconi, ITTIG - CNR - Florence, Italy
Naoki Fukuta, Shizuoka University, Japan
Frieder Ganz, University of Surrey, U.K.
Raúl García Castro, Universidad Politécnica de Madrid, Spain
Rosa M. Gil Iranzo, Universitat de LLeida, Spain
Fabio Grandi, University of Bologna, Italy
Sotirios K. Goudos, Aristotle University of Thessaloniki, Greece
Tudor Groza, University of Queensland, Australia
Francesco Guerra, University of Modena and Reggio Emilia, Italy
Alessio Gugliotta, Innova SpA, Italy
Peter Haase, Fluid Operations, Germany
Ivan Habernal, University of West Bohemia, Czech Republic
Armin Haller, CSIRO ICT Centre - Canberra, Australia
Carmem S. Hara, Universidade Federal do Parana, Brazil
Sven Hartmann, Clausthal University of Technology, Germany
Shun Hattori, Muroran Institute of Technology, Japan
Xin He, Airinmar Ltd., UK
Ralf Heese, Freie Universität Berlin, Germany
Tracy Holloway King, eBay Inc., U.S.A.
Wladyslaw Homenda, Warsaw University of Technology, Poland
Carolina Howard Felicissimo, BRGC - Schlumberger, Brazil
Ching-Hsien (Robert) Hsu, Chung Hua University, Taiwan
Sergio Ilarri, University of Zaragoza, Spain
Prasad Jayaweera, University of Sri Jayewardenepura, Sri Lanka
Wassim Jaziri, ISIM Sfax, Tunisia
Achilles Kameas, Hellenic Open University, Greece
Katia Kermanidis, Ionian University - Corfu, Greece
Pavel Klinov, University of Ulm, Germany

Jaroslav Kuchar, Czech Technical University in Prague, Czech Republic
Kyu-Chul Lee, Chungnam National University - Daejeon, South Korea
Thorsten Liebig, derivo GmbH - Ulm, Germany
Antonio Lieto, University of Turin, Italy
Héctor Llorens Martínez, Nuance Communications, Spain
Sandra Lovrenčić, University of Zagreb - Varaždin, Croatia
Hongli Luo, Indiana University - Purdue University Fort Wayne, U.S.A.
Eetu Mäkelä, Aalto University, Finland
Maria Maleshkova, The Open University, UK
Erik Mannens, Ghent University, Belgium
Miguel Felix Mata Rivera, Instituto Politecnico Nacional, Mexico
Maristella Matera, Politecnico di Milano, Italy
Michele Melchiori, Università degli Studi di Brescia, Italy
Elisabeth Métais, Cedric-CNAM, France
Vasileios Mezaris, Informatics and Telematics Institute (ITI) and Centre for Research and Technology Hellas (CERTH) - Thermi-Thessaloniki, Greece
Michael Mohler, Language Computer Corporation in Richardson, U.S.A.
Shahab Mokarizadeh , Royal Institute of Technology (KTH) - Stockholm, Sweden
Mir Abolfazl Mostafavi, Université Laval - Québec, Canada
Ekawit Nantajeewarawat, Sirindhorn International Institute of Technology / Thammasat University, Thailand
Vlad Nicolicin Georgescu, SP2 Solutions, France
Lyndon J. B. Nixon, STI International, Austria
Csongor Nyulas, Stanford Center for Biomedical Informatics, USA
David A. Ostrowski, Ford Motor Company, USA
Peera Pacharintanakul, TOT, Thailand
Andrea Perego, European Commission - Joint Research Centre, Ispra, Italy
Livia Predoiu, University of Oxford, UK
Hemant Purohit, Wright State University, USA
Jaime Ramírez, Universidad Politécnica de Madrid, Spain
Isidro Ramos, Valencia Polytechnic University, Spain
Werner Retschitzegger, Johannes Kepler University Linz, Austria
German Rigau, IXA NLP Group. EHU, Spain
Tarmo Robal, Tallinn University of Technology, Estonia
Sérgio Roberto da Silva, Universidade Estadual de Maringá, Brazil
Alejandro Rodríguez González, Universidad Carlos III de Madrid, Spain
Marco Rospocher, Fondazione Bruno Kessler (FBK), Italy
Thomas Roth-Berghofer, University of West London, U.K.
Michele Ruta, Politecnico di Bari, Italy
Gunter Saake, University of Magdeburg, Germany
Melike Sah, Trinity College Dublin, Ireland
Satya Sahoo, Case Western Reserve University, USA
Minoru Sasaki, Ibaraki University, Japan
Michael Schmidt, fluid Operations AG, Germany
Kinga Schumacher, German Research Center for Artificial Intelligence (DFKI) - Berlin, Germany
Wieland Schwinger, Johannes Kepler University Linz, Austria
Floriano Scioscia, Politecnico di Bari, Italy
Kunal Sengupta, Wright State University - Dayton, USA

Md. Sumon Shahriar, Tasmanian ICT Centre/CSIRO, Australia
Sofia Stamou, Ionian University, Greece
Vasco Soares, Instituto de Telecomunicações / Polytechnic Institute of Castelo Branco, Portugal
Mari Carmen Suárez-Figueroa, Universidad Politécnica de Madrid (UPM), Spain
Cui Tao, Mayo Clinic - Rochester, USA
Saïd Tazi, LAAS-CNRS, Université de Toulouse / Université Toulouse1, France
Andreas Textor, RheinMain University of Applied Sciences, Germany
Zoltán Theisz, evopro Innovation LLC, Hungary
Tina Tian, Manhattan College, U.S.A.
Tania Tudorache, Stanford University, USA
Christina Unger, CITEC - Bielefeld University, Germany
Roland Wagner, Johannes Kepler Universität Linz, Austria
Shenghui Wang, OCLC Leiden, The Netherlands
Wai Lok Woo, Newcastle University, UK
Sule Yildirim Yayilgan, Gjøvik University College, Norway
Fouad Zablith, American University of Beirut, Lebanon
Filip Zavoral, Charles University in Prague, Czech Republic
Yuting Zhao, The University of Aberdeen, UK
Hai-Tao Zheng, Tsinghua University, China
Amal Zouaq, Royal Military College of Canada, Canada

Copyright Information

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

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

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

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

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

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

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

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

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

Table of Contents

Parallel Search Through Statistical Semantic Spaces for Querying Big RDF Data

Alexey Cheptsov and Axel Tenschert
High Performance Computing Center Stuttgart,
University of Stuttgart
Stuttgart, Germany
e-mail: cheptsov@hlrs.de, tenschert@hlrs.de

***Abstract*—With billions of triples in the Linked Open Data cloud, which continues to grow exponentially, challenging tasks start to emerge related to the exploitation and reasoning of Web data. A considerable amount of work has been done in the area of using Information Retrieval (IR) methods to address these problems. However, although applied models work on the Web scale, they downgrade the semantics contained in an RDF graph by observing each physical resource as a 'bag of words (URIs/literals)'. Distributional statistic methods can address this problem by capturing the structure of the graph more efficiently. However, these methods are computationally expensive. In this paper, we describe the parallelization algorithm of one such method (Random Indexing) based on the Message-Passing Interface technology. Our evaluation results show super linear improvement.**

Keywords-Statistical Semantics; Random Indexing; Parallelization; High Performance Computing; Message-Passing Interface;JUNIPER.

I. INTRODUCTION

We live in a big data world, which is already estimated to be of the size of several Zetta (10^{21}) Bytes. However, the most considerable growth has seen the linked (open) data domain. Recent years have seen a tremendous increase of structured data on the Web with public sectors such as UK and USA governments opening their data to public (e.g., the U.S.'s *data.gov* initiative [1]), and encouraging others to build useful applications. At the same time, Linked Open Data (LOD) [2] project continues stimulating creation, publication and interlinking the RDF graphs with those already in the LOD cloud. In March 2009, around 4 billion statements were available in Resource Description Framework (RDF) format [3], while in September 2010 this number increased to 25 billion, and continues to grow every year exponentially. This massive amount of data requires effective exploitation and is now a big challenge not only because of the size but also due to the nature of this data. Firstly, due to the varying methodologies used to generate these RDF graphs there are inconsistencies, incompleteness, but also redundancies. These are partially addressed by approaches for assessing the quality, such as through tracking the provenance [4]. Secondly, even if the quality of the data would be at a high level, exploring and searching through large RDF graphs requires familiarity with the structure, and knowledge of the used ontology schema. Another challenge is reasoning over these vast amounts of data. The languages used for expressing formal semantics (RDF etc.) use the logic that does not scale to the amount of information and the setting that is required for the Web. The approach suggested by Fensel and van Harmelen [5] is to merge retrieval process and reasoning by means of selection or subsetting: selecting a subset of the RDF graph that is relevant to a query and sufficient for reasoning.

A considerable amount of work has been done in the area of using Information Retrieval (IR) methods for the task of selection and retrieval of RDF triples, and also for searching through them. The primary intention of these approaches is location of the RDF documents relevant to the given keyword and/or a Unified Resource Identifier (URI). These systems are semantic search engines such as Swoogle [6] or Sindice ([7], [8]). However, although these models work on the Web scale, they downgrade the semantics contained in an RDF graph by observing each physical resource as a 'bag of words (URIs/literals)'. More sophisticated IR models can capture the structure more efficiently by modelling meaning similarities between words through computing the distributional similarity over large amount of text. These are called statistical semantics methods and examples include Latent Semantic Analysis [9] and a more modern technique – Random Indexing, which is based on the vector space concept [10]. In order to compute similarities, these methods first generate a semantic space model. Both generating this model, and searching through it (e.g., using cosine similarity), are computationally expensive. The linear feature of searching through the large semantic space model is a huge bottleneck: for the model representing 300 million documents calculating cosine similarity in order to find similar terms can take as long as several hours, which is currently not acceptable for the problem domain specialists.

In this paper, we describe a parallelization approach for the Random Indexing search algorithm, suggested by Sahlgren [10]. We also discuss some techniques that allowed us to reduce the execution time down to seconds on the way to achieving a Web scale. The paper is structured as follows. In Section II, we present the use cases in which this work has been applied. An explicit description about the applied parallelization strategy and the modifications made to the Random Indexing algorithm are presented in Section III. Moreover, we give a thorough evaluation about the algorithm's performance and scalability on a distributed shared-memory system in Section IV. Finally, Section V presents conclusions and discusses main outcomes as well as future work directions.

II. USE CASES

In this section, we briefly describe two use cases that are taking advantage of the parallelization of the cosine similarity algorithm used by statistical semantics methods, which is the main topic of this paper. Cosine similarity [10] is a measure of similarity between two vectors of n dimensions, which is finding the cosine of the angle between them. If the cosine is zero, the documents represented by vectors are considered dissimilar, while one indicates a high similarity. We present the query expansion use case, which is used to improve the recall when searching, e.g., Linked Life Data (4 billion statements), followed by a subsetting scenario used to reduce the execution time when reasoning over the FactForge repository [11], which contained 2 billion statements at time of performing the experiment.

A. *Query Expansion*

Query expansion is used in Information Retrieval extensively with the aim to expand the document collection that is returned as a result to a query. This method employs several techniques, such as including lemmas and synonyms of the query terms, in order to improve precision and recall. It works by expanding the initial query thus covering larger portion of documents. In this context, finding synonyms is a very important step and one way to achieve this is by employing statistical semantics methods. These methods operate on a set of documents and therefore, we need to lexicalise an RDF graph in a way that will preserve the semantics and “relatedness” of each node with those in its neighbourhood, into an abstraction, which we call a virtual document.

In order to generate virtual documents from an RDF graph, we first select the relevant part of the original graph and subdivide it into a set of potentially overlapping subgraphs. The next step is lexicalisation in order to create virtual documents from these subgraphs. Finally, we generate the semantic index from the virtual documents. The details of how each of these steps is performed significantly influences the final vector space model. For example, in the selection and subdivision step, all or just a part of the ontology could be selected; the subgraphs could be individual triples, or RDF molecules (the set of triples sharing a specific subject node), or more complex/bigger subgraphs. In the lexicalisation step, the URIs, blank nodes, and literals from an RDF subgraph are converted to a sequence of terms. When generating the semantic index, different strategies for creating tokens and performing normalisation have to be applied to typed literals, string literals with language tags, and URIs.

Once the semantic index has been generated, it can be used to find similarities between URIs and literals. We use the ranked list of similar terms for URIs/literals that occur in certain kinds of SPARQL queries [5] to make the query more generic and also return results for entities that are semantically related to those used in the original query.

Thus, the application of query expansion through the use of statistical semantics method is feasible for those SPARQL queries that are not returning all relevant hits. In other words, query expansion here is aimed to improve recall, which is done by adding terms that are similar to the given ones in the original query.

B. *Subsetting*

For reasoning at web-scale, subsetting becomes a key, because most well-known reasoning algorithms can only operate on sets several orders of magnitude smaller than the Web. Getting subsetting algorithms to work is then of capital importance.

There is evidence that by sticking to smaller datasets, computer and cognitive scientists may be optimizing the wrong type of models. Basically, there is no warranty that the proven best performing model on thousands of entities is also the best performing model when datasets are four orders of magnitude larger [12].

III. BASICS OF PARALLEL RANDOM INDEXING ALGORITHM

Random Indexing and other similar algorithms can be broken down into two steps:

1) *generating a semantic index (vectors), and*
2) *searching the semantic index.*

Both parts are quite computationally expensive, however, the first part is a one-off step, which does not have to be repeated and the semantic index can be updated to follow changes in the documents if they happen. The second step, however, affects the end user, and therefore is a huge bottleneck for real-time applications. Hence, our focus is optimisation of the search part of the Random Indexing algorithm. Usually, search is performed over all vectors in the semantic index. Thereby the vectors are analysed independently of each other, i.e., in the arbitrary order.

This basically means that the search can be efficiently improved, when performed on several computing nodes in parallel instead of the “vector-by-vector” (i.e., sequential computation) processing in the current realisation. Practically, the whole vector space domain is decomposed into sub-domains each of which is processed in a separate block/program instance on a different machine. The division of the vectors between the blocks is defined by the domain decomposition [13] (Figure 1). Depending on the realisation, a synchronisation is required among the blocks, e.g., to collect the partial outputs of each block and produce the final result. Generally, the synchronisation step is expensive, and much attention should be paid to the correct implementation of the synchronisation in order to ensure the minimum overhead. In the next section, we describe the major parallelization strategies, enabling the full utilisation of multiple computing nodes as well as the optimal synchronisation between the distributed tasks.

Although a simple multi-threading approach would be extremely efficient in terms of the performance and easy in terms of the implementation efforts [14], it is not sufficient for achieving the Web-scale due to the limited number of CPU cores/nodes interconnected by a shared-memory bus in the currently available computing architectures (current shared-memory architectures offer a maximum of 8 to 16 interconnected cores).

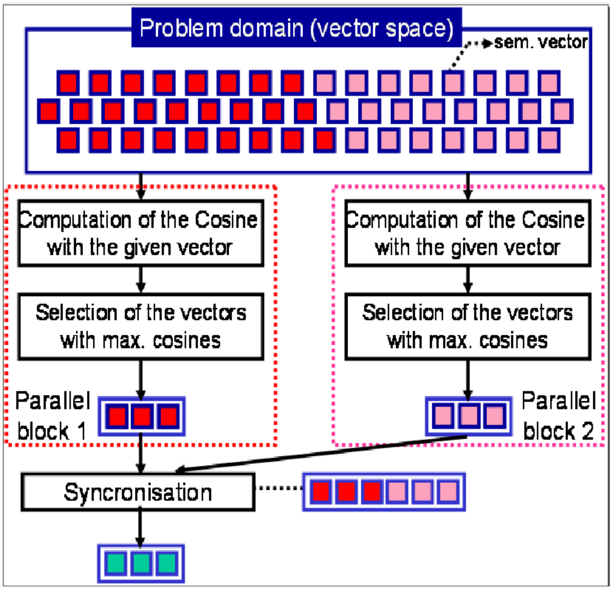

Problem domain (vector space)
sem. vector
Computation of the Cosine with the given vector
Selection of the vectors with max. cosines
Parallel block 1
Computation of the Cosine with the given vector
Selection of the vectors with max. cosines
Parallel block 2
Syncronisation

Figure 1. Domain decomposition based parallelisation of the Random Indexing algorithm.

Thus a distributed-memory parallelisation strategy is needed for Big Data. There are several parallelization strategies, differentiating in ways the synchronisation between the processes is implemented. The most promising for the Semantic Web in terms of performance gains are however the Message-Passing Interface (MPI) [15] and MapReduce [16].

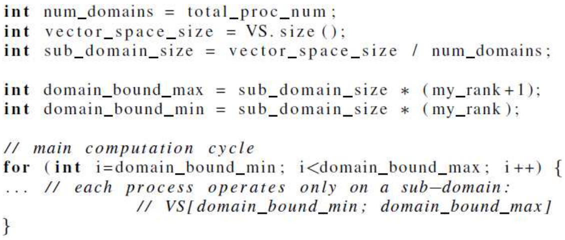
MPI is a wide-spread implementation standard for parallel applications, implemented in many programming languages, including Java. As the name suggests, the MPI processes communicate by means of the messages transmitted between two (a so called “point-to-point” communication) or among many (involving several or even all processes, i.e., a collective communication) compute nodes. Normally, one process is executed on a single computing node (however, the MPI standard does not limit the number of processes on one node). If any process needs to send/receive data to/from other processes, it calls a corresponding MPI function. Both point-to-point and collective communications available for MPI processes are documented in the MPI standard [15].

MapReduce is another popular framework for processing big datasets on certain kinds of distributable problems, originally introduced by Google [16] and currently followed by Yahoo in its Hadoop implementation. MapReduce is a promising parallelisation model for data centric applications. However it is quite restrictive with regard to the range of applications that it can be applied to. In this publication, we are focusing on practical aspect of applying the MPI-based distributed memory parallelization for the Random Indexing search algorithm. Due to the algorithmic complexity of splitting the execution workflow according to the map and reduction operation, the MapReduce-based approach [16] will be presented in a separate publication.

IV. Implementation With The Message-Passing Interface (MPI) and Evaluation

A. *Parallelisation of Airhead Search*

Airhead is an open source implementation of Random Indexing in the S-Space package by University of California [17]. Parallelization of the search operation in Airhead was performed by applying the domain decomposition to the semantic vector space, whereby number of domains corresponds to the number of computing nodes the application is running on. Thus, each process performs computation only on a part (sub-domain) of the vector space of size (m/n), where m is the size (dimensionality) of the vector space, and n is the number of processes (and sub-domains, accordingly). The boundary elements of the vector space to be computed by each process are calculated dynamically based on the process rank and the total number of processes provided by MPI, as demonstrated in Figure 2.



```
int num_domains = total_proc_num;
int vector_space_size = VS.size();
int sub_domain_size = vector_space_size / num_domains;

int domain_bound_max = sub_domain_size * (my_rank+1);
int domain_bound_min = sub_domain_size * (my_rank);

// main computation cycle
for (int i=domain_bound_min; i<domain_bound_max; i++) {
... // each process operates only on a sub-domain:
               // VS[domain_bound_min; domain_bound_max]
}
```

Figure 2. Specification of sub-domains. Each process calculates its respective sub-domain of the vector space based on its Rank and the number of processes in the group.

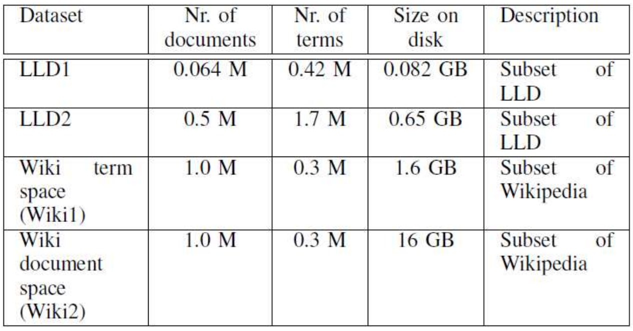
B. *Performance Evaluation on Cluster*

For the evaluation, a testbed based on the BW-Grid [18] cluster (Intel Xeon CPU architecture, 2 Quad-Core CPUs and 16 GB RAM per computing node), provided by the High Performance Computing Center Stuttgart, was used. Configuration of 1, 2, 4, 8, and 16 computing nodes were benchmarked to evaluate the scalability of the developed algorithms on the target architecture. In our tests, we mainly considered two different datasets coming from well-known semantic repositories (see test sets’ parameters in Table I):

a) Linked Life Data (LLD) repository: a large integrated repository, which contains over 4 billion RDF statements from various sources covering the biomedical domain. We investigated two subsets of the LLD that contain major terms (for pharmacological scenarios) and relations between them.

b) FactForge repository: contains schemata and ontologies from DBPedia, lingvoj, the CIA Factbook, Wordnet, Geonames, Freebase and musicbrainz. After full materialisation, it contains 404 million resources. We used the DBpedia/Wikipedia section of this space (we will refer to it as Wikipedia from now on, since they are parallel and have the same number of concepts at around 4 M). After filtering out redundant concepts, we kept only 1M documents. After some parameter exploration, we settled on n=1000. That is, the random vectors are 1000-dimensional.

TABLE I. BENCHMARKED DATASETS.



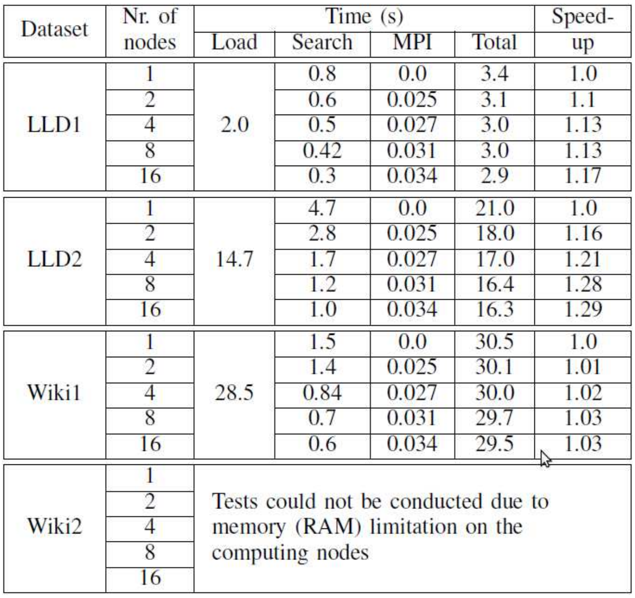
Dataset	Nr. of documents	Nr. of terms	Size on disk	Description
LLD1	0.064 M	0.42 M	0.082 GB	Subset of LLD
LLD2	0.5 M	1.7 M	0.65 GB	Subset of LLD
Wiki term space (Wiki1)	1.0 M	0.3 M	1.6 GB	Subset of Wikipedia
Wiki document space (Wiki2)	1.0 M	0.3 M	16 GB	Subset of Wikipedia

As the first step, we investigated the scalability and stability of the parallelised algorithm on the cluster, increasing the number of nodes involved in the computation, for different problem (dataset) sizes. The time for loading the datasets from the disk (i.e., the whole vector file has to be loaded into the memory of each node), the actual search operation as well as the overhead of the inter-node communication was in the focus of our measurements (Table II).

The evaluation reveals that our concern about the large impact for loading file from the disk time on the overall application performance was feasible. For all investigated use cases, the load time was considerably higher than the search time. Providing the bad scale of the load operation, the maximum speed-up achieved on 16 clusters computing nodes was only 1.29. Moreover, the experiments with the largest available semantic space (Wiki2) were impossible to be conducted due to exceeding the available RAM on the test bed. For the first test case, although the parallelization has been properly implemented, its usability for datasets with the large number of referenced documents and small amount of dependencies has not been proved. This is because the amount of computation for the search operation was relatively small as compared with the total execution time. Nevertheless, the second variant based on the split of datasets (Figure 1), demonstrated its value in terms of both performance and scalability for the diverse problem sizes (Table III). The LLD1 set has been excluded because of its small size.

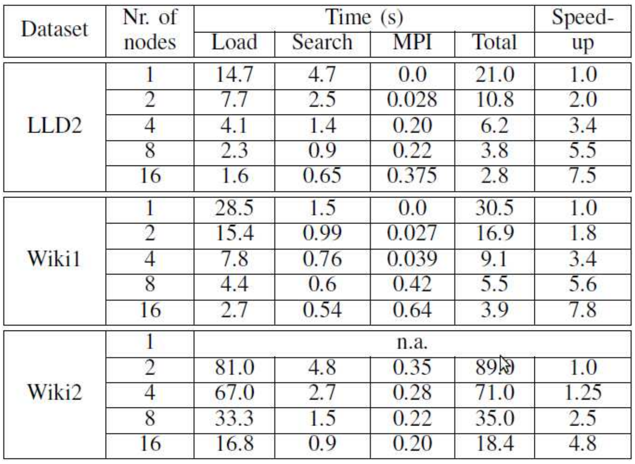
Despite the increasing communication overhead (caused by MPI operations), which is due to more complex communication pattern (as described in the previous publication [21]), the evaluation reveals a significant performance improvement for both load and search operations (see Figure 3). Generally, the use cases taking advantage of the dataset fragmentation show an improvement in time of approximately 85% (i.e., and average speed-up of approx. 7.0 has been achieved) over the non-parallel realisation. This clearly shows that our parallelization technique can be used to benefit Random Indexing applications significantly. Moreover, the technique facilitates applying Random Indexing for the datasets that have not been analysed before due to the limitations of non-parallel test beds.

TABLE II. PERFORMANCE CHARACTERISTICS GROUPED BY DATASET AND NUMBER OF COMPUTING NODES.



Dataset	Nr. of nodes	Time (s)				Speed-up
		Load	Search	MPI	Total	
LLD1	1	2.0	0.8	0.0	3.4	1.0
	2		0.6	0.025	3.1	1.1
	4		0.5	0.027	3.0	1.13
	8		0.42	0.031	3.0	1.13
	16		0.3	0.034	2.9	1.17
LLD2	1	14.7	4.7	0.0	21.0	1.0
	2		2.8	0.025	18.0	1.16
	4		1.7	0.027	17.0	1.21
	8		1.2	0.031	16.4	1.28
	16		1.0	0.034	16.3	1.29
Wiki1	1	28.5	1.5	0.0	30.5	1.0
	2		1.4	0.025	30.1	1.01
	4		0.84	0.027	30.0	1.02
	8		0.7	0.031	29.7	1.03
	16		0.6	0.034	29.5	1.03
Wiki2	1	Tests could not be conducted due to memory (RAM) limitation on the computing nodes				
	2					
	4					
	8					
	16					

TABLE III. PERFORMANCE CHARACTERISTICS FOR FRAGMENTED DATASETS, THE NUMBER OF FRAGMENTS CORRESPONDS TO THE NUMBER OF COMPUTING NODES.



Dataset	Nr. of nodes	Time (s)				Speed-up
		Load	Search	MPI	Total	
LLD2	1	14.7	4.7	0.0	21.0	1.0
	2	7.7	2.5	0.028	10.8	2.0
	4	4.1	1.4	0.20	6.2	3.4
	8	2.3	0.9	0.22	3.8	5.5
	16	1.6	0.65	0.375	2.8	7.5
Wiki1	1	28.5	1.5	0.0	30.5	1.0
	2	15.4	0.99	0.027	16.9	1.8
	4	7.8	0.76	0.039	9.1	3.4
	8	4.4	0.6	0.42	5.5	5.6
	16	2.7	0.54	0.64	3.9	7.8
Wiki2	1	n.a.				
	2	81.0	4.8	0.35	89.0	1.0
	4	67.0	2.7	0.28	71.0	1.25
	8	33.3	1.5	0.22	35.0	2.5
	16	16.8	0.9	0.20	18.4	4.8

C. Discussion and Future Directions

As described in the previous section, the performance of the complex search algorithm greatly benefits from the "correct" implementation of the corresponding parallelization paradigm. Correct, here, does not solely mean that MPI has been successfully applied to the Random Indexing search algorithm in order to enable usage of large shared-memory systems, but rather that the algorithm itself has been modified in order to obtain highest performance and scalability - the concept of domain decomposition [13] has been applied to the algorithm to allow the processing

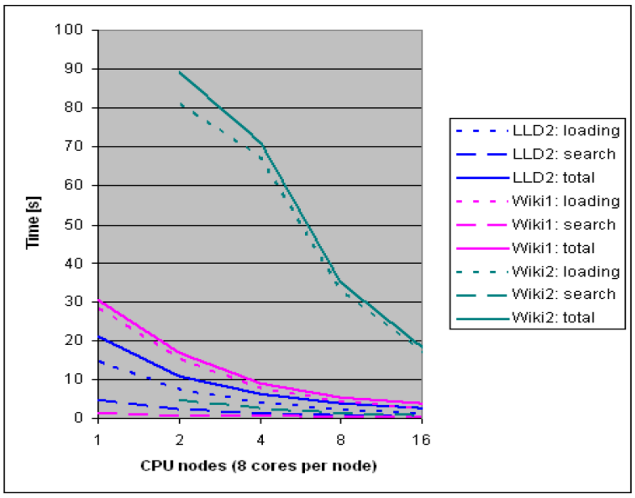
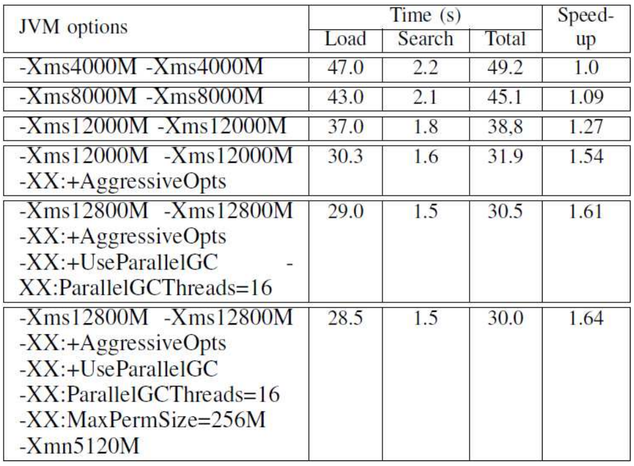


Figure 3. Performance results for decomposed datasets.

of large vector space files (>= 16 GB) and (2) to obtain scalable computation through processing (i.e., the search and in particular the load operation) of smaller subsets of the vector space file concurrently, i.e., distributing the processing to multiple nodes. However, there are other factors, which do influence the overall performance of a parallel application, too.

Developers can also tune their applications at runtime by using advanced settings for the Java Virtual Machine (JVM) [19]. For this reason, we have also experimented with different settings for the JVM during our tests. We have performed 30 runs of the parallel Airhead search using 6 different JVM settings (each setting has been repeated 5 times) to estimate the optimal configuration for our machines. Due to the fact that all our machines are equipped with equivalent hardware and software, the explicit tests were solely carried out on one particular node with the assumption that the settings are optimal for all other nodes within the cluster, too. A summary of our test runs and the speed-ups achieved is provided in the Table IV.

TABLE IV. PERFORMANCE RESULTS FOR PARALLEL AIRHEAD SEARCH ALGORITHMWITH VARYING JVM SETTING ON WIKI1 DATASET.



JVM options	Time (s)			Speed-up
	Load	Search	Total	
-Xms4000M -Xms4000M	47.0	2.2	49.2	1.0
-Xms8000M -Xms8000M	43.0	2.1	45.1	1.09
-Xms12000M -Xms12000M	37.0	1.8	38,8	1.27
-Xms12000M -Xms12000M -XX:+AggressiveOpts	30.3	1.6	31.9	1.54
-Xms12800M -Xms12800M -XX:+AggressiveOpts -XX:+UseParallelGC - XX:ParallelGCThreads=16	29.0	1.5	30.5	1.61
-Xms12800M -Xms12800M -XX:+AggressiveOpts -XX:+UseParallelGC -XX:ParallelGCThreads=16 -XX:MaxPermSize=256M -Xmn5120M	28.5	1.5	30.0	1.64

As shown in Table IV, a properly configured JVM significantly improves the overall performance of an application. In our scenario, we could optimize the performance of the parallelized application for approximately 40% using the proper JVM settings. Based on these tests, we were also able to determine the best suited JVM configuration for our environment as well as learned more about the optimal values for individual JVM parameters(e.g., the total amount of heap size, the number of threads used for garbage collection, etc.), which can be used for other Java applications as well.

An alternative promising approach is suggested by the JUNIPER (”Java platform for hIgh PErformance and Realtime large scale data management”) project [20]. JUNIPER is an EU-FP7 project that aims to establish a development platform for new-generation data-demanding applications. The JUNIPER approach is to exploit synergies between all major parallelization technologies (such as MPI, MapReduce, COMPSs, etc.) and elaborate new paradigms in data centric parallel processing that will balance flexibility and performance of data processing applications in heterogeneous computing architectures. A possibility to combine diverse parallelization technology within a single application, as offered by JUNIPER, would also be of a huge advantage for Random Indexing algorithms, e.g., to implement the semantic space generation with MapReduce and the search with MPI. In our following research, we are going to investigate the benefits of this “heterogeneous” approach for the Airhead package.

V. CONCLUSION AND FUTURE WORK

This paper presented an evaluation of our approaches to parallelize the Airhead library for Random Indexing, which can be used to significantly improve information retrieval methods, in particular those that use the cosine similarity for searching a large vector space model. We use an effective parallel programming paradigm, namely MPI, to exploit parallelism for the RI algorithm in order to take advantage of large-scale distributed shared-memory systems and thus to improve its performance. We evaluated the parallelized algorithm on different hardware and software configurations (i.e., we varied the amount of computational nodes as well as the input datasets) with promising results. The algorithm improves performance in all of the presented experiments. However, if each process (i.e., node) has to load the full dataset at once, the overall speed-up is relatively small and the algorithm does not scale very well while increasing the number of machines. For this reason, we implemented a way to split the input dataset into smaller junks, which can be independently and concurrently processed by each node. This feature significantly decreased the processing time of the load operation and thus improved the overall performance. Moreover, we are now able to process datasets with billions of statements because we are not directly limited by the system’s memory anymore. In addition, we experimented with different Java Virtual Machine settings in order to optimize and fine-tune the application for the given runtime environment. Most importantly, these results suggest that we need both parallel algorithms and Java Virtual

Machine optimizations to effectively utilize machines (not necessarily parallel systems but any common personal computer) for our Semantic Web applications. Finally, we demonstrated the effectiveness of the parallelized algorithm and its usage and benefits within an interesting for biomedical domain application scenario. In the future, we will investigate further possibilities to optimize our code (e.g., using a different MPI implementation for Java) as well as compare the actual MPI-based parallelization with the MapReduce implementation. In particular, methods to cross-fertilize the advantages of diverse programming models in a common application workflow will be explored. Data modelling techniques, investigated by the JUNIPER platform [22], will be the major technology to enable such across-fertilization of the parallelization technologies. The new technologies that JUNIPER works out will be applied to the most challenging Big Data domains, in particular to Semantic Web.

ACKNOWLEDGMENT

This research has been supported in part by the EUFP7 project JUNIPER. We would like to thank the main developer of the Airhead library - David Jurgens - for offering a use case and his great support.

REFERENCES

[1] U.S.'s data.gov intiative website. [Online]. http://www.data.gov/. [retrieved: April, 2013].

[2] Linked Data project website. [Online]. http://linkeddata.org. [retrieved: April, 2013].

[3] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data - the story so far", International Journal on Semantic Web and Information Systems (IJSWIS), 5(3), 2009, pp. 1-22.

[4] O. Hartig and J. Zhao, "Using web data provenance for quality assessment", in Int. Workshop on Semantic Web and Provenance Management, Washington D.C., USA, 2009, pp. 29-34.

[5] D. Fensel and F. van Harmelen, "Unifying reasoning and search to web scale", IEEE Internet Computing, vol. 11(2), 2007, pp. 95-96.

[6] L. Ding et al., "Swoogle: a search and metadata engine for the semantic web", in Proc. the thirteenth ACM international conference on Information and knowledge management CIKM '04, New York, NY, USA, 2004, pp. 652-659.

[7] G. Tummarello, R. Delbru, and E. Oren, "Sindice.com: Weaving the open linked data", in Proc. the 6th International Semantic Web Conference, Busan, Korea, 2007, pp. 552-565.

[8] E. Oren et al., "Sindice.com: A document-oriented lookup index for open linked data", International Journal of Metadata, Semantics and Ontologies, vol. 3, 2008, pp. 37-52.

[9] T. K. Landauer, P. W. Foltz, and D. Laham, "Introduction to latent semantic analysis", Discourse Processes, vol. 25, 1998, pp. 259-284.

[10] M. Sahlgren, "An introduction to random indexing", in Proc. Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE 2005, 2005, pp. 1-8.

[11] Fact Forge semantic repository website. [Online]. http://factforge.net/. [retrieved: April, 2013].

[12] A. Halevy, P. Norvig, and F. Pereira, "The unreasonable effectiveness of data", IEEE Intelligent Systems, vol. 24, 2009, pp. 8-12.

[13] T. F. Chan and T. P. Mathew, "Domain decomposition algorithms", Acta Numerica, vol. 3, 1994, pp. 61-143.

[14] S. Akhter and J. Roberts, "Multi-core programming: Increasing performance through software multi-threading", Intel Press, Santa Clara, Tech. Rep., 2006.

[15] W. Gropp and A. S. E. Lusk, Eds., Using MPI: Portable Parallel Programming with the Message-Passing Interface. Cambridge: MIT Press, 1994.

[16] R. Lammel, "Google's mapreduce programming model - revisited", Science of Computer Programming, vol. 70,1, 2008, pp. 1-30.

[17] D. Jurgens, "The S-Space package: An open source package for word space models", in Proc. the ACL 2010 System Demonstrations, 2010, pp. 30–35.

[18] High Performance Computing Center Stuttgart's BW-Grid cluster description. [Online]. https://wickie.hlrs.de/dgrid/index.php/Hardware. [retrieved: April, 2013].

[19] J. Shirazi, Ed., Java Performance Tuning. Sebastopol: O'Reilly & Associates, Inc, 2002.

[20] Juniper project website. [Online]. Available: http://juniperproject.org. [retrieved: April, 2013].

[21] A. Cheptsov and M. Assel, "Distributed Parallelization of Semantic Web Java Applications by Means of the Message-Passing Interface", M. Resch et al. (eds.), High Performance Computing on Vector Systems 2011, Springer Verlag Berlin Hedelberg 2012, pp. 51-64.

[22] A. Cheptsov and B. Koller, "JUNIPER takes aim at Big Data", inSiDE - Journal of Innovatives Supercomputing in Deutschland, vol. 11, No. 1, Spring 2011, pp. 68-69.

Recovery of Temporal Expressions From Text: The RISO-TT Approach

Adriano Araújo Santos
Programa de Pós-Graduação da Universidade Federal de Campina Grande - UFCG
Faculdade de Ciências Sociais Aplicadas - FACISA
Universidade Estadual da Paraíba - UEPB, Paraíba, Brazil
adriano@copin.ufcg.edu.br

Ulrich Schiel
Departamento de Sistemas e Computação – Universidade Federal de Campina Grande - UFCG
Campina Grande, Paraíba, Brazil
ulrich@computacao.ufcg.edu.br

***Abstract*— The necessity of managing the large amount of digital documents existing nowadays, associated to the human inability to analyze all this information in a fast manner, led to a growth of research in the area of development of systems for automation of the information management process. Nevertheless, this is not a trivial task. Most of the available documents do not have a standardized structure, hindering the development of computational schemes that can automate the analysis of information, thus requiring jobs of information conversion from natural language to structured information. For such, syntactic, temporal and spatial pattern recognition tasks are needed. Concerning the present study, the main objective is to create an advanced temporal pattern recognition mechanism. We created a rules dictionary of temporal patterns, developing a module with an extendable and flexible architecture for retrieval and marking. This module, called RISO-TT, implements this pattern recognition mechanism and is part of the RISO project (Retrieval of Information with Semantics of Contexts). Two experiments were carried out in order to evaluate the efficiency of the approach. The first one was intended to verify the extendibility and flexibility of the RISO-TT architecture and the second one analyses the efficiency of the approach, based on a comparison between the developed module and two consolidated tools in the academic community (Heidelime and SUTime). RISO-TT outperformed the rivals in the temporal expression marking process, which was proved through statistical tests.**

Keywords-temporal expressions extractor; temporal pattern recognition; natural language processing

I. INTRODUCTION

The large amount of digital documents existing nowadays, resulting from the unconstrained access and freedom to publish given by the Web to people [1] defeated the human capability to analyze all existing information. So, there is an increasingly need for automating the access, research, and management of information in order to generate valuable sources of knowledge [9].

Computers, in turn, can process structured or semi-structured information. Nevertheless, since most of the available information are non-structured [9], the present challenge is to allow computers to process information in natural language by converting this natural language to structured information, thus allowing a higher level of automation of computational processes.

So, Natural Language Processing (NLP) emerges as a possible solution for this challenge, because it is characterized as a set of computational techniques for analysis of texts in one or more linguistic levels, with the purpose of simulating the human linguistic processing. Among such techniques, there is the Recognition of Mentioned Entities (RME) [4], which aims to locate and classify atomic elements of a text, according to a pre-defined set of categories.

Information Extraction (IE) is the task of retrieving information from large volumes of documents or texts, structured or free [17]. For Zambenedetti [17], a well-developed information extraction technology allows the rapid development of extraction systems for new tasks that would have the same performance level of tasks performed by humans, a level not reached yet.

This paper addresses the process of extracting temporal expressions from texts, which is an activity that has become a significant research and development field in Computer Science, motivated by the large number of applications that explore temporal information extracted from texts. As examples of such applications we may cite Geographic Information Systems, automatic question, and answer applications and text summarization systems. With the use of temporal expression extraction techniques, applications perform tasks in a higher automation level [12].

There are several approaches for the recognition of temporal expressions in texts. Saquete [12] enumerates as main approaches: a) rules-based; b) Machine Learning and c) Combining rules, and machine learning. Regardless of the adopted approach, the output is a scheme of standardized temporal annotations. The schemes TIDES 2005 (Translingual Information Detection, Extraction, and Summarization) [11] and TimeML [14] are the most adopted.

Considering some limitations of the evaluated existing tools, we developed a new temporal expressions extractor, Retrieval of Semantic Information from Textual Objects Temporal Tagger (RISO-TT), as part of a project called RISO (Retrieval of Information with Semantics of Contexts) [18].

We carried out an experiment to prove the extensibility and flexibility of the system, as well as to check whether RISO-TT presents some competitive advantage and brings any contribution for this research area.

In the following section other approaches of temporal expressions extraction are presented. Section III introduces the RISO-TT approach to temporal expressions extraction and after that, the results of the proposal are analyzed on section IV Verification and Validation. Contributions and future work are discussed in the final section V Conclusions.

II. Related works

Developed at the Center for Computational Language and Education Research, University of Colorado, the ATEL (Automatic Temporal Expression Labeler) [5] adopts a statistical approach to detect temporal expression in English and Chinese languages.

The system used a training database made available by TERN (Time Expression Recognition and Normalization) with a set of temporal terms and, for each sentence found in a processed document, the term is marked with tags.

ITC-IRST, Centro per la Ricerca Scientifica e Tecnológica, Povo, Italy, has developed Chronos System [10]. It is a rules-based approach, separating the identification of temporal expressions into recognition (detection) and interpretation of values (normalization). Chronos is based on linguistic analysis (tokenization, tagging and pattern recognition).

Another system, TempEx, has been eveloped by MITRE Corporation with a Perl application for recognition and interpretation of temporal expressions according to the TIMEX2 2001 specifications, TempEx is characterized as one of the firsts of this kind [8].

TempEx is able to recognize absolute times (E.g., March 15th, 2013) and relative ones (e.g., born after World War II), and the computation of the normalization is based on the publication date of the document, which means that the algorithm uses meta-information from the very document to compute the normalization of relative times [8].

Developed by the University of Georgetown, GUTime [16] is an extension of the TempEx tagger [8], recognizing and normalizing temporal expressions in TIMEX3 standard.

An important feature of this system is that it enables shifting temporal expressions, causing computations to be performed with basis on an input date [8]. GUTime has incorporated a set of ACE TIMEX2 expressions, including duration, a variety of temporal modifiers and European date formats [16].

The DANTE (Detection and Normalization of Temporal Expressions) system [9] has a modular architecture which consists, basically, of two modules: recognition and interpretation.

The temporal expression recognition module was developed by using of the JAPE grammar [Cunningham et al. 1999] which consists of a set of <condition, action> rules

The interpretation module scans sentence by sentence a document, searching for patterns that match a pre-defined one (knowledge base).

TERSEO (Temporal Expression Recognition System applied to Event Ordering) was developed by the Research Group on Natural Language Processing and Information Systems, University of Alicante. The system generates annotations in TIMEX2 standard.

At first, according to Saquete [12], TERSEO was developed as a knowledge base system, intended for automatic recognition and normalization of temporal expressions in Spanish texts. It uses the translation of the temporal expressions to temporal models already defined in the first version to obtain, automatically, the temporal expressions from other languages [12].

TIPSem [7] deals with six different tasks related to the treatment of multilingual temporal information proposed by TempEval-2, (Evaluating Events, Time Expressions, and Temporal Relations). These tasks are classified as A, B, C, D, E and F, where the task A consists in defining temporal extensions, task B consists in classifying the events defined by TimeML (Markup Language for Temporal and Event Expressions) and the remaining task are related to categorization of different temporal links.

Heideltime [13] is a rule-based system based intended to extract and normalize temporal expressions in several languages. It uses the TIMEX3 annotation standard, and there are, presently, versions in English and German languages. The marking of temporal expressions in HeidelTime depends on the domain where the documents are inserted, such as news, reports, colloquial, or science (e.g., biomedical studies).

The SUTime is a library for recognizing and normalizing temporal expressions developed by Stanford University. It is a system developed in Java and based on deterministic rules designed to be extensible.

In its development was used TokensRegex framework, a generic framework for defining patterns on the text and mapping of semantic objects and makes use of regular expressions for the recognition of temporal expressions [3].

For the analyzed tools, we also observed that most of the temporal expression extraction tools are neither flexible nor extensible, and the recognition of compound temporal expressions or not allowed.

III. RISO TEMPORAL TAGGER (RISO-TT)

RISO-TT is the temporal expression extractor of the RISO Project (acronym, in Portuguese, of Semantic Information Retrieval from Textual Objects). It differs from the other temporal extraction tools for considering more complex signs and grammatical associations in the process of identifying temporal expressions.

The complex temporal expressions considered result from grammatical associations, which determine time intervals and not just temporal tokens. Compound temporal expressions are more accurate because they allow the specific understanding of the time expressed in the text.

A compound temporal expression is a structure formed by closed intervals (e.g., *from the beginning of January 10th to July 20th*), or semi-open intervals (*since 1968*). Also grammatical associations formed by prepositions, adverbs, numbers, and temporal tokens are considered, such as *in December*, or *every day*, and several other relation of terms in a semantic temporal expression.

To exemplify a compound temporal expression and to show why is it necessary to be identified, consider the expression "*from December 10, 2011 to December 10, 2012*". This sentence refers to a specific period which can be represented by 12/10/2011 < X < 12/10/2012, where X is the temporal variable referred by the expression.

RISO-TT does not depend on fixed standards and third-party software in its architecture. Both cases can cause problems in the future, since standards evolve, and a considerable architectural change can lead to serious development problems.

A. Architecture

RISO-TT is a rules-based system and since it was conceived to become extensible and flexible it uses a configuration file which determines the connection from the internal logic to the rules-base. To insert a new standard (or rule) in the rules base, the new standard is simply added to the configuration files. For this change to be realized, it just needs that RISO-TT be run again with a document that contains, in its content, the corresponding expressions. Figure 1 presents the RISO-TT architecture.

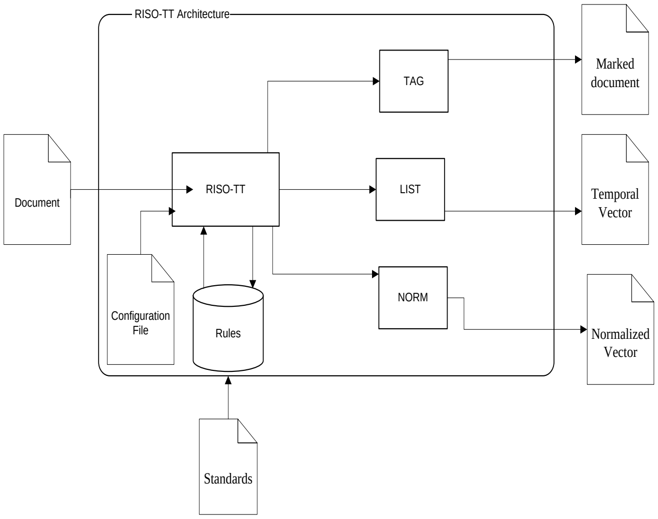


Figure 1. RISO-TT architecture.

A standard is a sequence of (temporal or grammatical) semantically associated terms, which can assign a value to a temporal expression. Typical standards are: preposition, adverbs, seasons of the year, dates, hours, and regular expressions.

A rule is an ordered sequence of standards (temporal and/or nominal) that characterizes the formation of temporal expressions. A rule takes into consideration the position of terms that form an expression. For example, the rule *Day Month Year* is different from the rule *Month Day Year*.

With the use of rules, the temporal standards were extended in such a way that complex structures among the grammatical relations and classical temporal expressions can be recognized as a single expression. With this, expressions like "*from December 10, 2011 to December 10, 2012*" are classified as a single temporal expression, and not as two independent temporal tokens.

B. Processed Outputs

The processing of a document in RISO-TT generates three documents:

- Marked document (TAG): the document given as input generates a document marked with the RISOTime tags and type attributes (e.g., <RISOTime type=Pre-EBT>On September 1, 1939</RISOTime>). The value assigned to the type attribute is the name of the rule of which the expression found is part.
- Temporal Vector (LIST): a document with a list of the temporal expressions found in the document (e.g., EBT-N -> from 499 to 493 BC) is created.
- Normalized Vector (NORM): Another document with a list of the temporal expressions found in the input document and their normalized values (e.g., On September 1, 1939 <--> 1-09-1939).

C. Example of Document Processing in RISO-TT

The following sentence is part of the document "16_SpanishCivilWar" from WikiWars:

... On 7 March, the Nationalists launched the Aragon Offensive. By 14 April they had pushed through to the Mediterranean, cutting the Republican-held portion of Spain in two. The Republican government tried to sue for peace in May, but Franco demanded unconditional surrender; the war raged on. In July, the Nationalist army pressed southward from Teruel and south along the coast toward the capital of the Republic at Valencia but was halted in heavy fighting along the XYZ Line, a system of fortifications defending Valencia. The Republican government then launched an all-out campaign to reconnect their territory in the Battle of the Ebro, from 24 July until 26 November.

We can find many types of temporal expressions in these sentences and simple expressions such as *On 7 March* are found by common Temporal Expression Extractors. But some types of temporal expression are more complex (e.g., *from 24 July until 26 November*). Theses sentences we called of compound temporal expressions.

The RISO-TT finds simple and compound temporal expressions and, if an expression is not detected, a corresponding rule identifying this kind of expressions can be added to the rules base.

The example text above generates the following marked document to by the function TAG:

<RISOTime type=Pre-EBT>On 7 March</RISOTime>, the Nationalists launched the Aragon Offensive. <RISOTime type=Pre-EBT>By 14 April</RISOTime>, they had pushed through to the Mediterranean, cutting the Republican-held portion of Spain in two. The Republican government tried to sue for peace <RISOTime type=Pre-EBT>in May </RISOTime>, but Franco demanded unconditional surrender; the war raged on. <RISOTime type=Pre-EBT>In July</RISOTime>, the Nationalist army pressed southward from Teruel and south along the coast toward the capital of the Republic at Valencia but was halted in heavy fighting along the XYZ Line, a system of fortifications defending Valencia. The Republican government then launched an all-out campaign to reconnect their territory in the Battle of the Ebro, <RISOTime type=I>from 24 July until 26 November</RISOTime>.

where 'type=Pre-EBT' means the association of a *Preposition* and a *Basic Temporal Structure*. "type=I**"** is the Intervals rule.

The function LIST applied to the text generates:
Pre-EBT -> On 7 March
Pre-EBT -> By 14 April
Pre-EBT -> in May
Pre-EBT -> In July
I -> from 24 July until 26 November

Finally the normalization of these expressions gives
On 7 March <--> 7-03-XXXX
By 14 April <--> 14-04-XXXX
in May <--> Pattern not identified yet
In July <--> Pattern not identified yet
from 24 July until 26 November <--> 24-07<X<26-11

where:
XXXX: is the unknown year.

IV. VERIFICATION AND VALIDATION

A. Verification

The verification process was responsible for answering the question about extensibility and flexibility of RISO-TT, which asks: "*Is the model of the RISO Temporal Tagger flexible and extensible?*".

To answer this question, we carried out three tests with the WikiWars corpus, with three different rules configurations, where the adjustment made in each version was based on patterns not found in the previous version.

To exemplify this process, imagine that a document *d* has a set of Temporal Expressions (TE) and that this document was marked by a Temporal Tagger (TT), resulting in a document *d'*. This document d' is composed of a set of temporal marks defined as $TM = \{m_1, m_2, .., m_n\}$, where each m_i is an expression marked with basis in the set of temporal rules $R = \{p_1, p_2, ..., p_n\}$. A temporal standard p_i is formed by a set of temporal expressions. In this case, $TM \subseteq TE$

Analyzing document *d'*, we noticed that there are temporal expressions *E'* that were not marked by *TT*; and this occurs because the rule *p* that is able to identify a temporal expressions *E'* does not belong to the set *R* of rules. That is, $p \notin R$.

Once the new rule has been inserted in the Rules file and the Configuration file has been updated, a new test has been carried out with R'=R ∪ {p}, obtaining a new *d''* determining *MT'* such that MT ⊆ MT'.

B. Validation

In order to validate the development of RISO-TT and analyze its performance, we realized a comparative experiment. We selected two temporal marking tools for this comparison: Heideltime and SUTime. The tests were performed and the results computed and compared with the ideal markings mold, made available by WikiWars.

The WikiWars corpus has a mold with the temporal markings that exist in all documents that compose it. Based on these documents, the results of the markings by the tools chosen for experiment were compared and evaluated according to the number of correctly marked expressions, the missing expressions and those incorrectly marked.

Based on the information found, we computed the Accuracy, Coverage and F-Measure of the samples. The results are presented in Table III.

C. Data Analysis

Right after tests, the first task carried out was the evaluation of normality of the resulting data. For this activity, we ran the Shapiro-Wilk test, obtaining the results displayed in Table I.

TABLE I. SHAPIRO-WILK NORMALITY TESTS

	W	P-value
Heideltime	0.652	0,00508
SuTime	0.942	0.2176
RisoTT	0.9831	0.9574

We notice that the p-value obtained from the Heideltime data characterizes a non-normalized sample. However, evaluating the data, we noticed that they concern three samples which, in the document marking process, had not a good result. The format in question is for dates with three characters (e.g., 200 AD). This format was not recognized by Heideltime and, so, due to the non-

Copyright (c) IARIA, 2013. ISBN: 978-1-61208-293-6

expressive number of documents, will be considered as outlier in the research.

We performed the statistical test based on the trust interval of 95%. For this, we computed the sample mean, standard deviation, and standard error. Based on this, the results found were presented in Table II.

TABLE II. TRUST INTERVALS

	Average Sample – Standard Error	Average Sample	Average Sample + Standard Error
Heideltime	0.7187674	0.7792521	0.8397369
SuTime	0.7566779	0.7842650	0.8118521
RisoTT	0.8970079	0.9139180	0.9308280

Based on the BoxPlot presented in Figure 2, it is possible to conclude that we cannot state whether there is superiority of one of the tools Heideltime and SUTime, since the trust intervals intersect each other. This occurs maybe due to the fact that both tools use the same marking standard and are based only on the knowledge based available in the standard. However, it is possible to state that there is a statistically proved superiority of the results of RISO-TT, compared to the others.

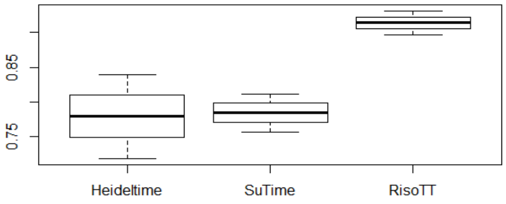


Figure 2. BoxPlot of the Trust Interval

We believe that this superiority is due to the number of relations between the mapped temporal standards and their relations, defined by the rules, in RISO-TT. The processing of information, due to this number of relations, is slower than the other tools and this may cause the making of temporal expressions to be more detailed than the others.

V. CONCLUSION AND FUTURE WORK

The development of RISO-TT is part of the RISO [18] research project and the experiments proved that it is extensible and flexible, and its performance was superior of other existing approaches.

A. *Contributions*

Considering the information presented throughout this paper, the main contributions of RISO-TT are:

- Flexible and extensible architecture based on standards and rules configurable by means of XML files;
- Independence of third-party software;
- Temporal Expression Recognition based on rules priority analysis;
- Possibility of creating complex structures of temporal and grammatical associations;
- Extends the standards with the possibility of arrangements and associations with other non-temporal expressions;
- Normalization of complex temporal standards, taking intervals between temporal tokens into consideration.

As future work we list:

- Concerning the temporal expression normalization process, incomplete time of an expression in a sentence may be completed by metadata about the document or other times of the current phrase or paragraph.
- The temporal expressions recognizer could be integrated with a spatial expressions recognizer.
- Recognition and treatment of ambiguities of the temporal expressions found in the document (e.g. May (Mouth) or may (Verb)).

With these documents indexing procedures integrated in the RISO project a Semantic Query Processor is under development to take into account this rich indexing structure of documents in order to optimize the quality of the information retrieval process.

B. *Model Packaging*

The RISO-TT Project is available in the RISO website [18].

REFERENCES

[1] R. Baeza-Yates and B. Ribiero-Neto. “Modern Information Retrieval”. Boston: Addison-Wesley Longman,1999.

[2] H. Cunningham. “JAPE: a Java Annotation Patterns Engine”. Research Memorandum CS–99–06, Department of Computer Science, University of Sheffield, May, 1999.

[3] A. X. Chang and C. D. Manning, “SUTIME: A Library for Recognizing and Normalizing Time Expressions”. 8th International Conference on Language Resources and Evaluation (LREC 2012), 2012.

[4] E. Ferneda. “Processamento da linguagem natural”. Available in: <http://www.marilia.unesp.br/Home/Instituicao/Docentes/EdbertoFerneda/MRI-06%20-

%20Processamento%20da%20Liguagem%20Natural.pdf>. Accessed in: 12 April. 2012.

[5] K. Hacioglu,Y. Chen and B. Douglas, "Automatic time expression labeling for english and chinese text". In: GELBUKH, A. F. Computational Linguistics and Intelligent Text Processing, 6th International Conference, CICLing'05, Lecture Notes in Computer Science, Mexico City, Mexico, February. Springer, 2005, pp. 548–559.

[6] H. S. Llorens. "A Semantic Approach to Temporal Information Processing (PhD Dissertation)" - University of Alicante, Departamento de Lenguajes y Sistemas Informáticos, Alicante, 2011.

[7] H. S., E. Llorens and B. Navarro. "TIPSem (English and Spanish): Evaluating CRFs and Semantic Roles in TempEval-2". In: Proceedings of the 5th International Workshop on Semantic Evaluation, Association for Computational Linguistics - ACL 2010, Uppsala, Sweden, 15-16 July, 2010, pp. 284–291.

[8] I. Mani, G. Wilson, B. Sundheim, andL. Ferro. "Guidelines for Annotating Temporal Information". In Proceedings of HLT 2001, First International Conference on Human Language Technology Research, J. Allan, ed., Morgan Kaufmann, San Francisco, 2001.

[9] P. Mazur and R. Dale. "WikiWars: A New Corpus for Research on Temporal Expressions". In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, MIT, Massachusetts, USA, 9-11 October, 2010, pp. 913–922.

[10] M. Negri and L. Marseglia. "Recognition and Normalization of Time Expressions: ITC-irst ar TERN 2004 MEANING - Developing Multilingual Web-scale Language Technologies", 2004.

[11] J. Pustejovsky et al. "TimeML:Robust Specification of Event and Temporal Expressions in Text". In IWCS-5, Fifth International Workshop on Computational Semantics, Tilburg, The Netherlands, January, 2003.

[12] E. Saquete. "ID 392:TERSEO + T2T3 Transducer. A System for Recognizing and Normalizing TIMEX3". In: Proceedings of the 5th International Workshop on Semantic Evaluation, 2010.

[13] J. Strotgen and M. Gertz, "Heideltime: High Quality Rule-based Extraction and Normalization of Temporal Expressions". In: Proceedings of the 5th International Workshop on Semantic Evaluation, Association for Computational Linguistics - ACL 2010, Uppsala, Sweden, July 15-16, 2010, pp. 321–324.

[14] Timeml Working Group. "Guidelines for Temporal Expression Annotation for English for TempEval 2010". 2010.

[15] M. Verhagen. "Temporal closure in an annotation environment". Language Resources and Evaluation, no. 39, , May, 2005, pp. 211–241.

[16] M. Verhagen et al. "Automating temporal annotation with TARSQI". In: Proceedings of the 43th Annual Meeting of the Association for Computational Linguistics, Ann Arbor, USA, 2005.

[17] C. Zambenedetti. "Extração de Informações sobre Bases de Dados Textuais". 2002. 144 f. Dissertação (Mestrado) - Universidade Federal do Rio Grande do Sul, Porto Alegre, 2002.

[18] RISO. Avaliable in: https://sites.google.com/a/copin.ufcg.edu.br/riso-t/projeto>. [retrieved: July, 2013]

TABLE III. ACCURACY, COVERAGE AND F-MEASURE RESULTS

	Accuracy			Coverage			F-Measure		
Documents	**Heideltime**	**Sutime**	**Riso-TT**	**Heideltime**	**Sutime**	**Riso-TT**	**Heideltime**	**Sutime**	**Riso-TT**
01_WW2	0,86227545	0,8882	0,994	0,847059	0,841176	0,970588	0,854599	0,864048	0,982143
02_WW1	0,89177489	0,85	0,984	0,777358	0,769811	0,958491	0,830645	0,807921	0,971319
03_AmCivWar	0,85135135	0,7857	0,911	0,84	0,733333	0,96	0,845638	0,758621	0,935065
04_AmRevWar	0,86896552	0,8601	0,946	0,857143	0,836735	0,959184	0,863014	0,848276	0,952703
05_VietnamWar	0,84729064	0,8191	0,938	0,702041	0,665306	0,865306	0,767857	0,734234	0,900212
06_KoreanWar	0,87301587	0,7724	0,963	0,738255	0,637584	0,865772	0,8	0,698529	0,911661
07_IraqWar	0,89035088	0,8018	0,967	0,821862	0,704453	0,825911	0,854737	0,75	0,89083
08_FrenchRev	0,86363636	0,8067	0,94	0,76	0,691429	0,897143	0,808511	0,744615	0,918129
09_GrecoPersian	0,51724138	0,7477	0,893	0,232558	0,620155	0,837209	0,320856	0,677966	0,864
10_PunicWars	0,88235294	0,8085	0,922	0,263158	0,666667	0,824561	0,405405	0,730769	0,87037

Statistics-Based Graphical Modeling Support for Ontologies

Andreas Emrich
Institute for Information Systems (IWi)
German Research Center for AI (DFKI)
Saarbrücken, Germany
andreas.emrich@dfki.de

Frieder Ganz
CCSR
The University of Surrey
Guildford, Surrey, UK
f.ganz@surrey.ac.uk

Dirk Werth, Peter Loos
Institute for Information Systems (IWi)
German Research Center for AI (DFKI)
Saarbrücken, Germany
{dirk.werth,peter.loos}@dfki.de

***Abstract*—Models are a conceptualization and aperture of the real world. The asynchronous characteristics of this model construction poses a significant problem especially in highly dynamic and evolving environments. Hence, models need to be permanently checked against the data they represent. From this new challenges for modeling tools arise: Contemporary modeling tools must be able to anticipate environmental events and changes and to provide appropriate support for knowledge engineers. This paper presents a conceptual approach to process events, collect usage statistics and leverage this information for automatic and semi-automatic modeling support of ontologies. In a prototypical evaluation, a plugin for Protegé is developed, that allows for editing ontologies and visualizing new insights according to captured statistics. The paper concludes with two distinct show cases for business process modelling and sensor networks.**

Keywords*—*user modeling support; ontology modeling; ontology evolution; usage evaluation

I. Introduction

Conceptual models formally describe the real world to foster sensemaking and communication [1]. The process of model description takes a certain amount of time. The same applies for the time period of model usage. Consequently, models have to be checked against their real world equivalents permanently [2]. Especially in highly dynamic environments, the pace of environmental changes can be overwhelming for modelers [3], i.e., they need tool support to aid them with anticipating these environmental changes [4].

For the semantic web, notions such as ontology learning, ontology evolution, etc. paint a vision of self-adapting, knowledge-based model bases, which comprehend, reflect and adapt towards their environment. A core challenge in this respect is, that many modeling decisions cannot be decided automatically. For instance, if we capture interactions for a swim guide application for the iPhone, and we get a high correlation on ”good swimming experience” and ”good weather”, but also on ”good swimming experience” and ”bad weather”, it is hard to decide automatically, if the weather does matter or not. The specific domain knowledge of the knowledge engineer could help to resolve this problem. This example demonstrates, how data acquisition methods and human modeling activities are co-dependent from each other. Hence, an approach is needed that unifies these divergent perspectives.

In the context of this paper, we present a conceptual approach for a statistics-driven modeling support that presents an application programming interface (API) for capturing context events and transferring them to our ontology base and for leveraging this data .The paper demonstrates a novel concept, that will be constructed according to a design-oriented research methodology [5]. In a prototypical implementation, we show a prototype plugin for Protegé that supports statistics-enabled introduction of new concepts and properties to an existing ontology. The prototype features support for ontology classes but not yet for instances. In the show case section we discuss the benefits and potentials of this approach in the fields of business process modelling and sensor networks. Overall, this should serve as a starting point for future research to adopt the depicted concept and to improve modeling practice in various domains.

II. Related Work

Over the years, many different editors for OWL-based ontologies have emerged. While many support a tree-based or form-based editing of ontologies and quite a lot of them also support visualization techniques, only a few support graphical editing of the ontologies.

The approach of Dimitrova et al. [6] provides a basic assistance for modeling ontologies based on defined linguistic rules that have been applied to English text. Populous [7] uses a pattern-based approach to transform table-based data into ontology content. Tools like Cicero [8] or Collaborative Protegé [9] focus on non-automatic, collaborative support by creating workplace for collaborative ontology construction. Other tools such as GrOWL [10] or Protegé plugins such as OntoGraf [11], OntoViz [12], OWLViz [13], etc. merely support a declarative graphical editing of ontologies.

Overall, none of the described approaches supports a direct feedback loop from actually monitored usage data with the ontology models at hand. Although, process modeling and sensor data modeling have their specifics, the same shortcomings can be observed in these areas:

1) Process-specific support: The research field of process mining aims at constructing process models from mining process event logs automatically. Although, there has been extensive work in literature and industrial practice regarding this topic, until now there are only a few approaches that reflect the representational bias of process mining [14] or support semi-automatic approaches to use the mined knowledge to construct models manually. A further problem of contemporary approaches is the static mining process itself and the lack of the consideration of interaction. An exception in this area is the work of Hammori et al. [15], in which the authors create a permanent loop between monitored interactions and the modeling tool.

2) Sensor-specific support: In the domain of semantic sensor networks, ontologies are usually defined by ontology engineers only. In our previous work [16] we provided an automatic static creation of a semantic representation of sensor networks. Sensors publish their meta information such as sensor capabilities, energy status and neighbouring nodes to a centralised entity (i.e data sink, gateway) where the information is linked and stored in an ontology. However this approach does not contemplate the higher meanings of the data and does not provide a perceptual view of the sensors environment but only about the sensor devices itself . Recently some novel approaches try to combine machine learning methods to either bootstrap or refine ontologies and represent the meaning of the raw data in a semantic representation.

In the work of Stocker et al. [17] a system is introduced to detect and classify different types of road vehicles passing a street with the help of vibration sensors and machine learning algorithms. The objectives of the work are to acquire knowledge, represented in an ontology by abstracting from the physical sensor layer and the sensor data layer via classification methods.

The outcome of the classification process is then transferred into an ontology representation. The authors use rule based inference to map the outcome of the classifier to the ontology. The ontology consists of concepts such as feature of interest (vehicle type) and observation result time. For each classified car, an individual is created in the ontology with the particular context information.

In the work of Barnaghi et al. [18] abductive reasoning is used to analyse raw sensor data and eventually infer through ruling out obsolete explanations what ontological concept the data refers to.

However, none of the approaches use a supportive semi-automatic approach in which engineers and intelligent algorithms can complement each other. In this paper, we introduce an hybrid approach that on the one hand works autonomously but also supports the decision making process for domain specialists.

III. CONCEPT

The goal of this research, is to provide a tool to support ontology modelling and management by incorporating live data originating from ontology usage. Based on the analysis of related approaches in literatures (cf. Section II), some key requirements can be formulated, that aid to design such a system:

1) **Various event notification modes**: Based on different modelling needs, also different interaction modes need to be supported:
    - *Batch operation*: Historical Events from log files or databases can be populated and taken over into the tool at once.
    - *Real Time interaction*: Every single event that is monitored in real time will be pushed directly to the tool.

2) **Different degrees of model adaptation**: Based on the criticality, importance of certain artefacts or the statistical significance of their correlation, different degrees of adaptation support will be provided by the tool:
    - *Automatic adaptation*: In case statistical significance of an unknown relationship is present, such relationships will be explicitly modelled as properties interlinking the two respective concepts with the relationship.
    - *Semi-automatic adaptation*: Statistical significance of a new relationship or concept is only partly given, this will be modeled and highlighted to be revised manually by a domain expert.

A. Overview of the Approach

The approach chosen aims at being generic, in order to cater the needs of different use cases that expose a strong need for dynamic adaptation and the demand for a deep analysis of interactions with the models. Figure 1 depicts the influences relevant for the considerations made in this approach:

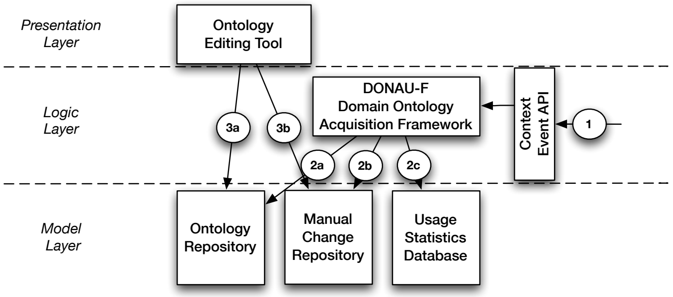


Fig. 1. Ecosystem for Statistics-Driven Ontology Modeling Support

The figure explains the general infrastructure, we provide for ontology evolution support.

1) *Event notification*: A generic context API serves as an entry point for any kind of event that needs to be associated with model data (case 1). A shallow data representation for exchange of such event notification is chosen. In technical terms, this API provides RESTful Web Service interfaces in order to be easily accessed through many different applications and programming languages.

2) *Change reasoning*: The central component for all ontology evolution processing and reading and writing the usage statistics database is the DONAU-F or DONAU framework (Domain Ontology Acquisition Framework) component. It loads the underlying ontologies and infers on the facts stated there. The DONAU-F component provides a plugin infrastructure for data-specific extensions of reasoning mechanisms. See Section V for some examples related to business process modelling

and sensor networks. In terms of the quality and significance of statistical correlations, the changes are either propagated to the ontology base automatically (case 2a) or are marked for later manual editing in the ”manual change repository” (case 2b). In any case, all event statistics are synchronized with the usage statistics database (case 2c), which serves as data repository for all further adaptations taken in the future by DONAU-F.

3) *Ontology editing*: The ontology editing tools load the ontologies in their usual manner, but also highlights relationships and concepts that have been inferred based on our component. It distinguishes between automatically inferred concepts and relationships and such that require a specific modeling action by the knowledge engineer. Section IV describes, how this is implemented in a prototypical version for Protegé.

Overall, every system undergoes changes, that are mainly influenced by the actual interactions with these systems. All these interactions should be covered in a consistent manner, in order to analyse them and derive possible adaptations for the ontology design. The main idea is to correlate observed behaviour, identify artefacts and discover relationships amongst them. According to the statistical significance of these relationships, automatic or manual adaptation plans can be triggered. The core monitoring analyser components fulfils this task, feeds the associated statistics database and eventually causes the ontology to change on its own, or marks certain discovered relationships for later manual editing.

B. Definition of Monitored Aspects

As the main principle has been described, it is still unclear, how such monitoring can be implemented. For a data definition the following aspects have to be considered:

- *defined concepts*: In order to monitor behaviour, it is essential to name aspects clearly, that are monitored, in order to deliver starting point for further analysis. E.g. in our swim app example: If the concepts weather and swimming are already defined, it is easy to analyze a relationship amongst them.
- *wildcards*: For unknown aspects, it is essential to name them according to some variable name, in order to correlate them later on and to name them.
- *defined relationships*: In order to capture the nature of dependencies, it is important to have named relationships that can be either validated or invalidated through the observed behaviour.
- *open relationship interlinking*: The possibility of linking aspects arbitrarily must not be impeded by a superponed model. This is important to ensure that unstructured scenarios are possible as well.

IV. Prototypical Implementation

The described concept has been implemented as a prototype for the ontology modeling software Protegé. To bring about the described changes, the OWLViz [13] plugin has been extended by a new view that supports statistics-enabled, graphical editing of ontologies. The view consists of a graphical editing panel that displays discovered concepts and relationships.

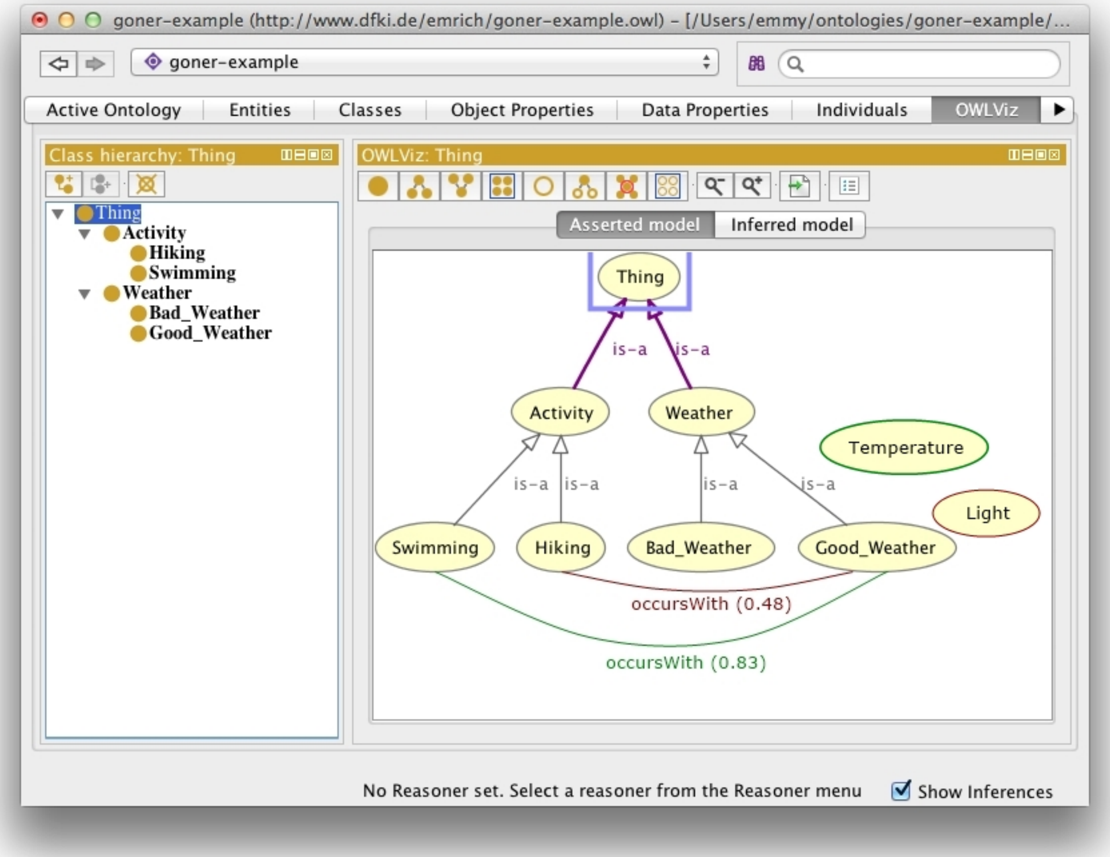
The screenshot in Figure 2 visualizes, how the plugin works. In the graphical editing panel, ontology classes are represented by ovals and properties, that link them are represented with connecting lines. Newly discovered classes and properties are displayed either in green or in red. Green indicates, that the given concept or relationship was significant enough for automatic detection, whereas red indicates that the co-occurrences in the event logs hint at a possible concept or relationship, but because of the low significance it could not be confirmed. The size of the ovals and the thickness of the connecting lines is associated with the relative importance gathered from the underlying statistics, i.e., if a newly discovered class A is used more often than another class B, A is being displayed bigger than B. In our example, the concepts ”Temperature” and ”Light” are new. As ”Temperature” occurs more often in the associated statistics, the term is represented with a bigger oval. In the given example, there are no discovered relationships between these discovered concepts and already established concepts that are significant enough to be displayed, i.e., in order to be displayed, the statistical significance of such relationships has to surpass a defined threshold. In our example, this is 0.3, i.e., concepts have to cooccur in more than 30% of the cases in order to be displayed. The same principle is applied to properties: The stronger a correlation among two classes through a linking property is, the thicker the line is that represents his property. In the given example, there is obviously a stronger correlation between ”Swimming” and ”Good Weather” than between ”Hiking” and ”Good Weather”. Apparently, the relationship between ”Swimming” and ”Good Weather” is significant enough, to infer an automatically discovered relationship.

What cannot be seen in the screenshot is, that statistics are also maintained for already existing concepts and relationships. Although there is no dedicated view for that in our current version of the prototype, a future ”Ontology Management View” should enable to reassess the importance and relevance of certain concepts and relationships. By and large, this can help to ensure that an ontology retains a certain size and thus helping to reduce computation time for querying the ontology base.

A. System Model

The main design goal for our solution is, that it provides an open infrastructure for ontology evolution support, which is independent from the underlying ontologies and the programming language of the source systems. Figure 1 shows the main architecture components of our ontology evolution infrastructure.

It is straightforward to monitor and analyse concepts and relationships that are already defined by the user through the Ontology Editing Tool. However, unstructured information collected via the Context Event API require mechanisms to 1) create new labelled concepts and relationships that reflect the work flow of the underlying processes and 2) to validate and or invalidate existing semantic knowledge.



No Reasoner set. Select a reasoner from the Reasoner menu

Fig. 2. Graphical Editing View of the DONAU-F plugin

In this section, we discuss DONAU-F, an inference framework to detect wildcard concepts through clustering that groups events similar to their occurrence and context with the help of the k-means clustering algorithm. Furthermore, to detect new linking or invalidate existing linking between concepts, a Markov model is used to create a probability distribution of the temporal relation between different concepts.

1) Wildcard Concept Extraction: In our prototypical implementation, we use a k-means clustering mechanism that groups certain events based on their properties (occurrence, meta information, time) into groups that can either lead to new concepts or be mapped to existing ones. We define a certain threshold that indicates if a new mapping between cluster and concept can be populated without manual revision or if an ontology engineer has to be considered and the new concept therefore has to be highlighted in the editing tool.

2) Open Relationship Interlinking Extraction: Our approach exploits the frequency of events and their temporal occurrence to construct a Markov chain that represents the likelihood of temporal relations and correlation between events. The system counts the occurrence of events and creates a frequency distribution table. The created Markov chain represents the probability of the transition from one event to another event. The chain let us infer if events occur more frequent after certain events and therefore are in some relationship that is going to be represented in the ontology.

Figure 3 shows how temporal relationships are discovered levaring the depicted Markov chain.

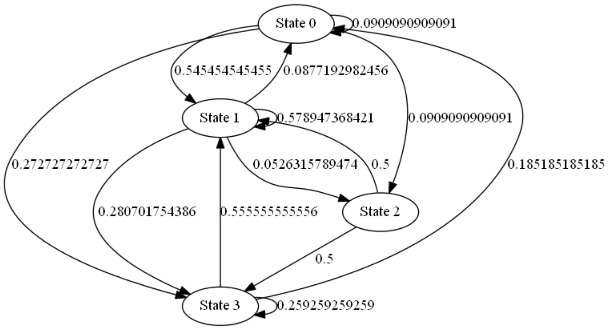


Fig. 3. Relations between different wildcard concepts and their temporal likelihood to occur before/after each other

In our implementation, we are able to detect, represent and highlight relations between concepts through temporal properties such as *occursAfter*, *occursBefore* and *occursSame*. Through hierarchical clustering [19] the system is able to relate concepts through properties such as *isA* and *similarTo*.

V. Show Cases

The presented concept and tool support could improve the model management in many different areas. In order to explain the benefits approach, the two areas of business process modeling and sensor networks are explored for application potentials of our approach.

A. Business Process Modeling

For business process modeling the representation of the operational business of a company is a key requirement. However, the business environment is very dynamic and constantly changing. In terms of business process reengineering, according to which the organizational workflows should be reengineered throughout the modeling activities, a big issue is, that the actual workflows and processes change while the modeling efforts are being done. In consequence, once the modeling is finished, it cannot be ensured, that the representation of the actual processes is still correct.

A business process does not only have workflow-related aspects, but also has to consider data, resources, organizational aspects etc. A possible notation for business processes, which covers all these aspects, is the extended event-driven process chain (eEPC). It comprises events, functions, organizational units, resources, documents, etc. as possible design artefacts. The business process modeling notation (BPMN) concentrates on the workflow issues, organizes roles with pools and lanes and introduces ”artifact” for all other aspects.

In the context of business process modeling, our approach could have the following positive impacts:

- *discover process variants*: Especially in processes with many execution alternatives and degrees of freedom concerning execution, it is hard to explicitly model the process without limiting the user. By analyzing process executions in the crowd, process variants can be discovered and the process modeler in charge can easily decide, whether to accept newly emerged process variants as standard operating procedures or to discard them.
- *discover new process steps*: If a process model in terms of an ontology exists, and new unknown process steps are discovered, this might indicate, that the process actually is not compliant to the model, which could have its origin in an insufficient model or an erroneous execution.
- *discover new process responsibilities, resources, etc.*: In the same manner, new responsibilities can be determined. In terms of organizational units this means, that a process can also be executed by other organizational units than originally planned, or it can be identified as a situation, that is not desired. The same applies for associated resources, inputs, outputs of process steps, etc. As many process steps, e.g. a contract checking, rely on deep consistency checks of associated resources and such consistency checks rely on a defined model, it is obvious, that this method helps to ensure that all relevant artefacts are considered by such consistency checks.
- *validate existing relationships*: As mentioned earlier, a defined model does not necessarily reflect the current reality of a system. Therefore, existing model relationships should be constantly monitored and analyzed in order to be validated or invalidated. E.g. if we have have a business process with several subsequent process steps, and according to that process model step B always follows step A, but statistics shows, that in 60% of the cases step C follows step A, the process model needs to be changed accordingly. In terms of evaluation mechanisms this implies, that we need a plugin mechanism as described in Section IV-A to enable a deep semantic analysis of concepts and relationships in an ontology.

Especially in unstructured and not fully modeled scenarios, the approach seems to have its merits and enable an easier maintenance and evolution of business process models. Future research should catch up with process mining and process evolution work in order to achieve the vision described here.

B. Sensor Networks

Sensor Networks are exploited to capture and share data from the physical world and integrate it into software systems. Recently, there has been an increasing trend in research fields such as pervasive and ubiquitous computing and especially in the Smart-Home, -Office domains, where making the gathered data available to the end-user is crucial.

One of the main challenge that remains is to make the usually raw unstructured sensor data understandable for the user and/or machine-interpretable [20]. The Semantic Sensor Web [21] is one approach that allows to annotate, represent and map gathered sensor data to semantic concepts and also represent their relations via properties. Despite the data-centric focus, also device and physical information such as sensor device meta information and hierarchies, environmental attributes and network layouts can be modelled [22].

However the nature of sensor networks, phenomena and data gathered is volatile. Sensor devices can be faulty, parameters and attributes of events can change and other adhoc issues, that alter the association between model and reality can occur. Moreover the vast amount of information produced leads to an information overload that can not be managed by single experts.

The dynamics of sensor networks and observed phenomena has to adapted in the model. We identified several use-cases where the proposed approach can be used to facilitate the adaptation between real state of the network and environment, and the semantic model.

- *Knowledge Acquisition from unstructured raw sensor data*: Events monitored by sensor networks can be modelled by domain experts in the initial ontology as defined items such as ”bad weather” or ”good weather”. However, with the upcoming deluge of data and the information overload for human ontology engineers, this process can be outsourced to the DONAU framework. DONAU can be used to support and facilitate the construction of an initial model or

Copyright (c) IARIA, 2013. ISBN: 978-1-61208-293-6

to refine the representation based on the statistics gathered.

- *Outlier Detection*: The DONAU framework cannot only be used to acquire knowledge that is expected in the domain, but also to infer new insights. Occurring events that cannot be related to existing defined items can be marked as wildcard concepts, and eventually highlighted in the Protege Plugin for further inspection.
- *Network Topology Tracking*: In case that the network topology of sensor networks are modelled in a semantic representation, the approach can be used to monitor changes and update the ontology. The Context API can be used to retrieve health information from particular nodes, in case nodes are not responding or communicating failure, changes can be reflected in the ontology.

VI. Conclusion and Outlook

The synchronization of models and their equivalents represent a core problem of ontology management and information modeling, especially in highly dynamic environments such as the Internet of Things or business information systems. This paper presented a conceptual approach for the capturing of event data within the DONAU framework, that enables to identify automatic changes to an ontology or to provide recommendations for model adaptations to ontology engineers. A prototypical implementation in the ontology modeling tool Protegé demonstrated, how this support functionalities can enhance the graphical modeling of ontologies. Furthermore, we explicated, how the depicted approach can improve the graphical modeling and model management in the domains of business process management and sensor networks.

In terms of evaluation, this paper has shown a first prototypical evaluation as a proof-of-concept. It demonstrates the potential of the depicted approach. However, further evaluations are needed in the future to evaluate the quality of recommendations from an information retrieval (IR) perspective using common IR metrics such as precision or recall and from a user perspective with help by structured user walkthroughs and qualitative questionnaires.

The approach presented in this paper has its focus on a class and not an instance level at the moment. For the consideration of the instance level more aspects have to be considered, as the relationship properties might not be only class-to-class or instance-to-instance but also class-to-instance relationships. Moreover, complex properties are not considered at the moment, that relate to more than two involved artefacts. Clustering methods could guide the way, how to find the most appropriate subsets of artefacts that constitute such relationships and hence are recommended in the tool. Furthermore, aspects as costs or priority as described by Maedche et al. [23] have not been considered so far, but could help to improve future versions. In a similar manner, existing model relationships could be permanently reevaluated regarding their significance. This could help in areas, where relatively compact models are needed, e.g. for high-performance reasoning. The authors plan to release the presented prototypical implementation as an open source software project, in order to provide a tool and code base for a new generation of ontology editing tools, that allow for the seamless integration of the modeling world and actual running systems.

Moreover, the concept that has been proposed is not only limited to ontology modeling and editing but also can help to find potential improvements. In terms of tool support, future implementations could transfer these results to more domain-related tools such as business process modeling suites, etc. Furthermore, the gathered statistics could be part of business intelligence applications for model governance. E.g., non-relevant relationships could be dropped from the model in order to speed up associated analysis. Moreover, dependency analysis of certain events as shown in [24] could help to estimate the impacts of proposed modeling changes and could be a vital feedback for ranking mechanisms. Besides modeling, the applied principles could also be transferred to other problem classes, such as the navigation of ontologies. Based on the research presented in [25] new mechanisms could be developed, that offer a relevance-based navigation of ontologies.

References

[1] Y. Wand and R. Weber, “Research Commentary: Information Systems and Conceptual Modeling?A Research Agenda,” *Information Systems Research*, vol. 13, pp. 363–376, 2002.

[2] G. Y. G. Yang, M. B. Dwyer, and G. Rothermel, “Regression model checking,” *2009 IEEE International Conference on Software Maintenance*, pp. 115–124, 2009.

[3] B. A. Rajabi and S. P. Lee, “Modeling and analysis of change management in dynamic business process,” *International Journal of Computer and Electrical Engineering*, vol. 2, pp. 181–189, 2010.

[4] J.-C. B. A. Francois, D. J. Blackwood, and P. W. Jowitt, “Decision mapping : understanding decision making processes,” *Civil Engineering and Environmental Systems*, vol. 19, pp. 187–207, 2002.

[5] A. R. Hevner, S. T. March, J. Park, and S. Ram, “Design science in information systems research,” *MIS Quarterly*, vol. 28, no. 1, pp. 75–105, 2004. [Online]. Available: http://www.hec.unil.ch/ypigneur/HCI/articles/hevner04.pdf

[6] V. Dimitrova, R. Denaux, G. Hart, C. Dolbear, I. Holt, and A. G. Cohn, “Involving Domain Experts in Authoring OWL Ontologies,” *Proceedings of the 7th International Semantic Web Conference ISWC2008*, vol. 401, pp. 1–16, 2008. [Online]. Available: http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-401/iswc2008pd_submission_33.pdf

[7] S. Jupp, M. Horridge, L. Iannone, J. Klein, S. Owen, J. Schanstra, K. Wolstencroft, and R. Stevens, “Populous: a tool for building OWL ontologies from templates,” *BMC Bioinformatics*, vol. 13, no. Suppl 1, p. S5, 2011. [Online]. Available: http://www.biomedcentral.com/1471-2105/13/S1/S5

[8] K. Dellschaft, H. Engelbrecht, M. Barreto, S. Rutenbeck, and S. Staab, “Cicero: Tracking Design Rationale in Collaborative Ontology Engineering,” *The Semantic Web Research and Applications*, vol. 5021, no. 5021, pp. 782–786, 2008. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-68234-9_58

[9] N. F. Noy and T. Tudorache, “Collaborative Ontology Development on the (Semantic) Web,” *Nature Biotechnology*, 2008.

[10] S. Krivov, R. Williams, and F. Villa, “GrOWL: A tool for visualization and editing of OWL ontologies,” *Web Semantics Science Services and Agents on the World Wide Web*, vol. 5, no. 2, pp. 54–57, 2007. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S1570826807000157

[11] “OntoGraf Project Website.” [Online]. Available: http://protegewiki.stanford.edu/wiki/OntoGraf

[12] “OntoViz Project Website.” [Online]. Available: http://protegewiki.stanford.edu/wiki/OntoViz

Copyright (c) IARIA, 2013. ISBN: 978-1-61208-293-6

[13] "OWLViz Project Website." [Online]. Available: http://www.co-ode.org/downloads/owlviz/

[14] W. M. P. V. D. Aalst, "On the Representational Bias in Process Mining," pp. 2–7, 2011. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5990012

[15] M. Hammori, J. Herbst, and N. Kleiner, "Interactive workflow miningrequirements, concepts and implementation," *Data & Knowledge Engineering*, vol. 56, no. 1, pp. 41–63, 2006. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0169023X05000273

[16] F. Ganz, P. Barnaghi, F. Carrez, and K. Moessner, "Context-aware management for sensor networks," in *Proceedings of the 5th International Conference on Communication System Software and Middleware*. ACM, 2011, p. 6.

[17] M. Stocker, M. Rönkkö, and M. Kolehmainen, "Making sense of sensor data using ontology: A discussion for road vehicle classification," in *International Congress on Environmental Modelling and Software*. iEMSs, 2012.

[18] P. Barnaghi, F. Ganz, C. Henson, and A. Sheth, "Computing perception from sensor data," in *Sensors, 2012 IEEE*. IEEE, 2012, pp. 1–4.

[19] A. Maedche and S. Staab, "Ontology learning for the semantic web," *Intelligent Systems, IEEE*, vol. 16, no. 2, pp. 72–79, 2001.

[20] O. Corcho and R. García-Castro, "Five challenges for the semantic sensor web," *Semantic Web*, vol. 1, no. 1, pp. 121–125, 2010.

[21] A. Sheth, C. Henson, and S. S. Sahoo, "Semantic sensor web," *Internet Computing, IEEE*, vol. 12, no. 4, pp. 78–83, 2008.

[22] M. Compton, P. Barnaghi, L. Bermudez, R. García-Castro, O. Corcho, S. Cox, J. Graybeal, M. Hauswirth, C. Henson, A. Herzog *et al.*, "The ssn ontology of the w3c semantic sensor network incubator group," *Web Semantics: Science, Services and Agents on the World Wide Web*, 2012.

[23] A. Maedche, B. Motik, L. Stojanovic, R. Studer, and R. Volz, "Managing Multiple Ontologies and Ontology Evolution in Ontologging," *Intelligent Information Processing IFIP The International Federation for Information Processing*, vol. 93, pp. 51–63, 2002.

[24] A. Emrich, D. Werth, and P. Loos, "Real-time semantic process change impact analysis," in *Innovative Unternehmensanwendungen mit In-Memory Data Management 2011. Innovative Unternehmensanwendungen mit In-Memory Data Management (IMDM-11), December 2, Mainz, Germany*, W. Lehner and G. Piller, Eds. Springer, 2011, pp. 91–100.

[25] A. Emrich, A. Chapko, J. Gehlen, F. Ganz, D. Werth, P. Loos, and D. Campus, "Personalized and Context-Aware Visualization and Navigation of Ontologies for Mobile Services," in *Proceedings of the 12th International Conference on Human-Computer Interaction with Mobile Devices and Services*, Lisbon, 2010.

Semantic Approach for Finding Suitable Commercial Business Location

Felix Mata
Computing Mobile Laboratory
IPN-UPIITA
Mexico City, Mexico
mmatar@ipn.mx

Omar Juarez, Consuelo Garcia, Roberto Zagal
Systems Department
IPN-ESCOM
Mexico City, Mexico
{omarjg82@gmail.com, varinia400@hotmail.com, rzagalf@ipn.mx}

***Abstract*—Nowadays, in order to find the best suitable location for a commercial business, one needs to apply a market analysis in several steps. This research improves this process by applying spatial semantics. We define a methodology to find a list of candidates for the best commercial business location. Some parameters involved in spatial semantics and market analysis are: floating population, taxes cost, market competition, transportation, land use, geographical area, among many others. All of them should be evaluated with different methods. A list of semantic rules are defined to find possible location, ontology is built based on the geographic and market conditions. Each one of places found complies with law restrictions, administrative and geographic requirements defined as input parameters. Additionally, it is possible to relax the parameters' weight in order to obtain new results for possible locations. Proofs include results obtained in conjunction (semantic and market analysis) that shows better results that when these criteria are used independently.**

Keywords-Spatial Semantics; geographic domain conceptualization; axiomatic relation.

I. INTRODUCTION

Nowadays, finding the adequate location for establishing a business, is a task that includes review aspects of market (e.g., business market competition, economy, geographic conditions, etc.). This process is not done by just one person; it is made by the collaborative work of specialists from different disciplines (marketers, economists, geographers). Then, the possibility of automating this task is possible, if we consider that: 1) ontologies can represent the knowledge involved in this context (spatial and market conditions), 2) semantic analysis allows to explore ontologies and define rules based on certain constraints, 3) the processing of spatial semantics is used to find locations, based on semantic rules and geographic conditions. Thus, these three statements are integrated as a methodology to find a list of possible locations for a business. In addition, users can assign higher or lower relevance to each parameter (qualitative or quantitative configuration) in order to generate a new list of possible places for locating a business. For example, suppose the following query:

Q_1= "find a business location in a county without market competition in a radius of 5 km". It can be relaxed qualitatively by the user in parameter market competition. Then, the system generates new queries as follows:
Q_{1a} = "find a business location in a county with low market competition in a radius of 10 km".
Q_{1b}: "find a business location in a county without market competition of first impact in a radius of 5 km".

The case study focuses on three types of business: stationer' shops, drugstores and restaurants. The methodology includes definition of semantic processing algorithms, geographic queries implementation and the design and implementation of an ontology. The research is focused on assisting in making a decision about the possible locations for establishing a business. The implementation was done into a Web system. The rest of paper is organized as follows: Section II describes the related work; Section III explains the methodology definition; Section IV shows the obtained results and, finally, the conclusions and future work are outlined in Section V.

II. RELATED WORK

The use of an ontology as a resource of knowledge is well known from tasks such as extracting the semantics, the meaning, of natural texts to entity recognition (people, places, companies, and prices) [14]. Ontology is also employed as a method for identifying categories, concepts, relations, and rules [5], [6], [8]. In our work, we use this principle to find the best locations for a business. On the other hand, the use of ontology in market domain is not new. Researchers have suggested employing market mechanisms for the allocation of Web Services [16], while Lamparter and Schnizler [15] presented an architecture of an ontology-driven market for trading Semantic Web Services. Auction schema is enriched by a set of components enabling semantics based matching, as well as price-based allocations. Meersman [13] uses the market information into the knowledge system of a company in order to contribute to the process of product innovation, competence development and relevant social interaction. Nevertheless, these previous

Copyright (c) IARIA, 2013. ISBN: 978-1-61208-293-6

works do not consider the geographic parameters that are relevant in the market analysis.

Ontology is also used to inform designers of data models and information systems to make them better equipped for handling geographic concepts [1], [3], [7], while Smith and Mark [5] report the results of a series of experiments designed to establish how non-expert subjects conceptualize geospatial phenomena. Subjects were asked to give examples of geographical categories in response to a series of differently phrased elicitations. The results yield an ontology of geographical categories, which consists of a catalog of the prime geospatial concepts and categories. When combined with query languages, domain ontologies favor the design and development of domain-based search engines and their application to different areas such as in transportation [9]. A categorization of GIS (Geographic Information System) tasks and GIR (Geographic Information Retrieval) queries has been suggested in [10], the approach being applied to heterogeneous sources, including multimedia sources. Moreover, Biletskiy and Ranganathan [17] explain an ontology and rule-based framework for the development of business domain applications, which includes semantic processing of externalized business rules and, to some extent, externalization of application logic. The framework also includes a rule learning system to semi-automate the generation of information extraction rules from source documents. Summarizing, the use of semantic processing, and the ontology exploration, have been used in other domains, but not in combination with the geographic aspects in market analysis. Thus, in this paper, we show a methodology to be applied in order to get a useful application for real life.

III. METHODOLOGY FOR QUERYING AND SEMANTIC PROCESSING

The methodology is composed of three phases: 1) Ontology modeling, 2) Ontology building, and 3) Semantic processing. Ontology modeling consists of capturing knowledge regarding market context for a set of specific business and their involved geographical parameters. For that, one is required to get a set of official data sources with: 1a) geographic data of places to analyze, 1b) statistics regarding to economy, soil use, population, among others, 1c) mechanisms used in the market to get the type and behavior of it. These components (1a, 1b, 1c) are transformed into semantic rules and concepts. The ontology building consists of reordering the concepts into a taxonomy and establishing the context for each concept following the methodology described in [2]. Finally, semantic processing is responsible to apply the semantic rules and measure the relevance of query parameters. This means, assign a lesser or greater importance to one parameter such as: the price of rent or the proximity of a communication route. This is achieved using the mechanisms of the market in previous step 1c). Basically, these mechanisms are surveys and quests made in others studies.

III.A Modeling Ontology

The modeling consists of taking the knowledge from one conceptual domain to a logical domain. The knowledge is acquired and abstracted, then is translated to logical model to obtain a design (classes and properties). To identify the domain of ontology, we classified the analysis based on semantics and market analysis for the location of a business in: 1) market analysis, 2) proximity analysis and 3) analysis of buildings infrastructure.

1. Market Analysis. It refers to the analysis of potential customers and demand generators. It is composed of three steps:

1a) Compute the size of local market. It is obtained by processing the total population living in the area, the floating population (people who do not live in the area but work or study in it), and income and expense levels by social stratum.

1b) Analyze the market competition. It consists of calculating the number of businesses in the local market to determine if the market is saturated or free. There may be commercial firms who are sacrificing their profit margin just to have a presence in the area (this is a market strategy).

1c) Detect demand generators: identifying corporate office buildings, schools, hospitals, recreation and leisure, and so on.

2. Proximity analysis. It refers to the analysis of the proximity of a building with relevant points (cultural, historical places, downtown, etc.). We establish minimum and maximum distance from 1 km to 5 km for walking distance, and from 5 km to 10 km driving distance.

2a) Locate the areas where most businesses are concentrated. Identify if in the area there are several centers and commercial corridors (including historical centers).

2b) Identify the areas with high vehicular and pedestrian traffic.

2c) Identify the roads surrounding the area, road type, high or low level of vehicular and pedestrian traffic.

3. Analysis of buildings. It refers to an analysis of the characteristics of the buildings: number of levels, construction material (concrete, wood, etc.), parking included, stairs or elevator, number of doors, windows.

III.B Ontology

The ontology was built based on data obtained from INEGI, an official Mexican organism that produces and manage the geographical data in Mexico [4] and documents from other sources such as [11], [12] and with an approach similar to [14] although it is possible to use Open Street Map for the mapping data, but the tabular data of economics parameters should be acquired from other sources as was described at the beginning of Section III. We identify the relevant concepts in a business by following the procedures to measure the behavior of a market in Mexico, (SNIIM in Spanish) [18].

The concepts and terms are abstracted based on definitions of Royal Academy of the Spanish language. Some examples are:

• Offer: Set of property or goods being offered on the market at one price and at a specific time.

• Demand: Overall amount of purchases of goods and services performed or provided by a community.

• Market: Set of consumers able to purchase a product or service.

• Product: Profit.

Moreover, we identify properties and relationships of each one of the concepts. Then, we obtain the taxonomy and classes shown in Fig. 1.

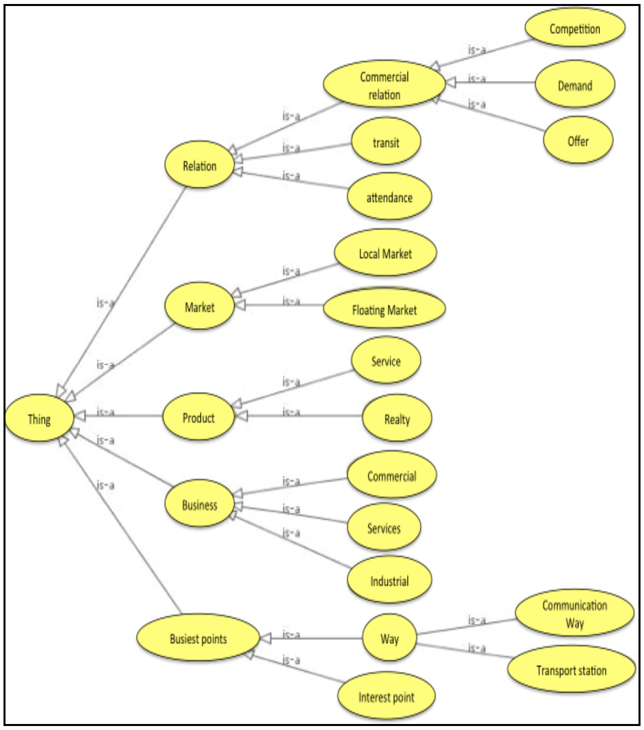


Figure 1. Taxonomy of Ontology.

For the definition of concepts and final model of ontology, the knowledge acquired and surveys were used. Some rules and terms used in market context are as follows:

a) Check type and amount of products offered by the business. b) Check floating market, demand products of business and the importance of offer and demand. c) Check the places that people of floating market attend and how often they attend. d) Check the ways and stations where the floating market is often attended. f) Evaluate the influence of interest points and attractions in the area.

Regarding the surveys conducted during the research, we define a range of values to qualify characteristics of a building or property (see Table 1).

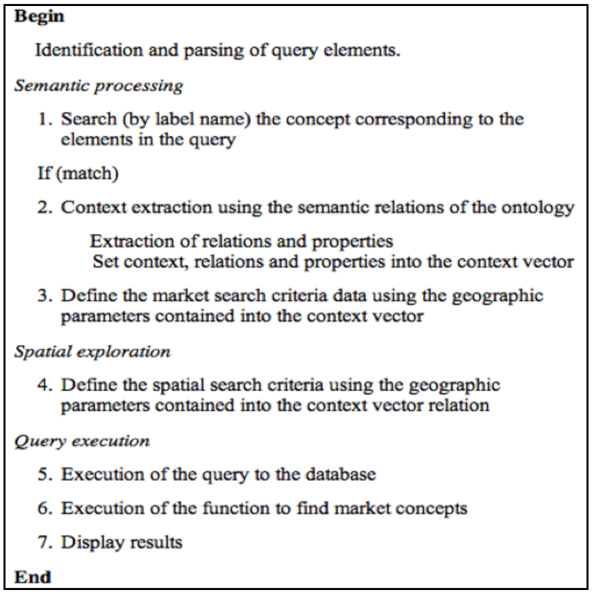
TABLE I. TABLE RELATION QUALIFYING- WEIGHT VALUES

Qualifying	Weight value
Without importance	0
Without importance but required	25
desirable	50
necessary	75
required, essential	100

We used reasoner FaCT ++, for instance creation and classification, and we explored the ontology using OWL (Ontology Web language) Java API (Application Program Interface).

III.C Semantic Processing

The query is received and the processing starts: the query is analyzed and contextualized through the exploration of the ontology. Next, the rules of the context are applied (market analysis constraints). This results in the parameters that define a geographic query in market domain. Then, the ontology is explored to find a matching between query elements and the concepts stored into ontology. Each element of a query is searched within the ontology (e.g., by label names). When a concept is found into the ontology, then its context is extracted. A context is formed by the neighboring concepts of a matched concept and its semantic relations are extracted and stored into a vector. The process works according to the following algorithm depicted in pseudo code (see Fig. 2):



```
Begin
  Identification and parsing of query elements.
Semantic processing
  1. Search (by label name) the concept corresponding to the
     elements in the query
  If (match)
  2. Context extraction using the semantic relations of the ontology
        Extraction of relations and properties
        Set context, relations and properties into the context vector
  3. Define the market search criteria data using the geographic
     parameters contained into the context vector
Spatial exploration
  4. Define the spatial search criteria using the geographic
     parameters contained into the context vector relation
Query execution
  5. Execution of the query to the database
  6. Execution of the function to find market concepts
  7. Display results
End
```

Figure 2. Ontology exploration algorithm.

In Fig. 2, the algorithm illustrates how the search parameters are used to find relevant information according to the steps involved in market analysis. A specific Web service has been developed in order to match a concept name with the term name commonly used in market analysis. The matching results are used as a parameter for other queries that are generated.

Moreover, the semantic processing includes an interpretation for each parameter. For example, in the case of stores, it is desirable that they are located in a corner (it is a defined rule), while for a book store, for example, it is more important that it is located near any school.

Table 2 shows the possible semantic values for some parameters in a market context.

TABLE II. SEMANTIC VALUES FOR MARKET PARAMETERS

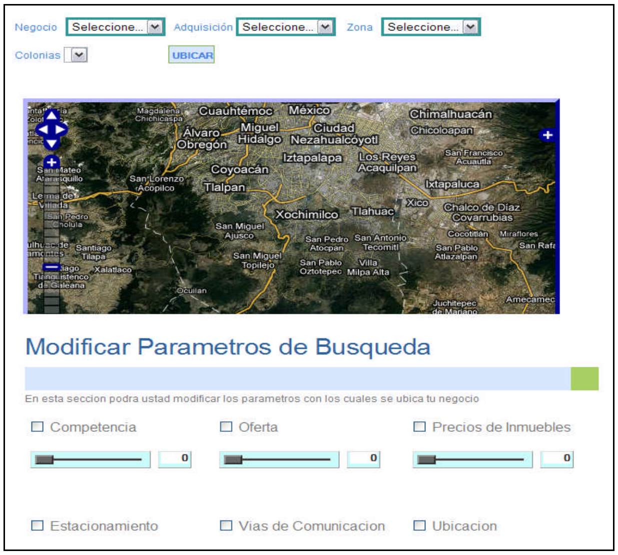
Parameter	Possible semantic values
Influence of the zone	Low, medium, high.
Market competition	First impact, high impact, small impact, based on percentage.
Transportation	Fast, without traffic.
Population	Floating, fixed.

In Table 2, the possible values or range of values for each parameter is showed. For example, when a business requires a form of transportation, the evaluation is done in two ways: fast, in terms of time and distance, considering the traffic during the transportation.

IV. TEST AND RESULTS

The tests are made by the methodology implementation on a Web system, considering the three steps of methodology: semantic modeling, semantic building and semantic processing. A list of best suitable locations for a particular business on specific geographic area is generated. Then, users can set these restrictions and display the location of the commercial business according to requirements and constraints discussed in previous sections. The business types considered are: stationer' shops, drugstores and restaurants.

For testing the interface, the user sets the query through capture controls and options. The web interface is in Spanish, in a local system (see Fig. 3).



Negocio Seleccione... Adquisición Seleccione... Zona Seleccione...

Colonias UBICAR

Modificar Parametros de Busqueda

En esta seccion podra ustad modificar los parametros con los cuales se ubica tu negocio

☐ Competencia 0 ☐ Oferta 0 ☐ Precios de Inmuebles 0

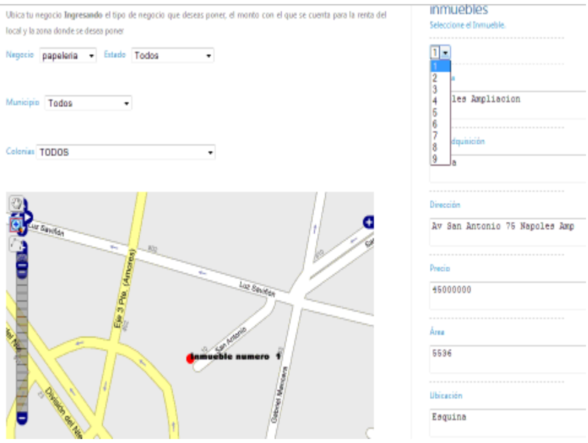
☐ Estacionamiento ☐ Vias de Comunicacion ☐ Ubicacion

Figure 3. Input parameters (interface Web system).

As Fig. 3 shows, business requirements are defined through preset options: commercial business type, population size, neighborhood, type of acquisition, among others.

We consider the following query:

Q_{1a} = "find a business location in Mexico City with market competition in a radius of 10 km located in corner, or into the street, rent or buy". The result is the list of properties that meet the criteria, one of which is shown in Fig. 4.



Ubica tu negocio Ingresando el tipo de negocio que deseas poner, el monto con el que se cuenta para la renta del local y la zona donde se desea poner

Negocio papeleria Estado Todos

Municipio Todos

Colonias TODOS

inmuebles

Seleccione el Inmueble.

1

1 2 3 4 5 6 7 8 9

les Ampliacion

dquisición

Dirección: Av San Antonio 76 Napoles Amp

Precio: 45000000

Área: 5536

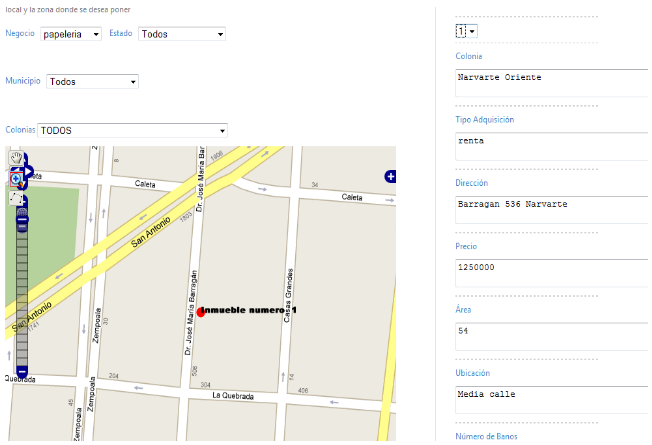
Ubicación: Esquina

Figure 4. Best Rated Property.

Now, if the query Q_1 is modified on the relevance of the parameters: parking and roads near. Then, the new Q_{1b} query is generated. The result is now a new list of locations.

Q_{1b}: "find a business location in a county without market competition of first impact in a radius of 5 km, located in corner, or into the street, rent or buy"".

Note, that the most appropriate business location from the previous test, was located now in eighth place (see Fig. 5).



local y la zona donde se desea poner

Negocio: papeleria ▾ Estado: Todos ▾

Municipio: Todos ▾

Colonias: TODOS ▾

1 ▾

Colonia: Narvarte Oriente

Tipo Adquisición: renta

Dirección: Barragan 536 Narvarte

Precio: 1250000

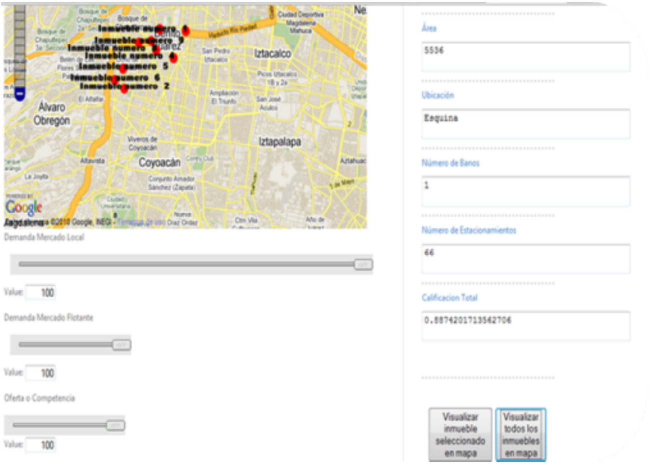
Área: 54

Ubicación: Media calle

Número de Banos

Figure 5. The best location for commercial business from Q_{1A}.

It is possible to modify / relax the parameter values to find new locations, changing the relevance of the parameters: market competition, offer, property rates, parking, roads, and location. This is achieved at the interface through sliders. For example, consider the query Q_{1A}, with the following changes: the weight factor assigned is set 50% to market competition. It generates new results shown in Fig. 6.



Demanda Mercado Local — Value: 100

Demanda Mercado Flotante — Value: 100

Oferta o Competencia — Value: 100

Área: 5536

Ubicación: Esquina

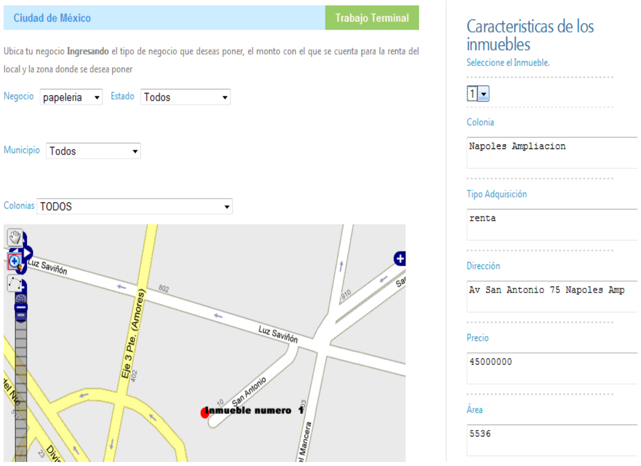
Número de Banos: 1

Número de Estacionamientos: 66

Calificacion Total: 0.8874201713562706

Figure 6. Results for new generated query Q_{1A}.

As we can see in Fig. 6, it is possible to relax the values for any parameter (change or configure to not be taken into account in the analysis process). Next, a new map will be displayed with the new results. Another test was conducted based on the query Q_3: “find a location for a restaurant in a corner and available for purchase, for any price, with fast transportation available”. The result of this query Q_3 is showed in Fig. 7.



Ciudad de México | Trabajo Terminal

Ubica tu negocio **Ingresando** el tipo de negocio que deseas poner, el monto con el que se cuenta para la renta del local y la zona donde se desea poner

Negocio: papeleria ▾ Estado: Todos ▾

Municipio: Todos ▾

Colonias: TODOS ▾

Caracteristicas de los inmuebles

Seleccione el Inmueble.

1 ▾

Colonia: Napoles Ampliacion

Tipo Adquisición: renta

Dirección: Av San Antonio 75 Napoles Amp

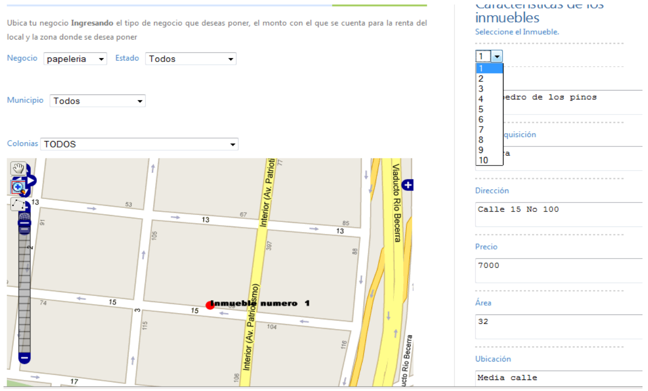
Precio: 45000000

Área: 5536

Figure 7. Best commercial locations for query Q_3.

As shown in Fig. 7, the result is only one location that complied with the values of parameters indicated.

Another test conducted for the query Q_4: "Find location for a stationer’ shop, with a low floating population, located to 3 km from a zone influenced commercially." The results obtained show 10 possible locations, the most relevant being located in: Calle 15 No 100. This is shown in Fig. 8.



Ubica tu negocio **Ingresando** el tipo de negocio que deseas poner, el monto con el que se cuenta para la renta del local y la zona donde se desea poner

Negocio: papeleria ▾ Estado: Todos ▾

Municipio: Todos ▾

Colonias: TODOS ▾

Caracteristicas de los inmuebles

Seleccione el Inmueble.

1 ▾ (1, 2, 3, 4, 5, 6, 7, 8, 9, 10)

…edro de los pinos

…quisición: …a

Dirección: Calle 15 No 100

Precio: 7000

Área: 32

Ubicación: Media calle

Figure 8. Rated building or property for query Q_4.

In Fig. 8, the best rated property for query Q_4 is shown; the relevancy is calculated based on the parameters values, if a property has the value requested, then it is considered relevant, but if not, then the value is evaluated, if it is lesser or higher, the relevance of it is assigned based on evaluation.

For example, the query Q_4 the influenced commercial zone around the all suitable locations is shown in Fig. 9.

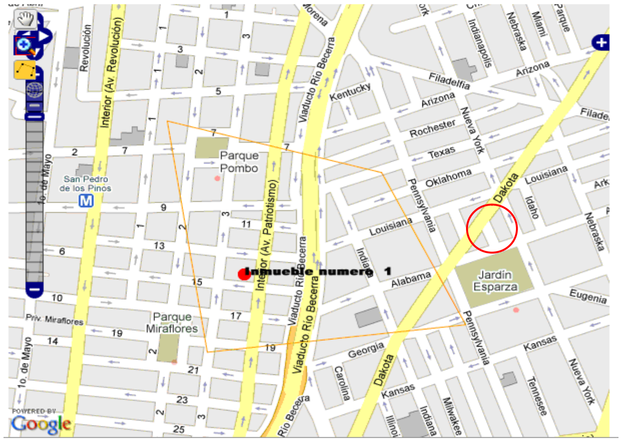


Figure 9. Zone of influence for the best location of query Q_4.

Fig. 9 shows a polygon (in orange) that represents the commercial zone that influences the locations found and evaluated as good candidates for query Q_4.

V. Conclusion

The research presented in this paper introduces an approach to perform geo-market analysis in order to find best possible places to locate a business. The methodology can be applied in other places using the corresponding geographic and statistical data of each country and procedures with applicable laws in a particular market. This can support people skilled in the branch to the location of a business and people without experience.

Semantic processing is driven by a domain ontology. The queries used are contextualized in order to closely relate information to market analysis and business location.

The approach is based on a method that matches user-defined queries with ontological concepts according to the market domains. The approach has been experimented on an illustrative case study applied in business from Mexico City.

Further work will consider larger information spaces and document collections like Web, and an integration of several distributed databases as well as additional semantic analysis. The domain ontology can also be enriched by considering additional spatial relations and by application-based conceptualizations.

Acknowledgment

The authors of this paper wish to thank the Mobile Computing Laboratory-UPIITA, Instituto Politécnico Nacional in Mexico, IPN-UPIITA, IPN-ESCOM, COFAA-IPN, IPN-SIP and ICyTDF for their support.

References

[1] Camara, G., Monteiro, A., Paiva, J. and Souza, R., 2000. Action-driven ontologies of the geographical space: beyond the field-object debate. In: M.J. Egenhofer and D.M. Mark, eds. Proceedings of the first international conference on geographic information science, October 2000, GIScience, Savannah, GA. Georgia: AAG, 52–54.

[2] Quintero R., Guzmán G., Menchaca R., Torres M., and Moreno M. 2012. An ontology-driven approach for the extraction and description of geographic objects contained in raster spatial data. *Expert Syst. Appl.* 39, 10 (August 2012), 9008-9020.

[3] Fonseca, F., et al., 2002. Using ontologies for integrated geographic information systems. Transactions in Geographical Infomation Science, 6 (3), 231–257.

[4] INEGI, 1996, Diccionario de datos topograficos 1:50 000 (Vectorial). Agnascalientes: Instituto Nacional de Estadistica Geografia e Informatica. Technical report. ISO/IEC, 2005. ISO/IEC 18025:2005(E) Information technology - environmental data coding speci- fication (EDCS) [online]. Available from: http://standards.sedris.org/ [retrieved: March 2013].

[5] Smith, B. and Mark, D., 2001. Geographical categories: an ontological investigation. International Journal of Geographical Information Science, 15 (7), 591–612.

[6] Sorrows, M. and Hirtle, S., 1999. The nature of landmarks for real and electronic spaces. Lecture Notesin Computer Science, 1661, 37–50.

[7] Timpf, S., 2002. Ontologies of Wayfinding: a traveler’s perspective. Networks and Spatial Economics, 2 (1), 9–33.

[8] Tversky, B. and Hemenway, K., 1984. Objects, parts, and categories. Journal of Experimental Psychology: General, 113 (2), 169–193.

[9] Huang, B., Claramunt, C., Spatiotemporal data model and query language for tracking land use change. Transportation Research Record, 107-113 (2005)

[10] Larson, R.R.: Geographic Information Retrieval and Spatial Browsing, University of California, Berkeley. Available from: https://sherlock.sims.berkeley.edu/geo_ir/PART1.html [retrieved: March 2010].

[11] MexicoTop. articulo locales comerciales [in spanish]. Available from: http://www.mexicotop.com/article/Locales+comerciales [retrieved: February 2013].

[12] MINDMAPS. Geomarketing Estudios de Mercado. Available from: http://www.mindmaps.com/ [retrieved: March 2013].

[13] Davor M. 2007. Market driven product ontologies. In *Proceedings of the 2007 OTM confederated international conference on On the move to meaningful internet systems - Volume Part I* (OTM'07), Robert Meersman, Zahir Tari, and Pilar Herrero (Eds.), Vol. Part I. Springer-Verlag, Berlin, Heidelberg, 275-283.

[14] Grefenstette G. 2010. Use of semantics in real life applications. In *Proceedings of the 19th ACM international conference on Information and knowledge management* (CIKM '10). ACM, New York, NY, USA, 5-6.

[15] Lamparter S. and Schnizler B.. 2006. Trading services in ontology-driven markets. In*Proceedings of the 2006 ACM symposium on Applied computing* (SAC '06). ACM, New York, NY, USA, 1679-1683.

[16] Li Z, Zhao H. and Ramanathan S. Pricing web services for optimizing resource allocation an implementation scheme. In *Proc. of the Web2003, Seattle, WA*, 2003.

[17] Biletskiy Y. and Ranganathan G. 2010. A semantic approach to a framework for business domain software systems. *Comput. Ind.* 61, 8 (October 2010), 750-759.

[18] SIIM (in spanish). Available from: http://www.economia-sniim.gob.mx/nuevo/ [Accessed: March 19, 2013].

A Layered Model for Knowledge Transfer

Felix Schiele, Fritz Laux
Fakultät Informatik
Reutlingen University
Reutlingen, Germany
Felix.Schiele@Reutlingen-University.de
Fritz.Laux@Reutlingen-University.de

Thomas M Connolly
School of Computing
University of the West of Scotland
Paisley, UK
Thomas.connolly@uws.ac.uk

***Abstract*— Knowledge transfer is very important to our knowledge-based society and many approaches have been proposed to describe this transfer. However, these approaches take a rather abstract view on knowledge transfer, which makes implementation difficult. In order to address this issue, we introduce a layered model for knowledge transfer that describes the individual steps of knowledge transfer in more detail. This paper gives a description of the process and also an example of the application of the layered model for knowledge transfer.**

Keywords-knowledge transfer; transfer of knowledge; knowledge conversion; impart knowledge.

I. Introduction

In our knowledge-based society, the relevance of knowledge transfer is increasing. Knowledge management and the understanding of economic coherency can help an organization to handle the challenges of an increasingly fast-evolving environment [1]. The transfer of knowledge from one person to another is of major importance for enterprises [2]. The Socialization, Externalization, Combination, and Internalization (SECI) Model of Nonaka and Takeuchi [3] is an approach that supports organizations in the handling of the important knowledge resource and describes knowledge conversions between internal and external knowledge. However, the SECI Model does not contain precise descriptions of knowledge transfer. This paper aims to introduce a model for knowledge transfer that makes problems emerging during the transfer visible and explainable, and facilitates its implementation through a more detailed and clearer structuring.

This paper is structured as follows: Section II discusses and provides working definitions of data, information and knowledge. Section III discusses existing communications models and Section IV proposes a model of knowledge transfer that aims to reduce errors on each of the knowledge levels. Section V draws conclusions and discusses future directions.

II. Data, Information, Knowledge, Conversation, and Communication

As mentioned by Nonaka [4], the terms information and knowledge are sometimes used interchangeably even though they have different meanings. In her study on the wisdom hierarchy, Rowley [5] pointed out that it is especially important to define the concepts of data, information, and knowledge. Since this paper focuses on the transfer of knowledge, the following section presents definitions to distinguish the terms data, information and knowledge. Having examined various definitions the authors will present their own definitions, which are based on some of the previously introduced ones.

A. Data

Hasler Roumois [6] stated that data consist of symbols that are combined into words by using syntax. The words receive a semantic meaning when they are associated to things. Davenport and Prusak [7] describe data as the raw material for information without an intrinsic meaning. A data set can contain facts about an event or thing. This is also the view of Wormell cited in Boisot and Canals [8] that data are alphabetic or numeric signs that without context do not have any meaning. Rainer [9] characterized data items as "*an elementary description of things, events, activities, and transactions that are recorded, classified, and stored but are not organized to convey any specific meaning*." Ackoff [10] viewed data as "*symbols that represent properties of objects, events and their environment. They are products of observation*." Frické [11] criticized the opinion of those who say that data have to be true, which means that the statement of the data must be true. The following example confirms Frické's criticism: consider a data set containing incorrect or imprecise data, then according to the others this data would not be considered data. Weggeman [12] differentiates between hard and soft data. If the measuring technique and the measurement that created the data are unequivocal, Weggeman describes it as hard data, otherwise the data are softer. Weggeman's classification requires, however, knowledge about the data and the things they represent which is beyond the scope of data, instead part of the scope of information.

1) Definition: data

Data consist of symbols that are combined into words by using syntax. Data are produced by humans or machines. They can be the result of observations of the real world, descriptions of abstract things, or the result of processing existing data. Data cannot be true or false since this decision is beyond the scope of data.

B. *Information*

In the definition of information, there are two fundamentally different theories. The more technical approach characterizes information as data where context has been added [13]. In the more philosophical approach it depends on the receiver whether something is information or only data. Hasler Roumois [6] stated that when people recognize the meaning of data and consider their relevance they become information. Similarly, Davis and Olson [14] view information as data that has been processed into a form that is meaningful to the recipient. Dretske [15] noted about information: "*Roughly speaking, information is that commodity capable of yielding knowledge, and what information a signal carries is what we can learn from it. If everything I say to you is false, then I have given you no information*". However, the recipient of the message may receive the meta information that the other person is lying, Dretske stated. Weggeman [12] provides the example that an author will look at his book as information whereas others may consider it initially as a collection of data. It is up to the receiver to consider whether the data are relevant or not. Weggeman argues that data becomes information even if it is irrelevant to the recipient, because the assessment is a form of recognition that leads to information. As stated in the example from Dretske, the recipient may receive meta information. For this analysis the receiver had to compare the message with his personal knowledge base. If he already knew the content, this may lead to reinforcement by the additional confirmation through the message. Therefore, the authors agree with Dretske that the receiver may achieve meta information, but in this case the data does not become information. Rainer and Cegielski [9] described information as organized data that have meaning and value to a recipient.

1) *Definition: information*

Data becomes information when a person receives data, decodes them, recognizes the meaning and considers them relevant. If the data do not contain anything new for the receiver, the data do not become information. However, they may result in meta information, such as confirmation of the known.

C. *Knowledge*

For the processing of information the existing knowledge is of crucial importance. Wormell, cited in Boisot and Canals [8], believes knowledge is enriched information by a person's or a system's own experience; it is cognitive based; it is not transferable, but through information we can communicate about it. Dretske represents the relation of information and knowledge as follows: "*Knowledge is identified with information-produced (or sustained) belief, but the information a person receives is relative to what he or she already knows about the possibilities at the source*" [15]. About knowledge Polanyi [16] said: "*I shall reconsider human knowledge from the fact that we can know more than we can tell*". Thus he shows that knowledge has a secret or tacit part and not everything a person knows can be passed. Polanyi describes explicit knowledge, which in turn can be expressed in formal, semantic language, and tacit knowledge, which is personalized and therefore hard to express [17]. According to Nonaka [18] explicit knowledge is knowledge that can be articulated into formal language, such as words, mathematical expressions, specifications and computer programmes, and can be readily transmitted to others. This is in contrast to tacit knowledge, which is personalised and based upon experience, context and the actions of an individual; tacit knowledge resides in individuals who may be unaware that they possess such knowledge. There is also implicit knowledge, which refers to knowledge that is revealed in task performance without any corresponding phenomenal awareness; implicit knowledge is often expressed unintentionally. This characteristic is described as type dimension of knowledge [19]. For this article, the explicit type of knowledge represents the most important knowledge type, because it is the knowledge that can be easily externalized. Weggeman [12] firmly believes that information and knowledge only exist inside the person whereas data can exist outside a person. Davenport and Prusak [7] describe knowledge as bound to a person: "*It [knowledge] originates and is applied in the mind of the knowers*." The transformation from information to knowledge takes place when the information is linked to the existing knowledge through a thinking process [6]. The authors propose the term knowledge base as the collection of all facts, rules, and values which are represented in the brain of a person. Spitzer [20] depicts that through the learning process links are created or dissolved in the brain, which results in changes of the knowledge base. Spitzer [20] points out that messages, which have the quality of relevance and novelty, can be memorized easily.

1) *Definition: knowledge*

Information becomes knowledge if a thinking process occurs in which the information is linked to the existing knowledge and is stored persistently. The quality of information being relevant and new, insofar as there is a difference to the existing knowledge, encourages the permanent memorization of information. Based on the input by the information, the knowledge base of the person may be extended or restructured.

D. *Knowledge Conversion*

Nonaka and Takeuchi [3] described the conversation of knowledge in their SECI Model. For this work externalization and internalization of knowledge are of particular importance. Nonaka and Takeuchi describe the internalization as conversion from explicit to tacit knowledge and the externalization as conversion from tacit to explicit knowledge. The authors use the concepts of externalization and internalization with respect to the conversion of data to knowledge and vice versa. Externalization enables a person to converse parts of the personal knowledge base, making them accessible to others. For example, if someone writes down what he knows, everyone except him will refer to this as data. Internalization will happen when a reader receives new knowledge by reading and learning from it.

Transfer and persistent storage require an externalization of knowledge in a recognized and structured language. The various levels of messages are related to levels of semiotics, which are syntactic, semantic and pragmatic. Krcmar [21] states that syntax declares the rules according to which characters can be combined to words and which can be combined to sentences. The relation between words and objects represented by the words as the relationship between characters is denoted by semantics. The intention of a person sending words as a message is explained as pragmatic.

E. *Communication*

The protagonist of systems theory, Luhmann [22], explained communication as a process consisting of three steps of selection. In the first step, the sender decides which information he wants to pass on. In the second step, he selects a single message from many possible messages. In the last step, the recipient selects the information out of the message thereby completing the communication. Based on Luhmann’s work, Berghaus [23] describes several results, which can occur if a sender is forwarding a message to a receiver.

- Case 1: The receiver picks up the message and interprets it in the desired way.
- Case 2: The receiver picks up the message but interprets it differently.
- Case 3: The receiver does not recognize the message as a message.

Only one of the three cases achieves the desired result. In this paper the second case and the various reasons for the error in communication will be considered in more detail. The third case plays a minor role as it is assumed that the message is detected as a message because only the messages presented as data are considered.

III. Related Work: Communication Models

A. *Schema of Social Communication*

Figure 1 shows Aufermann’s [24] model for social communication in which two parties are involved. The sender encodes the statement he intends to submit in a message. Therefore, he uses his own character set to encode the message. The message is sent via a medium to the recipient whereby spatial and temporal distance is overcome. When receiving the message the recipient will use his own character set for the decoding of the message.

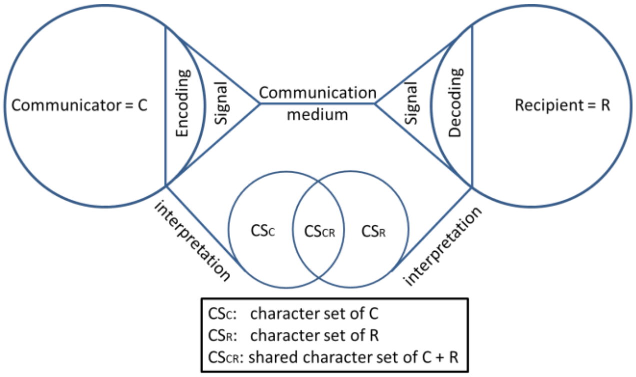


Figure 1. Schema of Social Communication [24] (German)

The model illustrates the important point of the character sets used by sender and recipient and the need to use only those characters that are within the shared character set.

B. *A Mathematical Theory of Communication*

In Shannon’s description of the operation of a communication system, the sender is named “information source” and the receiver is called “destination” [25]. Shannon has investigated the frequency of characters contained in a message, and compared the expected and the actual occurrence of a character. Using the ‘entropy’ Shannon invented a key figure to measure the information contained in a message. Due to the technical use of the model, specifically the control of missiles, the emphasis is on the transmission of the signal [26]. In addition to Aufermann’s schema of social communication, Shannon’s model describes the influence of the transmission of a signal by a noise source.

C. *Four Forms of Knowledge Conversion*

The SECI Model, developed by Nonaka and Takeuchi [3] is focused on the knowledge conversions during knowledge transfer. The description of four conversions takes place at an abstract level showing the particularities of each conversion. However, a detailed description of the individual conversions is missing. Nonaka and Takeuchi describe socialization as a direct knowledge transfer from the tacit knowledge of one person to the tacit knowledge of another person, enabled by action and observation. However, this abstract view does not show exactly how knowledge is transferred in this case. A situation in which socialization happens may arise when master and apprentice work together. Even though the master does not express his knowledge intentionally he externalizes it through his action. Based on the perceived action and the results of action, the apprentice will unconsciously obtain knowledge by internalization.

D. *A Hierarchical Modelling Approach to Intellectual Capital Development*

Ammann [19] describes knowledge conversions from one person to another, in which the different types of knowledge are taken into account. In addition to the knowledge conversions described in the SECI Model the conversion from latent or conscious knowledge to explicit knowledge is described. Even though Ammann’s approach represents knowledge transfer in greater detail, this approach does not give a precise description of how the transmission works.

IV. Model of Knowledge Transfer

A message is a possible way to impart knowledge. The correct interpretation of the message may be prevented by interferences that can affect the message. As described by Shannon the disruption may be caused by a noise source disturbing the medium transmitting the message. In addition to the interferences from the outside that may influence the transport medium, the personal knowledge base of the sender and the receiver may also affect the transfer. The influence of the transfer through the personal knowledge of sender and receiver can take place in four layers. The interpretation of the message depends on the elements that are used and whether they are part of the knowledge base of the receiver and equivalent to the elements of the sender's knowledge base.

A. *Layers that Influence the Transfer*

The four layers that influence the transfer of a message from one person to another are code, syntactic, semantic and pragmatic layer. The concept of a knowledge transfer through different layers was influenced by the OSI Reference Model [27]. Figure 2 illustrates the transfer of a message from the sender to the receiver passing through the four layers.

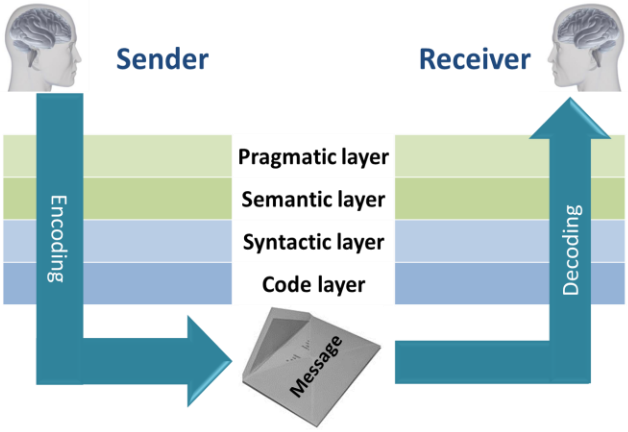


Figure 2. Knowledge Transfer through four layers

1) *Code Layer*

At the lowest level of the layer for transfer is the code. The code consists of symbols or signs that represent the smallest unit, which forms the basis of the higher layers. In the case of written language, which is the focus here, the smallest elements are the characters, σ, taken from an alphabet Σ. In the case of spoken language it would be phonemes, or in sign language gestures.

2) *Syntactic Layer*

The second layer is constituted by the syntax that contains rules for the combination of signs or symbols. In written language, L, the characters σ are combined to form words ω by the use of production rules P.

3) *Semantic Layer*

The third layer contains the semantics that establish the relation between words ω and meaning m. This relation, called semantics s(ω, m), connects the word to its meaning, which can be a real world entity or an abstract thing.

4) *Pragmatic Layer*

The top layer is the pragmatic layer. Pragmatics p(s, c) connects the term represented in semantics with a concept c. The concept contains the course of action and the aims and moral concepts that are represented in the human brain. They influence the thinking and acting of the sender.

B. *Process of a Knowledge Transfer via Messages*

The premise of the following example is the desire of a person, called sender, to communicate something to another person, called receiver. Even if the model is general, the focus is on the written notification.

1) *Sender: Pragmatic Layer*

The core of the message is represented in the pragmatic layer. The aims and moral concepts of the sender do not only affect the externalization of the message, but also the assumptions he makes about the receiver.

2) *Sender: Semantic Layer*

This layer contains all words ω and their relation to the objects. The sender must choose appropriate words that are available in his personal knowledge base. Appropriate means, not only the term which fits best, but also which refer to the knowledge of the recipient.

3) *Sender: Syntactic Layer*

This layer contains the rules P according to which the sentences and terms are made. The words ω chosen to carry the meaning are wrapped in sentences.

4) *Sender: Code Layer*

To transfer the message as written communication the sender has to write the words ω by using characters σ that are part of an alphabet Σ of a language.

5) *Transfer: Message*

The communication medium (e.g. letter, email) transmits the data from the sender to the receiver.

6) *Receiver: Code Layer*

The receiver will view the message and read the characters σ, if he knows them. In the case where the message contains characters from an alphabet unknown to the receiver, the transfer might be disrupted. With only small deviations of the used characters a reconstruction might be possible, otherwise it can lead to misinterpretation or stop the decryption.

7) *Receiver: Syntactic Layer*

The receiver will compose the characters σ to words ω and sentences if they are part of a language L he knows. As in the decoding of the code small difference can be compensated under favourable circumstances, otherwise misinterpretation or stopping the decryption are the consequences.

8) *Receiver: Semantic Layer*

Almost simultaneously with the combination of words and sentences the receiver will put the terms in relation to the things for which they stand. The more the receiver knows the context and the sender of the message, the easier it is to capture the meaning of the text.

9) *Receiver: Pragmatic Layer*

In a final step the receiver will interpret the message in relation to his own aims and values. The things the receiver knows about the sender as well as the assumptions regarding the receiver that are influenced by the sender's own values and aims, play an important role in the decoding of the message.

C. Influence of Overlapping Knowledge

Knowledge about the receiver is an important requirement for a successful and lossless transfer of a message. The better the sender knows the receiver, the easier he can encode the message. A proper encoding of the message can be done by using elements that exist identically in the personal knowledge base of the sender as well as in the personal knowledge base of the receiver. If the receiver is unknown, only assumptions can be made to support the selection. The other way around it is easier for the receiver to decode the message if he knows the sender of the message very well. Figure 3 visualizes the overlapping of the knowledge in different layers.

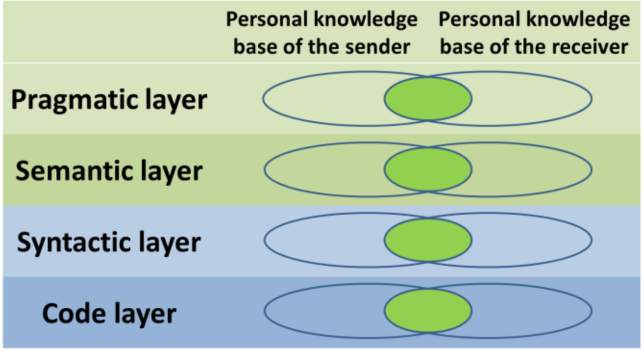

Personal knowledge base of the sender
Personal knowledge base of the receiver
Pragmatic layer
Semantic layer
Syntactic layer
Code layer

Figure 3. Overlapping Knowledge

D. Example of Knowledge Transfer

A challenge in knowledge transfer is the different knowledge base of sender and receiver. In companies, this situation may occur when a business analyst explains a modelled process to a technician in a department. The business analyst, an expert in business process modelling (BPM), will interview the employees of the department to review the department’s processes. During the interview he will make notes and sketches, which he subsequently transfers to business process models.

The business analyst will show and explain the modelled processes to the departmental employees to check that everything has been modelled properly so that model and practised processes are consistent. When explaining the model to the technician, the business analyst must take into account that the technician might not have (sufficient) knowledge of a business process modelling language. We assume that the business analyst and the technician speak the same language and have had similar schooling. Consequently, symbols that exist in their knowledge base are nearly equal although the business analyst might know additional symbols such as those used in the business process modelling languages. This consensus also occurs in the syntactical layer, which contains rules to build words, and the semantic layer, where things are represented through words. The largest differences in the knowledge base are probably found in the pragmatic layer. The basic concepts of aim and moral, that are shaped by education, culture, and environment, may be similar for both. However, the business analyst might have a larger knowledge base in the respective aims and concepts of BPM, while the technician might have a larger knowledge base in the respective aims, processes, and concepts of his special field.

The business analyst, after seeing that the technician has not mastered a business process modelling language, will avoid using terms and concepts unknown to the technician. When explaining the model, the business analyst will introduce the necessary symbols, terms, and concepts to explain the process. He can try to use simple explanations and he can bring in additional information that facilitates the interpretation of the message. The interpretation of the symbols is dependent on the knowledge base of the interpreting person. The interpretation can be facilitated by restrictions; in this example, the terms used for the process are terms from the domain of the department as well as from BPM. The context the terms are used in thereby facilitates the correct interpretation of the process.

V. Conclusion and Future Directions

Knowledge transfer is affected by many different parameters. Because of the relevance of knowledge transfer, it is important to understand the impact of the different parameters. The sociologists Luhmann and Aufermann deal with communication aspects but they neglect the issue of implementation. Shannon’s model focuses on the technical implementation but is restricted to the layers of code and syntax. The model of Nonaka and Takeuchi deals with organizational knowledge and knowledge conversion, but the practical transmission is not considered in detail. Ammann describes knowledge conversions in more detail. However, this model is still too abstract to facilitate implementation. The approach presented in this paper addresses these issues by introducing a model with different layers. The intention behind introducing the layers is to reduce errors on each of the knowledge levels. Thus the process of knowledge transfer is divided into several steps, which can be examined separately. This makes it easier to detect and identify errors and facilitates the prevention of misinterpretation.

The model is to be used for knowledge transfer in the area of business processes. The important knowledge of a company, describing the procedures for the production of products and services, is incorporated in business processes. Due to the fact that business processes represent important corporate knowledge they are an interesting area of application. With respect to the description of the various levels of the model, an appropriate representation will be used. The application of the model on business processes aims to reduce errors both in modelling and analysing business processes.

References

[1] K. Dalkir, Knowledge Management in Theory and Practice: MIT Press, 2011

[2] T. H. Powell and V. Ambrosini, “A Pluralistic Approach to Knowledge Management Practices: Evidence from Consultancy Companies,” Long Range Planning, vol. 45, no. 2–3, 2012, pp. 209–226, http://www.sciencedirect.com/science/article/pii/S002463011200009X, [accessed June 2013].

[3] I. Nonaka and H. Takeuchi, The knowledge-creating company. Oxford: Oxford University Press, 1995.

[4] I. Nonaka, “A dynamic theory of organizational knowledge creation,” Organization Science, vol. 5, 1994, pp. 14–37, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.115.2590, [accessed July 2012].

[5] J. Rowley, “The wisdom hierarchy: representations of the DIKW hierarchy,” J. Information Science, vol. 33, no. 2, 2007, pp. 163–180.

[6] U. Hasler Roumois, Studienbuch Wissensmanagement: Grundlagen der Wissensarbeit in Wirtschafts-, Non-Profit- und Public-Organisationen, translated: Record of Study Knowledge management: Foundations of knowledge work in business-, non-profit- and public-organization, 2nd ed. Zürich: Orell Füssli, 2010.

[7] T. H. Davenport and L. Prusak, Working knowledge: How organizations manage what they know. Boston, Mass: Harvard Business School Press, 2000.

[8] M. Boisot and A. Canals, “Data, information and knowledge: have we got it right?,” Journal of Evolutionary Economics, vol. 14, no. 1, 2004, pp. 43–67.

[9] R. K. Rainer and C. G. Cegielski, Introduction to information systems, 3rd ed. Hoboken, N.J, Chichester: Wiley; John Wiley [distributor], 2011.

[10] R. L. Ackoff, “From Data to Wisdom,” Journal of Applied System Analysis, vol. 16, 1989, pp. 3–9.

[11] M. Frické, “The knowledge pyramid: a critique of the DIKW hierarchy,” J. Information Science, vol. 35, no. 2, 2009, pp. 131–142.

[12] M. Weggeman, Wissensmanagement: Der richtige Umgang mit der wichtigsten Ressource des Unternehmens, translated: Knowledge management: The right way to deal with the most important resource of the company, 1st ed. Bonn: MITP-Verl, 1999.

[13] J. S. Valacich, C. Schneider, and L. M. Jessup, Information systems today: Managing in the digital world, 4th ed. Upper Saddle River, N.J: Prentice Hall, 2010.

[14] G. B. Davis and M. H. Olson, Management information systems: Conceptual foundations, structure, and development, 2nd ed. New York: McGraw-Hill, 1985.

[15] F. I. Dretske, Knowledge and the flow of information. Stanford, CA: CSLI Publications, 1999.

[16] M. Polanyi, The tacit dimension. London, England: Cox & Wyman Ltd, 1966.

[17] I. Nonaka, P. Byosiere, C. C. Borucki, and N. Konno, “Organizational Knowledge Creation Theory: A First Comprehensive Test,” Organization Science, 1994.

[18] I. Nonaka, “The Knowledge-Creating Company,” Harvard Business Review, vol. Nov/Dec91, no. Vol. 69 Issue 6, 1991, pp. 96–104.

[19] E. Ammann, “A Hierarchical Modelling Approach to Intellectual Capital Development,” in Electronic Journal of Knowledge Management Volume 8 Issue 2, C. Bratianu, Ed, 2010, pp. 181–191.

[20] M. Spitzer, Lernen: Gehirnforschung und die Schule des Lebens, translated: Learning: Brain Research and the School of Life, 1st ed. München: Spektrum Akademischer Verlag, 2007.

[21] H. Krcmar, Informationsmanagement, translated: “Information management”, 5th ed. Berlin, Heidelberg: Springer, 2010.

[22] N. Luhmann, Soziale Systeme: Grundriss einer allgemeinen Theorie, translated: Social Systems: Outline of a general theory, 1st ed. Frankfurt am Main: Suhrkamp, 1987.

[23] M. Berghaus, Luhmann leicht gemacht: Eine Einführung in die Systemtheorie, translated: Luhmann made easy: An Introduction to Systems Theory, 3rd ed. Köln: Böhlau, 2011.

[24] J. Aufermann, Kommunikation und Modernisierung: Meinungsführer und Gemeinschaftsempfang im Kommunikationsprozess, translated: Communication and modernisation: opinion leaders and community reception in the communication process, München-Pullach: Verlag Dokumentation, 1971.

[25] C. E. Shannon, “A mathematical theory of communication,” SIGMOBILE Mob. Comput. Commun. Rev, vol. 5, no. 1, 2001, p. 3.

[26] A. Roch, Claude E. Shannon: Spielzeug, Leben und die geheime Geschichte seiner Theorie der Information, translated: Claude E. Shannon: Toys, life and the secret history of his theory of information, 1st ed. Berlin: Gegenstalt, 2009.

[27] H. Zimmermann, “OSI Reference Model--The ISO Model of Architecture for Open Systems Interconnection,” Communications, IEEE Transactions on, vol. 28, no. 4, 1980, pp. 425–432, http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=1094702, [accessed May 2013].

Data Types in UML and OWL-2

Jesper Zedlitz
Christian-Albrechts-Universität
Kiel, Germany
Email: jze@informatik.uni-kiel.de

Norbert Luttenberger
Christian-Albrechts-Universität
Kiel, Germany
Email: n.luttenberger@email.uni-kiel.de

***Abstract*—Both OWL-2 and UML static class diagrams lend themselves very well for conceptual modeling of complex information systems. To ease the choice between either of these languages it worthwhile to clarify the differences and similarities in the representation of different kinds of datatypes (primitive types, enumerations, complex datatypes, and generalization of datatypes) in static UML data models and OWL-2 ontologies. Where similarities allow a transformation of datatypes from one language into the other, we describe a possible transformation.**

***Keywords*—UML; OWL; datatypes; conceptual modeling**

I. Introduction

Though nowadays the Web Ontology Language (OWL) is mostly considered as a language for knowledge representation, it can also be used as a language for conceptual modeling of complex information systems, i.e., as a language for representing the entities of a certain domain and for expressing the meaning of various, usually ambiguous terms and to identify the relationships between these. In this respect, OWL can be seen as a direct “competitor” to static Unified Modeling Language (UML) class diagrams, which are—e.g., in the ISO 191xx series of standards—often used for this purpose. Known approaches for UML-to-OWL (and reverse) transformations having this “conceptual modeling focus” mostly neglect the subtle problem of datatype mapping, i.e., the mapping of the OWL type system to the UML type system and reverse. This aspect should however not be ignored, especially as OWL-2 comes with an elaborate support for datatype properties. In this paper, we focus on these datatypes: We highlight the different representation of datatypes in UML and OWL-2 and present possibilities and limitation of transforming datatype definitions written in language into the other language.

The paper is organized as follows. We start with an overview of existing approaches for transformations between UML and OWL-2 that take datatypes into account. Section III gives an general overview of different kind of datatypes. The next section shows how these datatypes can be represented in UML and OWL-2. In Section V, we present how datatypes defined in one language can be transformed into the other language. How we use the transformation described before is shown in Section VI. Section VII concludes and points out fields of future work.

II. Related work

Several publications discuss the relation of UML and OWL in general [1] [2] and transformations of UML class diagrams into OWL ontologies [3] [4]. The revision of OWL-2 made it necessary to rework the transformations from UML to OWL-2 [5] [6]. However, datatypes play only a minor role in these publications. The approaches have in common that UML attributes with a primitive type are transformed into OWL-2 data properties. Enumerations become enumerated datatypes (`owl:oneOf` resp. `DataOneOf`) and vice versa.

Tschirner et al. [7] note that datatypes can be structured. However, this fact and its impact on a transformation is not further discussed in the article. It is noteworthy that—in difference to all other approaches—enumerations are transformed into sets of individuals instead of set of literals.

A special kind of extensible enumeration defined in the ISO 19103 standard “Codelist” and its transformation from UML to OWL-2 has been discussed in [8].

III. Introduction to datatypes

In general, a datatype consists of three components: the value space, the lexical space, and a well-defined mapping from the lexical into the value space. The value space is the—possible infinite—set of values that can be represented by the datatype. The lexical space describes the syntax of the datatype’s values. The mapping is used to map syntactically correct values to elements of the value spaces. It is possible that—even infinite—many syntactically different values are mapped to the same element of the values space.

Primitive datatypes do not have an internal structure. Examples of primitive types are character strings, logical values, and numbers.

Enumerations are a special kind of datatypes with no internal structure. In contrast to general primitive types the lexical space and the value space of an enumeration are equal-sized, well-defined finite sets. The mapping from lexical to value space is a one-to-one mapping. An example for an enumeration datatype are the English names of the days of the week which consist of seven possible values.

In contrast to primitive data types **complex data types** have an internal structure. These are some examples for complex data types:

- a person’s name consisting of given name and family name
- a physical measurement consisting of value and unit of measurement
- an address consisting of street name, house number, postal code and city name

Generalization of datatypes can be defined similarly to the generalization of element types. If a datatype A generalizes a datatype B each date that is instance of B (i.e., its lexical representation belongs to the lexical space of B and its value belongs to the value space of B) is also instance of datatype A. For example the integers generalize natural numbers. Each natural number is also an integer.

IV. Representation of Datatypes

A. *Unified Modeling Language (UML)*

Besides a few pre-defined primitive types UML allows the definition of additional datatypes in class diagrams. These can be primitive types, complex datatypes, and enumerations. In UML, datatypes—similar to classes—can have owned attributes (as well as operations which are not discussed here). Therefore, they can be used to describe structure. Figure 1 shows examples for the three kind of datatypes.

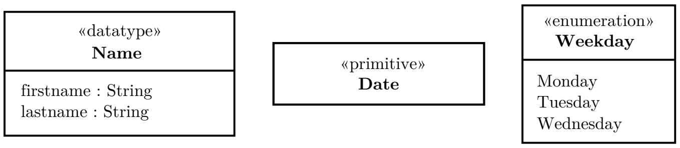


Fig. 1. Examples for datatypes in UML. Left: user-defined datatype with two components. Center: user-defined primitive datatype. Right: Enumeration with three allowed values.

In contrast to instances of classes "any instances of that data type with the same value are considered to be equal instances."[9, p. 63] Although the graphical representations of datatypes in general (instances of *DataType*) as well as primitive types (instances of *PrimitiveType* and enumerations (instances of *Enumeration*) in particular look similar to the representation of classes (instances of *Class*) they are different elements of the meta model as shown in Figure 2.

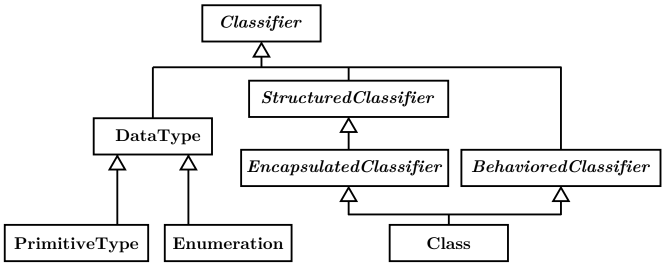


Fig. 2. Extract from the UML meta model, showing the difference between classes and datatypes.

In UML, generalizations are defined for *Classifier* and therefore also for *DataType*. Thus inheritance/generalization relations between datatypes can be defined in a UML class diagram.

B. *Web Ontology Language (OWL-2)*

In OWL-2 three different kinds of datatypes can be distinguished:

1) `rdfs:Literal` as base datatype
2) datatypes of the OWL-2 datatype map, which is basically a subset of the XML Schema datatypes [10].
3) datatypes that have been defined within an ontology using `DatatypeDefinition`

The value space of the base datatype `rdfs:Literal` is the union of the value spaces of all other datatypes. The OWL-2 datatype map adopts the value space, lexical space, and the restrictions for user-defined datatypes from the XML Schema specification. Sets of values (instances of datatypes)—so called *Data Ranges*—can be defined by combining datatypes via common set-theoretic operations. A set of values consisting exactly of a pre-defined list can be described by using `DataOneOf`. A `DatatypeRestriction` allows to define a set of values by restricting the value space of a datatype with *constraining facets*. The OWL-2 datatype map defines which restrictions are allowed. For example a number datatype can be restricted by: less equal, greater equal, equal, and greater.

An OWL-2 datatype is defined by assigning an Internationalized Resource Identifier (IRI) to a `DataRange` using a `DatatypeDefinition` axiom. According to the OWL-2 DL specification this IRI must have been declared as the name of a datatype.

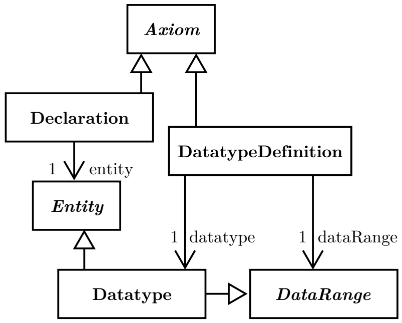


Fig. 3. Extract relevant for datatypes from the OWL-2 meta model.

The abstract syntax (see Figure 3) shows that a datatype is linked indirectly (via an instance of `DatatypeDefinition`) with its value space (an instance of a subclass of `DataRange`. Therefore, it is possible to use a datatype with no assigned values space. By definition this datatype has the value space of `rdfs:Literal`.

Subclasses of `DataRange` (e.g., `DataUnionOf`) which are used for the definition of value sets (and therefore datatypes) have references to `DataRange`. `Datatype` is a subclass of `DataRange`, too. Thus, arbitrarily nested constructions of datatype-defining elements are possible.

V. Transformation of Datatypes

A. *Primitive Types*

Three cases have to be considered for the UML → OWL-2 transformation of primitive types:

1) The datatype is one of the four pre-defined datatypes "Boolean", "Integer", "String", or "UnlimitedNatural".
2) The datatype is one of the XML Schema datatypes.
3) The definition of the (user-defined) datatype is part of the UML-model.

Since OWL-2 uses the datatype-definitions from XML Schema a datatype in case (1) can be transformed into its corresponding datatype from XML Schema. Primitive types can be recognized by the fact that they are contained in a packet “UMLPrimitiveTypes”.

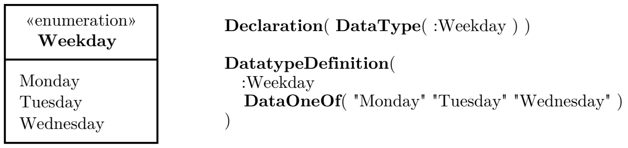
The transformation in case (2) is even more obvious because a datatype is used that is also present in OWL-2. The name of the package containing the primitive types depends on the UML type library used. A common package name is “XMLPrimitiveTypes”. This name can be used to recognize primitive types falling under case (2). The XML Schema datatype can be referenced in the ontology by adding the XSD namespace to the type’s name.

For user-defined datatypes in case (3) a new datatype is defined in the ontology by using a `Datatype` axiom. OWL-2 datatypes—like all OWL-2 model elements—are identified by unique IRIs. Therefore, an appropriate IRI must be generated during the transformation. In UML, elements (including datatypes) are uniquely identified by their name and package hierarchy. Therefore, a combination of package and datatype name can be used for the IRI.

For the transformation OWL-2 → UML primitive types are difficult. OWL-2 offers a variety of possibilities to define new datatypes. However, some primitive types—and probably the most common ones—can be transformed. The primitive types of OWL-2 derive from the XML Schema datatypes. There are established UML-libraries for the XML datatypes. Therefore, it is sufficient to include such a library into the transformation process. An instance of a primitive type contained in the library can be looked up by the IRI of the OWL-2-datatype and references as necessary.

B. Enumerations

As mentioned in Section II, several authors have already discussed how to transform enumerations: In OWL-2 the data range `DataOneOf` is suitable for defining a datatype with a fixed pre-defined value space. Each lexical value of the `DataOneOf` data range is transformed into an *EnumerationLiteral* instance and vice-versa. OWL-2 as well as UML support the specification of datatypes for the elements of an enumeration: An OWL-2 *Literal* instance has a *datatype* attribute, an UML *EnumerationLiteral* instance has a *classifier* attribute referencing the datatype.



```
Declaration( DataType( :Weekday ) )

DatatypeDefinition(
  :Weekday
  DataOneOf( "Monday" "Tuesday" "Wednesday" )
)
```

Fig. 4. Example for the transformation of an enumeration.

For the transformation OWL-2 → UML one has to consider the fact that in OWL-2 the data range `DataOneOf` can be used without a `DatatypeDefinition` which assigns a name to it. Since an UML *Enumeration* necessarily needs a name it can be generated based on the literals contained in the data range.

C. Complex Data Types

OWL-2 datatypes consist of exactly one literal and are therefore not further structured. Since OWL-2 is built upon the Resource Description Framework (RDF) there is the theoretical possibility to use a blank node and the RDF-instruction `parseType="Resource"` to implement complex data as shown in this listing:

```
<rdf:RDF xml:base="http://example.com/persons/"
   xmlns="http://example.com/persons/"
   xmlns:owl="http://www.w3.org/2002/07/owl#"
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">

   <owl:Ontology rdf:about="http://example.com/persons/"/>

   <owl:Class rdf:about="Person" />

   <owl:NamedIndividual rdf:about="Timmi">
      <rdf:type rdf:resource="Person"/>
      <hasName rdf:parseType="Resource">
         <first>Timmi</first>
         <last>Tester</last>
      </hasName>
   </owl:NamedIndividual>
</rdf:RDF>
```

However, neither the OWL-1 nor the OWL-2 specification mention `parseType="Resource"`. Therefore, it is probably not a valid construct for OWL-2. Even if this notation was valid for OWL-2 and an element type could be assigned to such an anonymous individual, the definition of the element type would be indistinguishable from the definition of a “normal” element type.

The UML → OWL-2 transformation of complex datatypes, i.e., datatypes with owned attributes, is similar to the transformation of UML classes with owned attributes into OWL-2 classes and properties. There are two characteristics of UML datatypes that have to be considered:

1) Values do not have an identity.
2) Every value exists only once.

Since the transformation is similar to the transformation of classes the instances of the resulting element in the ontology will be individuals. In OWL-2 every (typed) individual must have a name. Therefore, the semantics for characteristic (1) is changed: In UML, the instance of the datatype does not have an identity. The corresponding individual in OWL-2 is assigned with an IRI by which it can be referenced (and also identified).

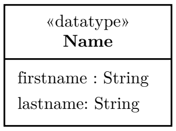
Characteristic (2) requiring that every value must exist not more than once can be ensured by using `HasKey` axioms. For every UML datatype D with owned attributes $a_1 \ldots a_n$ that is transformed into a OWL-2 class C with data property $dp_1 \ldots dp_n$ the following axiom is added to the ontology:

HasKey(C () ($dp_1 \ldots dp_n$))

This axiom ensures that every occurrence of an individual with the same values for $dp_1 \ldots dp_n$ is one and the same individual.

D. Generalization of datatypes

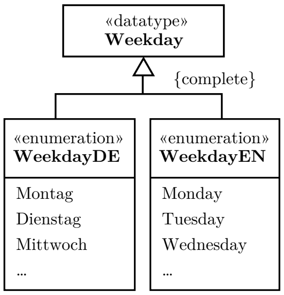
In general, the transformation of a datatype generalization in a UML class diagram is not possible since OWL-2 has no support for inheritance/generalization of datatypes. In the special case of a complete generalization of datatypes with



```
Declaration( Class( :Name ) )

Declaration( DataProperty( :Name_firstname ) )
DataPropertyDomain( :Name_firstname :Name )
DataPropertyRange( :Name_firstname xsd:string )

Declaration( DataProperty( :Name_lastname ) )
DataPropertyDomain( :Name_lastname :Name )
DataPropertyRange( :Name_lastname xsd:string )

HasKey( :Name () ( :Name_firstname :Name_lastname ))
```

Fig. 5. Example for the transformation of a complex datatype.

no internal strcuture (e.g., enumerations) a transformation is possible: While the generalization of UML classes can be transformed into an OWL-2 `ObjectUnionOf` class expression this is not possible for datatypes. As the name suggest, an `ObjectUnionOf` can only be used for classes. Instead an instance of `DataUnionOf` is used. The sub-datatypes combined in the `DataUnionOf` constitute a new data range. Using a `DatatypeDefinition` axiom a name is assigned to this set of datatypes. This name is the name of the super-datatype from UML. Figure 6 shows an example for such a transformation.



```
Declaration( Datatype( :Weekday ) )
Declaration( Datatype( :WeekdayDE ) )
Declaration( Datatype( :WeekdayEN ) )

DatatypeDefinition( :WeekdayDE
    DataOneOf( "Montag" "Dienstag" "Mittwoch" ... )
)

DatatypeDefinition( :WeekdayEN
    DataOneOf( "Monday" "Tuesday" "Wednesday" ... )
)

DatatypeDefinition(
   :Weekday
    DataUnionOf( :WeekdayDE :WeekdayEN )
)
```

Fig. 6. Example for the transformation of a generalization relation between datatypes.

VI. Implementation

We have implemented the transformations presented in this paper as part of two model-to-model transformations written in Meta Object Facility (MOF) 2.0 Query/View/Transformation (QVT) Relations language [11]. In contrast to other approaches working on transfer formats (e.g., XML Metadata Exchange (XMI) and a XML-based syntax for OWL-2) our transformation is specified using the meta-models of UML and OWL-2 only. Therefore, the transformation is independent of any concrete syntax.

QVT-Relations is a declarative model-to-model transformation language. In addition to a textual syntax it also has a visual syntax. An example of the visual syntax is shown in Figure 7. During a transformation execution so called *trace classes* and their instances are automatically created to record what occurred that execution. These characteristics of QVT-Relations make it possible to analyze the transformation easily. Additional to manual/visual automated tests can be performed since the output of one transformation is again a model that can serve as the input of the other transformation.

VII. Conclusion and Future Work

In this paper we have focused on the datatypes of static data models—often neglected when working on transformation between UML and OWL-2. We showed differences and similarities in the representation of datatypes in UML and OWL-2. Where similarities allow a transformation of datatypes from one language into the other we have described a possible transformation and highlighted the tricky points and/or limitations.

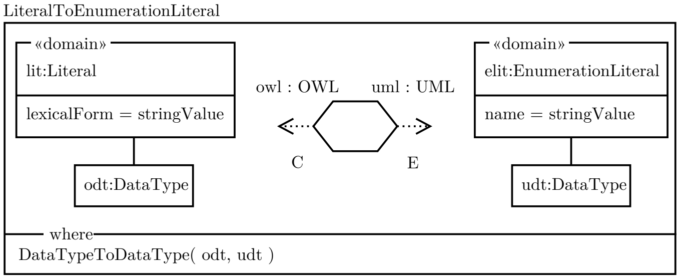


Fig. 7. Example of a QVT-Relations rule from the OWL-2 → UML transformation that maps OWL-2 *Literals* to UML *EnumerationLiterals*.

References

[1] G. Schreiber, “A UML Presentation Syntax for OWL Lite,” 2002. [Online]. Available: http://www.swi.psy.uva.nl/usr/Schreiber/docs/owl-uml/owl-uml.html

[2] K. Kiko and C. Atkinson, “A Detailed Comparison of UML and OWL,” Mannheim, 2008, technical report.

[3] L. Hart, P. Emery, B. Colomb, K. Raymond, S. Taraporewall a, D. Chang, Y. Ye, E. Kendall, and M. Dutra, “OWL Full and UML 2.0 Compared,” 2004. [Online]. Available: http://www.omg.org/docs/ontology/04-03-01.pdf

[4] OMG, “Ontology Definition Metamodel,” Object Management Group, 2009. [Online]. Available: http://www.omg.org/spec/ODM/1.0/

[5] S. Höglund, A. Khan, Y. Liu, and I. Porres, “Representing and Validating Metamodels using the Web Ontology Language OWL 2. TUCS Technical Report No. 973,” Turku 2010. [Online]. Available: http://tucs.fi/publications/attachment.php?fname=TR973.full.pdf

[6] J. Zedlitz, J. Jörke, and N. Luttenberger, “From UML to OWL 2,” in *Proceedings of Knowledge Technology Week 2011*, D. Lukose, A. R. Ahmad, and A. Suliman, Eds., Berlin/Heidelberg, 2012, pp. p. 154–163.

[7] S. Tschirner, A. Scherp, and S. Staab, “Semantic access to INSPIRE,” in *Terra Cognita 2011. Proceedings of the Terra Cognita Workshop on Foundations, Technologies and Applications of the Geospatial Web*, R. Grütter, D. Kolas, M. Koubarakis, and P. D., Eds., 2011, pp. p. 75–87. [Online]. Available: http://ceur-ws.org/Vol-798/proceedings.pdf

[8] J. Zedlitz and N. Luttenberger, “Transforming Between UML Conceptual Models and OWL 2 Ontologies,” in *Proceedings of the Terra Cognita Workshop on Foundations, Technologies and Applications of the Geospatial Web, in conjunction with the 11th International Semantic Web Conference (ISWC 2012)*, D. Kolas, M. Perry, R. Grütter, and M. Koubarakis, Eds., 2012, pp. p. 15–26. [Online]. Available: http://ceur-ws.org/Vol-901/paper2.pdf

[9] OMG, “Unified Modeling Language, Superstructure Version 2.4,” 2011. [Online]. Available: http://www.omg.org/spec/UML/2.4/Superstructure

[10] XMLSchema-2, “XML Schema Part 2: Datatypes,” 2004. [Online]. Available: http://www.w3.org/TR/xmlschema-2/

[11] OMG, “Meta Object Facility (MOF) 2.0 Query/View/Transformation Specification Version 1.1,” 2011. [Online]. Available: http://www.omg.org/spec/QVT/1.1/PDF/

An Overview of Ontology Engineering Methodologies in the Context of Public Administration

Bernd Stadlhofer, Peter Salhofer, Augustin Durlacher
Institute of Information Management
FH JOANNEUM, University of Applied Sciences
Graz, Austria
bernd.stadlhofer@fh-joanneum.at, peter.salhofer@fh-joanneum.at, augustin.durlacher@fh-joanneum.at

Abstract— **This paper gives an overview of latest ontology engineering methodologies that are analyzed in terms of a representative set of criteria and aspects. The portfolio of criteria considers general structural aspects of ontology development (such as strategy for building ontologies) as well as project management aspects (such as recommended process model or the consideration of collaborative construction). While the study criteria principally stay generic we particularly try to include possible characteristics of the E-Government domain. Whereas the study shows that none of the discussed methodologies is fully mature to serve as a domain expert centered ontology engineering methodology in the context of electronic service provisioning in public administration, it also outlines the potential of the discussed methodologies to which extend they can contribute to a new methodology in this field.**

Keywords-Ontology Engineering Methodologies; E-Government; Comparative Study.

I. Introduction

In the context of public administration, ontology engineering is mainly involved when applying Semantic Web technologies to E-Government and to the electronic public service provisioning process respectively. Actually, this domain has become an important field of research aiming at enhancing transparency, interoperability as well as citizen orientation of public agencies [1]. In fact, developing formal ontologies is a complex task that requires having significant skills in software and knowledge engineering in order to being able to design, implement, and maintain ontologies. Beyond this, it requires domain expertise in order to verify the correctness of domain specific ontologies. Whereas a domain expert (in the context of public administration we use the term domain expert synonymously with legal expert) possesses in-depth knowledge of the specific domain to be modeled she or he is very likely not to have sufficient ontology engineering skills at the same time. In order to reduce the complexity of ontology engineering for domain experts methodological guidelines assisted by intelligent tooling have to be applied.

When reviewing literature on semantic E-Government initiatives listed in [1], only one initiative (i.e., [2]) mentions explicitly a specific ontology engineering methodology for designing semantic models. Hence, relevant ontologies in the E-Government sector still tend to be built rather on an ad hoc basis than following a well-defined engineering process supported by adequate tools (similar observations are documented in [3]). As a consequence, the actual ontology engineering is rather done by software engineers than by domain experts. This circumstance also fosters the effect that E-Government projects, generally, suffer from unsustainable activities in the organizational environment, e.g., external stakeholders (such as knowledge and software engineers) leave after a project ends [4].

An essential aspect in the context of public administration is the consideration of legal certainty. In fact, in ontology engineering for the public administration sector, legislation and enforcement of law on all governmental levels has to be ensured. This requires collaboration with a variety of different legal experts. In many cases, constraints probably might have to be weakly encoded, supported by textual explanations and links to further information and supporting bodies. This is necessary to reflect the special demands of a legal system and to safeguard legal certainty [5]. An ontology engineering methodology in the context of public administration should explicitly include steps for domain experts that deal with this circumstance. Hence, not only the validation of the formal model consistency has to be considered but also steps for a simple validation of legal aspects have to be applied.

Consequently, it is our claim to support domain experts of public administration with sufficient guidelines, which enables them to design, implement, verify and maintain their semantic artifacts by themselves. A first step towards this goal is to review the state of the art in this field of research. Hence, the aim of this paper is to give an overview of general ontology engineering methodologies available, having a focus on impacts on E-Government in particular. The resulting overview should give valuable input to our overall research goal, which is to establish a "domain expert centered ontology engineering methodology in the context of electronic service provisioning in public administration".

The remainder of this paper is organized as follows: In Section II, the specific criteria and aspects for the study are introduced followed by a description of a selected list of methodologies in terms of the presented analysis aspects in Section III. Section IV briefly compares the selected methodologies according to the chosen criteria. Section V lists related work in the field of evaluating ontology engineering methodologies. In Section VI, the work is concluded with the summarization of the analysis' results.

II. Analysis aspects

In this section, the concrete aspects and criteria considered for the study are discussed. The portfolio of criteria considers general structural aspects of ontology development (such as strategy for building ontologies) as well as project management aspects (such as recommended life cycle) and aspects directly related to the E-Government domain (such as collaborative construction among public authorities or consideration of legal certainty):

1) Strategy for building ontologies: With this aspect it is examined which strategy is used to develop ontologies. Is it, a) application-dependent, which means that resulting ontologies are designed for and usable by a specific semantic application only, b) application-independent, which means reuse of resulting ontologies is maximized by developing general-purpose descriptions, or c) application-semidependent, where possible scenarios of ontology use are somehow limited [6]. Generally, concerning the strategy for building ontologies it can be argued that "… the more an ontology is independent of application perspectives, the less usable it will be. In contrast, the closer an ontology is to application perspectives, the less reusable it will be" [8]. Thus, there is always a trade-off between application-dependent approaches that typically add some extra value to ontologies, since they can be immediately used in a particular context, and more general application-independent strategies that allow for simplified reuse in different contexts.

2) Recommended process: This aspect examines the existence or recommendation of specific process models one has to go through in order to model ontologies, e.g., being aligned along the general waterfall phases or following iterative, cyclical or agile development models.

3) Consideration of collaborative construction: Modeling an ontology of some public administration domain generally requires numerous authorities to be involved. Since these experts from different public agencies are typically locally distributed, it is simply not possible to develop and maintain all relevant information at one central point [5], particularly when the modeled domain is rather complex.

4) Tool-support: Does a specific tool explicitly support the methodology in question? In terms of enabling domain experts to model ontologies, providing context is a major goal in order to reduce complexity of the modeling process [9]. Specific methodological guidelines combined with intelligent and human-centered tooling should overcome a possible lack in engineering skills.

5) Target group: For what group of people is the methodology primarily designed? Traditionally an ontology engineering methodology is intended for knowledge and ontology engineering experts whereas domain experts are only involved in the knowledge elicitation phase. In contrast, we consider domain experts as the primary target group that should have a maximum of responsibility during the whole ontology life cycle. Only domain experts possess the respective knowledge to be modeled and the expertise to ensure legal certainty of resulting artifacts. Centralizing domain experts in the ontology engineering process should also boost sustainable development of semantic initiatives in E-Government.

Whereas aspects 1 and 2 rather aim at enabling a structured discussion of the selected methodologies, aspects 3 to 5 represent methodological requirements, which have been identified by conducting a number of expert interviews with representatives of public agencies on municipal and federal level in Austria. The first three aspects have been derived from existing comparative studies (i.e., [6]). However, what is new in our approach is to specifically focus on analyzing aspects of human-centered computing in ontology engineering. Beyond these 5 aspects, we initially identified some more (e.g., reuse), however, analysis showed that the investigated methodologies could not be differentiated along these aspects. According to Lutz and Stelzer [7], only criteria that enable differentiations between the target objects should be used in comparative studies. This is why such aspects have been removed from the final analysis.

III. Methodologies

Literature research resulted in a list of 20 documented ontology engineering methodologies mainly reported by Casellas [10]. The methodology developed by Uschold and King [11] can be considered the first approach towards developing a methodology for building ontologies. This methodology builds the foundation for many other approaches that have emerged over the last couple of years. In this study we discuss a selected list of these methodologies based on meeting the requirements represented by aspects 3 to 5 (c.f. Section II).

Initially, we planned to exclude all methods that do not meet all of these 3 requirements to the following extend.

- Aspect 3: There have to be at least recommendations for a collaborative development process.
- Aspect 4: There is at least one modeling tool available that explicitly supports the methodology.
- Aspect 5: The methodology has to focus on the domain expert as the major target group in the modeling process.

In fact, one single methodology in question (i.e., Methodology 4) fulfills all three of these requirements in a reasonable way. Consequently, we revised the exclusion criteria that to be included in the study the methodology under question at least has to address one out of these three methodological requirements.

A. *Methodology 1*

Holsapple and Joshi [12] present a "Collaborative Approach to Ontology Design". The authors discuss the fundamental importance of ontological commitment, which is "…the agreement by multiple parties (persons and software systems) to adopt a particular ontology when communicating about the domain of interest. … Working toward ontological commitment should not be an afterthought, but rather an integral aspect of ontological engineering. This contention underlies the collaborative approach to ontology design we advocate." [12].

The methodology suggests four phases in the ontology engineering process (aspect 2): Preparation, Anchoring, Iterative Improvement, and Application. In the preparation phase design criteria are defined, boundary conditions and evaluation standards are determined. This aims at both, guiding development of the ontology and assessing the degree of its success. In the anchoring phase an initial version of the ontology is created serving as an anchor to help focus the attention of collaborators. In phase 3 the approach uses an adaption of the Delphi method, which is a formal technique for integrating the individual opinions of a group of experts on some topic. "This gives a systematic way for gathering perspectives and critiques on an ontology as a basis for iterative improvement" [12]. Finally, in phase 4, the ontology is explored in various ways in order to prove the ontology utility. Thereby, the authors do not report about a concrete dependence on a specific semantic application, which leads to an application-independent methodology (aspect 1).

By using the Delphi method a clear and structured collaboration process is introduced. However, this form of collaboration tends to be rather inflexible and heavyweight as feedback collection is coordinated centrally by a control board and not interactively and immediately shared by all participants. The validation of legal certainty is not explicitly addressed. This can be defined as evaluation standard in phase 1 and therefore also be included in iterative Delphi rounds, though (aspect 3). As the name suggests the methodology concentrates on the collaboration aspect only and does not include any guidelines concerning the actual modeling process, which is therefore naturally conducted by classical ontology engineers (aspect 5). Tool-support does not exist (aspect 4).

B. Methodology 2

DIstributed, Loosely-controlled and evolvInG Engineering of oNTologies (DILIGENT) represents a methodology that focuses on the evolution of ontologies instead of the initial design. Thus, the methodology supports an evolutionary lifecycle (aspect 2). It focusses on user-centric ontology development and provides integration of automatic agents in the process of ontology evolution [13].

The process starts with various stakeholder-groups (domain experts, users, knowledge engineers, ontology engineers) building together an initial version of the ontology. The initial version results from a rather quick consensus about some high-level terms among all participants. Subsequently, users start to work with the ontology and locally adapt (by sub-classing) it to their specific needs. A control board collects change requests to the shared core ontology. The control board then analyses the various local ontologies, tries to find similarities and introduces a new version of the shared ontology. The control board also regularly revises the shared ontology in terms of not diverging too far. Ontology engineers are responsible for maintaining the ontology based on the board's decisions. Users can then locally update the local ontologies in terms of reusing new terms instead of using their previously defined local terms [13].

With this approach reuse should be maximized among all users whereas not narrowing usage in different application scenarios (aspect 1).

The authors argue that decentralized knowledge management systems are getting increasingly important and therefore emphasize distributed and collaborative construction (aspect 3).

Domain experts in a distributed setting are supported by a fine-grained methodological approach based on the Rhetorical Structure Theory [14]. A standard Wiki is used to allow a traceable discussion. Snapshots of the ontology agreed on are imported to the Wiki, in order to visualize the ontology and ease the discussion of it (aspect 4).

DILIGENT involves numerous different user groups in the engineering process, namely domain experts, users, knowledge engineers and ontology engineers. In the revision phase domain experts are responsible for evaluating an ontology from a domain point of view (does it represent the domain, or does it contain factual errors?). This may also include the validation of legal certainty as necessary for public administration. In fact, the methodology was also applied at the development of an ontology for professional legal knowledge [15]. However, the actual ontology implementation is still intended for ontology engineers (aspect 5).

C. Methodology 3

Very similar to DILIGENT (Section III.B), the Human-Centered Ontology Engineering Methodology (HCOME) [16] supports the development and evaluation of "living" ontologies in the context of communities of knowledge workers. The authors mention common impediments for knowledge workers (or domain experts) to participate actively in ontology engineering: they are unfamiliar with formal representation languages and knowledge engineering principles as well as with methods and techniques for constructing and synthesizing ontologies. The main goal of HCOME therefore is to empower domain experts to evolve their formal conceptualizations in their day-to-day activities. Thus, this methodology focuses on the active participation of domain experts in the ontology life cycle (aspect 5). For this purpose, the authors also developed a Human Centered Ontology Engineering Environment (HCONE), which directly supports the development of ontologies following the HCOME methodology (aspect 4).

The methodology proposes specification, conceptualization and exploitation as the three life cycle phases of ontology engineering. All involved tasks are performed iteratively, until a consensus has been reached between the participants (aspect 2). In the specification phase, knowledge workers are joining groups aimed at developing shared ontologies. Workers are discussing requirements in a shared space, produce documents and agree on the aim and the scope of a new ontology [16]. Consequently, the aspect of collaborative development is directly addressed by this methodology (aspect 3). In the conceptualization phase, workers can follow any approach to the development of ontologies in their personal space. In the exploitation phase shared ontologies can be used in the context of specific ontology-driven applications and settings. However, the overall methodology is application-independent, as it doesn't give a recommendation for a specific semantic application to use (aspect 1). The evaluation and further development of personal ontologies are achieved via a recorded structured conversation in order to enable the tracking of changes and decisions.

D. Methodology 4

The "Integrated Modeling Methodology" [17] principally guides the process of creating application domain dependent parts of an organizational learning system named Advanced Process- Oriented Self- Directed Learning Environment (APOSDLE). The methodology consists of four main phases: Scope & Boundaries, Knowledge Acquisition, Modeling of Domain, and Modeling of Learning Goals. Validation & Revision is included as individual activity in all of the main phases (aspect 2). The resulting semantic artifacts are directly applied and exploited in the APOSDLE system [26], which leads to an application-dependent approach (aspect 1). Domain experts are considered to be an important stakeholder group and mostly included in the knowledge acquisition phase. The knowledge acquisition is performed with well-known state-of-the-art techniques like, interviews, card sorting, laddering, and concept/step/section listing. The authors thereby mention the problem that domain experts are often rarely available and scarcely motivated towards modeling [17] (aspect 5).

The methodology is explicitly supported by the so-called Modeling WiKi (MoKi), which allows users to describe semantic artifacts in an informal but structured manner using natural language. The subsequent automatic translation into formal models does not require the users to have in-depth formal modeling skills (aspect 4). The Wiki nature of the MoKi naturally enables a collaborative tool that provides support for domain experts with hardly any knowledge engineering skills to model domains directly. However, the methodology suggests that domain experts, knowledge engineers and experts (coaches) collaboratively work in a rather agile modeling process (aspect 3).

E. Methodology 5

Klischewski and Ukena [2] present a methodology that aims at the design of semantic E-Government services driven by user requirements. The authors suggest a step-by-step design process that signals public administration authorities to focus on the intended common understanding of citizens concerning the description of public administration services' interfaces. Generally, the authors describe the aim of the design of semantic structures in E-Government as: to support informational needs during service processing, to capture domain knowledge and to support technical implementation.

In contrast to other approaches that focus on knowledge-driven or domain-driven design, this methodology focuses on requirements-driven design that should emphasize what users or providers will consider as valuable information [2].

The proposed seven steps for the development of semantic E-Government services are: Identify informational needs, identify required information quality, create glossary of topics and terms, create controlled vocabulary, group and relate terms, design an ontology, implement semantics (aspect 2). While these steps themselves are generic, the authors also give some concrete examples how the specifics of E-Government are addressed.

As already mentioned, the methodology itself is rather generic. Nevertheless the authors use Web Service Modeling Ontology (WSMO) [27] as semantic execution environment in their pilot scenario. However, the authors do not exclude any other semantic execution environments as, e.g., Web Ontology Language for Web Services (OWL-S) [28], which leads to an application-independent approach (aspect 1).

The authors mention the fact that in service provisioning of the public administration domain a large number of different authorities might be involved. The aspect of collaborative construction (aspect 3) is not covered by the proposed methodology, though.

Besides IT specialists also domain experts of public administration are identified as an important stakeholder group who are responsible for establishing a common understanding of the service interface, analyzing information demand and quality requirements as well as determining topics, terms and relations to be used. This methodology is directly intended for the public administration sector. Consequently it should also consider essential legal aspects of respective public administration domains. However, the methodology does not include any validation step where domain experts could ensure legal certainty. The actual ontology design is conducted by classical ontology engineers (aspect 5).

The authors do not mention any tool support for the methodology (aspect 4).

F. Methodology 6

Developing Ontology-Grounded Methods and Applications (DOGMA) [18] represents an ontology engineering methodology that is aimed at building both highly reusable and usable ontologies. Concerning aspect 1, this is the only methodology that covers both application-dependence as well as application-independence in one approach and highlights the importance of developing reusable as well as usable ontologies. This goal is reached by introducing a shared ontology base that consists of "plausible" domain axiomatizations and application axiomatizations. Application axiomatizations consist of a selected set of lexons from the ontology base and a specified set of rules to constrain the usability of these lexons [18].

Development is supported by the so-called DOGMA Studio Workbench (aspect 4) that also provides plugins for a community layer that aims at supporting the DOGMA-MESS methodology [19]. DOGMA-MESS emphasizes on providing guidelines for collaborative and interorganizational ontology engineering (aspect 3). Thereby, the authors discuss that “a viable methodology requires not building a single, monolithic domain ontology by a knowledge engineer, but supporting domain experts in gradually building a sequence of increasingly complex versions of interrelated ontologies over time”.

The process of ontology building is hierarchically structured. Every domain has a so-called Upper Common Ontology that is maintained by the core domain expert. The most important artifacts of this ontology are templates that describe a common knowledge definition. Over time, templates should become more numerous and should evolve during multiple iterations of development. Templates are then specialized into organizational specializations by the domain experts representing different organizations. The authors present a so-called Lower Common Ontology for negotiating the meaning of specific terms (aspect 2).

The authors refer to the importance of domain experts in interorganizational ontology engineering and also include human-centered aspects in the respective software tool (aspect 5).

G. *Methodology 7*

In contrast to other ontology engineering methodologies, the NeOn methodology [20] does not define a rigid process to follow, but instead, it suggests a variety of pathways for developing ontologies. It defines nine different scenarios, a glossary of processes and activities, two ontology life cycle models (waterfall life cycle model, iterative-incremental life cycle model) as well as a set of methodological guidelines for different processes and activities (aspect 2) [20].

The authors discuss the fact that due to the increase of online available ontologies ontology development is more and more becoming a reuse-centric process. Consequently, ontology development can be characterized as the construction of a network of ontologies, managed by different people and different organizations. Thus, the proposed methodology particularly aims at providing support for the collaborative construction of ontology networks (aspect 3) [20].

The methodology is intended for the classical ontology engineer who is defined as software developer or ontology practitioner involved in the development of ontologies. Hence, the methodology does not include any guidelines for non-experienced domain experts to autonomously develop ontologies. However, the methodology includes a well-elaborated evaluation activity, which could also incorporate safeguarding legal certainty (aspect 5).

The NeOn toolkit provides explicit support for developing ontologies following the proposed methodology (aspect 4).

As it is the aim of the authors to define a generic framework for the development of ontologies, it is completely application-independent (aspect 1).

IV. Results

As shown in Table I, it can be observed that most of the methodologies suggest a rather generic and application independent approach to ontology engineering. In contrast, Methodology 4 is developed for a specific domain (organizational learning) and system (APOSDLE) aiming at a rapid application of developed ontologies. Additionally, Methodology 6 discusses that both usability and reusability of ontologies are important. Hence, this methodology focuses on application independence as well as application dependence.

The recommended processes and life cycles range from classical waterfall development, to iterative and incremental development. In our opinion, in this context no approach can be seen as better than another. Whereas most of the presented methodologies recommend only one procedural model Methodology 7 defines several of them. The method describes use-cases that should help to identify the most appropriate process for a given situation.

Most of the investigated methodologies name collaborative construction of ontologies as an essential goal. Methodology 1, Methodology 2, Methodology 3, and Methodology 4 include explicit assistance aiming at structured conversations between all participants. Whereas the first two follow a rather centralized approach with a control board that manages inputs from participants, the other methodologies prefer a more interactive and agile approach resulting in faster response times.

TABLE I. Summary of study results

Methodology	Asp. 1	Asp. 2	Asp. 3	Asp. 4	Asp. 5
1	Appl. ind.	Iterative	Yes	No	OE
2	Appl. ind.	Iterative	Yes	Wiki	OE/DE
3	Appl. ind.	Iterative	Yes	HCONE	DE
4	Appl. dep.	Agile	Yes	Moki	OE/DE
5	Appl. ind.	Waterfall	No	No	OE/DE
6	Appl. ind. Appl. dep.	Iterative	Yes	DOGMA Studio	OE/DE
7	Appl. ind.	Waterfall Iterative	Yes	NeOn Toolkit	OE

OE: Ontology Engineer; DE: Domain Expert

Many methodologies identify the domain expert as a crucial participant in the ontology engineering process. For example, Methodology 6 discusses that “… an interorganizational ontology needs to be modeled not by external knowledge engineers, but by domain experts themselves. Only they have the tacit knowledge about the domain and can sufficiently assess the real impact of the conceptualizations and derived collaborative services on their organization. …” However, it is interesting to observe that only Methodology 3, Methodology 4, and Methodology 6 offer explicit support for domain experts to model the respective ontologies, or at least parts of it, autonomously.

V. RELATED WORK

A very comprehensive comparative study that presents the most representative methodologies used in ontology development at that time was conducted by Fernandez-Lopez [6]. The study analyses methodologies against the IEEE Standard for Developing Software Life Cycle Processes (1074-1995). The author already mentions a criterion for collaborative and distributive construction but comes to the conclusion that none of the publications at that time cover this aspect explicitly.

A very similar study has been conducted by Fernández-López and Gómez-Pérez [21] that additionally introduces a methodology categorization. The categorization includes methodologies for building ontologies from scratch, methodologies for reengineering ontologies and methodologies for collaborative construction.

The study by Beck and Pinto [22] gives a rather informal overview of methodologies for ontologies. The paper emphasizes aspects like "consider reuse" and also mentions life cycles and typical ontology engineering activities (e.g., specification, conceptualization, formalization, implementation, or maintenance).

Corcho, Fernández-López, and Gómez-Pérez [23] additionally describe ontology tools and ontology languages available at that time. The authors come to the conclusion that future work should be driven towards the creation of a common workbench that supports ontology development during the whole life cycle, ontology management, ontology support as well as methodological support for building ontologies.

Sandkuhl [24] provides an analysis of ontology development methodologies in the context of small and medium-sized enterprises (SMEs). The study focuses on reducing development time for building ontologies. Thus, the study analyses aspects like completeness of the methodology, life cycle coverage and reuse of already existing ontologies.

Kim and Choi [25] present an evaluation of ontology development methodologies with CMM-i. Although the idea of taking CMM-i (a very comprehensive framework for organizations to assess their development and maintenance processes) as evaluation framework sounds promising the actual study does not present many valuable results.

Casellas [10] presents the latest approach in the field of comparing ontology engineering methodologies. The article can be seen as a recommendation in terms of listing most of the relevant methodologies currently available. However, the analysis tends to focus only on the followed life cycle of the studied methodologies and lacks in taking more analysis criteria into account.

VI. CONCLUSION

Considering aspects and requirements of the E-Government domain described in Section I and II, we come to the conclusion that none of the analyzed methodologies is fully mature to serve as a "domain expert centered ontology engineering methodology in the context of electronic service provisioning in public administration". In fact, there is one single candidate (i.e., Methodology 4) that technically addresses all methodological requirements (see aspects 3 to 5, Section II) in an acceptable way. Unfortunately, this methodology was developed for a different domain (i.e., organizational learning). Hence, its domain-dependence and application-dependence make a direct exploitation for the E-Government domain very difficult. However, aspects and general guidelines of this as well as of some other methodologies can definitely contribute to a future methodology in this field.

Methodology 7 suggests different activities and processes depending on a specific situation and does not follow a rigid workflow for every situation. In fact, a public administration subsumes a variety of different domains (e.g., welfare, health, buildings and constructions, education). This methodology may take into account that not each public domain, each public service or each modeling activity may fit into one single process model, as a potential advantage.

Methodology 6 proposes that not only reusability but also usability of ontologies in specific ontology-driven applications is important. This contributes to the situation that domain experts in general often lack in sufficient abstraction abilities. Hence, in order to be able to validate consistency and reasonableness of resulting ontologies domain experts should be able to check the consequences of modeling decisions in respective applications immediately.

As ontology engineering in public administration generally involves numerous experts from different agencies, expert knowledge is usually scattered over the involved participants. Enabling a structured conversation among all participants is crucial for an ontology engineering methodology in this context. On the one hand, this should lead to a collaborative construction of the domains in question and on the other hand should also assure the differentiation of responsibilities and roles. Aspects of Methodology 1, Methodology 2, Methodology 3, and Methodology 4 may valuably contribute to this requirement.

Many methodologies include a revision or validation activity in the proposed process. Whereas in an engineering-centered development approach validation activities usually deal with formal issues of the ontology a domain expert-centered approach should additionally emphasize factual aspects. For example, Methodology 2 asks questions like "does it represent the domain, or does it contain factual errors?" In the public administration sector, also legal aspects have to be considered in this respect. This circumstance is hardly ever addressed by the studied methodologies.

Methodology 4, Methodology 6 and Methodology 3 already include some general guidance for inexperienced ontology engineers. In this respect, the authors of [17] mention the problem that domain experts are often rarely available and scarcely motivated towards modeling. Consequently, we infer that firstly, much more effort is necessary to develop methods and tools that further reduce the complexity of ontology engineering, and secondly, future research has to pay special attention to improving the commitment of domain experts.

To conclude, one important implication of introducing semantic technologies to the E-Government sector is to increase transparency of the decision making process as well as to increase the citizen orientation of public agencies. Proposing an ontology engineering methodology in the context of public administration would definitely be a next step for an increased matureness of such semantic initiatives. The requirements-driven approach to ontology engineering as proposed by Methodology 5 that starts with the viewpoint and desires of the citizens who want to consume public services already addresses this issue and may therefore also contribute to a future methodology in this field.

REFERENCES

[1] V. Peristeras, K. Tarabanis, and S. K. Goudos, “Model-driven eGovernment interoperability: A review of the state of the art,” Computer Standards & Interfaces, vol. 31, no. 4, Jun. 2009, pp. 613–628.

[2] R. Klischewski and S. Ukena, “Designing semantic e-Government services driven by user requirements,” in Proceedings of ongoing research project contributions and workshops 6th International EGOV Conference Trauner Verlag Linz Austria, 2007, pp. 1–8.

[3] J. V. F. Dombeu and M. Huisman, “Combining Ontology Development Methodologies and Semantic Web Platforms for E-government Domain Ontology Development,” International journal of Web & Semantic Technology, vol. 2, no. 2, 2011, p. 14.

[4] R. Klischewski and L. Lessa, “Sustainability of e-Government Success: an Integrated Research Agenda,” in International Federation for information Processing (IFIP) e-Government Conference 2012, 2012, pp. 153–162.

[5] D. Liebwald, “Knowledge Representation and Modelling Legal Norms : The EU Services Directive,” in Proceedings of the Third Workshop on Legal Ontologies and Artificial Intelligence Techniques, Jun. 2009.

[6] M. Fernandez-Lopez, “Overview Of Methodologies For Building Ontologies,” in Proceedings of the IJCAI99 Workshop on Ontologies and ProblemSolving Methods Lessons Learned and Future Trends CEUR Publications, 1999, vol. 1999, pp. 1–13.

[7] H. Lutz and D. Stelzer, Informationsmanagement: Grundlagen, Aufgaben, Methoden (German), 2nd ed. Munich: Oldenbourg Wissenschaftsverlag, 2011.

[8] M. Jarrar and R. Meersman, “Ontology Engineering -The DOGMA Approach,” in Advances in Web Semantics I, vol. Volume LNC, Springer, 2008, pp. 7–34.

[9] T. D’Entremont and M.-A. Storey, “Using a degree of interest model to facilitate ontology navigation,” 2009 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC), Sep. 2009, pp. 127–131.

[10] N. Casellas, “Methodologies, Tools and Languages for Ontology Design,” in Legal Ontology Engineering - Methodologies, Modelling Trends, and the Ontology of Professional Judicial Knowledge, Springer, 2011, pp. 57–109.

[11] M. Uschold and M. King, “Towards a Methodology for Building Ontologies,” in Proceedings of IJCAI95’s Workshop on Basic Ontological Issues in Knowledge Sharing, 1995, vol. 82, no. 1, pp. 74–82.

[12] C. W. Holsapple and K. D. Joshi, “A collaborative approach to ontology design,” Communications of the ACM, vol. 45, no. 2, 2002, pp. 42–47.

[13] D. Vrandecic, S. Pinto, C. Tempich, and Y. Sure, “The DILIGENT knowledge processes,” Journal of Knowledge Management, vol. 9, no. 5, 2005, pp. 85–96.

[14] W. C. Mann and S. A. Thompson, “Rhetorical Structure Theory: A Theory of Text Organization,” Ablex Publishing Corporation, 1987.

[15] C. Tempich, D. Vrandečić, and R. Benjamins, “OPJK modeling methodology,” 2005.

[16] K. Kotis and G. A. Vouros, “Human-centered ontology engineering: The HCOME methodology,” Knowledge and Information Systems, vol. 10, no. 1, 2005, pp. 109–131.

[17] C. Ghidini, M. Rospocher, B. Kump, V. Pammer, A. Faatz, and A. Zinnen, “Integrated Modelling Methodology - Version 2 . 0,” 2009.

[18] P. Spyns, Y. Tang, and R. Meersman, “An ontology engineering methodology for DOGMA,” Applied Ontology vol. 2, no. 3, 2008, pp. 13–39.

[19] A. De Moor, P. De Leenheer, and R. Meersman, “DOGMA-MESS : A Meaning Evolution Support System for Interorganizational Ontology Engineering,” Engineering, vol. 4068, 2006, pp. 189–202.

[20] M. C. Suarez-Figueroa, A. Gomez-Perez, and M. Fernandez-Lopez, “The NeOn Methodology for Ontology Engineering,” in Ontology Engineering in a Networked World, M. C. Suárez-Figueroa, A. Gómez-Pérez, E. Motta, and A. Gangemi, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 9–34.

[21] M. Fernández-López and A. Gómez-Pérez, “Overview and analysis of methodologies for building ontologies,” The Knowledge Engineering Review, vol. 17, no. 2, 2002, pp. 129–156.

[22] H. Beck and H. S. Pinto, “Overview of Approach , Methodologies , Standards , and Tools for Ontologies,”. unpublished.

[23] O. Corcho, M. Fernández-López, and A. Gómez-Pérez, “Methodologies, tools and languages for building ontologies. Where is their meeting point?,” Data & Knowledge Engineering, vol. 46, no. 1, 2003, pp. 41–64.

[24] K. Sandkuhl, “Towards a Methodology for Ontology Development in Small and Medium-Sized Enterprises,” in IADIS Conference on Applied Computing, 2005, pp. 369–376.

[25] J. A. Kim and S. Y. Choi, “Evaluation of Ontology Development Methodology with CMM-i,” in 5th ACIS International Conference on Software Engineering Research, Management & Applications (SERA 2007), 2007, pp. 823–827.

[26] Advanced Process- Oriented Self- Directed Learning Environment project site, http://www.aposdle.tugraz.at [retrieved: June, 2013].

[27] Web Service Modeling Ontology project site, http://www.wsmo.org [retrieved: June, 2013].

[28] OWL-S: Semantic Markup for Web Services W3C Member Submission, http://www.w3.org/Submission/OWL-S/ [retrieved: June, 2013].

A Profile-Policy Integration for A Personalized Lifestyle

Ozgu Can, Okan Bursa, Murat Osman Unalir
Department of Computer
Engineering, Ege University
Izmir, Turkey 35100
Email: {ozgu.can, okan.bursa, murat.osman.unalir}@ege.edu.tr

***Abstract*—As the amount of information increases and access to this information gets easier, the need for personalized systems is inevitable. A personalized system gives users the efficiency to meet their specific preferences by increasing usability and decreasing the unwanted content. The core component of creating a qualified personalized system is defining semantically rich user profiles. We propose to integrate user profiles with policy management concept to provide a rule-based personalization. The main contributions of this work are: developing a profiling methodology to define semantically rich user profiles and generating a profile-based policy management in order to satisfy the demands of a personalized system. We demonstrated our empirical approach for the health care domain to build a personalized lifestyle model. This user-adaptive system will also give the user a significant time reduction when searching specific items for a special user profile type by restricting the options based on the same profile when compared to a non-adaptive system.**

Keywords-Profile Management; Personalization; Policy Enforcement; Healthcare Systems

I. INTRODUCTION

The promising advantages of online networks create impressive occasions for users. The success of these occasions should be improved by adapting web services to each user's characteristics and behaviors. Personalized systems are the key component to achieve this improvement. As the amount of information increases, making decisions about information becomes difficult. Thus, a personalized system gives users the efficiency to meet their preferences.

Personalization is the process of giving decisions among the given choices according to the user's behavior, needs, preferences, interests and demographics. Hence, users can reach personalized contents such as customized web pages, advertisements, music albums and restaurants that match their profiles. Profiles can be used to describe a wide variety of knowledge about people [1] and this knowledge can have many levels according to the depth of the user information.

User-profile based personalization is the process of making decisions based upon stored user profile information. We use Friend-Of-A-Friend (FOAF) [2] ontologies to store static user profiles. Gruber [3] defines ontology as an explicit specification of a conceptualization. Ontologies are used to represent information in a machine-readable fashion and to model specific domain information by defining objects, concepts and relationships. A FOAF profile is a machine-readable page, which is describing a person, her activities and her relations to other people and objects.

FOAF is also a consistent and common vocabulary to describe the demographical information. Most of the ontological information on the web generally use personal FOAF files. There are more than 13,120,000 people using FOAF to describe their personal profiles [4]. We propose a user profiling methodology with multi-metamodeling by using FOAF profiles. This methodology gives us the opportunity to create a complex and personal profile, which is a demand for an effective personalized system.

Policies are used to control access to resources. Policy management in Semantic Web is used to define declarative rules for accessing a resource and to allow users to interpret and comply with these rules. Integrating profiles into policies is improving personalization under the influence of policy management.

In order to qualify personalization successfully, we are integrating user-profile based personalization with policy management. In this paper, we propose a personalized system to help users choose an item from a large set of items of the same type by filtering this large set using policies according to their defined user profiles. We demonstrate our empirical approach for food domain to meet the requirements of health care domain.

Today, many people care about their health. Therefore, they pay extra attention to what they eat, what ingredients do their meals include and how many calories do their meals have. In order to satisfy this demand, we focus on the food domain ontology to perform the profile-based policy concepts in a personalized system. A personalized system that we propose in our case study can serve several objectives:

– nutrition information to preserve health,
– caution for people who have specific conditions, such as allergies or diabetes,
– ingredient information of meal courses,
– calorie control mechanism to restrict a person's daily calorie intake.

Health care is an information-rich domain and needs to be handled in care. User profiling in such a delicate topic requires more abstraction and variation than a regular FOAF file. This variation in profiles gives more efficiency in building policies to achieve rule-based personalization.

Copyright (c) IARIA, 2013. ISBN: 978-1-61208-293-6

Our profile methodology has the capability to describe the health domain profiles. We can describe several profiles using these profiles, such as diabetic profile, diet profile, individualized ingredient profile and personal profiles where personalization needs a complex domain knowledge, such as health. Profiles are the key ingredients to tailor a profile-based policy management to restrict personalized rules.

This paper is organized as follows: Section 2 describes the user personalization and explains our profiling methodology. Section 3 expresses policy representation and policy ontology concepts. Also, it clarifies the connection between profile and policy ontologies. In Section 4, a case study is presented. Additionally, the food domain ontology concepts, profile and policy examples are demonstrated in this section. Related Work is given in Section 5. Finally, Section 6 concludes and gives the future direction of our work.

II. User Personalization

The profile of a person is an abstract description of the person's demographic, social and behavioral condition. A profile is a representation of a person's daily or permanent properties. In their life time, people change their minds and their situation also changes, due to different conditions. Thus, a static profile, which consists of these properties, has to adapt itself. Demographic properties like a person's occupation, age or school can not change rapidly due to their static nature. However, on a daily basis, a person can have different moods, different roles and different social choices. For example, a person, who is a doctor, can have many daily roles, such as being a mother, a parent or a child. She may want to use different preferences and different identifications for each of these roles. But, as she is a person, she also has demographical properties. So, for all these situations, we have developed a profiling methodology to represent a person's daily profiles by using demographic, social and behavioral properties. This methodology consists of a domain ontology, profile ontology and a metaprofile ontology to represent profile attributes and general descriptions.

A profile is the representation of demographic properties of a person. Let us state a profile as p, a user as u and a FOAF profile of a person as f. As we can call a FOAF profile as a base, we can define many profiles inside the base by using the $hasProfile$ property, $F(u) = hasProfile(P)$. These profiles are meaningful when P has properties, which are included by F. So, we can add data type, D, and object type properties, O, to this definition. $f \in F$, $p \in P$, $d_n \in D$, $o_n \in O$;

$$f(u) \quad = \left\{ \begin{array}{c} hasProfile(p_1), .., hasProfile(p_n), \\ d_1, d_2, d_3, ..., d_n, \\ o_1, o_2, ..., o_n \end{array} \right\} \qquad (1)$$

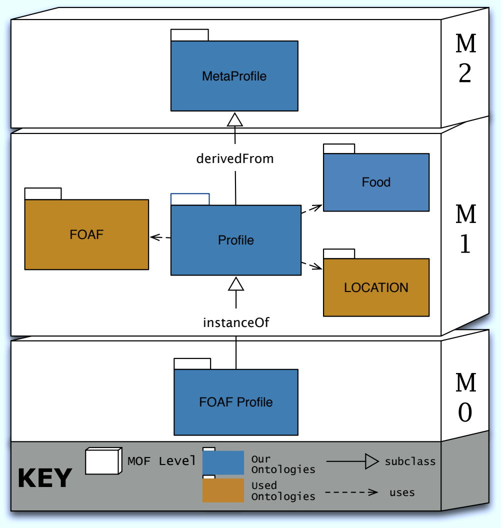


Fig. 1: Profile Methodology

In our ontology metamodel, as seen in Figure 1, we propose a new metamodel based on OMG's Meta-Object-Facility(MOF)[5]. In our metamodel, FOAF documents are our individuals. Inside FOAF documents, we use definitions and structures that are defined inside M1 level: Profile, FOAF definition, Location and Food ontologies. Profile ontology uses MetaProfile's ontological definitions. Metaprofile is independent from the domain. So, it consists of the basic properties to represent a social profile and its ancestors, a behavioral profile and its properties, and demographic properties of a person. These representations need to be designed inside a person's profile. Thus, we aim to define indicators. A profile indicator, pi , is the key property that defines a profile. $p \in P$, $pi \in PI$, $d_n \in D$, $o_n \in O$;

$$p = \{pi, d_1, ...d_n, o_1, ..., o_n\} \qquad (2)$$

As an example, a diabetic profile is meaningful when a person has diabetes or regulations including diabetes inside the profile. Another example is a diet profile, which needs to include the definition of diet or maximum amount of calorie that a person should consume during the day. We develop three types of indicators. The first one is a point-based profile indicator, PB, which helps to define a basic profile property that has a singular value or individual.

$$PB = (d_1) \vee (o_1) \qquad (3)$$

The second one is a range-based profile indicator, RB, which helps to define a range literal value with minimum and maximum values.

$$RB = (d_{1_{min}}, d_{1_{max}}) \qquad (4)$$

The third one is a set-based profile indicator, SB, which includes a set of individuals. Profiles with set-based profile indicator could have individuals only described in this set-based profile indicator.

$$SB = \{o_1, o_2, o_3, o_4\} \tag{5}$$

Set-based profile indicator can be homogeneous or heterogeneous.

$$\begin{gathered} SB_{ho} = \{o_1, o_2, o_3\} \ : \\ \forall o \in SB_{ho} \mid o \in O_1 \\ SB_{he} = \{o_1, o_2, o_3\} \ : \\ \exists o_1 \in O_1 \land o_2, o_3 \in O_2 \land O_1 \neq O_2 \end{gathered} \tag{6}$$

This methodology gives us the ability to construct general profiles that can be explicitly defined in people's attributes. Thus, we can categorize people based on their profiles and represent these group profiles. Group profiles, G, are a generalization of a community of people based on their profile attributes. A group profile, $g \in G$, needs at least a profile identifier to describe itself. Also, later, this property will be the key to add user profiles into this group profile.

$$\begin{gathered} g = \{a_1, a_2, a_3, ..., a_n\} : \\ \exists a \in G \mid a \in PI \land PI \subset SB \cup RB \cup PB \end{gathered} \tag{7}$$

Group profiles can have these three profile identifier types: set-based, range-based and social. These identifiers are based on the key attribute(s) that they are constructed by. Moreover, a group profile may need two or more profile identifiers to describe itself. In this case, we define a set of profile identifiers.

$$g = a \land b : a, b \in PI \land PI \subset SB \cup RB \cup PB \tag{8}$$

Group profiles enable us to describe a policy for communities and persons based on their group or personal profiles.

III. Policy Representation

A policy is a declarative rule set that is based on constraints to control the behavior of entities. Policy rules define a declarative information on what an entity can do or cannot do. A policy consists of an entity, a constraint and a deontic object. An entity is the subject of the policy and a constraint defines the condition on a policy rule. A deontic object defines the concepts of permission, prohibition, obligation and dispensation. Permission is what an entity can do, prohibition is what an entity can not do, obligation is what an entity should do, and finally dispensation is what an entity need no longer do.

There are some general requirements that any policy representation should satisfy regardless of its field of applicability: expressiveness, simplicity, enforceability, scalability and analyzable [6]. In this work, by taking these requirements and ease of use criteria into consideration, we used Rei [7] policy language to represent policies. Rei policy language is composed of seven ontologies: ReiPolicy, ReiMetaPolicy, ReiEntity, ReiDeontic, ReiConstraint, ReiAnalysis, and ReiAction.

A. Policy Ontology

In a policy ontology, a policy is shown with a triple as *(S, O, A)*, in which *S* is subject, *O* is object and *A* is action. The subject indicates the entity that wants to access a resource, the object indicates the resource, which is going to be accessed, and the action indicates an operation, which the entity wants to achieve on a resource. The set of subjects, objects and actions is represented as $S = \{s_1, s_2, ..., s_i\}$, $O = \{o_1, o_2, ..., o_j\}$ and $A = \{a_1, a_2, ..., a_k\}$, respectively. The set of deontic objects, which are used to form policy rules is represented as
$DO = \{Permission, Prohibition, Obligation, Dispensation\}$

B. Connecting Profile Ontology with Policy Ontology

In order to integrate profiling methodology into policy management, we substitute the set of subjects with the set of profiles, $P = \{p_1, p_2, ..., p_n\}$. When creating policies using Rei policy language, the subject of the policy is related with `entity:Variable` class. `entity:Variable` is a class of `ReiEntity` ontology. While creating a profile-based policy ontology, instances of an action's actors are now profile instances of the profile ontology. Thus, profile instances are used instead of the instances of `entity:Variable` class as the subject of the policy. As a result, policy subjects are comprised of semantically rich profile ontology.

The OWL representation of a `Vegetarian` profile defined in `entity:Variable` class is as follows:

```
<owl:Thing rdf:about="#Vegetarian">
  <rdf:type rdf:resource="&ReiEntity;Variable"/>
</owl:Thing>
```

The OWL representation of a `Vegetarian` profile defined in profile ontology is as follows:

```
<owl:Thing rdf:about="#Vegetarian">
  <rdf:type rdf:resource="&Profile.owl;Vegetarian"/>
</owl:Thing>
```

In a profile-based policy management [8], policy rules are assigned to profiles. Subjects are assigned with profiles and access rights to objects are given to profiles. Profile-based policy determines the ideal behaviors of the user using the user profile information. Figure 2 shows the policy components of the model.

A subject is represented by a profile and a profile is comprised of the profile ontology, which uses metaprofile ontology. An action and an object are based on domain ontology. Profile, action and object triple is used to form policy objects. Policy objects are used to create policy ontology, which is also based on the metapolicy ontology.

IV. Case Study

In this section, we present a case study for personalization by using policy management based on profiling methodology. The following conditions are some examples for personalization:

1. A diabetic person who is looking for a restaurant, she can be permitted or prohibited for her meal course preferences according to her health condition.

2. A professor who has an obligation for beverages, like not

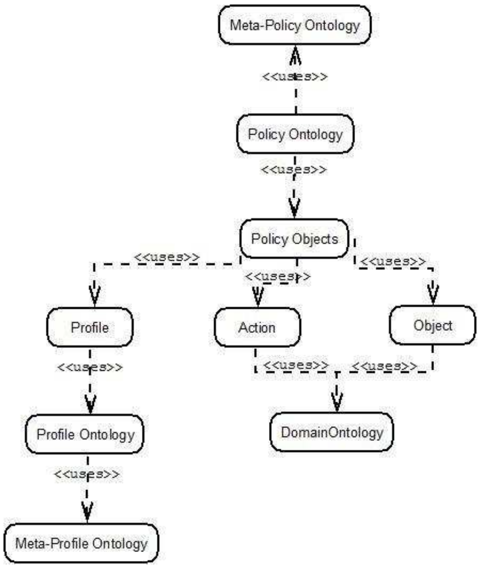


Fig. 2: Policy components of the model

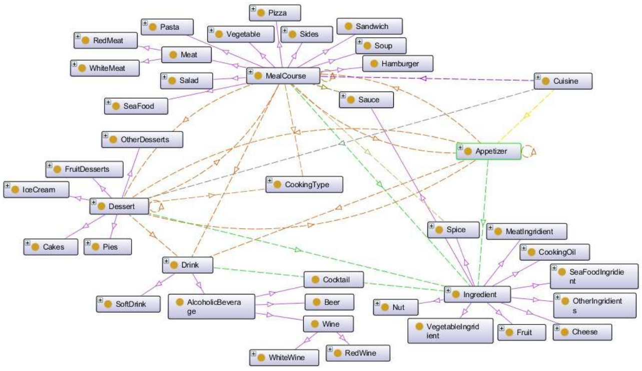
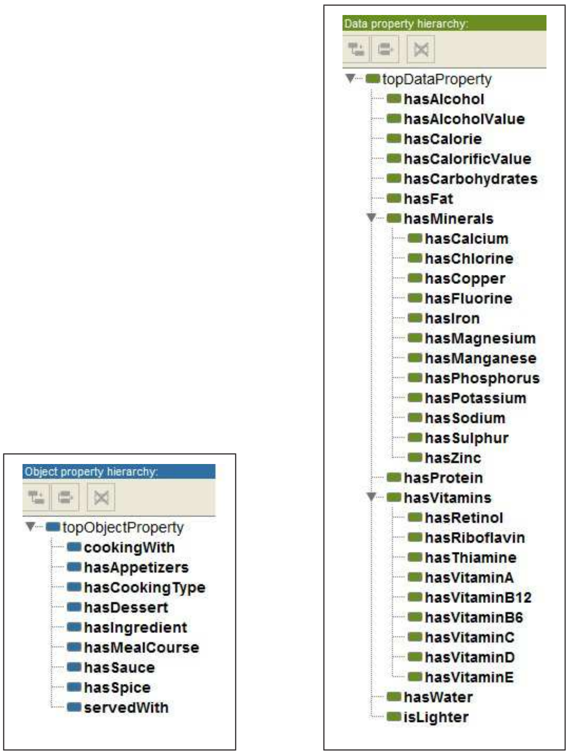


Fig. 3: Class hierarchy of the food domain ontology


Object property hierarchy:

- topObjectProperty
  - cookingWith
  - hasAppetizers
  - hasCookingType
  - hasDessert
  - hasIngredient
  - hasMealCourse
  - hasSauce
  - hasSpice
  - servedWith

Data property hierarchy:

- topDataProperty
  - hasAlcohol
  - hasAlcoholValue
  - hasCalorie
  - hasCalorificValue
  - hasCarbohydrates
  - hasFat
  - hasMinerals
    - hasCalcium
    - hasChlorine
    - hasCopper
    - hasFluorine
    - hasIron
    - hasMagnesium
    - hasManganese
    - hasPhosphorus
    - hasPotassium
    - hasSodium
    - hasSulphur
    - hasZinc
  - hasProtein
  - hasVitamins
    - hasRetinol
    - hasRiboflavin
    - hasThiamine
    - hasVitaminA
    - hasVitaminB12
    - hasVitaminB6
    - hasVitaminC
    - hasVitaminD
    - hasVitaminE
  - hasWater
  - isLighter

Fig. 4: Object and Data type properties of the food ontology

drinking alcoholic beverages, when she is in a foreign country at a conference.
3. A person who is on a low calorie diet for a particular day may demand to be prohibited from choosing meal courses that have a high calorie content.
4. A peanut allergic person may demand to be prohibited from meal courses that include peanut.
5. A vegetarian person would like to know the meal courses that have vegetarian ingredients.

According to these examples given above, we build a food domain ontology. We use different sources to gather location and profile information. Unfortunately, we could not find any food ontology that combines all these ontologies together. So, we developed our own food ontology to overcome this problem. The next section explains in detail our domain ontology.

A. *Domain Knowledge*

As our case study needs a domain ontology to express the examples that are mentioned above, we build a food domain ontology. Figure 3 shows the class hierarchy of the food domain ontology.

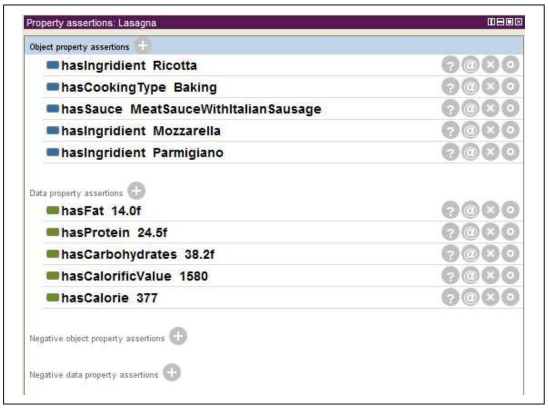
Each item in a restaurant menu can be an individual of the food domain ontology. Each individual of appetizer, meal course, drink and dessert has an ingredient information, which has tied to `Ingredient` class with `hasIngredient` object property. Additionally, each individual has nutrition summary information defined with data properties. The nutrition summary values are taken from `fatsecret` [9] web site. Figure 4 shows object and data type properties of the food ontology, respectively.

Figure 5 shows an example of `Lasagna` individual of `MealCourse` class.

Besides the food domain ontology, a location ontology needs to be developed in order to provide a semantic connection between a place and this place's food menu. For this purpose, we selected the schema.org's [10] ontology and adapted this ontology to our case study. Schema.org's ontology has a property to describe a menu item, but it is a general definition, which ranges to a `string` or the `Thing` class. Furthermore, a connection between the menu and the food domain ontology is a necessity. The relationship between the location ontology and the food domain ontology can be seen in Figure 6. This connection gives us the opportunity to build a profile-based policy description to handle the problems in our case study examples.

B. *Profile Examples*

The following examples define the profiles mentioned in the case study. These profiles are based on the profile methodology



Property assertions: Lasagna

Object property assertions

hasIngridient Ricotta

hasCookingType Baking

hasSauce MeatSauceWithItalianSausage

hasIngridient Mozzarella

hasIngridient Parmigiano

Data property assertions

hasFat 14.0f

hasProtein 24.5f

hasCarbohydrates 38.2f

hasCalorificValue 1580

hasCalorie 377

Negative object property assertions

Negative data property assertions

Fig. 5: `Lasagna` individual of `MealCourse` class

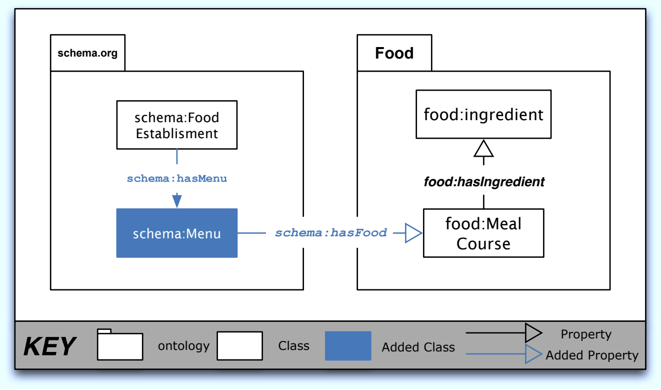


Fig. 6: Location and Food Ontology Relations

that we described in Section 2.

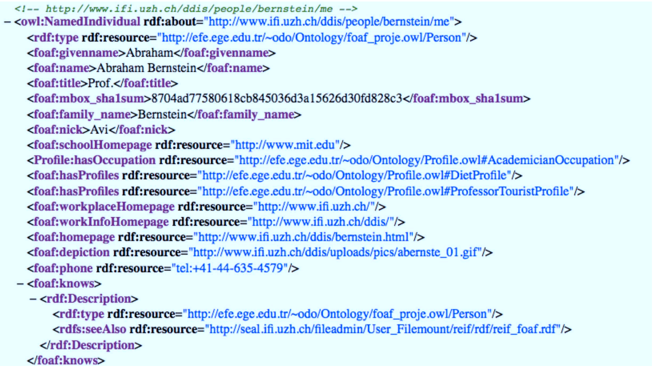
First profile has diabetes, so she possesses a `Diabetic` profile. Diabetic profile has a rule, which states that *a diabetic person can not drink an alcoholic beverage*.

$$\begin{gathered} diabetic = canDrink(n), \\ \forall n \in NonAlcoholicBeverage \cup \\ SetBasedProfileIndicator \cup Demographic \\ \rightarrow n \notin AlcoholicBeverage \end{gathered} \tag{9}$$

The second profile is that of a special professor profile who does not want to drink any alcoholic beverage when she is attending a conference in a foreign country.

$$\begin{gathered} professorAbroad = canDrink(n) \wedge visits(c), \\ \forall n : n \in NonAlcoholicBeverage \cup \\ SetBasedProfileIndicator \cup Demographic \wedge \\ \in Country \rightarrow n \notin AlcoholicBeverage \wedge \\ c \neq homeCountry(professorAbroad) \cup \\ SetBasedProfileIndicator \cup GeoDemographic \end{gathered} \tag{10}$$

The third profile is a behavioral profile, which describes the diet of a person. This profile has a range-based profile indicator to describe a `lowCalorie` profile, which has a minimum and maximum range in calorie calculation.



```
<!-- http://www.ifi.uzh.ch/ddis/people/bernstein/me -->
- <owl:NamedIndividual rdf:about="http://www.ifi.uzh.ch/ddis/people/bernstein/me">
    <rdf:type rdf:resource="http://efe.ege.edu.tr/~odo/Ontology/foaf_proje.owl/Person"/>
    <foaf:givenname>Abraham</foaf:givenname>
    <foaf:name>Abraham Bernstein</foaf:name>
    <foaf:title>Prof.</foaf:title>
    <foaf:mbox_sha1sum>8704ad77580618cb845036d3a15626d30fd828c3</foaf:mbox_sha1sum>
    <foaf:family_name>Bernstein</foaf:family_name>
    <foaf:nick>Avi</foaf:nick>
    <foaf:schoolHomepage rdf:resource="http://www.mit.edu"/>
    <Profile:hasOccupation rdf:resource="http://efe.ege.edu.tr/~odo/Ontology/Profile.owl#AcademicianOccupation"/>
    <foaf:hasProfiles rdf:resource="http://efe.ege.edu.tr/~odo/Ontology/Profile.owl#DietProfile"/>
    <foaf:hasProfiles rdf:resource="http://efe.ege.edu.tr/~odo/Ontology/Profile.owl#ProfessorTouristProfile"/>
    <foaf:workplaceHomepage rdf:resource="http://www.ifi.uzh.ch/"/>
    <foaf:workInfoHomepage rdf:resource="http://www.ifi.uzh.ch/ddis/"/>
    <foaf:homepage rdf:resource="http://www.ifi.uzh.ch/ddis/bernstein.html"/>
    <foaf:depiction rdf:resource="http://www.ifi.uzh.ch/ddis/uploads/pics/abernste_01.gif"/>
    <foaf:phone rdf:resource="tel:+41-44-635-4579"/>
  - <foaf:knows>
    - <rdf:Description>
        <rdf:type rdf:resource="http://efe.ege.edu.tr/~odo/Ontology/foaf_proje.owl/Person"/>
        <rdfs:seeAlso rdf:resource="http://seal.ifi.uzh.ch/fileadmin/User_Filemount/reif/rdf/reif_foaf.rdf"/>
      </rdf:Description>
    </foaf:knows>
```

Fig. 7: FOAF profile of Prof. Bernstein

$$\begin{gathered} lowCalorieProfile = hasMood(lowCalorie) \\ lowCalorie = hasMaximum(maximumCalorie) \wedge \\ hasMinimum(minimumCalorie), \\ lowCalorie \in RangeBasedProfileIndicator \cup Mood \end{gathered} \tag{11}$$

The fourth profile is a `peanutAllergic` profile who has an allergic reaction to peanuts. This profile needs to be defined based on `hasIngredient` object property that defines ingredients of a meal course.

$$\begin{gathered} peanutAllergic = hasAllergic(p) \\ p = hasIngredient(\text{"}peanut\text{"}), \\ \forall p : p \in PeanutAllergicFood \cup \\ SetBasedProfileIndicator \cup Demographic \end{gathered} \tag{12}$$

The last profile is a `Vegetarian` profile who only eats vegetarian food.

$$\begin{gathered} vegetarian = canEat(f) \\ f = hasIngredient(i), \forall i : i \in VegetarianFood \wedge \\ \forall f : f \in MealCourse \cup \\ SetBasedProfileIndicator \cup Demographic \end{gathered} \tag{13}$$

Figure 7 shows a professor who has a FOAF profile as mentioned in the second example of the case study. The Professor has many different profiles inside his FOAF profile. So, when he travels abroad for a conference and wants to have a light lunch according to his daily diet, there will be some restrictions on the lunch menu of the restaurant he choses. As seen from Figure 7, he has a `professorConference` profile and a `dietProfile`. These profiles have preference restrictions on alcoholic beverages and the total calorie limit for his lunch menu.

C. *Policy Examples*

This section demonstrates policy examples and their Semantic Web Rule Language (SWRL) [11] rules for the related case study examples. The following example shows a prohibition for a `Diabetic` profile. According to this rule, if a `Diabetic` profile chooses `ScillianScampi` from

Appetizer, she will be prohibited, because cookingWith property has Chardonnay individual, which has true value for its hasAlcohol property.

$$\begin{gathered} Profile(?Diabetic) \wedge Appetizer(?x) \\ \wedge cookingWith(?x, ?y) \wedge hasAlcohol(?y, true) \Longrightarrow \\ Prohibition\_ \\ orderScillianScampi(?Diabetic, ?x) \end{gathered} \tag{14}$$

In the second policy example, professorAbroad profile will be permitted when she chose FruitPunch that has false value for its boolean hasAlcohol data property.

$$\begin{gathered} Profile(?professorAbroad) \wedge Drink(?x) \\ \wedge cookingWith(?x, ?y) \wedge hasAlcohol(?y, true) \\ \wedge hasAlcohol(?x, false) \Longrightarrow \\ Permission\_ \\ orderFruitPunch(?professorAbroad, ?x) \end{gathered} \tag{15}$$

The following policy example gives an obligation to the lowCalorie profile according to the profile's daily calorie range for one meal course defined in the profile that has a range between minimum 400 and maximum 500. Thus, when she chose Herb-GrilledSalmon, if its hasCalorie property is less than the maximum calorie defined for lowCalorie profile, then she will be permitted to order Herb-GrilledSalmon, otherwise she will be prohibited.

$$\begin{gathered} Profile(?lowCalorie) \\ \wedge hasCalorie(?Herb - GrilledSalmon, ?x) \\ \wedge hasMaximumCalorie(?lowCalorie, ?y) \\ \wedge isLessThan(?x, ?y) \Longrightarrow \\ Obligation\_ \\ orderHerb - GrilledSalmon \\ (?lowCalorie, ?Herb - GrilledSalmon) \end{gathered} \tag{16}$$

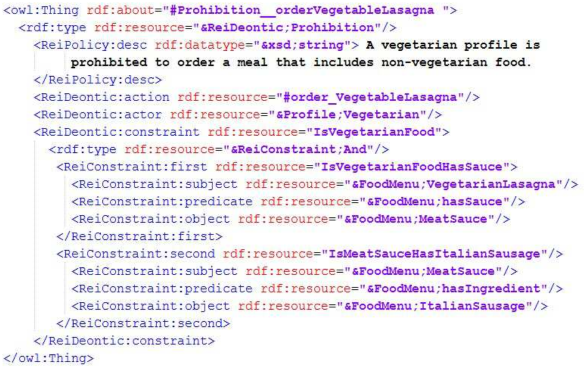
A prohibition will be given to the peanutAllergic profile when she chose PumpkinPie, which has Peanut value for its hasIngredient property.

$$\begin{gathered} Profile(?peanutAllergic) \\ \wedge hasIngredient(?x, ?Peanut) \\ \Longrightarrow Prohibition\_ \\ orderPumpkinPie(?peanutAllergic, ?x) \end{gathered} \tag{17}$$

The last policy example prohibits the Vegetarian profile when she chose VegetableLasagna, because the course's hasSauce property's value is MeatSauce, which also has ItalianSausage value for its hasIngredient property. Figure 8 shows the OWL representation of this policy.

$$\begin{gathered} Profile(?Vegetarian) \\ \wedge hasSauce(?x, ?y) \\ \wedge hasIngredient(?y, ?ItalianSausage) \\ \Longrightarrow Prohibition\_ \\ orderVegetableLasagna(?Vegetarian, ?x) \end{gathered} \tag{18}$$

All these profile definitions and their integration with policies are described manually by the domain experts.



```
<owl:Thing rdf:about="#Prohibition__orderVegetableLasagna ">
  <rdf:type rdf:resource="&ReiDeontic;Prohibition"/>
    <ReiPolicy:desc rdf:datatype="&xsd;string"> A vegetarian profile is
        prohibited to order a meal that includes non-vegetarian food.
    </ReiPolicy:desc>
    <ReiDeontic:action rdf:resource="#order_VegetableLasagna"/>
    <ReiDeontic:actor rdf:resource="&Profile;Vegetarian"/>
    <ReiDeontic:constraint rdf:resource="IsVegetarianFood">
      <rdf:type rdf:resource="&ReiConstraint;And"/>
       <ReiConstraint:first rdf:resource="IsVegetarianFoodHasSauce">
         <ReiConstraint:subject rdf:resource="&FoodMenu;VegetarianLasagna"/>
         <ReiConstraint:predicate rdf:resource="&FoodMenu;hasSauce"/>
         <ReiConstraint:object rdf:resource="&FoodMenu;MeatSauce"/>
       </ReiConstraint:first>
       <ReiConstraint:second rdf:resource="IsMeatSauceHasItalianSausage"/>
         <ReiConstraint:subject rdf:resource="&FoodMenu;MeatSauce"/>
         <ReiConstraint:predicate rdf:resource="&FoodMenu;hasIngredient"/>
         <ReiConstraint:object rdf:resource="&FoodMenu;ItalianSausage"/>
       </ReiConstraint:second>
    </ReiDeontic:constraint>
</owl:Thing>
```

Fig. 8: OWL representation for Vegetarian profile policy

D. *Practical Application*

In our scenario, we used Prof. Bernstein's FOAF profile [12] as our FOAF Person. Firstly, we changed the FOAF URI in order to access the metalevel profile and the hometown property. As an example in our scenario, the professor uses his FOAF profile to order meals through the system. When he attends a conference in a foreign country, he chooses his AcademicianTourist profile, which has an restriction on alcoholic beverages. Besides, he is also on a diet. Thus, his diet profile must be active. His FOAF profile can be seen in Figure 7.

As he is an academician, he has an AcademicianProfile. When he attends a conference, he has ProfessorTouristProfile and also DietProfile. During the conference, he wants to dine in a good restaurant with his colleagues and his colleagues offer to go to a place named with Winter Garden. But first, he wants to check the menu of Winter Garden and uses his mobile application. After he loads his FOAF profile and policy definitions, his mobile application checks the restrictions connected to his profiles. The process of using restrictions with profiles needs an ontology parser and rule engine. This overall architecture can be seen in Figure 9.

The mobile application can query the SPARQL [13] endpoint to get the restricted or granted menu items from the Ontology DB. As we have not developed a mobile application yet, our mock-up for mobile application interface can be seen in Figure 10.

In this interface, granted menu items are green, and restricted menu items are red and not selectable. The mobile application queries the Ontology DB by using the SPARQL. An example query is given in Figure 11.

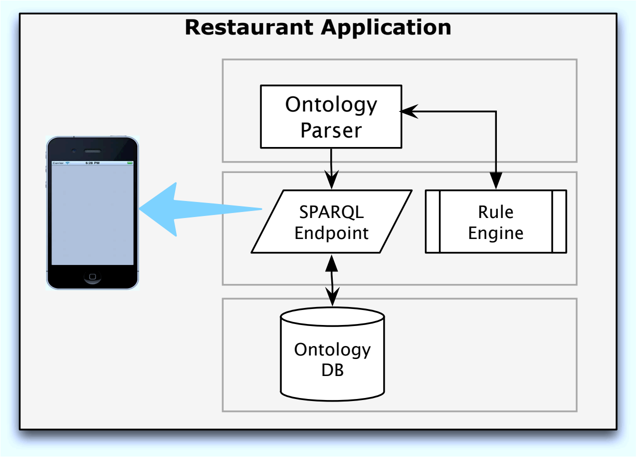


Fig. 9: General view of the architecture

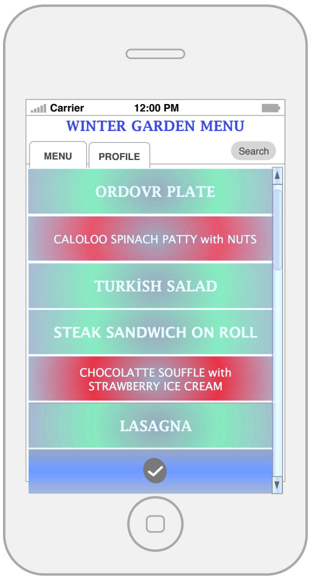


Fig. 10: Mobile application example

V. Related Work

Different domains need different profiling methodologies. The spectrum of profile description in the literature is wide. A profile is a storage that keeps the usable properties of a user and profiling is storing this user information. In [14], user profiles are used as a static storage document for basic information, calendar for daily meetings and so on. As the system becomes more complex, developing a profiling methodology also becomes a complex task. In order to provide services such as recommendation [15] and location based personalization [16], a profile can include different types of properties like online social network information, last visited web page and last clicked advertisement information.

User profiles can be used as a static document but it is more convenient as a dynamic and social projection, which saves a person's daily activities, social roles and preferences.

```
SELECT DISTINCT ?Person ?Profile ?prohibition ?ingredient
WHERE{?person rdf:type foaf:Person.
?person foaf:hasProfiles ?profile.
?restaurant rdf:type location:Restaurant.
?restaurant  menu:hasMenu ?menu.
?menu menu:hasFood ?MealCourse.
?MealCourse ?ObjectProperty ?ingredient.
?prohibition rdf:type reideontic:prohibition.
?prohibition reideontic:actor ?profile.
?prohibition reideontic:reiconstraint ?foodconstraint.
?foodconstraint rdf:type reiconstraint:And.
?foodconstraint ?numberOfConstraint ?firstconstraint.
?foodconstrain reiconstraint:predicate ?ObjectProperty.
?firstconstraint reiconstraint:object ?ingredient.
}
```

Fig. 11: SPARQL Example

In [17], a profiling methodology has been developed to store user preferences. The study presents a User Profile Ontology based on user characterization. This ontology provides an extensible user profile model that focuses on the modeling of dynamic and static user aspects. On the contrary to [17], preference handling needs a complex methodology to extract the possible preferences from domain knowledge and cover these preferences inside appropriate preference types as proposed in [18].

User profiling is also an asset for Quality of Service. In [19], user profiling is a solution for a group of workers who need to be authorized based on different authorization grants. Authorization based on group and individual user profile is a good solution. Besides, it is a strict solution and very hard to change or adapt to different domains. These profile definitions are convenient to be used in small data environments.

However, when data gets bigger, profiling becomes a tough problem. Likewise in [20], profiling is designed inside social networks and a general profile is constructed. As social networks emerge in time exponentially, profiling data emerges elsewhere, so that, describing policies with such a big data becomes a problem. In our work, we are proposing an abstraction to profiling methodology by using metamodel levels [5]. Thus, handling such a huge data becomes less problematic.

User profiles can be integrated into policy management mechanisms. There are various developed policy languages. KAoS [21], Rei [7] and Ponder [22] are the most common policy languages. KAoS is a DAML/OWL policy language. It is a collection of policy and domain management services for web services. KAoS distinguishes between authorizations and obligations. Rei is a policy specification language based on OWL-Lite. It allows users to express and represent the concepts of rights, prohibitions, obligations, and dispensations. Ponder is a declarative, object-oriented policy language for several types of management policies for distributed systems and also provides techniques for policy administration. Ponder has four basic policy types: authorizations, obligations, refrains and delegations. Tonti [6] gives a comparison of these three policy languages.

A framework that offers tools to specify adaptation policies in the form of rules on profile attributes is presented in [23]. However, this is not sufficient to achieve the development of semantically rich applications. In [8], profile-based policy management is studied in order to make use of semantically rich policies in terms of the personalization scope.

An ontology-based solution to personalized clinical management is presented in [24]. The proposed ontology provides a solution for the personalized care challenges in home-based telemonitoring scenarios, and aims to model the tasks specified within a patient profile. Unlike this work, we use FOAF to specify profiles and integrate them with policies for personalization. A health care domain ontology is developed in [25] and access control policies are created based on this domain to manage patient's health records.

Since there are numerous ontology developers, there are also several food ontologies developed. A food-oriented ontology was developed in [26]. Additionally, Cantais [27] proposes a health care domain designed as a part of PIPS (Personalized Information Platform for Health and Life Services) project. However, both of these works do not fulfill the semantics of our scope and the relationship that we need to establish between location and food ontology. Thus, we built a new food menu ontology to achieve the semantically rich data representation.

VI. Conclusion and Future Work

Personalization should lead users to reach person specific information and customization by preventing unwanted content. User profiles are used to construct the personalized content by determining the user's choices and behaviors. User profiles can also be used as subjects for policy management. If profiles are well defined, they can give the exact and direct information about user's behaviors. We proposed an empirical approach to user profiling and built a profile-based policy management model to demonstrate a qualified personalized lifestyle system. We developed a new food ontology to be able to calculate calorie measures. Calorie measures of a menu can be calculated with this information and this makes the policy enforcement possible with the help of individuals profile selections. Profile, as means of a user's daily life role, is used to personalize policies to be able to define different policy rules for different daily situations. We showed policy and profile definitions of our case study examples. We also explained the policy rules and how we enforced these rules by using SPARQL. Our profiling methodology gives a richer user information to policy framework to provide a rule-based personalization. This information is useful to simulate real world problems into policy management.

As part of our future work, we will add new features to the food domain ontology and build a visual tool that allows users to create their profiles and make their meal choices from the restricted menu list. We are currently working on completing our mock-up based mobile application. Therefore, we will be able to gather user experience feedbacks of the methodology. A comparison between an user-adaptive and a non-adaptive system in the measurement of time that is spent for searching a specific item for a specific profile type will also be experimented. Additionally, we will automatize our food ontology's calorie extraction by using FatSecret's Platform API [28].

References

[1] E. Rich, Users are Individuals: Individualizing User Models, International Journal of Man-Machine Studies, Vol.18, pp. 199-214, 1983.

[2] D. Brickley, Friend-Of-A-Friend (FOAF), http://www.foaf-project.org, Last Access: July 3, 2013.

[3] T. R. Gruber, A Translation Approach to Portable Ontology Specification, Knowledge Acquisition 5(2), 1993, pp. 199-220.

[4] J. Goldbeck, M. Rothstein, Linking Social Networks on the Web with FOAF, Proceeding AAAI'08 Proceedings of the 23rd national conference on Artificial intelligence, Vol 2, 2008, pp. 1138-1143.

[5] OMG's Meta Object Facility (MOF) 2012, http://www.omg.org/mof/, Last Access: July 3, 2013.

[6] G. Tonti, J. M. Bradshaw, R. Jeffers1, R. Montanari, N. Suri, and A. Uszok, Semantic Web Languages for Policy Representation and Reasoning: A Comparison of KAoS, Rei, and Ponder, International Semantic Web Conference 2003, Vol. 2870, 2003, pp. 419-437.

[7] L. Kagal, T. Finin, and A. Joshi, A Policy Language for A Pervasive Computing Environment, In IEEE 4th International Workshop on Policies for Distributed Systems and Networks, 2003, pp. 63-74.

[8] O. Can, O. Bursa, and M. O. Unalir, Personalizable Ontology-Based Access Control, Gazi University Journal of Science, 23(4), 2010, pp.465-474.

[9] FatSecret, Calories Nutrition All Things about food and diet, http://www.fatsecret.com/calories-nutrition, Last Access: July 3, 2013.

[10] Schema.org Schema.org with its RDFS, http://schama.rdfs.org, Last Access: July 3, 2013.

[11] W3C, SWRL: A Semantic Web Rule Language Combining OWL and RuleML W3C Member Submission 21 May 2004, http://www.w3.org/Submission/SWRL/, Last Access: July 3, 2013.

[12] FOAF rdf, http://www.ifi.uzh.ch/ddis/people/bernstein/foaf.rdf, Last Access: July 3, 2013.

[13] W3C, SPARQL:SPARQL Query Language for RDF W3C Recommendation 15 January 2008, http://www.w3.org/TR/rdf-sparql-query/, Last Access: July 3, 2013.

[14] H. Chen, F. Perich, D. Chakraborty, T. Finin, and A. Joshi, Intelligent agents meet semantic web in a smart meeting room, In Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2004), Volume 2, Washington DC, USA, 2004, pp. 854-861,

[15] S. E. Middleton, N. R. Shadbolt, and D. C. De Roure, Ontological User Profiling in Recommender Systems, ACM Transactions on Information Systems, Vol. 22, 2004, pp. 54-88.

[16] K. W. -T. Leung, D. L. Lee, and Wang-Chien Lee, Personalized Web search with location preferences, IEEE 26th International Conference on Data Engineering (ICDE), 2010, pp. 701 - 712.

[17] K.L.Skillen, L.Chen, C.D. Nugent, M.P. Donnelly, W. Burns, and I. Solheim, Ontological User Profile Modeling for Context-Aware Application Personalization, UCAmI'12 Proceedings of the 6th International Conference on Ubiquitous Computing and Ambient Intelligence,, 2012, pp. 261-268,

[18] D. Tapucu, O. Can, O. Bursa, and M. O. Unalir, Metamodeling Approach to Preference Management in the Semantic Web, 4th Multidisciplinary Workshop on Advances in Preference Handling (M-PREF 2008) (In conjunction with AAAI 2008), Chicago, Illinois, July 13-14, 2008.

[19] J. C. Royer, R. Willrich, and M. Diaz, User Profile-Based Authorization Policies for Network QoS Services, Seventh IEEE International Symposium on Network Computing and Applications, 2008, pp. 68-75.

[20] Z. Iqbal and J. Noll, Toward User-Centric Privacy-Aware User Profile Ontology for Future Services, Third International Conference on Communication Theory, Reliability, and Quality of Service (CTRQ), Athens, TBD, Greece, 2010, pp. 249 - 254.

[21] A. Uszok, J. M. Bradshaw, and R. Jeffers, KAoS: A Policy and Domain Services Framework for Grid Computing and Semantic Web Services, In Trust Management (iTrust 2004), Volume 2995 of Lecture Notes in Computer Science, 2004, pp. 16-26.

[22] N. Dulay, E. Lupu, M. Sloman, and N. Damianou, A Policy Deployment Model for The Ponder Language, In Proceedings of IEEE/IFIP International Symposium on Integrated Network Management (IM 2001), 2001, pp. 14-18.

[23] A. Agostini, C. Bettini, N. Cesa-Bianchi, D. Maggiorini, and D. Riboni, Integrated Profile and Policy Management for Mobile-Oriented Internet Services, Technical Report Firb-Web-Minds N. TR-WEBMINDS-04, 2003.

[24] N. Lasierra, A. Alesanco, S. Guillen, J. Garcia, A three stage ontology-driven solution to provide personalized care to chronic patients at home, Journal of Biomedical Informatics, 46, 2013, pp. 516529.

[25] J. Calvillo, I. Romana, L.M. Roa, Empowering citizens with access control mechanisms to their personal health resources, International Journal of Medical Informatics, 82, 2013, pp. 58-72

[26] C. Snae, M. Brckner, FOODS: A Food-Oriented Ontology-Driven System, 2nd IEEE International Conference on Digital Ecosystems and Technologies (DEST 2008), 2008, pp. 168-176.

[27] J. Cantais, D. Dominguez, V. Gigante, L. Laera, and V. Tamma, An Example of Food Ontology for Diabetes Control, Proceedings of the International Semantic Web Conference 2005 Workshop on Ontology Patterns for the Semantic Web, 2005.

[28] FatSecret Platform API 2013, http://platform.fatsecret.com/api/, Last Access: July 3, 2013.

Copyright (c) IARIA, 2013. ISBN: 978-1-61208-293-6

Towards Purposeful Reuse of Semantic Datasets Through Goal-Driven Summarization

Panos Alexopoulos and José-Manuel Gómez Pérez
iSOCO, Intelligent Software Components S.A.
Madrid, Spain,
e-mail: {palexopoulos, jmgomez}@isoco.com

***Abstract*—The emergence in the last years of initiatives like the Linked Open Data (LOD) has led to a significant increase of the amount of structured semantic data on the Web. Nevertheless, the wider reuse of such public semantic data is inhibited by the difficulty for users to decide whether a given dataset is actually suitable for their needs. This is because semantic datasets typically cover diverse domains, do not follow a unified way of organizing the knowledge and may differ in a number of dimensions. With that in mind, in this paper, we report our work in progress on a goal-driven dataset summarization approach that may facilitate better understanding and reuse-oriented evaluation of available semantic data.**

***Keywords*-Semantic Data Reuse; Semantic Data Summarization.**

I. Introduction

The emergence in the last years of initiatives like the Linked Open Data (LOD) [1] has led to a significant increase of the amount of structured semantic datasets on the Web. Nevertheless, while this increased availability of such datasets yields various opportunities for organizations and technical professionals to derive added value from them, their wide heterogeneity and underlying complexity makes their practical use and exploitation quite difficult and challenging. For that, solutions that can enable the better understanding and easier consumption of semantic datasets are of crucial importance.

The typical use case scenario we consider in this paper assumes some organization that wants to reuse public semantic datasets to i) enrich with them its own data so as to make the latter more usable and increase its usability and value and ii) utilize the enriched data within knowledge intensive applications for particular purposes (e.g., decision support). Such tasks are typically performed by knowledge engineers and the common problem associated to them is the so called **knowledge acquisition bottleneck**, namely, the high amount of time and effort required to acquire and maintain the needed knowledge [2].

Our position is that the reuse of existing public semantic data can be a promising way to (partially) alleviate the knowledge acquisition problem. One reason for that is that the volume and diversity of public semantic datasets are increasing at high rates [1], resulting into a large amount of both generic and domain-specific knowledge that is available to use for various application scenarios. Another advantage of the reuse approach is that the maintenance and evolution of these datasets is the responsibility of their publishers, thus reducing the required efforts and costs for this task in the organization's side.

As an example of this, consider a sport news organization that wants to create and maintain a knowledge base about the Spanish football league (teams, rosters, results, etc.). The pace at which this knowledge changes is quite fast (e.g., team rosters change at least every year, sometimes even more frequently), meaning that the organization needs to have a dedicated team that constantly monitors these changes and updates the knowledge base. As much of this information is already available in public semantic datasets and, more importantly, it is (almost) always up to date, it would be better for the organization to reuse this data instead of creating it from scratch and having to maintain it.

Nevertheless, an important problem that inhibits the wider reuse of such public semantic data is the difficulty for knowledge engineers to decide whether a given dataset is actually suitable for their needs. This is because semantic datasets typically cover diverse domains, do not follow a unified way of organizing the knowledge and differ in a number of features including size, coverage, granularity and descriptiveness. This makes the task of assessing whether a dataset satisfies particular requirements (e.g., covering adequately a particular domain) and/or comparing different datasets to select which one is more suitable for a given purpose quite difficult.

For instance, in the example mentioned above about data related to the Spanish football league, one may find such data in DBPedia[12] and Freebase[11]. To evaluate these sources, the knowledge engineer needs to examine and assess a variety of factors including i) the domain's coverage, namely, the degree to which the containing data cover the Spanish football league (e.g., one of the sources might not contain adequate data for a given year), or ii) the dataset's consistency, namely, the absence of contradictions in the data (e.g., there might be statements suggesting that a player is currently playing for two clubs).

As a way to tackle this problem, we envision the development of a framework that will enable users to derive **semantic data summaries**, namely useful descriptions, measures and indicators that provide a landscape yet informative view on a dataset that enables the assessment of the latter's potential value. This task of semantic data summarization is rather overlooked in the research community and has only been addressed by a few works, e.g., [3] [4] [5], each of which generates dataset summaries according to different data features and by applying different criteria.

Yet, the problem with these approaches is that they treat the summarization task in an application and user independent way by producing generic summaries whose usefulness is limited to an all-purpose very high level overview of the data. By contrast, in our scenario, we are interested in facilitating the generation of requirements-oriented and task-specific summaries that may be significantly more helpful to the knowledge engineers and data practitioners in their task to locate semantic data to reuse and exploit.

To that end, in this paper, we report our work in progress on a goal-driven data summarization framework that may be used to examine and evaluate the suitability of semantic data sources for reuse in particular application domains and scenarios. Within this framework users are able to define and execute custom summarization processes to generate useful dataset summaries. A custom summarization process can be seen as an orchestration of primitive predefined parameterizable data analysis processes each of which may deal with a different aspect of the data. More importantly, such a process is linked to a particular goal/problem/need that it is supposed to serve, thus forming a reusable knowledge component that can be shared among multiple users with similar needs.

The structure of the rest paper is as follows: In the next section, we outline the key aspects of our approach and the basic components of our summarization framework. In Section III, we discuss a particular small-scale application of our framework in a dataset evaluation scenario, and, in Section V, we conclude and outline our future work plans.

II. Semantic Dataset Summarization Framework

Our proposed summarization framework aims to enable its intended users to answer the following question: *"Given an application scenario where semantic data is required, how suitable is a given existing dataset for the purposes of this scenario?"*. To answer this question, users normally need to be able to: i) explicitly express the requirements that a dataset needs to satisfy for a given task or goal and ii) automatically measure/assess the extent to which a dataset satisfies each of these requirements and compile a summary report.

To implement these two capabilities, we follow a *checklist-based* approach. Checklists are practically lists of action items arranged in a systematic manner that allow users to record the completion of each of them and they are widely applied across multiple industries, like healthcare or aviation, to ensure reliable and consistent execution of complex operations [6]. In our case, we apply checklists to define and execute custom dataset summarization tasks in the form of lists of goal-specific requirements and associated summarization processes. In the following paragraphs, we explain how such tasks and processes may be represented, created and used.

A. Summarization Task Representation

To represent custom summarization tasks according to the aforementioned checklist paradigm, we adopt the *Minim model* [7] that allows us to represent for concrete instances of summarization tasks the following information:

- The **Goals** the dataset summarization task is designed to serve. In the Minim's terminology [7], these are called constraints and they are used to denote the purpose of the summarization task and the intended use of the produced summary. This is important as different tasks may have different purposes (e.g., the requirements for checking whether a dataset is appropriate for disambiguation may be different from those required for question answering) and, thus, the goal-related information is crucial for selecting an already defined task in a given application scenario.
- The **Requirements** (or checklist entries) against which the summarization task evaluates the dataset. For example, we may wish to assess whether a dataset contains particular information about a given domain or topic or that it satisfies particular quality criteria (e.g., consistency). The number and nature of the requirements depend practically on the goal of the summarization task and thus they may be substantially different among different application scenarios.
- The **Data Analysis Operations** that the summarization task employs in order to assess the satisfaction of its requirements. In the Minim's terminology, these operations are called rules and practically they take many forms, from simple execution of queries to complex data processing and analysis algorithms like graph analysis or topic modeling. The assessment of a given requirement may require the execution of multiple operations while the same operation may be used to assess multiple requirements.

B. Summarization Task Creation

To create a summarization task one needs to define its goal(s), its requirements and the associated to these operations. Some high-level requirements that we have already identified and they may be used for multiple goals are the following:

- **Evaluate the dataset's coverage of a particular domain/topic**: This requirement aims to measure the extent to which a dataset describes a given domain or topic. This can be at schema level (e.g., how many and which concepts or relations are defined), at instance level (e.g., how many and which instances of a given concept or relation does the dataset have) or with more complex operations (e.g., comparison with a corpus).

- **Evaluate the dataset's labeling adequacy and richness**: This requirement aims to measure the extent to which the dataset's elements (concepts, instances, relations etc.) are accompanied by representative and comprehensible labels, in one or more languages. This can be useful to assess two things: i) the comprehensibility of the data, i.e., the ease with which human consumers can understand and utilize the data and ii) the quality and usefulness of a dataset as a term thesaurus.
- **Evaluate Connectivity**: This requirement checks the existence of paths between concepts or entities, i.e., whether it is possible to go from a given concept to another on the graph and in what ways. This is can be an important aspect of a dataset related, for example, to its ability to answer queries involving particular related entities.

Each of the above requirements can be implemented by means of one or more data analysis operations. Some operations we have already defined for our framework are the following:

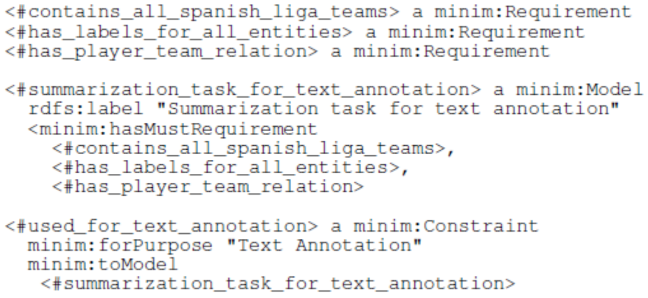
- Check the existence of a particular element (concept, relation, attribute, instance, axiom) in the dataset or of a relational path between particular concepts or instances.
- Measure the number of ambiguous entities in the dataset.
- Measure the number of labeled entities.

C. Dataset Summary Generation

For the generation of goal-specific dataset summaries, we are currently developing a tool that may take as input one or more datasets and a summary goal and run on them specified summarization tasks that correspond to this goal. The output of this tool should be a detailed report about the input datasets, describing whether and to what extend do they satisfy each requirement. The next section provides a concrete example of this output in the context of an actual use case where we applied our framework.

III. Framework Application

A concrete scenario where we applied our framework involved the assessment of public datasets for the purposes of reusing them within a semantic annotation system. In particular, we wanted to annotate texts describing football matches from the Spanish League by means of an in-house ontology-based semantic entity recognition system whose effectiveness depends on the characteristics and quality of the available domain knowledge. For that, we wanted the dataset to be reused to i) contain information about all the current teams of the Spanish football league, ii) all its entities to have at least one associated label and iii) to relate teams with the players that current play in them.



```
<#contains_all_spanish_liga_teams> a minim:Requirement
<#has_labels_for_all_entities> a minim:Requirement
<#has_player_team_relation> a minim:Requirement

<#summarization_task_for_text_annotation> a minim:Model
  rdfs:label "Summarization task for text annotation"
  <minim:hasMustRequirement
    <#contains_all_spanish_liga_teams>,
    <#has_labels_for_all_entities>,
    <#has_player_team_relation>

<#used_for_text_annotation> a minim:Constraint
  minim:forPurpose "Text Annotation"
  minim:toModel
   <#summarization_task_for_text_annotation>
```

Figure 1. Example of Formal Summarization Task Definition

To perform this assessment, we used the model of section II to define a custom summarization task that could help us assess the degree to which some existing datasets satisfied these requirements. A snapshot of the formal definition of the task where the task, its goal and its requirements are defined, is shown in Figure 1.

We executed this task against DBPedia and Freebase, automatically producing the summary report of table I. As one can see the system provides a yes/no answer as to whether each dataset satisfies each requirement but also additional information on why this may or may not be the case (e.g., the percentage of missing labels). The first reason this latter feature is important is that a requirement might not be satisfied because the relevant threshold might have been set too high (e.g., the requirement for 100% labeling). Thus, by showing the actual satisfaction score, the user may decide to relax his/her constraints for the given requirement, especially when there is no dataset fully satisfying it. The second reason is that a requirement might seem to be satisfied, yet that might not be actually true for reasons pertaining to the system's underlying methods and/or the datasets. For example, a closer inspection of the current roster relation in Freebase's website reveals that its instances do not adhere to the semantics of the relation as there are player-team pairs that are no longer valid. Thus, the generated summaries allow users to judge further the suitability of the datasets and refine the requirement rules.

IV. Related Work

Most approaches for semantic data summarization focus on deriving generic goal-independent summaries that provide a high level overview of the data and highlight some of its aspects. For instance, in [3], summaries have the form of questions that can be answered by the dataset, while in [4] summaries consist of the most representative concepts of an ontology, determined based on cognitive and statistical criteria. Nevertheless, these types of summaries are not linked to particular goals nor are they parameterizable.

Relevant to ours work may be also found in the area of semantic data quality where various approaches attempt to define quality criteria and metrics for semantic data. SemRef

Table I
EXAMPLE OF A GOAL-DRIVEN DATASET SUMMARY

Requirement	**DBPedia**	**Freebase**
Spanish League Coverage	YES	YES
At least one label per entity	NO (5% of the entities has no labels)	YES
Player-Team Relation	YES (*“dbpprop:currentclub”*)	YES (*“http://freebase.com/soccer/football_team/current_roster”, “http://freebase.com/soccer/football_player/current_team”*)

[8], for example, defines such criteria for evaluating the quality of semantic metadata with respect to how well they describe a set of resources. A more generic framework is *Sieve* [9] that allows the definition and calculation of custom quality metrics over already available dataset metadata. In that sense it is similar to our approach as it is parameterizable and goal-driven. Nevertheless, our framework goes one step further by allowing also the definition of generation methods for this metadata (in the form of the data analysis operations), thus covering a wider set of use cases.

Finally, checklist-based approaches have been recently used in biology [10] and in scientific workflows [7], though not yet, to the best of our knowledge, for the task of summarizing and evaluating semantic datasets for reuse purposes.

V. Conclusion and Future Work

In this paper, we presented our ongoing work on a framework for the definition and execution of goal-driven semantic data summarization tasks, as a way to enable organizations and practitioners to take better decisions on whether existing datasets are suitable for their purposes. The framework follows the checklist paradigm and uses a formal ontological model to represent summarization tasks by means of goals, requirements and data analysis operations. Our immediate future works include further technical development of the framework, especially in relation to the management of the datasets (a list of available datasets needs to be created and maintained from sites like http://linkeddata.org/datasets, while local endpoints should be created for datasets that currently lack ones). Moreover, additional high-level requirements and data analysis operations will be defined, as well as a User Interface for the definition and generation of semantic data summaries.

Acknowledgment

The research leading to this results has received funding from the People Programme (Marie Curie Actions) of the European Union’s 7th Framework Programme P7/2007-2013 under REA grant agreement n^o 286348.

References

[1] C. Bizer, T. Heath, and T. Berners-Lee, “Linked data - the story so far,” International Journal of Semantic Web Information Systems, vol. 5, no. 3, 2009, pp. 1-22.

[2] S. Szumlanski and F. Gomez, “Automatically acquiring a semantic network of related concepts,” in Proceedings of the 19th ACM international conference on Information and knowledge management. New York, NY, USA, ACM, 2010, pp. 19-28.

[3] M. d’Aquin and E. Motta, “Extracting relevant questions to an RDF dataset using formal concept analysis.” in K-CAP, M. A. Musen and O. Corcho, Eds. ACM, 2011, pp. 121-128.

[4] S. Peroni, E. Motta, and M. dAquin, “Identifying key concepts in an ontology through the integration of cognitive principles with statistical and topological measures,” in Third Asian Semantic Web Conference, Bangkok, Thailand, 2008.

[5] V. Presutti, L. Aroyo, A. Adamou, B. Schopman, A. Gangemi, and G. Schreiber, “Extracting core knowledge from linked data,” in Proceedings of the Second Workshop on Consuming Linked Data, COLD2011, Workshop in conjunction with the 10th International Semantic Web Conference 2011 (ISWC 2011). CEUR-WS.

[6] B. Hales and P. Pronovost, “The checklista tool for error management and performance improvement,” Journal of Critical Care, vol. 21, no. 3, Sep. 2006, pp. 231-235.

[7] K. Belhajjame, M. Roos, E. Garcia-Cuesta, G. Klyne, J. Zhao, D. De Roure, C. Goble, J. M. Gomez-Perez, K. Hettne, A. Garrido, “Why workflows break - understanding and combating decay in taverna workflows.” in Proceedings of the 2012 IEEE 8th International Conference on E-Science. IEEE Computer Society, 2012, pp. 1-9.

[8] Y. Lei, V. Uren, and E. Motta, “A framework for evaluating semantic metadata,” in Proceedings of the 4th international conference on Knowledge Capture, K-CAP 2007, 2007, pp. 135-142.

[9] P. N. Mendes, H. Muhleisen, and C. Bizer, “Sieve: Linked Data Quality Assessment and Fusion,” in 2nd International Workshop on Linked Web Data Management (LWDM 2012) at the 15th International Conference on Extending Database Technology, EDBT 2012, March.

[10] C.F. Taylor, N.W. Paton , K. S. Lilley, P. A. Binz, R. K. Julian Jr, A.R., Jones, W. Zhu, R. Apweiler, R. Aebersold, E. W. Deutsch, M-. J. Dunn, A. J. Heck, A. Leitner, M. Macht, M. Mann, L. Martens, T. A. Neubert, S.D. Patterson, P. Ping, S.L. Seymour, P. Souda, A. Tsugita, J. Vandekerckhove, T. M. Vondriska, J. P. Whitelegge, M- R. Wilkins, I. Xenarios, J.R. Yates 3rd, H. Hermjakob., “The minimum information about a proteomics experiment (MIAPE).” Nature biotechnology, no. 8, Aug. 2007, pp. 887-893.

[11] Freebase, http://www.freebase.com,Accessed: 16/09/2013.

[12] DBPedia, http://dbpedia.org, Accessed: 16/09/2013.

Bridging Semantics Through Ontologies

Tameem Chowdhury
School of Design, Engineering and Technology
University of South Wales
Newport, Wales.
Email: tameem.chowdhury@southwales.ac.uk

Christopher Tubb
School of Design, Engineering and Technology
University of South Wales
Newport, Wales.
Email: Christopher.tubb@southwales.ac.uk

Stilianos Vidalis
Faculty of Computing, Engineering and Technology
Stafford University
Stafford, England.
Email: Stilianos.vidalis@staffs.ac.uk

***Abstract*— Semantic metadata enables contextual and relevant data to be identified for a particular entity. The use of ontologies creates a bridging mechanism, whereby semantic metadata can be referenced and validated to ensure that relevant and useful information is collected. This also ensures trust and logic can be attained in search functionality. The paper explores the foundations of the research for the design of an Information Gathering tool for the Business Intelligence Domain. The aim of the project is to effectively present next to real-time knowledgeable answers to runtime user generated queries for extracting business intelligence. The tool will collect information from disparate sources and requires the implementation of semantics to safeguard the future of knowledge discovery and reuse. This paper summaries the research and conceptualisation for our Information Gathering tool using semantic metadata to be utilised in the area of Business Intelligence.**

Keywords-Semantics; Metadata; Ontology; Business Intelligence.

I. INTRODUCTION

“The World Wide Web was originally built for human consumption, and although everything on it is machine-readable, this data is not machine-understandable. It is very hard to automate anything on the web, and because of the volume of information the web contains, it is not possible to manage it manually” [1].

As the technological growth exponentially increases, the vastness of data and information available for consumption and reuse is equally daunting. Incorporating semantics, specifically semantic metadata, into search functionality and classification, relevance and precision can be enhanced. In order to successfully implement semantic metadata, ontologies can be utilised and these principles can be applied for conducting knowledge extraction for gaining Business Intelligence (BI). The paper discusses the fundamentals of semantic metadata and ontology and how their application will benefit the Intelligence Gathering Using Semantic Metadata and Ontology (IGUSMON) project, currently work in progress. The aim of the tool is to provide next to real-time knowledgeable answers to runtime user generated queries, from disparate sources, in noncritical multimedia systems focusing on BI. We present the design, which combines ideas discussed in “The Semantic Web” [2] with theory proposed from the study of nature, most notably for our research, Swarm Intelligence [3] and proposes how they can be applied to extract knowledge for BI.

The outline for the paper is as follows: Section II will discuss the fundamentals of semantic metadata and the advantages of having well defined concepts for appropriation. It further explores Swarm Intelligence and how the theory studied and documented from research into particular natural systems can help design an efficient computer system, with the ability to utilise logic in its decision-making. Section III presents the design of the IGUSMON project algorithm and analyses the benefits and limitations that may be encountered during the development phase. Related and existing work is also identified.

II. SEMANTIC METADATA AND ONTOLGICAL FUNDAMENTALS

For the design of an Intelligence Gathering tool, the difference between simple information, assets and actual intelligence required definition and identification. Information encapsulates a wide range of concepts and phenomena. They relate to both the processes and material states, which are closely interrelated. Information can be:

- “A product, which encompasses information as an object, as resource, as commodity.

- What is carried in a channel, including the medium channel itself.
- The Contents." [4][5][6][7]

Information can be an asset to stakeholders and or a particular entity, for example, data companies such as Acxiom, TargusInfo and BlueKai [8]. An asset can be defined as a single item of ownership having exchange value [9][10][11][12][13]. Information assets are physical, hardware, software, data, communications, administrative and personnel resources of a computing system [14]. Every information asset contains some sort of information that we can analyse and extract intelligence from.

Intelligence can be defined as a specialised form of knowledge, an activity, and an organisation. As knowledge, intelligence informs leaders, stakeholders or entities, uniquely aiding their judgment and decision-making. As an activity, it is the means by which data and information are collected, their relevance to an issue established, interpreted to determine likely outcomes, and disseminated to individuals and organisations who can make use of it, otherwise known as consumers of intelligence [15]. This becomes more complicated depending on the situation and the stimuli that we are observing and impacts how we extract different intelligence. The application and usability of this intelligence simply depends upon the search criteria and purpose for the collection. For the objectives of the IGUSMON project, collected information will be referenced against ontologies, which will be specifically created for BI, to filter relevant intelligence according to the subjects identified.

An important factor when collecting information that will be classified, as intelligence is the need for accuracy and trust, since the World Wide Web or information environment, unfortunately and inevitably provides a wealth of misinformation. The United States Department of Defense (DoD) has defined the Information Environment (IE) as:

"The aggregate of individuals, organisations and systems (resources) that collect, process, disseminate, or act on information." [16]

Akin to reality, the virtual space is the new realm of warfare and dissemination of misinformation. Clausewitz and Tzu [17][18] theorised about warfare and military mentality and strategy in their respective works, and although the context is different, the theory can still be applied to virtual information warfare. Through the implementation of consistent semantic metadata and well-defined ontologies, BI will be collected, structured, efficiently stored and organised; ensuring they can also be easily retrieved and analysed when required. Furthermore the threat of misinformation can be minimised and or eliminated and trust attributed to the extracted knowledge. Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource. Metadata is often called data about data or information about information [19]. It is utilised in the classification, archiving and most importantly the retrieval of information, data, and resources and assets. If the metadata is maintained and organised correctly, the availability and retrieval is exponentially increased [20][21].

Jokela [20] identifies thirteen categorisations of metadata, of which we have identified the three main types of metadata that will be utilised in the IGUSMON project:

- Descriptive Metadata describes a resource for purposes such as discovery and identification.
- Structural Metadata indicates how compound objects are put together, for example, how pages are ordered to form chapters.
- Administrative Metadata provides information to help manage a resource, such as when and how it was created, file type and other technical information and who can access it [19][20][21].

Metadata is utilised in a variety of different situations by varying institutions. The Police Force, Military facilities, Governments, Libraries, Museums, Internet search engines, Public and Private Sector companies are just a few examples of where metadata is applied and incorporated into everyday tasks and utilised on a daily basis [22]. Foulonneau and Riley [21] add: "Metadata allows various functions to be performed on digital resources, for example, discovery, interpretation, preservation, management, representation and the reuse of objects."

Semantics is the branch of linguistics and logic concerned with meaning. The two main areas are logical semantics, concerned with matters such as sense, reference, presupposition and implication, and lexical semantics, concerned with the analysis of word meanings and relations between them [23]. Semantic Metadata, or meaningful and useful data, are essential in today's information oriented world of discovery and provide the foundations for developing our ontologies.

Simply defining ontology is exigent and requires some background into its lexicology and etymology. Originally the term is from philosophy and denotes a systematic account of existence. In computer science and Artificial Intelligence, ontology is an explicit specification of a conceptualisation and states what exists can be represented [24].

Jokela [20] concurs: "Ontologies are conceptual models that map the content domain into a limited set of meaningful concepts." Formal ontology aims to provide a specification of the meaning of terms within a vocabulary. When conceptualising ontological expressions, the design needs to ensure that the continuants and participants are not stochastically determined [25].

By defining ontologies based on a particular domain [26], the algorithm [27][28] within the Intelligence Gathering tool will facilitate the return of intelligence in a structured manner and only for information predefined within our ontologies for BI. Figure 1 presents a breakdown of the thinking required behind ontology design and will form the foundations for developing our BI ontologies.

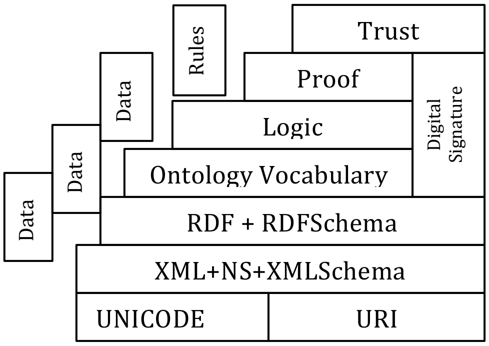


Fig. 1. Visualisation of layers for Ontology Creation [2][29]

The combination of utilising the semantic metadata with the creation of ontologies focusing on the intelligence domain, integrated by an algorithm, will enable the system to simulate and implement logic in its decision-making. The notions put forth by Dumontier and Hoehndorf [25] will also be considered, ensuring that Entities or Subjects can be combined with meaningful Continuants or Objects' respectively [22]. The algorithm will utilise web spiders to collect the data, and use swarming agents to enable communication between the different system components, which include the ontologies.

Swarm intelligence [3][30] theories, developed through research and study into natural systems, are often implemented and utilised in the design of robotic agents. "Theories of Self-Organisation (SO), were originally developed for the contextual benefit of physicists and chemists to describe the emergence of macroscopic patterns" [31][32]. However, SO can be extended to social insects and describe how complex collective behaviour may emerge from interactions along individuals that exhibit simple behaviour but contribute towards the same task. Recent research reflects that SO is indeed a major component of a wide range of collective phenomena in social insects and designers of robotic agents have applied this natural inspiration in the realisation of different robotic agents and artificial systems [22][33]. Social insects have limited cognitive abilities, and therefore the simplicity can be applied to the design of robotic agents, that mimic their behaviour at some level of description [3][31].

The systems of nature and their behaviours are theories, in the continuous processes of study and research and the accuracy of the exact biological science of their physical behaviour is not of importance for our purposes. "Algorithms do not have to be designed after accurate or true models of biological systems; efficiency, robustness and flexibility are the driving criteria, not biological accuracy" [3]. This is why we often use the term biologically inspired. The modelling of social insects by means of SO can help design artificial distributed problem solving devices- swarm-intelligent systems. Although biologically inspired swarm intelligence has an appeal to those developing such systems, it is however, fair to say that very few applications of swarm intelligence have been developed. One of the main reasons for this relative lack of success resides in the fact that swarm-intelligent systems are hard to 'program', because the paths to problem solving are not predefined but emergent, resulting from interactions among individuals and between individuals and their environment, as much as from the behaviours of the individuals themselves [3]. There are two types of emergence, light and strong. Light emergence, where the final behaviour can be deduced from the rules, is in contrast to strong emergence. There are philosophical arguments regarding this; however it is always easier to take a system and analyse how the behaviour results from the interacting rules, than it is in all but trivial cases, to engineer behaviour from simple interacting rules. Therefore, using a swarm-intelligent system to solve a problem requires a thorough knowledge not only of what individual's behaviours must be implemented but also of what interactions are needed to produce such or such global behaviour [3]. This is where ontologies are introduced into the design of our system.

The reduction of the behaviour of these agents can be expressed in equations [3] and have been applied in applications in the areas of Robotics, Information Operations, Evolutionary Computing, Neural Networks, Agent Management and others [30]. Watson adds, "Agent properties can be utilised in: Learning; Social Learning; Environmental Learning; Histories; Cognition and Communications" [30].

III. IMPLEMENTING ONTOLOGIES WITH SEMANTIC METADATA WITHIN THE IGUSMON PROJECT

Web spiders enable the search and retrieval of specific information from the contents of a particular webpage or website. Furthermore, spiders can be programmed to search vast datasets without the need for continuous human interaction. Once the spider is deployed it can crawl from webpage to webpage, through the extraction of hyperlinks and therefore create a list of searchable content. Spiders can implement intelligence gathering through the collection of specific information from disparate sources, relationally stochastic and orthogonal. They can be programmed for the required level of independency, and will function by examining the semantic metadata of the digital resource. The web spiders provide an excellent mechanism for gathering the required websites and the corresponding semantic metadata for the target search, which will then enable the other features of the system to mine and structure the data for presentation in the form of a knowledgeable answer [22].

The research is in its infancy and the following architecture and design described is the conceptualisation of our algorithm for intelligence gathering using semantic metadata. Figure 2 illustrates the conceptual design of the Information Gathering tool, which demonstrates how the web spiders will act as a mechanism for gathering the raw data, before sending the extracted semantic metadata back to the database for validation with the predefined ontologies. Once the extracted data is verified, a data-mining [34] algorithm structures the data into information before returning it as a knowledgeable answer to the Query Management System.

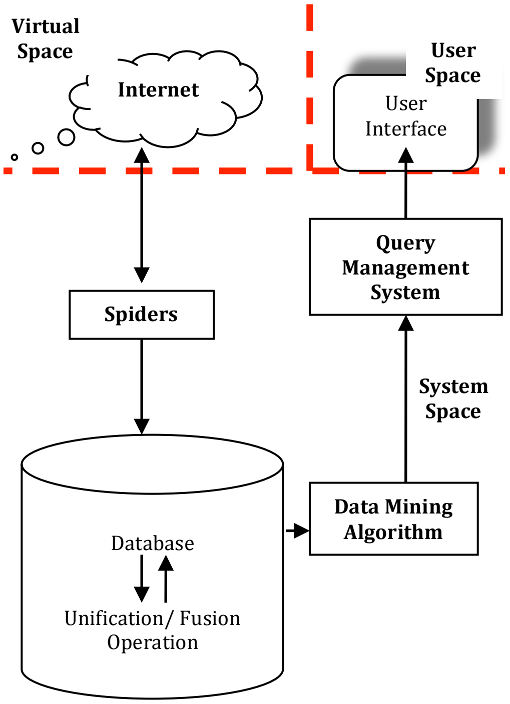


Fig. 2. IGUSMON Project System Architecture [22]

Venturing deeper into the mechanics of the Intelligence Gathering tool and specifically to the core elements of the design, Figure 3 illustrates the System Architecture and the critical elements of the system, as well as how the swarming agents communicate. The Query Management System will signal the release of the web spiders from the spider deployment module via the database and a swarming agent. Collected information will be verified for relevant intelligence within the Validation Module via ontology checks. However, before the semantic metadata reaches the Validation Module, a final check will be conducted via a worker agent against the Irrelevant Data module, where discarded information from previous extractions, that did not produce positive intelligence results relating to a query, are stored.

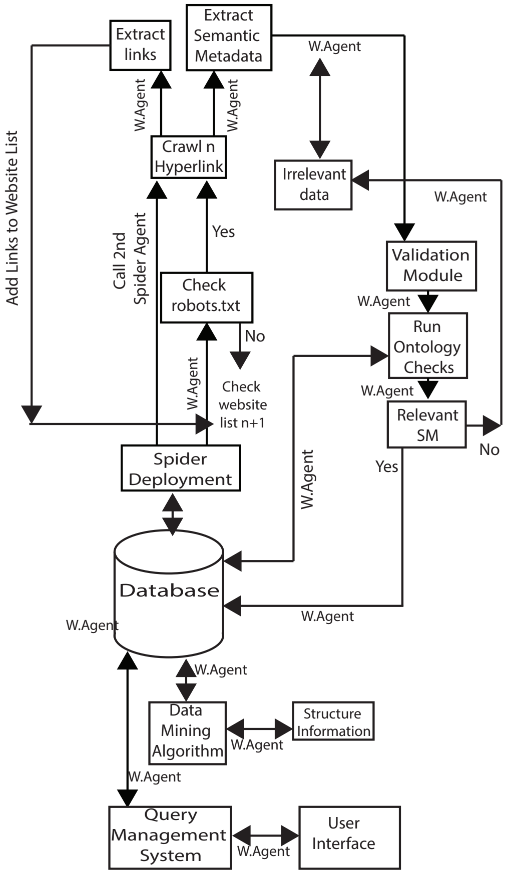


Fig. 3. Implementation of Swarming Intelligence Behaviour with the IGUSMON Project System Architecture

As stated earlier, all information may prove to be intelligence and can be utilised depending upon a particular objective or query; therefore all extracted semantic metadata will be stored, either in the database or the Irrelevant Data module. The information gathered will be filtered through a data- mining [34] algorithm and the architecture of the Data Mining Algorithm will incorporate Floridi's [5] Mathematical Theory of Communication (MTC) in the design, illustrated in Figure 4.

The architecture of the algorithm differs from related work in that it focuses only on extracting semantic meta-data for filtering against our BI ontologies. Furthermore, the application of swarming worker agents within the system ensures that multiple tasks are conducted concurrently. The benefits of this focus are anticipated to ensure vast datasets can be quickly referenced and utilised for extraction. The direct integration of the semantic metadata with the ontologies will ensure that relevant knowledge can be extracted. An obvious limitation to this method will be determined by how much of the relevant data is attributed with semantic metadata. Even though semantic web methods have been proposed for over a decade now, data does exist that was created before and after, which seldom or minimally focuses on semantics. However this does not mean that semantic metadata is limited; with the technological growth and vast amounts of growing data, this limitation is becoming less finite. The other limitation that will impact our research will be the reach of the algorithm. When the conceptualisation of the algorithm is developed, the testing will focus on a finite number of websites for extraction, due to available computing power and time constraints. As mentioned, the focus of the IGUSMON project is currently immersed within this area and development is in progression; some of the design elements proposed may change as the modules are created and tested for feasibility.

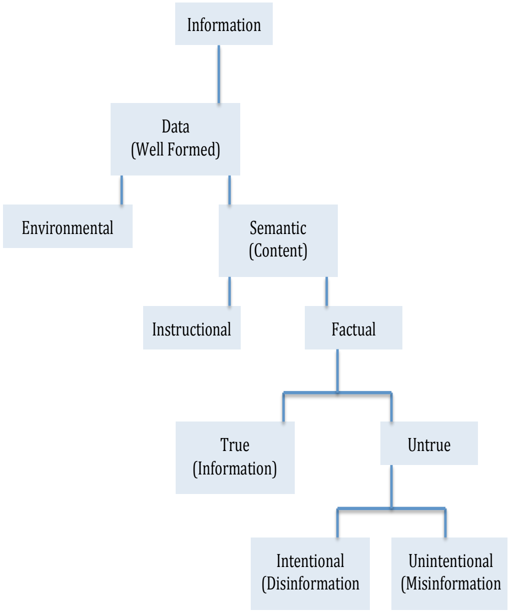


Fig. 4. Mathematical Theory of Communication [5][22]

A. Related Work

The foundations of the research are attributed to Berners-Lee et als. *"The Semantic Web"* [2], Gruber's [24] research on ontologies and Bonabeau et als. [3] research on Swarm Intelligence. Further related work within our research focused on where semantic metadata and ontological mapping has been incorporated within the design, collection and extraction. Jokela [20] implements the use of semantic metadata within media content. Whereas, Stefanov and Huang's [35] research focuses on metadata context management. Vlachidis et al. [36] attribute and refer to this concept of utilising semantic metadata as Semantic Annotation within their research. They incorporate Semantic Annotation within their mechanism responsible for connecting natural language and formal conceptual structures, observing that the incorporation of semantic metadata could enable new information accessibility and enhance existing methods and systems. The IGUSMON project focuses on applying these methods in the area of BI.

IV. CONCLUSIONS AND FUTURE WORK

Implementing ontologies into the design and application of the IGUSMON project, enables relevant information to be defined within a strict set of requirements, so that precise retrieval can be achieved. The sheer volume of information assets or intelligence that can be gathered through search today is overwhelming; the focus on semantic metadata ensures that ontologies can be developed to conceptualise subjects and objects and ultimately enable us to simulate logic in the search for valuable intelligence. The development of the algorithm and the creation of the ontologies for BI have begun. The intention for demonstrating the successful completion of the algorithm and architecture will be through the use of a user interface, enabling users to submit runtime generated queries. The design of the algorithm and overall architecture of the tool, will ensure that if the ontologies are modified, there will be minimal disruption and ensures that any expansion of search parameters can be integrated. Semantics enable contextual and relevant intelligence to be gathered; the extensibility of the database storing the ontologies ensures that additional information and specifically triplets, can be incorporated when a limitation is identified. This is a key factor since the web spiders will retrieve information specified by their defined semantic metadata, and as linguistics and modern languages have taught us throughout history, the semantics of words and expressions are always evolving to reflect changes in society.

ACKNOWLEDGMENT

Tameem Chowdhury would like to acknowledge the European Social Fund for the funding provided through The Knowledge Economy Social Skills (KESS) Ph.D. Scholarship Scheme, enabling the opportunity to conduct this research and development for the IGUSMON Project.

REFERENCES

[1] W3C. RDF Syntax. [www] http://www.w3.org/TR/PR-rdf-syntax/, 1999. [retrieved: July 2013].
[2] T. Berners-Lee, J. Hendler, and O. Lassila. "The semantic web." Scientific American, 284(5): pp. 34–43, 2001.
[3] E. Bonabeau, M. Dorigo, and G. Theraulaz. "Swarm Intelligence From Natural to Artificial Systems." New York: Oxford University Press, 1999.
[4] M. Menou. "The impact of information ii: Concepts of information and its value." Information Processing & Management, 31(4): pp. 479–490, 1995.
[5] L. Floridi. "Semantic conceptions of information." In Edward N. Zalta, editor, The Stanford Encyclopedia of Philosophy. Spring 2011 edition, 2011.

[6] N. Belkin. “Information concepts for information science.” Journal of Documentation, 34(1): pp. 55-85, 1978.

[7] B. Brookes. “The foundations of information science. Part i. philosophical aspects.” Journal of Information Science, 2(3-4): pp. 125–133, 1980.

[8] E. Pariser. The Filter Bubble. Penguin Group, 2011.

[9] K. Choong. “Intellectual capital: definitions, categorization and reporting models.” Journal of Intellectual Capital, 9(4): pp. 609–638, 2008.

[10] P. Kostagiolas and S. Asonitis. “Intangible assets for academic libraries: Definitions, categorization and an exploration of management issues.” Library Management, 30(6/7): pp. 419–429, 2009.

[11] L. Joia. “Measuring intangible corporate assets: Linking business strategy with intellectual capital.” Journal of Intellectual Capital, 1(1): pp. 68–84, 2000.

[12] L. Kaufmann and Y. Schneider. “Intangibles: A synthesis of current research.” Journal of Intellectual Capital, 5(3):366–388, 2004.

[13] M. Harvey and R. Lusch. “Protecting the core competencies of a company: Intangible asset security.” European Management Journal, 15(4): pp. 370–380, 1997.

[14] T. Chowdhury, S. Vidalis, and C. Tubb. “Proactively defending computing infrastructures through the implementation of live forensic capture in corporate network security.” In The 3rd International Conference on Cybercrime, Security and Digital Forensics. 2013. In press.

[15] N. Hendrickson. “Critical thinking in intelligence analysis.” International Journal of Intelligence and Counter Intelligence, 21(4): pp. 679–693, 2008.

[16] S. Vidalis and O. Angelopoulou. “Deception and manoeuvre warfare utilising cloud resources.” In The 3rd International Conference on Cybercrime, Security and Digital Forensics. 2013. In press.

[17] C. Von Clausewitz. On War. Princeton: Princeton University Press. 1976.

[18] S. Tzu. The Art of War by Sun Tzu Special Edition (Translated and annotated by Lionel Giles). El Paso Norte Press, 1910.

[19] NISO. Understanding Metadata. National Information Standards Organization: Bethesda, 2004.

[20] S. Jokela. Metadata Enhanced Content Management In Media Companies. Acta Polytechnica Scandinavica, Mathematics and Computing Series No. 114, Espoo: Finnish Academies of Technology, 2001.

[21] M. Foulonneau and J. Riley. Metadata for Digital Resources. Oxford: Chandos Publishing (Oxford) Limited, 2008.

[22] T. Chowdhury, S. Vidalis, and C. Tubb. “An ontological approach to intelligence gathering using semantic metadata.” The Journal of Communication and Computer US. 2013. In press.

[23] Oxford University Press. [www] http://oxforddictionaries.com/definition /semantics, 2012. [retrieved: January 2012].

[24] T. Gruber. “A translation approach to portable ontology specifications.” Knowledge Creation Diffusion Utilization.” 5(April): pp.199–220, 1993.

[25] M. Dumontier and R. Hoehndorf. “Realism for scientific ontologies.” *6th International Conference on Formal Ontology in Information Systems*. volume 209, pp. 387–399. IOS Press, 2010.

[26] M. Oldfield. Domain Modelling. [www] http://www.aptprocess.com, 2002. [Retrieved: March 2012].

[27] S. Dasgupta, C. Papadimitriou and U. Vazirani. Algorithms. McGraw-Hill, 2006.

[28] T. Cormen, C. Leiserson, R. Rivest and C. Stein. Introduction to Algorithms. The MIT Press: Cambridge, Massachusetts and London, England. 3rd edition, 2009.

[29] P. Mankato. The Semantic Web - An Overview. [www] www.youtube.com, 2011. [Retrieved: September 2012].

[30] L. Watson. Swarming In Information Warfare. Edith Cowan University, Perth, Western Australia, 2002. [Retrieved: October 2012].

[31] H. Haken. Synergetics. Berlin: Springer-Verlag, 1983. (Cited in E. Bonabeau, M. Dorigo, and G. Theraulaz. “Swarm Intelligence From Natural to Artificial Systems.” New York: Oxford University Press, 1999).

[32] G. Nicolis and I. Prigogine. “Self-Organiization in Non-Equilibrium Systems.” NewYork, NY: Wiley & Sons, 1977. (Cited in E. Bonabeau, M. Dorigo, and G. Theraulaz. “Swarm Intelligence From Natural to Artificial Systems.” New York: Oxford University Press, 1999).

[33] J. Deneubourg, S. Goss, N. Franks, and J M. Pasteels. “The blind leading the blind: Modeling chemically mediated army ant raid patterns.” Journal of Insect Behavior, 2(5): pp. 719–725, 1989.

[34] B. Palace. Data Mining. Technology Note prepared for Management 274A, Anderson Graduate School of Management at UCLA [www] http://www.anderson.ucla.edu, 1996. [Retrieved: May 2012].

[35] S. Stefanov and V. Huang. “A semantic web based system for context metadata management.” Metadata and Semantic Research, 46: pp. 118–129, 2009.

[36] A. Vlachidis, C. Binding, D. Tudhope, and K. May. “Automatic metadata generation in an archaeological digital library: Semantic annotation of grey literature.” 2012. [www] hypermedia.research.glam.ac.uk [Retrieved: October 2012].

Effectiveness Gain of Polarity Detection Through Topic Domains

Faiza Belbachir
University of Toulouse, IRIT UMR 5505 CNRS
118, route de Narbonne
F-31062 Toulouse cedex 9. France
Faiza.Belbachir@irit.fr

Malik M. S. Missen
The Islamia University of Bahawapur
Departement of Computer Science
and IT. Pakistan
Saad.missen@gmail.com

***Abstract*—Most of the work on polarity detection consists in finding out negative or positive words in a document using sentiment lexical resources. Indeed, some versions of such approaches have performed well but most of these approaches rely only on prior polarity of words and do not exploit the contextual polarity of words. Sentiment semantics of a term vary from one domain to another. For example, the word "unpredictable" conveys a positive feeling about a movie plot, but the same word conveys negative feeling in context of operating of a digital camera. In this work, we demonstrate this aspect of sentiment polarity. We use TREC Blog 2006 Data collection with topics of TREC Blog 2006 and 2007 for experimentation. The results of our experiments showed an improvement (95%) on polarity detection. The conclusion is that the context plays a role on the polarity of each word.**

Keywords-opinion; polarity; blogs; information retrieval; query categorization.

I. Introduction

Opinion retrieval aims at relating documents that are both relevant to the query (topic) and express opinions about it. It suffers from problems that are different from the ones that occur in classical information retrieval where the subject is identified only by keywords [14][15].

The opinion conveyed by a text can be expressed by very subtle and varied words, therefore it is often difficult to exactly determine it. The classification of sentiments (polarity) is a sub-task in opinion detection [23][27]. It consists in determining whether an opinion in a given document is positive or negative, which has been challenged at Text Retrieval Conference (TREC) Blog Track since 2006 [28]. The approaches explored by track participants can be devised in two types of approaches for opinion and polarity detection. Some of them are based on the lexicon of opinion words, others on machine learning [17][20].

The first type of approach uses a lexicon of opinion words. This lexicon can be general (such as SentiWordNet [21], General Inquirer [22], Subjective Lexicon [25]), built manually or generated automatically from the corpus (words that contain an opinion are taken directly from the corpus). Each word in the lexicon is associated with opinion and polarity scores. These scores are exploited by different approaches to compute the opinion (or polarity) score of a document. A simple method is to assign a score equal to the total number of words containing an opinion (or polarity) in the document [4][20].

The second type of approach is based on machine learning. This type of approach has two aspects: the level of the features (it is the characteristics of opinion word that determine whether a document contains opinions or not), and the type of classifier. The main features that are used are: single words, bi-grams, trigrams, part of speech and the main classifiers that are used in the polarity detection are: SVM, Naive Bayes, Logistic Regression [5][20]. Other works use a mixed approach (machine learning and lexicon) [13][14].

However, most of previous work do not take into account the context of words. The context can be defined by negation, word senses, syntactic role of words around the given word, intensifiers (or diminishers), or the domain of the topic. The prior polarity of a word is sometimes subject to changes under its context. The new polarity of the word defined by its context is called its contextual polarity. Let us take examples to illustrate what contextual polarity is:

- Negation: Polarity assigned to the term happy is positive, but if this term is preceded by negation word such as "not" or "never", its polarity changes and becomes negative.
- Word sense: the word "Car" has different meanings. For example it means "a motor vehicle with four wheels; usually propelled by an internal combustion engine" or "the compartment that is suspended from an airship and that carries personnel and the cargo".
- Intensifiers: "very bad" (intensifiers), "little problem" (diminishers).
- Domain of topic: the word "unpredictable" gives a positive feeling while writing a movie plot but the same word is negative about the features of a digital camera.

The above examples show that a word changes meaning (polarity) according to several characteristics (Negation, Word sense, Domain of topic). These characteristics are part of polarity context. We are interested in one part of the polarity context, it is the domains of the topic. Our basic assumption is that a word changes its polarity from one topic to another, e.g., "unpredictable". To investigate this question we propose to categorize the topics into classes (domain), so that an opinion word has the same polarity for all topics of the same class. Then, we determine the polarity for each class.

In this paper, we show the impact of the context in the polarity detection by conducting experiments on data sets of various domains. We use TREC Blog (Text Retrieval Conference) 2006 Data collection with topics of TREC Blog 2006 and 2007 for experimentation purposes [19]. We use a machine learning system and simple features as number of positive words, number of negative words, number of neutral words, and the number of adjectives in a text to the polarity detection. We categorize the topics into six classes (Films, Person, Organization, Event, Product, Issue), and show that this categorization improves the opinion detection. The goal isn't to use sophisticated level of linguistic analysis but it is to show the impact of topic domain on polarity detection.

The remainder of this paper is organized as follows. In the Section 2, we present the related work. In Section 3, we describe the Text Retrieval Conference (TREC). Sections 4, 5 and 6 describe our experiments. We, then, conclude the paper and give some remarks about the related future work.

In this work, we have evaluated the effectiveness of using topic domains on sentiment detection using a standard data collection. It is found that using topical knowledge of topics helps increasing effectiveness of sentiment detection.

II. RELATED WORK

Few works exist that have proposed approaches to identify the contextual polarities in opinion expressions [7][9][12]. Yi, Nasukawa, Bunescu and Niblack [9] use a lexicon and manually developed high quality patterns to classify contextual polarity. Their approach shows good results with high precision (75-95%) over the set of expressions that they evaluate.

Popescu and Etzioni [7] use an unsupervised classification technique called relaxation labeling [10] to recognize the contextual polarity of words. They adopt a three-stage iterative approach to assign final polarities to words. They use features that represent conjunctions and dependency relations between polarity words.

Suzuki, Takamura and Okumura [12] use a bootstrapping approach to classify the polarity of tuples of adjectives and their target nouns in Japanese blogs. Negations (such as "only" and "not") were taken into account when identifying contextual polarities. The problem with the above approaches is their limitation to specific items of interest, such as products and product features, or to tuples of adjectives and nouns.

In contrast, the approach proposed by Wilson, Wiebe and Homan [11] classifies the contextual polarity of all instances of the words in a large lexicon of subjectivity clues that appear in the corpus. Included in the lexicon are not only adjectives, but nouns, verbs, adverbs, and even modals. They dealt with negations on both local and long-distance levels. Besides this, they also included clues from surrounding sentences. It was the first work to evaluate the effects of neutral instances on the performance of features for discriminating between positive and negative contextual polarity.

III. TEXT RETRIEVAL CONFERENCE TREC

Text Retrieval Conference (TREC) was stated in year 1992 with the sponsor of U.S. Department of Defense and U.S. National Institute of standards and Technology (NIST). The objective of the TREC is to support and encourage IR by providing an infrastructure for evaluation of text retrieval methodologies. This infrastructure is composed by: a test data collection (Table I), a set of queries (Table II) and a set of relevance assessments (qrels) (Table III).

Table I
TREC BLOG 2006 COLLECTION DETAILS [28]

Characteristic	**Value**
Number of Unique Blogs	100,649
RSS	62%
Atom	38%
First Feed Crawl	06/12/2005
Last Feed Crawl	21/02/2006
Number of feed Fetches	753,681
Number of Permalinks	3,215,171
Number of Homepages	324,880
Total Compressed size	25 GB
Total Uncompressed size	148 GB
Feeds (Uncompressed)	38.6 GB
Permalinks (Uncompressed)	88.8 GB
Homepages (Uncompressed)	20.8 GB

Table II
STANDARD TREC BLOG TOPIC FORMAT

```
<top>
<num> Number: 851 </num>
<title> March of the Penguins </num>
<desc> Description:
Provide opinion of the film documentary "March of the Penguins".
</desc>
<narr> Narrative:
Relevant documents should include opinions concerning the film
documentary "March of the Penguins".
Articles or comments about penguins outside
the context of this film documentary are not relevance.
</narr>
</top>
```

Many tracts are considered by TREC as blog Track, many tasks are defined in this Track for example: Opinion Finding Retrieval Task and Polarity Opinion Finding Retrieval Task. Several data collection with their relevance judgments(baseline) for different IR tasks were provided par TREC. For the blogs Track, TREC has released two data collections: Blog 2006 and Blog 2008. From 2006 to 2009,

Table III
TREC BLOG RELEVANCE JUDGEMENTS LABELS

Label	Caption	Description
-1	Not Judged	A label of -1 means that this document was not examined at all due to offensive URL or Header
0	Not Relevant	The post and its comments are not at all relevant to the topic
1	Relevant	The post or its comments contain some information about the topic but no opinion found about the topic concerned
2	Relevante, Negative Opinions	The post is relevant and contain a negative sentiment for the topic
3	Relevant, Mixed Positive and Negative Opinions	The post is relevant and contain both positive and negative opinions about the topic
4	Relevant, Positive Opinions	The post is relevant and explicitly positive about the topic

Table IV
JUDGMENT OF ANNOTATORS FOR DIFFERENT TOPICS

ANNOTATOR 1	ANNOTATOR 2						
	TV	PE	OR	EV	PR	IS	TOT
TV	12	0	0	3	0	1	**16**
PE	0	20	1	0	0	0	**21**
OR	0	0	14	3	0	2	**19**
EV	0	1	0	7	0	2	**10**
PR	0	0	1	0	13	3	**17**
IS	0	0	0	2	0	15	**17**
TOT	12	21	16	15	13	23	**100**

Table V
THE FINALE TOPIC CATEGORIZATION

CLASS	TOPICS 2006	TOPICS 2007	TOT
TV	9	3	**12**
PE	11	10	**21**
OR	9	8	**17**
EV	3	10	**13**
PR	5	10	**15**
IS	13	9	**22**

TREC has been providing 50 new topics each year. For our work, we choose to evaluate experimentation using TREC blog 2006 data collection with topics of year 2006 and 2007.

IV. CATEGORIZATION OF TOPICS

We propose to classify the topics of TREC blogs 2006 and TREC blogs 2007 into six classes: TV (TV), Person (PE), Organization (OR), Event (EV), Product (PR), Issue (IS). This categorization was built manually and inspired by [20] (Table IV).

Each topic of TREC blog 2006 and 2007 was marked by two people (PHD students) called annotators. In the instructions, annotators were asked:

- to read the descriptions, the title of each topic.
- to assign one class among the available classes.

We showed that there is small disagreement (Kappa = 0.77) between the annotators: for the topics of year 2007 "15 disagreements" and only one for 2006. To solve the disagreements of the two annotators, a third annotator was asked to classify these topics. Table V shows the results.

We conducted experiments on the polarity detection using this topic categorization. We worked only with relevant documents of these topics. We then analyzed the effects of this categorization. We performed experiments in two phases. In the first phase, we performed experiments of polarity detection without categorization of topics. In the second phase, we use the categorization of topics to detect polarity. The result of those experimentations was compared with the relevance judgment of TREC.

V. POLARITY DETECTION WITHOUT CATEGORIZATION OF TOPICS

We used a logistic regression model for our experiments. We chose some simple and common features of polarity detection (number of positive words, number of negative words, number of neutral words, and the number of adjectives), as already used in [1]. The experiments for the polarity detection without categorization of topics are devised in three different environments. All experiments and their parameters are explained below:

A. *First experiment*

The experiment was performed using the same features as those explained above. A cross-validations were performed for topics of 2007. The evaluation measures being used to report results are MAP (Mean Average Precision) and P@10 (Precision at 10 documents). More these measures are higher, more the detection of polarity is better. Table VI shows the results of polarity finding MAP and Precision. In this experiment, the data used in the learning phase are much larger than the data used for the testing, because of that the results are not significant. Therefore, before discussing other causes that could improve these results, we conduct another experiment using a small number of learning data for experiments without topics categorization.

Table VI
RESULTS OF THE FIRST EXPERIMENTATION

RUN	POS		NEG	
	MAP	P@10	MAP	P@10
EXPERIMENT 1	0.099	0.200	0.065	0.060

B. *Second experiment*

In this context, the learning data was reduced from 40 to 22 topics. 22 is the maximum number of topics in a group categorization (Table IV) and the choice of topics of the test was done in numerical order: the first test was done for the topics from 901 to 910, the second for the topics from 911 to 920, the third for topics from 921 to 930, the fourth for topics from 931 to 940 and the fifth for topics from 941 to 950.

Table VII
RESULTS OF THE SECOND EXPERIMENTATION

RUN	POS		NEG	
	MAP	P@10	MAP	P@10
EXPERIMENT 2	0.163	0.200	0.062	0.058

The problem that can arise is that the topics of the same class may be in the test and in the learning, which should be avoided. Therefore, we conduct another experiment using a third parameter.

C. *Third experiment*

For this experiment, we wondered about performance when an item of an unknown class has to be processed for polarity detection. To test robustness, we designed an experiment for which we train the classifier on all classes (e.g., Event, Product, TV, Person, Organization) but one (e.g., Issue which acts as the unknown class). For the testing phase, we submitted topics of "Issue" class to the classifier and measured performance. This process was repeated for all the 6 classes. Then, we averaged the results, which are showed in Table VIII. Notice that this intends to evaluate our classifier in the worst situation.

The results of this last experiment are even worse than other results. This leads to the conclusion that a model learned from a data of this topic is not suitable for data of another topic. Next, we present our experimentation with the categorization of topics, and compare the results.

Table VIII
EXPERIMENTS ON THE POLARITY WITHOUT CLASSIFICATION OF TOPICS

RUN	POS		NEG	
	MAP	P@10	MAP	P@10
EXPERIMENT 3	0.055	0.072	0.036	0.054

VI. Detection of polarity with categorization of topics

In this section, we used the same features as those used for the detection of polarity without categorization, namely: number of positive words, number of negative words, number of neutral words, and the number of adjectives. A group of topics has been created for each class. We considered the topics of TREC 2006 and TREC 2007 classified in six classes (films, person, organization, event, product, issue). (N-1) cross validation was performed among topics in each group, using a Logistic Regression model [17], where *N* is the number of classes (6).

The comparison between "with categorization of topics" and "without categorization of topics" (that is, in the worst case) intends to show the benefit of topic categorization. The results are shown in Table IX, where MAP and P@10 are averaged across all topics.

Table IX
COMPARISON OF RESULTS FOR THE POLARITY

	POS		NEG	
	MAP	P@10	MAP	P@10
WITHOUT CATEGORIZATION	0.055	0.072	0.036	0.054
WITH CATEGORIZATION	0.109	0.146	0.068	0.068
% IMPROVEMENT	98.18	102.77	85.24	26.87

These results show that the classification of topics has improved the results for all experiments that have been made. We considered the last experiment (the third experiment in Section 4) as the baseline for comparisons because it represents the worst case situation (with a new class to process). A considerable improvement (98.18% Map) can be noted in the results. These results showed that the categorization of topics can improve the detection results of the polarity. It should be noted that the purpose of this work was not to improve the previous work to detect the polarity, but rather to analyze the effects of classification on the task of detecting opinions.

Figures 1 and 2 show an improved measurement of each MAP TREC topic 2007 for positive and negative polarities. These figures showed that the topics for which significant improvement (was validated through t-test (with $p < 0.05$)) was found in both polarities, are those belonging to classes "Event","Issue" or "Person"(902, 907, 908, 924, 938, etc.).

Copyright (c) IARIA, 2013.   ISBN: 978-1-61208-293-6

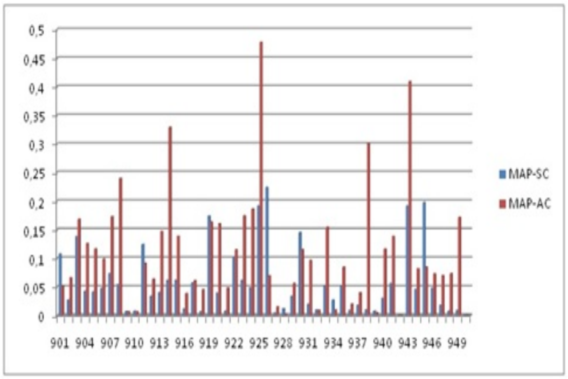


Figure 1. The result of Positive MAP in the various topics of TREC, using the two approaches: with categorization "MAP-AC" and without categorization "MAP-SC".

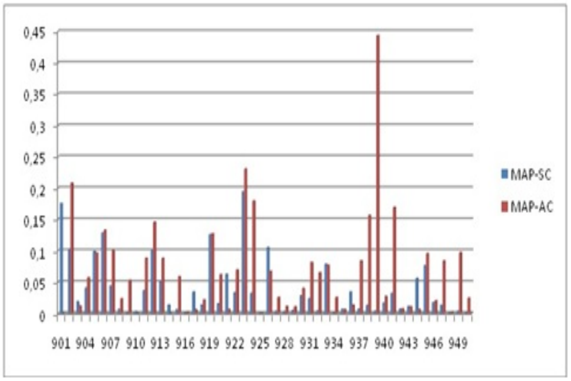


Figure 2. The result of Negative MAP in the various topics of TREC, using the two approaches: with categorization "MAP-AC" and without categorization "MAP-SC".

Tables X and XI show the improvement of few topics of this classes. The MAP of positive and negative words for the first approach ("without categorization") is very low compared to the second approach ("with categorization").

Table X
THE RESULT OF FEW TOPIC (TREC) FOR POSITIVE WORDS

TOPIC	MAP-SC	MAP-AC	Improvement %
902	0.027	0.066	144.444
907	0.073	0.172	134.690
908	0.0541	0.239	342.513
924	0.047	0.186	288.726

One reason why a significant improvement is obtained in these classes may be due to the number of topics in the training data sets. Topics number of class "Issue" and class "Person" are, respectively, 22 and 21.

Table XI
THE RESULT OF FEW TOPIC (TREC) FOR NEGATIVE WORDS

TOPIC	MAP-SC	MAP-AC	Improvement %
902	0.102	0.207	102.239
907	0.044	0.100	127.272
908	0.006	0.023	270.312
924	0.031	0.179	463.836

However, this justification does not hold for the class "Events" where we have 13 topics in total which is less than the class "Org" (17 topics) and the class "Prod (15)". One possible reason could be the classification itself of the topics. We observed that most conflicts encountered during the categorization of topics were to decide between the topics classified as an "Event" and the topics classified as "Issue". For example, it was difficult to decide whether the "Speech" of the president is an "Issue" or an "Event".

VII. Conclusion and future work

Our work focuses on the detection of polarity in blogs. We assume that the context plays a role on the polarity of each word. One word changes meaning (polarity) when used in different subjects. We proposed two approaches. The first approach uses simple features to determine the polarity. The second approach introduces a categorization of topics and documents relevant to these topics. For each class we use the simple features and Logistic Regression classifier. A comparison of these two methods is made, the second method gives better results than the first with more than 95% improvement. The conclusion is that the domain context improves the result for the polarity detection.

In our work, the ranking of the topics was built manually; in the future, we propose to use categorization algorithms of machine learning (i.e., Support Vector Machine (SVM) [29]) and directory services (Yahoo, Dmoz, etc.) and use different features for each class to improve the polarity detection.

References

[1] L. Hoang, S. Lee, G.Hong, J. Lee, and H. Riml, "A Hybrid Method for opinion finding task", (KUNLP at TREC Blog Track), 2008.

[2] C. Yejin, C. Claire, R. Ellen, and P. Siddharth, "Identifying sources of opinions with conditional random fields and extraction patterns", in Conference on Human Language Technology and Empirical Methods in Natural Language Processing, Vancouver, British Columbia, Canada, 2005, pages 355-362.

[3] D. Hannah, C. Macdonald, J. Peng, B. He and I. Ounis, University of Glasgow at TREC 2007, " Experiments in Blog and enterprise Tracks with Terrier", in TREC: Proceedings of the Text Retrieval Conference, 2007.

[4] K. Yang, N. Yu, and H. Zhang, WIDIT in TREC 2007 Blog Track, "Combining Lexicon-Based Methods to Detect Opinionated Blogs", in TREC: Proceedings of the Text Retrieval Conference, 2007.

[5] M. M. S. Missen and M. Boughanem, "Sentence-level opinion-topic association for opinion detection in blogs", ACIS-ICIS, 2009, pages 733-737.

[6] P. Kolari, A. Java, T. Finin, T. Oates, and A. Joshi, "Detecting spam blogs: A machine learning approach", in proceeding AAAI, 2006, pages 1351-1356.

[7] A. Popescu and O. Etzioni, "Extracting product features and opinions from reviews", in HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, Morristown, NJ, USA, 2005, pages 339-346.

[8] Y. Suzuki, H. Takamura, and M. Okumura, "Application of semi-supervised learning to evaluative expression classification", in Proceedings of CICLing-06, the 7th international conference on Computational Linguistics and Intelligent Text Processing, Mexico City, MX, 2006, pages 502-513.

[9] J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack, "Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques", in Proceedings of the Third IEEE International Conference on Data Mining (ICDM-03), Washington, DC, USA, 2003, pages 427-434.

[10] R. Hummel and S. Zucker, "On the foundations of relaxation labeling processes", Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1987, pages 585-605.

[11] T. Wilson, J. Wiebe, and P. Homann "Recognizing contextual polarity in phrase-level sentiment analysis", in HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, Morristown, NJ, USA, 2005, pages 347-354.

[12] Y. Suzuki, H. Takamura, and M. Okumura, "Application of semi-supervised learning to evaluative expression classification", in Proceedings of CICLing-06, the 7th international conference on Computational Linguistics and Intelligent Text Processing, Mexico City, MX, 2006, pages 502-513.

[13] Y. Lee, S. Na, J. Kim, S. Nam, H. Jung, and J. Lee, "KLE at TREC 2008 Blog Track: Blog Post and Feed Retrieval", in TREC Proceedings of the Text Retrieval Conference, 2008.

[14] T. Huifeng, T.Songbo and C. Xueqi "A survey on sentiment detection of reviews", Journal Expert System with Application 36, 2009, volume Special Publication, pages 10760-10773.

[15] R. Santos, B. He, C. Macdonald and I. Ounis "Integrating proximity to subjective sentences for blog opinion retrieval", in ECIR, Toulouse, France, 2009, pages 325-336.

[16] S. Na, Y. Lee, S. Nam, and J. Lee, "Improving opinion retrieval based on query-specific sentiment lexicon", in ECIR 09, Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval, Berlin, Heidelberg, 2009, pages 734-738.

[17] C. Fautsch and J. Savoy, "UniNE at TREC 2008: Fact and Opinion Retrieval in the Blogosphere", in TREC Proceedings of the Text Retrieval Conference, 2008.

[18] C. Macdonald and S. Ounis, "Overview of the TREC 2007 Blog Track", in Proceedings of the TREC 2007.

[19] E. Voorhees and D. Harman, "TREC Experiment and Evaluation in Information Retrieval". Information Retrieval Journal, Springer, 2008, pages 473-475.

[20] G. Zhou, Joshi H., and C. Bayrak, "Topic categorization for relevancy and opinion detection". In TREC Proceedings of the Text Retrieval Conference, 2007.

[21] A. Esuli and F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining", in Proceedings of the 5th Conference on Language Resources and Evaluation (LREC-06), Genao, Italy, 2006, pages 417-422.

[22] Z. Zhang, Q. Ye, R. Law, and Y. Li, "Automatic detection of subjective sentences based on Chinese subjective patterns", in Computer and Information Science, Springer Berlin Heidelberg, pages 29-36.

[23] Y. Choi and C. Cardie, "Adapting a Polarity Lexicon using Integer Linear Programming for Domain-Specific Sentiment Classification", in Proceedings of Conference on Empirical Methods in Natural Language Processing, Singapore, 2009, pages 590-598.

[24] S. Kim and E. Hovy, "Identifying opinion holders for question answering in opinion texts", in Proceedings of AAAI Workshop on Question Answering in Restricted Domains, Pittsburgh, Pennsylvania 2005.

[25] O. Vechtomova, "Using Subjective Adjectives in Opinion Retrieval from Blogs". In Proceedings of Text Retrieval Conference, TREC 2007.

[26] C. Macdonald and I. Ounis, "The TREC Blogs06 collection: creating and analysing a blog test collection". In Proceedings of Text Retrieval Conference, TREC 2006.

[27] H. Yulan, L. Chenghua, and A. Harith, "Automatically extracting polarity-bearing topics for cross-domain sentiment classification", Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Oregon, Portland, 2011, pages 123-131.

[28] I. Ounis, M. Rijke, C. Macdonald, G. Mishne, and I. Soboroff, "Overview of the TREC-2006 Blog Track", In Proceedings of Text Retrieval Conference, TREC 2006.

[29] C. Corinna and V. Vladimir, "Support-Vector Networks", In Proceeding of Kluwer Academic Publishers, 1995, pages 273-297.

Novel Models and Architectures for Distributed Semantic Data Management

Kuldeep B.R. Reddy
Indian Institute of Technology Madras
Chennai, India
brkreddy@cse.iitm.ac.in

***Abstract*—Semantic data management refers to a range of techniques for the manipulation and usage of data based on its meaning and its rapid growth gives rise to the problems of building novel models and architectures for its distributed management allowing efficient query processing and reasoning. The first part of the work proposes an actor model for distributed semantic data management based on the concept of liquid architectures proposing an actor programming framework and execution environment to store, query and reason over structured RDF data. The motivation being to provide a low latency, high throughput distributed platform for semantic data. The second part of the work proposes a pay-as-you-go model and architecture for providing OWL-based semantics as a service including ontology construction, alignment and noise removal from text documents according to the query workload using hadoop map-reduce framework. The third part of the work proposes a query model, including four initial approaches, to generate interactive suggestions as an aid to the user for better formulation of SPARQL queries.**

***Keywords*-distributed semantic data management; actor-based systems; ontology learning; ontology noise removal; ontology alignment; hadoop map-reduce; interactive sparql querying**

I. Introduction

The goal of this paper is to present ideas for research projects on distributed semantic data management and its querying. The paper sketches initial approaches towards designing novel models and architectures for it.

Section II proposes an actor model for distributed semantic data managment based on the concept of liquid architectures. The actor model abstraction, including a programming model and run-time system, has been used to deploy web services on the cloud and this paper proposes to extend it to store semantic data which would allow decentralized query processing and reasoning. The motivation behind it is to develop a low latency, high throughput distributed platform specifically for semantic data. Sections A,B,C,D explain the actor model, proposed architecture, related work and future work respectively.

Section III proposes a pay-as-you-go model to provide semantics as a service which produces the research problems of query specific ontology construction from text, noise removal and alignment over hadoop map-reduce framework. Sections A,B,C,D explain the owl-based semantics as a service, related work, pay-as-you-go framework and architecture and future work.

Section IV introduces the problem of a query model to generate interactive SPARQL query suggestions over distributed semantic data to allow the user to formulate better queries and presents initial three approaches towards it. Sections A,B,C,D explain the initial four approaches including Ontology-based suggestions, Query-log based suggestions, cache based query reformulation suggestions and exploration based suggestion.

II. LiquidRDF : An Actor Model for Distributed Semantic Data Management

This part of the work proposes an Actor Model for Distributed Semantic Data Management. The foundation of the proposed model is based on the Actor Model of Computation. The motivation for the proposed model is to allow the development of low latency and high throughput platform and allow decentralized SPARQL query processing and reasoning. Actor based Distributed Systems are built on the concept that everyone are actors which are able to abstract away the individual hosts and do not share any memory, instead communicating only through messages. The work proposes an architecture of the model based actor programming framework and run-time system where each actor is a liquid RDF store providing a SPARQL endpoint, which for instance can be programmed in rdfstore-js, which is a JavaScript implementation of RDF stores, according to the concept of liquid architectures and a run-time system based on the one proposed in [8]. The following subsections sketch the actor model, programming model and the run-time system to store and query distributed semantic data.

A. Proposed Actor Model

The Actor model is based on the concept that everything is an Actor [3]. An Actor as a computational entity with a behavior such that in response to each message received can concurrently: Send a finite number of messages to other Actors, Create a finite number of new Actors and Designate the behavior to be used for the next message received. Communications with other Actors occur asynchronously. Actors abstract away the individual host. A number of actor languages such as STAGE [4], have been proposed to build actor based distributed systems.

The information workbench provides the front end user interface which allows the client to pose queries and develop applications. The information workbench, proposed in [5], is realized as a Web Application with AJAX-based front end and a pure Java back-end. At its core is a semantic data store which stores RDF data in triple stores that are accessed through Sesame [6] APIs. The FedX model, proposed in [7], implements the semantic store as a federation-based model of SPARQL endpoints. The proposed work makes use of the liquid architectures principles, as proposed in [8] to model each RDF store providing a SPARQL end-point as an actor allowing the development of semantic web applications on the distributed semantic data as well providing a decentralized query processing capabilities.

B. Proposed Architecture

In the proposed programming framework, each actor is a liquid RDF store, which can be programmed in rdfstore-js, that is a JavaScript implementation of RDF stores. A run-time framework is proposed which is based on the actor programming language execution environment which takes care of naming system for the actors. as well as enabling communication amongst them. The problem of changes in distributed storage, processes and the traffic load as a result of it is part of future work.

1) Programming Framework: Actors In the proposed model, each actor is a liquid RDF store providing a SPARQL end-point. The proposed architecture treats each actor both as a semantic service provider and consumer. Each actor has a service description describing the contents of the liquid RDF store written in the form of SPARQL graph patterns. Actors communicate with each other in the RDF data format. The work also plans to add a feature to actors which would allow them to take control of a set of actors and coordinate them to attain specific goals.

Actor Script To begin with, the work plans to use rdfstore-js [9], which is a JavaScript implementation of RDF store with support for SPARQL queries, to program the liquid RDF stores in the proposed architecture.

2) Run-Time Framework: Theater A theater represents the execution environment of the actors. Each system has a theater running on it, which has multiple threads or processes of frames. The runtime should be able to compile each actor as an independent entity but executed on a separate frame on a separate thread or process.

Manager The manager is the module which runs on all the systems and locates the actors and allows communication of messages between them.

Migration & Load Balancing This module will take care of actor allocations on different frames and their migration across the systems.

C. Related Work

An architecture of a worldwide computing framework was proposed in [10], which consists of a actor programming language, a distributed run-time system and a middleware architecture for load balancing. The proposed model in this work can be considered as an adaptation of this framework for the semantic web that would allow leveraging the over-arching standards of the web and the semantic web via RDF, SPARQL, and JavaScript to devise a viable platform for application development. There has been related work on storing RDF data on existing distributed platforms as in RDF on hadoop [23], p2p systems [21] and agent based systems [22]. This work on the other hand proposes a distributed sys-tems designed specifically for semantic data and comparing it with schemes to store on existing distributed platforms is part of future work.

D. Future Work

1. Devising SPARQL query optimization strategies on the proposed architecture for decentralized SPARQL query processing in the proposed model where each actor evaluates a part of the SPARQL query and transmits the partially bound SPARQL query to other actors. Possible approaches towards optimization can be based on the reputation-based message routing model.
2. Devising RDFS reasoning strategies on the proposed architecture to enable inferring new information and pre-senting additional results for the SPARQL queries including optimizations for forward chaining and backward chaining approaches.
3. Application Development on the proposed architecture.

III. Pay-as-you-Go framework and architecture for OWL-based semantics as a service

This part of the work proposes a Pay-As-You-Go frame-work and architecture for providing OWL-based Semantics as a Service in the cloud. The model is related to the Data as a Service model which is based on the concept that the data is treated as a product and can be provided as a service, whereas in this model the meaning of the data based on OWL ontologies is treated as a product and provided as a service. Schemes to interpret words have traditionally been based on ontologies and the proposed work addresses the problem of OWL ontology management on the cloud in a pay-as-you-go fashion and offering the interpretation of words based on them as a service. The documents are stored in the cloud and the ontologies are built from them in a pay-as-you-go manner. The ontologies are gradually refined and aligned as needed by the query workload. The following subsections sketch the proposed service, pay-as-you-go model and architecture and the associated associated research problems.

A. OWL-based Semantics as a Service

Cloud computing has emerged as a paradigm which deliv-ers hosted services over the internet [11]. Cloud computing relies on sharing of resources to achieve economies of scale. These services have traditionally been classified into three categories : Infrastructure as a Service, Platform as a Service, Software as a Service.

Here, the cloud provider provides OWL-based Semantics as a Service to the client. Semantics refers to the study of interpretation of words and the idea behind the service is that the meaning of words can be provided as a service over the cloud. Schemes to interpret words have traditionally been based on ontologies. An ontology formally represents knowledge as a set of concepts within a domain, and the relationships between pairs of concepts. The work presents the problem of OWL ontology management on the cloud including the building and maintaining of OWL ontologies in a pay-as-you-go fashion and offering the interpretation of words based on them as a service. The proposed service also looks to provide the user with an option to construct an ontology using the ontologies already present in the web using map-reduce.

Related Work Ontology as a Service was proposed in [12], which is based on sub-ontology extraction and merging, whereby multiple sub-ontologies are extracted from various source ontologies, and then these extracted sub-ontologies are merged to form a complete ontology to be used by the user. The proposed work on the other hand proposes a pay-as-you-go framework where the ontologies are constructed and merged from the text documents as per the queries of the user. In addition, the proposed work looks to remove noise in the ontologies and provides a SPARQL query engine in the cloud and therefore represents a more comprehensive solution to the problem.

B. *Pay-as-you-go framework*

Pay-As-You-Go framework has traditionally been used for data management [13]. The idea behind the approach has been to provide some services immediately and gradually form tighter integrations as needed. The factors that lead to the concept have been very large volumes of the information that would require significant cost in order to integrate them upfront. This framework combines the process of integration with the query processing and iteratively forms the connections and refines them as per the query workload.

The proposed work utilizes this concepual framework for managing OWL ontologies in the cloud. The documents are stored in the cloud and the ontologies are built from them in a pay-as-you-go manner. The ontologies are gradually refined and aligned as needed by the query workload. The benefit of this approach is that the entire ontology need not be built upfront thus saving costs and the query processing time is also reduced as only the parts of the ontology which are frequently accessed are learned as a result the query is processed over a much smaller part of the ontology.

Figure 1 illustrates the pay-as-you-go framework for OWL ontology management in the cloud. The user interface accepts the query as input from the user. The parser module parses the query converting its constituent terms into keywords. The ontology construction module takes the keywords as input and uses it to set a context for which the ontology is built from the documents. It also refers the existing ontologies to avoid constructing the part which has already been built earlier. The built ontology is aligned and merged with the existing ontologies using the alignment module, the noise removal module is invoked here to detect and remove the inconsistencies and unsatisfiable concepts. The RDF converter module converts the extracted ontology to the RDF and merges it with existing RDF graph stores it across the nodes. The RDF graphs are repartitioned according to the query workload if necessary. The SPARQL query executor module executes the SPARQL over the converted RDF graph and returns the answers back to the user.

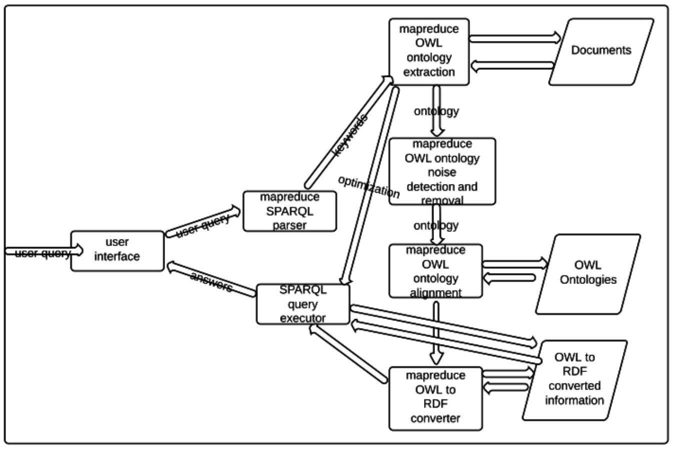

mapreduce OWL ontology extraction
Documents
ontology
keywords
mapreduce OWL ontology noise detection and removal
optimization
mapreduce SPARQL parser
ontology
user query
user query
user interface
mapreduce OWL ontology alignment
OWL Ontologies
answers
SPARQL query executor
OWL to RDF converted information
mapreduce OWL to RDF converter

Figure 1. Pay-As-You-Go Framework for OWL ontology management in the cloud

C. *Pay-as-you-go architecture*

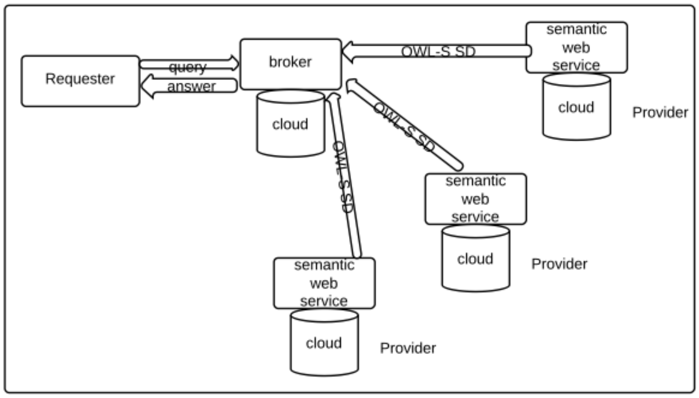


Figure 2. Pay-As-You-Go Architecture of the proposed system

Figure 2 illustrates the pay-as-you-go architecture of the proposed system. Each node illustrates a semantic web service which is implemented using a pay-as-you-go framework in the cloud as described earlier. Each web service releases a service description using OWL-S [14] specification. The OWL-S ontology has three main parts: the service profile, the process model and the grounding. The service profile is used to describe what the service does. The process model describes how a client can interact with the service. The service grounding specifies the details that a client needs to interact with the service. A Broker has emerged as an important component in web services infrastructure by facilitating the discovery and mediation amongst the web services for the client. The proposed architecture implements the broker in the cloud incorporating the idea that the OWL-S service descriptions of the web services are refined over time in a pay-as-you-go manner as per the query workload.

D. Future Work

1. Query-based learning of OWL ontology from documents in the cloud.
The proposed work aims to build a module to learn OWL ontology from the text documents for the given query in the cloud. The idea is to build the OWL ontology in stages as the queries fed as input to the system using the map-reduce framework. The benefit of this approach is that only the relevant parts of the OWL ontology are constructed thus saving the query processing costs. Similar approach to contruct relevant parts of the ontology to the given query was proposed in [15], which involves the user and case-based reasoning. The goal of this module is to devise an approach which constructs the relevant ontology automatically by identifying the topic from the keywords in the queries without involving the user using the map-reduce framework in hadoop and also recognizes that a part of the OWL ontology required for the new query has already been constructed earlier and therefore builds only the required parts.

2. Query-based noise removal in OWL ontology in the cloud.
The proposed work aims to detects and removes noise while the ontology using the information present in the query in the cloud. Noise considered in the work are the inconsistencies and the unsatisfiable concepts present in the ontologies. The existing approaches to detect and remove them have been presented in [16]. The goal of this part of the work is to devise a scheme which treats the additional structural information present in the query as valuable and uses it during the noise removal process to resolve inconsistencies and detect erroneous concepts in the OWL ontology using the map-reduce framework in hadoop.

3. Query-based OWL ontology alignment in the cloud.
The proposed work aims to map and align the ontologies in the cloud taking the help of the additional structural information present in the query. The current approach presented in [17] is based on particle swarm optimization. An ontology alignment is defined as a set of correspondences between ontological entities, i.e. classes, properties, and individuals, of two ontologies. The goal of this module is to devise an approach which looks to utilize the additional structural knowledge present in the query during the alignment process using the map-reduce framework in hadoop.

4. SPARQL query engine for OWL ontology in the cloud.
The proposed work aims to build a SPARQL query engine for querying OWL ontology using the map-reduce framework in hadoop. One of the approaches to query OWL ontology using SPARQL on a stand-alone machine works by computing closure of the entire OWL ontology then converting it to RDF graph and then processing SPARQL queries on the resulting RDF graph. The goal of this module is to devise an approach to identify only a portion of the OWL ontology related to the query, compute its closure and convert it to the RDF graph on the cloud and merge it with the existing RDF graphs and process the query on it. The proposed approach will reduce the execution time considerably as the query is executed on a smaller OWL ontology. The RDF graphs can be partitioned using the METIS graph partitioning algorithm and stored accross different nodes and they can be modified as per the query workload.

5. Pay-as-you-go cloud-based OWL-S broker for semantic web services.
The proposed work aims to build a cloud-based OWL-S semantic web services broker. The model here is that the service descriptions of the web services in the OWL-S format are continously being refined and updated as per the query workload. An architecture of the broker was presented in [18], the proposed work aims to extend the broker implementation to the cloud while making changes to the broker protocol of advertisement and mediation which would allow taking into account the pay-as-you-go refinements of OWL-S service descriptions.

IV. Query Model for Interactive SPARQL query suggestions over distributed semantic data

This part of the paper presents an overview of the research project dealing with generating suggestions to the SPARQL queries given by the user which are executed over distributed semantic data. We present four approaches for it. In the first approach, we recognize that a triple pattern is erroneous in the query by comparing it with the RDFS/OWL ontology. The suggestions to correct the errors in the triple patterns are presented to the user. In our second approach, we make use of the SPARQL query log to produce suggestions. In our third approach, we propose to suggest modifications to the SPARQL query being executed to the user in order to speed up its execution. In the fourth approach, we explore the distrubuted semantic data surrounding the entities discovered during the query execution process looking for paths which are equivalent to the a predicate in the query and present them as suggestions to the user.

A. *Ontology-based suggestions*

In the first approach, we make use of RDFS/OWL ontology to generate suggestion to correct possible errors in the query triple patterns. The approach works by taking each predicate at level i in the SPARQL BGP and comparing its domain and range with the domains and ranges of its preceding and subsequent predicate at levels i-1 and i+1. If their intersection results is empty set, we recognize this predicate as the one which requires correction. We then look to replace the predicate with the predicates from the ontology being used by the user whose domain and range matches that of the preceding and subsequent predicates in the query. With this approach, we get a number of queries with different predicates chosen from the ontology all of which are presented to the user as suggestions. Future work also consists of how the quality of service can be maintained.

B. *Query-log based Suggestions*

This section addresses the problem of generation SPARQL query suggestions as they are being partially entered by the user using the query log. We represent the set of queries entered by the user in the form of a single directed graph. The nodes of this graph contain the triple patterns and the edges contain weights which represent the number of times the triple pattern has been given as part of a query. We normalize the different variable names given in queries into a single set, with each variable representing the nodes at each stage in the query log graph. We maintain a list containing the different combinations of the two variables in the descending order of their occurrance for each node in the query log graph.

When the new query is being entered by the user, the triple patterns entered are matched with the nodes in the query log graph. To produce suggestions, the outgoing edges for the matched node in the graph are collected and sorted in the decreasing order of their weights and the triple patterns which are attached to the edges presented to the user as the next possible triple pattern and the two variables for the triple pattern are picked from the node's associated variable list. We also propose the use of an ontology during the process of generation of suggestions by re-ordering the triple patterns based on how close they are semantically to the user's original query. The semantic distance is computed using the least common ancestor method between two concepts in the ontology.

C. *Cache based Query Reformulation suggestions at Run-time*

This section presents an approach to generate suggestions in order to speed up the execution of SPARQL queries. Our approach proceeds by suggesting the modifications to the query at run-time to the user. The query is modified by either replacing the predicates in it with another set of predicates chosen from a query which was issued earlier or adding new set of predicates from an earlier query whose results are present in the cache. This allows the remaining results for a part of the query to be picked from the cache itself.

In order to replace or add the new predicates in the current query, we take a sample of the intermediate results of the query and compare it with sample of the intermediate results of previous queries during the query execution [19]. If the number of matches exceeds a pre-determined threshold we generate the new query and suggest the modification to the user to either replace it with the new predicates or add the new ones instead, such that the intermediate results are now picked from the cache. The scheme poses the question of how to access the relevant samples of intermediate results of previous query executions efficiently which we plan to address in the future.

D. *Query reformulations suggestion based on exploration*

In the fourth approach, we explore the distributed semantic data looking for paths which are equivalent to the concerned predicate in the query and present them as suggestions to the user. We consider an equivalent path as one which connects the same pair of entities which the concerned predicate in the query connects. Of course, there will be many sets of equivalent paths which requires heuristics to determine the set which can be used to expand the query. The heuristic we use in this paper computes the number of times the query predicate and an probable equivalent path occur together. If the number of times a probable equivalent path occuring with the query predicate exceeds a certain threshold, we confirm it as an equivalent path and use it to expand the query. We also propose the use of an ontology to optimize the search process of semantically equivalent paths. The semantic distance is computed using the least common ancestor method between two concepts in the ontology. Future work also consists of making use of statistics and ontology matching techniques to precisely determine the equivalent path.

E. *Conclusions*

In this part of the paper, we presented an overview of the research project dealing with generating suggestions to SPARQL queries given by user which are executed over the web of linked data. We also proposed four approaches for this problem. Related work is being done in the field of traditional databases to generate suggestions to SQL queries [20] and we believe extending it to SPARQL taking into consideration the semantic web concepts of reasoning and implicit information represents a new research direction worth exploring.

REFERENCES

[1] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data – the story so far," *International Journal on Semantic Web and Information Systems*, vol. 5, no. 3, pp. 1–22, 2009.

[2] O. Hartig and J.-C. Freytag, “Foundations of traversal based query execution over linked data.” in *HT*, E. V. Munson and M. Strohmaier, Eds. ACM, 2012, pp. 43–52. [Online]. Available: http://dblp.uni-trier.de/db/conf/ht/ht2012.htmlHartigF12

[3] C. Hewitt, “Actor model of computation: Scalable robust information systems,” Tech. Rep. v24, July 2012, cite arxiv:1008.1459Comment: improved syntax. [Online]. Available: http://arxiv.org/abs/1008.1459v24

[4] J. Ayres and S. Eisenbach, “Stage: Python with Actors,” in *International Workshop on Multicore Software Engineering (IWMSE)*, April 2009. [Online]. Available: http://pubs.doc.ic.ac.uk/actors-in-python/

[5] P. Haase, M. Schmidt, and A. Schwarte, “The information workbench as a self-service platform for linked data applications.” in *COLD*, ser. CEUR Workshop Proceedings, vol. 782. CEUR-WS.org, 2011. [Online]. Available: http://dblp.uni-trier.de/db/conf/semweb/cold2011.htmlHaaseSS11

[6] J. Broekstra, A. Kampman, and F. van Harmelen, “Sesame: A generic architecture for storing and querying RDF and RDF Schema,” in *The Semantic Web – ISWC 2002: First International Semantic Web Conference Sardinia, Italy*, 2002, pp. 54–68.

[7] K. Hose, R. Schenkel, M. Theobald, and G. Weikum, “Database foundations for scalable RDF processing,” in *Reasoning Web*, 2011, pp. 202–249.

[8] D. Bonetta and C. Pautasso, “Towards liquid service oriented architectures.” in *WWW (Companion Volume)*. ACM, 2011, pp. 337–342. [Online]. Available: http://dblp.uni-trier.de/db/conf/www/www2011c.htmlBonettaP11

[9] A. Garrote. (2011) antoniogarrote / rdfstore-js. [Online]. Available: https://github.com/antoniogarrote/rdfstore-js

[10] T. Desell, K. E. Maghraoui, and C. Varela, “Load balancing of autonomous actors over dynamic networks,” in *In Proceedings of the Hawaii International Conference on System Sciences, HICSS-37 Software Technology Track*, 2004, pp. 1–10.

[11] B. P. Rimal, E. Choi, and I. Lumb, “A taxonomy and survey of cloud computing systems,” in *Proceedings of the 2009 Fifth International Joint Conference on INC, IMS and IDC*, ser. NCM ’09. Washington, DC, USA: IEEE Computer Society, 2009, pp. 44–51. [Online]. Available:

[12] A. Flahive, D. Taniar, and W. Rahayu, “Ontology as a service (oaas): a case for sub-ontology merging on the cloud,” *The Journal of Supercomputing*, pp. 1–32, 2011. [Online]. Available: http://dx.doi.org/10.1007/s11227-011-0711-4

[13] M. Franklin, A. Halevy, and D. Maier, “From databases to dataspaces: a new abstraction for information management,” *SIGMOD Rec.*, vol. 34, no. 4, pp. 27–33, Dec. 2005. [Online]. Available: http://doi.acm.org/10.1145/1107499.1107502

[14] D. Martin, M. Burstein, E. Hobbs, O. Lassila, D. Mcdermott, S. Mcilraith, S. Narayanan, B. Parsia, T. Payne, E. Sirin, N. Srinivasan, and K. Sycara, “OWL-S: Semantic Markup for Web Services,” Tech. Rep., Nov. 2004. [Online]. Available: http://www.w3.org/Submission/OWL-S/

[15] N. B. Mustapha, H. B. Zghal, M.-A. Aufaure, and H. H. B. Ghzala, “Semantic search using modular ontology learning and case-based reasoning.” in *EDBT/ICDT Workshops*, ser. ACM International Conference Proceeding Series. ACM, 201. [Online]. Available: http://dblp.uni-trier.de/db/conf/edbtw/edbtw2010.htmlMustaphaZAG10

[16] P. Haase and J. Vlker, “Ontology learning and reasoning – dealing with uncertainty and inconsistency,” in *Uncertainty Reasoning for the Semantic Web I*, ser. LNCS. Springer Berlin / Heidelberg, 2008, vol. 5327, pp. 366–384.

[17] J. Bock, A. Lenk, and C. Dänschel, “Ontology Alignment in the Cloud,” in *Proceedings of the 5th International Workshop on Ontology Matching (OM-2010)*, vol. 689. http://ceur-ws.org: CEUR Workshop Proceedings, November 2010, pp. 73–84.

[18] M. Paolucci, J. Soudry, N. Srinivasan, and K. Sycara, “A broker for owl-s web services,” in *First International Semantic Web Services Symposium, AAAI Spring Symposium Series*, 2004.

[19] M. Yang and G. Wu, “Caching intermediate result of sparql queries.” in *WWW (Companion Volume)*. ACM, 2011, pp. 159–160.

[20] J. Fan, G. Li, and L. Zhou, “Interactive sql query suggestion: Making databases user-friendly,” in *Proceedings of the 27th International Conference on Data Engineering, ICDE 2011, April 11-16, 2011, Hannover, Germany*. IEEE Computer Society, 2011, pp. 351–362.

[21] Filali. Imen, Bongiovanni. Francesco, Huet. Fabrice and Baude. Francoise, “A Survey of Structured P2P Systems for RDF Data Storage and Retrieval.” in *Transactions on Large-Scale Data and Knowledge-Centered Systems III*. Springer Berlin / Heidelberg, volume 6790, 2011.

[22] Fuhr. Norbert, Klas. Claus-Peter, “Combining RDF and Agent-Based Architectures for Semantic Interoperability in Digital Libraries” in *Proceedings of the DELOS Workshop on Interoperability in Digital Libraries*.2001.

[23] Husain. Mohammad Farhan, Doshi. Pankil, Khan. Latifur, Thuraisingham. Bhavani, “Storage and Retrieval of Large RDF Graph Using Hadoop and MapReduce” in *Cloud Computing*. . Springer Berlin / Heidelberg, volume 5931, 2009.