Semi-Automatic Ontology Engineering in Business Applications

Felix Burkhardt*, Jon Atle Gulla**, Jin Liu*, Christian Weiss*, Jianshen Zhou*

*T-Systems Enterprise Services
Goslarer Ufer 35
10589 Berlin, Germany

** Norwegian University of Science and Technology
Department of Computer and Information Science
NO-7491 Trondheim, Norway

[Felix.Burkhardt | Jin.Liu | Christian.Weiss | Jianshen.Zhou] @t-systems.com
Jon.Atle.Gulla@idi.ntnu.no

Abstract: Search technology can be applied to many applications and is in the heart of always growing information society. A promising new development is the use of ontologies to enable semantic modeling of data and user queries. Generation and maintenance of ontologies are a costly thing though. We propose a framework to build and maintain ontologies in a semi-automatic way. The article motivates the development from application point of view, relates to work known from the literature and introduces the so-called ontology workbench.

1 Introduction

As semantic web technologies including semantic search is nowadays not only a vision anymore but results in many applications, this article describes the first steps in a project for (semi-) automatic building, maintaining and extending ontologies. Ontologies provide the basis for the ontology-based semantic search and play a major role for allowing semantic access to data resources. However the human effort for creating, maintaining and extending ontologies is normally very high.

In order to reduce effort for engineering and managing ontologies, we have developed a general framework for ontologies learning from text. This framework will be applied as a fist step in the media domain, but is planned to be applicable in arbitrary domains, e.g. automotive applications, the travel domain or customer technical support.

This paper is organized as follows. In section 2 we give an overview of semantic web based technology applications and recall the beneficial effects using ontologies in various domains. In section 3 we refer to related work, concentrating on existing approaches to computer aided ontology engineering.. Section 4 describes our ontology workbench which is the nucleus of ontology building.

2 Envisaged Applications

Semantic technology is an application-driven technology. In one of the most recent studies [MD08], more than 100 application categories have been examined, where semantic related technologies can be applied to. Generally speaking, semantics as a leading technology in the evolution of the internet and in the information society, should help to understand and manage the opinions freely expressed by people and make them not only understandable to humans but also to computers. This is the most essential focus in many applications, as machine-to-machine communication will play more and more an important role in helping people to search, find and evaluate desired information.

Semantically enabled search and management technologies have still been characterized as an early market till now. The majority of current investment is more for R&D effort than for operational deployments. More and more companies, however, are seriously considering the gradual introduction of the new technologies. Similar to the classic motivations for investment, there exist three basic elements for measuring the business value:

1. Cost saving: This is to raise the efficiency. Purpose is to do the same job faster, cheaper and with fewer resources than it was done before.

2. Return on assets: This is to increase the effectiveness. Doing a better job than you did before, improving the productiveness and performance.

3. Return on investment: This is to create new and /or value-added services by changing some existing business aspects and/or adding new strategic advantage.

In the scope of the project where the planned ontology workbench will be developed, we'll address all three business values described above. In a first step, we'll concentrate on a selected use case with the application of advanced semantic and search technologies in the media entertainment domain

3 Related Work

As building, maintaining and evolution of ontologies is in the focus of applied research since long, numerous ontology learning approaches have been proposed and many toolsets have being developed over recent years. One can say that most of the ontology engineering methodologies follow a common approach. A minimal assembly is studying the feasibility, analyzing the requirements, extract the concepts and deployment. Please see among others [NVC04], [HV05], [GBI04] and [VVSH07]. Most of the ontology learning approaches combine a certain level of linguistic analysis with machine learning algorithms to find potentially interesting concepts and relations between them. The conceptualization is a non trivial process and involves the development of the domain model, the formalization of the model and the implementation [PM04]. Ontology learning toolsets may generate candidate concepts and relationships, but human labor is needed to verify the suggestions and complete the ontologies.

There are still very few ontology learning toolsets in active use in industrial ontology engineering projects. Most ontologies are constructed using traditional modeling approaches with teams of ontology experts and domain experts working together. This is partly due to the strategic issues involved in ontology engineering, but it also seems that current ontology learning tools do not have the reliability or credibility needed in large-scale ontology engineering projects

To follow and get an overview of current technologies in ontology engineering we refer to the comprehensive surveys of ontology learning techniques which are given in [BCM05], [Ci06], and [CVS06]. In the following we would like to mention shortly some existing ontology learning toolsets that are comparable to our approach.

Text2Onto[1] is a framework for data-driven change discovery by incremental ontology learning. It uses natural language processing and text mining techniques in order to extract an ontology from text and provides support for the adaptation of the ontology over time as documents are added or removed. Text2Onto was developed by the AIFB[2] initially in context with the European SEKT project (Semantically-Enabled Knowledge Technologies) [3] [CV05]. Text2Onto uses mainly libraries of Eclipse, Gate, Kaon, Lucene and Google.

OntoGen[4] is a system for data-driven semi-automatic topic ontology construction. The topic ontology is a set of topics connected with different types of relations. Each topic includes a set of related documents. OntoGen was also developed in the European-funded project SEKT by the project partner Jožef Stefan Institute, Slovenia[5] [FGM05].

1 Text2Onto - http://www.aifb.de/WBS/jvo/text2onto/
2 AIFB - http://www.AIFB.de/
3 SEKT - http://sekt.semanticweb.org/
4 OntoGen – http://ontogen.ijs.si/
5 Jožef Stefan Institute - http://wordnet.princeton.edu/

OntoLT[6], developed by the DFKI[7] by Buitelaar, P., D. Olejnik, M. Sintek and others [BOS04], is a plug-in for the mostly used ontology development tool Protégé, which supports the interactive extraction and/or extension of ontologies (concepts and relations) from a linguistically annotated text collection.

OntoLearn is a system for (semi-)automated ontology learning from domain texts, which has been developed by the Department of Computer Science at the University of Rome "La Sapienza"[8] [NVC04]. The key task performed by OntoLearn is the semantic interpretation or semantic disambiguation of terminology through machine learning and natural language processing. OntoLearn tries to identify the correct sense for each term and the relation between terms by building domain concept trees using data from the WordNet[9] knowledge base.

4 The Ontology Workbench

Whereas traditional ontology workbenches base their extraction on long chains of linguistic and statistical components, our workbench has a more interactive approach to ontology learning. Central to this architecture is a set of ontologically structured indices that are specifically designed to support numerous ontology extraction techniques.

The document collection used in the ontology learning process is first run through a chain of linguistic components. These include tokenization, stopword removal, parts of speech tagging, lemmatization, and noun phrase recognition. We assume that potential concepts are noun phrases that consist of either consecutive nouns or foreign phrases. This pre-processing phase is set up for both German and English documents, and the result is a set of documents ready for indexing and statistical analysis.

For some domains or data sources, we can beforehand define some overall high-level concepts that we afterwards use to filter the documents fed into the analysis. These concepts, which typically correspond to pre-defined headings or document structures, tend to mark out instance data that can be extracted directly with pattern-based techniques. For example, the heading Cast on www.imdb.com's movie pages tells us that the following lines list the actors (instances of concept ACTOR) and role names (instances of ROLE) in the movie.

If there are some high-level concepts defined beforehand, we use this information to construct specific indices in Lucene for each defined concept. We also construct indices for the full text available, and for all extracted noun phrases from the linguistic pre-processing.

6 OntoLT - http://olp.dfki.de/OntoLT/OntoLT.htm/
7 DFKI - http://www.DFKI.de/
8 La Sapienza - http://www.uniroma1.it/
9 WordNet - http://wordnet.princeton.edu/

Having built indices of sufficient sizes, we can start applying ontology extraction techniques on the indexed information. A particular analysis consists of the following three steps:

1. Definition of document set to use. To extract more specialized terminology, it is often necessary to restrict the document base to a subset of the whole collection and use the rest as a contrastive reference set. The definition is done with a structured query like GENRE: Drama, which selects drama movies for the analysis and uses all other movies as reference data.

2. Focus of analysis. The analysis may be focused on concepts or instances that are reflected in the index structure, or it may include all texts or all recognized noun phrases.

3. Choice of learning technique(s). After the document set and focus have been decided, a learning technique is chosen and the analysis is executed. The results are displayed and may be stored for later combination with other techniques or analyses.

Currently, we use a tf.idf-based technique for extracting potential concepts from the document collection. The ranking can be adjusted with different variations of the tf.idf score, and the analysis may be restricted to certain types of words, for example all noun phrases that contain at least two terms.

As relationships are more difficult to extract, we use several techniques that may be combined to achieve a satisfactory result. Using the index itself, we can characterize all identified concepts by means of large vectors of index terms. Calculating the cosine similarity score for every concept pair, we get a ranked list of relationships among concepts in the model. Another technique makes use of suffix tree clustering [ZA97] to establish relationships between related movies or other instances. Lastly, association rules may be used to extract relationships between concepts on the basis of document distribution measures [AIS93].

For every technique applied, the results may be stored and later combined with the results of other analyses. For example, we may accept only relationships that have been suggested by both the association rule component and the clustering component and have a combined score above a specified threshold. Research suggests that these hybrid approaches – where scores are combined from fundamentally different techniques - perform substantially better than the techniques in isolation [GBI08]. After experimenting with different techniques and combining results into acceptable lists of extracted concepts, instances and relationships, the user may generate an OWL ontology that also includes the statistical evidence for each ontology element.

We do not expect the generated ontology perfectly to reflect the terminology of the domain. Since the model is represented in OWL, however, an ontology editor like Protégé can afterwards be used to verify the model, correct any mistakes and complete the missing parts of the ontology. Our ontology learning components do not suggest any relationship names, for example, but these may be added manually if they are important to the subsequent application of the ontology.

5 Acknowledgements

The described work was done in scope of a project funded by the Deutsche Telekom Laboratories.

References

[AIS93] Agrawal, R., T. Imielinski, and A.N. Swami, Mining Association Rules between Sets of Items in large Databases, in Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data. 1993

[BCM05] Buitelaar, P., P. Cimiano, and B. Magnini – Ontology Learning from Text: An Overview. In: Paul Buitelaar, Philipp Cimiano, Bernardo Magnini (eds.) Ontology Learning from Text: Methods, Evaluation and Applications Frontiers in Artificial Intelligence and Applications Series, Vol. 123, IOS Press, July 2005

[BOS04] Buitelaar, P., Daniel Olejnik, Michael Sintek - A Protégé Plug-In for Ontology Extraction from Text Based on Linguistic Analysis. In: Proceedings of the 1st European Semantic Web Sym-posium (ESWS), Heraklion, Greece, May 2004

[Ci06] Cimiano, P. - Ontology Learning and Population from Text: Algorithms, Evaluation and Applications. Springer. 2006

[CV05] Cimiano, P. and J. Völker - Text2Onto – A Framework for Ontology Learning and Data-Driven Change Discovery. In Proc. NLDB, 2005

[CVS06] Cimiano, P., J. Völker, and R. Studer - Ontologies on Demand? A Description of the State-of-the-Art, Applications, Challenges and Trends for Ontology Learning from Text. Information, Wis-senschaft und Praxis, 2006. 57(6-7): p. 315-320

[FGM05] Fortuna, B., Grobelnik, M., Mladenič, D. - Visualization of text document corpus. Informatica (Slovenia) 29(4): 497-504 (2005)

[GBI08] Gulla, J. A. and T. Brasethvik. "A Hybrid Approach to Ontology Relationships Learning." Accepted for 13th International Conference on Applications of Natural Language to Information Systems, London, 2008.

[HV05] Haase, P. and Völker, J.: Ontology Learning and Reasoning - Dealing with Uncertainty and Inconsistency. ISWC-URSW 2005: 45-55

[MD04] Mills Davis: The Business Value of Semantic Technologies. September, 2004. http://www.knowledgefoundations.com/pdf-files/BusinessValue_v2.pdf

[MD08] Mills Davis: Semantic Wave 2008 Report: Industry Roadmap to Web3.0 & multibillion dollar market opportunities. http://www.readwriteweb.com/archives/semantic_wave_2008_free_report.php 2008

[MS01] Maedche A. and Staab, S.: Ontology Learning for the Semantic Web IEEE Intelligent Systems archive, Volume 16 , Issue 2, 2001

[NVC04] Navigli, R., Paola Velardi, Alessandro Cucchiarelli and Francesca Neri. - Extending and Enriching WordNet with OntoLearn, Proc. of The Second Global Wordnet Conference 2004 (GWC 2004), Brno, Czech Republic, 2004

[PM04] H.S. Pinto, J.P. Martins: Ontologies: How can they be built? Knowledge and Information Systems, 6(4), pp. 441-464, 2004

[VVSH07] Völker, J., D. Vrandecic, Y. Sure, and A. Hotho – Learning Disjointness. In Proceedings of 4th European Semantic Web Conference, The Semantic Web: Research and Applications (ESWC 2007), Innsbruck, June, 2007.

[ZA97] Zamir, O., Etzioni, O., Madani, O., & Karp, R.. Fast and intuitive clustering of Web documents. In Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, 1997