



# **SEMAPRO 2012**

The Sixth International Conference on Advances in Semantic Processing

ISBN: 978-1-61208-240-0

September 23-28, 2012

Barcelona, Spain

## **SEMAPRO 2012 Editors**

Diletta Romana Cacciagrano, University of Camerino, Italy

Petre Dini, Concordia University, Canada / China Space Agency Center, China

# SEMAPRO 2012

## Forward

The Sixth International Conference on Advances in Semantic Processing (SEMAPRO 2012), held on September 23-28, 2012 in Barcelona, Spain, considered 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.

With the progress on ontology, web services, semantic social media, semantic web, deep web search /deep semantic web/, semantic deep web, semantic networking and semantic reasoning, SEMAPRO 2012 constituted the stage for the state-of-the-art on the most recent advances.

We take here the opportunity to warmly thank all the members of the SEMAPRO 2012 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 that dedicated much of their time and efforts to contribute to the SEMAPRO 2012. We truly believe that thanks to all these efforts, the final conference program consists of top quality contributions.

This event could also not have been a reality without the support of many individuals, organizations and sponsors. We also gratefully thank the members of the SEMAPRO 2012 organizing committee for their help in handling the logistics and for their work that is making this professional meeting a success. We gratefully appreciate to the technical program committee co-chairs that contributed to identify the appropriate groups to submit contributions.

We hope the SEMAPRO 2012 was a successful international forum for the exchange of ideas and results between academia and industry and to promote further progress in semantic processing.

We hope Barcelona provided a pleasant environment during the conference and everyone saved some time for exploring this beautiful city.

**SEMAPRO 2012 Chairs**

**SEMAPRO Advisory Chairs**

René Witte, Concordia University - Montréal, Canada  
Wladyslaw Homenda, Warsaw University of Technology, Poland  
Bich-Lien Doan, SUPELEC, France

**SEMAPRO Industry Liaison Chairs**

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

**SEMAPRO Research Chair**

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

## **SEMAPRO 2012**

### **Committee**

#### **SEMAPRO Advisory Chairs**

René Witte, Concordia University - Montréal, Canada  
Wladyslaw Homenda, Warsaw University of Technology, Poland  
Bich-Lien Doan, SUPELEC, France

#### **SEMAPRO 2012 Industry Liaison Chairs**

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

#### **SEMAPRO 2012 Research Chair**

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

#### **SEMAPRO 2012 Technical Program Committee**

Nasser Alalwan, King Saud University - Riyadh, Saudi Arabia  
Riccardo Albertoni, Consiglio Nazionale delle Ricerche - Genova, Italy  
José F. Aldana Montes, University of Málaga, Spain  
Eckhard Ammann, Reutlingen University, Germany  
Sofia J. Athenikos, Amazon, USA  
Isabel Azevedo, ISEP-IPP, Portugal  
Ebrahim Bagheri, Athabasca University & University of British Columbia, Canada  
Khalid Belhajjame, University of Manchester, UK  
Helmi Ben Hmida, FH MAINZ, Germany  
Jorge Bernardino, ISEC - Polytechnic Institute of Coimbra, Portugal  
Christopher Brewster, Aston University - Birmingham, UK  
Diletta 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  
Sana Châabane, ISG - Sousse, Tunisia  
Delroy Cameron, Wright State University, USA  
Sam Chapman, Knowledge Now Limited, UK  
Smitashree Choudhury, UK Open University - Milton Keynes, UK  
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  
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  
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  
Raimund K. Ege, Northern Illinois University, USA  
Enrico Francesconi, ITTIG - CNR - Florence, Italy  
Raúl García Castro, Universidad Politécnica de Madrid, Spain  
Rosa M. Gil Iranzo, Universitat de Lleida, Spain  
Gregor Grambow, Aalen University, Germany  
Fabio Grandi, University of Bologna, Italy  
Sotirios K. Goudos, Aristotle University of Thessaloniki, Greece  
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 - Plzen, Czech Republic  
Armin Haller, CSIRO ICT Centre - Canberra, Australia  
Shun Hattori, Muroran Institute of Technology, Japan  
Xin He, Airinmar Ltd., UK  
Ralf Heese, Freie Universität Berlin, Germany  
Steffen Henicke, HU-Berlin/Berlin School of Library and Information Science, Germany  
Wladyslaw Homenda, Warsaw University of Technology, Poland  
Carolina Howard Felicissimo, BRGC - Schlumberger, Brazil  
Prasad Jayaweera, University of Sri Jayewardenepura, Sri Lanka  
Wassim Jaziri, ISIM Sfax, Tunisia  
Katia Kermanidis, Ionian University - Corfu, Greece  
Jaroslav Kuchar, Czech Technical University in Prague, Czech Republic  
Kyu-Chul Lee, Chungnam National University - Daejeon, South Korea  
Thorsten Liebig, derivo GmbH - Ulm, Germany  
Sandra Lovrenčić, University of Zagreb - Varaždin, Croatia  
Maria Maleshkova, The Open University, UK  
Maristella Matera, Politecnico di Milano, Italy  
Elisabeth Métais, Cedric-CNAM, France  
Vasileios Mezaris, Informatics and Telematics Institute (ITI) and Centre for Research and Technology Hellas (CERTH) - Thessaloniki, Greece  
Małgorzata Mochól, T-Systems Multimedia Solutions GmbH - Berlin, Germany  
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  
Lyndon J. B. Nixon, STI International, Austria  
Csongor Nyulas, Stanford Center for Biomedical Informatics, USA  
Carlos Pedrinaci, The Open University, UK  
Andrea Perego, Università degli Studi dell'Insubria - Varese, Italy  
Livia Predoiu, University of Magdeburg, Germany  
Hemant Purohit, Wright State University, USA

Jaime Ramírez, Universidad Politécnica de Madrid, Spain  
Isidro Ramos, Valencia Polytechnic University, 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  
Thomas Roth-Berghofer, University of West London, UK  
Michele Ruta, Politecnico di Bari, Italy  
Melike Sah, Trinity College Dublin, Ireland  
Satya Sahoo, Case Western Reserve University, USA  
Minoru Sasaki, Ibaraki University, Japan  
Kinga Schumacher, German Research Center for Artificial Intelligence (DFKI) - Berlin, Germany  
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  
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  
Tian Tian, Manhattan College, New York, USA  
Tania Tudorache, Stanford University, USA  
Roland Wagner, Johannes Kepler Universität Linz, Austria  
Shenghui Wang, OCLC Leiden, The Netherlands  
Wai Lok Woo, Newcastle University, UK  
Filip Zavoral, Charles University in Prague, Czech Republic  
Yuting Zhao, The University of Aberdeen, UK  
Hai-Tao Zheng, Tsinghua University, China

## 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 to 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 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

A Robust Scene Change Detection Using Mode Distribution in H.264 <i>Young-Suk Yoon, Won-Young Yoo, and Young-Ho Suh</i>	1
Towards an Agile E-Commerce <i>Daniela Wolff, Marc Schaaf, and Stella Gatzju Grivas</i>	5
Enabling High Performance Computing for Semantic Web Applications by Means of Open MPI Java Bindings <i>Alexey Cheptsov</i>	11
Ontological Representation of Knowledge Related to Building Energy-efficiency <i>German Nemirovski, Alvaro Sicilia, Fatima Galan, Marco Massetti, and Leandro Madrazo</i>	20
Ontology Search Engines: Overview and recommendations <i>Isabel Azevedo, Carlos Carvalho, and Eurico Carrapatoso</i>	28
Semantic Supply Chain Management <i>Katalin Ternai and Ildiko Szabo</i>	35
Semantics on the Cloud: Toward an Ubiquitous Business Intelligence 2.0 ERP Desktop <i>Diletta Romana Cacciagrano, Emanuela Merelli, Leonardo Vito, Andrea Sergiacomi, and Serena Carota</i>	42
Using DBPedia to Bootstrap new Linked Data <i>Alexiei Dingli and Silvio Abela</i>	48
Word Sense Disambiguation Based on Distance Metric Learning from Training Documents <i>Minoru Sasaki and Hiroyuki Shinnou</i>	54
Searching Documents with Semantically Related Keyphrases <i>Ibrahim Aygul, Nihan Cicekli, and Ilyas Cicekli</i>	59
Optimizing Geographical Entity and Scope Resolution in Texts Using Non-Geographical Semantic Information <i>Panos Alexopoulos and Carlos Ruiz</i>	65
Bi-Directional Ontology Updates Using Lenses <i>Andreas Textor</i>	71
Capturing Knowledge Representations Using Semantic Relationships An Ontology-based Approach <i>Ruben Costa, Paulo Figueiras, Luis Paiva, Ricardo Jardim-Goncalves, and Celson Lima</i>	75
Higher Education Qualification Evaluation	82



*Ildiko Szabo*

Consolidation of Linked Data Resources upon Heterogeneous Schemas 87  
*Aynaz Taheri and Mehrnoush Shamsfard*

Strategies for Semantic Integration of Energy Data in Distributed Knowledge Bases 92  
*Alvaro Sicilia, Fatima Galan, and Leandro Madrazo*

A Semantic Environmental GIS for Solid Waste Management 97  
*Miguel Felix Mata Rivera, Roberto Eswart Zagal Flores, Consuelo Varinia Garcia Mendoza, and Diana Gabriela Castro Frontana*

Comparing a Rule-Based and a Machine Learning Approach for Semantic Analysis 103  
*Francois-Xavier Desmarais, Michel Gagnon, and Amal Zouaq*

Hyponym Extraction from the Web based on Property Inheritance of Text and Image Features 109  
*Shun Hattori*

# A Robust Scene Change Detection Using Mode Distribution in H.264/AVC

Young-Suk Yoon, Won-Young Yoo, and Young-Ho Suh

Contents Protection & Management Research Team  
Electronics and Telecommunications Research Institute  
138 Gajeongno, Yuseong-gu, Daejeon, 305-350, Korea  
{ys.yoon, zero2, syh}@etri.re.kr

**Abstract**—In this paper, we propose a novel scene change detection (SCD) scheme which is available for a semantic video retrieval technique. Using the rate-distortion optimization (RDO) technique used in the H.264 reference software, we have developed an efficient SCD scheme based on the analysis of the mode distribution between intra modes and inter modes. In order to enhance the accuracy of detecting the scene changes, we have also modified the RD function used in RDO technique. Simulation results on several digital videos including abrupt and gradual scene changes show that the proposed scheme provides enhanced performance over previous works.

**Keywords**—Scene change detection; Video retrieval; Mode distribution; H.264/AVC

## I. INTRODUCTION

Multimedia users would like to search a digital video trying to find out in a lot of related digital videos and wish to be recommended ones similar to a query video. Moreover, contents providers need to protect their digital videos from illegal users in order to avoid an infringement of copyright. However, it is difficult to manage and handle massive amount of digital videos including many frames. Therefore, we need to analyze digital videos into their features.

In general, a video sequence can be divided into spatial and temporal features for the efficient analysis such as browsing, indexing, retrieving, monitoring, editing, and authoring digital video. First, the spatial feature depicts edge, texture information, and spatial complexity of a frame. Next, the temporal feature represents the time continuity and discontinuity for scenes, the motion of objects, and optical flows. Especially, a scene change expresses the gap between a scene and the next another for a digital video. A set of scene change is the semantic and reliable information which is used to differentiate videos from each other.

Fig. 1 illustrates frames of a digital video based on a time domain. Video sequences consist of a lot of consecutive frames. A scene is a set of frames connected according to a semantic context of a digital video. Herein,  $Scene_{n-1}$  and  $Scene_n$  have semantically different frame configurations and contexts, and a scene change exists at a boundary between two scenes.

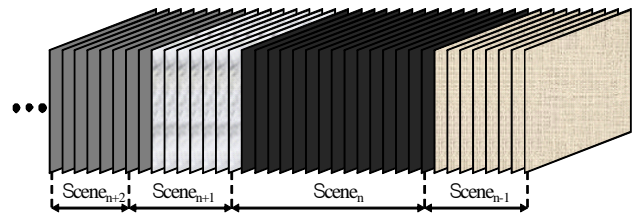


Figure 1. Hierarchical structure of a video sequence

Many researchers have proposed various methods for scene change detection such as pixel-based, histogram-based, edge-based, statistics-based, compression-based, and hybrid methods. Pixel-based methods used the difference of pixel values between successive frames [1]. Gray or color histogram-based schemes compared histograms for neighboring frames, respectively [2]. Edge-based methods employed either object segmentation or edge detection scheme and estimated the degree of change compared with outlines of consecutive frames [3]. Statistics-based methods employed many statistical inferences used in signal processing [4]. Compression-based methods utilized a concept which describes more information is needed to encode a frame when there is a scene change [5]. Hybrid methods apply more than two methods to scene change detection in order to obtain better detecting performance [6].

However, it is not easy to correctly detect a scene change which is necessary for a semantic-oriented video retrieval. In this paper, we present H.264/AVC based mode type [7] classification method for reliable detection of various scene changes. Analyzing the mode distribution between inter and intra modes, we measure temporal and spatial correlation of each frame and utilize the correlation ratio to determine a scene change. Furthermore, we have modified the RD function by using mean removed sum of squared difference (MRSSD) in order to achieve robust scene change detection for illumination change. For the scene change detection between two similar background shots or simple background shots, we have also proposed a selective mode counting (SMC) technique. Moreover, we do not consider the mode information of macroblocks which regions are homogenous or are located in the boundary area.

## II. MODE DECISION IN H.264/AVC

The latest video coding standard, H.264/AVC, uses variable block sizes ranging from  $4 \times 4$  to  $16 \times 16$  in interframe coding. To achieve the highest coding efficiency, H.264/AVC uses rate-distortion optimization (RDO) technique which maximizes coding quality and minimizes resulting data bits. The RDO mode decision method finds the optimal prediction mode in terms of rate distortion. This method computes rate-distortion (RD) cost based on the actual rate and distortion after successive processes, transform, quantization, entropy coding, and reconstruction. The RD cost is defined [8] as

$$J(s, c, M | QP, \lambda_{MODE}) = SSD(s, c, M | QP) + \lambda_{MODE} \cdot R(s, c, M | QP) \quad (1)$$

where  $s$  and  $c$  are the source video signal and the reconstructed video signal, respectively.  $QP$  is the quantization parameter and  $\lambda_{MODE}$  is the Lagrange multiplier.  $SSD(s, c, M | QP)$  is the sum of the squared differences between  $s$  and  $c$ .  $M$  indicates a MB mode. In Eq. (1),  $R(s, c, M | QP)$  is the number of bits associated with the given  $M$  and  $QP$ . H.264/AVC encoder calculates the RD cost of every possible mode and chooses the best mode having the minimum RD cost.

H.264/AVC adopts the highest number of modes than any other video coding standards. For a P frame, a macroblock can be coded in the middle of the possible modes {SKIP,  $16 \times 16$ ,  $16 \times 8$ ,  $8 \times 16$ ,  $8 \times 8$ ,  $8 \times 4$ ,  $4 \times 8$ ,  $4 \times 4$ ,  $I4 \times 4$ ,  $I16 \times 16$ }. We, thus, classify those mode set into two categories, such as inter mode and intra mode as follows:

INTER MODE {SKIP,  $16 \times 16$ ,  $16 \times 8$ ,  $8 \times 16$ ,  $P8 \times 8$ }  
 INTRA MODE { $I16 \times 16$ ,  $I4 \times 4$ }

## III. PROPOSED METHOD

In the conventional scene change detection algorithms including [5], the sequence of sum of absolute difference (SAD) values between frames have been computed and used to detect scene changes. In addition, they have also considered statistical properties, such as mean value and standard deviation used to define a continuously updating automated threshold. In general, they make a decision for scene change when a high SAD value is observed between frames. However, as shown in Fig. 2, we can observe high SAD values during rapid movement, abrupt illumination changes and transition effects such as zoom in/out, fade in/out, dissolve etc. Moreover, we can have scene changes with very different value levels. A scene break, where both scenes have similar background, does not give a peak as high as if they had different ones. Consequently, a proper decision

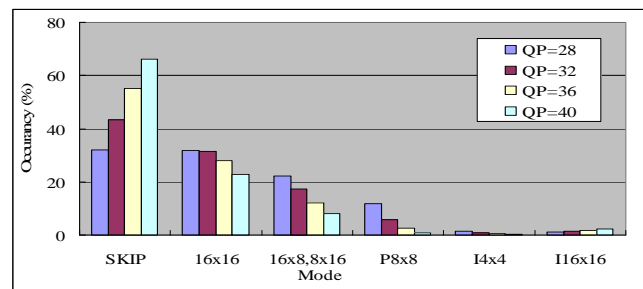


Figure 2. Various kinds of scenes from the test sequences

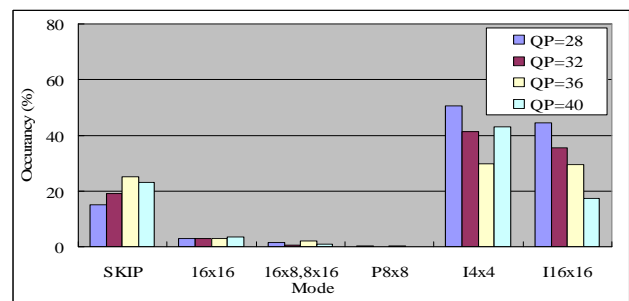
function is needed, which can take into account characteristics of the scene without a previous input.

For a new decision function, we utilize the RDO technique in H.264/AVC because the distribution of the best mode obtained from RDO effectively represents the relationship between temporal correlation and spatial correlation in each frame. Since temporal correlation is even higher than spatial correlation for a frame which does not belong to scene change. *INTRA MODE* rarely occurs (about 2%) in the inter frame as shown in Fig. 3. However, *INTRA MODE* frequently occurs where scenes change. These characteristics have already been investigated and verified in [7] and our previous research work [8].

Fig. 3 shows the best mode distribution according to the existence of scene changes. Therefore, using mode distribution between *INTRA MODE* and *INTER MODE*, we can efficiently find the exact time positions where the real scene changes occur.



(a) Mode distribution in non-scene change frame



(b) Mode distribution in scene change frame

Figure 3. The mode distribution

### A. Propocessing

First of all, the proposed system with video decoders using FFmpeg software [9] decodes input video data into frame sequences as shown in Fig. 4. It then normalizes temporally decoded sequences and spatially resized frames to detect scene change in the same circumstances (Frame size: 320x240, Frame rate: 10fps). Temporal normalization reduces processing time and diminishes effect of continuous scene change. Furthermore, spatial normalization decreases computational complexity to detect scene change.

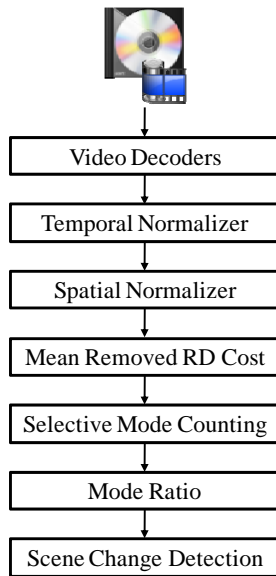


Figure 4. Block Diagram of Proposed Method

### B. Mean Removed RD Cost

The severe illumination changes seem to be scene-cuts and increase the number of false hits. In order to design robust scene change detection scheme for illumination change, we have modified the RD function by using mean removed sum of squared difference as follows

$$MRSSD = \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \{(x_{ij} - m_x) - (y_{ij} - m_y)\}^2 \quad (2)$$

where  $x_{ij}$  and  $y_{ij}$  represent the pixel intensity of original block and motion compensated block, respectively. In Eq. (2),  $m_x$  and  $m_y$  are the average pixel intensity in each block.

$$m_x = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N x_{ij}, \quad m_y = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N y_{ij} \quad (3)$$

Therefore, we have changed the RD cost function using Eq. (1) and (2) as

$$J(s, c, M | QP, \lambda_{MODE}) = MRSSD(s, c, M | QP) + \lambda_{MODE} \cdot R(s, c, M | QP) \quad (4)$$

### C. Selective Mode Counting (SMC)

For the scene change detection between two similar background shots or simple background shots as shown in Fig. 2, we use a selective mode counting (SMC) technique. It does not consider the mode information of macroblocks; the regions of which are homogenous or are located in the boundary area. The homogeneity of a macroblock is checked by RD cost as follows

$$J(s, c, M | QP, \lambda_{MODE}) < \overline{J_P(s, c, SKIP | QP, \lambda_{MODE})} \quad (5)$$

where  $\overline{J_P(s, c, SKIP | QP, \lambda_{MODE})}$  represents the average RD cost for SKIP mode in a previous frame. The boundary area in each macroblock is simply regarded as lines located in the most right, left, top, and bottom of each frame.

### D. Mode Ratio (MR) and Scene Change Detection

Finally, we can obtain the mode ratio (MR) from Eq. (5) and determine that a scene change occurs if MR is larger than the given threshold (TH). For simplicity, we have fixed the TH into 60 in our experiments.

$$MR = \frac{Count\_INTRA}{Count\_INTER + Count\_INTRA} \times 100 > TH \quad (6)$$

In Eq. (6),  $Count\_INTRA$  and  $Count\_INTER$  are the number of valid intra modes and inter modes, respectively.

## IV. EXPERIMENTAL RESULTS

In this section, we validate the proposed scheme for detecting various scene changes. Simulations were carried out using H.264/AVC reference software JM 12.4 [10] and coding parameters used are shown in Table 1. However, the proposed system does not encode frame sequences into bit streams for H.264/AVC, but utilizes only the information of inter and intra modes.

TABLE 1 SIMULATION CONDITIONS

Reference Software	JM 12.4 [10]
Profile	Baseline
RDO Mode	Fast High Complexity Mode
GOP Structure	I P P P . . .
Reference Frames	2
Search Range	$\pm 16$
FME	UMHexagonS

We used video sequences of music video and commercial advertisement contents as shown in Fig. 2. They were chosen because they have scenes with intense motion, change of light conditions, high complexity and different types of scene changes.

Its resolution is 320x240 and its length is 2,500 frames. The number of true scene changes in this sequence was 70.

We obtained them ‘manually’ by watching the video and counting them.

The performance of the proposed method is evaluated by comparing with other methods and the ground truth. For this reason, the “recall” and “precision” ratios are defined as follows

$$R = \frac{N_c}{N_c + N_M} \times 100 (\%) \quad (7)$$

$$P = \frac{N_c}{N_c + N_F} \times 100 (\%) \quad (8)$$

where  $N_C$ ,  $N_F$ , and  $N_M$  are the number of correct detections, the number of false ones, and the number of missed ones, respectively.

Table 2 shows the results of the various scene change detection algorithm. For the first two cases, after having the SAD values for the whole sequence, the fixed threshold and Dynamic threshold in [5] were chosen optimally to minimize the number of missed and false detections. For the last two cases, RD function in H.264/AVC and our proposed schemes are used, respectively.

TABLE 2 PERFORMANCE OF PROPOSED METHOD

	$N_C$	$N_F$	$N_M$	$R(\%)$	$P(\%)$
Fixed Threshold	56	29	14	65.88	80.00
Adaptive Threshold	60	17	10	77.92	85.71
H.264/AVC RD	60	9	2	96.77	86.95
Proposed Method	64	5	1	98.46	92.75

In Fig. 5, we also represent the distribution of SAD and MR for each frame from the experimental results. We can verify that the mode type classification technique using RD function is more appropriate to detect various scene changes than the conventional approach using SAD function. We also confirmed that our proposed method using MRSSD, SMC and MR enhances the recall and precision ratio about 2% and 6% compared with RDO technique in H.264/AVC, respectively.

## V. CONCLUSION

In this paper, we presented a reliable detection method for various abrupt and continuous scene changes through an analysis of the mode distribution between intra modes and inter modes in each frame. In order to enhance scene change detection ratio, we have adopted mean removed sum of

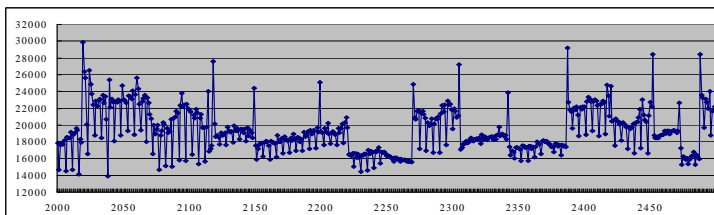
squared difference (MRSSD), selective mode counting (SMC) and mode ratio (MR) schemes. Based on these schemes, the proposed scene change detection technique works better than others in detecting dissolves with low variance frames, and decreases false hits induced by illumination change.

## ACKNOWLEDGEMENT

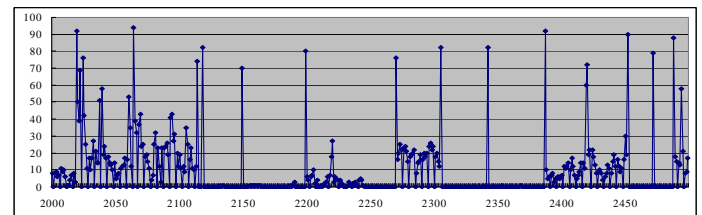
This research project was supported by the Government Fund from Korea Copyright Commission. [2012-cloud-9500 : Development of content-based usage control technology for clean cloud]

## REFERENCES

- [1] H. J. Zhang, A. Kankanhalli, and S. W. Smoliar, “Automatic Partitioning of Full-Motion Video,” *Multimedia Systems*, vol.1, no.1, pp. 10-28, 1993.
- [2] C. F. Lam and M. C. Lee, “Video Segmentation using Color Difference Histogram,” *Lecture Note in Computer Science 1464*, New York: Springer-Verlag, pp. 159-174, 1998.
- [3] W. J. Heng and K. N. Ngan, “High Accuracy Flashlight Scene Determination for Shot Boundary Detection,” *Signal Processing: Image Communication*, vol.18, no.3, pp. 203-219, Mar. 2003.
- [4] I. K. Sethi and N. Patel, “A Statistical Approach to Scene Change Detection,” *Proceedings of SPIE*, vol. 2420, pp. 329-338, Feb. 1995.
- [5] A. Dimou, O. Nemethova, and M. Rupp, “Scene Change Detection for H.264 using Dynamic Threshold technique,” *Proceedings of 5<sup>th</sup> EURASIP Conference on Speech and Image Processing, Multimedia Communications and Service*, Jun. 2005.
- [6] C. L. Huang and B. Y. Liao, “A Robust Scene-Change Detection Method for Video Segmentation,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol.11, no.12, pp. 1281-1288, Dec. 2001.
- [7] D. S. Turaga and T. Chen, “Estimation and Mode Decision for Spatially Correlated Motion Sequences,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol.11, no.10, pp. 1098-1107, Oct. 2001.
- [8] S. H. Kim and Y. S. Ho “A Fast Mode Decision Algorithm for H.264 using Statistics of the Rate Distortion Cost,” *Electronics Letters*, vol. 44, no. 14, pp. 849-851, 2008.
- [9] FFmpeg Revision 13712, available online at: <http://ffmpeg.mplayerhq.hu/>.
- [10] JVT reference software version 12.4, available online at: [http://iphome.hhi.de/suehring/tml/download/old\\_jm/](http://iphome.hhi.de/suehring/tml/download/old_jm/).



(a) Distribution of the sum of absolute difference between frames



(b) Distribution of the mode ratio in each frame

Figure 5. Distribution of SAD and MR

# Towards an Agile E-Commerce

Daniela Wolff, Marc Schaaf, Stella Gatzju Grivas  
 School of Business, Institute of Business Information Technology  
 University of Applied Sciences Northwestern Switzerland  
 Olten, Switzerland  
 {daniela.wolff, marc.schaaf, stella.gatzjugrivas}@fhnw.ch

**Abstract**—Creating an agile e-commerce is still a challenging issue. The alignment between business and IT plays a key role as changing business demands must be implemented immediately. To avoid misunderstandings and to lead to a better business-IT alignment we provide an ontology model, describing enterprise objects and their relations. As business rules guide or influence business behaviour, we use business rules which can be easily created and changed by business people. A system works directly with the ontology and the business rules. So, if changes occur, the business can express their changes and the IT system can react accordingly. As it easier for business people to express their knowledge and needs in a semi-formal way, we present a 3-phase procedure which helps to transform the semi-formal expressed knowledge into the formal representation needed for IT systems.

**Keywords**-context-awareness; rule based; complex event processing.

## I. INTRODUCTION

Continuously changing challenges, like shorter product cycles, increasing customer expectations, changing regulations, forces today's enterprises to be more agile [1][20]. Henbury regards agile enterprises as capable of rapid adaptations in response to unexpected and unpredicted changes and events, market opportunities and customer requirements [14]. As e-commerce becomes the preferred way of doing business [10], an enterprise must be able to adapt its e-commerce immediately when changes occur.

The adaptation of the e-commerce requires the

- 1) definition of the business model, i.e., knowledge about users, products and business rules
- 2) dynamic adaption of the business model according to happenings in the environment (events). The dynamic adaptation of this business model leads to the adaption of the Information Technology (IT) to match new business strategies, goals and needs, the so-called business-IT alignment.
- 3) personalization, for example the analysis of the user behavior. While the user is navigating through the web site his clickstream is observed and according to his interest and behaviour the web pages are personalized [3].

For the definition of the business model, a common approach is the use of adaptive hypermedia and adaptive Web systems. Adaptive software systems are based on a business model representing user knowledge, goals, interests and other features to distinguish among different users. The challenge for the

adaption of the business model is the use of different languages by different actors in the alignment process. For instance, IT managers can read and understand UML but such languages may not provide adequate information for business people [15].

As ontologies promote a common understanding among people [19], we present an ontology describing business objects and rules. This ontology is used as the knowledge base for the e-commerce and web site adaptation. If changes occur, the business user can express his changes in the ontology and the system uses the updated knowledge base. This approach supports enterprises, especially e-commerce, to be more agile and to be able to react to changing environments immediately.

This paper is structured as follows. First, we introduce a simplified scenario, which is used in this paper to show our approach. Then we describe the knowledge base. As business users can express their needs better in a semi-formal way we propose a method which enables users to express their needs using a structured template which can easily transformed into the formal representation. Finally, we show the benefit of the model-based approach for the business-IT alignment.

## II. SCENARIO

To explain our approach, we use a simplified scenario of a book store. The book store provides information about books, authors, and search functionality. A customer can register himself and can enter further information.

The store distinguishes between the four browsing strategies proposed by [18]: direct buying, search/deliberation, Hedonic browsing and knowledge building. A visitor using the direct buying strategy has a specific product in mind which he wants to buy. His browsing pattern is therefore very focused and targeted. Visitors using the search and deliberation strategy are also focused with a future purchase in mind. Their objective is to acquire relevant information to help make a better choice. The hedonic browsing is dominated by exploratory search behaviour and therefore more sessions are spent viewing the broader product category level pages than product information. The visitor using the knowledge building strategy is acquiring relevant product information potentially useful in the future. They tend to focus more on information pages.

The strategy of the book store could for example be to help the users following the direct buying strategy by the providing of relevant information related to the content they have already visited. For instance, if a user searches for a

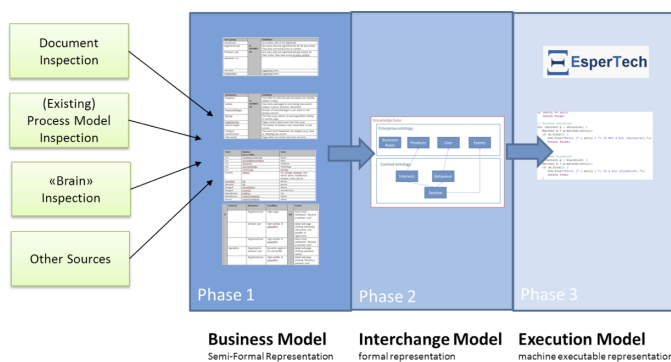


Figure 1. 3-Phase-Procedure.

specific criminal story set in Cologne from Frank Schätzing, the users navigates to the web page containing a list of all books of Frank Schätzing and then he starts looking for the criminal story. To reduce the list of results he enters into a search field Cologne. The System recognizes from his click path that he is interested in criminal stories of Frank Schätzing and Cologne. The system retrieves that Frank Schätzing has written the criminal story “Tod und Teufel” playing in Cologne and provide the user a link to this book. With this also, people who do have a specific book in mind can be inspired. The click stream is analysed and interpreted and related books are given as links on the web page. To stay agile, the book store does not want to fix business objects, like products, for instance in a database, and the link between these objects. Additionally, coding the business rules, expressing actions which should be triggered depending on specific conditions, like sending an E-Mail when a user gets often to restricted web pages, in java or other programming languages leads to inflexible enterprises as IT experts are required to implement changing rules. The bookstore must have the freedom to change the product catalogue according to the customers’ needs and also have the ability to change the user groups and the actions which should be triggered according to their behaviour and interests. Additionally, a business user should have the freedom to express changing business rules in an easy way.

In the next section, we present an ontology with which an enterprise can be described. To make it easier for business people, we propose a 3-phase-procedure where business people can express their knowledge in a semi-formal way, which can be transformed into a formal and afterwards if necessary into an executable form.

### III. PERSONALIZATION ONTOLOGY

In [8], we described a method for ontology development in the e-government field, which we have adapted for the e-commerce area. The method comprises four levels of formalisation: informal (knowledge captured in natural language), semi-formal (knowledge represented in a semi-formal way in structured templates), formal (knowledge formalized in OWL [5] and SWRL [6]) and executable form (knowledge formalized in e.g., Esper-Rules [9] and Java).

The method consists of three phases (Figure 1):

Phase 1 – Defining the business model: capturing user groups, product catalogue and business rules in a semi-formal way.

Phase 2 – Defining the interchange model: transform the semi-formal expressed terms, facts and rules into OWL and SWRL.

Phase 3 – Defining the execution model: transform the formal model into a machine executable form.

These phases are described in more detail in the next sections.

#### A. Business Model

Writing down the terms, facts and rules (all together called ‘business rules’) in a semi-formal way so that business people can easily understand them because they are close to ‘normal English’ and IT people can understand them as well as they are clearly structured. We use the templates provided by Barbara von Halle [23]. The first step is about defining the user groups. As an example, we use the browsing strategies introduced in the last section. Using the template a user group can be defined as follows:

Behavior	IS DEFINED AS	Definition
Search/ Deliberation		a strategy which intends to acquire relevant information to help make a more optimal choice.

Secondly, as the various user groups provide different navigation patterns, the measures for how to recognize the user groups have to be also specified. For instance to recognize a hedonic browser, the page types must be analysed. A person who uses the hedonic browsing strategy focuses on category pages. So, the number of category pages is very high. Additionally, he visits a lot of product pages. However, he does not repeat visiting a web page very often. So, first of all these page types must be defined.

Parameter	IS DEFINED AS	Definition
Product page		As a page describing a specific product

After expressing these parameters the conditions by which a user can be assigned to a specific user group can be defined. For the hedonic browsing strategy it might look like the following template.

IF	Condition	THEN	Consequence
	The focus of a session is on requesting category pages, the category variety is high and the product variety is high, repeat viewing is low		Hedonic browsing

Thirdly, to find out in which products a person is interested in, the product catalogue must be described. With this information the user behaviour and his interests are expressed. For this the IS DEFINED AS template can be used. As products are related somehow to other produces another template is necessary expressing those relations.

Term	Relation (IS A / VERB)	Term
a criminal book	Is A	book

As actions should be triggered if a user is interested in a specific product and shows a specific behaviour, these actions must be defined. For this reason we use also the templates. We reuse the definition templates to define the action.

IF	Interest	Behavior	THEN	Actions
	Book, Author and other specific attributes	Direct buying		Show list of related books, related to the author and the other specific attributes

### B. Interchange Model

The second phase of designing the personalisation ontology is focussed on the transformation of the models into a precise, machine understandable form. The purpose of this formal model derived in this phase is twofold: First, the semi-formal representation chosen in the first step can be easily understood by business people but has the disadvantage that it cannot be executed by a computer because the rules can be ambiguous. In order to be validated and executed, the user groups, product catalogue and the business rules have to be represented in a language with well- defined semantics. Second, there can be different run-time environments for the execution of business processes. The interchange format shall serve as a common language from which the execution formats can be derived unambiguously, if possible even automatic.

To fulfil these purposes the interchange format must have a clear and precise semantics. The enterprise objects and the business rules are represented in OWL and SWRL. Because of the partially ambiguous business models the transformation is not automated: the business objects (user groups, product catalogue) have to be transferred into OWL and the rules into SWRL manually. However, the development of a semantic representation from the semi- structured representation of the business models is straightforward. The interchange model consists of two main ontologies (Figure 2): the enterprise ontology and the context ontology. The enterprise ontology describes the enterprise itself, the product catalogue, the user groups, the actions. The context ontology provides information about the users’ session, like navigation path, historical events, current behaviour and interest.

This interchange model is used as the knowledge base to retrieve information about the user and his interests and to provide him with relevant information. For instance, if the system notices that a visitor has visited web pages about Frank Schätzing, criminal story and Cologne, he can retrieve from the database through the relations between the different topics, that the visitor might be interested in the book “Tod und Teufel”. If another person visits the web page about “Tod und Teufel” and “Mordshunger” the system can assume that the person is interested about criminal stories from Frank Schätzing and can provide him with a proper list.

1) *Enterprise Ontology*: Enterprise ontologies have been developed also with the intention “to assist the acquisition, representation, and manipulation of enterprise knowledge;

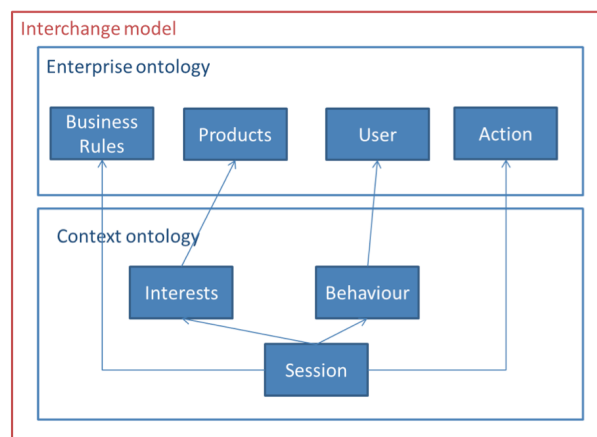


Figure 2. Interchange model. This figure shows the main concepts of the ontologies.

structuring and organizing libraries of knowledge [...]” [21]. Usually, enterprise ontologies are created to define and organize relevant enterprise knowledge like processes, organization structure and strategy [13].

Ushold et al. designed a particular ontology, the “Enterprise Ontology”, which aims to provide “a collection of terms and definitions relevant to business enterprise to enable coping with a fast changing environment [...]” [21]. The TOVE (Toronto Virtual Enterprise) project is being carried out by the Enterprise Integration Laboratory (EIL) at the University of Toronto. It provides a generic, reusable knowledge model providing a shared terminology for the enterprise. While both ontologies focus on business processes, there are common semantic concepts in both projects [11].

We use the enterprise ontology as a basis, but extend it by some sub classes. As the business rules implement the business strategy, trigger various events and are one of the main drivers of an enterprise the business rules must be made explicit. This allows changing the rules immediately. Therefore, we add a concept business rules to the strategy part of the enterprise ontology. The rules themselves are expressed using SWRL, which combines OWL and RuleML [12].

Another important part of the enterprise are the products. Semantically enriched and precise product information can enhance the offering of information. Product information consists of product properties and the relationship between products. For the description of the product catalogue existing domain ontologies, like wine or pizza, can be taken. A meta level for a product ontology can be found in [16]. This meta level helps to create a product ontology for each enterprise. Figure 3 illustrates a simplified ontology describing books for our book store scenario.

To express the various behaviour patterns of a user, we use the organisation part of the enterprise ontology and added user to the existing concept “Stakeholder”. For the analysis of the behaviour, in particular the shopping strategies, we rely on the different shopping strategies as proposed by [18]. They can be recognized by using different browsing patterns. To find the different page types Moe categorizes pages as category pages, product pages, home page or information pages. During



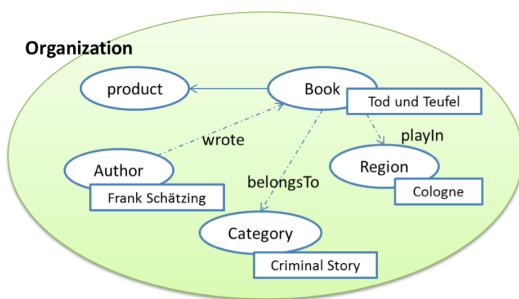


Figure 3. Simplified product description.

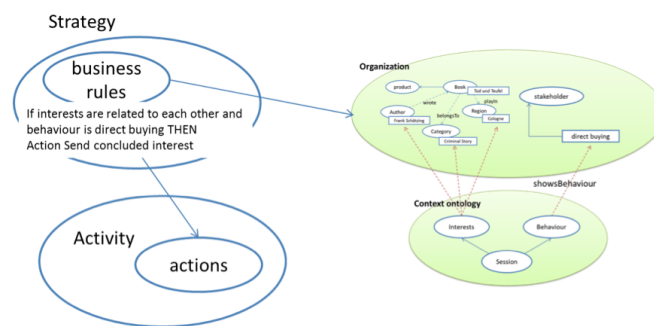


Figure 5. Business rules combining context and actions.

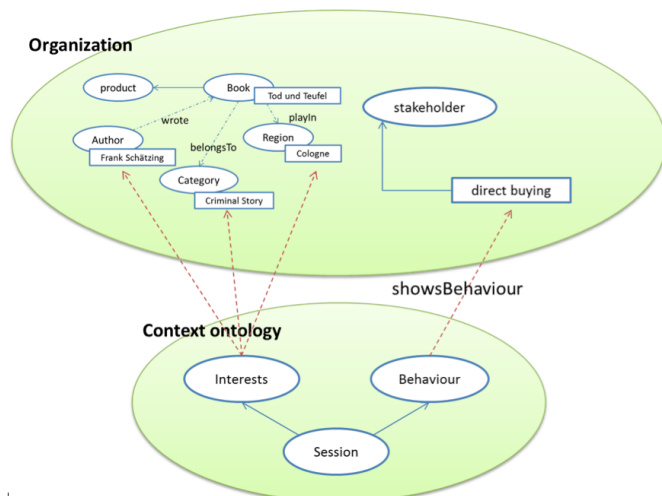


Figure 4. Relation context and organization ontology.

the user’s visit the percentage is identified, how often he/she visits an information page, category page or product page. If the user visits informational pages more often, the shopping strategy is more likely to be a knowledge building. Whereas, a visitor who visits a lot of category pages, seems to follow the hedonic browsing strategy. These user categories are added to the enterprise ontology as sub classes of user.

There exist two main parts of adaptation: internal adaptation, and external adaptation. Internal adaptation supports the usability. Usability can be associated with the aspects: content, ease of use, promotion, made-for-the-medium, and emotion [2]. [4] identified the following internal adaptations: adaptation of content and services delivered by accessed pages, adaptation of navigation, adaptation of whole hypertext structure and adaptation of presentation properties. External adaptation means the context is used to adapt external applications, like newsletter or e-mail services. For the external adaptation, we use web services which are described using OWL-S [17].

2) *Context Ontology*: The behaviour and the user’s interests can be analyzed while a visitor is navigating through a web site. This represents the situation of a visitor. According to Dey and Abowd [7], context is all information, which can be used to characterize the situation of an entity. Therefore, we use a context ontology which helps to interpret the users’ current situations. This ontology consists of three main concepts: session, interest and behaviour.

As shown in Figure 4, the concepts *interests* and *behaviour*

are linked to the organization ontology: the behaviour is related to the *Stakeholder*-concept and the *interests* concept relates to the product concept. During a session the links are continuously updated.

The ontology is split into two parts to distinguish between the static and dynamic parts. Whereas, the information provided in the enterprise ontology is more or less static, the context ontology provides dynamic and user specific information.

As business rules trigger events when a specific condition is met, rules combine both the context ontology and the enterprise ontology. So, rules use the whole knowledgebase.

We use rules to combine the context and the (re)actions. On the condition part of the rules the context is defined. This context is analyzed during run time by a rule engine. If a specific context is kept the rules trigger the appropriate actions. Figure 5 shows a simplified rule, combining the context with a service. In this case the rule expresses that if a visitor uses the direct buying strategy and is interested in a specific book, the visitor should get a list of concluded interests (in our case the book “Tod und Teufel”).

If changes in the enterprise environment occur the business user is able to adapt the product catalogue and the business rules immediately (directly in the ontology or using the intermediate step of the semi-formal language).

### C. Execution Model

In phase 3, parts of the interchange model created in phase 2 will be migrated into machine executable forms if necessary. As already described the interchange model can be used as a knowledge base, because the model helps to recognise the users’ behaviour and interests and a system can retrieve relevant information in the knowledge base.

However, some parts of the model must be transformed into another executable representation. The interchange model contains information specified by the business people. But, it does not contain information about technical details, like cookie or session ids, what actions a user performed or which agent a user uses. This information must be expressed in a machine executable form.

In our system, we consider each request (action) that is sent by the visitor’s browser to the content management system as an event. Each of those events is linked to a certain visitor and contains annotations that describe the requested content or the

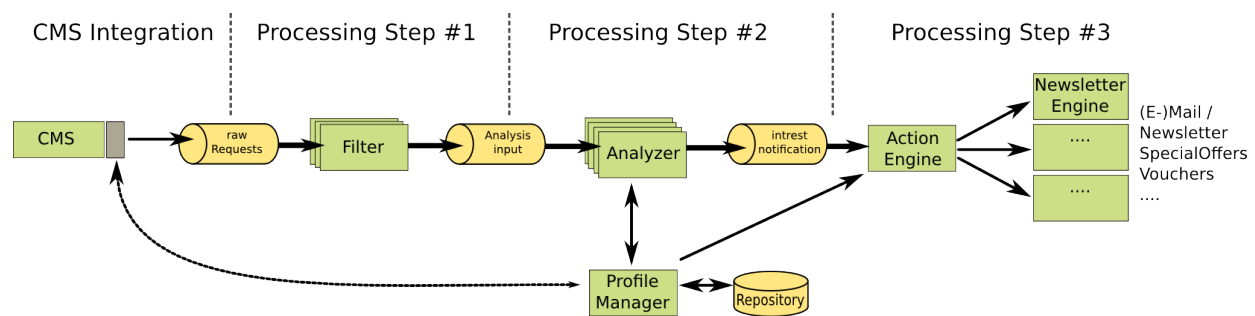


Figure 6. The used processing system with its components and the different processing steps.

requested action. The content annotation is stored in the CMS as meta information for the available content and is based on the concepts of the enterprise ontology. An event thus has the following simplified structure:

UserCookieID	e.g. 156978GH
PageTags	e.g. author:FrankSchaetzing
PageType	e.g. productDetail
RequestAction	e.g. addToShoppingCart
UserAgent	e.g. Firefox
SessionID	e.g. S47111147

These events are processed in three steps (Figure 6) where the first step filters the requests to remove requests that are for example generated by search engine bots. In the second step the requests are analyzed with the help of a set of rules that extend the context ontology of the corresponding user as needed. For example, for the event shown above, a rule would detect the adding of the particular wine to the shopping cart and would thus add a concept to the context ontology that represents the interest of this user in the given wine. The third processing step allows the rule based triggering of external action like sending an email to a customer. However, this part will not be discussed in further detail in this paper.

#### IV. STATUS OF THE IMPLEMENTATION

We use the presented combination of semantic technologies with event processing systems in a dynamic website personalization engine which we are currently developing in a joint research project with the Wyona AG ([www.wyona.com](http://www.wyona.com)). The aim of this engine is to build up a context ontology for a website visitor while he/she is browsing through a website like for example an online shop. The generated context ontology is used to support the user in his search for products that suite his interests and to generate personalized advertisement campaigns.

Due to this outlined process, a context ontology is built for each user while she/he is still browsing through the website. To allow the CMS to utilize the gathered information, a profile manager provides a simple query interface. The CMS can use this interface to retrieve related content to the current users' request. For such a query, the CMS specifies the current visitors ID together with the tags that annotate the page that the user currently requests. The profile manager uses this information to deduct tags from the context ontology together with the enterprise ontology as discussed in Section

3.2.2. The results are handed back to the CMS which in turn uses those tags to select content that might be interesting for the current visitor. Thus, our current realization approach follows the concept of event-driven architectures to realize a rapid processing of the visitors requests. With the help of the aforementioned process for the ontology and rule definition based on a simple table based schema, we aim to allow non-IT specialists to change and optimize the behavior of the fairly complex processing system.

#### V. CONCLUSION

As e-commerce becomes the preferred way of doing business an enterprise must be able to adapt their e-commerce immediately when changes occur. We provide an ontology model which enables business users to express their knowledge about their products, user groups, actions and business rules. If changes occur the business user can adapt the business rules and the business objects immediately (directly in the ontology or using the intermediate step of the semi-formal language).

The 3-phase procedure is currently done manually however we are going to develop a system, with which an automatic transformation is possible. We intend to use the resulting descriptions as the knowledge base for a self adapting web shop system to verify the usefulness of the generated information. Furthermore it is planned to evaluate the usability of the presented concepts together with business users.

#### VI. ACKNOWLEDGEMENTS

This project was funded by the CTI under the project number 11628.1 PFES-ES (Semantic OBDE).

#### REFERENCES

- [1] T. Allweyer, "Die Organisation des Wissens, Wie japanische Unternehmen eine brachliegende Ressource nutzbar machen." Witten, W3L GmbH, 2. Nachdruck, 2007.
- [2] R. Agarwal and V. Venkatesh, "Assessing a Firm's Web Presence: A Heuristic Evaluation Procedure for the Measurement of Usability", *Information Systems Research*, vol. 13, No. 2, 2002, pp. 168 - 186.
- [3] P. Brusilovsky and M. T. Maybury, "From Adaptive Hypermedia to the Adaptive Web", In *Communications of the ACM*, 45(5), 2002, pp. 31-33.
- [4] S. Ceri, F. Daniel, M. Matera, and F. M. Facca, "Model-driven development of context-aware Web applications", *ACM Trans. Internet Technology*, 7(2), 2007.
- [5] F. van Harmelen and D. L. McGuinness, Editors, "OWL Web Ontology Language Overview", W3C Recommendation, Available at: <http://www.w3.org/TR/owl-features/>, 2004, Retrieved: 26.08.2012.

- [6] I. Horrocks, P. F. Patel-Scheider, H. Boley, S. Tabet, B. Grosz, and M. Dean, "SWRL: A Semantic Web Rule Language Combining OWL and RuleML", W3C Member Submission, Available at: <http://www.w3.org/Submission/SWRL/>, 2004, Retrieved: 26.08.2012.
- [7] A. K. Dey and G. Abowd, "Towards a Better Understanding of Context and Context-Awareness", In Proceedings of CHI2000: Conference on Human Factors in Computing, The Hague, The Netherlands, 2000.
- [8] D. Feldkamp, K. Hinkelmann, and B. Thönssen, "Kiss-Knowledge-Intensive Service Support: An Approach for Agile Process Management", Proceedings of RuleML 2007, 2007, pp. 25-38.
- [9] "Esper - Complex Event Processing", Product Web Page, Available at: <http://esper.codehaus.org>, Retrieved: 26.08.2012.
- [10] P. Fingar, "Component-based Frameworks for E-Commerce", Communications of the ACM, Vol. 43(10), 2000, pp. 61-66.
- [11] C. Sechlenoff, R. Ivester, and A. Knutilla, "A Robust Process Ontology For Manufacturing Systems Integration", In Proceedings of 2nd International Conference on Engineering Design and Automation, Maui, 1998.
- [12] "The Rule Markup Initiative", Available at: <http://ruleml.org>, Retrieved: 26.08.2012.
- [13] A. Gomez-Perez, M. Fernandez-Lopez, and O. Corcho, "Ontological Engineering, with examples from the areas of knowledge management, e-commerce and the Semantic", Springer, London, 2004.
- [14] C. Henbury, "Two definitions of agility", Available at: <http://www.cheshirehenbury.com/agility/twodefinitions.html>, 2006, Retrieved 26.08.2012.
- [15] V. Hrgovic, W. Utz, and R. Woitsch, "Knowledge Engineering in Future Internet", In D. Karagiannis, Z. Jin (Eds.) Knowledge Science, Engineering and Management, Springer, Vienna, 2009.
- [16] T. Lee, I. Lee, S. Lee, S. Lee, and D. Kim, "Building an operational product ontology system", Electronic Commerce Research and Applications 5, 2006, pp. 16-28.
- [17] D. Martin, M. Burstein, J. Hobbs, O. Lassila, et al., "OWL-S: Semantic Markup for Web Services", W3C Member Submission, Available at: <http://www.w3.org/Submission/OWL-S/>, 2004, Retrieved: 26.08.2012.
- [18] W. M. Moe, "Buying, Searching, or Browsing: Differentiating Between Online Shoppers Using In-Store Navigational Clickstream", Journal of Consumer Psychology, 13(1&2), 2003, pp. 29-39.
- [19] N. F. Noy and D.L. McGuinness, "Ontology Development 101: A Guide to Creating Your First Ontology", Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001- 0880, 2001.
- [20] A. W. Scheer, O. Adam, A. Hofer, and F. Zangl, "Nach Cost Cutting, Aufbruch durch Innovation", In IM Fachzeitschrift für Information Management and Consulting, vol. 18, Sonderausgabe, 2003.
- [21] M. Ushold, M. King, S. Moralee, and Y. Zorgios, "The Enterprise Ontology", The Knowledge Engineering Review, Vol. 13, Special Issue on Putting Ontologies to Use, 1998.
- [22] R. Woitsch, D. Karagiannis, D. Plexousakis, and K. Hinkelmann, "Plug your Business into IT: Business and IT-Alignment using a Model-Based IT-Socket", eChallenges e-2009 Conference (eChallenges 09), Istanbul, IOS Press, 2009.
- [23] B. Von Halle, "Business Rules Applied, Building Better Systems Using the Business Rules Approach", Wiley and Sons, New York, 2002.

# Enabling High Performance Computing for Semantic Web Applications by Means of Open MPI Java Bindings

Alexey Cheptsov

*High Performance Computing Center Stuttgart (HLRS)*

*University of Stuttgart*

*70550 Stuttgart, Germany*

*Email: cheptsov@hlrs.de*

**Abstract**—The volume of data available in the Semantic Web has already reached the order of magnitude of billions of triples and is expected to further grow in the future. The availability of such an amount of data makes it attractive for Semantic Web applications to exploit High Performance Computing (HPC) infrastructures to effectively process such data. Unfortunately, most Semantic Web applications are written in the Java programming language, whereas current frameworks that make the most out of HPC infrastructures, such as the Message Passing Interface (MPI), only target C or Fortran applications. Attempts to port existing parallelization frameworks to the Java language prove to be very inefficient in terms of the performance benefits for applications. This paper presents an efficient porting based on the Open MPI framework.

**Keywords**—*High Performance Computing; Semantic Web; Data-Centric Computing; Performance; Scalability; Message-Passing Interface.*

## I. INTRODUCTION

The volume of data collected on the Semantic Web has already reached the order of magnitude of billions of triples and is expected to further grow in the future, which positions this Web extension to dominate the data-centric computing in the oncoming decade. Processing (e.g., inferring) such volume of data, such as generated in the social networks like Facebook or Twitter, or collected in domain-oriented knowledge bases like pharmacological data integration platform OpenPHACTS [1], is thus of a big challenge. Whereas there is a number of existing highly-scalable software solutions for storing data, such as Jena [2], the scalable data processing constitutes the major challenge for data-centric applications. The group of issues related to scaling the existing data processing techniques to the available volumes is often referred as the “Big Data” problem. Among those data-centric communities that address the Big Data, the Semantic Web enjoys a prominent position.

Semantic Data are massively produced and published at the speed that makes traditional processing techniques (such as reasoning) inefficient when applied to the real-scale data. The data scaling problem in the Semantic Web is considered in two its main aspects - horizontal and vertical scale. Horizontal scaling means dealing with diverse, and often

unstructured data acquired from heterogeneous sources. The famous LOD cloud diagram [3] consists of hundreds of diverse data sources, ranging from geospatial cartographic sources to governmental data, opened to the publicity, like Open Government Data [4]. Vertical scaling implies scaling up the size of similarly structured data. Along the open government data spawns over 851,000 data sets across 153 catalogues from more than 30 countries, as estimated in [5] at the beginning of 2012. Processing data in such an amount is not straightforward and challenging for any of the currently existing frameworks and infrastructures. Whereas there are some known algorithms dealing with the horizontal scaling complexity, such as identification of the information subsets related to a specific problem, i.e., subsetting, the vertical scaling remains the major challenge for all existing algorithms. Another essential property of the Big Data is the complexity. Semantic applications must deal with rich ontological models describing complex domain knowledge, and at the same time highly dynamic data representing recent or relevant information, as produced by streaming or search-enabled data sources. A considerable part of the web data is produced as a result of automatic reasoning over streaming information from sensors, social networks, and other sources, which are highly unstructured, inconsistent, noisy and incomplete.

The availability of such an amount of complex data makes it attractive for Semantic Web applications to exploit High Performance Computing (HPC) infrastructures to effectively process Big Data. As a reaction on this challenge, a number of major software vendors in the Semantic Web domain have been collaborating with high performance computing centers, and this trend is expected to grow in the nearest future [6]. Both commodity and more dedicated HPC architectures, such as the Cray XMT [7], have been in focus of the data-intensive Web applications. The XMT dedicated system, however, proved successful only for a limited number of tasks so far, which is mainly due to the complexity of exploiting the offered software frameworks (mainly non-standard pragma-based C extensions. Unfortunately, most Semantic Web applications are written in the Java programming language, whereas current frameworks

that make the most out of HPC infrastructures, such as the Message Passing Interface (MPI), only target C or Fortran applications. MPI is a process-based parallelization strategy, which is a de-facto standard in the area of parallel computing for C, C++, and Fortran applications. Known alternative parallelization frameworks to MPI that conform with Java, such as Hadoop [8] or Ibis [9], prove to be scalable though but are not even nearly as efficient or well-developed as numerous open-source implementations of MPI, such as MPICH or Open MPI [10]. We look at how to resolve the above-mentioned issues in a way that leverages the advances of the existing MPI frameworks.

The remainder of the paper is organized as follows. Section 2 gives an overview of the related work. Section 3 discusses the data-centric parallelization model based on MPI. Section 4 introduces our implementation of Java bindings for Open MPI. Section 5 gives examples of successful pilot scenarios implemented with our solution and discuss future work in terms of the development, implementation, and standardization activities.

## II. RELATED WORK

There are only a few alternatives to MPI in introducing the large-scale parallelism to Java applications. The most promising among those alternatives in terms of the performance and usability are solutions offered by IBIS/JavaGAT and MapReduce/Hadoop.

IBIS [11] is a middleware stack used for running Java applications in distributed and heterogeneous computing environments. IBIS leverages the peer-to-peer communication technology by means of the proprietary Java RMI (Remote Memory Invocation) implementation, based on GAT (Grid Application Toolkit) [12]. The Java realization of GAT (JavaGAT) is a middleware stack that allows the Java application to instantiate its classes remotely on the network-connected resource, i.e., a remote Java Virtual Machine. Along with the traditional access protocols, e.g., telnet or ssh, the advanced access protocols, such as ssh-pbs for clusters with PBS (cluster Portable Batch System)-like job scheduling or gsissh for grid infrastructures are supported. IBIS implements a mechanism of multiple fork-joins to detect and decompose the application's workload and execute its parts concurrently on distributed machines. While [9] indicates some successful Java applications implemented with IBIS/JavaGAT and shows a good performance, there is no clear evidence about the scalability of this solution for more complex communication patterns, involving nested loops or multiple split-joins. Whereas IBIS is a very effective solution for the distributed computing environments, e.g., Grid or Cloud, it is definitively not the best approach to be utilized on the tightly-coupled productional clusters.

MapReduce framework [8] and its most prominent implementation in Java, Hadoop, has got a tremendous popularity in modern data-intensive application scenarios. MapReduce

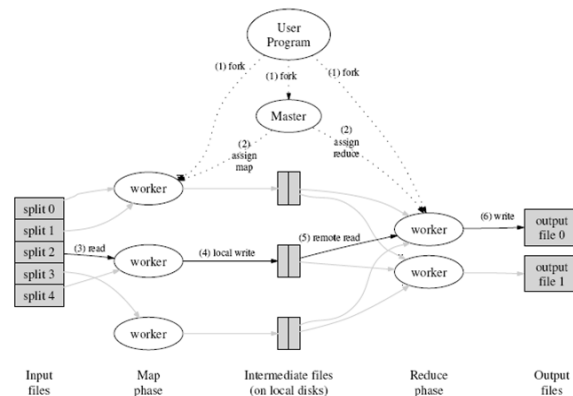


Figure 1. MapReduce processing schema

is a programming model for data-centric applications exploiting large-scale parallelism, originally introduced by Google in its search engine. In MapReduce, the application's workflow is divided into three main stages (see Figure 1): map, process, and reduce. In the map stage, the input data set is split into independent chunks and each of the chunks is assigned to independent tasks, which are then processed in a completely parallel manner (process stage). In the reduce stage, the output produced by every map task is collected, combined and the consolidated final output is then produced.

The Hadoop framework is a service-based implementation of MapReduce for Java. Hadoop considers a parallel system as a set of master and slave nodes, deploying on them services for scheduling tasks as jobs (Job Tracker), monitoring the jobs (Task Tracker), managing the input and output data (Data Node), re-executing the failed tasks, etc. This is done in a way that ensures a very high service reliability and fault tolerance properties of the parallel execution. In Hadoop, both the input and the output of the job are stored in a special distributed file-system. In order to improve the reliability, the file system also provides an automatic replication procedure, which however introduces an additional overhead to the inter-node communication. Due to this overhead, Hadoop provides much poorer performance than MPI, however offering better QoS characteristics related to the reliability and fault-tolerance. Since MPI and MapReduce paradigms have been designed to serve different purposes, it is hardly possible to comprehensively compare them. However they would obviously benefit from a cross-fertilization. As a possible scenario, MPI could serve a high-performance communication layer to Hadoop, which might help improve the performance by omitting the disk I/O usage for distributing the map and gathering the reduce tasks across the compute nodes.

### III. DATA-CENTRIC PARALLELIZATION AND MPI

By “data-centric parallelization” we mean a set of techniques for: (i) identification of non-overlapping application’s dataflow regions and corresponding to them instructions; (ii) partitioning the data into subsets; and (iii) parallel processing of those subsets on the resources of the high performance computing system. For Semantic Web applications utilizing the data in such well-established formats as RDF [13], parallelization relies mainly on partitioning (decomposing) the RDF data set on the level of statements (triples), see Figure 2a. The ontology data (also often referred as *tbox*) usually remains unpartitioned as its size is relatively small as compared with the actual data (*abox*), so that it is just replicated among all the compute nodes.

The Message-Passing Interface (MPI) is a process-based standard for parallel applications implementation. MPI processes are independent execution units that contain their own state information, use their own address spaces, and only interact with each other via interprocess communication mechanisms defined by MPI. Each MPI process can be executed on a dedicated compute node of the high performance architecture, i.e., without competing with the other processes in accessing the hardware, such as CPU and RAM, thus improving the application performance and achieving the algorithm speed-up. In case of the shared file system, such as Lustre [14], which is the most utilized file system standard of the modern HPC infrastructures, the MPI processes can effectively access the same file section in parallel without any considerable disk I/O bandwidth degradation. With regard to the data decomposition strategy presented in Figure 2a, each MPI process is responsible for processing the data partition assigned to it proportionally to the total number of the MPI processes (see Figure 2b). The position of any MPI process within the group of processes involved in the execution is identified by an integer  $R$  (rank) between 0 and  $N-1$ , where  $N$  is a total number of the launched MPI processes. The rank  $R$  is a unique integer identifier assigned incrementally and sequentially by the MPI runtime environment to every process. Both the MPI process’s rank and the total number of the MPI processes can be acquired from within the application by using MPI standard functions, such as presented in Listing 1. The typical data processing workflow with MPI can be depicted as shown in Figure 3. The MPI jobs are executed by means of the *mpirun* command, which is an important part of any MPI implementation. *mpirun* controls several aspect of parallel program execution, in particular launches MPI processes under the job scheduling manager software like OpenPBS [15]. The number of MPI processes to be started is provided with the *-np* parameter to *mpirun*. Normally, the number of MPI processes corresponds to the number of the compute nodes, reserved for the execution of parallel job. Once the MPI process is started, it can request its rank as well as the

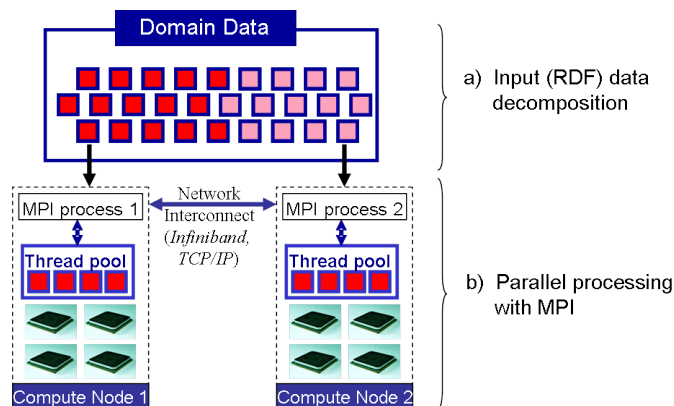


Figure 2. Data decomposition and parallel execution with MPI.

total number of the MPI processes associated with the same job. Based on the rank and total processes number, each MPI process can calculate the corresponding subset of the input data and process it. The data partitioning problem remains beyond the scope of this work; particularly for RDF, there is a number of well-established approaches discussed in several previous publications, e.g., horizontal [16], vertical [17], and workload driven [18] partitioning.

Since a single MPI process owns its own memory space and thus can not access the data of the other processes directly, the MPI standard foresees special communication functions, which are necessary, e.g., for exchanging the data subdomain’s boundary values or consolidating the final output from the partial results produced by each of the processes. The MPI processes communicate with each other by sending messages, which can be done either in “point-to-point” (between two processes) or collective way (involving a group of or all processes).

```

import java.io.*;
import mpi.*;

class Hello {
    public static void main(String[] args) throws
        MPIException
    {
        int my_pe, npes; // rank and overall number of MPI
            processes
        int N; // size of the RDF data set (number of
            triples)

        MPI.Init(args); // initialization of the MPI RTE

        my_pe = MPI.COMM_WORLD.Rank();
        npes = MPI.COMM_WORLD.Size();

        System.out.println("Hello_from_MPI_process" + my_pe +
            "_out_of_" + npes);
        System.out.println("I'm_processing_the_RDF_triples_
            from_" + my_pe/npes + "_to_" + (my_pe+1)/npes);

        MPI.Finalize(); // finalization of the MPI RTE
    }
}

```

Listing 1. Acquiring rank and total number of processes in a simple MPI application

More details about the MPI communication can also be

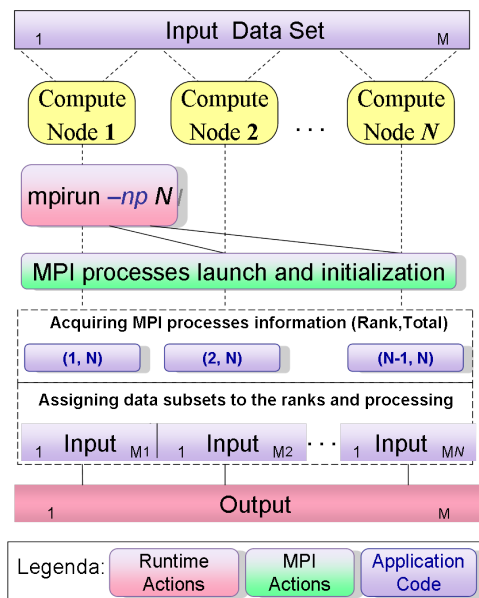


Figure 3. Typical MPI data-centric application's execution workflow

found in our previous publication [19].

#### IV. OPEN MPI JAVA BINDINGS

##### A. MPI bindings for Java

Although the official MPI standard's bindings are limited to C and Fortran languages, there has been a number of standardization efforts made towards introducing the MPI bindings for Java. The most complete API set, however, has been proposed by mpiJava [20] developers.

There are only a few approaches to implement MPI bindings for Java. These approaches can be classified in two following categories:

- Pure Java implementations, e.g., based on RMI (Remote Method Invocation) [21], which allows Java objects residing in different virtual machines to communicate with each other, or lower-level Java sockets API.
- Wrapped implementations using the native methods implemented in C languages, which are presumably more efficient in terms of performance than the code managed by the Java run-time environment.

In practice, none of the above-mentioned approaches satisfies the contradictory requirements of the Web users on application portability and efficiency. Whereas the pure Java implementations, such as MPJ Express [22] or MPJ/Ibis [9], do not benefit from the high speed interconnects, e.g., InfiniBand [23], and thus introduce communication bottlenecks and do not demonstrate acceptable performance on the majority of today's production HPC systems [24], a wrapped implementation, such as mpiJava [25], requires a native C library, which can cause additional integration and interoperability issues with the underlying MPI implementation.

In looking for a trade-off between the performance and the usability, and also in view of the complexity of providing Java support for high speed cluster interconnects, the most promising solution seems to be to implement the Java bindings directly in a native MPI implementation in C.

##### B. Native C Implementation

Despite a great variety of the native MPI implementations, there are only a few of them that address the requirements of Java parallel applications on process control, resource management, latency awareness and management, and fault tolerance. Among the known sustainable open-source implementations, we identified Open MPI [26] and MPICH2 [27] as the most suitable to our goals to implement the Java MPI bindings. Both Open MPI and MPICH2 are open-source, production quality, and widely portable implementations of the MPI standard (up to its latest 2.0 version). Although both libraries claim to provide a modular and easy-to-extend framework, the software stack of Open MPI seems to better suit the goal of introducing a new language's bindings, which our research aims to. The architecture of Open MPI [26] is highly flexible and defines a dedicated layer used to introduce bindings, which are currently provided for C, F77, F90 and some other languages (see also Figure 5). Extending the OMPI-Layer of Open MPI with the Java language support seems to be a very promising approach to the discussed integration of Java bindings, taking benefits of all the layers composing Open MPI's architecture.

##### C. Design and Implementation in Open MPI

We have based our Java MPI bindings on the *mpiJava* code [28]. *mpiJava* provides a set of Java Native Interface (JNI) wrappers to the native MPI v.1.1 communication methods, as shown in Figure 4. JNI enables the programs running inside a Java run-time environment to invoke native C code and thus use platform-specific features and libraries [29], e.g., the InfiniBand software stack. The application-level API is constituted by a set of Java classes, designed in conformance to the MPI v.1.1 and the specification in [20]. The Java methods internally invoke the MPI-C functions using the JNI stubs. The realization details for *mpiJava* can be obtained from [30][31].

Open MPI is a high performance, production quality, MPI-2 standard compliant implementation. Open MPI consists of three combined abstraction layers that provide a full featured MPI implementation: (i) OPAL (Open Portable Access Layer) that abstracts from the peculiarities of a specific system away to provide a consistent interface adding portability; (ii) ORTE (Open Run-Time Environment) that provides a uniform parallel run-time interface regardless of system capabilities; and (iii) OMPI (Open MPI) that provides the application with the expected MPI standard interface. Figure 5 shows the enhanced Open MPI architecture, enabled with the Java bindings support.

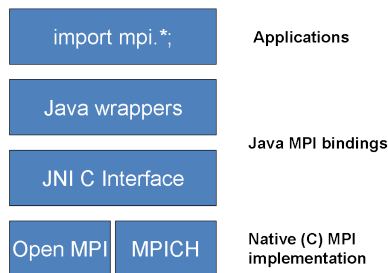


Figure 4. mpiJava architecture

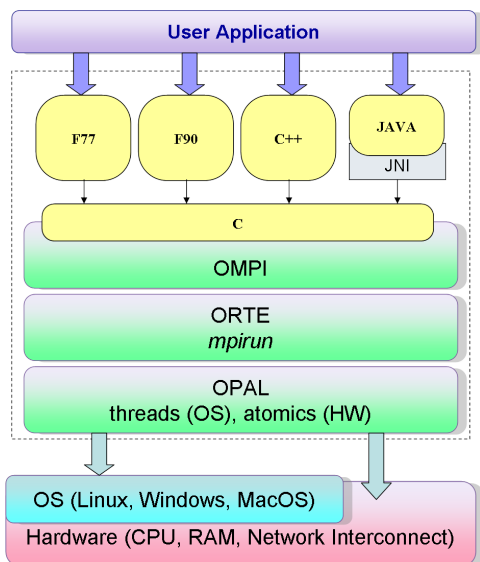


Figure 5. Open MPI architecture

The major integration tasks we performed were as follows:

- Extend the Open MPI architecture to support Java bindings
- Extend the previously available mpiJava bindings to MPI-2 (and possibly upcoming MPI-3) standard
- improve the native Open MPI configuration, build, and execution system to seamlessly support the Java bindings
- Redesign the Java interfaces that use JNI in order to better conform to the native realization
- optimize the JNI code to minimize its invocation overhead
- Create test applications for performance benchmarking

Both Java classes and JNI code for calling the native methods were integrated into Open MPI. However, the biggest integration effort was required at the OMPI (Java classes, JNI code) and the ORTE (run-time specific options) levels. The implementation of the Java class collection followed the same strategy as for the C++ class collection, for which the opaque C objects are encapsulated into suitable class hierarchies and most of the library functions are defined as

class member methods. Along with the classes implementing the MPI functionality (MPI package), the collection includes the classes for error handling (Errhandler, MPIException), datatypes (Datatype), communicators (Comm), etc. More information about the implementation of both Java classes and JNI-C stubs can be found in previous publications [30][24].

#### D. Performance

In order to evaluate the performance of our implementation, we prepared a set of Java benchmarks based on those well-recognized in the MPI community, e.g., NAS [32]. Based on those benchmarks, we compared the performance of our implementation based on Open MPI and the other popular implementation (MPJ Express) that follows a “native Java” approach. Moreover, in order to evaluate the JNI overhead, we reproduced the benchmarks also in C and ran them with the native Open MPI. Therefore, the following three configurations were evaluated:

- **ompic** - native C implementation of Open MPI (the actual trunk version), built with the GNU compiler (v.4.6.1),
- **ompJava** - our implementation of Java bindings on top of *ompic*, running with Java JDK (v.1.6.0), and
- **mpj** - the newest version of MPJ Express (v.0.38), a Java native implementation, running with the same JDK.

We examined two types of communication: point-to-point (between two nodes) and collective (between a group of nodes), varying the size of the transmitted messages. We did intentionally not rely on the previously reported benchmarks, e.g. [33], in order to eliminate the measurement deviations that might be caused by running tests in a different hardware or software environment. Moreover, in order to ensure a fair comparison between all these three implementations, we ran each test on the absolutely same set of compute nodes.

The point-to-point benchmark implements a “ping-pong” based communication between two single nodes; each node exchanges the messages of growing sizes with the other node by means of blocking Send and Receive operations. As expected, our *ompJava* implementation was not as efficient as the underlying *ompic*, due to the JNI function calls overhead, but showed much better performance than the native Java based *mpj* (Figure 6). Regardless of the message size, *ompJava* achieves around eight times higher throughput than *mpj* (see Figure 7).

The collective communication benchmark implements a single blocking message gather from all the involved nodes. Figure 8 shows the results collected for  $P = 2^k$  (where  $k=2-7$ ) nodes, with a varying size of the gathered messages. The maximal size of the aggregated data was 8 GByte on 128 nodes. Figure 9 demonstrates the comparison of collective gather performance for all tested implementations on the maximal number of the available compute nodes (128).



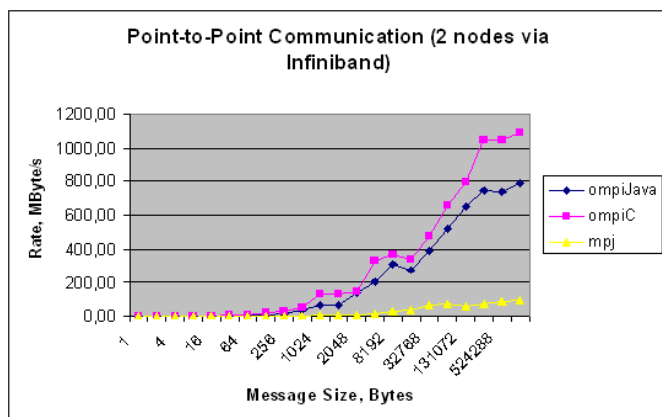
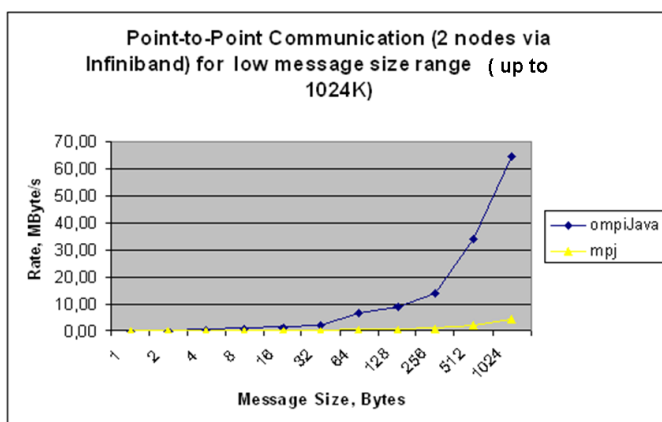
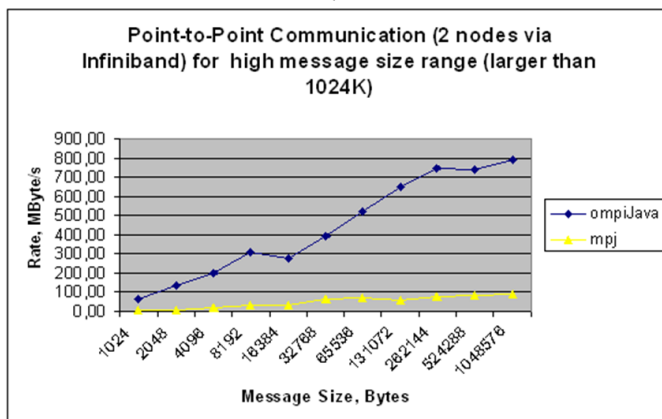


Figure 6. Message rate for the point-to-point communication



a)



b)

Figure 7. Comparison of the message rate for ompJava and mpj for a) low and b) high message size range

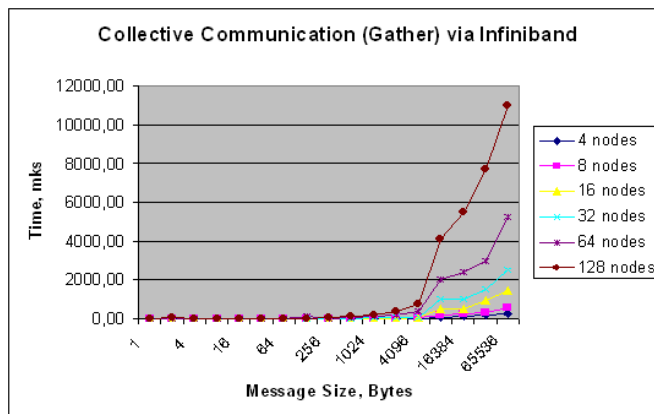


Figure 8. Collective gather communication performance of ompJava

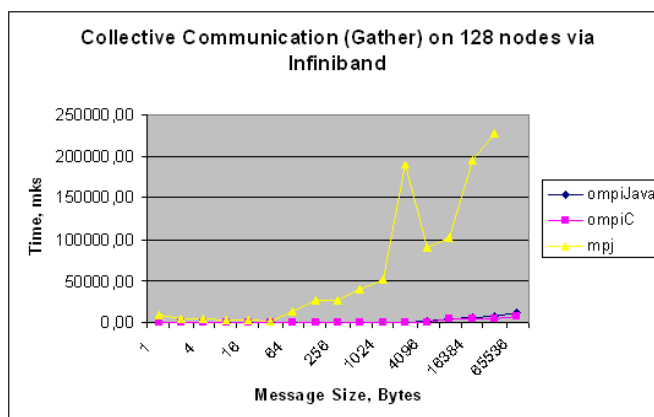


Figure 9. Collective gather communication performance on 128 nodes

Whereas the InfiniBand-aware *ompJava* and *ompC* scaled quite well, the native Java based *mpj* has shown very poor performance; for the worst case (on 128 nodes) a slow-down up to 30 times compared with *ompJava* was observed.

### V. MPI IMPLEMENTATION OF RANDOM INDEXING

Random indexing [34] is a word-based co-occurrence statistics technique used in resource discovery to improve the performance of text categorization. Random indexing offers new opportunities for a number of large-scale Web applications performing the search and reasoning on the Web scale [35].

The main challenges of the random indexing algorithms lay in the following:

- Huge and high-dimensional vector space. A typical random indexing search algorithm performs traversal over all the entries of the vector space. This means, that the size of the vector space to the large extent determines the search performance. The modern data stores, such as Linked Life Data or Open PHACTS consolidate many billions of statements and result in vector spaces

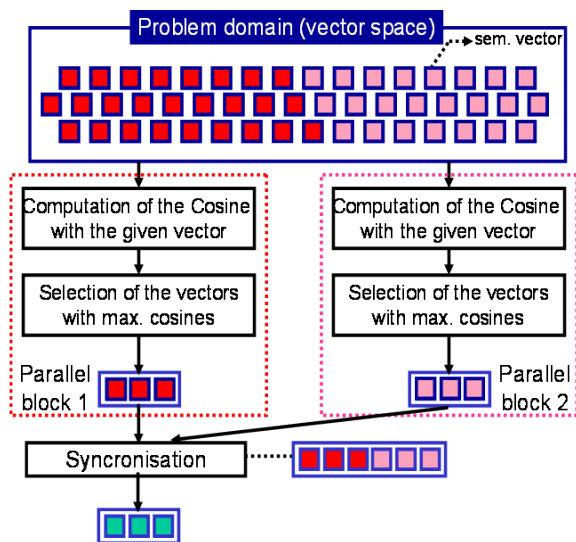


Figure 10. MPI-based parallel implementation of Airhead Search

of a very large dimensionality. Performing Random indexing over such large data sets is computationally very costly, with regard to both execution time and memory consumption. The latter poses a hard constraint to the use of random indexing packages on the serial mass computers. So far, only relatively small parts of the Semantic Web data have been indexed and analyzed.

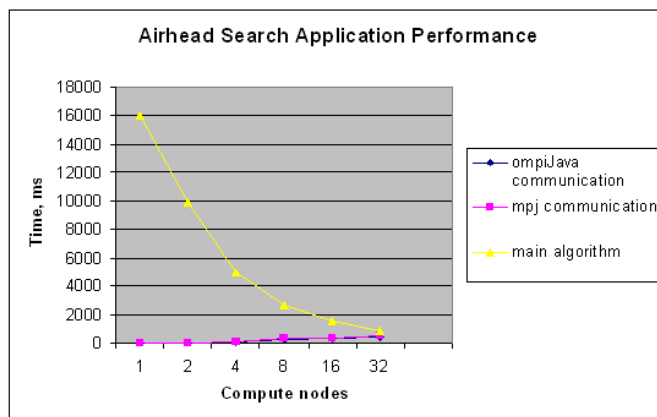
- High call frequency. Both indexing and search over the vector space is highly dynamic, i.e., the entire indexing process repeats from scratch every time new data is encountered.

In our previous work [36], we have already reported on the efforts done on parallelizing Airhead - an open source Java implementation of Random Indexing algorithm. Our MPI implementation of the Airhead search is based on a domain decomposition of the analyzed vector space and involves both point-to-point and collective gather and broadcast MPI communication (see the schema in Figure 10). In our current work, we evaluated the MPI version of Airhead with both *ompijava* and *mpj* implementations.

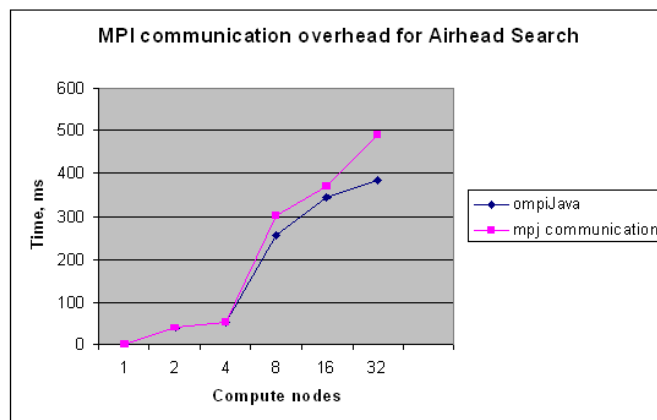
We performed the evaluation for the largest of the available data sets reported in [36] (namely, Wiki2), which comprises 1 Million of high density documents and occupies 16 GByte disk storage space. The overall execution time (wall clock) was measured. Figure 11a shows that both *ompijava* and *mpj* scale well until the problem size is large enough to saturate the capacities of a single node. Nevertheless, our implementation was around 10% more efficient over *mpj* (Figure 11b).

### VI. FUTURE WORK

Our future work will concentrate on promoting both MPI standard and our *ompiJava* implementation to Semantic Web



a)



b)

Figure 11. Airhead performance with *ompiJava* and *mpj*

applications as well as improving the current realization of the Java bindings in Open MPI. With regard to promotion activities, we will be introducing our data-centric and MPI-based parallelization approach to further challenging data-intensive applications, such as Reasoning [37]. Regarding this application, there are highly successful MPI implementations in C, e.g., the parallel RDFS graph closure materialization presented in [38], which are indicatively much more preferable over all the existing Java solutions in terms of performance. Our implementation will allow the developed MPI communication patterns to be integrated in existing Java-based codes, such as Jena [2] or Pellet [39], and thus drastically improve the competitiveness of the Semantic Web application based on such tools.

The development activities will mainly focus on extending the Java bindings to the full support of the MPI-3 specification. We will also aim at adding Java language-specific bindings into the MPI standard, as a reflection of the Semantic Web importance in supercomputing.

## VII. CONCLUSION

High Performance Computing is relatively a new trend for the Semantic Web, which however has gained a tremendous popularity thanks to the recent advances in developing data-intensive applications.

The Message Passing Interface seems to provide a very promising approach for developing parallel data-centric applications. Unlike its prominent alternatives MapReduce and IBIS, the MPI functionality is delivered on the library-level, and thus does not require any considerable development efforts in order to be implemented in the existing serial applications. Using MPI, the Semantic Web applications can take full advantage of modern parallel computing resources. For the RDF processing algorithms, MPI allows for achieving higher scalability and eliminates the need of approximation and dependency minimization in partitioning the work load, used in the previous known implementations as a workaround to overcome the performance limitations on the serial hardware.

We introduced a new implementation of the Java bindings for MPI that is integrated in one of the most popular open source MPI-2 libraries nowadays - Open MPI. The integration allowed us to deliver a unique software environment for flexible development and execution of parallel MPI applications, integrating the Open MPI framework's capabilities, such as portability and usability, with those of mpiJava, such as an extensive set of Java-based API for MPI communication. We evaluated our implementation for Random Indexing, which is one of the most challenging Semantic Web applications in terms of the computation demands currently. The evaluation has confirmed our initial considerations about the high efficiency of MPI for parallelizing Java applications. In the following, we are going to investigate further capabilities of MPI for improving the performance of data-centric applications, in particular by means of MPI-IO (MPI extension to support efficient file input-output). We will also concentrate on promoting the MPI-based parallelization strategy to the other challenging and performance-demanding applications, such as Reasoning. We believe that our implementation of Java bindings of MPI will attract Semantic Web development community to increase the scale of both its serial and parallel applications. The successful pilot application implementations done based on MPI, such as materialization of the finite RDFS closure presented in [38], offer a very promising outlook regarding the future perspectives of MPI in this community.

## ACKNOWLEDGMENT

The author would like to thank the Open MPI consortium for the support with porting mpiJava bindings as well as the EU-ICT Project LarKC [40], partly funded by the European Commission's ICT activity of the 7th Framework Programme (ICT-FP7-215535), for the provided pilot use case.

## REFERENCES

- [1] Openphacts eu project website. [Online]. Available: <http://www.openphacts.org> [retrieved: June, 2012]
- [2] P. McCarthy. Introduction to jena. IBM developerWorks. [Online]. Available: <http://www.ibm.com/developerworks/xml/library/j-jena> [retrieved: June, 2012]
- [3] Lod cloud diagram. [Online]. Available: <http://richard.cyganiak.de/2007/10/lod/> [retrieved: June, 2012]
- [4] Open government data website. [Online]. Available: <http://opengovernmentdata.org/> [retrieved: June, 2012]
- [5] R. Gonzalez. (2012) Closing in on a million open government data sets. [Online]. Available: [http://semanticweb.com/closinginona-millionopen-governmentdatasets\\_b29994](http://semanticweb.com/closinginona-millionopen-governmentdatasets_b29994) [retrieved: June, 2012]
- [6] A. Cheptsov and M. Assel, "Towards high performance semantic web experience of the larkc project," *inSiDE - Journal of Innovatives Supercomputing in Deutschland*, vol. 9(1), pp. 569–571, Spring 2011.
- [7] E. Goodman, D. J. Haglin, C. Scherrer, D. Chavarria, J. Mogill, and J. Feo, "Hashing strategies for the cray xmt," in *Proc. 24th IEEE Int. Parallel and Distributed Processing Symp.*, 2010.
- [8] J. Dean and S. Ghemawat, "Mapreduce- simplified data processing on large clusters," in *Proc. OSDI04: 6th Symposium on Operating Systems Design and Implementation*, 2004.
- [9] M. Bornemann, R. van Nieuwpoort, and T. Kielmann, "Mpi/ibis: A flexible and efficient message passing platform for java," *Concurrency and Computation: Practice and Experience*, vol. 17, pp. 217–224, 2005.
- [10] (1995) Mpi: A message-passing interface standard. Message Passing Interface Forum. [Online]. Available: <http://www.mcs.anl.gov/research/projects/mpi/mpi-standard/mpi-report-1.1/mpi-report.htm> [retrieved: June, 2012]
- [11] R. van Nieuwpoort, J. Maassen, G. Wrzesinska, R. Hofman, C. Jacobs, T. Kielmann, and H. Bal, "Ibis: a flexible and efficient java based grid programming environment," *Concurrency and Computation: Practice and Experience*, vol. 17, pp. 1079–1107, June 2005.
- [12] R. van Nieuwpoort, T. Kielmann, and H. Bal, "User-friendly and reliable grid computing based on imperfect middleware," in *Proc. ACM/IEEE Conference on Supercomputing (SC'07)*, November 2007.
- [13] (2004, February) Resource description framework (RDF). RDF Working Group. [Online]. Available: <http://www.w3.org/RDF/> [retrieved: June, 2012]
- [14] "Lustre file system - high-performance storage architecture and scalable cluster file system," White Paper, SunMicrosystems, Inc., December 2007.

- [15] Portable batch systems. [Online]. Available: [http://en.wikipedia.org/wiki/Portable\\_Batch\\_System](http://en.wikipedia.org/wiki/Portable_Batch_System) [retrieved: June, 2012]
- [16] A. Dimovski, G. Velinov, and D. Sahnaski, "Horizontal partitioning by predicate abstraction and its application to data warehouse design," in *ADBS*, 2010, pp. 164–175.
- [17] D. J. Abadi, A. Marcus, S. R. Madden, and K. Hollenbach, "Scalable semantic web data management using vertical partitioning," in *Proc. The 33rd international conference on Very large data bases (VLDB'07)*.
- [18] C. Curino, E. P. C. Jones, S. Madden, and H. Balakrishnan, "Workload-aware database monitoring and consolidation," in *SIGMOD Conference*, 2011, pp. 313–324.
- [19] A. Cheptsov, M. Assel, B. Koller, R. Kbert, and G. Gallizo, "Enabling high performance computing for java applications using the message-passing interface," in *Proc. The Second International Conference on Parallel, Distributed, Grid and Cloud Computing for Engineering (PARENG'2011)*.
- [20] B. Carpenter, G. Fox, S.-H. Ko, and S. Lim, "mpiJava 1.2: Api specification," Northeast Parallel Architecture Center. Paper 66, 1999. [Online]. Available: <http://surface.syr.edu/npac/66> [retrieved: June, 2012]
- [21] T. Kielmann, P. Hatcher, L. Boug, and H. Bal, "Enabling java for high-performance computing: Exploiting distributed shared memory and remote method invocation," *Communications of the ACM*, 2001.
- [22] M. Baker, B. Carpenter, and A. Shafi, "MPJ Express: Towards thread safe java hpc," in *Proc. IEEE International Conference on Cluster Computing (Cluster'2006)*, September 2006.
- [23] R. K. Gupta and S. D. Senturia, "Pull-in time dynamics as a measure of absolute pressure," in *Proc. IEEE International Workshop on Microelectromechanical Systems (MEMS'97)*, Nagoya, Japan, Jan. 1997, pp. 290–294.
- [24] G. Judd, M. Clement, Q. Snell, and V. Getov, "Design issues for efficient implementation of mpi in java," in *Proc. the 1999 ACM Java Grande Conference*, 1999, pp. 58–65.
- [25] B. Carpenter, V. Getov, G. Judd, A. Skjellum, and G. Fox, "MPJ: Mpi-like message passing for java," *Concurrency and Computation - Practice and Experience*, vol. 12(11), pp. 1019–1038, 2000.
- [26] E. G. et al., "Open MPI: Goals, concept, and design of a next generation MPI implementation," in *Proc., 11th European PVM/MPI Users' Group Meeting*, Budapest, Hungary, September 2004, pp. 97–104.
- [27] Mpich2 project website. [Online]. Available: <http://www.mcs.anl.gov/research/projects/mpich2/> [retrieved: June, 2012]
- [28] mpijava website. [Online]. Available: <http://sourceforge.net/projects/mpijava/> [retrieved: June, 2012]
- [29] S. Liang, Ed., *Java Native Interface: Programmer's Guide and Reference*. Addison-Wesley, 1999.
- [30] M. Baker, B. Carpenter, G. Fox, S. Ko, and S. Lim, "mpi-Java: An object-oriented java interface to mpi," in *Proc. International Workshop on Java for Parallel and Distributed Computing IPPS/SPDP*, San Juan, Puerto Rico, 1999.
- [31] M. Vodel, M. Sauppe, and W. Hardt, "Parallel high-performance applications with mpi2java - a capable java interface for mpi 2.0 libraries," in *Proc. The 16th Asia-Pacific Conference on Communications (APCC)*, Nagoya, Japan, 2010, pp. 509–513.
- [32] Nas parallel benchmark website. [Online]. Available: <http://sourceforge.net/projects/mpijava/> [retrieved: June, 2012]
- [33] Mpj express benchmarking results. [Online]. Available: <http://mpj-express.org/performance.html> [retrieved: June, 2012]
- [34] 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. 1–9.
- [35] 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.
- [36] M. Assel, A. Cheptsov, B. Czink, D. Damljanovic, and J. Quesada, "Mpi realization of high performance search for querying large rdf graphs using statistical semantics," in *Proc. The 1st Workshop on High-Performance Computing for the Semantic Web*, Heraklion, Greece, May 2011.
- [37] D. Fensel and F. van Harmelen, "Unifying reasoning and search to web scale," *IEEE Internet Computing*, vol. 11(2), pp. 95–96, 2007.
- [38] J. Weaver and J. A. Hendler, "Parallel materialization of the finite rdfs closure for hundreds of millions of triples," in *Proc. International Semantic Web Conference (ISWC) 2009*, A. B. et al., Ed., 2009.
- [39] E. Sirin, B. Parsia, B. C. Grau, A. Kalyanpur, and Y. Katz. Pellet: a practical owl-dl reasoner. *Journal of Web Semantics*. [Online]. Available: <http://www.mindswap.org/papers/PelletJWS.pdf> [retrieved: June, 2012]
- [40] Eu-fp7 project large knowledge collider (larkc). [Online]. Available: <http://www.larkc.eu> [retrieved: June, 2012]

# Ontological Representation of Knowledge Related to Building Energy-efficiency

German Nemirovski

Business and Computer Science  
Albstadt-Sigmaringen-University of Applied Sciences  
Albstadt, Germany  
nemirovskij@hs-albsig.de

Álvaro Sicilia, Fàtima Galán, Marco Massetti

Leandro Madrazo  
ARC Enginyeria i Arquitectura La Salle  
Universitat Ramon Llull  
Barcelona, Spain  
[asicilia, fatima, mmassetti, madrazo]@salleurl.edu

**Abstract** — Within the research project RÉPENNER, co-financed by the R+D+i Spanish National Plan, an information system to capture the energy-related data throughout the whole building life cycle is being developed. The purpose of the system is to provide improved quality information to the various stakeholders participating at the different stages of the building life cycle. This higher-quality information is derived from interlinking disparate data sources – proprietary and open – and from the application of mining techniques to the semantically modelled data. This paper describes the design and the most important features of the RÉPENNER global ontology, which is the core component of the information system is being developed. The ontology embraces knowledge originated from three realms: canonical domain knowledge, praxis-related usage cases and energy-related data stemming from various sources. The ontological design process – which includes the acquisition, unification, extension, formal specification and evaluation of the knowledge – is presented as a case study on knowledge discovery and engineering.

*Keywords-semantic web; ontology; taxonomy; information system; energy-efficiency; energy model.*

## I. INTRODUCTION

Due to rapid technological development and the imminent shortage of fossil energy resources required in nearly all technological areas, crucial decisions must be made in regards to the reduction of energy consumption. In recent years, particularly in the area of building construction, a great deal of meaningful data has been collected, the analysis of which can help in the decision-making process related to this domain. On aggregate, this data appears to be a real treasury for data mining and visualization methods, which might help to improve the energy performance of buildings. However, the available data is located in different data sources, heterogeneously structured and formatted. Thus, access to data in the right format and at the right time remains a substantial challenge for those who will develop services that help to improve the energy-efficiency of existing and planned buildings.

Two critical questions must be answered on the way towards the development of such services: 1) how to enable efficient querying over the entire space of distributed data, e.g., for the purposes of data mining; and 2) how to make the portfolio of all available data transparent to actors operating

at each phase of the building life cycle – from the design to construction and operation. Subsequently, an information system addressing these questions and tasks should provide lookup, browsing and data-transformation facilities which operate over the entire distributed data space.

The purpose of the RÉPENNER project [2] is to develop an ontology-based information system which supports decision-making processes and knowledge discovery by actors concerned with the energy management of buildings. In recent years, studies on data integration using ontologies have delivered substantial results. The main example which proves the feasibility of tasks solutions is the Linked Open Data project [1]. By September 2011, it had integrated 31 billion data records specified in the RDF format, which is the most popular language for the specification of ontology-related information.

This paper presents the design of the ontology which is a core component in the RÉPENNER information system. A comprehensive project description can be found in [2]. An important feature of this ontological design is the conceptualization of the domain knowledge determined from three different perspectives: first, the perspective of actors expressed through use case specifications such as energy consumption analysis and prediction; secondly, the perspective of canonical domain expertise expressed through standardization approaches in the field of energy performance of buildings; and thirdly, the perspective of data access expressed through models of data sources (e.g., entity relationship models).

The balance of the paper is structured as follows: Section II is dedicated to the description of background and methodology; in Section III, the process of knowledge acquisition is explained; the implementation and the ontology coding details are described in Section IV; Section V focuses on the goals and method of the ontological evolution; and lastly, in Section VI, some conclusions are summarized.

## II. METHODOLOGY

### A. Role of ontologies in data integration

The term "ontology" has been used in computer science since the early 1990s. One widely acknowledged definition was given by Gruber [3]. Ontology is an explicit

conceptualization of a knowledge domain whereby the basic ontology element, a concept, represents a term and its relationship to other terms from the vocabulary used in this domain. Therefore, ontological concepts are interrelated, e.g., "house" is a sub-concept of "building." Such relationships can be defined in the form of axioms, conceptual properties connecting concepts to each other or attributes connecting concepts to value domains, such as "integer" or "literal". A subsumption hierarchy of concepts interrelated by specialization/generalization or sub-concept/super-concept relationship lies at the core of the ontology. This is called taxonomy. Ontologies are formally specified using Description Logic formalisms, and are coded in machine-readable languages like OWL. These features make ontologies essential for the specification of vocabularies in such fields as natural language processing [4] and Semantic Web [5].

The general idea of using ontologies for the interlinking and querying of distributed data is based upon the property of ontological concepts to be represented by their instances. For example, the two records "residence of Nobel laureate in chemistry Carl Bosch" and "house, located in Schloß-Wolfsbrunnenweg 33, 69118 Heidelberg" can be specified as instances of one single concept titled "Villa Bosch". In this case, the semantic equality of these records becomes evident, not only for humans but also for artificial agents performing ontology-based information retrieval. Even if these records are stored in two different sources using different formats and data models, an artificial agent searching for occurrences of the concept "Villa Bosch" will be able to identify the concept/instance relations and retrieve both of them. Thus, the interoperability of heterogeneously structured data can be achieved by establishing references between data chunks and ontological concepts or, in other words, by revealing data semantics. This fact has made semantic modelling one of the most efficient technologies for the integration of distributed heterogeneously structured data. The Linked Open Data Project mentioned in the introduction follows a decentralized modelling approach based on this principle, and it uses shared identifiers (URIs) to interlink data distributed over the linked sources. Therefore, most open-link data sources are represented by an ontology and a single access point able to process queries formulated in a standard query language, e.g., SPARQL with respect to this ontology. However, the Linked Open Data approach faces two obstacles: 1) the structure of single data sources, i.e., the architecture of the corresponding ontologies, is usually unknown and, therefore, combining data stored in different sources requires discovering all sources where data may be located; 2) to discover data sources, one needs to interact with multiple endpoints offering a data querying interface [6]. In terms of openness and flexibility, such an approach works well. However, for the sake of efficiency and the completeness of the information thus retrieved, a centralized approach is preferable.

According to the centralized modelling approach, a single ontology is used as the main reference for all distributed data. The data of a single source either refers to concepts of the central ontology or to concepts of dedicated

source ontologies univocally mapped to the central ontology, by which each concept of the source ontology corresponds to one of the central ontology. In such a system, agents query data sources in interaction with a single, central end point, whereby all queries use the vocabulary of the central ontology. The process of indexing and looking up the entire distributed data space constitutes an integrated service of the information system. In this context, Calvanese [7] described an information integration scenario in which source models are mapped onto a central enterprise model specifying the entire knowledge over the distributed knowledge space. This approach was followed by Doerr [8], using the term "core ontology" to refer to an integrative ontology similar to the enterprise model. Uschold [9] defined the global ontology as either an intersection of local ontologies -- given that it encompasses concepts, properties and axioms shared by local ontologies -- or as a *union of elements from all local ontologies in the case of an intended application of the global ontology as one which would reference the entire space of terms*. Calvanese [10] introduced a formal framework which facilitates the efficient querying of integrated data corpus in a centralized manner.

The RÉPENER project follows the centralized approach of ontology-based data integration, adopting the terminology of Uschold [9]. Accordingly, ontologies which specify the data located in single sources are called local ontologies, while the central ontology serving as a target for the mapping of local ontologies, being defined as the union of their elements, is called global ontology.

#### B. Related approaches for energy data

Semantic technologies have already been applied to model energy information. However, and according to Keistead [11], "*there is not yet one widely used conceptualization for energy systems*".

However, there are ontologies developed in specific domains, such as in building usage and operation. Shah and Chao [12] created an electrical home appliance ontology which facilitates the occupant's awareness regarding energy consumption in the house. For the same purpose, a smart home knowledge base has been developed using semantic web standards [13]. Additionally, ontologies have been used in the process of designing a device platform to integrate different device standard models [14]. In this respect, semantic technologies have been applied for the purpose of ensuring the interoperability among device industry standards such as BACnet, KNX, LON, or EnOcean [15]. Ontology inference processes have been used to enhance a building management system based on ontology modelling [16]. More recently, Wagner proposed the semantic web as a foundation for the Smart Grid communication architecture [17].

Applications of semantic technologies to specific domains related to energy-efficiency in buildings – operation, interoperability, smart grid – are present in the literature, but they do not model the energy data generated by different applications throughout the building's life cycle. To our knowledge, one of the first attempts to model these data was carried out during the IntUBE project [18].

### C. General design strategies: collaboration and modularity

The design of formally specified ontologies has been an object of research since the early 1990s. Two significant works in this regard were carried out by Gruber [3] and Uschold and King [19]. The former defines the properties of ontological knowledge representation for the purposes of the engineering sciences. The latter deals with the design process of ontologies, being described as consisting of four phases: identifying ontology purposes, building the ontology, evaluating and documenting. In turn, the phase of ontology building is subdivided into three steps: 1) ontology capture, namely, definition, naming and description of key concepts and relationships between them 2) ontology coding, that is, using one of the formal languages or tools and 3) integrating existing ontologies. This approach has been further elaborated in work on this topic. A survey of up-to-date methodologies for ontological design was provided by Contreras and Martinez-Comenche [20].

Already in the 1990s, it became obvious that ontologies designed for practical industrial or medical application could be large and complex. Therefore, to overcome the complexity of ontology management, two approaches have emerged: collaborative ontological design supported by dedicated multi-user environments, as shown in Swartout [21], Sure [22], or Tudorache [23]; and reusing ontology elements, e.g., design patterns, as shown in Presutti [24] and Gangemi [25], or ontology modules as discussed by Cuenca Grau [26].

In contrast to the classic procedure described by Uschold and King [19], the design process of the RÉPENNER global ontology can be depicted as a sequence of iterations encompassing knowledge capture (conceptualization, concept naming, and description), and ontology coding using OWL 2 specification language and evaluation (Figure 1). Our approach followed this scenario. After each iteration, the ontology became more and more comprehensive.

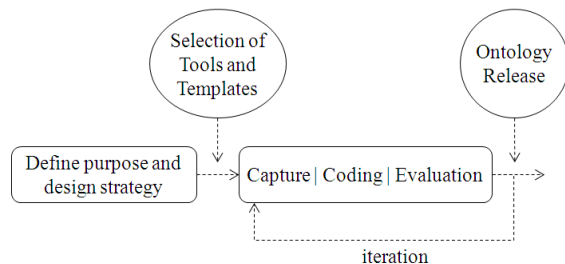


Figure 1. Design process of RÉPENNER global ontology.

The design process involved energy-domain experts and ontology engineers working at different locations in Germany and Spain. In diligence style [20], different ideas and proposals were generated in a distributed way through tools like Adobe Connect, Skype and Google Docs used as a platform for the project development. An Excel document was used as an instrument to capture the domain knowledge from the different realms in order to unify the terms and identify relationships between them. The resulting structure

was the base of the ontological design process (see Section III).

Based upon the approach of modular ontological design, RÉPENNER global ontology is built on certain selected modules of an upper-level ontology. In this way, each concept of the global ontology subsumes the concepts of the Suggested Upper Merged Ontology (SUMO). In this way, the foundational relationships and axioms which are valid for SUMO concepts remain valid for those of the RÉPENNER global ontology. Hence, the philosophical, engineering and linguistic issues incorporated by the SUMO ontology have been inherited by the RÉPENNER global ontology.

## III. INFORMATION CAPTURE

### A. Vocabulary acquisition

In each design iteration, the knowledge capture was carried out by: a) keeping in mind the purpose of the RÉPENNER global ontology, i.e., data management; b) taking into account the services to be performed by an information system for the energy-efficiency of buildings; and c) referring to the canonical knowledge structure of the domain of interest. This paradigm is reflected in the three-dimensional architecture of the term space (Figure 2), which became part of an informal knowledge specification aiming at determining the ontology vocabulary, including terms, relations data types and units of measure. One of the challenges in the vocabulary acquisition process is to avoid redundancy and terminology mismatching, which usually occur in the aggregation of heterogeneous information. To avoid this, a maximum number common terms for each dimension was identified.

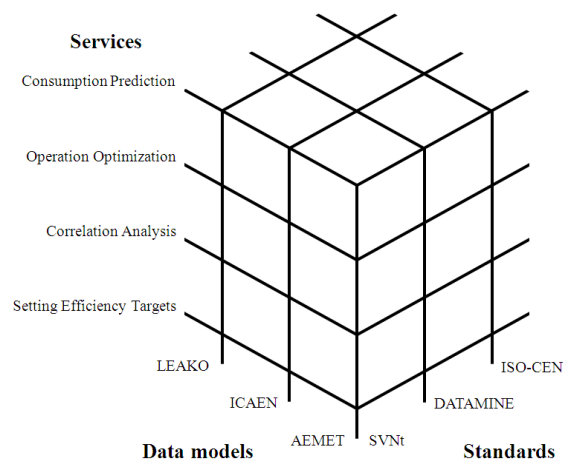


Figure 2. Three-dimensional architecture of the RÉPENNER term space.

The three dimensions of the term space mentioned above are illustrated in Figure 2. The first dimension comprises data sources containing two kinds of energy information: building information (building systems, energy consumption, energy demand, etc.) and contextual data (economic context, demographic context, climatic context, etc.). In the initial project phase, data from three sources was used: a) a

database of LEAKO, a Basque company handling installation, distribution, and HVAC control. The database contains consumption data for thermal (kwh) consumption for heating, hot water, gas and water consumption and indoor conditions, e.g., air temperature in several monitored residential buildings; b) a database of ICAEN, an organization of the Catalan government which gathers the energy certificates of newly planned buildings, including their simulated performance; and c) AEMET climate data from the Spanish Meteorological Agency. In this last case, the terms, relationships, data, and units of measure were extracted from the entity relationship models specifying the data sources.

The second dimension was built on the basis of standards and key parameters classifications, used to manage energy performance of buildings. The energy certification of buildings defined by DATAMINE project [34], the ISO CEN standards following the European Directive 2002/91/EC (e.g., ISO 13790:2008) and the Standard Network Variable Types from LonWorks (SNVTs), were utilized in the first two years of the RÉPENNER project. The terms were extracted out of document texts and tables.

The third dimension comprises services addressing support to stakeholders in the realms of their decision-making processes (design, maintenance). The first group of prototypically developed services consists of: a) a prediction service launched in the design phase, whose goal is to provide qualified information regarding the consumption and demand of a building construction; b) an operation optimization service for building managers to optimize the building's behaviour based on the reference data obtained from other buildings; c) a correlation analysis service to identify the key factors influencing energy consumption; and d) a service for setting the energy targets to be reached in the refurbishment of the existing buildings.

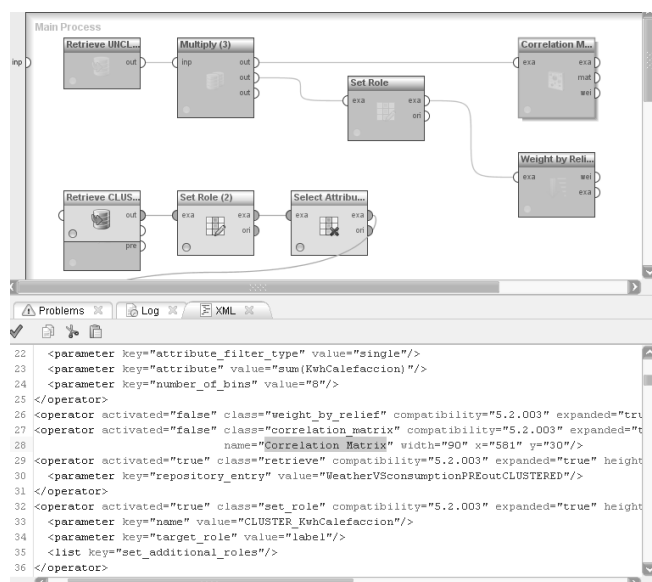


Figure 3. Data-mining process specification in RapidMiner software.

In this case, terms were extracted from the data-mining process specifications that were defined using RapidMiner software. In this software, processes are specified in XML and presented in a graphical editor, as shown in Figure 3. For obtaining terms for the energy model, a simulation of the above-described services took place by specifying the corresponding RapidMiner processes for propositionalized data from LEAKO and ICAEN databases. The terms were then extracted from the process specification (Figure 4).

Data definitions (DATAMINE Structure; ISO 13790:2008)				Input Data	Output Data				
<b>DATAMINE DATA STRUCTURE</b> Version 1.0 from 30th October 2006				<b>USE CASE</b> Search data level of filtering 1 Service data Importa nce 1 Importa nce 2					
No.	data field name	label	unit	definition	input type	option 1	option 2	option 3	
<b>B General Building Data</b>									
25	building location: city	bu_city			f				berdan viala del valles
27	building location: region	bu_region		if applicable, for each country a list of regions (respectively provinces, departments, Bundesländer ...)	p				
28	building location: climate zone	bu_climate		should be provided if national climate zones are defined a list should be provided	p		c2 (25)		
29	building erection yearperiod: first year	year1_buiding	year	year of erection (finishing) of the building. If not the concrete year but the approximate time period is known (e.g. building was erected some time between 1900 and 1920) insert here the first year of this time interval! (in the example: 1900)	f				2006?

Figure 4. Mapping DATAMINE terms onto the input/output parameters of services.

The result of the vocabulary acquisition has been documented in a series of Excel tables implementing relationships within the three-dimensional term space, being transparent for all participants of the collaborative knowledge-capture process. The DATAMINE data structure, which includes energy certificate data, general data of the building, building envelope data, energy demand and/or energy consumption has been used as the primary source of the terms. Figure 4 shows how DATAMINE field names (in the right part of the figure) are mapped onto the input and output terms of the data-analysis services (titled as “use cases” in the right part of the table). Three tables of this type are required for mapping the three dimensions in succession.

### B. Hierarchy of terms

In Section III.A, it has been shown how the terms, which originated at different realms of the three-dimensional term space, are mapped onto each other for the purpose of identifying a common vocabulary. Such dimensional mapping represents part of the energy model, which is the first step in the process of creating a formal ontology.

The other part of the energy model is a hierarchy of terms unified by the mappings. Such a hierarchy has been specified by means of the relationships *contains/part of*. The top level of the hierarchy is made of the domain names,



while the second level contains terms specifying sub-domains. This partitioning is extended up to the last hierarchy level, which contains terms associated with basic parameters such as envelope properties or heat-transfer coefficient.

Figure 5 shows the hierarchy of terms in a simplified form. The most important parts of the building energy domain, which is the core domain of the energy model, can be defined as follows:

- General project data: parameters which identify the project and define its generic characteristics such as location, use, project execution data, and site description;
- Performance: building performance indicators regarding energy use (energy demands, consumption of different energy carriers, e.g., gas or electricity, and different uses, e.g., heating, cooling, hot water, electricity and appliances), CO<sub>2</sub> emissions and indoor conditions (e.g., temperature and humidity);
- Building properties: geometric characteristics, construction systems and building services;
- Outdoor environment: climate characteristics and conditions of the physical environment which determine the building's performance: outdoor temperature, wind speed and direction, and solar radiation;
- Operation: usage and management of the building and its facilities for maintaining comfort levels (e.g., solar protection and thermostat regulation). It also includes the effects of the occupant activity in the indoor environment, such as thermal loads produced by occupants, lighting and appliances;
- Certification: information associated with building energy certificates. It includes indicators to qualify a building based on performance, e.g., according to a conventional scale as (A, B, C, etc.). It also includes the certification-process methodology.

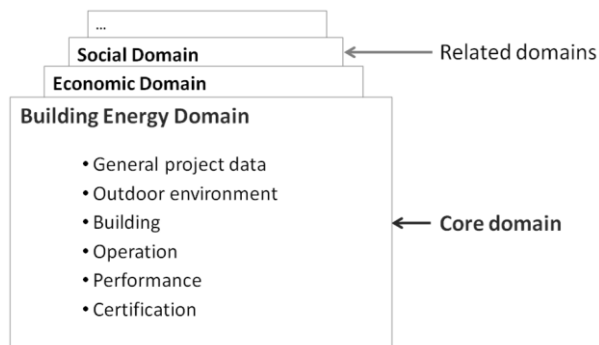


Figure 5. Energy model domains

Studies [27] have shown that the energy consumption of building correlates to socio-economic factors like real estate prices or the income levels of the inhabitants. To take this fact into consideration, we included the economic/social

domain into the building's energy model along with the building energy domain.

#### IV. ARCHITECTURE AND CODING

##### A. Ontology Architecture

As stated in Section II.C, the RÉPENER global ontology uses the Suggested Upper Merged Ontology (SUMO) at the upper level. The selection of SUMO for this role was made after comparing it to other foundational ontologies, such as DOLCE, PROTON, General Formal Ontology (GFO) and Basic Formal Ontology (BFO). SUMO scored well in such fields as simplicity of understanding, applicability for reasoning and inference purposes, and potential reuse in the Building Energy Domain, for instance, reusing concepts for specifying units of measure defined by the SI system (meter, watt, joule, etc.).

Some of the SUMO concepts subsume concepts of the RÉPENER ontology. For example, the concept *Building* is subsumed by SUMO's *StationaryArtifact* and SUMO's *Attribute* subsumes *BuildingProperty*, which in turn subsumes *BuildingGeometry*:

$BuildingGeometry \sqsubseteq BuildingProperty \sqsubseteq Attribute$

The resulting RÉPENER global ontology is a combination of two hierarchies: one of them is the taxonomy based on the concept of subsumption, where the upper level of the taxonomy is represented by generic SUMO concepts. The second hierarchy consists of the terms described in Section III.B, whereby building elements of this hierarchy are aggregative *has* or *includes* properties such as the property *hasGeometry* (Figure 6). The former hierarchy (Figure 5) is required for the formal reasoning, while the latter one (Figure 6) represents the knowledge from the perspective of the domain experts and users.

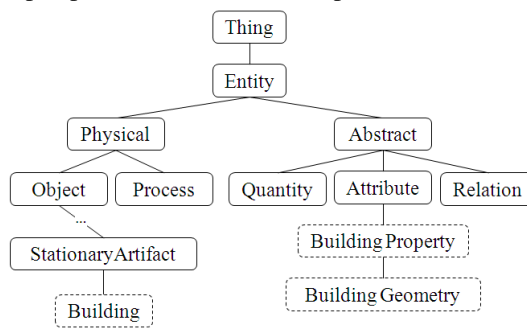


Figure 6. hierarchies as the basis structures of the RÉPENER global ontology.

##### B. Coding

OWL 2 has, in recent years, become a sort of default standard for ontology coding. As shown in Calvanese [28], the use of this specification language in its full version may be disadvantageous in terms of the computability of particular reasoning tasks, particularly those which require

conjunctive queries of large data volumes. Poggi [29] suggested a somewhat restricted  $DL-Lite_A$  formalism, which helps to overcome this obstacle. This approach was adopted by the RÉPENNER global ontology.

A detailed description of  $DL-Lite_A$  formalism is out of scope of this paper. Nevertheless, it is important to mention two of the most important features of an OWL-dialect that implements  $DL-Lite_A$ : 1) domain and range of properties can be specified only for functional data properties; and 2) definition of an object property connecting two OWL classes with each other, has to be modelled by means of axioms and not by specifying property's domain and range. For example, two following axioms in DL notation use subsumption ( $\sqsubseteq$ ), existence quantification ( $\exists$ ) and inversion ( $\bar{\phantom{x}}$ ) to express that the class `BuildingGeometry` relates to the class `Building` via the *hasGeometry* property.

```
Building  $\sqsubseteq$   $\exists$ hasGeometry
 $\exists$ hasGeometry  $\bar{\sqsubseteq}$  BuildingGeometry
```

In OWL the same is specified as follows:

```
<SubClassOf>
  <Class IRI="http://www.owl-
  ontologies.com/SUM0155.owl#Building"/>
  <ObjectSomeValuesFrom>
    <ObjectProperty IRI="#hasGeometry"/>
    <Class abbreviatedIRI=":Thing"/>
  </ObjectSomeValuesFrom>
</SubClassOf>
```

and

```
<SubClassOf>
  <ObjectSomeValuesFrom>
    <ObjectInverseOf>
      <ObjectProperty IRI="#hasGeometry"/>
    </ObjectInverseOf>
  <Class abbreviatedIRI=":Thing"/>
</ObjectSomeValuesFrom>
  <Class IRI="#BuildingGeometry"/>
</SubClassOf>
```

Although domains and ranges of properties are not explicitly specified in the code, if an ontology specification is valid, they can be inferred by reasoner software to be then visualized and viewed by the user.

## V. EVALUATION

Apart of the already mentioned work of Gruber [3], different views on essential ontology properties are described by Gómez-Pérez, [30], Obrst [31], and Gangemi [32]. After a comparative analysis of these approaches, we found that the following three criteria are of primary priority for the RÉPENNER global ontology:

- **Completeness:** in the RÉPENNER context this means that all terms and relations of the three-dimensional space of terms are explicitly specified in the ontology code or can be inferred by reasoning.

- **Intelligibility:** the ability of actors using the ontology and ontology-based applications in their decision-making process to understand the ontology structure.
- **Computational integrity and efficiency:** the ability of the ontology to support reasoning tasks such as conjunctive querying on high efficiency level, i.e., with a comparatively short response time.

Brank [33] described four types of evaluation approaches: 1. comparing ontologies with a “golden standard”, e.g., another ontology; 2. comparing ontologies with source data; 3. evaluating ontology application; and 4. evaluation by humans. In the RÉPENNER project, we have followed three of these approaches: we compared our ontology with source data, evaluated it by humans and evaluated it through the application of reasoners.

1. Comparing ontologies with source data to evaluate ontology completeness: a set of randomly selected items, such as fields and table names from databases of LEAKO and ICAEN or terms from the DATAMINE classification, are (manually) mapped by testers onto the current version of the ontology. If the mapping result for one item corresponds to the mapping in the energy model (Figure 4), the resulting coefficient, initially nullified, will be incremented by one. Success is measured as a percentage where 100% corresponds to the number of preselected items. We have carried out ten evaluations of this kind, selecting twenty different terms from the above-mentioned sources. Six of the twenty terms could be identified in the ontology. Three evaluations ended with the score 18, and one ended with a score of 17. Therefore, the total completeness of the ontology was rated at 95.5%, resulting from the following calculation:  $(20 \cdot 6 + 18 \cdot 3 + 17) \cdot 100 / (10 \cdot 20)$ . However, taking into account the fact that two of nine missing terms were intentionally omitted from the ontology, the completeness would be 96.5%.

2. Evaluation by humans aiming at the quantification of intelligibility: independent testers (who did not participate in the design process) are given the task of navigating the ontology or, in other words, finding a concept. The navigation is carried out in an ontology viewer developed for this purpose. The shortest navigation path from the top of the concept hierarchy (depending on the task, it can be the energy model or the concept taxonomy) is calculated in advance. The result of evaluation is measured as a percentage, where 100% corresponds to the number navigation steps equal to the number of edges in the shortest path minus one and 0% to this number plus 30, i.e., if a tester needed 30 clicks above the required minimum, his score was set to 0. The evaluation was carried out by two groups of testers. One group contained eight computer science students, and the other group contained five experts in the field of building energy. Each tester was offered three terms to find in the ontology. The surprising result of this evaluation was that the average score of these two groups did not differ a lot. The intelligibility of the ontology for

domain experts was 97.30%, while this metric for computer science students achieved the value 91.20%.

3. Evaluation of ontology application with the focus on computational integrity and efficiency: as stated above, in an ontology developed on the basis of *DL-Lite<sub>A</sub>*, the domains and ranges of properties specified using axioms can be inferred by reasoners. However, this method does not provide a measure for the quality of the ontology. Instead, it demonstrates the coding or conceptualization errors which have to be treated immediately.

However, for practical reasons the time required to complete the reasoning tasks is an important matter of consideration. This time strongly depends on i) the expressivity of the DL-Language used to specify the ontology; and ii) the number of axioms contained in an ontology. Our evaluation has shown that the former factor may be crucial for the performance of reasoning, while the latter one has only a moderate influence. For instance, an attempt to integrate QUDT ontology modules specifying units of measure vaulted the time of reasoning carried out on a machine equipped with an Intel i7 2600 CPU and 8GB RAM to three hours. We believe the explanation for this was the highly expressive OWL-profile used for the QUDT specification. The reasoning time for RÉPENNER global ontology using seven selected modules of SUMO upper-level ontology only (in this case, QUDT part was not imported) of a total size of 5.3 MB and containing 100 axioms as those described in IV.B achieved 1 minute 20 seconds on the same machine. When the number of axioms increased to 1,000 the reasoning time rose to 5 minutes 32 seconds (these measures are valid for the HermiT reasoner version 1.3.5). It should be mentioned that originally SUMO is specified using the KIF Knowledge Interchange Format (KIF) language, which has a high level of expressivity. When translated to OWL, however, SUMO modules lose many axioms which cannot be expressed in OWL one to one. Hence the translated version is on the  $\mathcal{EL}$  level [33].

## VI. CONCLUSION AND FUTURE WORK

This paper has presented a case study on knowledge discovery, as was carried out in the context of a particular domain (Building Energy Performance) and aiming at the fulfilment of a particular task (development of an information system for the decision-making process support of stakeholders participating at different stages of a building's life cycle). Within the case study, we have shown several stages of the process of knowledge discovery and engineering:

- 1) *Vocabulary acquisition* from different realms related to the domain and to the task of interest: services exposed by the information system to be developed; structure of data sources to be integrated and canonical domain knowledge in form of standardization documents;

- 2) *Mapping of terms* onto each other for the purpose of defining a common vocabulary for all of the realms;
- 3) *Specification of relationships* between terms, which is an important step from the definition of a vocabulary towards ontological design: in the course of a relationship definition, a term became a concept.
- 4) *Building taxonomy* of concepts by integration the SUMO ontology as the taxonomy's upper level. At this step, the abstract knowledge based on philosophic, linguistic and engineering postulates, as discovered and constructed by a third party, became part of the ontology being constructed;
- 5) *Formal specification* of the discovered knowledge, i.e., the elaboration of this knowledge towards a formal ontology. This operation makes the new knowledge available for exploitation, particularly in the context of data management and decision-making support on which the RÉPENNER project has its focus;
- 6) *Knowledge Evaluation*, using distinct criterion and methods.

This paper addressed issues related to knowledge discovery and ontological design, as were carried out in the context of the RÉPENNER project aiming at development of an information system supporting the stakeholder in all phases of a given building's life cycle. Nevertheless, the paper did not address the implementation aspect of the information system. Neither can we argue if the evaluation of the ontology hereby presented can replace the evaluation of the information services, to be developed. This is specified in a separate paper [2], which presents further motivation and the context for the RÉPENNER global ontology. However, we assume the existence of a strong correlation between the quality of the ontology and the quality of information services. The demonstration of such correlation will be the goal of further work. One of the most important tasks in this regard will be computability evaluation of the resulting information system using benchmarks for conjunctive queries addressing distributed data.

## ACKNOWLEDGEMENT

RÉPENNER is being developed with the support of the research program BIA 2009-13365, co-funded by the Spanish National R+D+i Plan 2009-2012.

## REFERENCES

- [1] T. Heath, and C. Bizer, *Linked Data: Evolving the Web into a Global Data Space* (1st ed.), Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 2011, pp. 1-136. Morgan & Claypool.
- [2] L. Madrazo, A. Sicilia, M. Massetti, and F. Galan, "Semantic modeling of energy-related information throughout the whole building lifecycle," In *Proceedings of the 9th European Conference on Product and Process Modelling ECPPM*. Iceland, 2012.

- [3] T. Gruber, "Towards principles for the design of ontologies used for knowledge sharing," *International Journal of Human Computer Studies*, Vol. 43 (5-6), 1995, pp. 907-928.
- [4] A. Kiryakov, B. Popov, I. Terziev, D. Manov, and D. Ognyanoff: "Semantic Annotation, Indexing, and Retrieval," *Journal of Web Semantics*, 2(1) 2004, pp. 49-79.
- [5] T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web. *Scientific American*", 284, 2001, pp. 34-43.
- [6] E. Oren, "Sindice.com: A document-oriented lookup index for open linked data," *International Journal of Metadata, Semantics and Ontologies*, 3(1), 2008, pp.37-52.
- [7] D. Calvanese, G. De Giacomo, M. Lenzerini, D. Nardi, and R. Rosati, "Description Logic Framework for Information Integration," 6th International Conference on the Principles of Knowledge Representation and Reasoning (KR-98). Italy, 1998.
- [8] M. Doerr, J. Hunter, and C. Lagoze, "Towards a Core Ontology for Information Integration," *Journal of Digital Information*, 4(1) 2003.
- [9] M. Uschold, "Creating, Integrating and Maintaining Local and Global Ontologies," *Proc. of 14th European Conf. on Artificial Intelligence (ECAI'00)*, Germany, 2000.
- [10] D. Calvanese, G. De Giacomo, D. Lembo, M. Lenzerini, R. Rosati, and M. Ruzzi, *Using OWL in Data Integration*, R. De Virgilio, F. Giunchiglia, and L. Tanca, (Eds.), *Semantic Web Information Management*, 2010 (Chapter 14, pp. 397-424), Springer-Verlag, Germany.
- [11] J. Keistead, and K.H. Van Dam, "A Survey on the Application of Conceptualisations in Energy Systems Modelling," In *Formal Ontologies Meet Industry: Proceedings of the 5th International Workshop*. Ed. P.E. Vermaas, V. Dignum, IOS Press, 2011, (pp. 50-62)
- [12] N. Shah and K. Chao, "Energy Conservation Recommending Semantic Services," In Alain Zarli (Ed.) conference CIB – W78, France, 2011.
- [13] M.J. Kofler, Ch. Reinisch, and W. Kastner, A semantic representation of energy-related information in future smart homes. In *Energy and Buildings*, 47(0), pp.169 - 179, 2012.
- [14] A. Noguero, N. Arana, and J. Martinez, "Enabling energy efficiency through device awareness using ontologies," In Alain Zarli (Ed.) conference CIB –W78, France, 2011.
- [15] K. Kabitzsch and J. Ploennings, "Ontology models and design patterns for building automation," In Alain Zarli (Ed.) conference CIB –W78, France, 2011.
- [16] J. Han, Y. Jeong, and I. Lee, "Efficient Building Energy Management System Based on Ontology, Inference Rules, and Simulation," *International Conference on Intelligent Building and Management. Proc. of CSIT vol.5 IACSIT Press*, Singapore, 2011.
- [17] A. Wagner, S. Speiser, A. Harth, "Semantic Web Technologies for a Smart Energy Grid: Requirements and Challenges," In *Proceedings of the 9th International Semantic Web Conference (ISWC2010)*, 2010.
- [18] H.M. Böhms, W. Plokker, B. Charvier, L. Madrazo, and A. Sicilia, "IntUBE energy information platform," in 8TH European Conference on Product and Process Modelling ECPPM, Ireland: CRC Press, 2010, pp. 339-344.
- [19] M. Uschold, and M. King, "Towards methodology for building ontologies," *Workshop on Basic Ontological Issues in Knowledge Sharing*, held in conjunction with IJCAI-95, Canada, 1995.
- [20] J. Contreras, and J. Martínez-Comeche, "Ontologías: ontologías y recuperación de información," 2008. Retrieved March 9, 2012, from [http://www.sedic.es/gt\\_normalizacion\\_tutorial\\_ontologias.pdf](http://www.sedic.es/gt_normalizacion_tutorial_ontologias.pdf)
- [21] B. Swartout, R. Patil, K. Knight, and T. Russ, "Toward distributed use of large-scale ontologies," In *Proceedings of the 10th Knowledge Acquisition Workshop (KAW'96)* Canada, 1996.
- [22] Y. Sure, M. Erdmann, J. Angele, S. Staab, R. Studer, and D. Wenke, "OntoEdit: Collaborative Ontology Development for the SemanticWeb," Horrocks and J. Hendler (Eds.) In *FIRST International Semantic Web Conference (ISWC 2002)* (Vol. 2342 of LNCS, pp. 221–235). Springer-Verlag Berlin, 2002.
- [23] T. Tudorache, N.F. Noy, and M.A. Musen, "Supporting collaborative ontology development in Protege," In *7th Intl. Semantic Web Conference, ISWC 2008*, Germany, 2008.
- [24] V. Presutti, A. Gangemi, S. David, G. Aguado de Cea, M.C. Suarez Figueroa, E. Montiel-Ponsoda, and M. Poveda, "Library of design patterns for collaborative development of networked ontologies," Deliverable D1.1.3, NeOn project, 2008.
- [25] A. Gangemi, and V. Presutti, *Handbook on Ontologies, chapter Ontology Design Patterns*, 2009, pp. 221. Springer.
- [26] B. Cuenca Grau, I. Horrocks, Y. Kazakov, and U. Sattler, "Modular reuse of ontologies: Theory and practice," *Journal of Artificial Intelligence Research*, 31, 2008, pp. 273-318.
- [27] K. Lomas, T. Oreszcyn, D. Shipworth, A. Wright, and A. Summerfield, "Carbon Reduction in Buildings (CaRB) - Understanding the Social and Technical Factors that Influence Energy Use in UK Homes," In *Proceedings of the RICS Annual Conference Cobra*, London, UK, 2006.
- [28] D. Calvanese, G. De Giacomo, D. Lembo, M. Lenzerini, and R. Rosati, "Tractable reasoning and efficient query answering in description logics: The DL-Lite family," *Journal of Automated Reasoning*, 39(3), 2007, pp. 385–429.
- [29] A. Poggi, D. Lembo, D. Calvanese, G. De Giacomo, M. Lenzerini, and R. Rosati, "Linking data to ontologies," *Journal on Data Semantics*, 2008, pp.133–173.
- [30] A. Gómez-Pérez, *Ontology evaluation*, In S. Staab, and R. Studer, (Eds.), *Handbook on Ontologies (1st ed.)*, 2004, (Chapter 13, pp 251-274). Springer.
- [31] L. Obrst, W. Ceusters, I. Mani, S. Ray, and B. Smith, *The evaluation of ontologies*, In C.J.O. Baker, and K.-H. Cheung, (Eds.), *Revolutionizing Knowledge Discovery in the Life Sciences*, 2007, (Chapter 7, pp. 139- 158). Springer.
- [32] A. Gangemi, C. Catenacci, M. Ciaramita, and J. Lehmann, "Ontology evaluation and validation: an integrated formal model for the quality diagnostic task," *Technical report, Laboratory of Applied Ontologies-CNR*, Italy, 2005.
- [33] J. Brank, M. Grobelnik, and D. Mladenic, "A survey of ontology evaluation techniques," *8th International Multi-Conference of the Information Society*, 2005, (pp. 166-169).
- [34] V. Corrado, S.P. Corgnati, M. Garbino, "Energy Consumption Data Collection with DATAMINE," *Energy, Climate and Indoor Comfort in Mediterranean Countries, AICARR (ITA), Climamed 2007*, 5-7 Settembre 2007, 2007, (pp. 803-816).

# Ontology Search Engines

## Overview and recommendations

Isabel Azevedo, Carlos Vaz de Carvalho  
 School of Engineering and GILT – Graphics,  
 Interaction and Learning Technologies  
 Polytechnic of Porto  
 Porto, Portugal  
 {ifp, cmc}@isep.ipp.pt

Eurico Carrapatoso  
 Faculty of Engineering  
 University of Porto  
 Porto, Portugal  
 emc@fe.up.pt

**Abstract**—The ability to find ontologies is a matter that has been receiving great attention each year, as it is time expensive to develop an ontology from the very beginning without using any work done earlier. In fact, this can be undesirable as many ontologies have been developed and their quality has been assured by different teams. However, currently ontology search engines need to be improved in order to incorporate other functionalities that are not common. This paper analyses tools that make easier the discovery of ontologies that cover some concepts, also providing some recommendations to facilitate the whole process.

**Keywords**—ontology; repositories; search engines, Semantic Web

### I. INTRODUCTION

Ontologies have been increasingly used in the context of the Semantic Web and they have been applied in different areas and projects. On the other hand, the reuse of ontologies is a step that has been proposed in many methodologies for ontology development [1].

Ontology search engine is a tool that does not require an active action from ontology developers, as it automatically searches for and indexes the ontologies they discover. Some examples are Swoogle [2], Watson [3], Sindice [4] and Falcons [5]. They vary in the metadata provided for each ontology, as there is no standard for ontology metadata and exchange.

This work began as part of a broader one that aimed at the development of an ontology reuse module that was incorporated in a repository of educational resources [6, 7] to improve their characterisation and findability, using semantics throughout these processes.

In this paper, ontology search engines are analyzed as tools that help users in the selection of useful ontologies, which are always dependent on the particular application that is envisaged. Thus, evaluation of ontologies in order to identify the suitable ones is out of the scope of this work. The study focuses in a number of aspects, and many of them are not semantic issues, but affect their usefulness.

The rest of the paper is organized as follow. In section 2 the theme of ontology search engines is expanded, and three of them are analysed. In the third section, they are compared through the results returned for some queries, exploring their

similarities, but also some differences. In this section, the results are analyzed, substantiating some suggestions. Finally, the last section provides some concluding remarks and general recommendations for the improvement of Semantic Web Search Engines.

### II. SEARCH ENGINES ANALYSIS

Ontology search engines accept queries in a format that varies from one tool to another. They usually provide results in an XML file. Their broader designation is Semantic Web Search Engines (SWSE), as they provide Semantic Web documents (SWD). However, this latter designation applies to a range of documents, besides ontologies, that fall into two categories: pure SWDs (PSWDs), and embedded SWDs (ESWDs), such as HTML documents with their associated metadata [8].

Different from other types of platforms that can be used to find suitable ontologies, such as ontology repositories, which sometimes only provide browse functionalities, ontology search engines permit a greater degree of automation.

The great amount of results provided by some SWSEs, which do not have concept or ontology search functionalities, disregard their consideration for ontology reuse based on concepts. For instance, a query on Sindice with the term ‘Table’ returns more than 800,000 results, much more than those returned by other SWSEs (see Table 1). However, a great part of them are not ontologies.

From the list previously mentioned, the more ontology-based search engines are Swoogle, Watson and Falcons, which are described in sections A, B and C, respectively.

They all allow human submission of Semantic Web documents. Also, their architectures include crawling, indexing and analyzing blocks, which are important components of any SWSE.

There are Swoogle’s statistics available at its Web site. It has indexed more than 3,800,000 Semantic Web documents and over 10,000 ontologies. It is mentioned in [9] that 11.7 million well-formed RDF/XML documents were crawled.

#### A. Swoogle

Swoogle was the first search engine dedicated to online semantic data and it remains one of the most popular SWSE.

Its development was partially supported by DARPA and NSF (National Science Foundation).

The current version of Swoogle is 3.1, which has been available since January 2006.

Swoogle’s architecture (see Fig. 1) has four major components:

- The Discovery component – It is responsible for collecting candidate URLs. It caches Semantic Web Documents. Swooglebot is the Swoogle’s Semantic Web Crawler that produces new candidates to be considered, but conventional search engines are also used for the same purpose. In addition, there is an option to submit sites and documents to be regarded;
- The Indexing component – It analyses the Semantic Web Documents (SWDs) found by the Discovery component and generates some metadata, which characterises the features associated with individual SWDs and Semantic Web Terms (SWTs), but also the relations among them. These metadata intend to improve searches;
- The Analysis component - It uses the metadata generated by the Indexing component to support ranking mechanisms;
- The Search Services module - It allows Swoogle to be used by agents and humans. It is mainly an interface component.

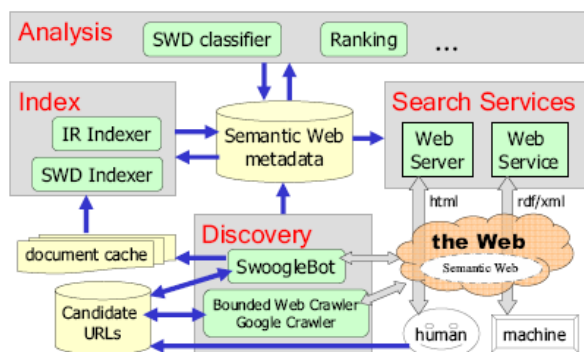


Figure 1. Swoogle 3.1 architecture (from [10]).

The Swoogle ranking method is based on the OntoRank algorithm, which is quite analogous to the PageRank algorithm (used by the Google search engine). Consider : page A which has n pages (T1, T2, ... Tn) with a link to it, a depicted in Fig. 2.

The PageRank of page A can be stated as follows [11]:

$$PR(A) = (1 - d) + d \sum_{i=1}^n \frac{PR(T_i)}{C(T_i)}$$

In the equation above, d is a normalising factor, whose value can vary from 0 to 1, C(Ti) is defined as the number o links that Ti points to. The PageRank of A (PR(A)) consider the PageRank of each Ti (PR(Ti)). OntoRank adapts the PageRank approach “to expose more ontologies which ar important to Semantic Web users” [10], using semantic relations between ontologies. Ding et al. detail the OntoRank method and compare it with the PageRank algorithm.

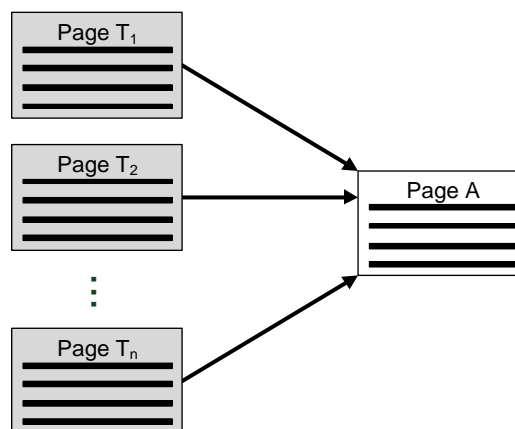


Figure 2. The idea behind PageRank algorithm (adapted from [11]).

It is noteworthy that the keywords specified in a query do not influence the ranking process, just the inclusion of a given document in the results set.

### B. Watson

The Watson development was partially supported by the NeOn [12] and the OpenKnowledge [13] projects.

The functions (collecting, analysing and querying) of the core components of Watson (see Fig. 3) do not differ significantly from those of Swoogle. These functions correspond to three different layers as follows:

- The ontology crawling and discovery layer is responsible for obtaining semantic data. Any document that cannot be parsed by Jena is disregarded as a way to guarantee that only documents that contain semantic data or ontologies are considered;
- The validation and analysis layer gathers metadata about the semantic data, which is also used for indexing purposes. In addition, semantic relations between ontologies are regarded for the retrieved ontologies (e.g., owl:imports, rdfs:seeAlso, namespaces, dereferenceable URIs) in order to detect other sources of ontologies;
- The query and navigation layer is related to the available query interfaces that allow using the Watson functionalities.

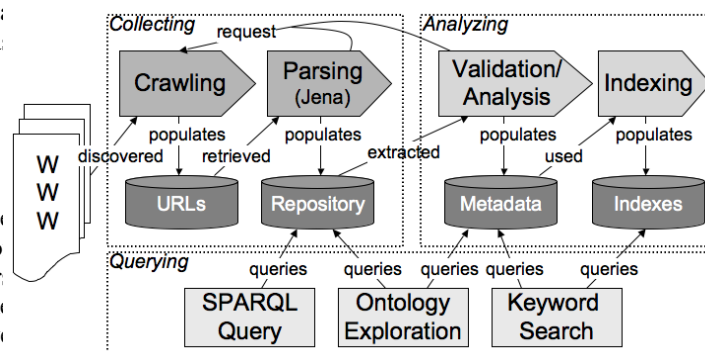


Figure 3. Watson architecture (from [3]).

In [14], it is mentioned that for ranking it is used “an initial set of measures that evaluate ontology complexity and richness”. Also, d’Aquin et al. [3] state that “the ranking mechanisms offered by Watson rely on a combination of simple, basic quality measures that are computed in the validation phase and stored along with the ontologies (i.e., structural measures, topic relevance, etc.)”. However, the exact ranking method used by Watson is unknown.

A distinctive characteristic of this SWSE in comparison to Swoogle and Falcons is the possibility to review ontologies or see how other users have reviewed it, a trend that have become popular in other areas and that led to the inclusion of user review sections in many different systems. In Watson, that functionally relies on Revyu.com, which is a web site where people can review and rate things.

### C. Falcons

The Falcons architecture has many components (see Fig. 4). The crawled documents are parsed and the URIs are then processed by the URI repository for further crawling. The quadruple (RDF triple plus the document URI) is stored. These data are processed by the meta-analysis component, which provides detailed ontological information to the metadata module. The indexer updates the next component, which is the basis of the keyword-based search functionalities. Objects are ranked in accordance to their

relevance to the query submitted and their popularity. Comprehensive information about all components is provided in [9].

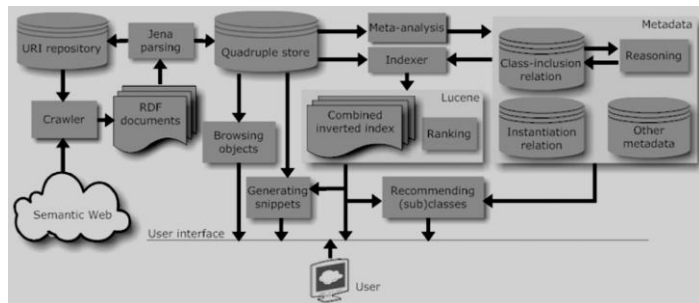


Figure 4. Falcons architecture (from [9]).

Users can use Falcons to search for objects, concepts, ontologies and documents. The object search option is useful when trying to find specific things. Concept search is useful to find classes or properties in ontologies. The option to search ontologies (see Fig. 5) provides a subset of the results returned using the option to search documents, and more metadata fields are considered.

The screenshot shows the Falcons search interface. At the top, there are tabs for 'Object', 'Concept', 'Ontology', and 'Document'. The 'Ontology' tab is selected, and the search term 'data\_model' is entered in the search box. The search button is labeled 'Search Ontologies'. Below the search box, the results are displayed as 'Ontologies 1 - 10 of 52 for your search data\_model'. The first result is 'http://open.vocab.org/terms/' with metadata: '- Metadata - 60 classes - 148 properties - Related ontologies'. Below this is a hierarchical ontology diagram for 'onto:affiliation'. The root node is 'onto:affiliation', which has several children: 'foaf:Person', 'Personal or Professional Affiliation', 'rdfs:Literal', 'Used to describe the affiliation, place ...', 'rdf:Property', and 'ext:terms'. The relationships are labeled with 'rdfs:domain', 'rdfs:label', 'rdfs:range', 'rdfs:comment', 'rdf:type', and 'rdfs:isDefinedBy'. The second result is 'http://linkedopencommerce.com/schemas/icecat/v1/' with metadata: '- Metadata - 6 classes - 29 properties - Related ontologies'. Below this is another hierarchical ontology diagram for 'onto:Product'. The root node is 'onto:Product', which has children: 'onto:Data Quality', 'onto:hasQuality', 'onto:hasShortSummaryDescription', 'onto:hasProductId', and 'onto:hasModelName'. The relationships are labeled with 'rdfs:range', 'rdfs:domain', and 'owl:equivalentClass'.

Figure 5. Ontology search with Falcons.

This visual layout of the results provided by Falcons is a distinctive characteristic of this SWSE in comparison to Watson and Swoogle. It lets users understand how the terms are included in each ontology from the results’ set.

### III. COMPARISON

Table I, Table II and Table III compare Swoogle, Watson and Falcons under the number of results using query terms

from different areas of engineering courses. As it was explained, the aim of reusing ontologies in a repository of engineering resources led to this work, and these terms in the tables characterise some engineering courses. The enumeration of the important terms corresponds to one of the recommended steps to follow when developing an ontology [15].

As Falcons does not correctly process the underscore character (see Fig. 5), even if the search strings are put in quotes, terms with this character were not considered in this search engine (in these situations “not applicable” is used in the tables). This point is expanded in the next section.

For the queries submitted to Watson, only classes and properties were considered and local names were regarded. The same options were used at Swoogle (using the `def` specifier). For Falcons the ontology search was used, but it is not possible to select exactly what is of interest, for instance, just classes and/or properties. Thus, the returned ontologies were manually inspected in order to consider just the

ontologies fulfilling the same characteristics used in the other search engines.

Table I shows the number of results for some search strings, comparing the results found by Swoogle, Watson and Falcons, but also the number of available results considering only the best ten ranked documents in the results set. Swoogle and Watson do not cope with different writing styles. For instance, the results found for ‘DataModel’ do not include those returned for ‘Data\_model’.

Table II provides the results obtained when some concepts from a Statistics course were considered. Table III shows the results found using some concepts from a Chemical course.

TABLE I. NUMBER OF RESULTS FOR SOME DATABASE CONCEPTS EXPRESSED IN DIFFERENT WAYS

Search string	Swoogle		Watson		Falcons	
	Number of results	Number of available results (Top Ten)	Number of results	Number of available results (Top Ten)	Number of results	Number of available results (Top Ten)
‘Distributed Database’	0	0	0	0	Not applicable	Not applicable
‘DistributedDatabase’	2	2	0	0	0	0
‘Distributed Databases’	3	2	0	0	Not applicable	Not applicable
‘DistributedDatabases’	0	0	0	0	0	0
‘Data_model’	13	5	1	0	Not applicable	Not applicable
‘DataModel’	11	7	1	1	0	0
‘DataModels’	0	0	0	0	0	0
‘Data models’	3	2	1	1	Not applicable	Not applicable
‘Table’	816	6	30	9	25	7
‘Tables’	77	9	4	1	1	1

TABLE II. NUMBER OF RESULTS FOR SOME STATISTICS CONCEPTS EXPRESSED IN DIFFERENT WAYS

Search string	Swoogle		Watson		Falcons	
	Number of results	Number of available results (Top Ten)	Number of results	Number of available results (Top Ten)	Number of results	Number of available results (Top Ten)
‘Sampling’	225	9	5	3	13	10
‘Samplings’	0	0	0	0	0	0
‘Probability’	232	6	10	6	6	5
‘Probabilities’	2	0	0	0	0	0
‘Linear regression’	1	1	0	0	Not applicable	Not applicable
‘LinearRegression’	10	2	2	0	0	0
‘LinearRegressions’	0	0	0	0	0	0
‘Linear regressions’	0	0	0	0	Not applicable	Not applicable
‘Probability distribution’	1	1	0	0	Not applicable	Not applicable
‘ProbabilityDistribution’	1	0	0	0	0	0
‘ProbabilityDistributions’	0	0	0	0	0	0
‘Probability distributions’	0	0	0	0	Not applicable	Not applicable



TABLE III. NUMBER OF RESULTS FOR SOME CHEMICAL CONCEPTS EXPRESSED IN DIFFERENT WAYS

Search string	Swoogle		Watson		Falcons	
	Number of results	Number of available results (Top Ten)	Number of results	Number of available results (Top Ten)	Number of results	Number of available results (Top Ten)
'Periodic table'	0	0	1	0	Not applicable	Not applicable
'PeriodicTable'	1	0	0	0	0	0
'PeriodicTables'	0	0	0	0	0	0
'Periodic tables'	0	0	0	0	Not applicable	Not applicable
'solution'	400	6	13	9	10	7
'solutions'	34	6	5	1	3	3
'Acid'	621	7	31	5	23	9
'Acids'	50	9	3	1	0	0
'Base'	1,625	4	27	6	79	5
'Bases'	48	5	3	1	0	0

From the experiments here documented it was found that:

- Search strings that can be considered very generic, such as 'Base', 'Table' or 'Solution' return many results. However, a great number of those results are not really for the envisaged area. For example, the results returned for the search string 'Base', included ontologies with classes for baseball, database, and space subjects, among others. Obviously, it does not mean that search engines did not function correctly, but if users can supply many keywords of possible interest (using the OR operator), it might be possible to consider each of them differently at least in the ranking process. Swoogle is the only one to allow the use of the logical operator OR, but each term used does not affect how the others are regarded;
- Although the common conventions of using the singular form in concept names and the CamelCase style to write compound words or phrases, followed by the W3C itself, these are not universally followed. The use of separator (underscore or no character in accordance with CamelCase naming convention) and singular or plural nominal word form in the submission of queries to SWSEs lead to different sets of results, which are not enclosed in the others;
- It was impossible to analyse all the results returned, but generally there is not an overlap in the top ten results provided by Swoogle, Watson and Falcons. It can be a result from the use of different ranking methods, but for some search strings, one SWSE provided no results, while the others returned. Although it is declared in that Watson uses a specialised crawler for Swoogle, it does not seem that it has been active.

Ontology versioning is "the ability to handle changes in ontologies by creating and managing different variants of it", and this subject is deeply analysed in [16]. Although Watson has some version control mechanisms and it is "able to detect some form of duplication of ontologies" [17], the same version of a given ontology can be returned by Watson, or even different versions of the same ontology. For instance, for the search string 'Base' the results returned by Watson

include some ontologies that correspond to different versions of the same file.

KANNEL is a framework for detecting and managing ontology relations for large ontology repositories [18]. It was used in conjunction with Watson, with interesting results [19]. It was noticed an improvement in the efficiency of search engines tasks, but also, in the satisfaction of the users involved in these activities. However, the use of KANNEL is not integrated in Watson at this moment.

Version detection problems were also identified in the results provided by Swoogle (see Fig. 6) and Falcons (see Fig. 7).

The screenshot shows the Swoogle search interface. The search bar contains 'def.distributed\_databases'. Below the search bar, there is a blue banner with the text 'list ontologies matching ontology search' and '1 - 3 of total 3 results for def.distributed\_databases in 0'. Below this, there are three search results listed with their URLs and metadata:

- <http://what.csc.villanova.edu/twiki/pub/Main/OWLFFileInformation/ComputingOntology-George2.rdf.xml.owl>  
[DEF]\_Databases, Distributed\_File\_Systems, Distributed\_Memories, Distributed\_Models, Distributed\_SemanticWebDocument, RDFXML, 2008-06-23, 613K, ontoRatio(0.99), metadata, cached
- <http://what.csc.villanova.edu/twiki/pub/Main/OWLFFileInformation/28Jul09.owl>  
[DEF]\_Concurrency\_Control, Distributed\_Data\_Storage, Distributed\_Databases, Distributed\_File\_Systems, SemanticWebDocument, RDFXML, 2009-07-29, 613K, ontoRatio(0.99), metadata, cached
- <http://cse.unl.edu/~scotch/SWont/acmCCS.owl>  
[DESC] numbers 4th level terms, such as "D.1.3.1 Distributed programming". The ACM CCS does not do [DEF]\_data\_structures, Distributed\_databases\_C.2.4, Distributed\_databases\_H.2.4, Distributed\_debugging, SemanticWebDocument, RDFXML, 2006-06-27, 272K, ontoRatio(0.98), metadata, cached

Figure 6. Top results found by Swoogle search using 'distributed\_databases' as search string (partial view).

In Fig. 6, the first two ontologies correspond to different versions of the same ontology. The older one appears before and at first perhaps because it had been much more used than the newer one, which affects their ranks. Detection of versioning relationships between documents from the Swoogle's Semantic Web archive was described in [20] and perhaps version control information will start to be considered.

The number of results is not the only criterion to be considered, but it is important as it should be easier to find appropriate ontologies in a large set. However, the results were analysed to determine by sampling if the top ten results were relevant, and they were. For instance, one of the results provided by Swoogle using 'Distributed\_Databases' as

search string is the computing ontology [21]. However, future studies need to examine the results in details to allow a further comparison at this level.

Some other aspects that were studied were:

- The existence of a limit number of queries accepted;
- The existence of multiple options to sort the results;
- The metadata provided by each returned ontology;
- The possibility of specifying many terms, all to be considered (use of logical operator AND);
- The possibility of specifying many terms to be considered alternatively (use of logical operator OR);
- The ability to dynamically discover semantic data depends on available APIs to access the semantic resources collected by Semantic Web search engines.

These points and others already discussed, as well as some statistical information are summarised in Table IV.

**distributed database system** - type of object, Class

• type: type of object  
 • label: **distributed database system**  
 • sameAs: **DistributedDatabaseSystem**  
 • comment: A specialization of <http://sw.opencyc.org/concept/Mx4rwBcktpwEbGdrcNEInstance0...>  
 • subClassOf: [Mx4rwBcktpwEbGdrcN5Y29ycA](http://sw.opencyc.org/concept/Mx4rwBcktpwEbGdrcN5Y29ycA)  
 • label: **DistributedDatabaseSystem**  
 • Pretty String: **distributed database systems**  
 • type: Class  
 • sameAs: **distributed database system**  
 • has sub Class: NIS  
<http://sw.opencyc.org/concept/Mx4rHWvdqEjkEdaKIQACs0uFOQ>

**DistributedDatabaseSystem** - Class, ObjectType, SubjectConcept  
<http://umbel.org/umbel/sc/DistributedDatabaseSystem>

**distributed database system** - type of object, Class

• type: type of object  
 • label: **distributed database system**  
 • sameAs: **distributed database system**  
 • comment: A specialization of <http://sw.opencyc.org/2008/06/10/concept/Mx4rwBcktpwEbGdrcN5Y29ycA> class="cyc\_term">DynamicIndexedInfoSource</a>. Each...  
 • subClassOf: [dynamic indexed info source](http://sw.opencyc.org/2008/06/10/concept/Mx4rHWvdqEjkEdaKIQACs0uFOQ)  
 • label: **DistributedDatabaseSystem**  
 • Pretty String: **distributed database systems**  
 • type: Class  
 • sameAs: **DistributedDatabaseSystem**  
 • sameAs: **distributed database system**  
<http://sw.opencyc.org/2008/06/10/concept/Mx4rHWvdqEjkEdaKIQACs0uFOQ>

Figure 7. Top results found by Falcons using ‘distributed\_database’ as search string (partial view).

Some common problems were detected. First, in the results set there was a number of ontologies that were no longer available. For instance, one of the three returned ontologies for the query ‘distributed\_databases’ has been unavailable (the first one - see Fig. 6) for more than two years. In that case it is known that this ontology has a newer version (whose URI is <http://what.csc.villanova.edu/twiki/pub/Main/OwlFileInformation/28Jul09.owl>). Thus, it does not seem that Swoogle has an efficient version control mechanism and, as stated before, Watson suffers from the same problem. However, due to recent versioning developments and experiments that used ontologies indexed by them, it is envisaged the changes will take place soon.

TABLE IV. SWOOGLE, WATSON AND FALCONS – A COMPARISON

Characteristic	Swoogle	Watson	Falcons
Available APIs	Yes	Yes	Yes <sup>1</sup>
Unlimited number of queries	No	Yes	Yes <sup>2</sup>
Multiple sorting possibilities	Yes	No	No
Provision of rich ontology metadata	Yes	Yes	Yes
Use of OR to specify possible terms	Yes	No	No
Use of AND to specify all terms	Yes	Yes	Yes
Possibility to see how other users considered the ontology or rated it	No	Yes	No
Number of crawled SWDs	>3,000,000 <sup>3</sup>	-	>11,700,000 <sup>4</sup>

<sup>1</sup>The RESTful APIs are described at <http://ws.nju.edu.cn/falcons/api/index.jsp>, but they were unavailable during the study here documented.

<sup>2</sup>It was not possible to test through the API.

<sup>3</sup>Data from July 2012.

<sup>4</sup>Data from August 2008 [9].

Another point to be improved in SWSEs is the use of wildcards, not their acceptance but how they are treated and the results of their usage. For instance, submitting a query string to Swoogle like “data\*model” provides the same results as a query string like ‘data\_model’, and they not include the results obtained with a query string like ‘datamodel’. Watson has a similar problem. A query submitted to Watson specifying ‘data\*model’ returns the same results returned by a query like “datamodel”, not including the results returned by a query like “data\_model”. Falcons also accepts wildcards but their effect is the same obtained by the use of the whitespace character or the logical operator AND.

Besides these aspects, the automatic detection of ontology relations, others than versioning, can simplify the results’ analysis. For instance, the automatic detection of the inclusion of concepts of one ontology in another one can be a useful functionality, but not yet common. More powerful indexing schemas to deal with similarity and relatedness between concepts at different levels should also be considered. For instance, a search using a query term such as ‘relational\_model’ will fail to provide ontologies with the concept ‘relational\_data\_model’, but they have a similarity score near 0.76 using Levenshtein distance [22].

Google Knowledge Graph [23] can be seen as preliminary step in order to provide structured results for keyword-based searches submitted to Google search engine, considering that there are “things, not strings”. Such approach in Ontology Search Engines could also help them regard, for instance, that table can be a piece of furniture, but also something meaningful in database area.

#### IV. CONCLUSION

The comparison of ontology search engines showed that lexical variations, such as the use of separators or not in the query terms and their specification in singular/plural form affect the results. Thus, although SWSEs have to be able to deal with diverse writing styles, currently they are not.

In addition, version control and object coreferencing detection are important for many applications, and also in ontology search engines, as it was discussed in the previous sections. However, at this moment Semantic Web Search Engines do not identify version ontology versions or do not show users this kind of information when they are trying to find ontologies. Changes are expected soon to Swoogle and Watson, as it was discussed.

Finally, a federated query service able to submit queries to multiple sources and a robust but flexible ranking strategy can benefit ontology developers as there is not a considerable overlap among results returned from different ontology search engines.

In addition, the data were collected from August 2010 to July 2011, it will be useful to consider how the results will vary from those reported here in the future, which can provide some insights into the way SWSEs crawl the Web and find new ontologies.

#### ACKNOWLEDGMENT

This work was supported in part by the PhD grants SFRH/BD/43437/2008 (Portuguese Foundation for Science and Technology) and SFRH/PROTEC/49362/2009 (Ministry of Science, Technology and Higher Education), and the project CASPOE - Characterization of Semantics and Pragmatics of Educational Objects (PTDC/EIA/65387/2006), with funding from the Portuguese Foundation for Science and Technology.

#### REFERENCES

- [1] K. Siorpaes and E. Simperl, "Human intelligence in the process of semantic content creation," *World Wide Web*, vol. 13, pp. 33–59, 2010.
- [2] L. Ding, T. Finin, A. Joshi, R. Pan, R. S. Cost, Y. Peng, P. Reddivari, V. Doshi and J. Sachs, "Swoogle: A search and metadata engine for the semantic web," *Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management*, 2004.
- [3] M. d'Aquin, M. Sabou, M. Dzbor, C. Baldassarre, L. Gridinoc, S. Angeletou and E. Motta, "Watson: A gateway for the semantic web," *European Semantic Web Conference (ESWC 2007)*, Innsbruck, Austria, 2007.
- [4] E. Oren, R. Delbru, M. Catasta, R. Cyganiak, H. Stenzhorn and G. Tummarello, "Sindice.Com: A document-oriented lookup index for open linked data," *International Journal of Metadata, Semantics and Ontologies*, vol. 3, 2008.
- [5] G. Cheng, W. Ge and Y. Qu, "Falcons: Searching and browsing entities on the semantic web," *Proceedings of the 17th International Conference on World Wide Web*, Beijing, China, 2008.
- [6] I. Azevedo, R. Seica, A. Ortiz, E. Carrapatoso and C. V. d. Carvalho, "A semantic approach for learning objects repositories with knowledge reuse," *EKAW 2010, Lecture Notes in Artificial Intelligence*, Lisbon, Portugal, 2010, pp. 580–589.
- [7] I. Azevedo, "Semantic and pragmatic characterisation of learning objects," *Informatics*. Porto, Portugal, Faculdade de Engenharia da Universidade do Porto. PhD thesis: 317 2012.
- [8] L. Ding and T. Finin, "Characterizing the semantic web on the web," *The 5th International Semantic Web Conference*, Athens, GA, USA, 2006.
- [9] G. Cheng and Y. Qu, "Searching linked objects with falcons: Approach, implementation and evaluation," *International Journal on Semantic Web and Information Systems*, vol. 5, pp. 50-71, July-September 2009 2009.
- [10] L. Ding, R. Pan, T. Finin, A. Joshi, Y. Peng and P. Kolari, "Finding and ranking knowledge on the semantic web," *The 4th International Semantic Web Conference*, Galway, Ireland, 2005.
- [11] L. Page, S. Brin, R. Motwani and T. Winograd, "The pagerank citation ranking: Bringing order to the web," *November 1999 1999*.
- [12] "Neon project," <http://www.neon-project.org>, retrieved: June, 2012.
- [13] "Openknowledge project," <http://www.openk.org>, retrieved: June, 2012.
- [14] M. d'Aquin, E. Motta, M. Sabou, S. Angeletou, L. Gridinoc, V. Lopez and D. Guidi, "Toward a new generation of semantic web applications," *IEEE Intelligent Systems*, vol. 23, pp. 20 - 28, May-June 2008 2008.
- [15] N. Noy and D. L. McGuinness, "Ontology development 101: A guide to creating your first ontology. Technical report ksl-01-05," *Stanford Medical Informatics*, Stanford 2001.
- [16] M. Klein and D. Fensel, "Ontology versioning on the semantic web," *The International Semantic Web Working Symposium (SWWS)*, Stanford, USA, 2001.
- [17] M. d'Aquin, C. Baldassarre, L. Gridinoc, M. Sabou, S. Angeletou and E. Motta, "Watson: Supporting next generation semantic web applications," *WWW/Internet 2007 Conference*, Villareal, Spain, 2007.
- [18] C. Allosca, "Automatic identification of ontology versions using machine learning techniques," in *Lecture notes in computer science*. vol. 6643, 2011, pp. 352-366.
- [19] C. Allosca, M. d'Aquin and E. Motta, "Impact of using relationships between ontologies to enhance the ontology search results," in *Lecture notes in computer science*. vol. 7295, 2012, pp. 453-468.
- [20] K. Viswanathan and T. Finin, "Text based similarity metrics and delta for semantic web graphs," *9th International Semantic Web Conference (Poster Session)*, Shanghai, China, 2010.
- [21] R. Kamali, L. Cassel and R. LeBlanc, "Keeping family of computing related disciplines together," *The 5th conference on Information Technology Education*, Salt Lake City, UT, USA, 2004, pp. 241 - 243.
- [22] V. Levenshtein, "Binary codes for correcting deletions, insertions, and reversals," *Doklady Akademii Nauk SSSR*, vol. 163, pp. 845–848, 1965.
- [23] Google Official Blog, "Introducing the knowledge graph: Things, not strings," <http://googleblog.blogspot.pt/2012/05/introducing-knowledge-graph-things-not.html>, 2012, retrieved: June, 2012.

# Semantic Supply Chain Management

Katalin Ternai

Department of Information Systems  
Corvinus University of Budapest  
1093 Budapest, Fővám tér 13-15., HUNGARY  
ternai@informatika.uni-corvinus.hu

Ildikó Szabó

Department of Information Systems  
Corvinus University of Budapest  
1093 Budapest, Fővám tér 13-15., HUNGARY  
iszabo@informatika.uni-corvinus.hu

**Abstract**—Small and medium enterprises formed into a supply chain are working in a multi-cultural and multilingual environment. The eBEST platform equips these enterprises and their associations with state-of-the-art software tools for ecosystem-wide business collaboration. To ensure effective collaboration through the whole supply chain visible communication, traceable workflow and process management are required by users. In the eBEST project the semantic interoperability was realized by ontological approach and is tested by pilots in nowadays.

**Keywords**—supply chain management; ontology based communication and workflow management

## I. INTRODUCTION

The efficiencies of Supply Chain Management (SCM) are often impaired by inconsistent exchange and sharing of information semantics among supply chain partners. Semantics-based technologies, especially ontologies have key role in Semantic Supply Chain Management - they are responsible for domain conceptualization, structuring knowledge embedded in business processes. The standardized ontologies for Supply Chain Management enhance the interoperability between the various Supply Chain Management systems. They also serve as a basis for building more specialized ontologies, for example, process ontology for building workflow models. To use ontologies in the development of Supply Chain Management systems results reusable, easy to integrate applications.

This paper aims at presenting an ontology-based SCM platform within the eBEST project (Empowering Business Ecosystems of Small Service Enterprises to Face the Economic Crisis) [8]. The project deals with equipping SMEs and SME associations with state-of-the-art software tools for ecosystem-wide business collaboration. The purpose of this paper is to discuss how ontologies may be used to raise interoperability and shared understanding in inter-organisational processes. Ontologies also play decisive role in turning process models into working software, providing a visual and textual representation of the processes, data, information, resources, collaborations and other measurements.

The paper will be structured as follows: In Section 2, theoretical overview about ontologies and Supply Chain Management is described. In Section 3, the SCM platform is presented in the light of supported ontologies. Finally, conclusion and future work are shown.

## II. THEORETICAL BACKGROUND

### A. Supply chain management

Supply chain management (SCM) is a rather practical-oriented than theoretical domain. Due to the multidisciplinary origin and the evolution way of this domain there isn't universal supply chain management definition. Mentzer et al. [16], Tan [19] and Cooper et al. [5] consider SCM as a management philosophy, whilst the next definition given by Council of Supply Chain Management Professionals emphasizes rather the activities and processes of SCM. This definition reflects better the eBEST approach than the others.

"Supply Chain Management encompasses the planning and management of all activities involved in sourcing and procurement, conversion, and all logistics management activities. Importantly, it also includes coordination and collaboration with channel partners, which can be suppliers, intermediaries, third-party service providers, and customers [6]."

Based on this definition we can distinguish two main groups of supply chain activities which are related to each other: planning and management of all activities; coordination and collaboration with channel partners.

There are several approaches to present SCM processes [7][18], etc. The most widely accepted framework for evaluating and comparing supply chain activities and their performance is the Supply Chain Operations Reference SCOR® model [18]. It is built on five primary management processes of Plan, Source, Make, Deliver and Return

The planning process provides companies a strategy for managing all the resources to satisfy the actual or forecasted demand with products or services.

The sourcing strategy is based on guarantee material availability in appropriate quantities at the right time for both internal purposes and for sales and distribution. Considering stocks and instruments providing production capacities companies can get an extent list up about the materials and tools, and they can start to choose the suppliers to deliver these goods.

The manufacturing flow process includes all activities which are responsible for making products and establishing manufacturing flexibility required by serving target markets.

The activities of deliver process in SCOR model are demand management, order management and warehouse management.

In the return process, the supply chain planners have to create and manage a flexible network on both supplier and customer side in order to handle the defective, excess products or recyclable/dangerous garbage.

Due to the complexity of supply chain, we can distinguish demand-side collaboration, supply-side collaboration or overall synchronization [16].

Barratt showed that “collaborative” culture is one of the major supporting elements of collaboration. It consists of the following elements: trust, mutuality, information exchange and openness/communication [1].

### B. Ontology

In our interpretation, semantic supply chain management means the support of main activities of SCM – managing business processes, collaboration and coordination with channel partners – by semantic technologies.

Ontologies have key role in Semantic Supply Chain Management; they are responsible for domain conceptualization, structuring knowledge embedded in business processes. Considering the scope of ontology-based applications (for example cooperative information systems, information retrieval, knowledge management, system analysis and design, etc.) we can distinguish the next three categories of ontology applications which are related to the above-mentioned SCM activities:

- Communication: between humans - informal, unambiguous ontology can be used for these purposes.
- Cooperation: between systems - it means translation among different tools, paradigms, languages and software instruments. In this case the ontology is the basis of data change.
- System design and analysis - the ontology can support the analysis and design of software systems with submitting a conceptual description.

The ontology approach has several advantages:

- Reusability: the ontology is the root of the formal description and coding of the most important entities, attributes, process and its internal relations. This formal description provides (maybe through automated translation procedure) the reusability and the common or shared use inside the given software.
- Knowledge acquisition: speed and reliability of knowledge acquisition can be accelerated, if ontology can be used for analysis or knowledge base creation.
- Reliability: automatic verification of consistency can be assured by the formal description.
- Specification: ontology enables the analysis of requirements and the determination of information systems specification.
- Standardization: top-level ontologies can be used well in different situations. New types of task and application

ontologies can be derived from these top-level models with specialization.

Ontologies have key role in semantic web [15]. More authors draw parallels between ontologies and the role of XML in data representation. Ontology describes not only data, but also the regularity of connection among data. Probably the most important description language of semantic web is the OWL (Web Ontology Language)[17] preferred by W3C[22].

In the process ontology, the goal is to be able to apply machine reasoning for the translation between the business process and executable process spheres, in particular for the discovery of processes, process fragments and for process composition [2]. Within process ontology two types of ontologies are utilized: domain ontologies and process specific ontologies. Domain ontologies support process modelling in terms of describing the actual data that is processed during process execution. Via this semantic description of the data, business process analysis can be semantically enhanced since the semantic meaning of the data is preserved during all phases of the process lifecycle [13].

### C. Supply Chain Management supported by ontologies

In SCM, the ontology development is to facilitate effective information change and knowledge sharing among collaborative supply chain partners [21], to model operational processes of supply chains and to capture and organize knowledge necessary for managing these workflows [3]. In agent-based approach ontologies can serve as a knowledge base to manage the agent behaviour through a conversation [12] or a base of workflow process modelling to facilitate customer service [21].

In the eBEST approach, ontologies serve as a standardized base to exchange data and to model workflow process. But the role of our ontologies firstly is to avoid communication problems like linguistic and translating problems, and to support the cooperation of companies community as an ecosystem by providing tools for modelling collaboration processes beyond built-in operational processes.

## III. THE E-BEST PLATFORM

The main objective of the eBEST project was the introduction of new collaboration practices in ecosystems of SMEs belonging to different industrial sectors. So, it aims at providing easily accessible ICT applications and services to enable community building, SME network constitution, and SME network operation for the network lead and its members. The eBEST platform supports the operation of digital business ecosystems. These are clusters of companies, small companies in particular, that collaborate within an operational context. These companies collaborate with each other in a multi-cultural and multilingual environment. In the first phase of the project, the collaboration habits and needs in the business ecosystems were analyzed and the

fundamental functional and non-functional requirements were determined [9]:

- Help shaping the ecosystem: to support the work of a group of Ecosystem Architects who are responsible for discovering, exploring and shaping interesting potential ecosystems.
- Find out collaboration units: to grasp business opportunities proposed by costumers and to participate or promote the definition of the relative distributed workflows.
- Increase company visibility: to form a supply chain where a company can play a supplier and a costumer role simultaneously implies the creation of visible profiles and offers provided by companies.
- Ease communication between companies: to provide a private workspace to manage documents and clearly defined concepts and terms both in offer and demand catalogue too in order to facilitate the exchange a variety of documents with the minimum need of human intervention.
- Support network planning: to find out the most convenient network configuration for each of the services to perform. The network planning algorithm assumes that a service is associated to a process model, the process is composed by activities and each activity can either be executed internally or assigned to candidate suppliers.
- Support internal resource scheduling: to provide easy but effective scheduling functions to assure that the tasks are allocated optimally to the available resources and to arrange their production by automatically optimise the usage of these resources. It is necessary to assure that exceptions are efficiently and timely handled to damp down perturbations.
- Semantic interoperability: to create a semantic repository where terms, their definitions and linguistic translations are collected in order to facilitate document transformation and contents translation.
- Trust building: customers and suppliers must feel confident that their transactions will not be intercepted or modified, that both sellers and buyers own the identity they claim, and that the transaction mechanisms are available, secure and legal.
- Technical issues: the platform must be perceived by companies as ready-to-use solution, and accessible by a simple web browser etc..

In the point of architectural view, the E-best platform proposes three interlinked software environments (presented by Figure 1) specifically conceived for networked small companies, supported by advanced suite of ICT services and applications [8]:

- Ecosystem shaping - offers the functions that every association can employ to promote the constitution and

characterize new company clusters, and improve their image over time, out of the ecosystem of its members.

- Collaboration framework - offers the functions that the single cluster can use to seize business opportunities, possibly identified by the association, and to prepare itself by designing the corresponding distributed processes.
- Operational framework - provides customers, companies and company clusters with a suite of operational functions enabling them to communicate, plan the distributed processes and schedule the internal resources.

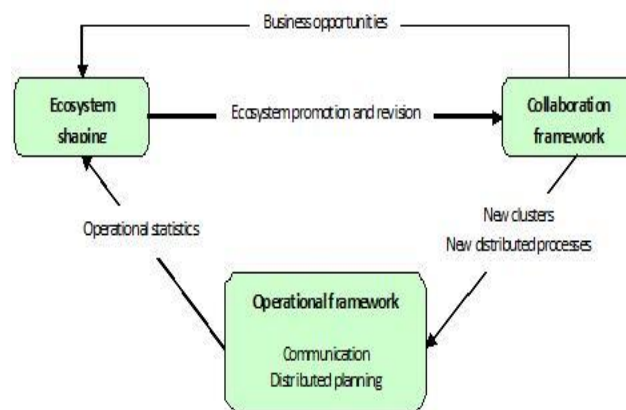


Figure 1 The eBEST platform [10]

In the point of operational view, these requirements demand the creation of a transparent communication framework and a traceable collaboration framework including workflow management and process management tools. This transparency of documents and processes is ensured by ontological approach.

#### A. Communication between ecosystems

To ensure an effective communication among the players of supply chain it is necessary to provide visible company profiles (in Company node), overall view about ecosystems (in Ecosystem node), documents with semantic contents related to the process elements (in Semantic node).

The following picture illustrates a general eBest environment with different ecosystems and configurations [11].

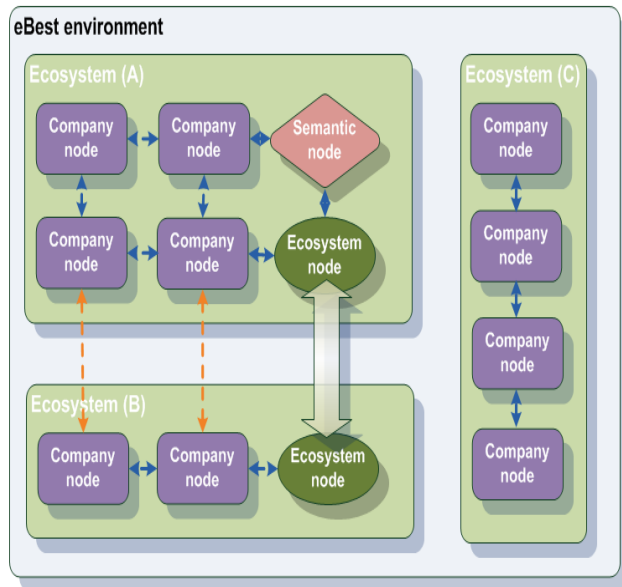


Figure 2. eBest environment

The functions of these nodes are the next ones:

- Semantic node functions. The eBest platform is asked to facilitate the business communications within an European context, hence it is necessary to pay a relevant attention to the linguistic and semantic issues.
- Ecosystem node functions. When many companies cope with each other, it is important to provide an unified view on their profiles and offers. Hence, this set of functions is mainly addressed to improve the ecosystem visibility and its capability of attracting potential new customers.
- Company node functions. This set of functions is conceived to let the companies interact with each other within an ecosystem.

*Semantic node functions*

The Semantic Node architecture is depicted by Figure 3 [11].

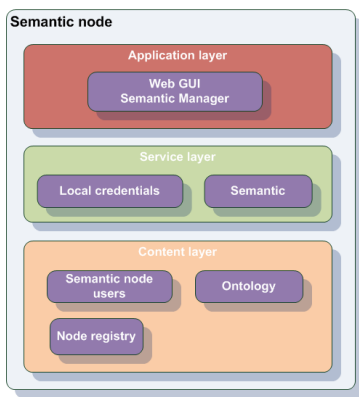


Figure 3. Semantic Node architecture

*Content Layer* handles users and roles having access to the semantic node, ontology information shared by all the

nodes within the same ecosystem: business documents structure, terms translations, service taxonomy and information of nodes belonging to the same ecosystem.

The primary goal of the eBest semantic repository is to provide a unique location where terms, definitions and linguistic translations are collected. Ontologies represent the building blocks of the eBest solution since static contents shared among the eBest nodes become defined. The ontologies information must be prepared before the eBest platform is deployed and any further change actually represents a platform update. Hence, ontologies shall be considered almost static objects, whose changes must be undertaken with care. eBest is based on three ontologies:

- Business document ontology. It defines which business documents are managed within the ecosystem and their specific data structures.
- Localization ontology. It defines the labels, with their relevant translations, addressed to feed the eBest application user interfaces.
- Offer taxonomy. It defines a hierarchy of terms conceived for the company offer classification. The offer taxonomy should take into account the trade-off between a wide taxonomy (very detailed classification with sparse samples) and a reduced taxonomy (generic classification with dense samples). Each taxonomy term is composed by a label and a definition.

Differently from ontologies, the catalogue vocabulary is a repository of terms that companies dynamically feed with the terms and definitions used to characterize their offers.

*Service Layer* provides user authentication and authorization function, terms translation management functions, the access to the catalogue vocabulary terms, published by companies, by means of distributed queries, the access to ontology items, and the download of ontology modules for being embedded in the eBest binary distributions.

*Application Layer* deals with translating terms, updating translations and browsing the ontology structures and term lists.

This node is responsible for feeding all platforms with semantic documents to foster effective and transparent communication, and to realize semantic interoperability.

*B. Communication through the operational process*

In the operational framework, the ontological support of the processes is to facilitate the exchange variety of documents through the supply chain. The next table consists of the main activities of the supply chain management and ontology elements related to them [14]. These elements are provided by the Semantic Node.

TABLE I. ONTOLOGY CONCEPTS RELATED TO SUPPLY CHAIN MANAGEMENT PROCESSES

<i>Process description</i>	<i>Ontology concepts</i>
<b>Network planning</b> It is handled in ecosystem shaping.	Company
<b>Sourcing</b> It is based catalogues which are divided offer and demand parts. Besides of the Catalogue of provided/requested services or goods, where the product families can be described by terms, attachments and parameters. Company profile specifies additional metadata about the company which can be used to filter out results of the search of prospective customers.	Company, Service, Parameter, Term, Attachment
<b>Ordering</b> Negotiation phase is implemented using Quotation concept, which is used to initiate ordering by Customer. Quotation consists of Configurations, which specifies the properties of the requested services or goods and conditions for payment and delivery time. Seller can simply confirm receipt of the Quotation and Customer can confirm unchanged Quotation by Purchase Order concept. Proposed changes to Quotation by the Seller are accomplished by direct modification of the Quotation where the Seller can indicate changed conditions or parameters.	Quotation, Order, Configuration, Parameter, Term, Attachment
<b>Fulfilment</b> In the current proposal, only proactive Despatch advice send from the Seller to the Customer is supported. Additionally Seller can send Order Progress concept where communicates order status and progress.	Dispatch Advice
<b>Billing</b> All invoicing types (prepayment invoice, pro-forma invoice and normal invoice) are implemented using the same Invoice concept.	Invoice
<b>Maintenance</b> Technical assistance or maintenance contracts are service contracts whose main feature is continuity over time. These contracts establishes cases, conditions, methods, times and costs of the activities that the Supplier is engaged to perform for guaranteeing the correct operation of a certain product or service as well as its repair or recovery after a fault.	Contract, Intervention call, Intervention report

These ontology concepts have a crucial role in their related processes. They carry unambiguous information for executing the ecosystem shaping and operational processes without any perturbations. Therefore, they ensure the visibility of companies and catalogues, and the transparent communication and cooperation. So they contribute the trust building among companies.

### C. Collaboration framework

The Collaboration Framework of the eBEST architecture serves as a shared environment for business ecosystem members, where member companies can cooperate on collective activity like tender management, event organisation, marketing and other areas of common interest. In order to fulfil this requirement, the ecosystems need a workflow management solution that can be freely customized for their specific needs. We have developed a process definition scheme and the actual software implementation for automated generation of collaboration workflow management [20].

The objective of this section is to present the technological innovation achieved during the development of the eBEST Collaboration Framework responsible for realizing the common environment for SMEs connected through a business cluster (business ecosystem) [20]. The theoretical focus is given to the experience derived from the creation of conceptual models, process model representation by ontology definition, and turning the process ontology output into workflow supporting application. The eBEST project ensured a practical framework for this transition, but the overall goal is to conceive a general implementation pattern. The building blocks of the proposed architecture are well-founded ideas, the innovation lies in utilizing them in a coherent theoretical and architectural framework.

We have defined an ontology based annotation scheme for planning collaborative business processes at a conceptual level that can be designed by non-IT personnel [20]. The annotation scheme is an extension of the OWL (Web Ontology Language), that determines the structure and attributes of the workflow processes defined by business process modelling.



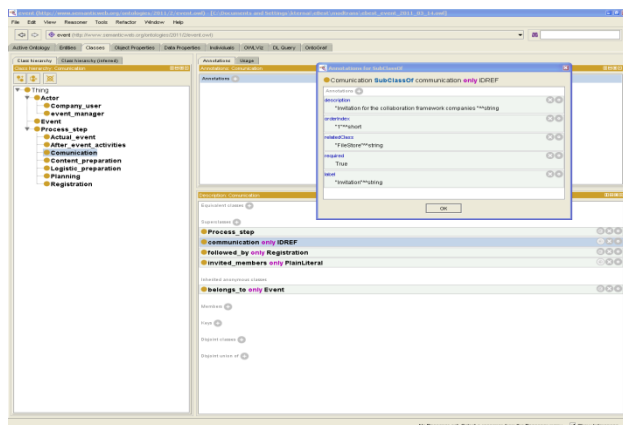


Figure 4. The event organisation process ontology in Protégé

We have developed the application framework which is able to interpret our workflow model and automatically generate the working software instance for workflow support [20].

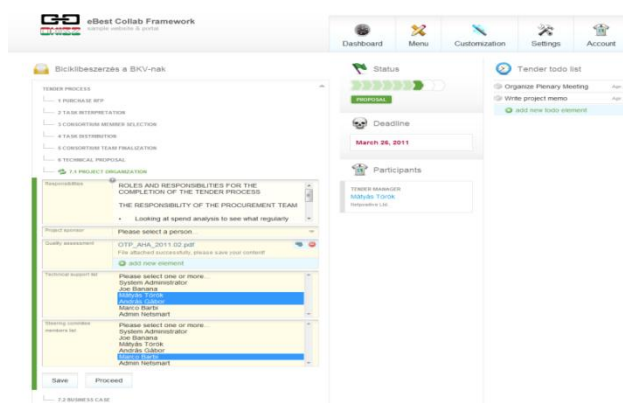


Figure 5. eBEST Collaboration Framework workflow interface [4]

We have validated the processes in real life circumstances with clusters utilizing the eBEST framework. The general idea is that business clusters themselves are empowered to design processes for their specific needs.

*D. Validation of the approach*

The performance of pilot experiments was foreseen to validate the eBEST approach. The relative software services with the twofold objective to demonstrate their effectiveness and collect hints for their best deployment to a wide population of the companies. The following table summarises the demo scenarios that were subject to experiment and grouped according to the eBEST components aiming to test.

TABLE II. PILOT SCENARIOS

Pilot	Ecosystem Shaping	Collaboration Portal	Operational Platform
SIRRIS, Belgium	X		
IDM-GAIA, Spain	X		
CCI KILKIS, Greece		X	
OKISZ, Hungary		X	
International		X	
TEL&CO, Italy			X
Fashion Contract, Italy			X
CEDEM, Italy			X
SCCI, Slovakia			X

The results from the pilot experiments are showed by the following although initial benefits:

- SME-AGs have the possibility to use a neutral and innovative support tool for shaping communities of potential partners as preliminary condition for their constitution as steady collaborative networks. This will impact on the competitiveness of networked companies and then on the survival and possible economic growth of entire ecosystems in the involved regions.
- Network leads have the possibility to achieve higher efficiency levels in coordinating the respective networks. The natural consequence is a stronger presence on the market and an increased trust in customers preferring direct and fast partnership with the lead SME on behalf of the entire network rather than managing dispersed relations with a number of individual suppliers.
- Network members have the possibility to collaborate with one or more networks or supply chains assuring fast responses to each of them without being affected by the need of adapting their legacy systems to the different leads. In other words they are finally in the condition to gain positions in the market according to their skills along with their reliability and collaboration efficiency.

In addition, the resulting collaboration framework and ICT platform have proved to be at the same time general enough for meeting the requirements of a variety of business ecosystem of service companies, and flexible enough to adapt to the single specific case.

IV. CONCLUSION AND FUTURE WORK

In this paper, an ontology-based Supply Chain Management platform was presented. The final product of the eBEST Framework is a software platform for SME clusters and associations, and the knowledge to use it at best in different operational conditions. The represented eBEST platform can provide several features with that SMEs can

overcome the above-mentioned semantic barriers. The Ecosystem Shaping helps to find collaboration partners and to constitute clusters and improve their image which facilitates the preparation of processes. The workflow management built in the system provides an appropriate base to harmonize the processes. With the help of Operational Framework the companies can associate to each other through a supply chain and accomplish the promised goals. The visibility of companies and catalogues, the exchange of standardized semantic documents within the business processes are provided by ontological concepts in Semantic Node. In the Collaboration Framework an application built on ontology-based annotation scheme is responsible for automatically generating software instance of a new collaborative business process. These semantic applications support the realization of system requirements (like visibility, semantic interoperability, managing business process in transparent manner etc.) and foster collaboration and coordination among channel partners.

#### V. REFERENCE

- [1] M. Barratt, "Understanding the Meaning of Collaboration in the Supply Chain," *Supply Chain Management*, vol. 9, no. 1, pp. 30–42, 2004.
- [2] J. Cardoso, M. Hepp, and M. D. Lytras, *The semantic web*. Springer, 2007.
- [3] C. Chandra, "Supply Chain Workflow Modeling Using Ontologies," in *Collaborative Engineering*, A. K. Kamrani and E. S. A. Nasr, Eds. Springer US, 2008, pp. 61–87.
- [4] Collaboration Framework.  
<http://ebesttest.tade.netpositive.hu/>
- [5] M. C. Cooper, D. M. Lambert, and J. D. Pagh, "Supply Chain Management: More Than a New Name for Logistics," *International Journal of Logistics Management*, The, vol. 8, no. 1, pp. 1–14, Jan. 1997.
- [6] Council of Supply Chain Management Professionals, "Supply Chain Management Definitions." [Online]. Available: <http://cscmp.org/aboutcscmp/definitions.asp>. [Accessed: 20-May-2012].
- [7] K. L. Croxton, S. J. García-Dastugue, D. M. Lambert, and D. S. Rogers, "The Supply Chain Management Processes," *International Journal of Logistics Management*, The, vol. 12, no. 2, pp. 13–36, Jan. 2001.
- [8] eBEST:Empowering Business Ecosystems of Small Service Enterprises to Face the Economic Crisis. The project co-funded by the European Commission, FP7-SME-2008-2 No. 243554. WWW page. <http://www.ebest.eu/>, accessed 7.08.2012.
- [9] eBEST Consortium, "Deliverable D2.1 Analysis of Exchanged Documents," 2010
- [10] eBEST Consortium, "Deliverable D3.1 Overall Platform Design," 2010.
- [11] eBEST Consortium, "Tentative platform requirements and architecture," 2010.
- [12] M. S. Fox, M. Barbuceanu, and R. Teigen, "Agent-Oriented Supply-Chain Management," *International Journal of Flexible Manufacturing Systems*, vol. 12, no. 2, pp. 165–188, 2000.
- [13] D. Karastoyanova, T. Van Lessen, F. Leymann, Z. Ma, J. Nitzsche, B. Wetzstein, S. Bhiri, M. Hauswirth, and M. Zaremba, "A reference architecture for semantic business process management systems," in *Multikonferenz Wirtschaftsinformatik*, 2008, pp. 1727–1738.
- [14] J. Liu, S. Zhang, and J. Hu, "A case study of an inter-enterprise workflow-supported supply chain management system," *Information & Management*, vol. 42, no. 3, pp. 441–454, Mar. 2005.
- [15] T. B. Lee, J. Hendler, O. Lassila, and others, "The semantic web," *Scientific American*, vol. 284, no. 5, pp. 34–43, 2001.
- [16] J. T. Mentzer, W. DeWitt, J. S. Keebler, Soonhoong Min, N. W. Nix, C. D. Smith, and Z. G. Zacharia, "Defining supply chain management," *Journal of Business Logistics*, vol. 22, no. 2, pp. 1–25, 2001.
- [17] OWL Web Ontology Language Guide. W3C Recommendation 10 February 2004.  
<http://www.w3.org/TR/owl-guide/>.
- [18] Supply Chain Council, "Overview of SCOR Process Reference Model." [Online]. Available: <http://supply-chain.org/SCOR-overview>. [Accessed: 16-Jan-2012].
- [19] K. C. Tan, "A framework of supply chain management literature," *European Journal of Purchasing & Supply Management*, vol. 7, no. 1, pp. 39–48, 2001.
- [20] K. Ternai and M. Torok, "A new approach in the development of ontology based workflow architectures," in *2011 17th International Conference on Concurrent Enterprising (ICE)*, 2011, pp. 1–10.
- [21] Y. Ye, D. Yang, Z. Jiang, and L. Tong, "Ontology-based semantic models for supply chain management," *The International Journal of Advanced Manufacturing Technology*, vol. 37, no. 11, pp. 1250–1260, 2008.
- [22] W3C: World Wide Web Consortium.  
<http://www.w3.org>.

## Semantics on the Cloud: Toward an Ubiquitous Business Intelligence 2.0 ERP Desktop

Diletta Cacciagrano\*, Emanuela Merelli\*  
\*UNICAM - School of Science and Technology  
Camerino, Italy  
{name.surname}@unicam.it

Leonardo Vito\*<sup>‡</sup>  
‡Logical System srl  
Jesi, Italy  
leonardo.vito@gmail.com

Andrea Sergiacomi, Serena Carota  
Marche Region,  
Ancona, 60125, Italy  
{name.surname}@regione.marche.it

**Abstract**—Adequate information management requires more than persistently storing data. Owl- $\mathcal{M}_c^{ea}N^{inc}G$  (to read either owl-mining or owl-meaning) is an expandable ‘Business Intelligence 2.0’ Enterprise Resource Planning (ERP) prototype, with the aim to lead Public Administration toward Business Intelligence and information maturity. Designed for the Marche Region, Owl- $\mathcal{M}_c^{ea}N^{inc}G$  allows transforming, analysing and mining distributed and heterogeneous knowledge through semantic-driven GUI (Graphical User Interface)-based components, integrated on a common semantic knowledge model and embedded in a Cloud-based middleware. Such an architecture puts Owl- $\mathcal{M}_c^{ea}N^{inc}G$  beyond the actual expert-oriented semantic computing and makes it a user-friendly environment, where also naive users can easily edit, monitor, execute and store transformation, analysis and mining operations as new, reusable and semantically consistent business process knowledge.

Capabilities of (i) encoding operational knowledge into a declarative format and (ii) producing new and complex operational knowledge by composition of simpler declarative one allows realizing in Owl- $\mathcal{M}_c^{ea}N^{inc}G$  processes of *externalization* (i.e., converting tacit knowledge into explicit one) and *combination* (i.e., creating new explicit knowledge from existing explicit one). An example of externalization on the top of the Marche Region’s data warehouse is proposed to show how exploiting Owl- $\mathcal{M}_c^{ea}N^{inc}G$  for converting implicit knowledge’s intangible character in its successful understanding and sharing.

**Keywords** - *Semantic Web; Ontology; Business Intelligence; Data Warehouse; Data Mining.*

### I. INTRODUCTION

Knowledge represents the intellectual principal of any company. This is particularly evident nowadays, where a nontrivial extraction of implicit, previously unknown, potentially useful information from data and its efficient use by effective Business Intelligence (BI) methods can undoubtedly promote business competition and opportunities.

Enterprise Resource Planning (ERP) systems are designed to provide such methods, with the aim of integrating all company business facets. The umbrella term ERP refers to the processes of data transformation (e.g., Extraction, Transformation and Loading—ETL), analysis (e.g., Online Analytical Processing—OLAP) and mining (e.g., querying and clustering), as well as to terms such as data quality, data enrichment, data warehouse (DW), data mart and operational data store.

ERP systems are multi-module software applications that help companies to manage important backbone operations. ERP’s major objectives are (i) integrating all company departments and functions onto a single system that can serve all of the company needs, and (ii) enabling companies to present one face to their customers via integrated business processes, DWs and easy access to updated operational data.

On the one hand, ERP provides a valuable conceptual basis. On the other hand, any ERP implementation has to address several factors: information distribution, semantics heterogeneity, impossibility to test and reuse logic from existing transformations (as it is buried in source-specific code), information redundancy (when the same source feeds different data marts, being extracted and transformed by separately coded routines), absence of constrained information (complex descriptions of terms are not retained in the DW dimension tables and, as a consequence, values matching particular criteria and additional information about a term cannot be found without directly inspecting the data source), lack of user support during the mining model specification phase.

Such factors often discourage companies from fully exploiting ERP solutions, restricting their use to trivial operations (e.g., for checking and conveying known information in a more digestible manner, confirming known trends and relationships, automatically providing data for a what-if analysis still dependent on experts’ manual judgments).

Major problems arise in the Public Administration (PA), where factors like low interoperability levels among information systems, budgetary restrictions, technological know-how deficits and a latent change resistance worsen the above scenario, moving PAs away from the idea to invest on ERP solutions and on BI techniques.

#### *A. Contribution of the paper: toward a ‘BI 2.0’ ERP desktop for PAs and private companies*

Until recently, theoretical research on applying ontology to data mining was carried out by several studies: for dealing with the issue of incorporating ontology in the knowledge discovery process [1], [2], [3], [4], for integrating OLAP and

information retrieval from DWs [5] and for multiple source integration for DW OLAP construction [6].

Taking inspiration from the literature, **Owl- $\mathcal{M}_i^{eaNocG}$**  (to read either *owl-mining* or *owl-meaning*) aims at providing an expandable ERP system with ‘BI 2.0’ [7] capabilities, where:

- Decisions, facts and context are developed through crowdsourcing.
- Data and reports incorporate context information supplied by users.
- Data have a more direct linkage with action. Exceptions, alerts and notifications are based on dynamic business rules that learn about user’s business and what he is interested in.
- User can directly act on information.
- Business decisions can be monitored and hypotheses about business tactics can be integrated into the decision support system.
- Visualizing data and complex relationships is easier and more intuitive models of info-graphics become mainstream.
- The ability to detect complex patterns in data through automated analytic routines or intelligent helper models is built into analytic applications.
- Finding information is easier and search results provide context. Anyone looking at the same data can see that context when viewed.
- Linkages with unstructured contents as well as a previously acquired knowledge base is the key to ensuring collective knowledge and collaboration.

**Owl- $\mathcal{M}_i^{eaNocG}$**  was born as a UNICAM ICT-outsourcing product for answering Marche Region’s demand of semantically unlocking earned information and ensuring high-quality and homogeneous internal decision making processes. However its modular software architecture - a mash up of Semantic Web, workflow techniques, Cloud and Agent computing embedded in a fully web- and GUI (Graphical User Interface)-based environment - makes **Owl- $\mathcal{M}_i^{eaNocG}$**  a low-cost and easily customizable solution for any PAs and private companies.

**Owl- $\mathcal{M}_i^{eaNocG}$**  consists of several fully semantic-driven and GUI-based components (currently, knowledge and workflow management, semantic annotation and visual query systems), integrated on a common semantic knowledge model and embedded in a Cloud-based middleware. This architecture makes **Owl- $\mathcal{M}_i^{eaNocG}$**  not only an innovative ERP system for transforming, analysing and mining distributed and heterogeneous knowledge, but also a user-friendly environment, where semantics helps naive users to edit, monitor, execute and store transformation, analysis and mining operations as new, re-usable and semantically consistent business process knowledge.

The semantic layer also allows filtering more specific search spaces, minimizing the possibilities of illegal settings of mining models, storing and sharing user’s mining work,

discriminating between usual and newly acquired knowledge.

The possibility of improving **Owl- $\mathcal{M}_i^{eaNocG}$**  through an incremental and non-invasive refinement process - thanks to the **Owl- $\mathcal{M}_i^{eaNocG}$**  Cloud-based platform where new BI components can be plugged in a compositional way - can lead therefore toward the realization of:

- An integrated knowledge space (instead of a set of isolated and heterogeneous knowledge resources) that will unify different perspectives and interpretations of knowledge resources and will enable their treatment on a far more fine grained level, allowing for more sophisticated applications and services.
- A collaborative BI working environment (instead of a single person decision making process) that will bring every user to the same level of effectiveness and productivity and will ensure more efficient knowledge sharing by providing, at the same time, the reliability and the consistency of the decision making process.
- A change management system (instead of ad-hoc management of changes) that will ensure harmonisation of requests for changes, resolution of changes in a systematic way and their consistent and unified propagation to the collaborative and knowledge space, in order to ensure the high quality of the decision-making process.
- A platform for proactive delivery of knowledge (instead of an one-way knowledge access) that enables creation of an adaptable knowledge sharing environment through learning from the collaboration between users and their interaction with the knowledge repository and supporting in that way full empowerment and acceptance of users. A strong involvement of employees and stakeholder representatives is crucial, since defining the corporate vision is often the first step toward manifesting strategic thinking in PAs and enterprises.
- An ubiquitous assistive mining environment for storing/changing/extending/generalizing mining rules, filtering more specific search spaces by concept-based queries, minimizing the possibilities of illegal settings of mining models, storing/re-using user’s work.

**Owl- $\mathcal{M}_i^{eaNocG}$**  can be tried at the link [http://resourceome.cs.unicam.it/eyeOS/\(11/08/2012\)](http://resourceome.cs.unicam.it/eyeOS/(11/08/2012)) (User: owlmining, Passw: tryowlmining) .

## B. Plan of the paper

The remainder of this paper is organized as follows: Section II presents the **Owl- $\mathcal{M}_i^{eaNocG}$**  overall architecture, giving details about each components - a declarative and operational knowledge management environment (*Resourceome*), a semantic annotation component (*DataSMart*) and a semantic-driven visual query editor (*OWLEye*). Finally, Section III closes the paper, with a sketch of the ongoing implementation results and intended future work.

## II. THE Owl- $M_i^{caNrcG}$ ARCHITECTURE

**Owl- $M_i^{caNrcG}$**  is conceived as a semantic ERP platform, pivoting on a semantic DW that stores ontology-based semantic annotations, along with semantic-driven mechanisms for the definition and the execution of transformation and meaning processes over the stored data (see Fig. 1). It is based on a pluggable architecture exploiting and integrating techniques from diverse areas such as Cloud Computing, databases, machine learning, cognitive science, Semantic web, and others.

Currently, **Owl- $M_i^{caNrcG}$**  embodies as services several fully semantic-driven and GUI-based components - namely, a declarative and operational knowledge management environment (*Resourceome*), a semantic annotation component (*DataSmart*) and a visual query editor (*OWLEye*) - pivoting on a common hybrid OWL/SKOS-based multi-layered knowledge model for the semantic annotation of *resources* and *activities* (see Fig. 2) [8]. A Cloud-based middleware (*EyeOS*) provides the needed integration mechanisms between each **Owl- $M_i^{caNrcG}$**  component and the knowledge base (modeled as in Fig. 2) and among **Owl- $M_i^{caNrcG}$**  components themselves, allowing also further meaning services (developed on the top of a knowledge model as in Fig. 2) to be plugged in **Owl- $M_i^{caNrcG}$**  without changing its current architecture.

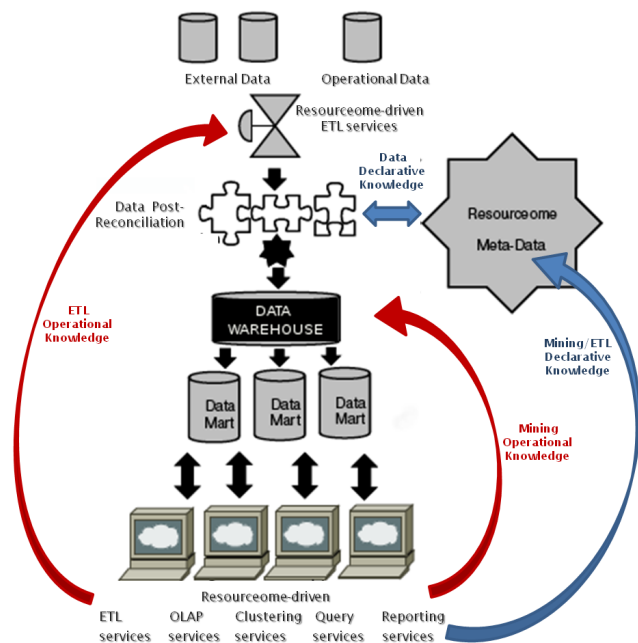


Figure 1: Owl- $M_i^{caNrcG}$  conceptual view.

### A. Resourceome + DataSmart: the semantic ERP kernel

*Resourceome* and *DataSmart* are the semantic core of **Owl- $M_i^{caNrcG}$** . Both components work on the top of a

specific knowledge model (Fig. 2) that, when applied to a knowledge base, allows contextualizing (i) resources w.r.t. a given domain and (ii) activities w.r.t. given resources. Usually, the knowledge base is represented by a DW, but can be also the integration of the DW with local and remote sources.

Requirement (i) is satisfied by splitting the Domain Ontology in [9] into two separate ontologies - a *Domain Ontology* conceptualizing the chosen domain instance and a *Resource Ontology* conceptualizing the resource space - and connecting them by abstract relations. Abstract relations also connect Domain and Resource Ontologies to a *Task Ontology* conceptualizing the activity hyperspace; such relations allow any activity to be linked to its working context and the involved roles and resources, thereby satisfying requirement (ii).

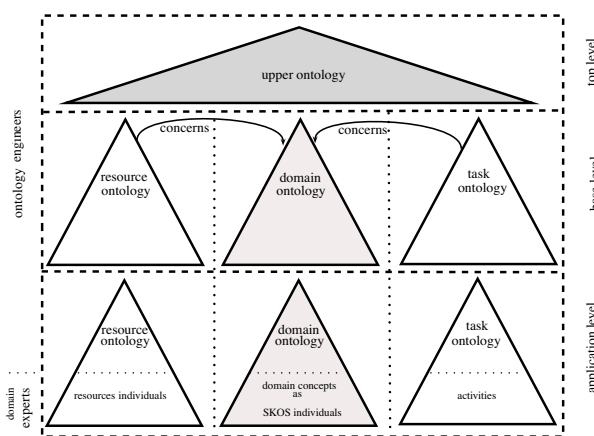


Figure 2: The knowledge model.

On the one hand, *DataSmart* is a *BioMart*[10]-based database federation system that makes it possible to present geographically distributed data sources as federated data in an integrated database, as well as to access and to cross-reference data from these data sources using a single user interface.

However, differently from *BioMart*, it can be also exploited as a data warehousing platform enabling ETL, OLAP and other mining operations (see Fig. 3). Most important, *DataSmart* is also a semantic annotation system based on a drag-and-drop interface, which allows imported data and attributes to be linked to a given knowledge model instance (see Fig. 4).

On the other hand, *Resourceome* [8] provides a web-based integrated environment for (i) managing distributed and heterogeneous knowledge as ontology concepts (e.g., as declarative knowledge); (ii) designing semantically consistent ETL/mining operations; (iii) running ETL/mining operations as distributed and mobile agent systems (e.g., as

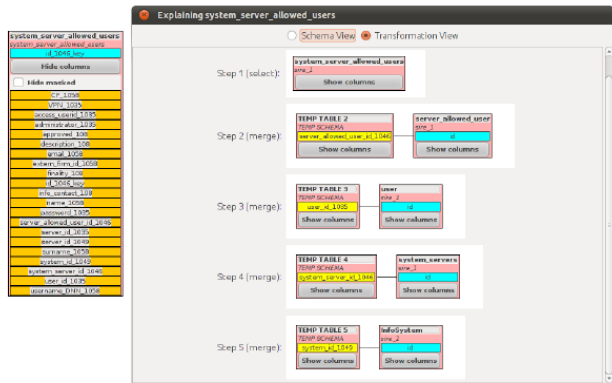


Figure 3: ETL-Transform and Load in DataSmart.

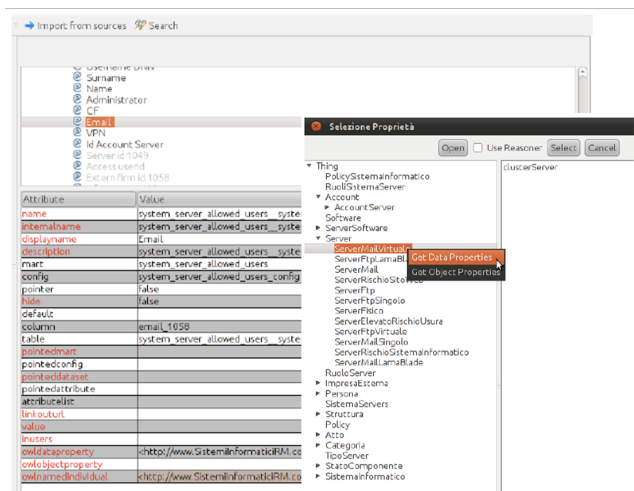


Figure 4: Semantic annotation by DataSmart.

operational knowledge); (iv) storing ETL/mining operations as ontology concepts (e.g., as declarative knowledge). Functionalities (i)-(iv) and (ii)-(iii) are provided respectively by a Knowledge Management System (KMS) and a Workflow Management System (WMS), both working on the knowledge base (modeled as in Fig. 2). Fig. 5 presents the final screenshot of an analysis process on financial data, edited and executed through Resourceome WMS and visualized by a Resourceome-driven reporting service.

*B. A dragging-and-dropping environment for conceptual queries*

Formulating non-ambiguous queries is often a too demanding task for users as they do not have the overview on the semantics of data stored in the system. Without complete comprehension of the schema and domain related knowledge, end users may develop a query based on their experience or intuition. Therefore, users' formulation of queries can possibly fall into some improper pits. This may lead to incorrect and redundant mining data space or mining

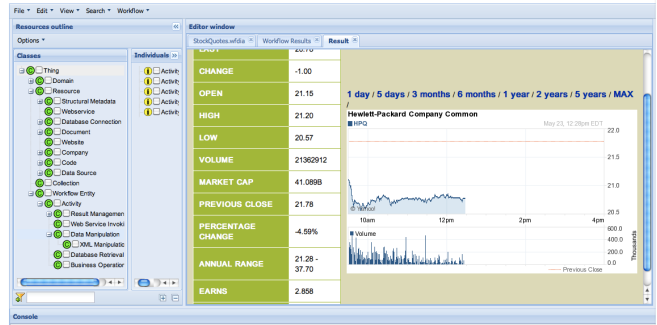


Figure 5: Example of analysis process output visualized by a Resourceome-driven reporting service.

results and waste the efforts accordingly.

The goal of OWLEye is to overcome this problem by providing an ontology-based information view of the data available in the knowledge base, integrated with a visual querying environment oriented to unskilled users.

OWLEye is equipped with a Query Design component allowing the graphical rendering of SPARQL queries by graphical constructs of the vSPARQL language [11]. This has necessarily entailed the development of a set of graphic notations - based on SPARQL syntax specification - supporting the visual representation of SPARQL query components.

Many of the vSPARQL constructs, once rendered, are selectable objects that can be edited using a popup menu. The menu allows users to define filtering, ordering and grouping information for the selected object. The design canvas itself can be zoomed and panned to view the entire query at different levels of resolution.

The possibility of browsing the knowledge model - embedded in the knowledge base through the DataSmart semantic annotation - insulates inexperienced users from the complexity of the query language and guides them in the process of query formulation. When the knowledge model is constructed correctly, the user can formulate semantically correct queries in a very intuitive way: dragging-and-dropping graphical elements allows user to browse the knowledge base and to select specific concepts of the knowledge model, while "stretching" edges permits to select properties and relations of interest (those associated to the stretched edges). Finally, query results can be visualized through several view layouts. An illustrative example of such features is given in Fig. 6.

**Example II.1.** We show how Owl-McNoel has been exploited to capture implicit knowledge from the Marche Region's knowledge base. As a single scenario cannot cover all the application possibilities, we focus on a specific Marche Region's request: knowing what *Struttura* (i.e., departments) are *StrutturaVulnerabile* (i.e., vulnerable department). For Marche's Region a department is considered vulnerable when at least five of its *SistemaInformaticoAmministrato*

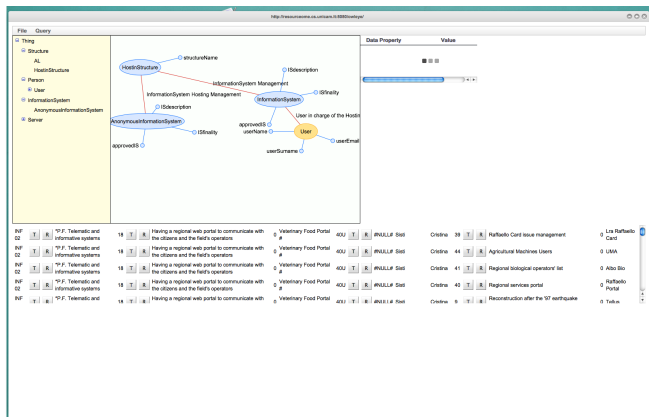


Figure 6: Example of query through OWLEye.

(i.e., information systems) manage personal and sensitive data.

This fact can be expressed by the OWL rule shown in Fig. 7, involving the concepts of *Struttura*, *SistemaInformativoAmministrato*, *PolicySistemaInformatico* and their relations. Fig. 8 shows the query formulated by OWLEye and the list of inferred vulnerable departments.

```

StrutturaVulnerabile ≡ Struttura ⊓
    ≥ 5 strutturaAmministraSistemaInformatico.(SistemaInformativo ⊓
        ∃ sistemaInformativoAssociatoPolicySI.(PolicySistemaInformativo ⊓
            ∃ policySistemaInformativoAssociatoPolicy.(Policy ⊓
                ∃ datiPersonaliPolicy.{true} ⊓ ∃ datiSensibiliPolicy.{true})))
    
```

Figure 7: Rule for inferring vulnerable departments.

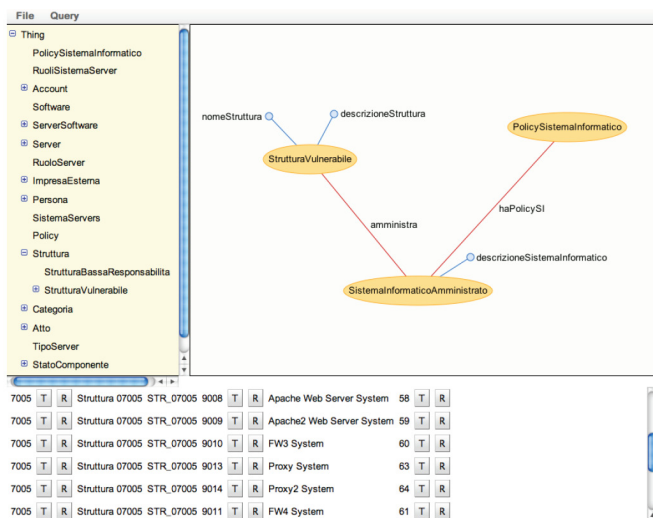


Figure 8: Query in OWLEye and list of inferred vulnerable departments.

### III. STATE-OF-THE-ART IMPLEMENTATION AND FUTURE WORK

There are a number of additional application areas for **OWL- $\mathcal{M}_i^{\text{co}}\mathcal{N}^{\text{inc}}\mathcal{G}$**  that we are exploring as part of our current and future work. In particular, we are studying the possibility to exploit **OWL- $\mathcal{M}_i^{\text{co}}\mathcal{N}^{\text{inc}}\mathcal{G}$**  for rule creation, information integration and knowledge acquisition.

It is well-known that SPARQL CONSTRUCT queries can be used for information integration and interoperability, since this kind of queries effectively modify and extend (perhaps multiple) knowledge bases according to the presence of information detected from one or more information sources. Since *OWLEye* supports the visualization of (among else) SPARQL CONSTRUCT queries, we argue that *OWLEye* can be used in **OWL- $\mathcal{M}_i^{\text{co}}\mathcal{N}^{\text{inc}}\mathcal{G}$**  also for editing rules and for representing the semantic mappings (or ontology alignments) between ostensibly disparate ontologies.

Another interesting point concerns the possibility to exploit **OWL- $\mathcal{M}_i^{\text{co}}\mathcal{N}^{\text{inc}}\mathcal{G}$**  for knowledge acquisition. Cluster mining is usually applied to discover groups in large amounts of data using large flat files as input source and, as a consequence, mining techniques are simply seen as tools trying to discover patterns.

As in the case of query-based mining, putting semantics into cluster mining allows to make explicit the conceptual knowledge structures of data, to take advantage of knowledge acquired in the previous knowledge discovery process stages, to provide users with further semantics that improves the understanding of the system, as well as to abstract from specific issues (platform, algorithms, parameters, etc).

For this reason, we plan to integrate in **OWL- $\mathcal{M}_i^{\text{co}}\mathcal{N}^{\text{inc}}\mathcal{G}$**  a *Resourceome*-driven clustering service equipped with a smart drag-and-drop based editor as *OWLEye*. Such a service shall embed a clustering algorithm with a level of accuracy similar to corpus-based ones but retaining the low computational complexity of path-based ones. At this aim, we are studying a weighted and ontology-based variant of the k-means algorithm [12], where weights are assigned on both data properties and relations and represent the importance level (see Fig. 9). The variant relies on a similarity measure defined as below:

$$sim_g = \left( \sum_k^m w_k (v_{k,i} - v_{k,j})^g \right)^{\frac{1}{g}}$$

where  $r_i = \{v_{i,1}, v_{i,2}, \dots, v_{i,k}\}$  denotes the value list of the  $i$ -th record in the dataset  $D = \{r_1, r_2, \dots, r_n\}$ ,  $S = \{a_{1,1}, \dots, a_{1,n_1}, a_{2,1}, \dots, a_{2,n_2}, \dots, a_{k,1}, \dots, a_{k,n_k}\} = \{a_1, a_2, \dots, a_N\}$  the attribute set (with  $\sum_{i=1}^k a_{i,n_i} = N$ ),  $a_{i,k}$  the  $k$ -th attribute of the  $i$ -th table,  $w_i \in (0..1]$  the weight of  $a_i$ .

Notice that  $sim_g$  can express the absolute distance ( $g = 1$ ), the euclidean distance ( $g = 2$ ) and the Chebyshev distance ( $g \rightarrow \infty$ ).

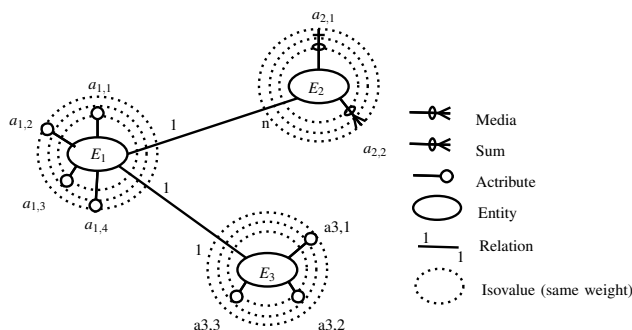


Figure 9: Clustering visualization system.

## ACKNOWLEDGEMENTS

A special thank to Sergio Villarreal and Constatino Giuliodori (Marche Region), Sauro Silvestrini (General Impianti Srl) and Pietro Liò (University of Cambridge), that gave an important contribution to the **Owl-M<sub>1</sub>NeG** design.

## REFERENCES

- [1] J. Aronis, F. Provost, and B. Buchanan, "Exploiting background knowledge in automated discovery," in *In the Proc. of the Second International Conference on Knowledge Discovery and Data Mining*, 1996, pp. 355–358.
- [2] S. Sharma, Osei-Bryson, and Kweku-Muata, "Framework for formal implementation of the business understanding phase of data mining projects," *Expert Syst. Appl.*, vol. 36, no. 2, pp. 4114–4124, Mar. 2009.
- [3] M. Domingues and S. Rezende, "Using taxonomies to facilitate the analysis of the association rules," in *In the Proc. of the Second international workshop on knowledge discovery and ontologies*, 2005.
- [4] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "The kdd process for extracting useful knowledge from volumes of data," *Commun. ACM*, vol. 39, no. 11, pp. 27–34, Nov. 1996.
- [5] T. Priebe and G. Pernul, "Ontology-based integration of olap and information retrieval," in *In the Proc. of the 14th intern. workshop on database and expert systems applications*, 2003, pp. 610–614.
- [6] W. Lin, M. Tseng, and C. Wu, "Ontology-incorporated mining of association rules in data warehouse," *Journal of Internet Technology*, vol. 8, no. 4, pp. 477–485, 2007.
- [7] G. S. Nelson, "Business Intelligence 2.0: Are we there yet?" in *In the Proc. of SAS Global Forum 2010*, Paper 040-2010, 2010.
- [8] D. Cacciagrano, F. Corradini, E. Merelli, L. Vito, and G. Romiti, "Resourceome: a multilevel model and a Semantic Web tool for managing domain and operational knowledge," in *The Third International Conference on Advances in Semantic Processing (SEMmapro 2009)*, P. Dini, J. Hendler, and J. Noll, Eds. IEEE Computer Society, 2009, pp. 38 – 43.
- [9] N. Guarino, *Formal Ontology in Information Systems: Proceedings of the 1st International Conference June 6-8, 1998, Trento, Italy*. Amsterdam, The Netherlands, The Netherlands: IOS Press, 1998.
- [10] J. Zhang, S. Haider, J. Baran, A. Cros, J. Guberman, J. Hsu, Y. Liang, L. Yao, and A. Kasprzyk, "BioMart: a data federation framework for large collaborative projects." *Database: the journal of biological databases and curation*, vol. 2011, no. 0, 2011.
- [11] M. Shaw, L. T. Detwiler, N. Noy, J. Brinkley, and D. Suci, "vsparql: A view definition language for the semantic web," *Journal of Biomedical Informatics*, vol. 44, no. 1, pp. 102 – 117, 2011.
- [12] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, L. M. L. Cam and J. Neyman, Eds., vol. 1. University of California Press, 1967, pp. 281–297.



## Using DBPedia to bootstrap new Linked Data

Alexiei Dingli  
 Department of Intelligent Computer Systems  
 Faculty of ICT  
 University of Malta  
 Msida MSD 2080, Malta  
 alexiei.dingli@um.edu.mt

Silvio Abela  
 Department of Intelligent Computer Systems  
 Faculty of ICT  
 University of Malta  
 Msida MSD 2080, Malta  
 sabe0004@um.edu.mt

**Abstract**—If the documents on the WWW were somehow structured, then machines can be made to extract meaning (or semantics) from the content and help us find more data that is relevant to what we search. There is an effort to find better ways to include machine-meaning in the documents already present on the WWW by using Natural Language Processing (NLP) techniques and Web technologies such as XML and RDF that are used to insert and represent the “meaning” in the extracted content. We propose an application that uses Information Extraction to extract patterns from a human readable text and use it to try and find similar patterns elsewhere by searching the WWW. This is done to bootstrap the creation of further data elements. These data elements are then stored in RDF format and reused in other searches. Our evaluation show that this approach gives an encouraging degree of success with a precision of 79% and a recall of 71%.

**Keywords**- RDF; Linked Data; Semantic Web; XML

### I. INTRODUCTION

The WWW has become a powerful modern medium that in some ways is surpassing the old-style media. This is true of advertising where in the United Kingdom online advertising has surpassed that of television [1] and in the United States, in 2010, a total of \$12 billion was spent on Web advertising [2]. This does not make the Web a structured medium, almost totally the opposite. [3] dreamt of a web where humans and machines share the contents and help each other in doing so. He said everyone should publish **linked data** that automatically links pieces of data together [4]. Thus, the aim of this research will be to increase Linked Data triples by using patterns extracted from parts of text to find similar patterns and generate new triples by using Information Extraction techniques.

### II. BACKGROUND

Our research will be divided in two: **Linked Data** and **Information Extraction**. We will show that there is a link between the two so information islands are linked together more.

#### A. Publishing Linked Data

With more linked data available we will be able to query the Web as a database [5]. The Web of Data will use names

for *things*. A book, a person, a car, a cat; they are all “things” having a URI as a name. To be able to publish Linked Data, Tim Berners-Lee listed the following four principles [6]:

- 1) Use URIs as names for things.
- 2) Use HTTP URIs so that people can look up those names.
- 3) When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL).
- 4) Include links to other URIs so that they can discover more things.

Since the beginning of the Linked Data project (<http://linkeddata.org/>) the number of datasets published as linked data now stands at 203 from a humble 12 in 2007 [7]. One of the first entities to use linked data was the British Broadcasting Corporation (BBC) that used linked data for all its programs in an effort to centralise the vast amount of information from all its programs micro sites [8].

#### B. RDF and DBPedia

The Resource Description Framework (RDF) is a simple graph model of the form **subject-predicate-object** hence triple. It was developed to describe web resources and machine readable information [5]. The graph has two nodes that may be either blank or a URI with a directed arc (the predicate) *always* a URI.

DBPedia is an effort to extract information from Wikipedia articles that have info boxes although the number of such articles is roughly a third of all the English articles [9]. The effort produced an astonishing 25 billion triples (<http://www4.wiwiss.fu-berlin.de/lodcloud>) but this is only a drop when compared with the vast amounts of information on the indexed Web (<http://www.worldwidewebsite.com>). The main difficulty to extract data from information on the WWW is that the latter is inconsistent, ambiguous, uncertain and whose corpus is constantly changing [10].

#### C. Natural Language Processing

This is the study of text processing within a computer system for a spoken or written language [11]. It touches on three main areas of understanding: Theoretical Linguistics,

Computational Linguistics and Artificial Intelligence with the greatest advances coming with the computer age. In 1950s Alan Turing hinted that machine translation needed to be unambiguous and that machines need to learn to think [12]. During the 1960's and early 1970's, ELIZA [13] and SHRDLU [14], were an early attempt at NLP and Artificial Intelligence (AI). Although from then, technology has mushroomed we are still far from an ideal situation due to the many nuances in spoken and written languages.

D. Extracting Information from the Web

Before we extract information we need to retrieve it. With Information Retrieval (IR) we get a subset of documents from a collection (the corpus) whilst with Information Extraction (IE) we extract facts or structured information from those documents [15]. IE is used in another field called Named Entity Recognition (NER) where special tools identify different types of semantics from words such as names, nouns, etc. NER tools determine what we take foregrounded while reading such as paragraphs or sentence endings [11]. NER tools contain resources such as Tokenisers, Part of Speech taggers (POS), Sentence Parsers and Semantic Taggers. These aid the system to successfully process a body of text.

III. METHODOLOGY

We aim to extract patterns from text and find similar patterns from the Web using GATE (http://gate.ac.uk) and JENA (http://jena.sourceforge.net).

The Gate system encompasses a host of well written tools providing a solid base for IE, and NER. It also boasts its own complete IE system called ANNIE (a Nearly New Information Extraction system). ANNIE makes use of JAPE (Java Annotations Pattern Engine). Gate uses

- Features: attribute/value pairs
- Corpora: collection(s) of document
- Documents: the input documents
- Annotations: directed acyclic graphs modelled as annotations using XML

We used ANNIE's resources to process the input into sentences, nouns, verbs etc. ANNIE does this by using JAPE which provides finite state transduction over the annotations based on regular expressions [16]. ANNIE sets a serial pipeline of text processing resources that inserts a unique ID, type and offsets within the text. Feature map.

By using JAPE grammars, ANNIE matches wanted patterns and then inserts these newly matched patterns into new feature maps. A JAPE grammar consists of a left hand side, containing annotation patterns, and a right hand side containing Java code to manipulate those patterns.

We also make use of Jena, a framework for reading and writing RDF data as XML. Jena also provides a backend triple database that can be either accessed from command

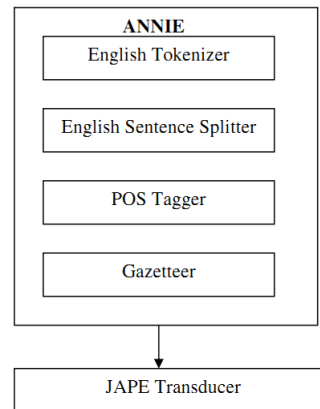


Figure 1. A typical ANNIE pipeline - source The GATE user manual

line or from a web interface called Fuseki. SPARQL is a query language for expressing RDF queries over triple-stores (http://www.w3.org). A SPARQL query may contain triples, conjunctions or disjunctions and some other optional patterns. Its syntax is similar to SQL with queries starting with a SELECT statement.

A. Overview

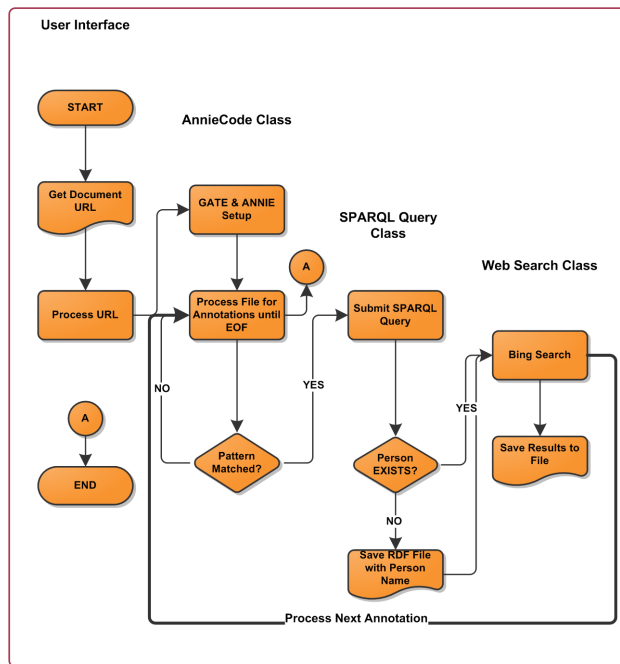


Figure 2. High-level System Flowchart

We propose a system that processes input from text documents to find patterns by using Natural Language Processing techniques. We use SPARQL queries over a database containing over 1.7 million triples and submit Web queries to find other similar patterns on the open WWW. To aid our system to find the required patterns we will also propose

two new JAPE grammars that match the patterns we need in the text input.

Our system presents a simple user interface that processes input documents using four stages:

- 1) The input stage whereby the system is given an initial URL of a document such as [http://en.wikipedia.org/wiki/Barack\\_Obama](http://en.wikipedia.org/wiki/Barack_Obama) which is then downloaded and stored in the system for future processing.
- 2) The processing stage where GATE is used to annotate text with specific rules or patterns. An example rule might be ...

Locate the sentence which contains a date, a location, a person and a born event

A date, location and a person are standard extraction patterns found in Gate. A born event is an additional pattern which we crafted to identify phrases such as "born on", etc.

The rule eventually extracts sentences such as "**Obama was born on August 4, 1961, in Hawaii**" since **Obama** will be recognised as a person, "**was born on**" is recognised by the born event extraction patterns, **August 4,1961** is a date and **Hawaii** is a location.

- 3) In the querying stage, we query our database to check if the data we just extracted exists in the database. If it doesn't, it is inserted in the database as a new RDF data item. An issue to consider is the validity of the data since different pages might return different results for the same person. In the example we're considering, we found that almost 2 million pages claim that "**Obama was born in Kenya**" and not in Hawaii. After removing the reposted articles and comparing the top patterns extracted, only one piece of data will prevail and that data will be inserted in the database. This is definitely not a foolproof way of ascertaining the truth however, it returns good results in most cases.
- 4) The Web search stage looks for patterns, similar to the ones in the database in order to extract new information. By similar we mean, that the data just retrieved is sent to the search engine but one of the patterns is omitted. This will retrieve other variances which might not have been covered. So using the data we just retrieved about Obama but omitting the born event, we discover a new piece of data such as

**"Obama's date of birth is August 4, 1961, in**

## Hawaii"

We can see that this data is very similar since all the data items match however the born event pattern is very different. Thus, we use this phrase to construct new patterns by adding wildcards such as

**\*'s date of birth is \*, in \***

This new query will then be fed into Bing and new pages are retrieved which are then sent to the input stage mentioned earlier and the cycle is repeated. By using this simple bootstrapping approach, new information can be discovered all over the web thus generating new linked data almost automatically.

## IV. TESTING

We tested our application with a small text file containing five seed samples that contain the four pattern parts we need to match in no particular order and in one complete short sentence.

- PersonEvent - *Kenny Matthieson DalGLISH*
- BornEvent - *the verb born or its tenses*
- DateEvent - *4 March 1951*
- LocationEvent - *Glasgow in Scotland*

A typical whole sentence pattern may be: *On 4 March 1951 in Glasgow in Scotland Kenny Matthieson DalGLISH was born.* These patterns are taken from Wikipedia™. During the initial test runs, our system retrieved quite a good number of new triples. We manually went through a sample of the triples and found that there were quite a few with errors or false positives.

## V. EVALUATION

We analysed our results both using a known metric and by manually going through the new triples to check for validity. As a metric we used Precision & Recall (P&R) with the following criteria:

$$Precision = \frac{Relevant\ triples\ retrieved}{Number\ of\ triples\ retrieved}$$

$$Recall = \frac{Relevant\ triples\ retrieved}{Number\ of\ triples\ in\ text}$$

This method of evaluation has been applied to the field of Information Retrieval such as extracting text summaries from news texts [17]. In IE, no evaluating algorithm is perfect due to contradictions in the texts and the nuances of the language [18].

Our initial evaluation of the results gave very low values of 0.21 and 0.13 for precision and recall respectively. In view of these results we had to determine what was causing

them to be so low.

We encountered errors due to different font encoding (ANSI, UTF-8), to characters that make the application throw exceptions etc. Other errors in the data resulted from:

- Annotations spanning sentences
- Partial annotations not being discarded
- False positives especially in DateEvents
- Order of annotations when passing through the Annie pipeline
- False positives overall

In order to get better results we rectified the above error instigators by modifying one of our Jape grammars and one of our most important methods so that we matched a more specific pattern in the left hand side of the grammar to ensure that the sentences retrieved *really* contain the four pattern parts we needed. We also added a more specific RegEx to match dates instead of relying on ANNIE's default Date type.

#### A. Modifications made

One of the main issues was that the annotations were spanning sentences and so we had to introduce a {Sentence} condition in our grammar. This made sure that we could select whole sentences as annotations to check within their span for our four part-pattern. The right hand side of our modified Jape grammar works by:

- 1) Retrieve a whole sentence annotation
- 2) Obtain an iterator over the annotation
- 3) Try to match the various pattern parts within the sentence annotation
- 4) Set a Boolean variable to true if a part is found
- 5) If all four pattern parts are found and all Flags are TRUE
  - put the parts together in a new type called AllParts
  - add a new rule called GetAllPartsRule

Other modifications were made in the performGateProcessing() method that now uses sentence annotations and then the algorithm searches over them for the pattern parts. This made sure that only the patterns within that sentence are selected.

These modifications gave their fruit as the results and the P&R values increased four-fold and five-fold for Precision and Recall respectively. In the following section we will only list the values recorded from the results tables.

#### B. Results after modifications

For Test 1 we used the same criteria as the initial test. This test gave the following results:

Table I  
MODIFICATION FOR PERFORMGATEPROCESSING() METHOD -  
SOURCE: AUTHOR

1. Get sentence annotation - put in sentence Set
2. Get sentence offsets - put in pattern Set
3. Retrieve the current annotation, i.e. the part of pattern needed
4. If GetAllParts is true
5. Get features and offsets
6. For (each type of pattern)
7. Check if the type's offsets are within the GetAllParts offsets
8. If (YES) retrieve the annotation with those offsets from the Set & Process accordingly
10. Remove consumed annotation from annotation set

#### Test 1

$$Precision = \frac{8 + 27 + 100 + 336}{8 + 27 + 100 + 389} = 0.89$$

$$Recall = \frac{8 + 27 + 100 + 336}{9 + 29 + 102 + 585} = 0.65$$

For subsequent tests, and to test the system's robustness, we used different seed files with information coming mainly from the IMDB website (<http://www.imdb.com>). The results of these tests gave a precision of 0.97, 0.61, 1, 0.7 and 0.54. The recall measure gave values of 0.97, 0.61, 0.83, 0.7 and 0.5. Taking all the above precision and recall results whilst summing for an average, we get the following:

$$Precision = \frac{4.71}{6} = 0.79$$

$$Recall = \frac{4.26}{6} = 0.71$$

These successful results were only possible with the code modifications we made after the initial runs of the application. This meant that when we used the same seed file during the development of the application and were getting good responses from the system we were failing to realise that the system was not as robust as it should have been. These shortcomings then manifested themselves after we used larger input files that gave many errors.

With these modifications the application retrieved less results, which was the only disadvantage we noticed. Apart from that we noticed that the results, although less, have more quality in that they are more correct. There are also less results that contain error or erroneous data and this can be verified from the Precision & Recall values we recorded. The quality of the triples we retrieved varied greatly although the quantity is less. We are of the opinion that less does not mean worse, we rather have less triples that are of very good quality rather than have large quantities whose quality is poor.

### C. A note on Precision & Recall

According to [19], Precision & Recall results are supposed to be inversely proportional. In [20] it was found that precision and recall applied to data from news sources (data having some structure), places P&R values in the high 80s or low 90s (percentage-wise). In another study by [21], similar to ours, it was found that precision and recall are not always inversely proportional but may also produce values close to our results.

## VI. ACHIEVEMENTS & LIMITATIONS

The main aim of our research was to try and increase the chances of new triples being extracted from unstructured text or documents.

### A. Achievements

In order to reach our set goals we produced an application that although having a simple design meets our goals' needs. The application makes use of embedded code from the GATE system and this permitted our application to generate the acceptable output we had. We also wrote two new JAPE grammars that helped us find our required four pattern parts and then use them to check whether these are contained in a given sentence that was also extracted from the same text. With the encouraging results we obtained after we modified our code, we have reached our objectives.

### B. Limitations

During our research and application development we also encountered some difficulties that we did not have control upon but which made our development a little more interesting.

Our application currently operates with both the SPARQL and the Web search query code hard-coded within the application's methods. We intended for these to be dynamic and left to the user to decide which of the pattern parts to use in the query. This would have given the user a better usage experience.

In this application we are also not inserting on updating the dataset in the database due to certain factors such as the retrieved person names' ambiguity. We are accepting each new triple at face value as long as they are valid and correctly built. For example, *Steven George Gerrard* and *Steven "Stevie" Gerrard* may be found as being two different persons when in actual fact they are not. To overcome this the application would need other grammars that would have been outside of the scope of this research.

We are also not checking for duplicate findings so that in our result files we can find more than one entry with

the same name. This happens because search engines may return similar result snippets that we are appending in one file and subsequently using these files as input for subsequent runs and tests.

## VII. FUTURE WORK

On the whole, the application produced good results. Because of this, there are two areas where it needs improvements. First and foremost, we need to test the application on other patterns. At the moment, the patterns used were limited to identifying birth date and locations whilst associating them to a person. This is very useful information however there are other areas which one could explore such as working relations, educational relations, family relations, etc. In particular, we should go a step further and explore ways of generating these initial patterns automatically thus ensuring that the system is fully automated. This can be done by delving further into the creation of personal ontologies automatically.

Secondly, we need to test the system on a larger dataset. In particular it would be ideal to test it on the whole DBPedia corpus. Obviously, this is not a trivial thing to do especially since we require massive investment in computing resources. However the recent proliferation of cloud computing means that these massive requirements are finally within reach. Thus, our next project will involve the utilisation of the cloud to generate even more Linked Data.

## VIII. CONCLUSION

Our research tried to exploit information extraction in order to generate linked data automatically thus realise Tim Berners-Lee's vision of the future web. A web made up of human readable documents together with documents that permit machines to also *understand* the documents they display and process. In doing so, the machine can be put in a better position to help its human user in doing more in a shorter span of time and in attaining more by the direct help of the machine itself.

Our system managed to extract patterns from a human readable text and use it to find similar patterns elsewhere by searching the WWW. This was achieved by using bootstrap techniques. The new data generated is then stored in RDF format and reused in other searches.

The evaluation conducted also produced extremely good results having an overall precision of 79% and a recall of 71%, thus encouraging us to look further into similar approaches. In conclusion, we demonstrated that semantic meaning can be extracted from natural text and that Linked Data can be generated with little effort. Our data can then be easily added to the growing datasets of the Linked Open Data cloud thus bring the Semantic Web a step closer to reality.

## REFERENCES

- [1] M. Sweeney, "Internet overtakes television to become biggest advertising sector in the UK," *The Guardian - newspaper*, 2009, last accessed 6th September, 2011.
- [2] PricewaterhouseCoopers, "Internet Advertising Revenue Report," Internet Advertising Bureau, Tech. Rep., 2010.
- [3] T. BernersLee, J. Hendler, and O. Lassila, "The semantic web - a new form of web content that is meaningful to computers will unleash a revolution of new possibilities," *Scientific American Magazine - May 2001*, 2001.
- [4] T. Berners-Lee, "Putting government data online," 2009, personal view only.
- [5] T. Berners-Lee, J. Hollenbach, K. Lu, J. Presbrey, and M. Schraefel, "Tabulator redux: Browsing and writing linked data," 2008.
- [6] T. Berners-Lee, "Linked data - design issues," 2006, author's personal view.
- [7] R. Cyganiak and A. Jentzsch, "Linking open data cloud," Richard Cyganiak and Anja Jentzsch, 2011, last accessed 3rd September, 2011.
- [8] G. Kobilarov, T. Scott, Y. Raimond, S. Oliver, C. Sizemore, M. Smethurst, C. Bizer, and R. Lee, "Media Meets Semantic Web - How the BBC uses DBpedia and Linked data to Make Connections." in *ESWC*, ser. Lecture Notes in Computer Science, L. Aroyo, P. Traverso, F. Ciravegna, P. Cimiano, T. Heath, E. Hyonen, R. Mizoguchi, E. Oren, M. Sabou, and E. P. B. Simperl, Eds., vol. 5554. Springer, 2009, pp. 723–737.
- [9] D. Lange, C. Böhm, and F. Naumann, "Extracting structured information from wikipedia articles to populate infoboxes," in *Proceedings of the 19th ACM international CONFERENCE on Information and knowledge management*, ser. CIKM '10. New York, NY, USA: ACM, 2010, pp. 1661–1664.
- [10] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "DBpedia: A nucleus for a web of open data," in *Proceedings of the 6th International Semantic Web Conference (ISWC)*, ser. Lecture Notes in Computer Science, vol. 4825. Springer, 2008, pp. 722–735.
- [11] P. Jackson and I. Moulinier, *Natural Language Processing for Online Applications: Text Retrieval, Extraction, and Categorization (Natural Language Processing, 5)*. John Benjamins Pub Co, 2002.
- [12] A. M. Turing, "Computing Machinery and Intelligence," *Mind*, vol. LIX, pp. 433–460, 1950.
- [13] J. Weizenbaum, "Eliza—a computer program for the study of natural language communication between man and machine," *Commun. ACM*, vol. 9, pp. 36–45, January 1966.
- [14] T. Winograd, "Procedures as a representation for data in a computer program for understanding natural language," *Cognitive Psychology*, vol. 3, no. 1, pp. 1–191, 1972.
- [15] S. Sarawagi, "Information extraction," *Found. Trends databases*, vol. 1, pp. 261–377, March 2008.
- [16] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, N. Aswani, I. Roberts, G. Gorrell, A. Funk, A. Roberts, D. Damljanovic, T. Heitz, M. A. Greenwood, H. Saggion, J. Petrak, Y. Li, and W. Peters, *Text Processing with GATE (Version 6)*, 2011.
- [17] J. Goldstein, M. Kantrowitz, V. Mittal, and J. Carbonell, "Summarizing text documents: sentence selection and evaluation metrics," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR '99. New York, NY, USA: ACM, 1999, pp. 121–128.
- [18] C. Welty and J. Murdock, "Towards knowledge acquisition from information extraction," in *The Semantic Web - ISWC 2006*, ser. Lecture Notes in Computer Science, I. Cruz, S. Decker, D. Allemang, C. Preist, D. Schwabe, P. Mika, M. Uschold, and L. Aroyo, Eds. Springer Berlin / Heidelberg, 2006, vol. 4273, pp. 709–722.
- [19] R. Baeza-Yates and G. H. Gonnet, *Modern Information Retrieval*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., 1999.
- [20] N. Indurkha and F. J. Demerau, *Handbook of Natural Language Processing, Second Edition (Chapman & Hall/CRC Machine Learning & Pattern Recognition)*. Chapman and Hall/CRC, 2010.
- [21] A. Dingli, F. Ciravegna, and Y. Wilks, "Automatic semantic annotation using unsupervised information extraction and integration," in *Proceedings of the Workshop on Semantic Annotation and Knowledge Markup, SEMANNOT2003, K-CAP 2003 Workshop, at Sanibel, 26 October 2003*, S. Handschuh, M.-R. Koivunen, R. Dieng, and S. Staab, Eds., 2003.

# Word Sense Disambiguation Based on Distance Metric Learning from Training Documents

Minoru Sasaki

*Dept. of Computer and Information Sciences  
Faculty of Engineering, Ibaraki University*

*4-12-1, Nakanarusawa, Hitachi, Ibaraki, 316-8511, Japan  
Email: msasaki@mx.ibaraki.ac.jp*

Hiroyuki Shinnou

*Dept. of Computer and Information Sciences  
Faculty of Engineering, Ibaraki University*

*4-12-1, Nakanarusawa, Hitachi, Ibaraki, 316-8511, Japan  
Email: shinnou@mx.ibaraki.ac.jp*

**Abstract**—Word sense disambiguation task reduces to a classification problem based on supervised learning. However, even though Support Vector Machine (SVM) gives the distance from the data point to the separating hyperplane, SVM is difficult to measure the distance between labeled and unlabeled data points. In this paper, we propose a novel word sense disambiguation method based on a distance metric learning to find the most similar sentence. To evaluate the efficiency of the method of word sense disambiguation using the distance metric learning such as Neighborhood Component Analysis and Large Margin Nearest Neighbor, we make some experiments to compare with the result of the SVM classification. The results of the experiments show this method is effective for word sense disambiguation in comparison with SVM and one nearest neighbor. Moreover, the proposed method is effective for analyzing the relation between the input sentence and all senses of the target word if the target word has more than two senses.

**Keywords**—word sense disambiguation, distance metric learning, similar example retrieval,

## I. INTRODUCTION

In natural language processing, acquisition of sense examples from examples that contain a given target word enables to construct an extensive data set of tagged examples to demonstrate a wide range of semantic analysis. For example, using the obtained data set, we can create a classifier that identifies its word sense by analyzing co-occurrence statistics of a target word. Also, we can construct a wide-coverage case frame dictionary automatically and construct thesaurus for each meaning of a polysemous word. To construct large-sized training data, language dictionary and thesaurus, it is increasingly important to further improve to select the most appropriate meaning of the ambiguous word.

If we have training data, word sense disambiguation (WSD) task reduces to a classification problem based on supervised learning. This approach is generally applicable to construct a classifier from a set of manually sense-tagged training data. Then, this classifier is used to identify the appropriate sense for new examples. A typical method for this approach is the classical bag-of-words (BOW) approach, where each document is represented as a feature vector

counting the number of occurrences of different words as features. By using such features, we can easily adapt many existing supervised learning methods such as Support Vector Machine (SVM) [2] for the WSD task. However, even though SVM gives the distance from the data point to the separating hyperplane, SVM is difficult to measure the distance between labeled and unlabeled data points.

In this paper, to solve this problem, we propose a novel word sense disambiguation method based on a distance metric learning to find the most similar sentence. In general, when words are used with the same sense, they have similar context and co-occurrence features. To obtain feature vectors that are useful to discriminate among word sense efficiently, examples sharing the same sense are close to each other in the training data while examples from different senses are separated by a large distance by using the distance metric learning method.

In this method, we apply two distances metric learning approach. One approach is to find an optimal projection which maximizes the margin between data points from different classes such as Local Fisher Discriminant Analysis (LFDA)[7][9], Semi-Supervised Local Fisher Discriminant Analysis (SELF) [8]. Another alternative is to learn a distance metric such that data points in the same class are close to each other and those in different classes are separated by a large margin such as Neighborhood Component Analysis (NCA) and Large Margin Nearest Neighbor (LMNN). We present the results of experiments using these two approaches of the proposed method to evaluate the efficiency of word sense disambiguation.

The rest of this paper is organized as follows. Section 2 is devoted to the introduction of the related work in the literature. Section 3 describes distance metric learning method. Section 4 illustrates the proposed system. Experimental results are presented in Section 5. Finally, Section 6 concludes the paper.

## II. RELATED WORKS

This paper proposes a method based on a distance metric learning for WSD. In this section, some previous research

using supervised approaches will be compared with our proposed method.

$k$ -nearest neighbor algorithm ( $k$ -NN) is one of the most well-known instance-based learning methods[1]. The  $k$ -NN classifies test data based on closest training examples in the feature space. One of the characteristics of this method is to calculate the similarity measure (e.g. cosine similarity) among instances. Therefore, this method can calculate a similarity measure between the new context and the training context, but do not consider the discriminative relations among the training data.

Support Vector Machines (SVM) has been shown to be the most successful and state-of-the-art approach for WSD[4][5]. This method learns a linear hyperplane that separates positive examples from negative examples from the training set. A test example is classified depending on the side of the hyperplane. Therefore, SVM has been successfully applied to a number of WSD problems. However, even though SVM gives the distance from the data point to the separating hyperplane, SVM is difficult to measure the distance between labeled and unlabeled examples. If the target word has more than two senses, This approach does not work so well.

### III. DISTANCE METRIC LEARNING

Distance metric learning is to find a new distance measure for the input space of training data, while the pair of similar/dissimilar points preserves the distance relation among the training data pairs. In the Distance metric learning, there are two types of leaning approaches: dimensionality reduction and neighborhood optimization. In this section, we briefly explain two distance metric learning approaches.

#### A. Metric Learning with Dimensionality Reduction

This approach employs a linear transformation which assigns large weights to relevant dimensions and low weights to irrelevant dimensions. This is commonly used for data analysis such as noise removal, visualization and text mining and so on. The typical methods of its approach are Local Fisher Discriminant Analysis (LFDA)[7][9] and Semi-Supervised Local Fisher Discriminant Analysis (SELF) [8]. LFDA finds an embedding transformation such that the between-class covariance is maximized and the within-class covariance is minimized, as shown in Figure 1.

This approach is efficient for representation of the relationship between data. However, problem arises when we apply this approach to predict new data. This method provides rotation of coordinate axes, not provide data points re-mapped to the original space, so that SVM generates a rotation of the hyperplane which is constructed in the original space. Therefore, there is little change in accuracy of performance compared to using the original feature space.

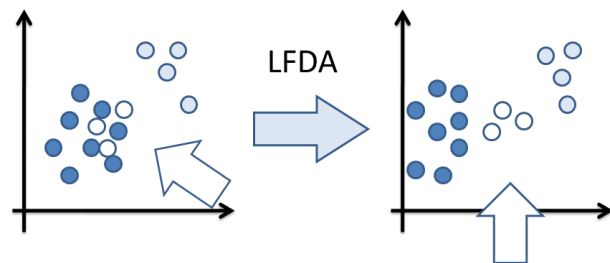


Figure 1. Local Fisher Discriminant Analysis

#### B. Metric Learning with Neighborhood Optimization

Alternative approach to distance metric learning is the method to learn a distance metric such that data points in the same class are close to each other and those in different classes are separated by a large margin. The two methods that implement this approach were developed, Neighborhood Component Analysis (NCA) [3] and Large Margin Nearest Neighbor (LMNN) [10].

1) *NCA*: NCA is a method for finding a linear transformation of training data such that the Mahalanobis distance between pairwise points is optimized in the transformed space. Given two data points  $x_i$  and  $x_j$ , the Mahalanobis distance between  $x_i$  and  $x_j$  is calculated by

$$d(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{A}\mathbf{x}_i - \mathbf{A}\mathbf{x}_j)^T (\mathbf{A}\mathbf{x}_i - \mathbf{A}\mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{M} (\mathbf{x}_i - \mathbf{x}_j), \quad (1)$$

where  $\mathbf{M} = \mathbf{A}^T \mathbf{A}$  is the distance metric that needs to be learned from the side information.

In this method,  $p_{ij}$  represents the probability of classifying the data point  $x_j$  to the data point  $x_i$  as neighbor as follows:

$$p_{ij} = \frac{\exp(-\|\mathbf{A}\mathbf{x}_i - \mathbf{A}\mathbf{x}_j\|^2)}{\sum_{k \neq i} \exp(-\|\mathbf{A}\mathbf{x}_i - \mathbf{A}\mathbf{x}_k\|^2)} \quad (2)$$

. Then, the probability  $p_i$  is defined as the sum of the probability  $p_{ij}$  of classifying the data points  $x_j$  into the class  $c_i$ .

$$p_i = \sum_{j \in C_i} p_{ij}, \quad (C_i = \{j | c_i = c_j\}) \quad (3)$$

The optimization function  $f(\mathbf{A})$  is defined as the sum of the probabilities of classifying each data point correctly. We maximize this objective function with respect to the linear transformation  $f(\mathbf{A})$ .

$$p_i = \sum_{j \in C_i} p_{ij}, \quad (C_i = \{j | c_i = c_j\}) \quad (4)$$

However, this objective function  $f(\mathbf{A})$  is not convex, so there is a possibility of getting stuck in local minima.

2) *LMNN*: LMNN is a method for learning a distance metric such that data points in the same class are close to each other and those in different classes are separated by a large margin, as shown in Figure 2.



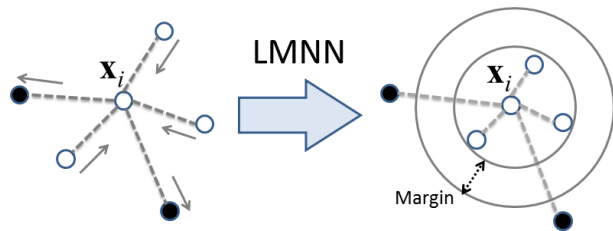


Figure 2. Large Margin Nearest Neighbor

In this method, the  $k$  neighbors of data  $\mathbf{x}_i$  are the  $k$  nearest neighbors that share the same label  $y_i$ , and the matrix  $\eta$  is defined as  $\eta_{ij} = 1$  if the input  $\mathbf{x}_j$  is a target neighbor of input  $\mathbf{x}_i$  and 0 otherwise. From these definitions, the cost function of LMNN is given by,

$$\varepsilon(\mathbf{A}) = \sum_{ij} \eta_{ij} \|\mathbf{A}\mathbf{x}_i - \mathbf{A}\mathbf{x}_j\|^2 + c \sum_{i|l} \eta_{ij} (1 - \eta_{il}) [1 + \|\mathbf{A}\mathbf{x}_i - \mathbf{A}\mathbf{x}_j\|^2 - \|\mathbf{A}\mathbf{x}_i - \mathbf{A}\mathbf{x}_l\|^2]_+, \quad (5)$$

where  $[\cdot]_+$  denotes the positive part, for example,  $[a]_+ = a$  if  $a > 0$ , and 0 otherwise, and  $c$  is some positive constant.

#### IV. WSD METHOD BASED ON DISTANCE METRIC LEARNING

In this section, we will describe the details of the WSD classifier using distance metric learning mentioned in the previous section.

##### A. Feature Extraction

At the first step, our method extracts a set of features; nouns and verbs that have co-occurred with the target word by morphological analysis from each sentence in the training and test data. Then, each feature set is represented as a vector by counting co-occurrence frequencies of the words. The set of word co-occurrence vectors forms a matrix for each target word.

##### B. Classification Model Construction

For the obtained this matrix, classification model is constructed by using distance metric learning method. The experiments in this paper use two learning methods such as NCA and LMNN to transform the data points. For the transformed data set using the NCA, we find optimal dividing hyperplane that will correctly classify the data points of the training data by using SVM. For the transformed data set using the LMNN, we apply one-nearest neighbor method in order to classify a new data point.

When the classification model is obtained by training data, we predict one sense for each test example using this model. When a new sentence including the target word is given, the sense of the target word is classified to the most plausible sense based on the obtained classification model. To employ the SVM for distinguishing more than two senses, we use

one-versus-rest binary classification approach for each sense. To employ the LMNN, we use the one-nearest neighbor (1-NN) classification rule to classify a test data set. The 1-NN method classifies a new sentence into the class of the nearest of the training data. Therefore, even if the target word has many senses, there is no need to repeat the classification process.

#### V. EXPERIMENTS

To evaluate the efficiency of the method of word sense disambiguation using the distance metric learning such as NCA and LMNN, we make some experiments to compare with the result of the SVM classification. In this section, we describe an outline of the experiments.

##### A. Data

We used the Semeval-2010 Japanese WSD task data set, which includes 50 target words comprising 22 nouns, 23 verbs, and 5 adjectives [6]. In this data set, there are 50 training and 50 test instances for each target word.

##### B. Evaluation Method

To evaluate the results of the methods using NCA and LMNN for the test data, we compare their performances with the results of simple SVM and 1-NN training. We obtain the total number of correct prediction of each target word using three methods: SVM, 1-NN, NCA+SVM and LMNN+1-NN. Moreover, we also obtain precision value of each method over all the examples to analyze the average performance of systems.

#### VI. EXPERIMENTAL RESULTS

##### A. Classification Performance

Table I and Table II show the results of the experiments of applying four methods. The proposed method using distance metric learning shows higher precision than the traditional one-nearest neighbor method. The distance metric learning provides an effective semantic relation between word senses so that this approach is effective for word sense disambiguation.

When NCA is applied to distance metric learning, the accuracy is increased on 9 words, decreased on ten words and the same on 31 words in comparison with SVM. Totally, NCA is not improved compared with SVM, because objective function of NCA tends to converge into a local optimum. To use the NCA for word sense disambiguation, further improvements are required for the prospective practical use. Examples of improvements include the use of a large data set, the use of other feature extraction methods or finding the optimal number of dimensions of projection etc.

When we use LFDA, we can not solve the generalized eigenvalue problem, since the co-occurrence matrix is very sparse. Hence, we apply SELF to their experiments instead of LFDA. The accuracy is increased on 1 word and the same

Table I  
EXPERIMENTAL RESULTS(1/2)

word	1-NN	SVM	SELF+ SVM	NCA+ SVM	LMNN+ INN
現場 (genba)	30	39	39	37	29
場所 (basyo)	48	48	48	48	48
取る (toru)	13	13	13	13	14
乗る (noru)	27	25	25	20	27
会う (au)	28	33	33	33	33
前 (mae)	24	31	31	29	27
子供 (kodomo)	26	18	18	21	26
関係 (kankei)	39	39	39	39	39
教える (oshieru)	15	9	9	9	13
勤める (susumeru)	20	16	16	16	27
社会 syakai)	40	43	43	43	42
する (suru)	18	21	21	23	20
電話 (denwa)	31	28	28	35	33
やる (yaru)	46	47	47	47	47
意味 (imi)	26	27	27	23	26
あげる (ageru)	15	18	18	18	17
出す (dasu)	18	14	14	17	26
生きる (ikiru)	47	47	47	47	47
経済 (keizai)	47	49	49	49	49
良い (yoi)	24	12	12	15	23
他 (hoka)	50	50	50	50	50
開く (hiraku)	45	45	45	45	45
もの (mono)	44	44	44	44	44
強い (tuyoi)	43	46	46	46	45
求める (motomeru)	39	38	38	38	39

on 49 words in comparison with SVM so that the experimental results of SVM and SELF are almost the same. LFDA obtains the optimal subspace that maximizes between-class and minimizes the within-class variance. However, this subspace is obtained by rotating and scaling the original coordinate space. Therefore, SVM produces the hyperplane equal to the transformation of it in the original space into the subspace obtained by LFDA.

When LMNN is applied to distance metric learning, precision of LMNN is slightly improved from 98.9% to 69.6% in comparison with SVM. It is possible to build a classification model that can perform better than NCA and SELF. Unlike NCA, we can obtain a global optimum solution by using LMNN so that we consider that LMNN is effective for word sense disambiguation.

### B. Efficiency of Distance Metric Learning

In traditional SVM classification, an additional process is required for extensive analysis on the relation between the new data and the training data. However, in the proposed method, we can perform such analysis easily. In contrast to SVM, we can retrieve the most similar sentence using one nearest neighbor for the input sentence.

To employ the SVM for classifying more than two senses, we solve multi-class classification problems by considering the standard one versus rest strategy. If the target word has more than two senses, it is difficult to compare the distance between the test data and its nearest neighbor. The LMNN

Table II  
EXPERIMENTAL RESULTS(2/2)

word	1-NN	SVM	SELF+ SVM	NCA+ SVM	LMNN+ INN
技術 (gijutu)	39	42	42	42	41
与える (ataeru)	21	29	29	28	25
市場 (shijou)	14	35	35	34	20
立つ (tatu)	18	26	26	22	16
手 (te)	41	39	39	39	40
考える (kangaeru)	49	49	49	49	49
見える (mieru)	19	26	26	23	23
一 (ichi)	45	46	46	46	46
入れる (ireru)	28	36	36	36	34
場合 (baai)	42	43	43	43	45
早い (hayai)	31	26	26	27	28
出る (deru)	22	30	30	30	28
入る (hairu)	20	25	25	26	34
はじめ (hajime)	38	30	30	33	44
情報 (johou)	39	40	42	37	32
大きい (ookii)	45	47	47	47	47
見る (miru)	39	40	40	40	40
可能 (kanou)	23	28	28	28	30
持つ (motu)	30	34	34	34	29
時間 (jikan)	43	44	44	42	44
文化 (bunka)	46	49	49	49	49
始める (hajimeru)	39	39	39	40	39
認める (mitomeru)	39	35	35	35	39
相手 (aite)	41	41	41	41	40
高い (takai)	26	43	43	43	43
precision	0.6544	0.6888	0.6896	0.6876	0.6964

method employs one nearest neighbor rule and can calculate the distance to its nearest neighbor for each sense. Therefore, the proposed method is effective for analyzing the relation between the input sentence and all senses of the target word. Also, this method is effective for identifying uncommon word senses of target words.

## VII. CONCLUSION

In this paper, we propose a novel word sense disambiguation method based on a distance metric learning to find the most similar sentence. To evaluate the efficiency of the method of word sense disambiguation using the distance metric learning such as NCA and LMNN, we make some experiments to compare with the result of the SVM classification. The results of the experiments show this method is effective for word sense disambiguation in comparison with SVM and one nearest neighbor. Moreover, the proposed method is effective for analyzing the relation between the input sentence and all senses of the target word if the target word has more than two senses.

Further work would be required to consider more effective re-mapping method of the training data to improve the performance of word sense disambiguation.

## REFERENCES

- [1] E. Agirre, O. Lopez, and D. Martínez, "Exploring feature spaces with svd and unlabeled data for word sense disambiguation," in *In Proceedings of the Conference on Recent*

*Advances on Natural Language Processing (RANLP '05)*, Borovets, Bulgaria, 2005.

- [2] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.
- [3] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov, "Neighborhood Component Analysis," in *Proceedings of Advances of Neural Information Processing*, 2004.
- [4] R. Izquierdo, A. Suárez, and G. Rigau, "An empirical study on class-based word sense disambiguation," in *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, ser. EACL '09, 2009, pp. 389–397.
- [5] R. Navigli, "Word sense disambiguation: A survey," *ACM Computing Surveys*, vol. 41, no. 2, pp. 10:1–10:69, Feb. 2009.
- [6] M. Okumura, K. Shirai, K. Komiya, and H. Yokono, "Semeval-2010 task: Japanese wsd," in *Proceedings of the 5th International Workshop on Semantic Evaluation*, ser. SemEval '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 69–74.
- [7] M. Sugiyama, "Local fisher discriminant analysis for supervised dimensionality reduction," in *Proceedings of the 23rd international conference on Machine learning*, ser. ICML '06. New York, NY, USA: ACM, 2006, pp. 905–912.
- [8] M. Sugiyama, T. Idé, S. Nakajima, and J. Sese, "Semi-supervised local fisher discriminant analysis for dimensionality reduction," *Machine Learning*, vol. 78, pp. 35–61, January 2010.
- [9] M. Sugiyama and S. Roweis, "Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis," *Journal of Machine Learning Research*, vol. 8, pp. 1027–1061, 2007.
- [10] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *The Journal of Machine Learning Research*, vol. 10, pp. 207–244, June 2009.

## Searching Documents with Semantically Related Keyphrases

Ibrahim Aygul, Nihan Cicekli  
 Department of Computer Engineering  
 Middle East Technical University  
 Ankara, Turkey

ibrahimaygul@gmail.com, nihan@ceng.metu.edu.tr

Ilyas Cicekli  
 Department of Computer Engineering  
 Hacettepe University  
 Ankara, Turkey  
 ilyas@cs.hacettepe.edu.tr

**Abstract** — In this paper, we present a tool, called SemKPSearch, for searching documents by a query keyphrase and keyphrases that are semantically related with that query keyphrase. By relating keyphrases semantically, we aim to provide users an extended search and browsing capability over a document collection and to increase the number of related results returned for a keyphrase query. Keyphrases provide a brief summary of the content of documents, and they can be either author assigned or automatically extracted from the documents. SemKPSearch uses a set of keyphrase indexes called SemKPIndex, and they are generated from the keyphrases of documents. In addition to a keyphrase-to-document index, SemKPIndex also contains a keyphrase-to-keyphrase index which stores semantic relation scores between the keyphrases in a document collection. The semantic relation score between keyphrases is calculated using a metric which considers the similarity score between words of the keyphrases, and the semantic similarity score between two words is determined with the help of two word-to-word semantic similarity metrics based on WordNet. SemKPSearch is evaluated by human evaluators, and the evaluation results showed that the evaluators found the documents retrieved with SemKPSearch more related to query terms than the documents retrieved with a search engine.

*Keywords-keyphrase extraction; semantic similarity; information retrieval; digital library.*

### I. INTRODUCTION

The number of documents available electronically has increased dramatically and the use of large document collections such as digital libraries has become widespread. Browsing a document collection and finding the documents of interest turns out to be more difficult. The full-text inverted indexes and ranking algorithms cause standard search engines often return a high number of results, and it is an overwhelming process to find whether a collection covers the useful information.

Gutwin et al. state that full-text indexing has several problems in browsing a collection [6]. First, although users can retrieve documents containing the words of user's query text, they usually use short topic phrases to explore a collection. The second problem stated by Gutwin et al. [6] is the result set. Standard search engines return a list of documents which is too specific for browsing purposes. Lastly, with the nature of browsing, the third problem is the query refinement, and standard engines do not support

constituting new queries. For the solution to these problems, Gutwin et al. propose a search engine "Keyphind", which is especially designed to help browsing document collections [6]. Keyphind uses keyphrase indexes in order to allow users to interact with the document collection at the level of topics and subjects. Keyphrases provide a brief description of a document's content and can be viewed as semantic metadata that summarize documents. Keyphrases are widely used in information retrieval systems [4] [5] [7] [9] [11] and other document browsing systems [8] [15]. With the help of the keyphrases of documents in the collection, the user can easily guess the coverage of documents and browse the relevant information.

In this paper, we present a keyphrase-based search engine, called SemKPSearch, using a set of keyphrase based indexes which is similar to the Keyphind index, for browsing a document collection. With the help of keyphrase indexes, the user can browse documents which have semantically related keyphrases with the query text. In this work, we extend the keyphrase index with a novel keyphrase to keyphrase index which stores the evaluated semantic similarity score between the keyphrases of the documents in a collection. To calculate similarity scores between keyphrases, we use the text semantic similarity measure given in [3], which employs a word-to-word similarity measure. We use a word-to-word semantic similarity metric [12] in the calculation of keyphrase similarities.

To evaluate SemKPSearch, we used a test corpus that is collected by Krapivin et al. [10]. The corpus has full-text articles and author assigned keyphrases. We also used the keyphrase extraction system KEA [16] to evaluate the system with automatically extracted keyphrases. We created keyphrase indexes for both author assigned and automatically extracted keyphrases. To determine the retrieval performance of SemKPSearch, we have evaluated SemKPSearch with Google Desktop search tool which uses full-text index. The evaluation is done by human testers, and evaluation results showed that SemKPSearch suggests valuable and helpful keyphrases that are semantically related with the query of the tester and the document retrieval performance is better than Google Desktop.

Section 2 describes the overall structure of SemKPSearch in addition to its index structure and generation. In Section 3, the evaluation methods and experimental results are presented. Section 4 concludes the paper and discusses the future work.

## II. SEARCHING WITH SEMANTICALLY RELATED KEYPHRASES

The searching and browsing interface of SemKPSearch is developed for querying documents in a digital library using their keyphrases. A keyphrase based index, SemKPIndex, is created for a document collection and SemKPSearch uses SemKPIndex for querying and browsing the collection in a user friendly interface. In SemKPSearch, browsing is also aided by suggesting keyphrases that are semantically related with the given query. As the documents in the collection are indexed by their keyphrases, semantically related keyphrases are indexed with a score which is calculated by employing a semantic similarity metric. We use two semantic similarity metrics to calculate a semantic similarity score between keyphrases.

The overall structure of SemKPSearch system is shown in Figure 1. A document collection with their keyphrases is the main input to SemKPSearch. If the documents in the collection do not have author assigned keyphrases, KEA [16] is employed to extract keyphrases. In addition to indexes between keyphrases and documents in SemKPIndex, each indexed keyphrase is compared to all other keyphrases and a similarity score is calculated, and then semantically related keyphrases are also stored in SemKPIndex. Using SemKPIndex on the SemKPSearch interface, the users query the document collection with topic like keyphrases, and the interface returns a set of document results that contains query term among their keyphrases. Besides the documents that contain query term in their keyphrases, SemKPSearch suggests semantically related keyphrases using SemKPIndex, and the users can expand search results by using these suggested keyphrases.

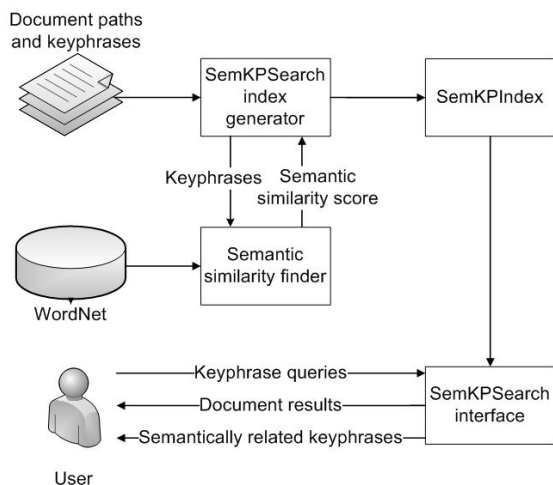


Figure 1. Overall structure of SemKPSearch system.

### A. SemKPIndex Structure

SemKPSearch uses a set of indexes, called as SemKPIndex, and it is composed of five indexes: keyphrase list, document to keyphrase index, keyphrase to document index, word to keyphrase index and keyphrase to keyphrase

index. The first four indexes are very similar to the structure of Keyphind index [6], and the fifth one is our new novel index structure. The last index is a keyphrase to keyphrase index which holds semantically related keyphrases.

Keyphrase list is a list of all keyphrases that are given with the documents in the collection. This index is used as a suggestion list that guides the user with possible keyphrases as the user enters the query terms.

Document to keyphrase index contains information for each document in the collection. Each keyphrase is kept with a relation score that shows the importance of the keyphrase to the owner document. If no relation score is given for the keyphrase, it is automatically calculated during index generation. Document to keyphrase index is used to improve the search results by showing each document with its keyphrases and to order the documents in the search result.

Keyphrase to document index is a mapping from all keyphrases to the paths of the owner documents. It is somehow the inverse of the document to keyphrase index. This index is used to retrieve the documents that have a given keyphrase among its keyphrases.

Word to keyphrase index contains all words in all of the keyphrases, and each entry corresponds to the keyphrases containing the entry word. This index is needed to show the user more results and more keyphrases to extend the search. For example, when the user searches “similarity”, in addition to the documents that contain the keyphrase “similarity”, the documents containing the keyphrases “semantic similarity”, “similarity measurement”, “similarity retrieval” will be retrieved by the help of this index.

Keyphrase to keyphrase index provides the main contribution in the study, and the aim of this index is to aid users in their searches by suggesting semantically related keyphrases with query terms. The index keeps semantic relations between keyphrases in the keyphrase list. During the index generation, a semantic relation score is calculated for each pair of keyphrases in the system, and the relations that exceed a predefined threshold value are stored in this index. Each entry is a mapping from a keyphrase to its semantically related keyphrase list. For example, the index entry for the keyphrase “face recognition” in the test collection contains its semantically related keyphrases such as “face recognition algorithm”, “shape recognition”, and “identification system” together with their semantic relation scores.

The keyphrase to keyphrase index gives the user a chance to see the semantically related keyphrases with the search terms. It also helps to extend search results with the suggested semantically related keyphrases. If the search term is a keyphrase in the index, the suggested related keyphrases are obtained from the index entry of that keyphrase. On the other hand, the suggested semantically related keyphrases are produced on the fly by comparing the search term with the keyphrases in the index when the search term is not available in the index.

### B. Generating SemKPIndex

SemKPSearch accepts a collection of documents and their keyphrases as inputs to the index generation process.

The keyphrases can be assigned by the authors or automatically extracted from the documents using a keyphrase extraction algorithm. The documents with their keyphrases are indexed one by one during index generation. For each document, the keyphrases of the document are added to keyphrase list. Then by using these keyphrases, other indexes are created.

The keyphrases of a document are added to document to keyphrase index together with their relation scores. If the keyphrases are found by the keyphrase extraction algorithm, their relation scores are also found. For the author assigned keyphrases, their relation scores are found relative to their positions in the keyphrase list of the document. The relation score of the  $i^{\text{th}}$  keyphrase of a document with  $n$  keyphrases is equal to  $1-(i/n)$ . Using this formula we assume that the author assigned keyphrases are given by the relevance order and the last keyphrases in the list are much less related with the document than the first keyphrases.

After creating document to keyphrase index, the keyphrases of the document are added to keyphrase to document index, and each entry in this index points to a document list sorted with relation scores. Index generation continues by adding each word of the keyphrases to the word to keyphrase index, and each word entry in the index points to a keyphrase list that gives a reference to keyphrases in which the word occurs.

After keyphrase list is created, a list of related keyphrases is created for each keyphrase in order to create keyphrase to keyphrase index. A semantic relation score is calculated for each pair of keyphrases, and top keyphrases which passes a predefined threshold semantic relation score are kept as a list of related keyphrases for each keyphrase. Each related keyphrase list is sorted with respect to the relation scores.

The semantic relatedness of two keyphrases can be calculated the same as the semantic similarity between two texts are calculated, and several methods to find the semantic similarity between two texts are discussed in the literature [3] [12] [13] [14]. The similarity between two keyphrases is based on the similarity of their words, and Corley and Mihalcea introduce a metric that combines word-to-word similarity metrics into a text-to-text semantic similarity metric [8]. In this approach, the value of the semantic similarity between two texts is calculated using the semantic similarities of words and inverse document frequencies of words. In our study, we use Corley and Mihalcea approach to calculate the semantic similarity between two keyphrases together with the WordNet based word-to-word similarity metric proposed by Li et al. [12].

In order to find the semantic similarity between two keyphrases using the discussed similarity metrics, first, we create a similarity matrix for the words of the keyphrases. All words of one keyphrase are compared to each word of the other keyphrases, and a similarity score for two words is found. Since keyphrases are short texts, it is not feasible to detect part of speech tags of a bunch of words. Besides, keyphrases of documents generally consist of nouns or verbs. Thus, for word comparisons, words are compared using their noun and verb senses in WordNet and whichever sense pair

produces higher similarity score, it is chosen as the similarity score of those words.

### III. EVALUATION

In order to evaluate the retrieval performance and the related keyphrase suggestions of SemKPSearch, we used a test corpus that is collected by Krapivin et al. [10]. The corpus contains 2304 papers from Computer Science domain, which were published by ACM between 2003 and 2005. It has full-text of articles and author assigned keyphrases.

We created two SemKPSearch indexes for the test corpus. The first index was created with author assigned keyphrases and the other index was created with KEA extracted keyphrases. In order to extract keyphrases automatically using KEA, 30 documents were randomly selected from the corpus and their author assigned keyphrases were given to KEA to build its training model. Then for each document in the corpus, KEA extracted 5 keyphrases which were up to 2 to 5 words. These keyphrases were selected to be used in the creation of the index. Since a one word length keyphrase may be too general, we chose keyphrases with at least 2 words in order to be able to obtain more precise keyphrases. In addition to these two SemKPIndexes, a full text index over the same corpus was created by Google Desktop [1] in order to compare SemKPSearch with Google Desktop.

We used two different word-to-word semantic similarity metrics in the calculation of the semantic relatedness of keyphrases. The first one was Wu and Palmer [17] word-to-word similarity metric, and the other one was the word similarity measure introduced by Li et al. [12]. We have tested our system with these two word-to-word similarity metrics. Since the performance of the system was better when Li et al. semantic similarity was used, here we only give the performance results of the system with this metric. We called the two created SemKPIndexes as KEA\_Sim<sub>Li</sub> in which KEA extracted keyphrases and Li et al. similarity metric were used, and Author\_Sim<sub>Li</sub> in which author assigned keyphrases and Li et al. similarity metric were used.

The user evaluation was done by 8 human evaluators who were all computer scientists. Each evaluator evaluated the relevancy of the keyphrases suggested by SemKPSearch, and the documents retrieved by SemKPSearch and Google Desktop. They gave a relevance score between 0 and 4 (0:irrelevant, 1:poorly relevant, 2:partially relevant, 3:relevant, 4:completely relevant) to each retrieved document and to each suggested keyphrase according to their relevancy to the query term. Each evaluator created his own two sets of query terms by randomly selecting terms from the two given sets of query terms. The first set contains query terms which occur as keyphrases of the documents in the collection, and the second set contains query terms which do not occur as keyphrases in the collection. This means that there is no document which is indexed by a query term in the second set. The results reported here are the average scores of the 8 evaluators.

TABLE I. AVERAGE SCORES FOR THE FIRST K SUGGESTED KEYPHRASES

Index	Avg@1	Avg@3	Avg@5	Avg@10
KEA_Sim <sub>Li</sub>	3,34	3,21	3,04	2,80
Author_Sim <sub>Li</sub>	3,69	3,42	3,08	2,81

### A. Keyphrase Suggestion Success

The performance of the semantically similar keyphrase suggestion of the system is discussed by calculating the average score of the evaluator scores for the first 10 suggested keyphrases. Table 1 gives the average scores for the first k keyphrase suggestions where  $k \in \{1,3,5,10\}$ . According to the results in Table 1, Author\_Sim<sub>Li</sub> achieves better results than KEA\_Sim<sub>Li</sub>. This is an expected outcome, since author assigned keyphrases may be more meaningful from the automatically extracted keyphrases. Although, Author\_Sim<sub>Li</sub> index has better suggestion results, KEA\_Sim<sub>Li</sub> index results are still competitive. Considering that in real life applications most of the documents in a collection do not have author assigned keyphrases, we can argue that keyphrase suggestion can be done with the automatically extracted keyphrases. Of course, if author assigned keyphrases are available for a collection, they can be used for better performance. The average scores for the first 3 suggested keyphrases indicate that a big percentage of these 3 suggested keyphrases has a relevance score above 3. This means that the first three suggested keyphrases are relevant with the query term.

### B. Document Retrieval Success

In order to measure document retrieval success, SemKPSearch configured with KEA\_Sim<sub>Li</sub> index was compared to Google Desktop on the same document collection. The document retrieval performances of the two systems were compared with the relevance scores for the retrieved documents given by the evaluators. Each evaluator randomly selected query terms from a set of keyphrases appearing in the SemKPSearch index and a set of query terms not appearing in the index. During scoring SemKPSearch, if the result set contained less than 10 documents, the evaluators expanded the result set by using the suggested keyphrases until reaching 10 documents. If the query text was not indexed in SemKPIIndex, then semantically related keyphrases are calculated on the fly by comparing the query text to all keyphrases. Since our evaluation results indicate that the first three suggested keyphrases are very relevant with a given query term, the evaluators first used the documents retrieved for three suggested keyphrases for expansion in the suggestion order. If they did not reach ten documents, they used a single document from other suggested keyphrases.

Table 2 presents the average relevance scores, mean reciprocal rank (MRR) values and precision values for both systems. Table 2.a shows the evaluation results for the documents returned for keyphrase queries which were indexed by the evaluated SemKPIIndex. In other words there was at least one document such that the queried term is its keyphrase. Table 2.b shows the evaluation results for queries

that do not occur as keyphrases. The average relevance scores are the averages of the evaluator scores for documents. The reciprocal rank of a query result list is equal to  $1/rank_{fc}$  where  $rank_{fc}$  is the position of the first correct answer in the result list, and we treat the retrieved documents with scores 4 and 3 (completely relevant and relevant) as correct answers. The MRR value of a query set is the average of the reciprocal ranks of the queries in the set. The precision value is the percentage of correct answers in the retrieved document set.

TABLE II. EVALUATION RESULTS TO COMPARE DOCUMENT RETRIEVAL PERFORMANCE OF SEMKPSEARCH AND GOOGLE DESKTOP

#### a) Searching with keyphrases indexed in SemKPIIndex

first n docs.	SemKPSearch			Google Desktop		
	Avg. Score	MRR	Pre.	Avg. Score	MRR	Pre.
1	3,95	1,00	1,00	3,05	0,70	0,70
3	3,57	1,00	0,83	2,94	0,83	0,67
5	3,32	1,00	0,78	2,74	0,83	0,56
7	3,04	1,00	0,70	2,49	0,83	0,49
10	2,74	1,00	0,62	2,15	0,83	0,40

#### b) Searching with phrases not indexed in SemKPIIndex

first n docs.	SemKPSearch			Google Desktop		
	Avg. Score	MRR	Pre.	Avg. Score	MRR	Pre.
1	2,04	0,43	0,43	2,14	0,29	0,29
3	1,93	0,50	0,33	1,81	0,29	0,25
5	2,01	0,54	0,34	1,85	0,29	0,21
7	1,71	0,54	0,25	1,90	0,31	0,25
10	1,71	0,54	0,21	1,73	0,31	0,22

According to Table 2.a, the documents retrieved with SemKPSearch get higher average scores than the documents returned by Google Desktop. Since this table is for the evaluation of the results with the keyphrases indexed in SemKPIIndex, one can argue that this is the success of the keyphrase extraction algorithm. The results in the first orders get apparently high scores because they are the directly returned documents having the search term as one of their keyphrases. With a further analysis of the raw results we see that for all queried keyphrases, the number of directly returned documents is 2,4 out of 10 on the average, and 76% of the evaluated documents are returned by assisting the query with semantically related keyphrases. The average score for the documents that are retrieved by the suggested keyphrases is 2,47. On the other hand, the average score for the last 8 documents out of 10 retrieved by Google Desktop is 1,9. MRR and Precision values on Table 2.a are similar to the average scores, and SemKPSearch beats Google Desktop. Here we see that the MRR value for SemKPSearch is 1, which means that for all queries, SemKPSearch returned

a relevant document to the query term at the first place. Actually this result comes from the success of the keyphrase extraction algorithm KEA because the first document has always the query term as its keyphrase extracted by KEA. These values reasonably show us that using keyphrases of documents, the document retrieval with SemKPSearch is more successful than Google Desktop.

In Table 2.b, a slightly different result is seen for the documents returned for the phrases not indexed in SemKPIndex. The average scores are a bit lower for the SemKPSearch results. However MRR and precision values show that for the queries with phrases that are not indexed as a keyphrase of a document, related documents appear on the higher orders in SemKPSearch.

Although Keyphind system [6] is not tested with our data set, we can still compare it with the results of our system. Keyphind returns the documents if the searched keyphrase is available in its index. But, it does not return any documents if the searched keyphrase is not available in its index. For this reason, Keyphind system would not have returned any documents for the searched keyphrases in Table 2.b since those keyphrases would not have been in Keyphind index. On the other hand, our SemKPSearch system returns the documents using the semantically related keyphrases. If there are enough documents associated with the searched keyphrase in a digital library, the performance of SemKPSearch configured with KEA\_Sim<sub>Li</sub> index will be similar to the performance of Keyphind since both use KEA to extract keyphrases. When there are not enough documents associated with the searched keyphrase, Keyphind will return only associated documents while SemKPSearch returns additional documents using semantically related keyphrases in addition to the documents associated with the searched keyphrase.

In Table 2.a, the average number of returned documents that are directly associated with searched keyphrase is 2,4 out of 10 documents, the rest of the returned documents are associated with semantically related keyphrases. The average score of the documents associated with searched keyphrase is 3,78 and the average score of the documents associated with semantically related keyphrases is 2,47. With a further analysis, the average score of the first results associated with semantically related keyphrases is 3,47, and the average score for the first three results associated with semantically related keyphrases is 3,01. These results indicate that the first results associated with semantically related documents are actually related with the searched keyphrase. These results also indicate that Keyphind system would have returned only 2,4 documents on the average for the keyphrases in Table 2.a and its average score will be similar to our average score (3,78). But, SemKPSearch returns 3 more related documents associated with semantically related documents with average score 3,01.

#### IV. CONCLUSION

In this paper, we proposed SemKPSearch system which has a user friendly search and browsing interface for querying documents by their keyphrases in a digital library.

SemKPSearch indexes the documents with their keyphrases in SemKPIndex. Through the user interface of SemKPSearch, the user can search documents with topic like query phrases. SemKPSearch returns keyphrases that are semantically related to the query text, as well as the documents having keyphrases containing the query text. The user can continue to browse more documents with the suggested semantically related keyphrases or with the keyphrases of the retrieved documents. In this way, it is expected that the user can reach the related documents with the query text even if the documents do not contain the query term.

To calculate the semantic similarity between keyphrases, we propose to use a text-to-text semantic similarity metric that is proposed by Corley and Mihalcea [3]. This metric employs a word-to-word semantic similarity measure, and we used Li et al. word-to-word similarity measure [12]. Thus, the semantic similarity of the keyphrases is formulated as a function of the similarity of the words of the keyphrases.

The evaluation of the system was done by the human evaluators. The evaluators judged the quality of the results and the effectiveness of the suggested semantically related keyphrases. In order to evaluate the document retrieval performance, SemKPSearch system was compared to Google Desktop which is a full-text index based search engine. The evaluation results showed that the evaluators found the documents retrieved with SemKPSearch more related to the query term than the documents retrieved with Google Desktop. Besides the document retrieval, the semantically related keyphrase suggestions were also evaluated by the evaluators. According to the results obtained for the related keyphrase suggestions, it is feasible to use the automatically extracted keyphrases and to relate them with the keyphrase semantic similarity that we proposed.

#### REFERENCES

- [1] Google Desktop - Features. <http://desktop.google.com/features.html>, retrieved: January, 2012.
- [2] WordNet - About WordNet. <http://wordnet.princeton.edu>, retrieved: July, 2012.
- [3] C. Corley and R. Mihalcea. Measuring the semantic similarity of texts. In Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment, pages 13–18. Association for Computational Linguistics, 2005.
- [4] W.B. Croft, H.R. Turtle, and D.D. Lewis. The use of phrases and structured queries in information retrieval. In Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval, pages 32–45. ACM, 1991.
- [5] J.L. Fagan. The effectiveness of a nonsyntactic approach to automatic phrase indexing for document retrieval. Journal of the American Society for Information Science, 40(2):115–132, 1989.
- [6] C. Gutwin, G. Paynter, I. Witten, C. Nevill-Manning, and E. Frank. Improving browsing in digital libraries with keyphrase indexes. Decision Support Systems, 27(1-2):81–104, 1999.
- [7] S. Jones. Design and evaluation of phrasier, an interactive system for linking documents using keyphrases. In Proceedings of Human-Computer Interaction: INTERACT'99, pages 483–490, 1999.
- [8] S. Jones and G. Paynter. Topic-based browsing within a digital library using keyphrases. In Proceedings of the fourth ACM conference on Digital libraries, page 121. ACM, 1999.



- [9] S. Jones and M.S. Staveley. Phrasier: a system for interactive document retrieval using keyphrases. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pages 160–167. ACM, 1999.
- [10] M. Krapivin, A. Autaeu, and M. Marchese. Large Dataset for Keyphrases Extraction. Technical Report DISI-09-055, DISI, University of Trento, Italy, 2009.
- [11] Q. Li, YB Wu, R.S. Bot, and X. Chen. Incorporating document keyphrases in search results. In Proceedings of the Americas Conference on Information Systems (AMCIS), New York, 2004.
- [12] Y. Li, Z.A. Bandar, and D. McLean. An approach for measuring semantic similarity between words using multiple information sources. IEEE Transactions on knowledge and data engineering, pages 871–882, 2003.
- [13] Y. Li, D. McLean, Z.A. Bandar, J.D. O’Shea, and K. Crockett. Sentence similarity based on semantic nets and corpus statistics. IEEE Transactions on Knowledge and Data Engineering, pages 1138–1150, 2006.
- [14] R. Mihalcea, C. Corley, and C. Strapparava. Corpus-based and knowledge-based measures of text semantic similarity. In Proceedings of the 21st national conference on Artificial intelligence-Volume 1, pages 775–780. AAAI Press, 2006.
- [15] N. Wacholder, D.K. Evans, and J.L. Klavans. Automatic identification and organization of index terms for interactive browsing. In Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries, page 134. ACM, 2001.
- [16] I.H. Witten, G.W. Paynter, E. Frank, C. Gutwin, and C.G. Nevill-Manning. KEA: Practical automatic keyphrase extraction. In Proceedings of the fourth ACM conference on Digital libraries, page 255. ACM, 1999.
- [17] Z. Wu and M. Palmer. Verbs semantics and lexical selection. In Proceedings of the 32nd annual meeting on Association for Computational Linguistics, pages 133–138. Association for Computational Linguistics, 1994.

# Optimizing Geographical Entity and Scope Resolution in Texts using Non-Geographical Semantic Information

Panos Alexopoulos and Carlos Ruiz  
*iSOCO, Intelligent Software Components S.A.*  
 Av. del Partenon, 16-18, 28042, Madrid, Spain,  
 Email: {palexopoulos, cruiz}@isoco.com

**Abstract**—Assigning geographical meta-information to textual pieces of information in an automatic way is a challenging semantic processing task that has been getting increasing attention from application and research areas that need to exploit this kind of information. With that in mind, in this paper, we propose a novel ontology-based framework for correctly identifying geographical entity references within texts and mapping them to corresponding ontological uris, as well as determining the geographical scope of texts, namely the areas and regions to which the texts are geographically relevant. Unlike other approaches which utilize only geographical information for performing these tasks, our approach allows the exploitation of any kind of semantic information that is explicitly or implicitly related to geographical entities in the given domain and application scenario. This exploitation, according to our experiments, manages to substantially improve the effectiveness of the geographical entity and scope resolution tasks, especially in scenarios where explicit geographical information is scarce.

**Keywords**—*Location Disambiguation; Geographical Scope Resolution; Ontologies.*

## I. INTRODUCTION

With the rapidly increasing popularity of Social Media sites, a lot of user-generated content has been injected in the Web resulting in a large amount of both multimedia items and textual data (tags and other text-based documents) [1]. As a consequence, it has become increasingly difficult to find exactly the objects that best match the users' information needs. Besides, as more of those searches are performed from mobile applications, geographic intent and scope become indispensable as users expect a search system not only to know their current location, but to understand their entire geographic context. Therefore, it is crucial for the system to be able to infer what is the location (if any) implicit in their search and the user-generated content.

Thus, Geographical Intention Retrieval [2] concerns all kinds of techniques related to the retrieval of information involving some kind of spatial awareness. These methods can improve all kinds of services and applications that rely on geographical information, ranging from its quite straightforward use in map services, to more advanced techniques of personalization. For example, a user searching for cheap flights to Paris has the implicit intent of flying from his current location, although the latter was not stated.

This implicit geographic nature of user queries is called geographic intent.

On the other hand, a text or a query has a geographic scope. For example, a query for cheap flights from London to Paris would include both London and Paris in the geographic scope, but not locations in between. Similarly, a text describing the Eiffel tower will have the geographic scope of Paris, rather than of France, although both locations could be mentioned in the tag set.

Geo-location services enable retrieval of likely geographical locations for given keywords or text [3]. Most of them apply data mining and statistical techniques on big-scale data sets in the Internet, nevertheless they rely only in syntactic analysis, missing the benefits of exploiting the real meaning of a piece of text. This leaves them suffering issues such as disambiguation problems with locations with the same name (Paris, France vs. Paris, Texas) or locations named somehow similar to non-geographic concepts (such as Reading, UK).

On the other hand, semantic analysis, either built on top of statistical analysis or as a standalone approach, can improve the previous approach by extracting not only geographical entities from a text, but also other types of entities (people, companies, etc.) that can, via reasoning or inference techniques, extract further geographic information.

Of course, the main limitation of semantic approaches is the need for geographical knowledge bases as input to the system, typically a bottleneck in the whole process. Previous approaches have tried to build geographic knowledge on top of different kind of resources, including ad hoc ontologies, geo-gazetteers or more generic knowledge hubs such as Wikipedia. However, the reuse of Open Data is a key element for improving this approach, avoiding or at least limiting the initial entry barriers for geographical semantic analysis. In particular, the Linked Data initiative [4] provides a crucial starting point for building a large and reliable geographical centered knowledge base, with enough information from other type of entities to allow for a comprehensive coverage of most domains.

Given the above, in this paper, we focus on geographical analysis of textual information and we propose a novel ontology-based framework for tackling two problems:

- 1) The problem of **geographical entity resolution**,

namely the detection within a text of geographical entity references and their correct mapping to ontological uris that represent them.

- 2) The problem of **geographical scope resolution**, namely the determination of areas and regions to which the text is geographically relevant.

The distinguishing characteristic of this framework is that, unlike other ontology-based approaches which utilize only geographical information for performing the above tasks, it allows the exploitation of any kind of semantic information that is explicitly or implicitly related to geographical entities in the given domain and application scenario. In that way, it manages to significantly improve the accuracy of the the above tasks, especially in domains and scenarios where explicit geographical information is scarce.

The rest of the paper is as follows. Section II presents related works. Section III presents our proposed framework and its components while Section IV presents and discusses experimental results regarding the evaluation of the framework's effectiveness. Finally, conclusions are drawn in Section V.

## II. RELATED WORK

Most related approaches to our work originate from the area of geographical information retrieval [2], where several approaches based on information retrieval, machine learning or semantic techniques are proposed to resolve geographic entities and scope.

Andogah et al. [5] describe an approach to place ambiguity resolution in text consisting of three components; a geographical tagger, a geographical scope resolver, and a placename referent resolver. The same authors, in [6], also propose determining the geographical scope as means to improve the accuracy in relevance ranking and query expansion in search applications. However these processes only rely on limited geographical information rather than using some other data available.

More related to the process of attempting to discern whether a texts topic is location-related, Mei et al. [7] present methods for finding latent semantic topics over locations (states or countries) and Wang et al. [8] propose a Location-Aware Topic Model based on Latent Dirichlet Allocation [9].

Besides, some other general approaches related to location disambiguation and inference are based on a query expansion process that augments a user's query with additional terms in order to improve the results, plus a filtering process for determining the relevance of results to the original query. For that, different dimensions can be taken into account in terms of how the relevance should be measured, ranging from its accuracy in a particular context to the inner meaning between terms. There are two primary query expansion approaches [10], [11]: on the one hand, probabilistic approaches sample from terms that co-occur with the original query as the basis for the expansion in

a local or global context, and, on the other hand, the use of ontologies by semantic approaches for query expansion relies on the formal and strongly defined structure they introduce, exploiting the existent relations between different concepts and entities.

Following a strict semantic approach, Kauppinen et al. [12] present an approach using two ontologies (SUO - a large Finnish place ontology, and SAPO - a historical and geographical ontology) and logic rules to deal with heritage information where modern and historical information is available (e.g., new name for a place, new borders in a country). This method is combined with some faceted search functionalities, but they do not propose any method for disambiguating texts.

More related to the fact that the disambiguation of a location depends on the context (such as in "London, England" vs. "London, Ontario"), Peng et al. [13] propose an ontology-based method based on local context and sense profiles combining evidence (location sense context in training documents, local neighbor context, and the popularity of individual location sense) for such disambiguation.

## III. PROPOSED GEOGRAPHICAL SCOPE RESOLUTION FRAMEWORK

Our proposed framework targets the two tasks of geographical entity and scope resolution based on a common assumption: that the existence of both geographical and non-geographical entities within a text may be used as **evidence** that indicate which is the most probable meaning of an ambiguous location term as well as which locations constitute the geographical scope of the whole text.

To see why this assumption is valid, consider a historical text containing the term "Tripoli". If this term is collocated with terms like "*Siege of Tripolitsa*" and "*Theodoros Kolokotronis*" (the commander of the Greeks in this siege) then it is fair to assume that this term refers to the city of Tripoli in Greece rather than the capital of Libya. Also, in a historical text like "*The victory of Greece in the Siege of Tripolitsa under the command of Kolokotronis was decisive for the liberation from Turkey*", the evidence provided by "*Siege of Tripolitsa*" and "*Kolokotronis*" and "*Greece*" indicates that Tripoli is more likely to be the location the text is about rather than Turkey.

Of course, which entities and to what extent may serve as evidence in a given application scenario depends on the domain and expected content of the texts that are to be analyzed. For example, in the case of historical texts we expect to use as evidence historical events and persons that have participated in them. For that reason, our approach is based on the a priori determination and acquisition of the optimal evidential knowledge for the scenario in hand. This knowledge is expected to be available in the form of an ontology and it's used within the framework in order to perform geographical entity and scope resolution. In

particular, our proposed framework comprises the following components:

- A **Geographical Resolution Evidence Model** that contains both geographical and non-geographical semantic entities that may serve as location-related evidence for the application scenario and domain at hand. Each entity is assigned evidential power degrees which denote its usefulness as evidence for the two resolution tasks.
- A **Geographical Entity Resolution Process** that uses the evidence model to detect and extract from a given text terms that refer to locations. Each term is linked to one or more possible location uris along with a confidence score calculated for each of them. The uri with the highest confidence should be the correct location the term refers to.
- A **Geographical Scope Resolution Process** that uses the evidence model to determine, for a given text, the location uris that potentially fall within its geographical scope. A confidence score for each uri is used to denote the most probable locations.

In the following paragraphs, we elaborate on each of the above components.

#### A. Geographical Resolution Evidence Model

For the purpose of this paper, we define an ontology as a tuple  $O = \{C, R, I, i_C, i_R\}$  where

- $C$  is a set of concepts.
- $I$  is a set of instances.
- $R$  is a set of binary relations that may link pairs of concept instances.
- $i_C$  is a concept instantiation function  $C \rightarrow I$ .
- $i_R$  is a relation instantiation function  $R \rightarrow I \times I$ .

Given an ontology, the **Geographical Resolution Evidence Model** defines which ontological instances and to what extent should be used as evidence towards i) the correct meaning interpretation of a location term to be found within the text and ii) the correct geographical scope resolution of the whole text. More formally, given a domain ontology  $O$  and a set of locations  $L \subseteq I$ , a geographical resolution evidence model consists of two functions:

- A **location meaning evidence function**  $lmef : L \times I \rightarrow [0, 1]$ . If  $l \in L$  and  $i \in I$  then  $lmef(l, i)$  is the degree to which the existence, within the text, of  $i$  should be considered an indication that  $l$  is the correct meaning of any text term that has  $l$  within its possible interpretations.
- A **geographical scope evidence function**  $gsef : L \times I \rightarrow [0, 1]$ . If  $l \in L$  and  $i \in I$  then  $gsef(l, i)$  is the degree to which the existence, within the text, of  $i$  should be considered an indication that  $l$  represents the geographical scope of the text.

In order to determine the above functions for a given domain and scenario we need to consider the concepts whose

instances are directly or indirectly related to locations and which are expected to be present in the text to be analyzed. This, in turn, means that some a priori knowledge about the domain and content of the text(s) should be available. The more domain specific the texts are, the smaller the ontology needs to be and the more effective and efficient the whole resolution process is expected to be. In fact, it might be that using a larger ontology than necessary could reduce the effectiveness of the resolution process.

To illustrate this point assume that the texts to be analyzed are about American History. This would mean that the locations mentioned within these texts are normally related to events that are part of this history and, consequently, locations that had nothing to do with these events need not be considered. In that way, the range of possible meanings for location terms within the texts as well as the latter's potential scope is considerably reduced.

Thus, a strategy for selecting the minimum required instances that should be included in the location evidence model would be the following:

- First, identify the concepts whose instances may act as location evidence in the given domain and texts.
- Then, identify the subset of these concepts which constitute the central meaning of the texts and thus "determine" mostly their location scope.
- Finally, use these concepts in order to limit the number of possible locations that may appear within the text as well as the number of instances of the other evidential concepts.

For example, when building a location evidence model for texts that describe historical events, some concepts whose instances may act as evidence for locations expected to be found in these texts are related locations, historical events, and historical groups and persons that participated in these events. The most location-determining concept would be the Historical Event, so from all the possible locations, groups and persons we consider only those that are, directly or indirectly, related to some event. Indirectly means, for example, that while "Siege of Tripolitsa" is directly related to "Tripoli", it is indirectly related to Greece as well.

The result of the above process should be a location evidence mapping function  $lem : C \rightarrow R^n$  which given an evidential concept  $c \in C$  returns the relations  $\{r_1, r_2, \dots, r_n\} \in R^n$  whose composition links  $c$ 's instances to locations. Table I shows such a mapping for the history domain and in particular about that of military conflicts.

Using this mapping function, we can then calculate the location meaning evidence function  $lmef$  as follows. Given a location  $l \in L$  and an instance  $i \in I$ , which belongs to some concept  $c \in C$  and is related to  $l$  through the composition of relations  $\{r_1, r_2, \dots, r_n\} \in lem(c)$ , we derive i) the set of instances  $I_{amb} \subseteq I$  which share common names with  $i$  and ii) the set of locations  $L_{amb} \subseteq L$  which share common names with  $l$  and are also related to  $i$  through

Table I  
LOCATION EVIDENCE MAPPING FUNCTION FOR MILITARY CONFLICTS  
DOMAIN

Evidence Concept	Location Linking Relation(s)
Military Conflict	<i>tookPlaceAtLocation</i>
Military Conflict	<i>tookPlaceAtLocation, isPartOfLocation</i>
Military Person	<i>participatedInConflict, tookPlaceAtLocation</i>
Combatant	<i>participatedInConflict, tookPlaceAtLocation</i>
Location	<i>isPartOfLocation</i>

$\{r_1, r_2, \dots, r_n\} \in lem(c)$ . Then the value of the function  $lme f$  for this location and this instance is:

$$lme f(l, i) = \frac{1}{|L_{amb}| \cdot |I_{amb}|} \quad (1)$$

The intuition behind this formula is that the evidential power of a given instance is inversely proportional to its own ambiguity as well as to the number of different target locations it provides evidence for. If, for example, a given military person has fought in many different locations with the same name, then its evidential power for this name is low. Similarly, if a given military person's name is very ambiguous (i.e., there are many persons with the same name) then its evidential power is also low.

Using the same equation we can also calculate the geographical scope evidence function  $gse f$ , the only difference being that we consider the set  $L'_{amb}$  that contains all the locations related to  $i$ , not just the ones with the same name as  $l$ :

$$gse f(l, i) = \frac{1}{|L'_{amb}| \cdot |I_{amb}|} \quad (2)$$

The intuition here is that the geographical scope-related evidential power of a given instance is inversely proportional to the number of different locations it is related to.

### B. Geographical Entity Resolution

The geographical entity resolution process for a given text document and a location meaning evidence function works as follows. First, we extract from the text the set of terms  $T$  that match to some  $i \in I$  along with a term-meaning mapping function  $m : T \rightarrow I$  that returns for a given term  $t \in T$  the instances it may refer to. We also consider  $I_{text}$  to be the superset of these instances.

Then, we consider the set of potential locations found within the text  $L_{text} \subseteq I_{text}$  and for each  $l \in L_{text}$  we derive all the instances from  $I_{text}$  that belong to some concept  $c \in C$  for which  $lem(c) \neq \emptyset$ . Subsequently, by combining the location evidence model function  $lme f$  with the term meaning function  $m$  we are able to derive a location-term meaning support function  $sup_m : L_{text} \times T \rightarrow [0, 1]$  that returns for a location  $l \in L_{text}$  and a term  $t \in T$  the degree to which  $t$  supports  $l$ . If  $l \in L_{text}$ ,  $t \in T$  then

$$sup_m(l, t) = \frac{1}{|m(t)|} \cdot \sum_{i \in m(t)} lme f(l, i) \quad (3)$$

Using this function, we are able to calculate for a given term  $t \in T$  in the text the confidence that it refers to location  $l \in m(t)$ :

$$c_{ref}(l) = \frac{\sum_{t_j \in T} K(l, t_j)}{\sum_{l' \in m(t)} \sum_{t_j \in T} K(l', t_j)} \cdot \sum_{t_j \in T} sup_m(l, t_j) \quad (4)$$

where  $K(l, t) = 1$  if  $sup_m(l, t) > 0$  and 0 otherwise.

In other words, the overall support score for a given candidate location is equal to the sum of the location's partial supports (i.e., function  $sup_m$ ) weighted by the relative number of terms that support it. It should be noted that in the above process, we adopt the one referent per discourse approach which assumes one and only one meaning for a location in a discourse.

### C. Geographical Scope Resolution

The process of geographical scope resolution is similar to the entity resolution one, the difference being that we consider as candidate scope locations not only those found within the text but practically all those that are related to instances of the evidential concepts in the ontology. In that way, even if a location is not explicitly mentioned within the text, it still can be part of the latter's scope. More specifically, given a text document and a geographical scope evidence function  $gse f$  we first consider as candidate locations all those for which there is evidence within the text, that is all those for which  $gse f(l, i) > 0$ ,  $l \in L$ ,  $i \in I_{text}$ . We call this set  $L_{cand}$ . Then, for a given  $l \in L_{cand}$  we compute the scope related support it receives from the terms found within the text as follows:

$$sup_s(l, t) = \frac{1}{|m(t)|} \cdot \sum_{i \in m(t)} gse f(l, i) \quad (5)$$

Finally, we compute the confidence that  $l$  belongs to the geographical scope of the text in the same way as Equation 4 but with  $sup_s$  substituting  $sup_m$ :

$$c_{scope}(l) = \frac{\sum_{t_j \in T} K(l, t_j)}{\sum_{l' \in L_{cand}} \sum_{t_j \in T} K(l', t_j)} \cdot \sum_{t_j \in T} sup_s(l, t_j) \quad (6)$$

where  $K(l, t) = 1$  if  $sup_s(l, t) > 0$  and 0 otherwise.

## IV. EXPERIMENTAL EVALUATION

To illustrate the effectiveness of our proposed framework we performed two experiments on historical texts describing military conflicts. In the first experiment, we focused on correctly resolving ambiguous location references within the texts while in the second, on correctly determining the

texts' geographical scope. In both cases, we built a common location evidence model using an appropriate ontology, derived from DBPedia, comprising about 4120 military conflicts, 1660 military persons, 4270 locations, 890 combatants and, of course, the relations between them (conflicts with locations, conflicts with persons etc.). The model's location evidence mapping function was that of Table I and it was used to calculate the evidential functions  $lmef$  and  $gsef$  for all pairs of locations and evidential entities (other locations, conflicts, persons and combatants).

Table II shows a small sample of these pairs where, for example, James Montgomery acts as evidence for the disambiguation of Beaufort County, South Carolina because he's fought a battle there. Moreover, his evidential power for that location is 0.25, practically because there are 3 other military persons in the ontology also named Montgomery. Similarly, Pancho Villa acts as evidence for the consideration of Columbus, New Mexico as the scope of a text (because he's fought a battle there) and his evidential power for that is 0.2 since, according to the ontology, he's fought battles in 4 other locations as well.

Table II  
EXAMPLES OF LOCATION EVIDENTIAL ENTITIES

Location	Evidential Entity	lmef	gsef
Columbus, Georgia	James H. Wilson	1.0	0.17
Columbus, New Mexico	Pancho Villa	1.0	0.2
Beaufort County, South Carolina	James Montgomery	0.25	0.25

Using this model, we first applied our proposed geographic entity resolution process in a dataset of 50 short texts describing military conflicts. All texts contained ambiguous location entities but little other geographical information and, in average, each ambiguous location reference had 2.5 possible interpretations. For each such reference, we determined its possible interpretations and ranked them using the confidence score derived from Equation 4. We then measured the effectiveness of the process by determining the number of correctly interpreted location references, namely references whose highest ranked interpretation was the correct one.

Table III shows results achieved by our approach compared to those achieved by some well-known publicly available semantic annotation and disambiguation services, namely DBPedia Spotlight [14], Wikimeta [15], Zemanta [16], AlchemyAPI [17] and Yahoo! [18]. As one can see, the consideration of non-geographical semantic information that our approach enables, manages to significantly improve the effectiveness of the geographical entity resolution task.

For the second experiment, we applied our proposed geographic scope resolution process in two different datasets, all comprising 50 short military conflict related texts but

Table III  
GEOGRAPHICAL ENTITY RESOLUTION EVALUATION RESULTS

System/Approach	Effectiveness
Proposed Approach	72%
DBPedia Spotlight	54%
Wikimeta	33%
Zemanta	26%
AlchemyAPI	26%
Yahoo!	24%

with different characteristics. The first dataset comprised texts whose geographical scope was not explicitly mentioned within them and which contained little other geographical information. The second dataset comprised texts whose geographical scope related locations were explicitly and unambiguously mentioned within them but along with other geographical entities that were not part of this scope.

In both cases, we determined for each text the possible locations that comprised its geographical scope and ranked them using the confidence score derived from equation 6. We then measured the effectiveness of the process by determining the number of correctly scope resolved texts, namely texts whose highest ranked scope locations were the correct ones. As a baseline, we compared our results to the ones derived from Yahoo! Placemaker [19] geoparsing web service.

The results of the above process are shown in Table IV. As one can see, the improvement our method achieves in the effectiveness of the scope resolution task is quite significant in both datasets and especially in the first one where the scope-related locations are not explicitly mentioned within the texts. This verifies the central idea of our approach that non-geographical semantic information can significantly improve the geographical scope resolution process and in particular the subtasks of:

- 1) Inferring relevant to the text's geographical scope locations even in the absence of explicit reference of them within the text (first dataset).
- 2) Distinguishing between relevant and non-relevant to the text's geographical scope locations, even in the presence of non-relevant location references within the text (second dataset).

Table IV  
GEOGRAPHICAL SCOPE RESOLUTION EVALUATION RESULTS

System/Approach	Dataset 1	Dataset 2
Proposed Approach	70%	85%
Yahoo! Placemaker	18%	30%

## V. CONCLUSION

In this paper, we proposed a novel framework for optimizing geographical entity and scope resolution in texts by means of domain and application scenario specific non-geographical semantic information. First, we described how,

given a priori knowledge about the domain(s) and expected content of the texts that are to be analyzed, one can define a model that defines which and to what extent semantic entities (especially non-geographical ones) can be used as contextual evidence indicating two things:

- Which is the most probable meaning of an ambiguous location reference within a text (geographical entity resolution task).
- Which locations constitute the geographical scope of the whole text (geographical scope resolution task).

Then, we described how such a model can be used for the two tasks of geographical entity and scope resolution by providing corresponding processes. The effectiveness of these processes was experimentally evaluated in a comprehensive and comparative to other systems way. The evaluation results verified the ability of our framework to significantly improve the effectiveness of the two resolution tasks by exploiting non-geographical semantic information.

It should be noted that our proposed framework is not meant as a substitute or rival of other geographical resolution approaches (that operate in open domains, use geographical information and relevant heuristics and apply machine learning and statistical methods) but rather as a complement of them in application scenarios where text domain and content are a priori known and comprehensive domain ontological knowledge is available (as in the case of historical texts used in our experiments). In fact, given these two requirements for our approach's applicability, future work will focus on investigating how statistical and machine learning approaches may be used, in conjunction with our approach, in order to i) automatically build geographical resolution evidence models based on text corpora and ii) deal with cases where available domain semantic information is incomplete.

#### ACKNOWLEDGMENT

This work was supported by the European Commission under contract FP7- 248984 GLOCAL.

#### REFERENCES

- [1] A. M. Kaplan and M. Haenlein, "Users of the world, unite! the challenges and opportunities of social media," *Business Horizons*, vol. 53, no. 1, pp. 59–68, 2010.
- [2] C. B. Jones and R. S. Purves, "Geographical information retrieval," *Int. J. Geogr. Inf. Sci.*, vol. 22, no. 3, pp. 219–228, Jan. 2008.
- [3] J. Raper, G. Gartner, H. Karimi, and C. Rizos, "Applications of location-based services: a selected review," *J. Locat. Based Serv.*, vol. 1, no. 2, pp. 89–111, Jun. 2007.
- [4] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data - the story so far," *Int. J. Semantic Web Inf. Syst.*, vol. 5, no. 3, p. 122, 2009.
- [5] G. Andogah, G. Bouma, J. Nerbonne, and E. Koster, "Place-name ambiguity resolution," in *Methodologies and Resources for Processing Spatial Language (Workshop at LREC 2008)*, 2008.
- [6] G. Andogah, G. Bouma, and J. Nerbonne, "Every document has a geographical scope," *Data and Knowledge Engineering*, 2012.
- [7] Q. Mei, C. Liu, H. Su, and C. Zhai, "A probabilistic approach to spatiotemporal theme pattern mining on weblogs," in *Proceedings of the 15th international conference on World Wide Web*, ser. WWW '06. New York, NY, USA: ACM, 2006, pp. 533–542.
- [8] C. Wang, J. Wang, X. Xie, and W.-Y. Ma, "Mining geographic knowledge using location aware topic model," in *Proceedings of the 4th ACM workshop on Geographical information retrieval*, ser. GIR '07. New York, NY, USA: ACM, 2007, pp. 65–70.
- [9] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003. [Online]. Available: <http://dl.acm.org/citation.cfm?id=944919.944937>
- [10] J. Xu and W. B. Croft, "Query expansion using local and global document analysis," in *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR '96. New York, NY, USA: ACM, 1996, pp. 4–11.
- [11] J. Bhogal, A. Macfarlane, and P. Smith, "A review of ontology based query expansion," *Inf. Process. Manage.*, vol. 43, no. 4, pp. 866–886, Jul. 2007.
- [12] T. Kauppinen, R. Henriksson, R. Sinkkilä, R. Lindroos, J. Vtinen, and E. Hyvnen, "Ontology-based disambiguation of spatiotemporal locations," in *Proceedings of the 1st international workshop on Identity and Reference on the Semantic Web (IRSW2008), 5th European Semantic Web Conference 2008 (ESWC 2008)*. Tenerife, Spain: CEUR Workshop Proceedings, ISSN 1613-0073, June 1-5 2008.
- [13] Y. Peng, D. He, and M. Mao, "Geographic named entity disambiguation with automatic profile generation," in *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, ser. WI '06. Washington, DC, USA: IEEE Computer Society, 2006, pp. 522–525.
- [14] "Dbpedia spotlight," <http://dbpedia.org/spotlight>, accessed: 05/07/2012.
- [15] "Dbpedia spotlight," <http://www.wikimeta.com>, accessed: 05/07/2012.
- [16] "Zemanta," <http://www.zemanta.com>, accessed: 05/07/2012.
- [17] "Alchemy api," <http://www.alchemyapi.com>, accessed: 05/07/2012.
- [18] "Yahoo!" <http://developer.yahoo.com/search/content/V2/contentAnalysis.html>, accessed: 05/07/2012.
- [19] "Yahoo! placemaker," <http://developer.yahoo.com/geo/placemaker/>, accessed: 05/07/2012.

# Bi-Directional Ontology Updates using Lenses

Andreas Textor

*Distributed Systems Lab*

*RheinMain University of Applied Sciences*

*Unter den Eichen 5, D-65195 Wiesbaden, Germany*

*andreas.textor@hs-rm.de*

**Abstract**—Ontologies can be used to unambiguously describe the semantics of the entities of a domain. Furthermore, ontologies can also contain instances that represent states of real world systems. When an ontology is dynamically updated to reflect changes in the real world, or vice versa (reaction to new information added by a reasoner), data needs to be mapped in both directions. In many systems, this happens through an ad-hoc implementation. Maintaining translations in both directions can be complex and time-consuming. Also, it is often difficult to split a mapping into reusable components. In this paper we examine how *lenses*, an approach to the view update problem originating from database research, can be applied to value updates in ontologies. Lenses provide bi-directional composable translations from one model to another. The application of the approach in the domain of IT management, where an ontology is constantly updated with values from managed systems, is described as part of the ongoing project.

**Keywords**—ontology; ontology update; ontology mapping; lenses; view update

## I. INTRODUCTION

With the ever growing amount of data in all areas of computing and Information Technology, effective means for managing information become more and more important. Especially when data exists in many different heterogeneous sources and formats, integration of information and interoperability of applications that process the data are essential. Furthermore, as syntactic translation of data between different sources is often not sufficient, ontologies are increasingly used to capture semantic information. However, using ontologies comes with its own range of problems that need to be solved, in particular when a single ontology is not sufficient. When multiple (sub-) ontologies, possibly from different sources or authors, are used, they need to be integrated. This leads to research questions such as ontology merging and mapping, matching and ontology alignment, distributed querying and distributed reasoning and others. When information sources and formats external to the ontology need to be dynamically connected to the ontology (i.e., values and/or model structures need to be synchronized), this often results in large amounts of boilerplate code, which is hard to maintain and poorly reusable.

Both cases, regular ontology alignment and the alignment of an ontology with other external models are comparable problems. In ontology alignment, relations between vocab-

ularies of different ontologies are established, while in the alignment with external models relations between concepts in the ontology and concepts in the external model are defined. Depending on the type of model, such relations between concepts can be of the types one-to-one, one-to-many or many-to-one. When the ontology is not only used as a passive information store, but is dynamically updated with information from the external data source, and vice versa (i.e., new facts found by a reasoner are pushed to the external system), information needs to flow in both directions. Translations of data formats and structures need to be performed each time data flows corresponding to the mapping of the external format to the ontology. If we assume the ontology to be a domain model that formally captures the domain and uses this semantic basis to connect other ontologies to it, possibly from different domains, it creates a comprehensive information base. Updating an external system using data from this compound ontology can pose a loss of information, as it only captures a part of the ontology (e.g., an IT management system probably does not include accounting information). On the other hand, importing data from the external system into the ontology may require incomplete data to be complemented to “fit” the data model of the ontology.

This problem is known in database research as the View Update Problem [1]. To approach the problem in the context of ontologies, we examine how *lenses*, a structure for bi-directional composable translations can be adapted to ontologies, with a focus on modularity. Lenses are well examined for the application in database systems, but to be able to be used with ontologies, different requirements must be taken into account. Therefore, we first explain the idea of lenses and then the basic application of lenses to ontologies.

This paper is structured as follows: Section II briefly explains the concept of lenses, as it is defined for the database context. Section III examines existing work in the areas of ontology update, view update and lenses. In Section IV, the approach for the application of lenses in the context of ontology values is described. Section V describes the work in progress, where the approach is applied in the domain of IT management. The paper closes with a summary and future work in Section VI.



## II. EXPLANATION OF LENSES

Lenses were first proposed by Foster et al. in [1] to address the View Update Problem - how can changes made to views be propagated back into the underlying tables. The authors show that the abstract concept of lenses is not only applicable to database schemas, but to other data models as well, and give concrete lenses for the transformation of trees. The concept was further examined in the context of relational databases by Bohannon et al. in [2]. How lenses can be implemented and more use cases are given in [3].

The definition for lenses given in [2] is as follows:

**Definition [Lenses]:** Given schemas  $\Sigma$  and  $\Delta$ , a *lens*  $v$  from  $\Sigma$  to  $\Delta$  (written  $v \in \Sigma \leftrightarrow \Delta$ ) is a pair of total functions  $v \nearrow \in \Sigma \rightarrow \Delta$  (“ $v \nearrow$ ” is pronounced “ $v$  get”) and  $v \searrow \in \Delta \times \Sigma \rightarrow \Sigma$  (pronounced “ $v$  putback”).

Intuitively, a lens combines the pair of functions `get` and `putback`, as shown in the visual explanation in Figure 1, which is derived from [4]. The `get` and `putback` functions define the mapping between the original data source (e.g., the database tables) and the external model (e.g., database views). Together, they provide a different view onto the data, hence the name *lens*.

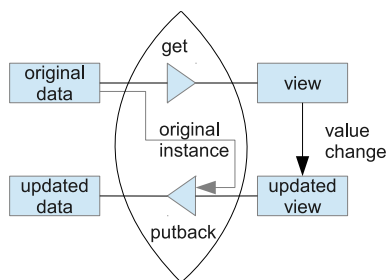


Figure 1. Visual explanation of a lens

The `putback` function is intended to be the inverse of the `get` function in a sense that the resulting lens is *reasonable*, i.e., that the `putback` function only revises the model structures and instances that are necessary for the change. For this reason the function not only depends on the updated structure from the external model, but also on the original structures and instances the change refers to.

To specify this requirement, the authors in [2] define so-called *well-behaved* lenses that must satisfy certain laws:

**Definition [Well-behaved lenses]:** Given schemas  $\Sigma$  and  $\Delta$  along with a lens  $v \in \Sigma \leftrightarrow \Delta$ , we say that  $v$  is a *well-behaved* lens from  $\Sigma$  to  $\Delta$  (written  $v \in \Sigma \leftrightarrow \Delta$ ) if it satisfies the laws GETPUT and PUTGET:

$$\begin{aligned} v \searrow (v \nearrow (I), I) &= I && \text{for all } I \in \Sigma && \text{(GETPUT)} \\ v \nearrow (v \searrow (J, I)) &= J && \text{for all } (J, I) \in \Delta \times \Sigma && \text{(PUTGET)} \end{aligned}$$

The GETPUT law, which is also called *Stability* in [4], states that the original model should not be changed, if

the external model is not changed. This means that the `putback` function should not touch (e.g., set to zero) fields in the original model, if they were not touched in the update operation. The PUTGET law (also called *Acceptability* in [4]) states that updates to the original model should be performed so that the next call of `get` yields exactly the previously put information. This law could be violated, if the `putback` function would write a value other than the one that was updated in the external model (e.g., if `putback` always writes a constant value). The result of a subsequent call of `get` would then be different than the updated value.

The lens laws assure one important property: The composability of lenses, i.e., the creation of new lenses through the composition of existing lenses, similar to function composition. This property allows the creation of separate lenses for each structural or data translation, which are then composed together to form the original specified mapping. When well-behaved lenses are chained together, Foster et al. [1] show that the resulting lens satisfies the lens laws as well.

## III. RELATED WORK

The approach presented here cuts different areas of research: ontology-based information integration, the View Update Problem, ontology updates and ontology mapping. Firstly, publications in which ontologies are employed to achieve information integration range over various domains, and usually describe a mapping of external data formats to the ontology. Representative for the problem at hand is [5], which describes an architecture where an ontology is used for mashups of streaming and stored data. They feature a semantic integration service that allows queries over independent heterogeneous data sources. This is implemented by providing individual mappings for each data source to a central ontology. However, it only works in one direction, as they do not specify how data is propagated back.

Updating ontologies still poses different questions than updating tables in a Relational Database Management System (RDBMS), because updated knowledge may not contradict existing knowledge (which was possibly deduced by a reasoner, rather than added manually), because it would render the ontology inconsistent. Belief update, and more specifically, ontology update, has been examined in several publications. For example, in [6], Löscher et al. propose an ontology update framework where ontology update specifications, which are similar to database triggers, describe certain change patterns that can be performed. Only when an update specification accounts for a change request, the request is accepted, otherwise it is denied. Most of the work on ontology updates is focused on changing the ontology structure, which poses a different problem than updating ontology values and is therefore not directly comparable to our approach. Ontology mapping has been discussed in many publications. Shvaiko and Euzenat [7] give a comprehensive overview of different ontology mapping approaches.

Scharffe and De Bruijn [8] propose requirements for a language to specify ontology mappings (which can also be bi-directional), while in [9], Belhadef gives a method for bi-directional ontology matching that relies on terminological, syntactical and structural comparisons. Again, this focuses on the ontology structure and is not directly comparable to our approach.

#### IV. APPROACH

In this section, we want to examine how the abstract concept of lenses can be applied in scenarios, where external data sources need to be synchronized with an ontology. The approach is orthogonal to existing works, as the goal is not to develop mappings, but an abstraction that allows mappings to be composed out of reusable smaller parts. Regardless of the actual domain, data model or mapping specification, this synchronization is usually implemented in a way that performs structural and value translations, according to the external model. For example, when data from an existing address book should be synchronized with an ontology that also contains other personal data, the ontology might have object properties and data properties that do not directly map to fields in the address book, and values with types such as date time, which might need to be converted from an internal representation to `xsd:dateTime` format, or strings, which might need an encoding conversion. Thus, with each conversion step between the external representation and the ontology, several sub-steps might be necessary. Instead of ad-hoc handling each sub-step in the data conversion implementation, the mapping should be modularized so that each sub-step is a separate entity, and one conversion step is just a composition of the individual sub-steps.

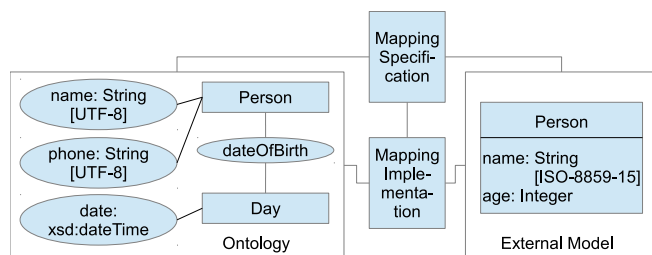


Figure 2. Mapping external data sources to ontologies

If we consider the example given in Figure 2, we can easily see what the mapping specification should look like: The `Person` class can be mapped to an `owl:Class`, string encodings must be translated, and the `age` property of the class should be converted to the right date format using a reference date. However, two problems arise, when the mapping implementation is straightforwardly derived or generated from the specification. First of all, if the mapping is implemented in a monolithic fashion, i.e., without further modularization into the sub-steps, the single conversion sub-steps (i.e., conversion of structure, data types, data values)

are neither reusable nor easily maintainable. Secondly, the specification and the implementation need to take of cases, when data is converted bi-directionally. When a `Person` instance record from the ontology is extracted and converted to an instance of the external `Person` type, the `phone` attribute is simply omitted. When the instance is then updated in the external model (i.e., the name is changed), and the corresponding ontology instance should be updated accordingly, it is desirable that the `phone` attribute from the original ontology `Person` instance remains unchanged. The mapping implementation therefore needs to consider existing `Person` instances in the ontology that represent the same external instance as well as newly inserted instances. Both problems, modularization of bi-directional translations, and the accounting for difference in structure and merging of existing and new fields, can be solved with the application of lenses. In this context, the ontology represents the original source of data, as it is intended to comprehensively aggregate the existing domain knowledge, while the external model can be compared to a database view, as it only covers a subset of the available information (hiding information is often the reason to define a view, while the external model only contains the data structures that are essential for the external system). In the definition of lenses, the `GETPUT` and `PUTGET` laws are necessary for the translation of data in both directions, but as the laws from the original lenses definition originate from database schemas rather than ontologies, the laws are not sufficient to guarantee the consistency of the ontology, because updating certain facts can lead to contradictions with existing facts. The process of changing beliefs to take into account a new piece of information about the world is called *belief change*. This problem has been extensively studied and most formal studies on belief change are based on the work of Alchourrón, Gärdenfors and Makinson (see, e.g., [10]). They specify postulates for contraction (i.e., removal of beliefs from a knowledge base) and revision (changing or updating beliefs in a knowledge base) operators, that must be satisfied by all rational belief change operators. Although belief change theory is not directly applicable to ontologies, Ribeiro and Wassermann [11] have shown that the theory can be applied to ontologies when certain postulates are adapted accordingly. The `PUTGET` law can be related to the *Closure*, *Success* and *Expansion* postulates. The *Closure* postulate ( $K * \alpha = Cn(K * \alpha)$ , where  $K$  is the knowledgebase,  $\alpha$  is the fact to be revised,  $*$  is the belief revision operator and  $Cn$  is the closure function) states that the knowledge base should be logically closed after the new fact is added, the *Success* postulate ( $\alpha \in K * \alpha$ ) states that new information should be successfully accepted, and the *Expansion* postulate ( $K * \alpha \subseteq K + \alpha$ ) states that the revised knowledgebase should not contain more facts than the result of  $K$  expanded by  $\alpha$  (i.e., the fact added without consideration of consistency). The *Consistency* ( $K * \alpha$  is inconsistent, only if  $\vdash \neg \alpha$ ), *Preservation* (If  $\neg \alpha \notin K$  then  $K + \alpha \subseteq K * \alpha$ ) and

*Extensionality* (If  $\alpha \equiv \beta$  then  $K * \alpha = K * \beta$ ) postulates do not apply to relational databases and are for this reason not reflected by the lens laws. In order to maintain consistency in the ontology when applying updates through lenses, the lens, which can be considered a revision operator, must therefore be implemented to satisfy the remaining postulates as well (if no other measures for maintaining consistency are taken).

As the composability of lenses makes it possible to create a library of lenses for common or very specific updates to ontologies, the revision postulates should be considered when lenses for ontology updates are created.

## V. APPLICATION

The concept of lenses is currently being used in the implementation of an ontology-based automated IT management system. We are working on the implementation of a concrete set of lenses for the translation between an OWL (Web Ontology Language) ontology representation and the external data source of a CIM environment. The Common Information Model (CIM, [12]) is an object-oriented model to represent entities and relationships of IT systems, and is used in IT management and storage management tools. A translation of CIM to OWL was previously examined in [13], and used in an architecture for automated IT management [14]. Preliminary results show that the application of lenses to implement the mapping between the ontology and the CIM environment, rather than the previously used prototypical monolithic implementation, can greatly contribute to the modularity and extensibility of the architecture.

## VI. CONCLUSION AND FUTURE WORK

In this paper, we have examined the abstract concept of lenses, an approach to the view update problem in databases, and its applicability to the context of ontologies. We have shown that for scenarios where an ontology serves as aggregation of domain knowledge that is dynamically updated with an external model, the ontology can be thought of as the original data source, while the external model can be thought of as a view. This allows the use of lenses for synchronization between the models. As ontology updates differ from updates of relational databases, we have examined how the lens laws relate to the postulates that belief revision operators must satisfy, and found that a lens that performs ontology updates can not solely rely on the lens laws, but must still follow the postulates. Future work therefore includes the completion of the formalisation of belief revision for ontology update lenses and the further evaluation of the approach in the domain of IT management.

## REFERENCES

[1] J. N. Foster, M. B. Greenwald, J. T. Moore, B. C. Pierce, and A. Schmitt, "Combinators for bi-directional tree transformations: A linguistic approach to the view update problem," in *In ACM SIGPLANSIGACT Symposium on Principles of*

*Programming Languages (POPL)*. ACM Press, 2005, pp. 233–246.

[2] A. Bohannon, B. C. Pierce, and J. A. Vaughan, "Relational Lenses : A Language for Updatable Views," in *Proceedings of the 25th ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 2006.

[3] A. Wider, "Towards Combinators for Bidirectional Model Transformations in Scala," in *Post-Proceedings of the 4th International Conference on Software Language Engineering (SLE'11)*. Braga, Portugal: Springer, 2011, pp. 1–10.

[4] B. C. Pierce, "The Weird World of Bi-Directional Programming," March 2006, ETAPS invited talk.

[5] A. J. G. Gray, R. García-Castro, K. Kyzirakos, M. Karpathiotakis, J.-p. Calbimonte, K. Page *et al.*, "A Semantically Enabled Service Architecture for Mashups over Streaming and Stored Data," in *Proceedings of the 8th Extended Semantic Web Conference*, 2011.

[6] U. Lösch, S. Rudolph, D. Vrandečić, and R. Studer, "Tempus fugit - Towards an Ontology Update Language," in *6th European Semantic Web Conference (ESWC 09)*, vol. 1. Springer, January 2009, pp. 278–292.

[7] P. Shvaiko and J. Euzenat, "Ontology matching: state of the art and future challenges," *IEEE Transactions on Knowledge and Data Engineering*, 2012.

[8] F. Scharffe and J. de Bruijn, "A Language to Specify Mappings Between Ontologies," in *Proceedings of the 1st International Conference on Signal-Image Technology and Internet-Based Systems (SITIS 2005)*, R. Chbeir, A. Dipanda, and K. Yétongnon, Eds., Yaounde, Cameroon, 2005, pp. 267–271.

[9] H. Belhadef, "A New Bidirectional Method for Ontologies Matching," *Procedia Engineering*, vol. 23, pp. 558–564, Jan. 2011.

[10] C. E. Alchourrón, P. Gärdenfors, and D. Makinson, "On the logic of theory change: Partial meet contraction and revision functions," *The Journal of Symbolic Logic*, vol. 50, pp. 510 – 530, 1985.

[11] M. M. Ribeiro and R. Wassermann, "First Steps Towards Revising Ontologies," in *Proc. of WONRO'2006*, 2006.

[12] Distributed Management Task Force, "Common Information Model (CIM)," <http://www.dmtf.org/standards/cim/>. Last access 2012-07-09.

[13] A. Textor, J. Stynes, and R. Kroegeer, "Transformation of the Common Information Model to OWL," in *10th International Conference on Web Engineering - ICWE 2010 Workshops*, ser. LNCS, vol. 6385. Springer Verlag, July 2010, pp. 163–174.

[14] A. Textor, F. Meyer, and R. Kroegeer, "Semantic Processing in IT Management," in *Proceedings of the Fifth International Conference on Advances in Semantic Processing (SEMAMPRO)*, 2011, pp. 87–90.

# Capturing Knowledge Representations Using Semantic Relationships

## An Ontology-based approach

Ruben Costa, Paulo Figueiras, Luis Paiva, Ricardo Jardim-Gonçalves  
 Centre of Technology and Systems  
 UNINOVA  
 Quinta da Torre, Portugal  
 rddc@uninova.pt, paf@uninova.pt,  
 luismpaiva@mail.telepac.pt, rg@uninova.pt

Celson Lima  
 Federal University of Western Pará  
 PC / IEG / UFOPA  
 Santarém, Brasil  
 celsonlima@ufpa.br

**Abstract**— Knowledge representations in the scope of this work are a way to formalize the content of documents using dependent metadata i.e. words in document. One of the challenges relates to limited information that is presented in the document. While past research has made use of external dictionaries and topic hierarchies to augment the information, there is still considerable room for improvement. This work explores the use of complex relationships (otherwise known as Semantic Associations) available in ontologies with the addition of information presented in documents. In this paper we introduce a conceptual framework and its current implementation to support the representation of knowledge sources, where every knowledge source is represented through a vector (named Semantic Vector - SV). The novelty of this work addresses the enrichment of such knowledge representations, using the classical vector space model concept extended with ontological support, which means to use ontological concepts and their relations to enrich each SV. Our approach takes into account three different but complementary processes using the following inputs: (1) the statistical relevance of keywords, (2) the ontological concepts, and (3) the ontological relations.

*Keywords*-Information Retrieval; Ontology Engineering; Knowledge Representation

### I. INTRODUCTION

Knowledge and its respective representation has been part of human activity since immemorial times. Mankind created ways to tangibly represent sources of knowledge in order to preserve such knowledge and to guarantee that it would be transmitted to and reused by future generations. Classical examples are Egyptian papyrus and Sumerians clay tablets.

With the evolution of the World Wide Web towards the semantic web, knowledge sources (KS) and their representations have jumped on the main stage since they play a key role in this arena. Meaning of things and the ability to precisely understand them has been the holy grail of major efforts targeting the settlement (at least partial) of the tangible semantic web. Various sorts of concepts and tools have been developed and tested, the journey is very promising but there is a long way forward.

Controlled Vocabularies (CV) [1] have been considered good means to achieve this goal and, as such, a myriad of

results & tools have been produced by researches around the world, based on the use of CVs. Among them, we are particularly interested in the use of ontological support to investigate the enrichment of knowledge representation of KS.

In this work, knowledge representation is expressed through the use of Semantic Vectors (SVs) based on the combination of the Vector Space Model (VSM) approach [2] and ontology-related features, namely ontological concepts and their semantic relations. Therefore, KS, in this work, are represented by SVs which contain concepts and their equivalent terms, weights (statistical, taxonomical, and ontological ones), relations and other elements used to semantically enrich each SV.

This paper is structured as follows. Section 2 defines the objectives and addresses the problem to be tackled. Section 3 presents the related work. Section 4 defines the process addressed by this work for knowledge representation. Section 5 illustrates the empirical evidences of the work addressed so far. Finally, section 6 concludes the paper and points out the future work to be carried out.

### II. RELEVANCE OF THE PRESENTED WORK

This paper proposes the development of a framework to support the semantic representation of KS, which will be assessed in building and construction sector. Main features of this work include the analysis of the links among concepts, and the KS they are representing as well as the enhancement of such links with semantic relations among concepts.

In order to understand the importance of semantic relations within KS from the building and construction, one can think, for instance, on two expressions/terms (considered as ontological concepts, for the sake of clarity): “Design Phase” and “Architect”. These concepts are not father and son (hierarchically related), but they are inherently connected through a semantic relation described as “has Design Actor”, i.e., a project’s design phase may have many actors associated with it; one of them is the “Architect”. Such relation may also be associated to a given weight, i.e., how strong is the influence of the actor “Architect” within a project “Design Phase”.

Considering the example explained above, when a user is searching for information regarding a project design phase,

two different types of results may be expected by the end user, since “Design Phase” concept could be strongly related with the “Architect” concept.

The idea presented here is to enrich the representation of KS used/created within project teams on a collaborative engineering environment with information extracted from a domain ontology. A variety of semantic resources ranging from domain dictionaries to specialized taxonomies have been developed in the building and construction industry. Among them are BS6100 (Glossary of Building and Civil Engineering terms produced by the British Standards Institution); bcXML (an XML vocabulary developed by the eConstruct IST project for the construction industry); IFD (International Framework for Dictionaries developed by the International Alliance for Interoperability); OCCS (OmniClass Classification System for Construction Information) , BARBi (Norwegian Building and Construction Reference Data Library); and e-COGNOS (CONSistent knowledge management across projects and between enterprises in the construction domain). For the purpose of this work, a domain ontology was developed and validated in conjunction with the support of domain knowledge experts, and also adopting several concepts from the initiatives presented above. One of the reasons that lead a development of a new ontology, was due to the fact that at the time there was no support for OWL regarding such initiatives.

One of the novelties addressed by this work is the adoption of the Vector Space Model (VSM) approach combined with the ontological concepts and their semantic relations. The idea behind the VSM is to represent each document in a collection as a point in a space (a vector in a vector space). Points that are close together in this space are semantically similar and points that are far apart are semantically distant. The user's query is represented as a point in the same space as the documents (the query is a pseudo-document).

This approach uses an approximation to the VSM to achieve knowledge representations of documents and queries, and to define a relationship between these representations, allowing comparisons among them. The documents are sorted in order of increasing distance (decreasing semantic similarity) from the query and then presented to the user [3].

Knowledge representation of documents, using the VSM, often comes in the form of semantic vectors. Semantic vectors are usually called matrixes of frequencies, as they define the probabilistic frequency of the existence of a concept on a document and, hence, the relevance of that concept on the representation of the document.

### III. RELATED WORK

In relation with the problem to be addressed by this work, Castells *et al.* [4] proposes an approach based on an ontology and supported by an adaptation of the Vector Space Model, just as in the presented work's case. It uses the TF-IDF (term frequency-inverse document frequency) algorithm [5], matches documents' keywords with ontology concepts, creates semantic vectors and uses the cosine similarity to

compare created vectors. A key difference between this approach and the presented work is that Castells' work does not consider semantic relations or the hierarchical relations between concepts (taxonomic relations).

On the other hand, Nagarajan *et al.* [6] proposes a document indexation system based on the VSM and supported by Semantic Web technologies, just as in the presented work. They also propose a way of quantifying ontological relations between concepts, and represent that quantification in documents' semantic vectors. There are some differences between this work and the presented approach, which does not distinguish between taxonomic and ontological relations, as our approach does.

### IV. THE PROCESS

The process being proposed by this work, is composed by several stages: the first stage (knowledge extraction) deals with the extraction of relevant words from KS, with the support of a text mining tool RapidMiner [7], and preforms a TF-IDF score for each relevant keyword within the corpus of KS that constitutes our knowledge base (knowledge sources repository); the second stage is the semantic vector creation, referred as Knowledge Source Indexation; and the third stage is document comparison and ranking processes, denominated Knowledge Source Comparison [8], as depicted in Figure 1.

The several stages that compose the process are illustrated with examples from a corpus with 70 KS related with the building and construction domain, where the creation of the semantic vectors example is described using an individual KS from the corpus.

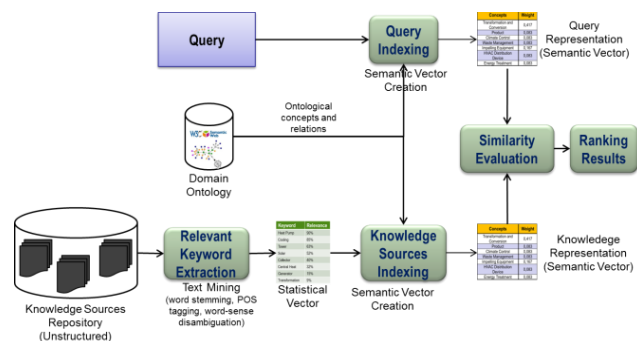


Figure 1. Document indexation and comparison

#### A. Knowledge Extraction

Although the use of text mining techniques is not the objective of this paper, it is worth to introduce some of text mining concepts, because the overall approach adopted here uses some of these concepts as an input to the knowledge representation mechanism.

Knowledge extraction is usually a process comprising three stages: word extraction, regular expressions filtering, and static vector creation.

Word extraction is the process in which words and expressions are extracted and divided through text-mining techniques. Regular expression filtering defines the process of removing expressions which have a great number of occurrences, but do not represent the knowledge within the

document (e.g. “and”, “the”, “when”). The last stage, statistic vector creation, is the process that builds the statistical representation of the documents in the form of a matrix composed by expressions, or keywords, and by the statistical weight of each keyword within the document, based on the TF-IDF score for each keyword within each KS.

Such structure is called statistical vector, and it is the main input for the presented work. Some frameworks and applications already treat knowledge extraction issues to the extent which our approach needs. Our approach uses RapidMiner to fulfil the needed knowledge extraction tasks and to create KS statistical vectors, which are then stored in a database.

It is important to mention that keywords presented in the statistical vector are composed by stemmed words (words that are considered a primitive form for a family of words, e.g. design: design, designer, designing, etc.). An example of such statistic vector for illustrative purposes is given in Table 1.

TABLE I. CONCEPTS AND WEIGHTS OF A DOCUMENT'S STATISTIC VECTOR

Keyword	Statistic weight (rounded values)
Agreement	0.550
Fund	0.376
Provis	0.317
Advanc	0.311
Record	0.250
Found	0.212
Feder	0.196
Local	0.166
Govern	0.153
Inspect	0.150
State	0.150
Ensur	0.144
Singl	0.116
modul	0.114
parti	0.114

### B. Semantic Vector Creation

Semantic vector creation is the basis for the presented approach, it represents the extraction of knowledge and meaning from KS's and the agglomeration of this information in a matrix form, better suited for mathematical applications than the raw text form of documents.

A semantic vector is represented by two columns: the first column contains the concepts that build up the knowledge representation of the KS, i.e. the most relevant concepts for contextualizing the information within the KS; the second column keeps the degree of relevance, or weight, that each term has on the knowledge description of the KS.

Our approach takes into account three different, but complementary procedures for building up the semantic vector, each of which is considered a more realistic iteration of the knowledge representation of a KS: Keyword-based, taxonomy-based and ontology-based semantic vectors.

Keyword-based semantic vectors are built upon the statistic representation of KSs in the form of expressions that

occur in the document, according to their emphasis and frequency of occurrence both locally (in the KS itself) and globally (in the document corpus' universe).

Table 2 depicts the weight of each ontology concept associated to each keyword within the statistic vector, where the first column corresponds to the ontology concepts that were matched to describe most relevant keywords extracted from the statistical vector, the second column indicates the most relevant keywords that were match to ontology equivalent terms, the third column corresponds the total ontology equivalent terms for each concept that was matched, and the fourth and last column, indicates the semantic weight for each ontology concept matched.

Taxonomy-based vectors push one notch further in the representation of KSs by adjusting the weights between expressions according to their taxonomic kin with each other, i.e., expressions that are related with each other with the “is a” type relation. If two or more concepts that are taxonomically related appear in a keyword-based vector, the existing relation can boost the relevance of the expressions within the KS representation.

Ontology-based vectors are the last iteration of the semantic vector creation process. The creation process for this type of vector uses the taxonomy-based vector as input to analyse the inherent ontological relation patterns between the input vector's expressions. These ontological relations define semantic patterns between concepts which can be used to enhance the representation of the document. For instance, if a vector has two concepts that are related to each other by an ontological relation, and if this ontological relation occurs frequently across the document corpus' universe, then the relevance of both concepts being together within the KS increases the weight of these concepts in the vector.

The major difference between taxonomy-based vectors construction and ontology-based vectors is that, taxonomy-based vectors take into account relations between concepts that are hierarchically related within the ontology tree (ex: father and son concepts). On the other hand, ontology-based vectors take into account relations between concepts that don't need to be hierarchically related but are semantically connected. Examples of the two types of vectors are described in the following sub-sections.

TABLE II. CONCEPTS AND WEIGHTS OF A DOCUMENT'S STATISTIC VECTOR

Concept	Keyword	Ontology keywords	Sem. weight
Presence_Detection _And_Registration	record	recording	0.189
Foundation	found	foundation	0.134
Association	feder	federation	0.124
Inspector	inspect	inspector, inspection	0.114
Territory	state	state	0.095
Issue	compli	complicatio n	0.087
Trainer	manag	manager	0.028

<b>Request</b>	request	request	0.063
<b>Consultant</b>	author	authority	0.057
<b>Management_Actor</b>	manag	manager, manageme nt actor	0.028
<b>Report</b>	report	report	0.025

1) *Keyword-based Semantic Vectors*

The next step deals with matching the statistical vector’s keywords with equivalent terms which are linked with the ontological concepts from the domain ontology. Equivalent terms for concept “Engineer” are shown in Figure 2. The matching process between equivalent terms presented on the domain ontology and the keywords within the statistical vector, is done by string matching. This approach may lead into some inconsistencies, since a keyword presented in the statistical vector may match two or more equivalent terms. This issue is being analysed and is considered to be part of future work.

It is worth also to mention that, the current process also addresses the introduction of new concepts and new semantic relations which are used to update the domain ontology. The process of updating the domain ontology is triggered every time new KS are introduced into the knowledge based. Algorithms for text processing (ex: association rules), are used to exploit new semantic relations between concepts or to update existing ones. This part of the process was intentionally not described here and is part of an on-going work.

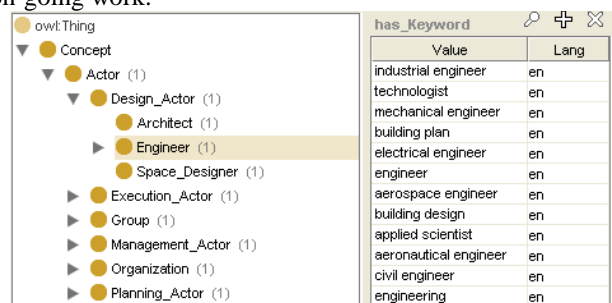


Figure 2. Ontological keywords and equivalent terms for concept "Engineer".

Each concept in the domain ontology has several keywords associated to it that present some semantic similarity or some meaning regarding that specific concept. Since keywords in the statistical vector comprise only stemmed words, several ontology-related keywords can be matched to one statistical vector’s keyword. Although this fact may lead to some inconsistencies in terms of knowledge reliability, in this case, and because the presented work uses a very specific domain, these issues are decreased and are to be tackled in the future work section.

For each ontological concept that was extracted, the weights of all keywords matched with that concept are summed in order to get the total statistical weight for that ontological concept.

The next step to be performed, deals with the attribution of semantic weights to each of the concepts. The presented

approach uses an approximation to the TF-IDF family of weighting functions [9], already used on other research works [4], to calculate the semantic weight for each concept resultant from the concept extraction process. The TF-IDF algorithm used is given by the expression:

$$w_x = \frac{w_{x,d}}{\max_y w_{y,d}} \cdot \log \frac{D}{n_x} \tag{1}$$

In Equation 1,  $w_{x,d}$  is the statistical weight for concept  $x$  in KS  $d$ ’ s statistical vector,  $\max_y w_{y,d}$  is the statistical weight of the most relevant concept,  $y$ , within the statistical vector of KS  $d$ ,  $D$  is the total number of KSs present in the KSs search space,  $n_x$  is the number of KSs available in such space which have concept  $x$  in their semantic vectors, and  $w_x$  is the resultant semantic weight of concept  $x$  for document  $d$ .

Statistical normalisation is performed over the keyword-based semantic vector’s weights, in order to obtain values between zero (0) and one (1).

This will be crucial for the upcoming vector comparison result ranking processes, because it will ease the computation processes needed and the attribution of relevance percentage to the results.

The keyword-based semantic vector is then stored in the database in the form  $[\sum_{i=1}^n x_i ; \sum_{i=1}^n w_{x_i}]$ , where  $n$  is the number of concepts in the vector,  $x_i$  is the syntactical representation of the concept and  $w_{x_i}$  is the semantic weight corresponding to concept.

2) *Taxonomy-based Semantic Vectors*

The taxonomy-based semantic vector creation process defines a semantic vector based on the relations of kin between concepts within the ontological tree. Specifically, the kin relations can be expressed through the following definitions [10]:

Definition 1: In the hierarchical tree structure of the ontology, concept A and concept B are homologous concepts if the node of concept A is an ancestor node of concept B. Hence, A is considered the nearest root concept of B,  $R(A,B)$ . The taxonomical distance between A and B is given by:

$$d(A,B) = |\text{depth}(B) - \text{depth}(A)| = |\text{depth}(A) - \text{depth}(B)| \tag{2}$$

In Equation 2,  $\text{depth}(X)$  is the depth of node  $X$  in the hierarchical tree structure, with the ontological root concept’s depth being zero (0).

Definition 2: In the hierarchical tree structure of the ontology, concept A and concept B are non-homologous concepts if concept A is neither the ancestor node nor the descendant node of concept B, even though both concepts are related by kin; If  $R$  is the nearest ancestor of both A and B, then  $R$  is considered the nearest ancestor concept for both A and B concepts,  $R(A,B)$ ; The taxonomical distance between A and B is expressed as:

$$d(A,B) = d(R,A) + d(R,B) \tag{3}$$

Figure 3 depicts the difference between homologous and non-homologous concepts.

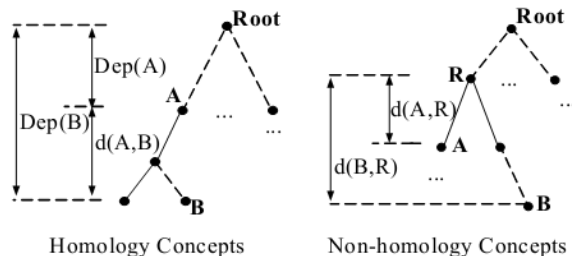


Figure 3. Homologous and non-homologous concepts (Li, 2009).

One of the major differences between our work and the work presented by Li [10], is that in our approach, the taxonomical weights between two concepts are not only related by their distance on the domain ontology, but also considering the relevance of the pair concepts A and B to each particular KS, i.e., if concepts A and B which are taxonomical related co-occur frequently, the taxonomical weight of such relation will be assigned a higher score.

### 3) Ontology-based Semantic Vectors

Other iteration of the semantic vector creation process is the definition of the semantic vector based on the ontological relations' which are defined in the domain ontology. Our system uses human input (knowledge experts in the building and construction domain) to establish final numerical scores on semantic relationships. The idea behind having a human intervene here is to let the importance of relationships reflect a proper knowledge representation requirement at hand. If the end-user is not interested in relationships between a project design phase and an architect actor, he should be able to rank those lower compared to other relationships. As an example, five ontological relations are shown in Table 3.

The first step is to analyse each ontological relation between concepts present on the input semantic vector. In this case, both keyword and taxonomy-based semantic vectors are used as inputs for this analysis. As in taxonomy-based semantic vector creation, there are two processes involved on the ontological relationship analysis: the first boosts weights belonging to concepts within the input semantic vector, depending on the ontology relations between them; the second adds concepts that are not present in the input vector, according to ontological relations they might have with concepts belonging to the vector [6].

In the first process (ontological relation between two concepts present in the input semantic vector),  $Co-Occurrence_{C_x C_y}$  is computed with Equation 4, but this time it will be taken into account the frequency of occurrence of the ontologically related concepts throughout the document corpus.

$$Co - Occurrence_{C_x C_y} = idf(C_x C_y) = \log \frac{D}{n_{C_x C_y}} \quad (4)$$

It is worth to notice, that an IDF calculus is performed but taking into account the ontological relation, i.e, the

frequency of such relation is calculated within the all document corpus.

As in taxonomy-based semantic vector creation, the new concept is added to the semantic vector only if the ontological relation importance is greater than or equal to a pre-defined threshold, for the same constraint purposes. The ontological relation's importance, or relevance, is not automatically computed; rather, it is retrieved from an ontological relation vector which is composed by a pair of concepts and the weight associated to the pair relation.

In the case of the second process (ontological relation between one concept within the input semantic vector and another concept not comprised in that vector), and again as in the taxonomy-based semantic vector creation process,  $C_x$  is not modified and  $C_y$  is added to the semantic vector, and its weight is computed as in Equation 5.

$$tw_{C_y} = w_{C_y} + \sum(all\ related\ C_xs) \left[ w_{C_x} * (TI_{C_x C_y}) \right] \quad (5)$$

TABLE III. EXAMPLES OF ONTOLOGICAL RELATIONS WITHIN ONTOLOGY.

Property	Subject	Object	Description
operates in	Actor	Project Phase	Actors operate in one or several particular project phases
is involved in	Actor	Project	Actors are involved in projects
has skills	Actor	Skill	Actors have some skills and expertise
has skill needs	Project	Skill	Projects need actors' skills and expertise
is decomposed in	Project	Task	Projects may be considered sets of tasks

## V. ASSESSMENT

This section illustrates the assessment process of our approach. Firstly, the knowledge source indexation process will be assessed. Secondly, an example of a query and its results is exemplified.

### A. Treating Queries

As mentioned earlier, queries are treated like pseudo-KSs, which means that all queries suffer an indexation process similar to the one applied to KSs. Initially, the query is divided into keywords and those keywords are then used to create a statistic vector for the query, equal to the statistic



term-frequency vector used for KS indexation. But, instead of passing the query through the knowledge extraction process the statistic vector is created by giving the same statistic weight to all keywords contained in the query. Such rule implies that the system assumes the same importance to all of the query's keywords.

For the purpose of this assessment, it was used a corpus of sixty five KS randomly selected, but all having a strong focus on the building and construction domain. Just as an example, a test query search for “door”, “door frame”, “fire surround”, “fireproofing” and “heating”, meaning that the user is looking for doors and respective components that are fireproof or that provide fire protection. In this case, keyword “door” is matched with concept “Door”, “door frame” is matched with “Door Component”, and so on, as shown in Table 5. Weights for matched ontological concepts are all equal to 0.2, because each concept only matches with one keyword. Hence, the semantic vector for this query will be the one of Table 4.

TABLE IV. EXAMPLE OF A QUERY'S SEMANTIC VECTOR.

#	Keyword	Ontology concept	Weight
1	Door	Door	0.2
2	door frame	Door Component	0.2
3	fire surround	Fireplace And Stove	0.2
4	Fireproofing	Fireproofing	0.2
5	Heating	Complete Heating System	0.2

### B. Comparing and Ranking Documents

Our approach for vector similarity takes into account the cosine similarity [11] between two vectors, i.e. the cosine of two vectors is defined as the inner product of those vectors, after they have been normalized to unit length. Let  $d$  be the semantic vector representing a document and  $q$  the semantic vector representing a query. The cosine of the angle  $\theta$  between  $d$  and  $q$  is given by:

$$\cos \theta = \frac{d \cdot q}{\|d\| \|q\|} = \frac{\sum_{k=1}^m w_{dk} w_{qk}}{\sqrt{(\sum_{k=1}^m w_{dk}^2)(\sum_{k=1}^m w_{qk}^2)}} \quad (6)$$

where  $m$  is the size of the vectors,  $w_{dk}$  is the weight for each concept that represents  $d$  and  $w_{qk}$  is the weight for each concept present on the query vector  $q$  [4], [10].

A sparse-matrix multiplication approach is used because the most commonly used similarity measures for vectors  $d$  and  $q$ , such as cosine, can be decomposed into three values: one depending on the nonzero values of  $d$ , another depending on the nonzero values of  $q$ , and the third depending on the nonzero coordinates shared both by  $d$  and  $q$ .

In this case, calculating  $f_1(d_k, q_k)$  is only required when both vectors have at least one shared nonzero coordinate. If the vectors do not possess any shared concept, i.e. a nonzero coordinate, the value for the function above is zero, and the vectors do not present any similarity. This also means that  $f_2$

and  $f_3$  do not need to be calculated, significantly reducing the computation needed [3].

On the other hand, even though the cosine method requires that both vectors have the same size, when using sparse-matrix multiplication the vectors' sizes do not necessarily have to coincide. If one vector is smaller than the other, then it means, in practice, that the smaller vector has zero values for all the concepts that are missing to reach the size of the bigger vector.

KS ranking is based on the similarity between KSs and the query. More specifically, and because the result of the cosine function is always 0 and 1, the system extrapolates the cosine function result as a percentage value.

The first result for the KSs tested is very satisfactory: the first search-resultant KS gives a relevance of 84% to the query, out of a total of sixty five KSs. The relevance of the KS corpus representation against the user query is presented in Table 5.

TABLE V. FIVE MOST RELEVANT RESULTS FOR THE USER QUERY.

Doc. Id	1	2	3	4	5	Query relevance%
190	0.093	0.093	0.077	0.077	0.0803	84
179	0.181	0.182	n.a.	n.a.	n.a.	57
201	0.121	0.122	0.013	0.013	n.a.	55
197	0.017	0.017	0.109	0.110	n.a.	52
172	0.045	0.045	0.035	0.037	0.012	48

It is easily comprehensible that, for the first result (doc. id 190), all concepts have higher semantic weight, with values near to 0.10 (or 10%). The second result presents high weights for the first two concepts, which means that it can have some relevance to the query, but its semantic vector does not contain the other three concepts of the query. This means that, although this KS has a good semantic reference to “Door” and “Door Component”, it does not have knowledge about the other three concepts. The last result, with 48%, has weights for all concepts of the query but they are very low (4% maximum). This means that although the KS might have some relevance to the query, after a manual inspection over KSs tested, the results reflect knowledge contained within such documents.

The results are presented by showing only the relevance percentage for each KS the database identifier of the KS and the name and type of the KS file.

## VI. CONCLUSIONS AND FUTURE WORK

Our contribution targets essentially the representation of KS which can be applied in various areas, such as semantic web, and information retrieval. Moreover, it can also support project teams working in collaborative environments, by helping them to choose relevant knowledge from a panoply of KS and, ultimately, ensuring that knowledge is properly used and created within organizations.

The results achieved so far and presented here do not reflect the final conclusion of the proposed approach and are part of an on-going work that will evolve and mature over

time, nevertheless preliminary results lead us to conclude that the inclusion of additional information available in domain ontologies in the process of representing knowledge sources, can augment such knowledge representations. Additional testing needed to be addressed, and other metrics for evaluating the performance of the proposed method (ex: precision and recall) needed to be implemented, in order to provide more concrete conclusions.

As future work, some improvements to the proposed approach within this work still needed to be carried out. As explained earlier, the corpus of KSs chosen to perform the assessment was adopting a randomly criteria. The fact that all documents are dealing with building and construction projects, make the scope very wide, which lead to a high level of noise introduced when creating statistical vectors adopting the TF-IDF approach. It is proposed as future work, to perform the creation of statistical vectors using a batch mode, where all KSs are previously grouped in clusters of domain area using clustering algorithms as the k-means algorithm. We believe that having documents previously grouped within clusters will reduce the level of noise introduced within the creation of statistical vectors.

Other operations for better enhance the semantic vectors can also be taken into account, for instance, union operations between taxonomical and semantic based vectors can also be seen as an approach for better represent KSs.

Additional work can also be driven on the building and construction domain ontology itself, which deals with the semantic features on knowledge representations. The domain ontology is seen as something that is static and doesn't evolve over time as organizational knowledge does. One possible approach to be adopted is to extract new knowledge coming from KSs (new concepts and new semantic relations) and reflect such new knowledge on domain ontology. The weights of such semantic relations should also be updated every time new KSs are introduced into the knowledge base. The idea is that, ontological concepts and relations should be inserted and managed dynamically, through a learning

process, in order to make possible for the ontology to learn, capture new concepts and relations from the KS corpus' universe and update relation importance between concepts, while new sources become available.

#### REFERENCES

- [1] C. Lima, A. Zarli, and G. Storer, "Controlled Vocabularies in the European Construction Sector: Evolution, Current Developments, and Future Trends," *Complex Systems Concurrent Engineering* ed. London : Springer, pp. 565-574, 2007.
- [2] G. Salton, A. Wong, and S. Yang, "A Vector Space Model for Automatic Indexing," *Communications of the ACM*, 18(11), pp. 613-620, 1975.
- [3] P. Turney, and P. Pantel, "From Frequency to Meaning: Vector Space Models of Semantics," *Journal of Artificial Intelligence*, pp. 141-188, 2010.
- [4] P. Castells, M. Fernández, and D. Vallet, "An Adaptation of the Vector-Space Model for Ontology-Based Information Retrieval" *IEEE Transactions on Knowledge and Data Engineering*, February, 19(2), pp. 261-272, 2007.
- [5] T. Yu and G. Salton, "Precision weighting—An effective automatic indexing method," *J. ACM* 23, 1, 76–88, 1976.
- [6] M. Nagarajan, A. Sheth, M. Aguilera, K. Keeton, A. Merchant and M. Uysal, "Altering Document Term Vectors for Classification - Ontologies as Expectations of Co-occurrence," *ReCALL*, p. 1225, 2007.
- [7] Rapid-I GmbH, (2012, April). Retrieved from <http://rapid-i.com/content/view/181/190/>
- [8] R. Costa, C. Lima, "An Approach for Indexation, Classification and Retrieval of Knowledge Sources in Collaborative Environments", Lisbon, 2011.
- [9] S. Jones, "A Statistical Interpretation of Term Specificity and its Application in Retrieval," *Journal of Documentation*, 60(5), pp. 11-21, 2004
- [10] S. Li, "A Semantic Vector Retrieval Model for Desktop Documents. *Journal of Software Engineering & Applications*," Issue 2, pp. 55-59, 2009.
- [11] M. Deza and E. Deza, "Encyclopedia of Distances," Heidelberg: Springer-Verlag Berlin Heidelberg, 2009.

# Higher Education Qualification Evaluation

Ildikó Szabó

Department of Information Systems  
Corvinus University of Budapest  
1093 Budapest, Fővám tér 13-15., HUNGARY  
iszabo@informatika.uni-corvinus.hu

**Abstract—** Qualification developed by the requirements of labor market is more competitive than the other ones. In Hungary, this issue constitutes one of the central elements of the higher education reform taking place nowadays. In this paper, a system in progress is presented, which aims at evaluating the learning outcomes of Business Informatics Bachelor's degree program at Corvinus University of Budapest versus the competences needed by the labor market, as appeared on a job recruitment portal. Ontology-based learning and matching domains are touched in the course of the development, so it is necessary to choose their appropriate tools to integrate them into a system. The tools written in Java constitute the base of the development.

*Keywords-competence; ontology learning; ontology matching*

## I. INTRODUCTION

The higher education reform is a long-term process in Hungary. One of the objectives of the government is to rationalize qualification obtained in the higher education in the light of requirements of the world of labor [10]. The research focus of the Ph.D. thesis is to examine in what measure the learning outcomes of *Business Informatics* Bachelor's degree program at Corvinus University of Budapest are matched to the requirements given by ICT job roles (job requirements).

The competences are the descriptors of learning outcomes and they serve as an appropriate tool to describe a job role, so the question is what the missing and surplus competences of this training program are. But, this concept has no universal definition, so the ontology approach serves as an appropriate method by providing an explicit, formal specification about this domain and it is capable of comparing the competences semantically and by considering the structure of related concepts too. The ontology learning approach is an appropriate method to build ontology dynamically and the ontology matching approach provides this semantic and structural comparison.

Two projects – OntoHR [12] and SAKE [13] – have already dealt with this problem. They aimed to identify the shortcomings of higher and vocational education learning outcome through matching a job role ontology based on competences retrieved from job role descriptions and a learning outcome ontology based on competences claimed and/or extracted from descriptions of a given training

program. SAKE project concerned several ICT job profiles and Business Informatics degree program, whilst the goal of OntoHR was to build an ontology-based selection and training system based on Information System Analyst (ISA) job role. One module of this system deals with the evaluation of the ICT degree programs. In these projects, the job role ontology reflected only a static moment of requirements of the labor market. In SAKE project, the job advertising documents were downloaded and tagged manually. In OntoHR, the ontology elements were extracted from the detailed descriptions of ISA job profile given by public organizations (e.g., O\*Net) or by projects concerned job analysis (e.g., EUQuaSIT). In the current research, a system is under development, which aims at formalizing the job requirements derived from IT/Telecommunication category of a popular job recruitment portal (Profession.hu) into the Job Role Ontology and matching this ontology to the Learning Outcome Ontology, which is created by the learning outcomes, and materials of Business Informatics Bachelor's degree program [17].

In Section 2, it is presented why the competence as a phenomenon gives the basis of this comparison. In Section 3, an incremental software development process is depicted, creating a prototype because there are not enough resources to implement all learning materials. Finally, conclusion and future work are shown.

## II. COMPARISON THROUGH COMPETENCES

In the previous work [17], it was shown that competence concept has several definitions in the literature due to contextual discrepancies of its usage, cultural traditions of the authors, and different epistemological foundations [19]; but, according to the presented definitions, common content elements (skills, knowledge and attitudes) were revealed. On the demand side of labor market (job demand), the importance of this concept was shown by the advantages of switching from job-based to competency-based organizational approach [8], by its strategic importance presented by Schoonover and Andersen [14], and by the role of updating competency models and job descriptions in talent specific succession planning [4].

On the supply side of labor market (education side), qualification frameworks based on competences (like European Qualifications Framework [5], Framework for

Qualifications of the European Higher Education Area [2]) give a guideline to develop the national framework like OKKR in Hungary [20].

Therefore, competence seems to be an appropriate base to achieve the comparison between the two sides of the labor market. (In English the competence and competency concept are distinguished. This paper follows the guideline of Hungarian public education that uses the first interpretation.)

### III. THE SYSTEM DEVELOPMENT

The learning outcomes of the above-mentioned degree program have not been changed since 2005, so the fundamental requirements related to the system are to adopt the changes occurred in job demand and to achieve the matching process with minimal human intervention. The system is capable of:

- collecting job requirements from the Internet in an automatic manner, extracting knowledge elements of them and forming these elements into the Job Role Ontology in a semi-automatic manner; formalizing the actual status into the Learning Outcome Ontology;
- achieving the matching process between Competence classes or its subclasses of both ontologies and evaluating the results.

We state that these requirements delineate into two development phases, the incremental system development methodology seems to be usable. Ontology learning is touched in the first stage and ontology matching in the second stage.

#### A. First development phase: Ontology building and learning

The objective of ontology learning is “to generate domain ontologies from various kinds of resources by applying natural language processing and machine learning techniques” [6]. The input of this phase is a collection of job requirements from the above-mentioned portal. A crawler was written in Java, to be responsible for ensuring this input - at given intervals.

Having examined the resulted collection, some problems were revealed. These problems and the related solutions are depicted in the next table:

TABLE I. PROBLEMS WITH THE JOB ADVERTISING COLLECTION

Problem	Solution
In one month approximately 500 advertisements usually bear. Among them, there are several identical documents or documents showing few discrepancies (for example the contact person’s name).	DOS Batch program find the same files, leave one file of them and delete the others.
HTML tags do not refer to its content. (For example: <h3>Requirement(s):</h3>)	Searching another patterns. For example: blocks assigned by colon.
The “requirement:” block is missing of certain job advertisements. If they exist, they contain only little information about competences.	The most job advertisements contain task description block, so competences have to be assigned to tasks (e.g. based on the knowledge elements of The Open Group Architecture Framework [18], an Open Group standard).
The documents are in XHTML formats, which are unstructured and customized by advertisers. It is ambiguous to process them.	After identifying the task: block it is necessary to create XML files from simple text. Java SAXparser() and DefaultHandler() classes can process XML files.

After these steps, a bouquet of XML files is created.

The first version of the Job Role Ontology is built on a collection derived from the first quarter of 2011. The meta-model of the ontologies is presented by Figure 1.

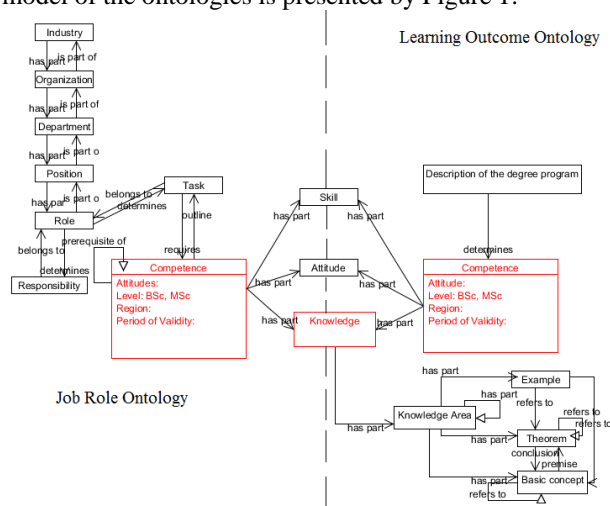


Figure 1. The meta-model of the Job Role Ontology and the Learning Outcome Ontology

In the meta-model of the *Job Role Ontology*, the *Industry*, *Organization*, *Department* and *Position* classes put the competences into an organizational context. Within an organization, the business processes consist of tasks that

roles and responsibilities belong to. In the backward direction, the *Role* as the parts of the *Position* class determines the entities of the *Task* and the *Responsibility* class. *Competence(s)* are required to execute a *Task*. The attitudes of the *Competence* class facilitate to execute the comparison at the appropriate Level, in the Period of Validity and in a given Region.

The meta-model of the Learning Outcome Ontology is an extended version of the OntoHR project's Educational Ontology [7] by the *Description of the Degree Program*, as sources of the competences, and by the attitudes of the *Competence* class.

The elements of competences mentioned in Section 2 (like *Skill*, *Attitude* and *Knowledge*) represent the basis of the comparison, but in the prototype we use only the *Knowledge* class to execute the comparison as we will see in the next section.

The Task class plays an important role in the construction of the Job Role Ontology. But too many positions and related tasks appear in the job advertisement collection, so we had to choose a position (like Software Developer position), its roles (Developer role and Contact Person role) and its related tasks (Designing the software development process, Preparing specification, Program coding, Program testing, Bug fixing and Communicating) to create the first version of the ontology. The knowledge extraction algorithm collects the concrete appearance of these tasks from the appropriate job advertisements and fitting them into this ontology. The steps of this algorithm are the following ones:

- To define a process whose tasks will be in the center of interest and formalize them into a first version of the ontology;
- To filter the job advertisements by their relevancy related to this process and the existence of tasks: block in order to cut this text block from the advertisements;
- To search expressions as patterns to describe a task (for example task – Communicating, expression: relation with customers or task – Designing the software development process, expression: design of an embedded software);
- To use these expressions like open sentences (for example (relation with; who) or (design of, something));
- To search the given words of the open sentences (e.g. relation, design) in the job advertisements and the nouns forming an expression with its preposition (e.g., with or of). Based on the position of these words in the text, we decide about that these nouns may be appropriate or not;
- To put the found expressions (e.g. relation with customers, or design of application), as subclass of the Task subclass related to the given expression,

and its original texts, as comments, into the ontology.

Having executed this algorithm, the first version of the Job Role Ontology is implemented in Protégé 4.2 [16] ontology development tool by a Protégé API written in JAVA. It is presented by Figure 2.

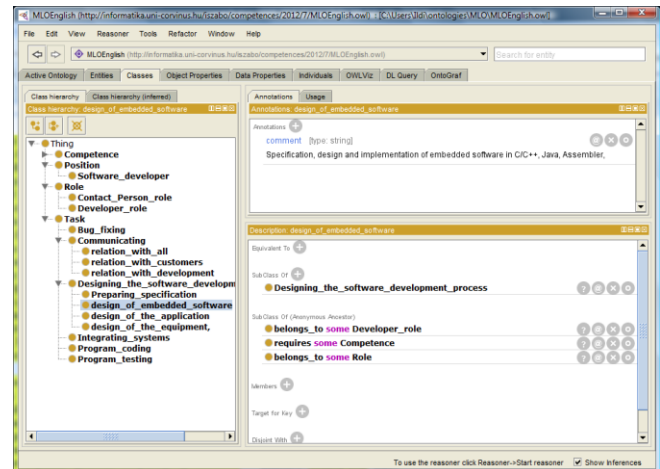


Figure 2. The implementation of the first version of the Job Role Ontology

As this figure illustrates, the first ontology version contains the main tasks of the determined process (software development process) that are expanded by the results of its open sentences. For example, the task is “communicating”, and it is expanded by the expression “relation with customer” or “relation with development” etc. due to the indirect object found in the open sentence (relation with; who).

After this step, the competence elements (mainly the knowledge elements) of TOGAF will be assigned to the appropriate Task subclasses in manual or semi-automatic manner. Based on the requirements appeared in the job advertisements (for example Generic knowledge in Unix / Linux, AIX or Windows), an algorithm will validate or complete these competence elements and determines the attitudes of the competences. The development of these algorithms is under way.

In this phase, we plan to evaluate the results given by these algorithms by the measure of needed human intervention.

Having constructed the Job Role Ontology by this approach and formalized the actual knowledge elements into the Learning Outcome Ontology, we can pass into the next stage.

#### B. Second development phase: Ontology matching

Alasoud, Haarslev and Shiri [1] define Ontology matching problem as follows: “given ontologies O1 and O2, each describing a collection of discrete entities such as classes, properties, individuals, etc., we want to identify

semantic correspondences between the components of these entities.”

In the first version of the prototype, the comparison between both sides will be executed through the Knowledge class, because the knowledge elements can be measured and can be assigned to the tasks more unambiguously than the other elements. This research concerns on finding the semantic and/or structural correspondences between the individuals of the Knowledge class of both ontologies.

In the research, ontology matching systems proposed by Choi [3] (Glue, Mafra, Lom, Qom, Onion, Omen) and offered by Noy [11] (Prompt, IF-Map) were investigated according to the following features:

- ontology matching is achieved in dynamic manner:
  - automatic, semi-automatic or non automatic working
  - the handling of changes occurred in the ontology
- reusability:
  - usage of different ontology format in matching process
  - type of matching method
  - modularity, integration with other systems
  - adaptability in Hungarian language environment.

Based on these characteristics MAFRA [9] and PROMPT [15] (or its built-in version into Protégé 4.2) ontology matching tools seem to be most suitable to achieve matching process. They are free downloadable, to execute from command prompt or a Java program automatically, to support RDF(S) or OWL languages and to handle changes occurred in the ontology through the usage of a semantic bridge or Protégé ontology editor. These are the most advantages of these programs compared to the others. However, they need human intervention as against IF-MAP. Nevertheless, the usage of algorithms from other systems (e.g., the one developed in OntoHR) can be taken into consideration.

In this phase, we plan to evaluate the results versus the results given by a human comparison.

#### IV. CONCLUSIONS AND FUTURE WORK

This two-phased incremental software development process creates a prototype, which is capable of building the Job Role Ontology from the actual job requirements and executing a matching algorithm to reveal same and different elements between this ontology and the Learning Outcome Ontology. In this prototype, only one position is implemented but it is extendable with others in same way.

Considering carefully the system’s requirements, detailed in the previous section, programming in Java seemed to be the most appropriate tool to develop the system. The main arguments are, that it provides a simply way to download contents from websites, it can be capable of running external commands (batch files) to create XML files, in order to put

knowledge elements extracted from these files into ontology format (like RDF or OWL 2.0 format in Protégé). MAFRA, PROMPT and Protégé 4.2 open source programs are written in Java, too.

The XML creator program, the Learning Outcome Ontology, the algorithm for extracting tasks from the job requirements and putting them into the first version of the Job Role Ontology are ready to use. The future work is to develop an algorithm to assign the TOGAF knowledge elements to the relevant tasks and to find an appropriate tool or algorithm to achieve the matching process. These features must be integrated into one system in order to achieve the same-time process execution.

#### ACKNOWLEDGMENT

The author wishes to express her gratitude to her supervisor Dr. Andras Gabor for the great topic and the powerful help provided during the development process.

#### V. REFERENCES

- [1] A. Alasoud, V. Haarslev, and N. Shiri, “An Effective Ontology Matching Technique,” in *Foundations of Intelligent Systems*, vol. 4994, A. An, S. Matwin, Z. Ras, and D. Slezak, Eds. Springer Berlin / Heidelberg, 2008, pp. 585–590.
- [2] Bologna Working Group on Qualifications Frameworks, “A framework for qualifications of the European Higher Education Area.” [Online]. Available: [http://www.bologna-bergen2005.no/Docs/00-Main\\_doc/050218\\_QF\\_EHEA.pdf](http://www.bologna-bergen2005.no/Docs/00-Main_doc/050218_QF_EHEA.pdf). [Accessed: 05.09.2012]
- [3] N. Choi, I.-Y. Song, and H. Han, “A survey on ontology mapping,” *SIGMOD Rec.*, vol. 35, pp. 34–41, Sep. 2006.
- [4] L. Egodigwe, “Pipeline to success,” *Black Enterprise*, vol. 36, no. 10, 2006.
- [5] European Commission, “The European Qualifications Framework (EQF).” [Online]. Available: [http://ec.europa.eu/education/lifelong-learning-policy/eqf\\_en.htm](http://ec.europa.eu/education/lifelong-learning-policy/eqf_en.htm). [Accessed: 05.09.2012].
- [6] P. Haase and J. Völker, “Ontology Learning and Reasoning — Dealing with Uncertainty and Inconsistency,” in *Uncertainty Reasoning for the Semantic Web I*, vol. 5327, P. da Costa, C. d’Amato, N. Fanizzi, K. Laskey, K. Laskey, T. Lukasiewicz, M. Nickles, and M. Pool, Eds. Springer Berlin / Heidelberg, 2008, pp. 366–384.
- [7] G. Kismihók, I. Szabó, and R. Vas, “Six Scenarios of Exploiting an Ontology Based, Mobilized Learning Environment,” *International Journal of Mobile and Blended Learning*, vol. 4, no. 1, pp. 45–60, 2012.
- [8] E. E. Lawler, “From job-based to competency-based organizations,” *J. Organiz. Behav.*, vol. 15, no. 1, pp. 3–15, Jan. 1994.
- [9] MAFRA, “Ontology Mapping FRAMework Toolkit”. [Online]. Available: <http://mafra-toolkit.sourceforge.net/>. [Accessed: 05.09.2012]
- [10] Ministry for National Economy, “Hungary’s Structural Reform Programme 2011-2014.” Mar-2011. [Online]. Available: <http://www.kormany.hu/download/b/23/20000/Hungary%27s%20S tructural%20Reform.pdf>. [Accessed: 05.09.2012].

- [11] N. F. Noy, "Semantic integration: a survey of ontology-based approaches," SIGMOD Rec., vol. 33, pp. 65–70, Dec. 2004.
- [12] OntoHR, "Ontology based competency matching", 504151 - LLP -1 - 2009-1-HU-LEONARDO-LMP. [Online]. Available: <http://ontohr.eu>. [Accessed: 05.09.2012].
- [13] SAKE, "Semantic enabled Agile Knowledge-based E-government", FP6 IST 027128. [Online]. Available: <http://www.sake-project.org/>. [Accessed: 05.09.2012]
- [14] S. C. Schoonover, H. Schoonover, D. Nemerov, and C. Ehly, "Competency-based HR applications: Results of a comprehensive survey," Arthur Andersen/Schoonover/SHRM, 2000.
- [15] Stanford University, "PROMPT". [Online]. Available: <http://protege.stanford.edu/plugins/prompt/prompt.html>. [Accessed: 05.09.2012]
- [16] Stanford University, "Protégé 4.2 Beta". [Online]. Available: [http://protegewiki.stanford.edu/wiki/Protege\\_4.2\\_Beta\\_Release\\_Notes](http://protegewiki.stanford.edu/wiki/Protege_4.2_Beta_Release_Notes). [Accessed: 05.09.2012]
- [17] I. Szabó, "Comparing the competence contents of demand and supply sides on the labour market," in Information Technology Interfaces (ITI), Proceedings of the ITI 2011 33rd International Conference on, 2011, pp. 345–350.
- [18] TOGAF, "The Open Group Architecture Framework". [Online]. Available: <http://www.opengroup.org/togaf/>. [Accessed: 05.09.2012]
- [19] J. Winterton, F. Delamare - Le Deist, and E. Stringfellow, "Typology of knowledge, skills and competences Clarification of the concept and prototype." Cedefop Reference series, 2006.
- [20] J. Temesi, Ed., Az Országos képesítési keretrendszer kialakítása Magyarországon. Budapest: Oktatókutató és Fejlesztő Intézet, 2011.

## Consolidation of Linked Data Resources upon Heterogeneous Schemas

Aynaz Taheri

NLP Research Lab, Computer Engineering Dept  
Shahid Beheshti University  
Tehran, Iran  
ay.taheri@mail.sbu.ac.ir

Mehrnoush Shamsfard

NLP Research Lab, Computer Engineering Dept  
Shahid Beheshti University  
Tehran, Iran  
m-shams@sbu.ac.ir

**Abstract**— Linked data resources have influential roles in conducting the future of semantic web. They are growing more and more, and the amount of published data is increasing at a fast pace. It causes some new concerns arise in the context of semantic web. One of the most important issues is the large amount of data that is produced as identical entities in heterogeneous data sources by different providers. This is a barrier to intelligent applications or agents that are going to utilize linked data resources. It prevents us from utilizing the potential capacity of web of data. Linked data resources are valuable when we could exploit them altogether. Therefore, we could obviously perceive the importance of linked data integration. In this paper, we propose a new approach for linked data consolidation. It helps us to have a consolidation process even between resources with heterogeneous schemas. In this approach, we are going to find more identical instances locally. This means that we direct our instance coreference resolution around the two instances which are certainly identical. The neighbors of two similar instances are a good source for our approach to proceed. In addition, these neighbors are beneficial for estimating some similarities between concepts of two heterogeneous schemas.

**Keywords**—Linked Data; Consolidation; Ontology; Schema; Instance.

### I. INTRODUCTION

Linked data has profound implications for the future of semantic web. Nowadays, the amount of published linked data is increasing and web of data is growing more and more. Linking Open Data (LOD) [24] project is the realization of web of data. Web of data includes billions of RDF [25] triples that are accumulated by different data providers. Accretion of data in Linking Open Data project is not the only challenge of publishing linked data; rather, matching and linking the linked data resources are also equally important and can improve the effective consuming of linked data resources. Linked data integration is one of the main challenges that become more important considering development of linked data. Without these links, we confront with isolated islands of datasets, which could not exploit knowledge of each other. The fourth rule of publishing linked data in [1] explains the necessity of linking URIs to each other. When there are possibilities of applying integrated linked data sources, information retrieval and utilizing linked data on the web would be thriving. Thus, we need identification and disambiguation of entities in different data sources. Unique entity identification in variant resources

causes reduction of problems about data processing in heterogeneous data resources.

We created a new approach for entity coreference resolution in linked data resources. The proposed approach receives two ontologies, two sets of instances as linked data sources and two similar concepts from two ontologies. Instance matching algorithm initiates its process among the instances of two similar concepts that are received from the inputs. In fact, the instance matcher is now assured of equality of these two concepts and knows that it can find identical instances among the instances of the two concepts. Our approach searches for finding identical instances by applying a new method that is explained in section 2. We use the properties of instances and their values to discover similar instances. In addition, neighbors of instances are the other significant features that we apply for identifying instances. Neighbors have prominent roles in the performance of our method. After finding identical instances, we continue the process locally around the identical instances. Identical instances are good points for our algorithm to proceed since searching around two identical instances raises the possibility of finding equal instances. Another great merit of finding similar instances in the neighborhood of identical instances is to help us contend with heterogeneous schemas. Section 3 explains about this issue.

This paper is structured as follows: Section 2 outlines the instance matching algorithm. Section 3 explains how instance matching of our approach helps us in overcoming difficulties of schemas heterogeneity. Section 4 discusses our experiments over one dataset. Section 5 points to some related works and finally, Section 6 concludes this paper.

### II. INSTANCE MATCHING

The process of instance coreference resolution needs to receive a pair of concepts from two ontologies. These two concepts are equal and we are going to find identical individuals among their instances.

#### A. Create a Net around the Instances

We introduce a new construction that is called *Net*, as the basis of our instance matching algorithm.

For two equivalent concepts that we receive as input, we must create Nets. For each instance of two concepts, we make one Net. If we have an instance that its URI is 'i', we explain how to create a Net for instance 'i'. For creating this Net, all of the triples whose subjects are instance 'i' are



extracted and added to the Net. Then, in the triples that belong to the Net, we find neighbors of instance 'i'. If instance 'j' is one of the neighbors of instance 'i', the same process is repeated for instance 'j'. Triples, whose subjects are instance 'j', are added to the Net, and the same process is repeated for neighbors of the instance 'j'. This process is actually like depth first search among neighbors of instance 'i'. To avoid falling in a loop and eliminating the size of search space, the maximum depth of search is experimentally set to 5. This depth gives us the best information about an instance and its neighbors. The Net that is created for instance 'i' is called  $Net_i$ . Starting point for this Net is instance 'i'.

The process of creating Nets is done for all of the instances of the two concepts. Creating Nets helps us in recognizing instance identities. Identities of instances are sometimes not recognizable without considering the instances that are linked to them, and neighbors often present important information about intended instances. In some cases in our experiments we observed that even discriminative property-value pairs about an instance may be displayed by its neighbors. Figure 1 shows an illustration about an instance that its neighbors describe its identity. This example is taken from IIMB dataset in OAEI 2010. Figure 1 shows  $Net_{item2117}$ . 'Item2117' is the starting point of this Net and is an Actor, Director and a character-creator. Each instance in the neighborhood of 'Item2117' describes some information about it. For example, 'Item7448' explains the city that 'Item2117' was born in and 'Item2705' explains the name of the 'Item2117'.

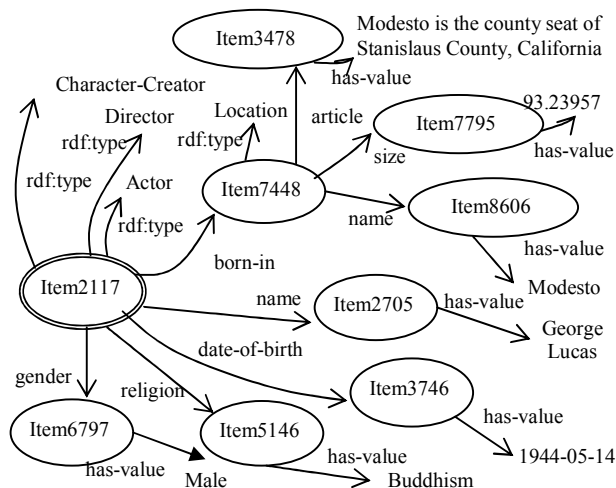


Figure 1. An Illustration of Net

Not only does creating Nets help us in discovering identities of instances, but also it helps us to find locally more similar instances. This issue is explained in the second part of Section B.

### B. Compute the Nets Similarities

In the previous step, Nets of two equal concepts were created. In this step, we must compare them.

#### 1) Finding identical instances

Each Net from one concept is supposed to be compared with all Nets of the other concept in order to find similar Nets. Starting points of two similar Nets would be equal. Each Net is composed of some triples that are extracted from the dataset. Therefore, triples of two Nets should be compared. In this process, only triples whose objects are data type values (and not instances) would participate in the comparison. Properties values are very important in comparison.

We use edit distance method for comparing string values of properties. Some properties explain comments about instances. In these situations, we used a token-based measure for computing similarity.

Similarity values of triples objects are added together for obtaining similarity value of two Nets. We applied a threshold for Edit Distance method. This threshold was found by making a benchmark and execution of edit distance algorithm based on the benchmark. We round the threshold to one decimal point and the value of threshold is 0.6.

After calculating similarity of properties values, we compute similarity of two Nets. Similarity of two Nets is dependent on similarity of their properties values. Triples in two Nets have specific importance depend on the depth of their subjects (instances that triples belong to) in the Nets. Depth of instances is estimated toward the starting point of the Net. When depth of an instance in a Net increases, its effectiveness on similarity computation of Nets decreases. The following triples belong to  $Net_{item2117}$  in Figure 1.

- 1 ('Item6797', has-value, Male)
- 2 ('Item3746', has-value, 1944-05-14)
- 3 ('Item7795', has-value, 93.23957)
- 4 ('Item3478', has-value, Modesto is the county seat of Stanislaus County, California)

The above triples describe some information about the starting point of  $Net_{item2117}$ . Two first triples explain that 'item2117' has male gender and date of his birth is 1994-05-14. Instances in the subjects of these two triples have depth equal to two. Two second triples explain that 'item2117' has born in a city that its size is 93.23957 and also is the county seat of Stanislaus County, California. Instances in the subjects of these two triples have depth equal to three. As you can see, the first two triples have more important role for determining the identity of 'item2117' than the second two triples. Gender of a person and date of his birth is more important than some comments about the city that he lives in. Nevertheless, this does not mean that existence of instances with greater depth are not beneficial in the Nets; rather, they are less important in identity recognition of the starting point of the Net than those with less depth.

In this regard, similarities of properties values are added with an particular coefficient. We use a weighted sum for computing similarity of Nets. The coefficients in this sum have inverse relations to the depth of the subject of triples in Net.

We normalize the sum of similarities of properties values in two Nets into a range of 0 and 1 by dividing the result to the sum of the numbers of triples in two Nets. After finding

the similarities between all the Nets of two concepts, we sort the similarity values in a list based on the descending order, and most similar Nets are selected respectively. An one to one relation is made between similar Nets. Nets with similarity values less than 0.5 are omitted. This threshold is obtained experimentally. We made a benchmark of our Nets and selected the best threshold which could represent us the similarity threshold.

When two Nets are selected as two similar Nets, we consider their starting points as identical instances. In this way, some identical instances could be found regarding to their properties and their neighbors.

#### 2) Finding identical instances in the vicinity of identical instances

We found some identical instances with utilizing their Nets. In this step, we continue the process of matching on those Nets of the previous step that led to discovering equal instances or in the other words, those Nets that have equal starting points. The strategy in this step is searching locally around the identical instances in order to find new equal instances. Seddiqui, et al. [20] created an algorithm for ontology matching and their algorithm is based on the idea that if two concepts of two ontologies are similar, then there is a high possibility that their neighbors are similar too. We use this idea but in instance level. This means that if two instances are identical, then there is possibility that their neighbors are similar too.

Suppose that 'i' and 'j' are two instances that are detected identical in the previous step. Their Nets are called  $Net_i$  and  $Net_j$ . In this step we describe how the approach finds more identical instances in  $Net_i$  and  $Net_j$ . For discovering similar instances in  $Net_i$  and  $Net_j$ , we compare instances in these two Nets. The process of comparing instances is similar to what mentioned in the first part of section B. Instances would be compared regarding their properties and values.

Finding identical instances of two concepts initially costs a lot in first part of section B because of considering all neighbors of an instance; later we can find locally more identical instances by paying low computational cost.

### III. COMPUTE CONCEPT SIMILARITIES IN SCHEMA LEVEL

After finding identical instances in the neighborhood of identical instances, now it is time to find similarities between concepts in two heterogeneous schemas. In this part, instance matcher gives feedback to us for finding similar concepts in schema level. If we find some similar instances such as 'm' and 'n' in the instances of  $Net_i$  and  $Net_j$ , concepts that 'm' and 'n' belong to them would be good candidates to be similar.

The approach repeats this step for every two similar Nets and considering to identical instances in two similar Nets, estimates similarities between concepts. We used a measure in order to find a similarity value between two concepts.  $C_1$  and  $C_2$  are two concepts that we made Nets for their instances and then compared their Nets.  $C_3$  and  $C_4$  are two concepts that we have concluded their similarity from the neighbor instances of  $C_1$  and  $C_2$  instances. Then, we define the similarity value of  $C_3$  and  $C_4$  based on the ratio of

neighbor instances of  $C_1$  and  $C_2$  instances that concluded similarity between  $C_3$  and  $C_4$ , to the number of Nets in  $C_1$  and  $C_2$ .

The approach gives us some similarity values between concepts of two ontologies. In the implemented approach, we did not apply any other methods for ontology matching. We used these similarity values and managed the matching process manually. In fact, similarity values conducted our matching process significantly. These equal concepts are inputs for the next execution of instance matcher.

### IV. EXPERIMENTS

We used a dataset of OAEI [5], a benchmarking initiative in the area of semantic web. We report the experimental results of our proposed approach on IIMB dataset in OAEI 2011. IIMB composed of 80 test cases. Each test case has OWL ontology and a set of instances. Information of test cases in IIMB track is extracted from Freebase dataset. IIMB divided test cases in four groups. Test cases from 1 to 20 have data value transformations, 21 to 40 have structural transformations, test cases from 40 to 60 have data semantic transformations and 61 to 80 have combination of these three transformations. All of these 80 test cases are supposed to be matched against a source test case. We choose IIMB 2011 test cases for the evaluation because this track of OAEI has all kinds of transformations and we could compare all aspects of our system against the other system. Moreover, the size of IIMB 2011 has increased greatly compared to last years and is more than 1.5 GB. Increased amount of the dataset size lets us evaluate scalability of our approach. Unfortunately, there has been just one participant in this track, CODI [10], with which we will compare our results. This shows the scalability difficulties in systems performance at large scale datasets.

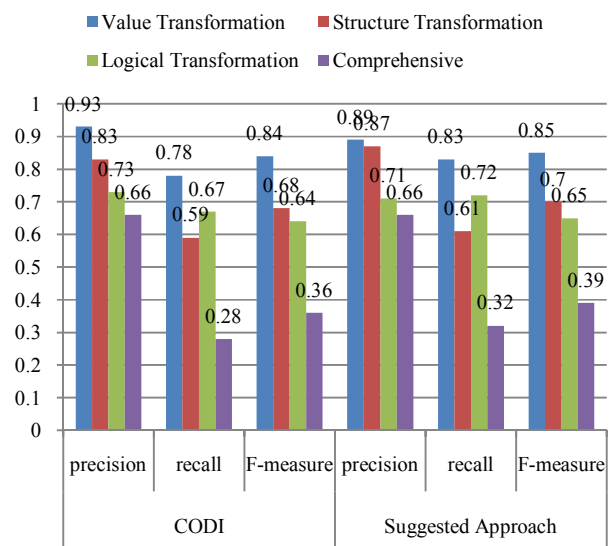


Figure 2. Results of OAEI'11 IIMB Track

We observe in Figure 2 that the recall values of our approach in four kinds of transformations are better than

CODI but this is not always true for precision value. The operations of our approach is clearly better than CODI in datasets with structure transformation considering three aspects of precision, recall and F-measure. This means that our approach is more stable in modifications such as removing, adding and hierarchal changing of properties.

## V. RELATED WORK

The problem of entity coreference resolution is not a new challenge. There are a large number of related works on this issue in the context of database and the problem is called record linkage. We state some of these works in the area of entity coreference resolution in the context of semantic web. Raimond, et al. [16] proposed a method for interlinking two linked data music-related datasets that have similar ontologies. Hassanzadeh and Consense [6] described how they interlinked a linked data source about movies with other data sources in LOD by applying some exact and approximate string similarity measures. In [22], a method for linking WordNet VUA (WordNet 3.0 in RDF) to DBpedia is proposed. Finding identical instances of foaf:person at social graph is explained in [17] by computing graph similarity. Hogan, et al. [7] proposed an approach that capturing similarity between instances is based on applying inverse functional properties in OWL language. Noessner, et al. [15] used a similarity measure for computing similarity of instance matching between two datasets with the same ontology. LN2R [18] is a knowledge based reference reconciliation system and combines a logical and a numerical method. Hogan and colleagues [8] proposed a method for consolidation of instances in RDF data sources that is based on some statistical analysis. ObjectCoref [9] is a self-training coreference resolution system based on a semi supervised learning algorithm. Song and Heflin [21] described an unsupervised learning algorithm in order to find some discriminable properties as candidate selection key. Zhishi.links [14] is a distributed instance matching system. It does not follow any special techniques for schema heterogeneity. It uses an indexing process on the names of instances. HMatch( $\tau$ ) [3] is an instance matcher and use HMatch 2.0 for TBox matching and then tries to capture the power of properties at instance identification. RiMOM [23], ASMOV [12] and AgreementMaker[4] are three ontology matching systems that recently equipped with instance matchers. CODI [10] is also a system for ontology and instance matching and is based on markov logic. Nikolov and colleagues proposed Knofuss architecture [13] that contains both schema and instance level. Linked Data Integration Framework (LDIF) [19] has two main components Silk Link Discovery Framework [11] and R2R Framework [2] for identity resolution and vocabulary normalization respectively.

What distinguish our approach from the aforementioned approaches is that our approach considers that the neighbors of an instance are important in order to find similarity between identical instances. We proposed a new approach for finding identical instances.

## VI. CONCLUSION AND FUTURE WORK

In this paper, we proposed a new approach for linked data consolidation. Instance resolution process starts after getting two equal concepts as input by instances matcher. Instance matcher creates Nets around the instances of two equal concepts and then compares these Nets. Our approach selects the Nets with most similarity value and considers that as similar Nets. Similar Nets have identical instances in their starting points. Instance matcher searches instances in the similar Nets in order to find identical instances around their equal starting points. After discovering instances with the same identity in Nets, instance matcher utilizes them and computes some similarity values between concepts in the schema level. It sends us most similar concepts as a feedback for starting the instance matching again.

Our future target includes utilizing some methods for schema matching in our approach. We could devise a schema matcher for our approach so that schema and instance matchers could perform consecutively. Furthermore, we must apply a better method for finding the threshold that is the final approver of two similar Nets. It is better to find a heuristic measure in order to find a dynamic threshold.

## REFERENCES

- [1] C. Bizer, T. Heath, and T. Berners-Lee, "Linked Data-The Story So Far," *Int. J. Semantic Web Inf. Syst.* vol. 5, no. 3, 2009, pp. 1-22.
- [2] C. Bizer and A. Schultz, "The R2R Framework: publishing and discovering mapping on the web," *Proc. 1<sup>st</sup> International Workshop on Consuming Linked Data*, Shanghai, China, Nov. 2010.
- [3] S. Castano, A. Ferrara, S. Montanelli, and D. Lorusso, "Instance matching for ontology population," *Proc. 16th symposium on advanced database systems*, Italy, Jun. 2008, pp. 121-132.
- [4] I. F. Cruz, C. Store, F. Caimi, A. Fabiani, C. Pesquita, F. M.Couto, and M. Palmonari, "Using AgreementMaker to align ontologies for OAEI 2011," *Proc. 6<sup>th</sup> International Workshop on Ontology Matching*, Bonn, Germany, 24 Oct. 2011.
- [5] J. Euzenat, C. Meilicke, H. Stuckenschmidt, P. Shvaiko, and C. Trojahn, "Ontology Alignment Evaluation Initiative: Six Years of Experience," *J. Data Semantics*, vol. XV, 2011, pp. 158-192.
- [6] O. Hassanzadeh and M. Consense, "Linked movie data base," *Proc. 2<sup>nd</sup> Link Data on the Web*, Madrid, Spain, Apr. 2009.
- [7] A. Hogan, A. Harth, and S. Decker, "Performing object consolidation on the semantic web data graph," *Proc. 1st Identity, Identifiers, Identification Workshop*, Banff, Canada, May 2007.
- [8] A. Hogan, A. Polleres, J. Umbrich, and A. Zimmermann, "Some entities are more equal than others: statistical methods to consolidate linked data," *Proc. 4th International Workshop on New Forms of Reasoning for the Semantic Web*, Crete, Greece, May 2010.
- [9] W. Hu, J. Chen, and Y. Qu, "A Self-training Approach for Resolving Object Coreference Semantic Web," *Proc. 20<sup>th</sup> International World Wide Web Conference*, India, Mar. 2011, pp. 87-96.
- [10] J. Huber, T. Sztyley, J. Noessner, and C. Meilicke, "CODI : Combinatorial Optimization for Data Integration – Results for OAEI 2011," *Proc. 6<sup>th</sup> International Workshop on Ontology Matching*, Bonn, Germany, Oct. 2011.
- [11] R. Isele, A. Jentzsch, and C. Bizer, "Silk server- adding missing links while consuming linked data," *Proc. 1<sup>st</sup> International Workshop on Consuming Linked Data*, Shanghai, China, Nov. 2010.
- [12] Jean-Mary, Y.R., Shironoshita, E.P., and Kabuka, M.R. "ASMOV: Results for OAEI 2010," *Proc. 5<sup>th</sup> International Workshop on Ontology Matching*, Shanghai, China, Nov. 2010.
- [13] A. Nikolov, V. Uren, E. Motta, and A. Roeck, "Overcoming schema heterogeneity between linked semantic repositories to improve coreference resolution," *Proc. 4<sup>th</sup> Asian Semantic Web Conference*, Shanghai, China, Dec. 2009, pp. 332-346.

- [14] X. Niu, S. Rong, Y. Zhang, and H. Wang, "Zhishi.links results for OAEI 2011," Proc. 6<sup>th</sup> International Workshop on Ontology Matching, Bonn, Germany, Oct. 2011.
- [15] J. Noessner, M. Niepert, C. Meinel, and H. Stuckenschmidt, "Leveraging Terminological Structure for Object Reconciliation," Proc. 7<sup>th</sup> Extended Semantic Web Conference, Heraklion, Greece, May 2010, LNCS 6089, pp. 334-348.
- [16] Y. Raimond, C. Sutton, and M. Sandler, "Automatic Interlinking of music datasets on the semantic web," Proc. 1<sup>st</sup> Link Data on the Web, Beijing, China, Apr. 2008.
- [17] M. Rowe, "Interlinking Distributed Social Graphs," Proc. 2<sup>nd</sup> Linked Data on the Web Workshop, Madrid, Spain, Apr. 2009.
- [18] F. Sais, N. Niraula, N. Pernelle, and M. Rousset, "LN2R a knowledge based reference reconciliation system: OAEI 2010 results," Proc. 5<sup>th</sup> International Workshop on Ontology Matching Shanghai, China, Nov. 2010.
- [19] A. Schultz, A. Matteini, R. Isele, C. Bizer, and C. Becker, "LDIF-Linked data integration framework," Proc. 2<sup>nd</sup> International Workshop on Consuming Linked Data, Bonn, Germany, Oct. 2011.
- [20] Md. Hanif Seddiqui, and M. Aono, "An Efficient and Scalable Algorithm for Segmented Alignment of Ontologies of Arbitrary Size," J. Web Sem. Vol. 7, Jan. 2009, pp. 344-356.
- [21] D. Song, and J. Heflin, "Automatically generating data linkage using a domain-independent candidate selection approach," Proc. of the 10th International Semantic Web Conference, Koblenz, Germany, Oct. 2011, LNCS 7031, pp. 649-664.
- [22] A. Taheri, and M. Shamsfard, "Linking WordNet to DBpedia," Proc. 6<sup>th</sup> Global WordNet Conference, Matsue, Japan, Jan. 2012, ISBN: 978-80-263-0244-5, pp. 344-348.
- [23] Z. Wang, X. Zhang, L. Hou, Y. Zhao, J. Li, Y. Qi, and J. Tang, "RiMOM Results for OAEI 2010," Proc. 5<sup>th</sup> International Workshop on Ontology Matching. Shanghai, China, 2010.
- [24] [www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData](http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData) [retrieved: July, 2012]
- [25] "RDF Vocabulary Description Language 1.0: RDF Schema", <http://www.w3.org/TR/rdf-schema/> [retrieved: July, 2012]

# Strategies for Semantic Integration of Energy Data in Distributed Knowledge Bases

Álvaro Sicilia, Fàtima Galán, Leandro Madrazo

ARC Enginyeria i Arquitectura La Salle

Universitat Ramon Llull

Barcelona, Spain

[asicilia, fatima, madrazo]@salleurl.edu

**Abstract** — This paper reports on the work that is currently being developed in RÉPENER, a research project co-financed by the Spanish National RDI plan 2009-2012. The objective of the project is to apply Semantic Web technologies to create an energy information system which puts together data from different sources, both private and public. To create such a system, it is necessary to integrate different data sources from different domains. Different strategies might be adopted, depending on the contents of the data sources involved. One of them is about adding new external data sources to create brand-new links between the existing ones. This is the strategy thus devised and implemented in this project. In this paper, a description of the process by which two databases with energy information have been linked using a SILK framework is provided.

**Keywords**-semantic integration; ontology design; object matching; energy data.

## I. INTRODUCTION

Energy-related information is dispersed in proprietary databases and open data sources. It is heterogeneous, since it is generated by different applications and for various purposes (modelling and simulation programs, monitoring systems), and it is compartmentalized by reference to the various stages of the building lifecycle; from design, to construction and operation. For this reason, energy information cannot be properly processed and analyzed because there is a lack of interoperability between applications and databases. Consequently, decision-making actors cannot exploit the benefits of correlative data from different stages and sources due to the lack of a common vocabulary and to the difficulties of accessing the data.

The application of Semantic Web technologies can help to overcome all of these limitations through the application of semantic data-integration processes. In recent years, studies on data integration using ontologies have delivered substantial results. The prime example, which proves the feasibility of tasks solutions, is the Linked Open Data project [1]. Therefore, semantic integration is a core issue in interoperability, particularly in a heterogeneous setting such as the World Wide Web, where different ontologies are used. Semantic integration inevitably leads to inter-ontological mapping, or ontology integration. As stated by Zhdanova [2] and Euzenat [3], ontology matching is a plausible solution to the problem of semantic heterogeneity in many applications. Once the matching is done, the conjunction of ontologies and

their interconnections facilitate an integrated access to heterogeneous energy data by providing: (i) a common vocabulary to unify different areas of knowledge or expertise today separated, (ii) an integrated way to explore energy information and its related data; and (iii) a compound bulk data with which to perform data analysis using data-mining techniques. The third feature can retrofit the information system by adding new data relations, which in turn enhance the exploration experience.

The purpose of the RÉPENER [4] project is to develop an ontology-based information system which supports decision-making processes and knowledge discovery by actors who deal in energy management with respect to buildings. The semantic information system which is being developed addresses the interoperability issues between different data sources using semantic technologies. Ontologies are designed using the OWL standard language, and data is exposed on the Internet using the RDF [5] following the Linked Open Data initiative. In this way, the interoperability problem is solved, because all data sources are described by means of a common language – which can be processed by humans and/or machines – using standard protocols. A comprehensive project description can be found in “in press” [6].

The feasibility of the data-integration process and the quality of the interrelationships amongst different data sources is a key issue. In the ideal scenario, data sources of different domains overlap in some concepts, and this allows one to create links between them. For example, on the one hand, a building repository can contain building instances which have a property location naming the city in which the buildings are located. On the other hand, a spatial data repository can contain landmarks with property names. Therefore, both data sources can be connected through the properties' locations and names. However, in an actual scenario, where data sources cannot be modified the process of connecting them is not that simple because there might be elements which do not overlap.

This paper presents the semantic integration process which has been carried out in the RÉPENER project with the objective of integrating data sources having non-overlapping elements.

The content of the following sections of the paper is summarized next. Some of the current strategies, procedures and tools used to perform the integration of data sources are discussed in Section II. In Section III it is described the work done to connect energy related data from different sources

and domains. Finally, the conclusions which can be drawn from the application of the procedures are summarized in Section IV.

## II. STRATEGIES FOR SEMANTIC INTEGRATION

Historically, data has been stored in relational databases, usually available in offline environments and published on the Internet through web applications which interconnect web documents instead of data instances. The Semantic Web concept coined by Tim Berners-Lee [7] was subsequently undertaken by the Linked Data movement, which has called for the creation of a web of data using Uniform Resource Identifiers (URIs) as the resource identification, Hypertext Transfer Protocol (HTTP) as the universal data retrieval mechanism, and the resource description framework (RDF) as a data model describing things in the world [1].

The web of data is comprised of several heterogeneous data sources which describe different domains using a vocabulary handled by domain experts. The interconnection between data sources is possible thanks to the addition of semantics to data, as achieved by means of metadata descriptions, thereby guaranteeing the interoperability between data. Furthermore, these semantic layers, jointly with the links between the data sources, enable applications to perform smart data analysis which can actually enrich the data. For example, an application could retrieve energy certifications of buildings as built from a repository, in a specific area, and then complement them with regional socioeconomic data from a statistical database. All of these would be for the purpose of applying the data-mining process to compare the energy rating with the economic level of the selected area. Finally, a user can use this improved information based on this comparison to make better decisions.

Data sources are incorporated in the web of data through semantic integration processes in two steps: 1) publishing semantic data which is translating relational data to RDF format, with the objective of exposing data through an SPARQL end point, and of releasing the ontology which models the data, and 2) interlinking data sources between them to create a network, which can subsequently be exploited.

### A. Data transformation strategies

To perform the integration of data sources, the source data must first be transformed from a relational database to the RDF language following the Linked Data principles. Data transformation may be implemented as a static ETL (Extract, Transform and Load) process or as a query-driven dynamic process. The static transformation process creates an RDF dump following the application of certain mapping rules. The most significant drawback of the static process is that the most recent data might not be considered. Contrastingly, the dynamic transformation process uses simple queries to access the latest data. A survey published by the W3C RDB2RDF incubator group has identified several tools which can carry out both transformations – static and dynamic – such as Virtuoso RDF View, D2RQ, R2O, or Triplify [8]. The survey concludes that there is not a

standard method for the representation of mappings between RDB and RDF and recommends, whenever possible, to implement on-demand mapping to access the latest version of the data. In February 2012, the RDB2RDF group published a R2RML (Relational database to RDF Mapping Language) [9], a recommendation which is currently being implemented in various projects.

### B. Linking strategies

The integration process involves interlinking objects of one or several data sources. Each data source usually uses different URIs to identify objects based on domain criteria, even if they describe the same real-world object. Therefore, links between these objects cannot be obtained in a straightforward way. For this reason, finding out that two data objects refer to the same real-world object is a key issue for data integration. Object matching methods (also known as instance consolidation, record linkage, entity resolution or link discovery) are focused on identifying semantic correspondences between objects of different data sources.

Object matches can be set manually or automatically. Typically, manual matching is carried out with small data sources, where it is important to ensure the high quality of the correspondences. If the data source is large, however, it is better to apply automated or semi-automated proposals [1]. The creation of links between objects can be handled with a domain-specific approach or with a universal approach, as stated by Ngonga Ngomo [10]. The second type of approach is not depending of the domain of the data sources, and, therefore, it can be applied to different scenarios. To perform this task, Ngonga Ngomo has identified different tools, among which the SILK framework stands out [11], and proposes an outperformed framework. Both frameworks – SILK and LIMES – generate links between RDF objects based on several similarity metrics (e.g., Levenshtein [12], Euclidean [13], or Jaro-Winkler [14]), which can be chosen and tailored by the user. From a technological point of view, both frameworks gather data from a SPARQL [15] end point, which is described in a configuration file containing the metrics applied and the object selection restrictions.

The aforementioned tools are designed for data sources which contain overlapped objects. In this case, the linking process can be carried out with these tools by setting up the configuration file to match them with the proper similarity metric. As has been stated previously, overlapped objects are not the usual case, and for this reason it is necessary to apply elaborated procedures. Accordingly, this research work proposes a data integration strategy which is based on complementing data sources with external data in order to enable a successful object matching process (Fig. 1).

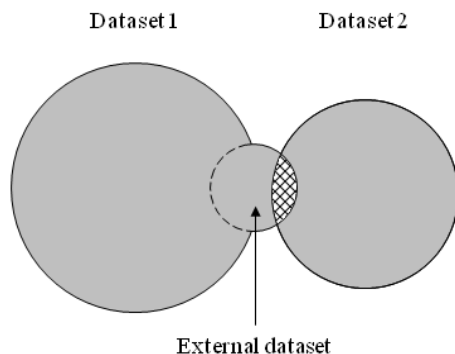


Figure 1. Elaborated data integration strategy.

The procedure of linking databases involves the attachment of external data properties to objects of a data source for, in a second step, applying the link-generation tools mentioned above. The first step is to identify the potential objects which can describe the same object of the world but do not have sufficient properties to be matched. For example, a data source DS1 which contains the monitoring data of a building whose location is described with a string property (e.g., the name of the place) and a data source DS2 that contains weather stations which are geo-located (e.g., longitude and latitude). Both objects might be connected through the property *hasWeatherStation*, where each building is linked to its closest station (Fig. 2). Because a string value and the set of longitude-and-latitude properties cannot be compared, a possible solution is to add *geo-localization* properties to the building objects or *place name* properties attached to weather-station objects.

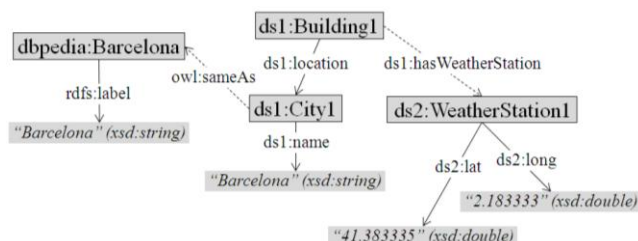


Figure 2. Object matching example.

Once the potential objects have been identified and analyzed, it is necessary to find new data properties in external data sources which can be used for object matching. The search process is carried out following the links between data sources, usually based on the *owl:sameAs* property. For specialized domains, it is important to be supported by a domain expert who can guide the exploration. In the previous example, city objects might have an *owl:sameAs* property linking DS1 with Dbpedia city objects which might have geo localization properties. When the links to external data sources are not available, it is necessary to generate them in order to access to the external data properties using link-generation tools. Finally, the data properties of external data sources are included in the existent data sources which can then be matched to other data sources.

The data sources are enriched by this method, which is based on the addition of new data properties gathered from external data sources instead of inferring new class relationships, taking into account the ontology itself as the enrichment step of the Linked Open Data life cycle.

This procedure has its weakness in the fact that external data must be added to the source-data sources. This is not always feasible. For instance, it occurs when the semantic data is generated through an RDF wrapper. In these cases, a possible solution is to publish RDF links in an intermediary RDF store (e.g. [16] service), which link-generation tools can then use to gather data.

### III. IMPLEMENTATION

This section describes the work which has been done to connect energy data sources applying the integration procedures previously described. The proprietary data sources used in the implementation have been provided by ICAEN – an organization of the Catalan government which gathers the energy certificates of newly planned buildings which include their simulated performance – and by Aemet, the Spanish meteorological agency which provides measurements made by a network of meteorological stations throughout Spain. The Ontology Engineering Group from the Universidad Politécnic de Madrid has published the Aemet data source through an SPARQL end point [17]. These data sources have been combined with the ultimate goal of assembling data from different sources and domains, thereby enabling the final user to understand the figures of buildings' energy certifications when applied to the building environment, particularly in the context of climate.

The semantic integration process is divided into two parts: the relational data transformation and the data interconnection.

#### A. Data transformation

The main purpose of data transformation is to create interconnected ontologies with the ability of integrating the different data sources. The data transformation embraces two actions: 1) the creation of an ontology which fits with the database, and 2) the transformation of data according to the ontology thus designed.

To create an ontology, the ICAEN data source [18] has been cleaned so as to eliminate unnecessary data, and consistency methods have been applied by energy domain experts. With the collaboration of energy domain experts and ontology engineers, an informal data structure has been developed in order to build an ontology which includes all the terms and categories identified. The ontology relies on a foundational ontology created for this project, which encompasses the building energy domain as well as other domains (social, economic). However, not all the data-source content has been contemplated in the ontology. For this reason, it has been necessary to define new concepts and relations to integrate this data in the ontology, as existing ontologies to be reused could not be found.

Once the ICAEN ontology has been created, the data is transformed to RDF. For the data transformation, a dynamic transformation process has been chosen, given the need for

access to the most recent data. The tool selected to carry out this work is D2RQ [19], because it supports database translation with high performance. It dynamically rewrites the SPARQL queries into SQL. This is a stable, lightweight solution. It represents mappings with an easily customizable D2RQ mapping language, enables configuration changes in real time, is independent of the database provider. It is currently being developed to support R2RML and it publishes data in HTML, RDF and through a SPARQL end point. The D2RQ tool has been configured to transform the ICAEN database according to the ontology thus designed.

**B. Data linking**

As a result of this transformation process, all of the data sources have become accessible through a SPARQL end point. The integration of the ICAEN and Aemet data sources has been accomplished, generating links between buildings' *ProjectData* and *WeatherStation* objects according to the *icaen:hasWeatherStation* property (Fig. 3) using the SILK tool [7].

The first attempt at integration was carried out by matching *ProjectData* and *WeatherStation* objects through the data properties: *icaen:ID\_Localitat* (the name of the city in which the building is located) and *aemet:stationName* (the name of the station, which is usually the same as the one of its closest city). The number of valid generated links was low, because nearly all station names were based on a mixture of the city name and internal terms, which interfered the matching. To solve this problem, we looked for external data sources which could provide additional data. The Linked GeoData repository – which provides spatial data such as roads, cities, mountains or points of interest – was selected as a source of external data. Linked GeoData objects are geo-located using latitude and longitude data properties, which are also used by weather stations. Therefore, it constitutes a feasible source of additional data.

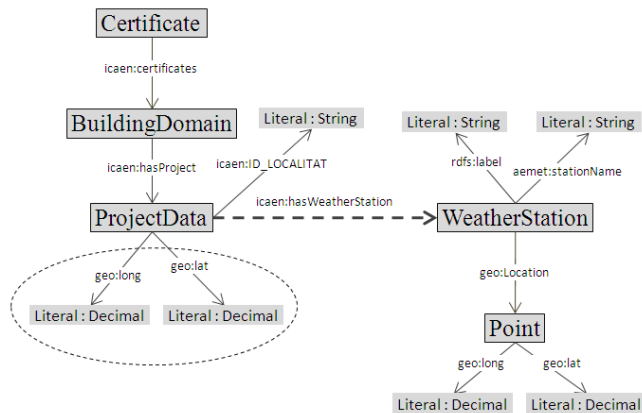


Figure 3. ICAEN and Aemet integration.

The SILK tool has been configured with the Levenshtein similarity function, which is best suited to compare string chains. ICAEN objects have been filtered by the *icaen:ProjectData* class, and Linked GeoData objects have been filtered by the *Igdo:City*, *Igdo:Town*, and *Igdo:Village* classes. The ICAEN data source contains 1,805 objects and

the Linked GeoData 720323. The SILK tool has found out 1,398 links between both data sources; thus, 77% of the buildings achieved links to a Linked GeoData place in less than an hour of execution time.

To attach geo-localization properties to ICAEN objects, a script which generates two RDF triples for each ICAEN object was developed: one for the latitude property and the other for the longitude one. The script queries the end points with SPARQL and generates a N-Triples file, which is later uploaded to the ICAEN data source.

Once the ICAEN objects contain geo localization properties from the Linked GeoData data source (the dotted circle in Figure 3), the SILK tool is called to generate links between the ICAEN and Aemet data sources. In this case, a geographical distance function is selected to use both the *geo:long* and *geo:lat* properties for the purpose of comparing objects. The Aemet data source contains only 260 weather stations, so the execution time is less than 4 seconds for generating 1,305 links, thereby covering 72% of the ICAEN buildings, or 93% of the ICAEN buildings which have links to linked GeoData objects (Table I).

TABLE I. LINK GENERATION COVERAGE

Icaen objects:	1805 (100%)
Linked to Linked GeoData:	1398 (77%)
Linked to Aemet:	1305 (72%)

**IV. CONCLUSIONS**

The work thus far developed facilitates semantic integration processes in linked data environments, taking advantage of existing links between data sources or by generating intermediary links. The procedure has been validated in a case study which demonstrates the feasibility of using external data sources to integrate semantic data.

It has been suggested that intermediary RDF stores can be used when it is not possible to add external data to a data source. As far as we know, this process is not possible using current link-generation tools, because they cannot integrate external data. Further work to be done in this regard is to have generation tools use federated queries in order to take advantage of the existing links.

Data sources will be released simultaneously with the user interface and project services by the end of the research project.

**ACKNOWLEDGEMENT**

RÉPENER is being developed with the support of the research program BIA 2009-13365, co-funded by the Spanish National RDI Plan 2009-2012.

**REFERENCES**

[1] T. Heath and C. Bizer, "Linked Data: Evolving the Web into a Global Data Space," Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 2011, pp. 1-136. Morgan & Claypool.



- [2] A. Zhdanova, "Towards a community-driven ontology matching," Proc. 3rd international conference on knowledge capture (K-Cap'05), ACM, Oct. 2005, pp. 221-222.
- [3] J. Euzenat, "Semantic technologies and ontology matching for interoperability inside and across buildings," Proc. 2nd CIB workshop on eeBuildings data models, Oct. 2011, pp. 22-34.
- [4] <http://arc.housing.salle.url.edu/repener> 09.08.2012
- [5] <http://www.w3.org/RDF/> 09.08.2012
- [6] L. Madrazo, A. Sicilia, M. Massetti, and F. Galan, "Semantic modeling of energy-related information throughout the whole building lifecycle," Proc. 9th European Conference on Product and Process Modelling (ECPM), Jul. 2012 .
- [7] T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web," Scientific American, vol. 284, May 2001, pp. 29-37, doi:10.1038/scientificamerican0501-34.
- [8] S. S. Sahoo, W. Halb, S. Hellmann, K. Idehen, T. Thibodeau, S. Auer, J. Sequeda, and A. Ezzat, "A Survey of Current Approaches for Mapping of Relational Databases to RDF," W3C publication, 2009. [http://www.w3.org/2005/Incubator/rdb2rdf/RDB2RDF\\_SurveyReport.pdf](http://www.w3.org/2005/Incubator/rdb2rdf/RDB2RDF_SurveyReport.pdf) (accessed July 16, 2012).
- [9] <http://www.w3.org/TR/r2rml/> 09.08.2012
- [10] A.-C. Ngonga Ngomo and S. Auer, "LIMES - A Time-Efficient Approach for Large-Scale Link Discovery on the Web of Data," Proc. 22nd International Joint Conference on Artificial Intelligence (IJCAI), IJCAI/AAAI, Jul. 2011, pp. 2312-2317.
- [11] J. Volz, C. Bizer, M. Gaedke, and G. Kobilarov "Discovering and maintaining links on the web of data," Proc. 8th International Semantic Web Conference (ISWC), Springer, Oct. 2009, pp. 650-665.
- [12] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," Soviet Physics Doklady, vol. 10, 1966, pp. 707-710.
- [13] E. Deza and Michel M. Deza, "Encyclopedia of Distances," Springer Berlin Heidelberg, 2009, pp. 94.
- [14] W. Winkler, "The state of record linkage and current research problems," Technical report, Statistical Research Division, U.S. Bureau of the Census, 1999.
- [15] <http://www.w3.org/TR/rdf-sparql-query/> 09.08.2012
- [16] <http://sameas.org/> 09.08.2012
- [17] [http://aemet.linkeddata.es/sparql\\_en.html](http://aemet.linkeddata.es/sparql_en.html) 09.08.2012
- [18] <http://www20.gencat.cat/portal/site/icaen> 09.08.2012
- [19] C. Bizer and R. Cyganiak, "D2R Server – Publishing Relational Databases on the Semantic Web," Poster at the 5th International Semantic Web Conference (ISWC), Springer, Nov. 2006.

## A Semantic Environmental GIS for Solid Waste Management

Miguel Felix Mata Rivera, Roberto Zagal Flores

Computing Mobile Laboratory  
UPIITA- IPN, ESCOM- IPN  
Mexico City

{mmatar@ipn.mx,@zagalmmx@gmail.com}

Diana Castro Frontana, Consuelo Garcia Mendoza

Soil Laboratory, Systems Department  
ENCB-IPN, ESCOM-IPN  
Mexico City

{dgcastro@ipn.mx,varinia400@hotmail.com}

**Abstract**— Nowadays, solid waste handling is a critic problem. Governments and specialists of different disciplines wrestle with environmental problems that poor waste handling generate. For example, in México city the sanitary landfills have been overstepped in their capacity. Then, they are inadequate for the collection and processing of municipal solid waste. We propose as a solution a multi-criteria approach based on semantics, in order to get the adequate place to built any waste handling facility. In this research, a methodology implemented in an environmental GIS system (EGIS) is shown. EGIS identifies and estimates several parameters required for planning or to dimension a waste handling facility (sanitary landfills). The approach proposed involves a multi-criteria solution that includes: environment considerations, administrative parameters, spatial analysis, constraints and Mexican regulations. All of them are combined and processed based on Mexican normative rules. In order to get a management of municipal solid waste (MSW) and a geo-environmental recommendation to locate sanitary landfills in places that comply with official regulations. The results are potential locations for a sanitary landfill site. In addition, information of possible financing sources is given to carry out waste handling projects accordingly. Methodology can be applied to other countries with similar problems regarding to sanitary landfills. Results obtained are better when semantics and multicriteria are combined that when they are used isolated.

**Keywords**-Solid Waste Management; Environmental GIS; Spatial Semantics

### I. INTRODUCTION

Today, the handling of Solid waste generation is a critic problem for Mexican municipalities. The Mexican Secretary for the Environment and Natural Resources (SEMARNAT by its acronym in Spanish) [3] have reported that there are only 82 authorized sanitary landfills for 2400 municipalities in the Mexican Republic. This is obviously insufficient to control the substantial quantity of waste collected. In Mexico, urban solid waste handling is under jurisdiction of municipalities. Solid Waste are those originated in the domestic and commercial activity of cities and towns. The waste produced by urban dwellers include garbage, old furniture and appliances. Packaging and waste from commercial activity, remains the care of the gardens, cleaning the streets.

Mexico is confronted with major problems in the management of solid waste. Owing to rapid industrialization and population growth in urban centers. There are programs at the municipal level for Prevention and Management of Municipal Solid Waste (MSW). They require several stages from analysis and diagnosis, assessment to strategy approval and publication of results. In order to qualify the feasibility to construct a new sanitary landfill that complies with Mexican regulations and norms.

Another problem is that the availability of information and tools for proper management of solid waste is missing. (Regulations, available technologies, cost analyses, viability studies). Among other issues are completely unknown by people involved. Therefore, the public or private sectors rejects projects of this nature by consequence. Several studies have already identified the causes for the inefficient waste handling en Mexico [4]. Main of them are: Poor application of the concepts involved in proper waste management; null knowledge of technical waste handling issues: idle recycling plants owing to high operation costs, sanitary landfills that get full very fast owing bad planning and operation, etc. In Mexico, several institutions grant financial aid for projects related to waste management. Nevertheless, it is required to present a cost-benefit analysis of their project to improve waste handling (municipalities have no people to generate it).

Therefore, it is necessary to develop systems that support each of the aspects. They can provide preventive measures and planning to the future. We proposed the development of an Environmental Geographic Information System (GIS). On it, each municipality manages information on waste handling projects. It is addressed to municipal authorities or environmental professionals. They might be working in waste handling projects. System is available for any other people interested in knowing data regarding waste generation. This information can be processed for any Mexican municipality.

### II. RELATED WORK

An environmental system is a unit whose elements interact together as they continually affect each other. They operate toward a common goal: take care of the environment around us. Landfill Management is a problem that has been treated from long time. A multicriteria approach has been

used in [1], where a system manages recollection, transportation, recovery and disposal activities. But the issues of a landfill construction are not treated. In addition, not semantics processing is used. In [9], a multi-criteria decision analysis for supporting the waste was formulated by integrating interval-parameter, mixed-integer, and chance-constrained programming methods. But is addressed to find a balance between cost and diversion rate of waste management. While in [10] research is focused on the optimum selection of the treatment and disposal facilities, their capacity planning and waste allocation under uncertainty associated with the long-term planning for solid waste management. The difference with our work is the planning; we take into account long, short and medium term.

Other works have treated Solid Waste, but at recollection level like in [4] the work is focused on location of the containers. As well as, the amount estimation of the deposited therein, and routes generation for recollection in municipality of Prat de Llobregat. A similar approach to our work is presented in [2], where Multi-Criteria Decision Making ontology with an inference engine is used. Opposed to our work the paper referenced is focused on distributed autonomous devices. Otherwise, logistic regression model have been used as a methodology in [5] as a part of studies related with MSW. Nevertheless, semantics processing was not used. In our research, spatial semantics is used to analyze data. In a similar way that a person who interprets qualitative variables. This approach has been used in works like [6]. Other related approach where semantics processing is used, can be found in [7, 8]. A survey in [11] reviewed the models of MSW generation and to propose beneficial design options concerning regional sampling and other factors, the final result is a relevance tree for methodology used in the works reviewed.

### III. METHODOLOGY

The methodology used in this work is described in four steps as follows:

- 1) *Define axioms and establish rules using Mexican legislation, specialist criteria and the mexican norm (NOM-083 by ist acronym in spanish)*
- 2) *Design and built ontology based on constraints from NOM-083 and define spatial semantics relations.*
- 3) *Define geographic operations (spatial analysis) to answer generated queries in semantic module*
- 4) *Design and implement Web services and mobile app.*

The methodology is applied into a system composed of four modules: 1) Semantic 2) Geographical, 3) Web, 4) Environment, 5) Mobile. General functionality is described as follows and posteriorly in detail:

**Environment module:** It is a Management Solid Waste (MSW) calculator. It receives data from user and processes them with data from official data sources, taking into account the requirements, constraints and statistical data. Such as: what is minimum distance required to built a MSW from an airport, from a natural protected area, moreover the module indicates if these distances can be relaxed if other studies are

made. As well as topographic and geologic aspects, among others. The module works using a semiautomatic process.

**Semantic module:** Determine the set of geographical operations and queries to find the adequate zones for sanitary landfill construction to any municipality. The data input are processed, filtered and analyzed in conjunction with a set of rules from Mexican norm (NOM-083-SEMARNAT-2003) and environmental constraints. The result is a set of attributive parameters such a as: what type of landfill sanitary should be built (in accordance with population size), the candidate geographic area, and constraints-descriptions of geographic objects within of area selected. Data are re- sent formatted as a geographic query to geographic module.

**Geographic module:** Here, the goal is to make the spatial analysis required to answer the geographic queries received by the semantics module. For example, flooding areas at 10 km to each possible landfill sanitary location. Output is a map with the potential zones to become a landfill sanitary.

**Web and Mobile module:** Displays area suitable for construction of a landfill. On maps interface. In addition, detailed reports with suggestions for financing the construction of an landfill sanitary. The mobile version offers the same functionality. It uses a set of Web services to communicate with the modules. Each module is able to respond itself, however it is necessary, the interaction of all modules, receiving the correct inputs and generating outputs when is necessary. Figure 1 shows the general operation of system.

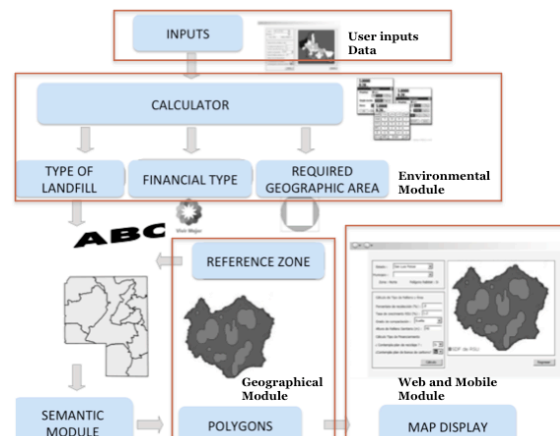


Figure 1. Working diagram ( inputs and outputs of system)

Figure 1 shows the flow process, and what operations belong to each module, letters ABC is the type of landfill. The system processed information from official sources such as Mexican Geography Institute (INEGI). These sources are used to estimate total waste generation, size of a landfill and population growth. In other cases, where official information was not available, data from specialized literature is integrated to the system. For example, the method for designing the size of a landfill for any

municipality was taken from publications by the Health Pan-American Organization [5]. Environment module receives the following input parameters: population size, subproducts suitable of being recycled, possible financing sources, etc. Output is a set of parameters: type of landfill sanitary, sort of financing and required geographic area. These parameters are interpreted by the semantic module (applying axioms) and sent as a geographic query to geographic module. For example, if output is: landfill sanitary of type three. The constraints applied to this sanitary type are retrieved from semantics relations of this concept. It can be the required area, the estimated growth population, estimated percentage of recollection per day. These constraints are formatted as a string array and sent to geographic module to be transformed into a geographical query (in this case, find areas of 20 km in municipality, because this area is required by landfill type 3). Then, geographical module determines which spatial analysis are required to perform in order to satisfy this query. The result is a map showing potential areas to landfill sanitary. The map display is performed via the web and mobile module which is rendered for appropriate visualization on each device. The general architecture, with some technical details describes the functionality; see Figure 2. In Section IV, the semantic processing is described in detail.

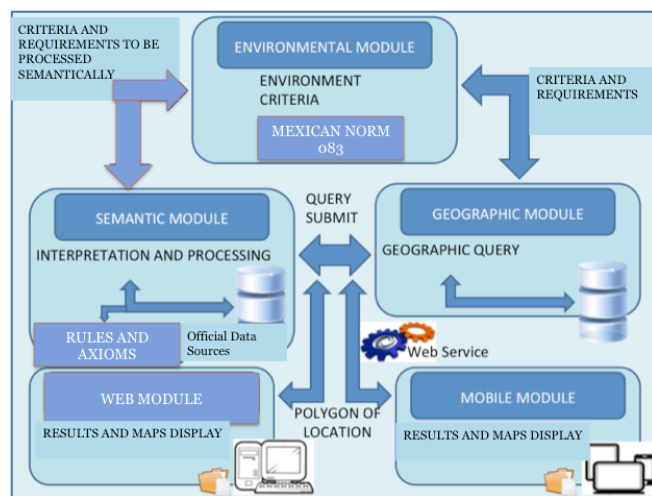


Figure 2. General architecture of system.

Figure 2 shows the data sources and scheme of communications. Data sources are divided into semantic and geographic module, it is due to heterogeneity, data are inserted by XML format. In the case of web module, the communication with semantic module is direct by sending the candidate geographic area (polygon of location). The mobile module is made by using a Web services. The complete process is described in Section IV.

#### IV. SEMANTIC AND GEOENVIRONMENTAL PROCESSING

Semantic processing involves applying several axioms to get the appropriate place to build a landfill sanitary.

Computing the growth population projection, estimation of solid waste generation, among other geographic and environment factors. There are constraints that can be applied to these factors. Some rules and axioms are defined in order to define which of them should be applied. Also, the spatial analysis should be performed. Moreover, in what cases some value factors can be relaxed. In addition, to find financial sources to build a landfill sanitary in municipalities with a small budget. The ontology contains 4 classes and 17 entities. The rules are grouped as follows:

- a) Geographic areas conditioned for a sanitary landfill construction. It means that these areas can be availability if some additional studies and formalities are made;
- b) Geographic areas not allowed. For example, wells at a certain distance, proximity of lakes, etc.,
- c) Minimum and recommended values of distances between geographic objects and possible areas of landfill construction.

Figure 3 shows the ontology with its classes, too, constraints, relations and concepts that allows to make the semantics processing for data inputs.

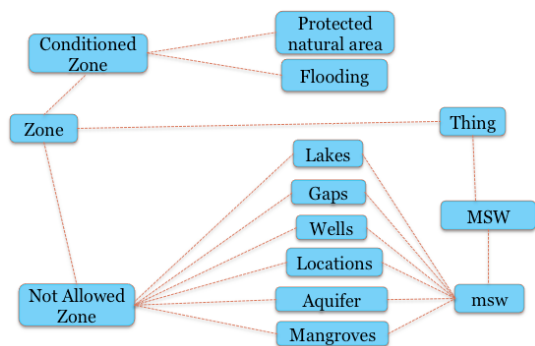


Figure 3. Graph of ontology.

In Figure 3 appears grouped by hypernymy, meronymy semantic relations and the concepts that are classified as not allowed zones. In the same way are grouped to conditioned zones. The following are some examples of axioms used in semantics processing:

**Axiom A:** If municipality has airport, then minimum distance to airport location is 20 km from landfill location. Flexibility: Aviary study. If an aviary study is presented, the minimum distance is reduced to 13 Km.

**Axiom B:** If landfill location is close to flooding then a long term of flooding (50 years).

**Axiom C:** If data total waste generation (twg) > 70 ton/day then the municipality can apply for resources from a Mexican infrastructure fund like PRORESOL (acronym in Spanish).

**Axiom D:** If twg > 100 tonnes/day then a “type A” landfill applies [3,2]. Depending on the geographical zone values of waste composition percentages are applied. Factor of twg is then multiplied by the percentages from this table [3]. The system analyses the input data to define which of them complies with the conditions and requisites requested by financing institutions to get sources of financing. In the case that not match was found the system indicates what studies or analysis are missing in order to get the funds. The

software Protégé [12] was used to build the ontology in similar form like in [6]. The knowledge model was implemented in OWL language. Ontology exploration is made using Hermit reasoner [13] with a OWL API. The semantic processing is performed in three steps: load ontology, runs the reasoning Hermit, relations are explored. The ontology returns a string array to be interpreted as a geographic query by PostgreSQL engine. An example of the output is shown as follows.

ARRAY[

```
[{"distance = 2000", "MinimumDistance = 1000", "exists = 1", "weight = 0.7", "name = Lakes"},
{"distance = 1600", "MinimumDistance = 500", "exists = 0", "weight = 0.7", "name = Pozos"},
{"distance = 0", "exists = 0", "weight = 0.8", "name = Wetlands"},
{"distance = 0", "exists = 0", "weight = 0.6", "name = Soiltype"},
{"distance = 2000", "MinimumDistance = 1000", "exists = 0", "weight = 0.7", "name = Estuaries"},
{"distance = 13000", "exists = 1", "weight = 0.3", "name = Airports"},
{"distance = 2000", "MinimumDistance = 1000", "exists = 1", "weight = 0.7", "name = Lagoons"},
{"distance = 0", "exists = 0", "weight = 0.7", "name = Mangroves"}]
```

As is shown into array, several parameters appear; the 'weight' attribute is computed based on the number of conditions complied and constraints flexibility. The scale used is from value 0 to value 1, where value 0 means that this place not qualify as a candidate for landfill construction.

The attribute 'exists' represents if the municipality has a manage plan and funds to built the sanitary landfill. The value 0 means that no funds are required and value 1 means that is required a financial plan in order to get funds. Each one of these parameters is parsed by geographic module and transformed into a geographic query.

For example, the attributes:

'MinimumDistance = 1000', 'name = Lakes',

are transformed in query:

```
QG1=SELECT ST_Buffer_Meters(the_geom, num_meters) FROM MunicipalitiesTable;
```

The semantic engine consists of a browser, semantic reasoner and query designer. The process is as follows: a polygon (candidate area to a landfill construction) is received from Web module in WKT format (Well-known text) with its location. The reasoner performs an exploration of ontology to find the concepts associated, the output is formatted as a string array. And it is sent as an input parameter to the query designer. It get browser parameters. Parse the strings array and transformed into queries. They are executed in PostgreSQL [14] such as "find locations match conditions". The array syntax is:

```
SELECT * FROM getHollows( ARRAY[ name=zzz, distance=#,
minimumdistance=#, weight=#, exists=#, 'ManagementPlan=#'...] , "polygon");
where:
```

*\_Name* is the name of the layer

*distance*: normal distance in meters between an element of a layer and MSW

*\_Minimum distance*: in meters for a feature and MSW.

(constraint added by the expert)

*\_weight*: is a weight (0 to 1) designated by the expert to identify which place has a greater weight than another.

*\_ManagementPlan*: exists (1) or not (0) a Management Plan.

The result is a list of objects containing geometry in WKT, the area of the polygon, the weight assigned to the layer, the layer name, the name of the polygon and the existence of the management plan. The function getLayersOntology() returns in a record two layers of possible locations: one with the normal distance and the other with the minimum distances (relaxed).

## V. EXPERIMENTS AND RESULTS

The study case for testing the system is with municipalities from San Luis Potosi (SLP) State, from Mexico. The attributes and values belong to these municipalities. A set of attributes is listed with values associated to define the size and type of landfill. Figure 4 shows a map from SLP State. On it is displayed the candidate areas to built a landfill sanitary considering the Mexican regulations and suggestions of specialists.

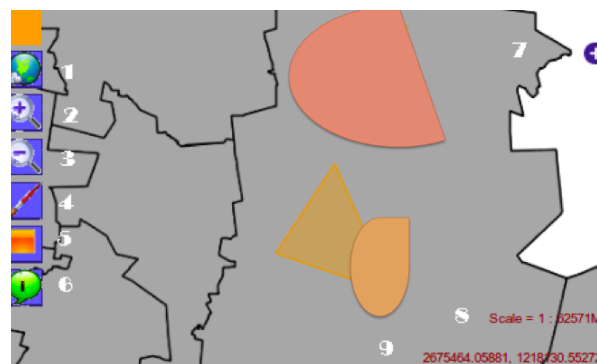


Figure 4. Potencial areas identified to be a landfill sanitary.

Figure 5 shows the areas generated. They represent the potential areas to build a sanitary landfill. Colors indicate the different levels of match. It means that an area match with the requirements from a 70% to 100%.

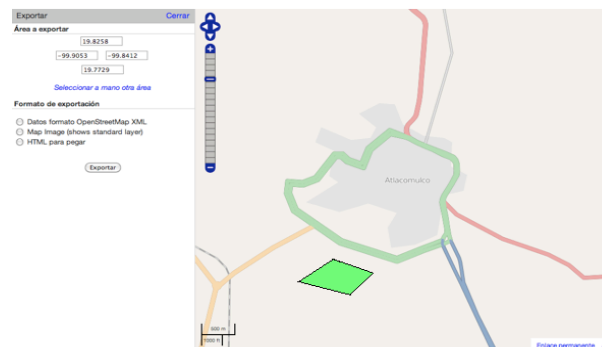


Figure 5. Potential area identified to become a landfill sanitary.

As Figure 5 shown, the descriptive data are obtained associated with the places showed on a map. In this case in green appears a rectangle drawn by user, the idea is to define if this area make with hand tool, is a good candidate to be a sanitary landfill.

Figure 6 shows the mobile interface of system, the initial screen is a formulary to get the input data: state, municipality, period of term (years of study), etc. Other data are completed automatically by the app.

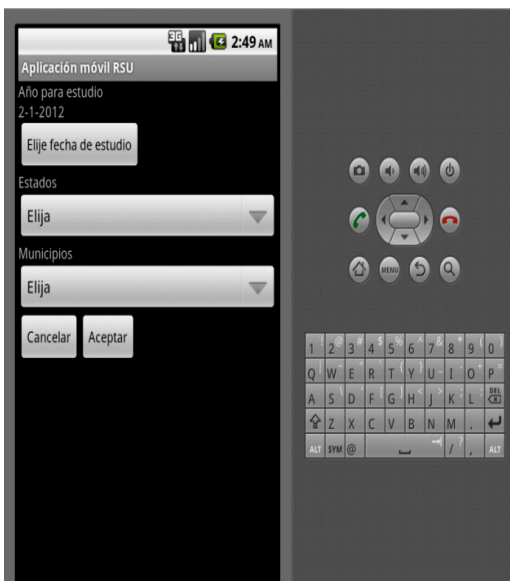


Figure 6. Input data into app in smartphone.

The process is started when data is entered and a connection with geographic module by web service is established. A mapping application mobile shown the result. See Figure 7.



Figure 7. Mobile version of calculator.

The green polygon in Figure 7 shows in this case the candidate area location to build a sanitary landfill in accordance with the user input's data.

## VI. CONCLUSION AND FUTURE WORK

In this approach, semantics processing is used to solve the cases when a sanitary landfill will be constructed in an

area where some constraints are applied. As an example when a sanitary landfill will be near an airport. The Mexican norm establishes a minimum distance of 20km. In this case, the constraint is processed semantically and is obtained that if is performed an aviary study. Then, the distance can be reduced to 13 km. In this scenery without semantics processing will not be possible to get this result.

In addition, is shown how the semantics processing is combined with a multicriteria approach. In order to find a point where constraints can be flexible and comply with the rule. The methodology can be applied to other similar sceneries.

The multicriteria approach is based on weight schemes in order to get relevance for the geographic candidate areas. Based on it, is possible to offer a list of candidates for sanitary landfill construction.

This research is part of a multidisciplinary project. The main contribution is to use a multi-criteria approach combined with semantics. Here is described the overall functionality.

Results are compared with the results obtained without semantics processing. For constraints related with distances. The results are better using semantics processing. Because other areas are found as candidates, while that using only multicriteria approach these areas do not appear as candidates. At this moment, a panel of users is testing the system. They have different levels of expertise (environmental engineers, computer engineers, geographers and students.) The future work considers to include all the constraints for Mexican norm to be processed using the approach mentioned.

## ACKNOWLEDGMENTS

The authors want to thank COFAA, EDD, and SIP-IPN for its support. This Project is part of a multidisciplinary project with numbers: 20113429 and 20120504.

## REFERENCES

- [1] Bazzani, Guido Maria, "Integrated solid waste management: a multicriteria approach". Sixth Joint Conference on Food, Agriculture and the environment. University of Minnesota, 1998, <http://ageconsearch.umn.edu/bitstream/14492/1/c6bazz01.pdf> [retrieved: July, 2012].
- [2] Ghadi Mahmoudi and Christian Müller-Schloer, "Semantic multi-criteria decision making SeMCDM", IEEE Symposium Series on Computational Intelligence, Symposium on Computational Intelligence in Multicriteria Decision Making, 2009, Nashville, USA
- [3] Official Mexican Norm, NOM-083-Semarnat-2003 <http://www.profepa.gob.mx/innovaportal/.../nom-083-semarnat-2003.pdf> [retrieved: July, 2012].
- [4] Sanz Conde and M. Mercedes, Universitat Politecnica de Catalunya. Departament d'Enginyeria del Terreny, Cartografica,

- <http://upcommons.upc.edu/handle/2099.1/7293>. [retrieved: May, 2012].
- [5] W. Perez and A. Tamayo, "El uso de las aplicaciones SIG para el manejo y tratamiento de residuos sólidos domiciliarios". <http://www.oterra.cl/descargables.html> [retrieved: March, 2012].
- [6] M. Moreno-Ibarra, "Semantic similarity applied to generalization of geospatial data". *Lecture Notes in Computer Science*. Vol. 4853. 2007. pp. 247-255
- [7] R. Quintero, *Representación semántica de datos espaciales raster*. Tesis Doctoral. Instituto Politécnico Nacional. 2007. pp. 77-108.
- [8] R. Quintero, M. Torres, M. Moreno, and G. Guzmán, "Metodología para generar una representación semántica de datos raster". T. Delgado, J. Capote (editors). *Semántica espacial y descubrimiento de conocimiento para desarrollo sostenible*. Ed. CUJAE. La Habana. 2009. pp. 119-145.
- [9] B. D. Xi, J. Su, G. H. Huang, X. S. Qin, Y.H Jiang, S. L. Huo, D. F. Ji, and B. Yao. 2010. Brief paper: "An integrated optimization approach and multi-criteria decision analysis for supporting the waste-management system of the city of Beijing, China". *Journal Engineering Applications of Artificial Intelligence*. Volume 23 Issue 4, June, 2010 pp. 620-631
- [10] Amitabh Kumar Srivastava and Arvind K. Nema, "Fuzzy parametric programming model for multi-objective integrated solid waste management under uncertainty". *Journal Expert Systems with Applications: An International Journal* aVolume 39 Issue 5, April, 2012 pp. 4657-4678
- [11] P. Beigl, S. Lebersorger and S. Salhofer, "Modelling of municipal solid waste generation: A review". *Waste Management*. 2008, v28 i1. 200-214
- [12] <http://protege.stanford.edu/> [retrieved: August, 2012].
- [13] <http://www.hermit-reasoner.com> [retrieved: August, 2012].
- [14] <http://www.postgresql.org/> [retrieved: July, 2012].

# Comparing a Rule-Based and a Machine Learning Approach for Semantic Analysis

Francois-Xavier Desmarais

École Polytechnique  
de Montréal

Montréal, PQ, Canada

Francois-Xavier.Desmarais@polymtl.ca

Michel Gagnon

École Polytechnique  
de Montréal

Montreal, PQ, Canada

Michel.gagnon@polymtl.ca

Amal Zouaq

Royal Military College of Canada  
Department of Mathematics and Computer  
Science

Kingston, ON, Canada

Amal.Zouaq@rmc.ca

**Abstract**—Semantic analysis is a very important part of natural language processing that often relies on statistical models and machine learning approaches. However, these approaches require resources that are costly to acquire. This paper describes our experiments to compare Anasem, a Prolog rule-based semantic analyzer, with the best system of the Conference on Natural Language Learning (CoNLL) shared task dedicated to a sub-task of semantic analysis: Semantic Role Labeling. Both CoNLL best system and Anasem are based on a dependency grammar, but the major difference is how the two systems extract their semantic structures (rules versus machine learning). Our results show that a rule-based approach might still be a promising solution able to compete with a machine learning system under certain conditions.

**Keywords**—*Semantic role labeling; evaluation; rule-based systems; machine learning.*

## I. INTRODUCTION

One of the most challenging tasks of natural language processing is semantic analysis (SA), which aims at discovering semantic structures in texts. Two schools of thought try to tackle this hard task:

**The Computational Semantics approach:** semantic analysis is often built on top of grammars describing lexical items through feature structures [3]. The aim is to extract a logical representation such as first-order logic and discourse representation structures (DRS). These types of grammars are often hard to build and maintain but they offer a wide coverage of various linguistic phenomena (e.g., co-reference resolution, negations, and long-distance dependencies). To our knowledge, such a wide coverage is only handled through this type of systems. Another problem is that very few if any datasets enable the comparison of these types of systems.

**The Machine Learning approach:** semantic analysis is decomposed into various tasks such as semantic role labeling (SRL) [4], co-reference resolution [11] and named entity extraction [1]. While machine learning, especially supervised approaches, proved to be successful in some of these tasks, it suffers from well-known shortcomings: Firstly, the algorithms depend highly on the availability of training corpora, which take a lot of resources to be developed. Secondly, the learned models often do not scale well on different datasets and domains, thus necessitating other training corpora.

The two aforementioned approaches involve a non-negligible effort either in terms of software development (computational semantics) or in terms of data availability (machine learning). The two obvious questions are whether one of these approaches is less costly than the other and whether one is more successful than the other. Trying to address the second aspect, this paper aims at providing insights on the following research question: **Can a rule-based semantic analyzer reach the same performance of a machine learning one?**

Despite the fact that machine learning systems have become prominent in some tasks such as syntactic parsing, there is no clear evidence that they are more efficient and less costly for the semantic analysis task.

In this paper, we compare two systems on a specific sub-task of semantic analysis, which is the identification of predicates and their arguments:

- *ANASEM*, our Prolog-based semantic analysis system which outputs Discourse Representations Structures based on dependency grammar patterns. We consider that ANASEM falls within the computational semantics approach;
- The *LTH Parser* [4], which is the winner of CoNLL (The Twelfth Conference on Computational Natural Language Learning) shared task, dedicated to SRL. This system is based on dependency parsing and machine learning.

These two systems are run on a subset of the CoNLL gold standard [12]. The paper explains in detail how we handled this comparison.

The paper is organized as follows. Section 2 provides a brief overview of the state of the art in semantic analysis. Section 3 is a description of our rule-based semantic analyzer Anasem. Section 4 presents the core of our methodology by explaining the adaptations we had to perform on our system to compare our results with CoNLL winner. Section 5 details the results obtained by Anasem. Finally, section 6 discusses the limitations of our approach.

## II. STATE OF THE ART

As aforementioned, the term “semantic analysis” might take various meanings depending on the targeted community. In this paper, we consider SA as the process of extracting predicates and arguments. There have been considerable efforts these last years in areas such as semantic role labeling [12], dependency-based representations [14]



and machine learning [4] to extract these semantic structures. Two main approaches are pervasive to state-of-the-art Natural Language Processing systems: statistical and machine learning techniques and rule-based techniques. Syntactic analysis seems to have evolved essentially towards statistical parsers [8]. However, rule-based approaches have proven successful in others tasks. For example, the best-performing system at the CoNLL 2011 shared task for coreference resolution [6] is a rule-based system. Similarly, in semantic analysis, the STEP 2008 shared task [2] reported on various systems among which Boxer [3] used a categorical grammar approach. A formal comparison of these systems, using a gold standard, is missing. To our knowledge, CoNLL 2008 Shared task is among the very few which offer such a gold standard. The participants at this competition were essentially machine learning systems including the first-ranked system, the LTH Parser [4], which relied on dependency analysis and classifiers for SRL. In this paper, our objective is to compare the performance of ANASEM with the LTH parser. To our knowledge, there was no previous tentative in recent semantic analyzers to compare a symbolic rule-based approach to a machine learning approach on the same corpus.

### III. ANASEM, A PROLOG-BASED SEMANTIC ANALYZER

Anasem [14] is a rule-based system written in Prolog and built on a modular pipeline made of 3 functionalities: syntactic parsing, canonical tree generation and pattern recognition.

#### A. Syntactic Analysis

The syntactic analysis is the first step in the pipeline, and like [4] it is based on dependency parsing. Anasem uses the Stanford parser [5], its dependency module [7] and its part-of-speech tagger [13] to perform the syntactic analysis. For instance, the sentence *They drank brandy in the lounge* returns the following result, where part-of-speech tags and dependencies are given (note that each word is given with its position in the sentence.)

Part of speech:

```
They/PRP drank/VBD brandy/NN in/IN the/DT
lounge/NN ./.
```

Syntactic analysis:

```
nsubj(drunk-2, They-1)
dobj(drunk-2, brandy-3)
prep(drunk-2, in-4)
det(lounge-6, the-5)
pobj(in-4, lounge-6)
```

#### B. Canonical Tree generation

The second step of the pipeline is to generate a canonical tree from the syntactic analysis to facilitate the subsequent step of pattern recognition. The dependency parse is coupled with parts-of-speech to create a Prolog term. This Prolog term represents a unified structure that can be processed recursively based on the principle of compositionality. Using our previous example, we obtain the following representation:

```
root/tree(token(drunk, 2)/v,
  [nsubj/tree(token(they, 1)/prp, []),
```

```
  dobj/tree(token(brandy, 3)/n, []),
  prep/tree(token(in, 4)/prep,
    [pobj/tree(token(lounge, 6)/n,
      [det/tree(token(the, 5)/d, [])]))]])
```

A final step is to modify the generated tree to facilitate patterns identification. Some important modifications are related to coordination and negation. Dependencies involving a coordinated form are duplicated and attached to every member of the coordination. For example, the parse tree for the sentence *John visited Paris and Roma* would be translated into a tree that corresponds to the sentence *John visited Paris and John visited Roma*. Another important transformation achieved at this step concerns negation. Instead of being dependent of the main verb of the clause, the negation is moved at the root of the clause.

#### C. Pattern recognition

Anasem pattern represents a syntactic rule that can be mapped to a semantic representation. Anasem contains about 60 patterns. Each part of the Prolog tree is analyzed in a recursive manner, thus implementing a pattern hierarchy (based on the rules appearance in Prolog). The output is a discourse representation structure [14]. Using the previous example, we obtain the following DRS:

```
-----
[id1,id2,e1,id3]
-----
entity(id1,they)
entity(id2,brandy)
event(e1,drank,id1,id2)
entity(id3,lounge)
in(e1,id3)
-----
```

This DRS introduces three entities and one event. Event e1 is a drinking event that involves two entities (the brandy and the persons who are drinking it). The DRS also expresses a relation between the event and its location (the lounge).

### IV. METHODOLOGY

This section describes the methodology followed to compare Anasem with LTH Parser.

#### A. CoNLL Corpus and Terminology Description

CoNLL Shared Task provided a corpus based on a subset of the Penn Treebank II [7] [12]. Two types of corpora were made available: a training corpus that contained a structured output with parts of speech, syntactic analysis and semantic representations, and a test corpus. In our case, the major problem with these corpora is the lack of compatibility with Anasem's output format (DRS, grammatical relationships, grammatical categories and semantic categories). Table I shows the subset of CoNLL format that was used in our adaptations. Each term is related to a part-of-speech, the position of its head in the sentence (value is 0 for the root of the sentence), and a grammatical relationship. The semantic representation starts with the semantic predicate (and its frame in PropBank [10] and NomBank [9]). Finally the last columns indicate semantic

arguments in the form of semantic categories labeled A0, A1, AM-TMP, etc. For every predicate there is a corresponding column, in the same order. For example, in Table I, predicate *happen.01* has arguments A1 and AM-TMP that correspond to *accident* and *as*, respectively.

TABLE I. AN EXAMPLE OF CONLL FORMAT

1	The	DT	2	NMOD	-	-	-
2	accident	NN	3	SBJ	-	A1	-
3	happened	VBD	0	ROOT	happen.01	-	-
4	as	IN	3	TMP	-	AM-TMP	-
5	the	DT	6	NMOD	-	-	-
6	night	NN	7	SBJ	-	-	A1
7	was	VBD	4	SUB	-	-	-
8	falling	VBG	7	VC	fall.01	-	-
9	.	.	3	P	-	-	-

B. Anasem Adaptation to CoNLL

Given the different terminology adopted by CoNLL, we had to modify two major modules of Anasem, namely the canonical tree generator and the semantic patterns that were using the Stanford nomenclature.

1) The Canonical Tree Generator

As aforementioned, Anasem uses the Stanford parser [5] to generate the canonical trees. To exploit our patterns, we had to keep Anasem's canonical tree representation while using CoNLL lexico-syntactic representations. These representations were available in the shared task corpora [12] designated hereafter as the gold standard (GS). We extracted the syntactic relations, the parts of speech and the head of each word from the GS (see Table II) and replaced Anasem's dependency relationships and parts of speech.

The sentence *The accident happened as the night was falling* was transformed into the trees illustrated in Table II.

As can be noticed, although there were similarities between the initial tree and the obtained tree, there were also some major differences. For example, some root nodes changed as shown by comparing the node *advcl falling/v* in our initial tree with the node *tmp as/prep* in the obtained tree. These differences can be explained by the fact that the Stanford parser and CoNLL have different syntactic representations. For example, the "auxiliary" is represented by the Stanford Parser with its head as the verb and the syntactic relation as "aux", while in CoNLL, the auxiliary is the head and its syntactic relation is called a "verb chain" (vc).

TABLE II. CANONICAL TREE TRANSFORMATION

Initial Canonical Tree
root happened/v
nsubj accident/n
det the/d
advcl falling/v
mark as/prep
nsubj night/n
det the/d
aux was/v
Tree using CoNLL Terminology

root happened/v
sbj accident/n
nmod the/d
tmp as/prep
sub was/v
sbj night/n
nmod the/d
vc falling/v
Modified Canonical Tree
root happened/v
sbj accident/n
nmod the/d
tmp falling/v
sbj night/n
nmod the/d
aux was/v
complm as/prep

We classified these differences into two major categories:

- a. Structural differences, which happen when word positions and heads inside the tree are different.
- b. Nominal differences, which happen when the terminology of the grammatical relations is different but the meaning is the same.

We had to adapt the canonical tree generator to deal with these differences. Most problems caused by structural differences were solved by creating a set of rules that we applied to the canonical tree (for instance AUX in our previous example). Nominal differences were then resolved by providing a mapping between Stanford grammatical relations and CoNLL relations and updating Anasem patterns accordingly.

2) Patterns Adaptation

Almost all the patterns had to be adapted to use the CoNLL grammatical terminology. Many patterns needed a nomenclature modification, for example the noun subject tagged *NSUBJ* in Stanford had to be renamed to *SBJ* to match CoNLL terminology. There were few exceptions such as the negative form which did not necessitate a change. Apart from the terminological changes, we experienced some mapping problems due to differences in granularities between the Stanford grammatical relationships hierarchy (which seems more fine-grained) and the CoNLL one. The grammatical relationship *NMOD* is a good example of this problem. For example, in CoNLL, determiner and adjectives are both classified as a *NMOD*, while Stanford has specific categories (*DET*, *AMOD*). In this case, we were able to perform mappings with the Stanford hierarchy by using parts of speech to differentiate the various possibilities.

Certain patterns were not used because their grammatical relations were not identified in CoNLL. For example, clausal complements, represented by the *CCOMP* relation in Stanford parser, are interpreted as generic complements in CoNLL.

Although there were many differences between Anasem's output (DRS) and CoNLL's output (Table I), attempts were made to automate the process, but they were unsuccessful due to too many special cases. Therefore, we had to select a subset of the original corpora, manually identify the

mappings and finally check the obtained tree transformations before being able to parse the obtained trees using the pattern recognition module.

### C. Corpus Selection and Comparison Methodology

We selected a subset of sentences from the original corpora to analyze them with our system and compare the results with LTH Parser. We established few rules to avoid some specific problems with few characters, which are not processed by the current version of Anasem, and to deal with the way CoNLL handle hyphenated words. These rules are as follow:

- A sentence must not contain the following characters (-&\$%()\_:\) as Anasem is not robust in front of this type of input.
- A sentence must not contain hyphenated words: this rule was due to the way CoNLL processes these words. The GS separates the words and the hyphen and considers each word independently while Anasem considers them as a single entity.
- A sentence must have between 5 and 30 words.
- A sentence must have at least 1 verb.

In particular, the last two rules were used to focus on the most representative and declarative sentences (e.g., with at least one verb). For instance, a sentence such as : “*At law school, the same*” was excluded from our evaluation.

Using these filtering rules, we extracted sentences from the CoNLL training corpus (dev set). Then, to avoid any bias, we randomized all the sentences with “www.random.org” and extracted the first 51. These sentences (dev set) were used to compare Anasem results with the gold standard semantic representations. We repeated the same process on the test corpus to extract a set of sentences to be used for a fair comparison between Anasem results and LTH Parser, which was trained on CoNLL training corpus. 50 sentences were extracted from the test corpus, with an overall of 101 sentences.

To be able to compare Anasem with CoNLL semantic representation, we had to establish a comparison methodology. Due to the differences between Anasem semantic representation (DRS containing entities, events, attributes, etc.), this comparison was essentially based on the unlabeled extraction of the semantic representations.

We ran Anasem on the test and dev corpora and the LTH Parser on the test corpus.

To evaluate ANASEM DRS, we extracted the predicates from the gold standard and we verified if these predicates were available in the generated DRS either as an `entity` or an `event`. Then for each of these predicates, we compared the arguments indicated in the gold standard with those in the DRS to correlate them with either an `event` argument (e.g., `event (e2, falling, id2)`) or a `complement` argument (e.g., `time (e1, e2)` or `in (id2, id3)`). This allowed us to compute recall for Anasem.

For the comparison between ANASEM and LTH Parser, we used the evaluation tool of the CoNLL competition

shared task. Since we aimed mainly at the identification of predicates and arguments (without providing a label), we slightly modified this tool to display the presentation of unlabelled arguments and predicates.

## V. RESULTS

In this section, we present the results of our experiments.

### A. Sentence-based results

We categorized our results into the following:

**All sentences:** These results include both fully and partially covered sentences on both corpora (development and test). Since we do not cover all the possible patterns yet in Anasem, it was anticipated that some sentences would be partially analyzed.

**Sentences with full predicate coverage:** These results describe the performance of Anasem over sentences that were successfully parsed at the syntactic level and whose predicates have all been discovered at the semantic analysis level. These sentences are qualified as fully covered sentences in our results.

All the results based on the GS are reported as unlabeled recall values only. In fact, even though we consider CoNLL corpus as a GS, our analysis is that this GS is not very well adapted for a fair computation of Anasem precision. In fact, Anasem covers semantic relationships that are not available in the GS but that are still valid. One case is that Anasem extracts all attributes when CoNLL GS seems to neglect some. For example in the phrase “...*his sweaty armpits*...”, CoNLL considers only “*his*” as an attribute of “*armpits*” (A1) while Anasem identifies also “*sweaty*” as an attribute, which seems reasonable. Another case is that CoNLL restricts the analysis to predicates with arguments from PropBank [10] and NomBank [9], and cannot identify predicates without arguments (such as Shipyard(X)), which might be criticized in the context of a global semantic analysis. Computing precision over this GS would affect negatively the precision of Anasem. However, to give the reader an idea about the performance of our system, we manually computed the precision over the DRS extracted from the test corpus (see Table III).

TABLE III. PRECISION RESULTS FOR ANASEM

Predicates	92 %
Arguments	80 %
Predicates and arguments	86 %

### All sentences

These results are obtained on the 101 sentences from the test and development corpora. Each sentence is broken down into predicates and arguments to allow for a more focused analysis of the results. Among these 101 sentences, there were 53 fully covered sentences, 37 partially covered sentences and 11 empty outputs. These outputs are due to failure at the semantic or syntactic level (e.g., an unknown pattern or an illegal Prolog term generation) with a total of

229 predicates and 404 arguments. For the same set of sentences, the gold standard indicated 298 predicates and 672 arguments. Overall, we obtained a recall of 77 % for the identification of predicates and 61 % for the arguments.

### Sentences with full predicate coverage

As mentioned previously, we tagged sentences with missing predicates as partially analyzed sentences. The missing items often resulted from some unimplemented pattern in our Prolog-based semantic analyzer. Thus, we decided to present the results of fully covered sentences separately, as we wanted to evaluate our success on identified patterns. There were 53 fully covered sentences in our corpus with 337 possible arguments among which we identified 271 arguments. This resulted in a significant improvement of 19% over the previous results with a recall of 80%.

Table IV summarizes the results obtained for the extraction of unlabeled predicates and arguments.

TABLE IV. RECALL VALUES FOR THE DETECTION OF PREDICATES AND THEIR ARGUMENTS (UNLABELED).

	Predicates	Arguments
All sentences	77%	61%
Fully covered sentences	100%	80%

We can notice that the recall for predicates and arguments identification is much better on fully covered sentences.

### B. Arguments recognition

These results focus on arguments rather than on the predicates. The objective is to assess the success of Anasem in correctly extracting arguments when the predicate is identified.

Table V shows the recall values for argument detection when we consider the correctly identified predicates, independently of the sentences (All detected predicates). In the 50 sentences corpus, 94 predicates were found. In the gold standard, there were 218 arguments associated to these predicates, from which 162 were found by Anasem. That is a 74 % rate of success.

TABLE V. RECALL VALUES FOR ARGUMENT DETECTION IF WE CONSIDER ONLY CASES WHERE PREDICATES WERE DETECTED.

	Arguments
All detected predicates	74%
Predicates detected in fully covered sentences	78%

Taking only the predicates that were detected in fully covered sentences, the figures are slightly better (106 out of 136 arguments), that is, a 78 % rate of success. This shows that choosing arguments from partially analyzed sentences tends to give worst results than with fully-covered sentences.

### C. Comparison with CoNLL Shared Task Best System

We also executed LTH Parser on the 50 sentences from the test corpus. LTH results were as follow for the unlabeled predicates and arguments: 136 predicates out of 142 and 250 arguments out of the 314 that were in the gold standard. This represents a 95 % recall rate for the predicates and 80 % recall value for the identification of the arguments (see Table VI). Results are significantly lower with Anasem especially for argument detection. However, we see that by considering only completely parsed sentences, the difference seems to vanish.

TABLE VI. COMPARISON OF RECALL VALUES IN % FOR ANASEM AND LTH.

	Predicates	Arguments
Anasem (All sentences)	72%	57%
Anasem (fully covered sentences)	-	79%
LTH	95%	80%

Since Anasem makes a distinction between core arguments (that correspond to A0, A1...A4 in CoNLL) and modifier arguments (all other arguments in CoNLL), we were interested to see whether the recall values are the same for both categories. Table VII shows the results, including LTH evaluation on the test corpus. Interestingly, we note that the difference between core and modifiers is greater in LTH Parser than in Anasem.

TABLE VII. RECALL VALUES FOR ARGUMENT DETECTION, DISTINGUISHING CORE ARGUMENTS AND MODIFIER ARGUMENTS.

	Core	Modifiers
All sentences results on test corpus - Anasem	57%	53%
Fully covered sentences results on test corpus - Anasem	81%	76%
Results on test corpus - LTH	84%	69%

## VI. DISCUSSION

### A. General Observations

There are a few things that we need to clarify regarding our results. In the CoNLL shared task, the participating systems were based on external lexicons, namely PropBank [10] and NomBank [9] to identify the arguments types. These lexicons identify arguments based on a word considered as a predicate. Each word-predicate is related to a frame that assigns generic categories such as A0 or A1 to the arguments.

The peculiarity of Anasem is that it is a standalone system, which does not rely on any external resource to identify predicates and arguments. Only dependency parses are used, which means that Anasem is able to extract these predicates and arguments but cannot annotate them with particular categories or types. Therefore, we relied on the

unlabeled extraction task of the competition and compared unlabeled results. This means that whenever we compare our output with the gold standard, we only compare the presence of the arguments and the predicates, regardless of the type of the arguments.

The results for all the sentences in the combined corpora were not outstanding (77 % for the predicates and 61 % for the arguments). On the one hand, this was not really surprising, considering the limited amount of patterns implemented in Anasem. On the other hand, with the fully covered sentences the results rivaled the winner's of the shared task of CoNLL (79% versus 80% argument wise). Our conclusion is that Anasem, though not complete yet, has a good potential to perform as well as a machine learning approach, while staying independent from a training corpus and domain. These results will have to be confirmed in future experiments on bigger corpora.

### B. Limitations

There are limitations in our experiments. To begin with, we used a sub-corpus of the original corpus of CoNLL, and it was a rather small part of it with only 101 sentences. This small number was largely due to the complexity of comparing our adapted output to the CoNLL format and to the time-consuming effort required to make the manual comparison. This manual comparison might have been biased due to possible potential errors by the expert performing the comparison. However, a clear methodology for comparing the representations was established upfront and was closely followed.

We are conscious that our limited set of sentences might affect positively our results if well-analyzed sentences are selected. However, these sentences were randomly selected as previously explained. Moreover, the opposite phenomenon might occur and amplify analysis errors if the wrong sentences (i.e., not well-analyzed sentences) are selected from the gold standard.

Another important point is that there are a number of errors that we identified in the CoNLL gold standard, and in general these errors affected negatively our experiments, and probably more strongly than if we had thousands of sentences. This is the main reason of dividing the results into fully covered and partially covered sentences.

## VII. CONCLUSION

This paper presented the results of a rule-based system for semantic analysis and compared it to the winner of the CoNLL shared task [12]. It shows that using a modular system with syntactic analysis based on dependency grammar can have comparable results with a machine-learning based analysis when the sentences are fully covered. It also demonstrates the difficulty of comparing systems based on various formalisms and lexicons. In future work, we plan to add new rules to our pattern recognition analyzer and to repeat the same types of experiments using a wider range of sentences. We also want to find a way to

measure the precision and the F1 scores. Our preliminary conclusion is that semantic analysis with ruled-based systems has its place among statistical and machine-learning approaches.

### ACKNOWLEDGMENTS

The authors would like to thank the Royal Military College of Canada (RMC) for its financial support through the Start-up Fund Program.

### REFERENCES

- 1] Baluja, S., Mittal, V. O. and Sukthankar, R., 2000. Applying Machine Learning for High-Performance Named-Entity Extraction.. *Computational Intelligence*, 16(4), pp. 586-595.
- 2] Bos, J., 2008. Introduction to the shared task on Comparing Semantic Representations. In: *Proceedings of the 2008 Conference on Semantics in Text Processing*. Venice: ACL, pp. 257-261.
- 3] Bos, J., 2008. Wide-Coverage Semantic Analysis with Boxer. In: J. Bos and R. Delmonte, eds. *Semantics in Text Processing. STEP 2008 Conference Proceedings*:College Publications., pp. 277-286.
- 4] Johansson, R. and Nugues, P., 2008. *Dependency-based syntactic-semantic analysis with PropBank and NomBank*. ACL, pp. 183--187.
- 5] Klein, D. and Manning, C. D., 2003. *Accurate Unlexicalized Parsing*. s.l., s.n., pp. 423-430.
- 6] Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M. and Jurafsky, D., 2011. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In: *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*:ACL, pp. 28-34.
- 7] Marcus, M., Santorini, B. and Marcinkiewicz, M. A., 1993. Building a large annotated corpus of English: the Penn Treebank.. *Computation Linguistics*, p. 19(2).
- 8] Marneffe, M.-C. d., MacCartney, B. and Manning, C. D., 2006. *Generating Typed Dependency Parses from Phrase Structure Parses*. s.l., s.n.
- 9] Meyers, A. et al., 2004. *The NomBank Project: An Interim Report*. Boston, ACL, pp. 24-31.
- 10] Palmer, M., Gildea, D. and Kingsbury, P., 2005. The Proposition Bank: A Corpus Annotated with Semantic Roles. *Computational Linguistics Journal*, p. 31:1.
- 11] Soon, W. M., Ng, H. T. and Lim, D. C. Y., 2001. A Machine Learning Approach to Coreference Resolution of Noun Phrases. *Computational Linguistics*, 27(4), pp. 521-544.
- 12] Surdeanu, M., Johansson, R., Meyers, A., Marquez, L. and Nivre, J., 2008. *The CoNLL-2008 Shared Task on Joint Parsing of Syntactic and Semantic Dependencies*. s.l., s.n.
- 13] Toutanova, K., Klein, D., Manning, C. and Singer, Y., 2003. *Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network*. s.l., s.n., pp. 252-259.
- 14] Zouaq, A., Gagnon, M. and Ozell, B., 2010. *Semantic Analysis using Dependency-based Grammars*. s.l., Bahri Publications, pp. 85-101.

# Hyponym Extraction from the Web based on Property Inheritance of Text and Image Features

Shun Hattori

*College of Information and Systems*

*Muroran Institute of Technology*

*27-1 Mizumoto-cho, Muroran, Hokkaido 050-8585, Japan*

*Email: hattori@csse.muroran-it.ac.jp*

**Abstract**—Concept hierarchy knowledge, such as hyponymy and meronymy, is very important for various Natural Language Processing systems. While WordNet and Wikipedia are being manually constructed and maintained as lexical ontologies, many researchers have tackled how to extract concept hierarchies from very large corpora of text documents, such as the Web, not manually, but automatically. However, their methods are mostly based on lexico-syntactic patterns as not necessary but sufficient conditions of hyponymy and meronymy, so they can achieve high precision but low recall when using stricter patterns or they can achieve high recall but low precision when using looser patterns. Therefore, we need necessary conditions of hyponymy and meronymy to achieve high recall and not low precision. The previous papers have assumed “Property Inheritance” from a target concept to its hyponyms and/or “Property Aggregation” from its hyponyms to the target concept to be necessary and sufficient conditions of hyponymy, and proposed several methods to extract hyponymy relations from the Web, based on property inheritance and/or property aggregation of text features such as meronyms and behavior. This paper proposes a method to acquire hyponymy relations from the Web, based on property inheritance of not only text features, but also image features for each conceptual word.

**Keywords**—*hyponymy; meronymy; concept hierarchy; Web mining; image analysis; property inheritance; typical image.*

## I. INTRODUCTION

Concept hierarchies, such as hyponymy (is-a) and meronymy (has-a) relations, are very fundamental for various Natural Language Processing (NLP) systems. For example, query expansion in information retrieval [1–4] or image retrieval [5], question answering [6], machine translation, object information extraction by text mining [7], Sense-based Object-name Search (SOS) [8], etc. Our appearance information extraction [7] is based on the heuristics that an appearance description about a target object-name (e.g., “kingfisher”) often has a pair of an appearance descriptor and its hypernym (e.g., “blue bird” and “beautiful bird”) or its meronym (e.g., “blue wings” and “long beak”).

While WordNet [9] and Wikipedia [10] are being manually constructed and maintained as lexical ontologies at the cost of much time and effort, many researchers have tackled how to extract concept hierarchies from very large corpora of text documents, such as the Web, not manu-

ally, but automatically [11–14]. However, their methods are mostly based on lexico-syntactic patterns as sufficient but not necessary conditions of concept hierarchies. Therefore, they can achieve high precision but low recall when using stricter patterns (e.g., “ $x$  such as  $y$ ” and “ $y$  is a kind of  $x$ ”) or they can achieve high recall but low precision when using looser patterns (e.g., “ $y$  is a/an  $x$ ”).

To achieve high recall and not low precision, our previous works [15–18] have assumed “Property Inheritance” from a target concept to its hyponyms (i.e., subordinate concepts for the target concept) and/or “Property Aggregation” from its hyponyms to the target concept to be necessary and sufficient conditions of hyponymy, and proposed several methods to extract hyponymy relations from the Web by text mining techniques, based on property inheritance and/or property aggregation of text features such as meronyms and behavior-words. The former assumption is to utilize the other semantic relations surrounding the subordinate (hyponymy) relation between a target concept and its hyponym candidate, i.e., superordinate relationships (hypernymy) and coordinate relationships (including synonymy and antonymy), and to improve a weighting of hyponymy extraction by using multiple property inheritances not only from the target concept to its hyponym candidate, but also between the other pairs of concepts (e.g., from a hypernym of the target concept to its hyponym candidate and/or from the target concept to a coordinate concept of its hyponym candidate). The latter assumption is to improve a weighting of property extraction by using property aggregation to each target concept from its typical hyponyms.

To make our previous method more robust, this paper utilizes not only Web text, but also Web images, and proposes a method to acquire hyponymy relations from the Web, based on property inheritance of not only text features, but also image features for each conceptual word.

The remainder of the paper is organized as follows. Section II proposes a method to extract hyponymy relations from the Web, based on property inheritance of not only text features, but also image features. Section III shows some experimental results to validate the proposed method. Finally, we conclude this paper in Section IV.

## II. METHOD

This section introduces our previously published basic method [15] to extract hyponymy relations from the Web by using not only lexico-syntactic patterns with a target word and its hyponym candidate as sufficient but not necessary conditions of hyponymy, but also “Property Inheritance” (of text features such as meronyms and behavior-words) from the target word to its hyponym candidate as their necessary and sufficient conditions. To make the basic method more robust, this section proposes a method to acquire hyponymy relations from the Web, based on property inheritance of not only text features, but also typical image features for each concept by using not only Web text, but also Web images.

Our methods for automatic hyponym extraction from the Web are based on the following basic assumption of “Property Inheritance”. Let  $C$  be the universal set of concepts (conceptual words). This paper assumes that if and only if a concept  $x \in C$  is a hypernym (superordinate) of a concept  $y \in C$ , in other words, the concept  $y$  is a hyponym (subordinate) of the concept  $x$ , then the set of properties that the concept  $y$  has,  $P(y)$ , completely includes the set of properties that the concept  $x$  has,  $P(x)$ , and the concept  $y$  is not equal (equivalent) to the concept  $x$ .

$$\text{isa}(y, x) = 1 \Leftrightarrow P(y) \supseteq P(x) \text{ and } y \neq x,$$

$$P(c) = \{p \in P \mid \text{has}(p, c) = 1\},$$

where  $P$  stands for the universal set of properties and  $\text{has}(p, c) \in \{0, 1\}$  indicates whether or not a concept  $c \in C$  has a property  $p \in P$ ,

$$\text{has}(p, c) = \begin{cases} 1 & \text{if a concept } c \text{ has a property } p, \\ 0 & \text{otherwise.} \end{cases}$$

In other words, if and only if a concept  $y$  is a hyponym of a concept  $x$ , then the number of properties that both concepts  $x$  and  $y$  share is equal to the number of properties that the superordinate concept  $x$  has (and is less than the number of properties that the subordinate concept  $y$  has).

$$\text{isa}(y, x) = \begin{cases} 1 & \text{if } \sum_{p \in P} \text{has}(p, y) \cdot \text{has}(p, x) = \sum_{p \in P} \text{has}(p, x), \\ 0 & \text{if } \sum_{p \in P} \text{has}(p, y) \cdot \text{has}(p, x) < \sum_{p \in P} \text{has}(p, x). \end{cases}$$

It is essential for automatic hyponym extraction from the Web based on the above basic assumption to calculate the binary value  $\text{has}(p, c) \in \{0, 1\}$  for any pair of a property  $p \in P$  and a concept  $c \in C$  accurately. However, it is not easy, and we can calculate only the continuous value  $\text{has}^*(p, c) \in [0, 1]$  by using Web text and/or Web images in this paper. Therefore, we suppose that the ratio of the number of properties that a concept  $y \in C$  inherits from a target concept  $x \in C$  to the number of properties that the

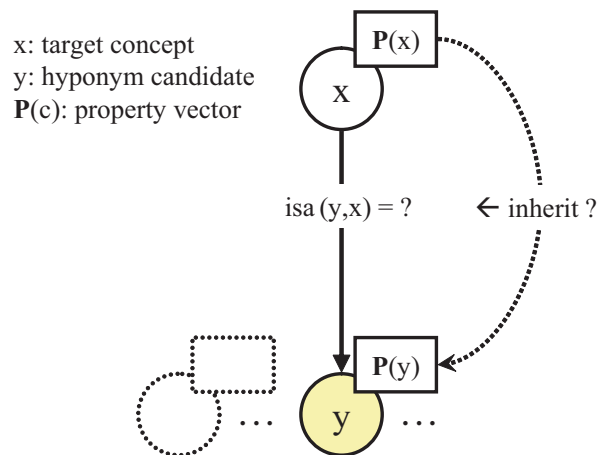


Figure 1. Hyponym Extraction based on Property Inheritance.

target concept  $x$  has,

$$\frac{\sum_{p \in P} \text{has}^*(p, y) \cdot \text{has}^*(p, x)}{\sum_{p \in P} \text{has}^*(p, x) \cdot \text{has}^*(p, x)},$$

can measure how suitable the concept  $y$  is for a hyponym of the target concept  $x$ ,  $\text{isa}^*(y, x)$ , as an approximation of whether or not the concept  $y$  is a hyponym of the target concept  $x$ ,  $\text{isa}(y, x)$ . Then, the concept  $y$  would be considered to be a hyponym of the target concept  $x$  when the ratio is enough near to one (or greater than a threshold value), while the concept  $y$  would be considered to be not a hyponym of the target concept  $x$  when the ratio is not near to one (or less than a threshold value).

When a target concept  $x \in C$  is given, our proposed method based on property inheritance executes the following four steps to extract its hyponyms from the Web. First, a set of candidates for its hyponyms of the target concept  $x$ ,  $C(x)$  is collected from the Web as exhaustively as possible. Second, the continuous value  $\text{has-txt}^*(p, c)$  or  $\text{has-img}^*(p, c)$  for each pair of a property (text or image feature)  $p \in P$  and a concept  $c \in C$  (the target concept  $x$  or its hyponym candidate  $y \in C(x)$ ) is calculated by analyzing not only Web text, but also Web images. Last, the continuous value  $\text{isa-PI}_n^*(y, x)$  for each pair of the target concept  $x$  and its hyponym candidate  $y \in C(x)$  is calculated based on property inheritance of the top  $n$  typical properties of the target concept  $x$  to its hyponym candidate  $y$ , and then a set of its top  $k$  hyponym candidates ordered by their weight would be outputted to the users.

### Step 1. Hyponym Candidate Collection

A set of hyponym candidates of the target concept  $x$ ,  $C(x)$  needs to be collected from the Web as exhaustively as possible and enough precisely. If  $C(x)$  should be set to

the universal set of concepts,  $C$ , its recall could equal to 1.0 (the highest) but its precision would nearly equal to 0.0 (too low). Meanwhile, if  $y \in C(x)$  is collected from some sort of corpus of text documents by using too strict lexico-syntactic pattern (e.g., “ $y$  is a kind of  $x$ ”), its precision is enough high but its recall is too low in most cases. Therefore, this paper uses not too strict but enough strict lexico-syntactic pattern of hyponymy to collect the set from the Web as exhaustively as possible and enough precisely. Any noun phrase  $y$  whose lexico-syntactic pattern “ $y$  is a/an  $x$ ” exists at least once in the title and/or summary text of the top 1000 search results by submitting a phrase “is a/an  $x$ ” as a query to Yahoo! Web Search API [19] is inserted into  $C(x)$  as a hyponym candidate of the target concept  $x$ .

### Step 2. Text Property Extraction

In our previous papers [15–18], typical properties  $p$  such as meronyms and behavior-words of each concept (the target concept  $x$  or its hyponym candidate  $y \in C(x)$ ) are extracted from only Web text as precisely as possible by using an enough strict lexico-syntactic pattern “ $c$ 's  $p$ ” as a sufficient condition of meronymy. The continuous value  $\text{has-txt}^*(p, c)$  of a text property  $p$  for each concept  $c$  is defined as follows:

$$\text{has-txt}^*(p, c) := \frac{\text{if}([\text{"c's p"}])}{\text{if}([\text{"c's s"}])} \in [0, 1],$$

where  $\text{if}([q])$  stands for the number (frequency) of Web images that meet a query condition  $q$  in such a corpus as the Web. This paper calculates it by submitting each query to Yahoo! Image Search API [20]. Note that  $\text{has-txt}^*(p, c)$  is not a binary value  $\{0, 1\}$  but a continuous value  $[0, 1]$ , so it cannot indicate whether or not a concept  $c$  has a property  $p$  but how typical the property  $p$  is of the concept  $c$ .

### Step 3. Image Property Extraction

This paper considers not only Web text, but also Web images, and extracts not only text features such as meronyms and behavior-words, but also image features of typical images as typical properties for each concept  $c$ . The top 100 search results by submitting a phrase “ $c$ ” as a query to Yahoo! Image Search API are reranked based on the VisualRanking algorithm [21] to acquire more typical images of the target concept  $c$ . The continuous value  $\text{has-img}^*(p, c)$  of an image feature  $p$  for each concept  $c$  is defined as follows by using the top  $k$  ( $= 10$ ) reranked images  $I_k(c)$ :

$$\text{has-img}^*(p, c) := \frac{\sum_{i \in I_k(c)} \text{prop}(p, i)}{k} \in [0, 1],$$

where  $\text{prop}(p, i)$  stands for the proportion of a HSV or SIFT [22] color-feature  $p$  in a Web image  $i$ .

### Step 4. Candidate Weighting by Property Inheritance

To filter out noisy hyponym candidates of the target concept  $x$ , each hyponym candidate  $y \in C(x)$  is assigned the weight  $\text{isa-PI}_n^*(y, x)$ , based on not only the inheritance

$\text{inherit-txt}_n^*(y, x)$  of the top  $n$  typical text features, but also the inheritance  $\text{inherit-img}_n^*(y, x)$  of the top  $n$  typical image features from the target concept  $x$ :

$$\begin{aligned} \text{isa-PI}_n^*(y, x) &:= (1 - \alpha) \cdot \text{inherit-txt}_n^*(y, x) \\ &\quad + \alpha \cdot \text{inherit-img}_n^*(y, x), \\ \text{inherit-txt}_n^*(y, x) &:= \frac{\sum_{p \in P_n^t(x)} \text{has-txt}^*(p, y) \cdot \text{has-txt}^*(p, x)}{\sum_{p \in P_n^t(x)} \text{has-txt}^*(p, x) \cdot \text{has-txt}^*(p, x)}, \\ \text{inherit-img}_n^*(y, x) &:= \frac{\sum_{p \in P_n^i(x)} \text{has-img}^*(p, y) \cdot \text{has-img}^*(p, x)}{\sum_{p \in P_n^i(x)} \text{has-img}^*(p, x) \cdot \text{has-img}^*(p, x)}, \end{aligned}$$

where  $\alpha \in [0, 1]$  stands for a certain combination parameter.

## III. EXPERIMENT

This section shows some experimental results to validate the proposed method to extract hyponymy relations from the Web, based on “Property Inheritance” of not only typical text features, but also typical image features for each concept, compared with a traditional lexico-syntactic pattern based hyponym extraction.

Figure 2 compares the average Precision-Recall curves by the proposed hybrid hyponym extraction ( $\alpha = 0.5, n = 10$ ) by using not only Web text, but also Web images, the previous hyponym extraction ( $\alpha = 0, n = 10$ ) by using only Web text, and a lexico-syntactic pattern based hyponym extraction for several kinds of target conceptual words such as “bird” and “flower”. The MAP (Mean Average Precision) of the proposed hybrid hyponym extraction is the best.

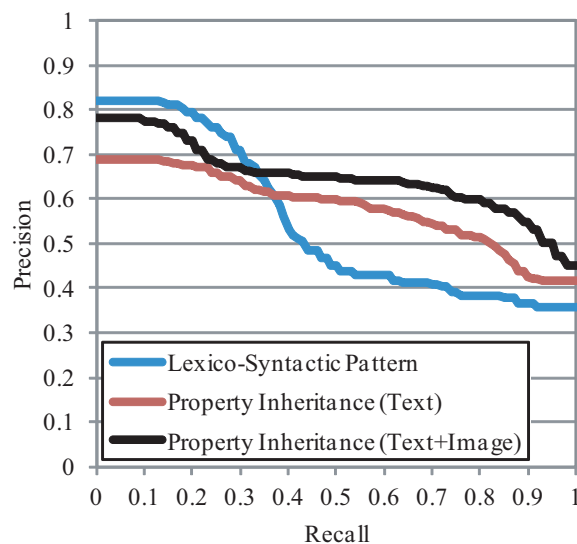


Figure 2. Precision-Recall of Hyponym Extraction based on Property Inheritance of Text and/or Image Features.



Table I  
TOP 18 HYPONYMS EXTRACTED FROM THE WEB FOR "PENGUIN".

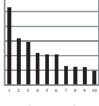
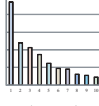
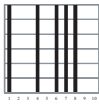
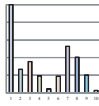
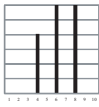
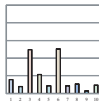
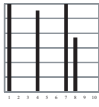
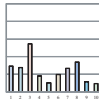
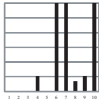
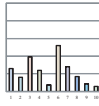
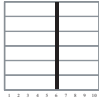
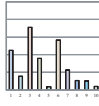
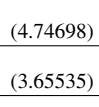
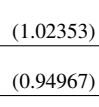
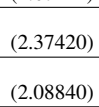
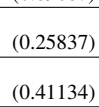
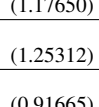
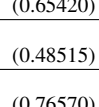
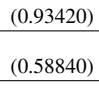
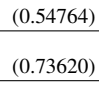
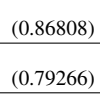
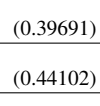




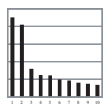
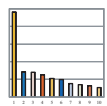
		1: photostream 2: iceberg 3: revenge 4: beak 5: poems 6: head 7: feet 8: nest 9: lair 10: eye	1: ■■■■ 2: ■■■■ 3: ■■■■ 4: ■■■■ 5: ■■■■ 6: ■■■■ 7: ■■■■ 8: ■■■■ 9: ■■■■ 10: ■■■■	penguin (—)	Top 10 Typical Text Features 	Top 10 Typical Color Features 
Rank	Syntactic Pattern	Text ( $\alpha = 0.0$ )	Image ( $\alpha = 1.0$ )	Text+Image ( $\alpha = 0.5$ )		
1	animal (196)	<b>gentoo penguin</b> (16.1158)	<b>gentoo penguin</b> (1.02559)	<b>gentoo penguin</b> (8.57070)		
2	favorite animal (128)	<b>yellow-eyed penguin</b> (11.0503)	<b>emperor penguin</b> (1.02353)	<b>yellow-eyed penguin</b> (5.72191)		
3	<b>tux</b> (86)	<b>little blue penguin</b> (7.66437)	<b>baby penguin</b> (0.94967)	<b>little blue penguin</b> (4.10788)		
4	book (50)	<b>king penguin</b> (6.78528)	<b>chinstrap penguin</b> (0.89687)	<b>king penguin</b> (3.63577)		
5	character (48)	<b>magellanic penguin</b> (6.53255)	pc (0.86006)	<b>magellanic penguin</b> (3.61665)		
6	<b>hoiho</b> (43)	<b>emperor penguin</b> (4.74698)	<b>african penguin</b> (0.85294)	<b>emperor penguin</b> (2.88526)		
7	pablo (43)	<b>baby penguin</b> (3.65535)	sutter (0.78754)	<b>baby penguin</b> (2.30251)		
8	friend (37)	<b>chinstrap penguin</b> (2.67442)	inch serving platter (0.784431)	<b>chinstrap penguin</b> (1.78565)		
9	<b>spheniscus mendiculus</b> (28)	mr. flibble (2.37420)	google (0.77023)	mr. flibble (1.31628)		
10	avatar (27)	<b>macaroni penguin</b> (2.08840)	<b>adelie penguin</b> (0.76570)	<b>macaroni penguin</b> (1.24987)		
11	hot dog (24)	favorite animal (1.25312)	political activist banksy (0.75514)	<b>royal penguin</b> (0.91535)		
12	uguin (22)	<b>royal penguin</b> (1.17650)	ty avalanche (0.75316)	favorite animal (0.86913)		
13	<b>galapagos penguin</b> (18)	<b>little penguin</b> (0.93420)	video (0.73873))	<b>adelie penguin</b> (0.84118)		
14	god (18)	<b>adelie penguin</b> (0.91665)	<b>tux</b> (0.73620)	<b>little penguin</b> (0.74092)		
15	<b>snares islands penguin</b> (17)	vigilance (0.86808)	<b>antarctic penguin</b> (0.73326)	<b>tux</b> (0.66230)		
16	heart (15)	misaki (0.79266)	<b>linux mascot tux</b> (0.71541)	<b>african penguin</b> (0.65259)		
17	poet (10)	wentworth miller (0.78618)	free pablo (0.70746)	vigilance (0.63249)		
18	<b>gentoo penguin</b> (9)	enemies (0.64338)	abbath (0.70085)	misaki (0.61684)		

Table II  
TOP 18 HYPONYMS EXTRACTED FROM THE WEB FOR “SUNFLOWER”.

Rank	Syntactic Pattern	Text ( $\alpha = 0.0$ )	Image ( $\alpha = 1.0$ )	Text+Image ( $\alpha = 0.5$ )		
		1: love 2: garden 3: field 4: seeds 5: life 6: smile 7: seed 8: head 9: leaves 10: spiral	1: ■■■■■ 2: ■■■■■ 3: ■■■■■ 4: ■■■■■ 5: ■■■■■ 6: ■■■■■ 7: ■■■■■ 8: ■■■■■ 9: ■■■■■ 10: ■■■■■	sunflower (——)	Top 10 Typical Text Features  (——)	Top 10 Typical Color Features  (——)
1	seed (208)	jill jack (480.541)	yellow (1.22165)	jill jack (240.390)	(480.541)	(0.23893)
2	favorite flower (52)	<b>tall sunflower</b> (213.538)	<b>girasol</b> (1.05447)	<b>tall sunflower</b> (106.943)	(213.538)	(0.34733)
3	district (42)	present invention (211.163)	<b>marigold</b> (0.86360)	present invention (105.940)	(211.163)	(0.71685)
4	navy blue field (23)	independent person (75.6619)	second parent sunflower plant (0.85420)	independent person (37.9542)	(75.6619)	(0.24643)
5	favorite thing (22)	<b>mirasol</b> (48.8920)	pairwise disjoint sets (0.83355)	<b>mirasol</b> (24.5911)	(48.8920)	(0.29011)
6	logo (21)	larva (42.6859)	sol (0.81621)	larva (21.4258)	(42.6859)	(0.16564)
7	yellow (12)	<b>common sunflower</b> (40.2172)	known prior art (0.75949)	<b>common sunflower</b> (20.4846)	(40.2172)	(0.75199)
8	hell (11)	favorite flower (35.4822)	<b>common sunflower</b> (0.75199)	favorite flower (17.8299)	(35.4822)	(0.17753)
9	sunbutter (11)	lead singer (19.1564)	inflorescence (0.73851)	lead singer (9.71413)	(19.1564)	(0.27188)
10	seal (10)	species (15.7655)	present invention (0.71685)	species (8.03862)	(15.7655)	(0.31178)
11	happiness (9)	aliya (13.6240)	imidazolinone herbicide (0.66606)	aliya (6.97572)	(13.6240)	(0.32740)
12	flower variation (8)	g-dragon (11.7593)	silver necklace (0.61189)	g-dragon (6.00615)	(11.7593)	(0.25297)
13	friend (7)	<b>jerusalem artichoke</b> (11.6205)	<b>maximilian's sunflower</b> (0.60568)	<b>jerusalem artichoke</b> (5.93293)	(11.6205)	(0.24531)
14	colour (6)	happiness (10.4684)	sunbutter (0.60099)	happiness (5.39702)	(10.4684)	(0.32564)
15	disjoint sets (6)	arapahoe (9.35538)	<b>helianthus annuus</b> (0.59916)	arapahoe (4.89790)	(9.35538)	(0.44043)
16	<b>jerusalem artichoke</b> (6)	mommy (6.20476)	size (0.59646)	mommy (3.25753)	(6.20476)	(0.31031)
17	pervenets (6)	fabric (5.60841)	disjoint sets (0.55639)	fabric (2.94874)	(5.60841)	(0.28907)
18	g-dragon (4)	larry (3.25074)	crepe back satin (0.55406)	larry (1.82185)	(3.25074)	(0.39296)

## IV. CONCLUSION

To achieve high recall and not low precision in automatic hyponym extraction from the Web, our previous work has assumed “Property Inheritance” from a target concept to its hyponyms and/or “Property Aggregation” from its hyponyms to the target concept to be necessary and sufficient conditions of hyponymy, and proposed several methods to extract hyponymy relations from the Web, based on property inheritance and/or property aggregation of text features such as meronyms and behavior-words. To make our previous method more robust, this paper has utilized not only Web text, but also Web images, proposed a method to acquire hyponymy relations from the Web, based on property inheritance of not only text features, but also image features for each conceptual word, and validated the proposed method by showing some experimental results.

## ACKNOWLEDGMENT

This work was supported in part by JSPS Grant-in-Aid for Young Scientists (B) “A research on Web Sensors to extract spatio-temporal data from the Web” (#23700129, Project Leader: Shun Hattori, 2011-2012).

## REFERENCES

- [1] Mandala, R., Tokunaga, T., and Tanaka, H.: “The Use of WordNet in Information Retrieval,” Proceedings of the COLING ACL Workshop on Usage of WordNet in Natural Language Processing, pp. 31–37 (1998).
- [2] Hattori, S., Tezuka, T., and Tanaka, K.: “Activity-based Query Refinement for Context-aware Information Retrieval,” Proceedings of the 9th International Conference on Asian Digital Libraries (ICADL’06), LNCS vol. 4312, pp. 474–477 (2006).
- [3] Hattori, S., Tezuka, T., Hiroaki, O., Oyama, S., Kawamoto, J., Tajima, K., and Tanaka, K.: “ReCQ: Real-world Context-aware Querying,” Proceedings of the 6th International and Interdisciplinary Conference on Modeling and Using Context (CONTEXT’07), LNAI vol. 4635, pp. 248–262 (2007).
- [4] Hattori, S.: “Alternative Query Discovery from the Web for Daily Mobile Decision Support,” Proceedings of the 5th IADIS International Conference on Wireless Applications and Computing (WAC’11), pp. 67–74 (2011).
- [5] Hattori, S.: “Hyponymy-Based Peculiar Image Retrieval,” Int’l Journal of Computer Information Systems and Industrial Management (IJCISIM), MIR Labs, vol. 5, pp. 79–88 (2012).
- [6] Fleischman, M., Hovy, E. and Echiabi, A.: “Offline Strategies for Online Question Answering: Answering Questions Before They Are Asked,” Proc. 41st Annual Meeting of the Association for Computational Linguistics, pp. 1–7 (2003).
- [7] Hattori, S., Tezuka, T., and Tanaka, K.: “Mining the Web for Appearance Description,” Proceedings of the 18th International Conference on Database and Expert Systems Applications (DEXA’07), LNCS vol. 4653, pp. 790–800 (2007).
- [8] Hattori, S. and Tanaka, K.: “Object-Name Search by Visual Appearance and Spatio-Temporal Descriptions,” Proc. of the 3rd Int’l Conference on Ubiquitous Information Management and Communication (ICUIMC’09), pp. 63–70 (2009).
- [9] Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J.: “Introduction to WordNet: An On-line Lexical Database,” International Journal of Lexicography, vol. 3, no. 4, pp. 235–312 (1993).
- [10] Völkel, M., Kröttsch, M., Vrandečić, D., Haller, H., and Studer, R.: “Semantic Wikipedia,” Proc. 15th Int’l Conference on World Wide Web (WWW’06), pp. 585–594 (2006).
- [11] Hearst, M. A.: “Automatic Acquisition of Hyponyms from Large Text Corpora,” Proc. 14th Int’l Conference on Computational Linguistics (COLING’92), vol. 2, pp. 539–545 (1992).
- [12] Morin, E. and Jacquemin, C.: “Automatic Acquisition and Expansion of Hypernym Links,” Computers and the Humanities, vol. 38, no. 4, pp. 363–396 (2004).
- [13] Kim, H., Kim, H., Choi, I., and Kim, M.: “Finding Relations from a Large Corpus using Generalized Patterns,” International Journal of Information Technology, vol. 12, no. 7, pp. 22–29 (2006).
- [14] Ruiz-Casado, M., Alfonseca, E., and Castells, P.: “Automatising the Learning of Lexical Patterns: An Application to the Enrichment of WordNet by Extracting Semantic Relationships from Wikipedia,” Data & Knowledge Engineering, vol. 61, no. 3, pp. 484–499 (2007).
- [15] Hattori, S., Ohshima, H., Oyama, S., and Tanaka, K.: “Mining the Web for Hyponymy Relations based on Property Inheritance,” Proceedings of the 10th Asia-Pacific Web Conference (APWeb’08), LNCS vol. 4976, pp. 99–110 (2008).
- [16] Hattori, S. and Tanaka, K.: “Extracting Concept Hierarchy Knowledge from the Web based on Property Inheritance and Aggregation,” Proc. of the 7th IEEE/WIC/ACM Int’l Conference on Web Intelligence (WI’08), pp. 432–437 (2008).
- [17] Hattori, S. and Tanaka, K.: “Extracting Concept Hierarchy Knowledge from the Web by Property Inheritance and Recursive Use of Term Relationships,” IPSJ Transactions on Databases (TOD40), vol. 1, no. 3, pp. 60–81 (2008).
- [18] Hattori, S.: “Object-oriented Semantic and Sensory Knowledge Extraction from the Web,” Web Intelligence and Intelligent Agents, In-Tech, pp. 365–390 (2010).
- [19] Yahoo! Web Search API, <http://search.yahooapis.jp/PremiumWebSearchService/V1/webSearch> (2012).
- [20] Yahoo! Image Search API, <http://search.yahooapis.jp/PremiumImageSearchService/V1/imageSearch> (2012).
- [21] Jing, Y. and Baluja, S.: “VisualRank: Applying PageRank to Large-Scale Image Search,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 11, pp. 1877–1890 (2008).
- [22] Lowe, D. G.: “Distinctive Image Features from Scale-Invariant Keypoints,” International Journal of Computer Vision, vol. 60, no. 2, pp. 91–110 (2004).