End-User Assisted Ontology Evolution in Uncertain Domains

Thomas Scharrenbach

Swiss Federal Institute for Forest, Snow and Landscape Research,
Zürcherstrasse 111, CH-8910 Birmensdorf, Switzerland
thomas.scharrenbach@wsl.ch

Abstract. Learning ontologies from large text corpora is a well understood task while evolving ontologies dynamically from user-input has rarely been adressed so far. Evolution of ontologies has to deal with vague or incomplete information. Accordingly, the formalism used for knowledge representation must be able to handle this kind of information. Classical logical approaches such as description logics are particularly poor in adressing uncertainty. Ontology evolution may benefit from exploring probabilistic or fuzzy approaches to knowledge representation. In this thesis an approach to evolve and update ontologies is developed which uses explicit and implicit user-input and extends probabilistic approaches to ontology engineering.

1 Introduction

The integration of datasources of different origin is quite a difficult task. Even though there exist standard mechanisms for querying like SQL, the underlying schemata may vary significantly. Traditional flat structures are very limited in the representation of the semantics of data. Over the recent couple of years, network based systems like ontologies have become more and more a standard in the semantic representation of data from and within different domains.

Ontologies allow for a sound definition of shared terms within and between different datasources. They are defined by the Web Ontology Language (OWL) recommended by the W3C. Currently, the Ontology layer is the highest layer of sufficient maturity within the Semantic Web [1]. As some sublanguages of OWL directly correspond to Description Logics (DL) traditional rule-based logical reasoning is straightforward and can be seen as state-of-the-art [2].

DL-based systems can model vague information only up to a certain degree by defining, e. g. disjoints, similarity relations etc. However, there are several cases where the explicit modelling of uncertainty is desirable [3]:

- While knowledge and knowledge representation usually are incomplete, there is no sound concept of vague information in DL systems: either something is asserted in the knowledge base or not. And if it is asserted it is true or false, nothing in-between.

A. Sheth et al. (Eds.): ISWC 2008, LNCS 5318, pp. 920–925, 2008.

- In addition to the presence of a logical consequence, it is desired to know how likely a certain event will occur. If e. g. we know that birds can fly with a probability of more than 0.9 and Tweety is a bird then the probability that Tweety has the ability to fly shall be higher than 0.9.
  $Bird(Tweety) \wedge Pr(canFly(Bird)) > 0.9 \rightarrow Pr(canFly(Tweety)) > 0.9$.
- In some cases, contradictions which would violate the consistency of a DL system, must be allowed up to a certain degree. Suppose a simple ontology of birds. Birds can usually fly with the exception of penguins. These are birds that cannot fly. The corresponding DL knowledge base would thus be inconsistent. $\{canFly(Bird), Penguin \sqsubseteq Bird, \neg canFly(Penguin)\}$. The simplest way to overcome this inconsistency is to split the concept of birds into two new concepts of flying and non-flying birds. A more elegant way is to assign the role *canFly* a probability in which the inconsistency is relaxed such as $Pr(canFly(Bird)) > 0.9$. This states that birds can fly with a probability of more than 0.9 but also allows for non-flying birds without structural changes.

These limitations can be overcome by extending the concept of ontologies with a probabilistic or fuzzy model. The need for such a model is even more acute when the *evolution* of ontologies is considered. The members of a community may want to develop an ontology further. Be it because the ontology is incomplete or because additional knowledge is created which is materialized in new concepts and facts. In such a case these concepts are typically related to the existing ones only in a weakly or undefined way which cannot be put in terms of the primitives of a classical logical formalism. Furthermore, they might introduce inconsistencies into the existing knowledge base which can be relaxated in a probabilistic or fuzzy model. Finally, one of the most interesting question is to what degree the construction or update of ontologies can be automated. This thesis investigates whether and how user-input can be used to automatically construct and/or update application ontologies for heteroneneous data sources by extending probabilistic and fuzzy approaches to description logic reasoning.

2 Related Work

Probabilitstic approaches to ontology engineering can be divided into two groups: approaches directly extending the ontology and approaches where the ontology is transformed into a different representation allowing for probabilistic modelling.

DL are a family of formal languages for structured terminological knowledge representation [2]. On the one hand they can describe the formal concepts of a domain and on the other hand they allow for first-order logic inference. The OWL languages OWL Lite and OWL DL are explicitely based on DL which makes the use of DL for reasoning in ontology based systems straightforward.

While DL provide formal logical representation and inference, uncertainty like "Birds can fly with probability of 0.9" cannot be modelled very well. Lukasiewicz proposed a probabilistic extension called *Probabilistic Description Logics* (PDL) [4]. Individuals can be assigned conditional probability constraints which are

asserted to a so-called PABox P_o for every probabilistic individual $o \in \mathbf{I}_P$. Analogously a PTBox $PT = (T, P)$ is defined holding a set of conditional constraints P for the knowledge base $\mathcal{K}$. PDL consist of a set of classical individuals $\mathbf{I}_C$ a set of probabilistic individuals $\mathbf{I}_P$, a PTBox, an ABox, and one PABox for every $o \in \mathbf{I}_P$. This concept allows for modelling quantified uncertainty and first software reasoning tools are available [5].

Classical DL cannot model vague concepts like "Tweety is young". Therefore, Straccia [6] introduced FuzzyOWL. The knowledge base is enriched by fuzzy role inclusion axioms, fuzzy concept inclusion axioms, fuzzy concept assertions and fuzzy role assertions. This induces a fuzzy RBox, a fuzzy ABox, and a fuzzy TBox, respectively. This method enables inferences of "vague rules". While first lacking methods for the reasoning process, recent progress has been made in this area and reasoning tools are available [7]. According to the fuzzy approach, measuring the level of uncertainty is not possible directly [8], at least not as straightforward as in the PDL case.

Although there exist many other approaches for probabilistic ontology based knowledge engineering systems like e. g. BayesOWL [9] the presented ones are considered as the most relevant for the research subject of this thesis. For a more detailed overview the reader is referred to [4].

Work on the update of ontologies has mainly been performed for classical DL based systems [10,11] and for agent systems [12]. The main challenge is to keep the knowledge base consistent which could efficiently only be achieved on the instance level so far. Recently, Haase and Völker proposed a scheme based on finding the minimal inconsistent subontology [13]. According to a confidence measure inconsistent resources are removed until there are no more inconsistencies left. This approachs allowing for the update of arbitrary resources removes contradicting information instead of modelling it.

The aspect of ontology update incorporating inconsistent information using probabilistic or fuzzy approaches has not been addressed yet and will be subject of this thesis.

3 Research Plan

The "Virtual Data Centre" (VDC) of the "Datenzentrum Natur Landschaft" (DNL) project is a collection of several enviromental databases mainly containing data from taxonomies, different land registers, legal documents etc. As such, though there exists no common scheme or explicit references, the data is strongly semantically correlated. The aim of this project is to model the semantics by a multi-lingual eco-ontology on the application level.

3.1 Current State of Research

In a first step, a bilingual eco-ontology was created by expert users from scratch. An open search was realized by an expansion scheme [14] and the reasoning is based on DL. Further research is performed with the objective to extend

the knowledge base by a so-called RCCBox representing composition tables for spatial inference based on the Region Connection Calculus (RCC) [15,16]. In late 2008, a first prototype shall be released for a test cycle of selected expert users at the Swiss Federal Office for the Enviroment.

3.2 Future Research

Creating the Baseline. In a first step, the ontology will be extended by a statistical model following both, the PDL and the FuzzyOWL approaches. While there do exist reasoning tools the main challenge lies within the estimation of the parameters for the underlying probability distributions and fuzzy sets. Both approaches will therefore be extended by methods for estimating and automatically updating the corresponding parameters. For this task, classical statistical text-classification approaches will be used like described in [17] which result in a probabilistic and a fuzzy ontology, respectively, acting as the baseline model for this thesis.

Incorporating User-Input. Though the data is semantically connected, in the baseline model these links are not yet established. Furthermore, the baseline model is assumed to be an incomplete representation of the data. Hence, user-input will be incorporated to obtain the required informaton.

One of the main applications of the DNL is the open search that will be used for gathering the desired input during the search process. This step is divided into two parts: Using explicit user-feedback on the one hand and using implicit user-input on the other hand [18]. In this context, the WSL Ontology Webeditor (WOW) is under development allowing for the explicit insertion of new resources into the ontology. While the incorporation of explicit user-input is straightforward, for the implicit input a search context has to be defined. The information of a failed query, i. e. the search terms, will be linked to following search terms and inserted into the knowledge base. In case of a successful query, the confidence of the corresponding resources will be increased.

Along with that, methods for the extension of the ontologies and the corresponding statistical models will be investigated enabling a sound an efficient update.

Handling Inconsistencies. In the first phase of research, the explicit extension will be restricted to the addition of new instances. Later on also the insertion will take place on concept and role level. Inconsistencies will then not be resolved by removing resources like in [19] but modelled explicitely by means of uncertainty. This way, information will not be pruned w. r. t. its relevance but will be kept inside the knowledge base itself. Not only will the proper presentation of the ontology for the insertion of new resources be part of this thesis' research, but also the aspect of how to offer the possibility to let the user specifiy the amount of vagueness for the extension. While the first may be adapted like presented in [20] the latter will be realized in terms of how likely the new individual matches

to the actual knowledge base. Within this context it will be interesting to see especially how the insertion will work for geo-spatial data. Particulary, modelling the update of geo-spatial approximations as described in [21] by probabilistic means.

Evaluation. For the evaluation of the performance of the developed methods, a reference dataset will be constructed in cooperation with expert end-users to measure the improvement of precision and recall for the updated ontologies. Since end-user-feedback will be available within the context of the DNL project, this will be used as well for the evaluation of how well the tested methods work for the evolution of the knowledge base.

4 Conclusion and Outlook

The problem of consistent evolution of probabilistic ontologies has not been addressed so far. This thesis investigates how to evolve, i. e. learn and update, a multi-source multi-langual eco-ontology from user-input. Therefore, probabilistic extensions of classical DL knowledge bases will be used with a focus on either Probabilistic Description Logics or FuzzyOWL. These approaches will be extended by an update scheme to incorporate implicit and/or explicit user-input into the knowledge base. Different aspects of how to obtain the desired information from the end-user for the extension of a knowledge base with an underlying statistical model will be investigated. While the extension will be at first restricted to instance level the extension to concept level will be explored based on the gathered results. For systematic evaluation, a reference dataset will be constructed as well as will be used explicit feedback from end-users.

Acknowledgments. I would like to thank Prof. Abraham Bernstein for supervising this thesis as well as Bettina Bauer-Messmer and Rolf Grütter for their support.

References

1. McGuinness, D.L., van Harmelen, F.: OWL Web Ontology Language overview. W3C recommendation, W3C (February 2004), `http://www.w3.org/TR/2004/REC-owl-features-20040210/`
2. Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., Patel-Schneider, P.F. (eds.): The Description Logic Handbook: Theory, Implementation, and Applications. Cambridge University Press, Cambridge (2003)
3. Bacchus, F.: Representing and Reasoning with Probabilistic Knowledge: a Logical Approach to Probabilities. MIT Press, Cambridge (1990)
4. Lukasiewicz, T.: Probabilistic Description Logics for the Semantic Web. Technical Report, Knowledge-Based Systems Group Tu Vienna (2007)
5. Clark & Parsia: Pronto - a Probabilistic Extension for OWL DL and Pellet, `http://pellet.owldl.com/pronto`

6. Straccia, U.: Towards a Fuzzy Description Logic for the Semantic Web. In: Gómez-Pérez, A., Euzenat, J. (eds.) ESWC 2005. LNCS, vol. 3532. Springer, Heidelberg (2005)
7. Stoilos, G., Stamou, G.: Extending Fuzzy Description Logics for the Semantic Web. In: 3rd International Workshop of OWL: Experiences and Directions, Innsbruck (2007)
8. Stoilos, G., Stamou, G., Tzouvaras, V., Pan, J., Horrocks, I.: Fuzzy owl: Uncertainty and The Semantic Web. In: 21st International Workshop on Description Logics (DL 2008), Galway (2005)
9. Ding, Z., Peng, Y.: A Probabilistic Extension to Ontology Language OWL. In: Proceedings of the 37th Hawaii International Conference On System Sciences (HICSS-37), Big Island, Hawaii (2004)
10. Winslett, M.: Updating Logical Databases. Cambridge University Press, Cambridge (1990)
11. Giacomo, G.D., Lenzerini, M., Poggi, A., Rosati, R.: On the Update of Description Logic Ontologies at the Instance Level. In: AAAI 2006 (2006)
12. McNeill, F., Bundy, A., Walton, C.: Facilitating Agent Communication Through Detecting, Diagnosing and Refining Ontological Mismatch. In: Proceedings of the KR 2004 Doctoral Consortium, AAAI Technical Report (2004)
13. Haase, P., Völker, J.: Ontology Learning and Reasoning - Dealing with Uncertainty and Inconsistency. In: Paulo, C.G., et al. (eds.) Uncertainty Reasoning for the Semantic Web I. Springer, Heidelberg (to appear, 2008)
14. Grütter, R., Bauer-Messmer, B., Frehner, M.: First Experiences with an Ontology-Based Search for Environmental Data. In: Proceedings of the 11th AGILE International Conference on Geographic Information Science (AGILE 2008), Girona, Spain (2008)
15. Grütter, R., Bauer-Messmer, B.: Towards Spatial Reasoning in the Semantic Web: A Hybrid Knowledge Representation System Architecture. In: Proceedings of the 10th AGILE International Conference on Geographic Information Science (AGILE 2007), Aalborg, Denmark (2007)
16. Grütter, R., Bauer-Messmer, B., Hägeli, M.: Extending an Ontology-Based Search with a Formalism for Spatial Reasoning. In: Proceedings of the 23rd Annual ACM Symposium on Applied Computing (ACM SAC 2008), Fortaleza, Brazil (2008)
17. McCallum, A., Nigam, K.: A Comparison of Event Models for Naive Bayes Text Classification. In: Proceedings of the AAAI 1998 Workshop on Learning for Text Categorization, pp. 41–48 (1998)
18. Bauer-Messmer, B., Grütter, R., Scharrenbach, T.: Improving An Environmental Ontology by Incorporating User-Input. In: EnviroInfo 2008 - Environmental Informatics and Industrial Ecology, Lüneburg, Germany (to appear, 2008)
19. Haase, P., Stojanovic, L.: Consistent Evolution of OWL Ontologies. In: Gómez-Pérez, A., Euzenat, J. (eds.) ESWC 2005. LNCS, vol. 3532. Springer, Heidelberg (2005)
20. Bernstein, A., Kaufmann, E.: Gino - a Guided Input Natural Language Ontology Editor. In: Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., Aroyo, L.M. (eds.) ISWC 2006. LNCS, vol. 4273, pp. 144–157. Springer, Heidelberg (2006)
21. Grütter, R., Scharrenbach, T., Bauer-Messmer, B.: Improving an RCC-Derived Geospatial Approximation by OWL Axioms. In: Sheth, A., et al. (eds.) ISWC 2008. LNCS, vol. 5318. Springer, Heidelberg (2008)