Probabilistic Information Integration and Retrieval in the Semantic Web

Livia Predoiu

Institute of Computer Science, University of Mannheim,
A5,6, 68159 Mannheim, Germany
livia@informatik.uni-mannheim.de

1 Research Context

The Semantic Web (SW) has been envisioned to enable software tools or Web Services, respectively, to process information provided on the Web automatically. For this purpose, languages for representing the semantics of data by means of ontologies have been proposed such as RDF(S) and OWL. While the semantics of RDF(S) requires a non-standard model-theory that goes beyond first order logics, OWL is intended to model subsets of first order logics. OWL consists of three variants that are layered on each other. The less expressive variants OWL-Light and OWL-DL correspond to the Description Logics $\mathcal{SHIF}$(D) and $\mathcal{SHOIN}$(D) [1], respectively, and thus to subsets of First Order Logics [2].

While RDF and OWL are W3C recommendations and hence a kind of standard, a lot of proposals emerged recently for representing Logic Programming (LP) variants on the Web. Such proposals are e.g. SWRL[1] and WRL[2]. Furthermore, a working group exists at the W3C for defining a rule interchange format[3]. Therefore, it can be expected that rule languages will play an important role in the SW. The Description Logics (DL) and the LP paradigm are orthogonal having just a small subset in common [3] and a comparison reveals a balanced amount of advantages and disadvantages of one compared to the other e.g. concerning the efficience of certain reasoning tasks [4].

The SW will consist of independent peers each providing information that describes overlapping domains by different ontologies or logic programs specified in different knowledge representation languages of the DL and the LP paradigm. In order to enable intelligent software tools to utilize the information represented by these peers in a coherent manner, the ontologies and logic programs need to be aligned by means of mappings. An number of approaches for learning mappings between ontologies exist already [5]. Most of them detect very simple matchings and can be used for learning mappings between ontologies and logic programs as well.

[1] http://www.w3.org/Submission/SWRL/

[2] http://www.w3.org/Submission/WRL/

[3] http://www.w3.org/2005/rules/wg.html

2 Problem Definition

Currently, mappings are mainly used deterministically. I.e. although in general, automatically learned mappings are closely connected to a confidence which expresses some kind of belief of the matcher that each mapping holds, the mappings are considered to be either true or false depending on some threshold level.

If the mappings that are found by a probabilistic matching approach, there is evidence that keeping the probabilities and using probabilistic inference for answering queries is likely to change and improve the outcome compared to a deterministic usage of mappings. This holds especially if we consider real world SW mapping scenarios where several ontologies are connected by mappings in a catenarian way (with mapping composition). Using mappings that have a probability lower than a threshold is likely to influence the results in such a way that previously ruled out results get a high probability. Also mappings that are found by non-probabilistic approaches are in general found with a number that expresses the confidence of its validity. This number can be interpreted probabilistically, e.g. by means of stating error probabilities. Thus, these findings are not limited to probabilistic mapping approaches.

This thesis aims for the development of a framework that enables a Semantic Web consisting of DL and LP knowledge bases being connected by mappings that are attached each by a probability that expresses the certainty of the validity of it. By means of such a framework and a probabilistic extension of a language that integrates LP and DL variants probabilistic information retrieval for the Web can be implemented.

3 Expected Contribution

For the development of a probabilistic information integration framework for the SW that integrates probabilistic mappings with probabilistic and deterministic ontologies and logic programs being mapped one to another by the mappings, a probabilistic SW Language and reasoning algorithms are required.

Hence, the expected contribution is

1. a probabilistic extension of a language that is capable of integrating the DL variants underlying OWL (or OWL 1.1 or one or more of its tractable fragments [6]) and variants of the LP paradigm.
2. distributed reasoning algorithms for this language that consider the inherently distributed nature of the information sources in the SW.
3. tools that implemement the language and the algorithms.
4. a framework that
   - integrates probabilistic and deterministic knowledge bases provided by peers in the SW by utilizing the language and reasoning algorithms mentioned above
   - provides facilities for (distributed) probabilistic information retrieval in order to enable efficient retrieval of the probabilistically integrated information

The advantages of such a framework are that it will be possible to

- express probabilistic knowledge in the SW
- integrate probabilistic and deterministic knowledge in the SW
- integrate DL and LP knowledge bases
- use the confidence of mappings (and thus the heuristics that matchers are using for discovering mappings) and improve the preciseness of information integration especially in settings that involve mapping composition.
- incorporate means to integrate preference between and trust in data sources and/or matchers
- use conflicting mappings to some extent
- perform information retrieval over distributed DL and LP knowledge bases

4 Related Work

A probabilistic framework for Information Integration and Retrieval on the SW does not exist yet. However, in [7] suggestions are made for such a framework. But the only substantial contribution to such a framework is a tool for learning mappings consisting of simple probabilistic Datalog (pDatalog) rules [8] between OWL ontologies. Ideas on how to reason with the ontologies and rules are missing.

There exist a couple of probabilistic extensions of SW languages that provide a tight integration on the formal level between a SW Language or a subset of it and a probabilistic model. Such a tight integration is needed for the framework that is intended to be developed in this thesis. Besides probabilistic extensions that just consider RDF or OWL, the following extensions are related to integrating DL and LP. pOWL Lite^{-} and its extension with equality, pOWL Lite^{EQ}, [9] are probabilistic extensions of a subset DLPs [10] basing on pDatalog. The resulting formalism is a subset of pDatalog. Information integration is not considered in the context of pOWL Lite$^{-}/^{EQ}$. The languages have been proposed solely for the purpose of expressing probabilistic OWL statements. However, as these languages are basing on DLPs which is a KR formalism lying in the common subset of DL and LP, probabilistic Information Integration can be realized with them. As oMap [7] discovers mappings consisting of simple pDatalog rules, an information integration setting is conceivable that combines the pOWL Lite$^{-}/^{EQ}$ languages with oMap.

Probabilistic Description Logic Programs is a KR formalism that integrates the DLs underlying OWL-Lite and OWL-DL with stratified Datalog [11] and disjunctive Datalog with Negation [12]. Its probabilistic model is based on Independent Choice Logic [13]. However, the interaction between the DL part and the LP part is limited. A less restricted probabilistic integration of the DLs underlying OWL-Lite and OWL-DL with disjunctive Datalog with Negation is expressed by tightly integrated probabilistic description Logic Programs (tiPDL) [14]. Currently, there are no reasoning tools available for these formalisms and reasoning in the general formalisms is very inefficient.

5 Approach and Methodology

Problem Definition: A framework for probabilistic information integration and retrieval for the SW which can be expected to consist of DL and LP knowledge bases does not exist yet, but is needed in order to make use of the uncertainty that is inherently present in each mapping.
Identification of Requirements:
The **language** for the framework has the following requirements on its *expressivity*: it needs to be capable of integrating the DL and LP variants that are important in the SW. It also requires a *tight integration with a probabilistic model.* In the scope of this thesis, DLPs have been extended with probabilities obeying the probabilistic model of Bayesian Logic Programs (BLPs) [15], yielding Bayesian DLPs. The resulting formalism is called Bayesian DLPs (BDLPs). For BDLPs, also a way to integrate probabilistic and deterministic ontologies and logic programs lying in the DLP fragment with probabilistic and deterministic mapping rules has been proposed in [4]. For reasoning, usage of the existing BLP reasoner Balios has been proposed as BLPs are a superset of BDLPs. While the integration of DLPs and BLPs is very tight and thus sufficient for our purposes, the expressivity of the DLP fragment is too limited. Currently, I am investigating the tiPDL [14] KR formalism mentioned above and subsets of it. Subsets that use the subset of ICL that lies in Bayesian Networks seem to be very promising for the purpose of the framework.
The requirements for **reasoning algorithms** in this framework are the *consideration of the inherent distribution* of the data over several peers. I.e. the reasoning algorithms to be developed should be able to select peers that are relevant to a specific query, merge the results of distributed reasoning resources and thus take advantage of parallel reasoning. Due to the high expressivity of the language, reasoning in general can be expected to be very inefficient. Therefore, the reasoning algorithms to be developed in this thesis will be approximate reasoning algorithms due to the natural requirement of *efficiency.*
The requirements for an **implementation of the framework** is the *creation of an infrastructure* for the framework. For this purpose, existing tools are intended to be reused. Thus, appropriate tools need to be evaluated in order to enable the choice of the ones that are best suited for the framework. Clearly, another requirement for the implementation is efficieny. Furthermore, methods that asses preference between and trust into data sources and matchers are needed.
Design: The design of the framework needs to enable fast and efficient access of the data sources to be integrated. It will be modular. However, its' specification depends on the results of the tool analysis for the implementation of the framework.
Evaluation: For the evaluation of the framework, ontologies (and mappings) from the Ontology Alignment Evaluation Initiative[4] can be used. A set of logic programs needs to be collected as well and a couple of mapping tools can be used for discovering a set of mappings between the logic programs and the logic

[4] http://oaei.ontologymatching.org/

programs and the ontologies. In order to show that the probabilistic approach of this thesis is appropriate for resolving conficting mappings, it can be evaluated against approaches that resolve conflicting mappings by repairing, e.g. [16]. It has also to be shown whether the usage of the confidence and probability values of the matchers improves the results of information integration in a setting that involves several ontologies and mapping composition. The information retrieval facilities will be compared with other current information retrieval tools.

References

1. Horrocks, I., Patel-Schneider, P.F., van Harmelen, F.: From $\mathcal{SHIQ}$ and RDF to OWL: the making of a Web Ontology Language. Journal of Web Semantics (2003)
2. Borgida, A.: On the relationship between description logic and predicate logic. Artificial Intelligence **82**(1-2) (2006)
3. Krötzsch, M., Hitzler, P., Vrandecic, D., Sintek, M.: How to reason with OWL in a logic programming system. In: Proc. of the Conference on Rules and Rule Markup Languages for the Semantic Web. (2006)
4. Predoiu, L.: Information integration with bayesian description logic programs. In: Proc. of 3rd IIWeb Workshop for Information Integration on the Web. (2006)
5. Euzenat, J., Shvaiko, P.: Ontology matching. Springer-Verlag, Heidelberg (2007)
6. Grau, B.C., Calvanese, D., Giacomo, G.D., Horrocks, I., Lutz, C., Motik, B., Parsia, B., Patel-Schneider, P.F.: OWL 1.1 Web Ontology Language Tractable Fragments. URL: http://www.w3.org/Submission/owl11-tractable/ (2006)
7. Straccia, U., Troncy, R.: Towards Distributed Information Retrieval in the Semantic Web: Query Reformulation Using the oMAP Framework. In: Proc. of the 3rd European Semantic Web Conference. (2006)
8. Fuhr, N.: Probabilistic Datalog: Implementing Logical Information Retrieval for Advanced Applications. Journal of the American Society for Information Science **51**(2) (2000)
9. Nottelmann, H., Fuhr, N.: Adding Probabilities and Rules to OWL Lite Subsets based on Probabilistic Datalog. Uncertainty, Fuzziness and Knowledge-Based Systems **14**(1) (2006)
10. Grosof, B.N., Horrocks, I., Volz, R., Decker, S.: Description Logic Programs: combining logic programs with description logic. In: Proc. of the 12th international conference on World Wide Web. (2003)
11. Lukasiewicz, T.: Stratified Probabilistic Description Logic Programs. In: Proc. of the ISWC Workshop on Uncertainty Reasoning for the Semantic Web. (2005)
12. Lukasiewicz, T.: Probabilistic Description Logic Programs. In: Proc. of the conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty. (2005)
13. Poole, D.: The independent choice logic for modelling multiple agents under uncertainty. Artificial Intelligence **94**(1-2) (1997)
14. Cali, A., Lukasiewicz, T.: Tightly Integrated Probabilistic Description Logic Programs. Technical report, Institut für Informationssysteme. TU Wien (2007)
15. Kersting, K., Raedt, L.D.: Bayesian Logic Programs. Technical report, Albert-Ludwigs University, Freiburg (2001)
16. Meilicke, C., Stuckenschmidt, H., Tamilin, A.: Repairing Ontology Mappings. In: Proc. of AAAI. (2007)