Pronto: Probabilistic Ontological Modeling in the Semantic Web

Pavel Klinov
University of Manchester
Manchester M13 9PL, UK
pklinov@cs.man.ac.uk

Bijan Parsia
University of Manchester
Manchester M13 9PL, UK
bparsia@cs.man.ac.uk

ABSTRACT

This demonstration illustrates the benefits of probabilistic ontological modeling for uncertain domains in the Semantic Web. It is based on Pronto - probabilistic OWL reasoner that allows modelers to complement classical OWL ontologies with probabilistic statements. In addition to Pronto's features and capabilities, a great deal of the demonstration will be devoted to presenting modeling patterns, typical pitfalls, desirable as well as incidental consequences of probabilistic reasoning. The testbed will be the prototype of the Breast Cancer Risk Assessment ontology that we have developed to evaluate Pronto.

1. INTRODUCTION

One of the limitations of ontological languages used in the Semantic Web, namely OWL, is the inability to handle uncertain knowledge. It is a serious obstacle to the expansion of the Semantic Web because many domains of human interest contain knowledge that cannot be represented with absolute certainty. One example of an uncertain domain is medicine, in particular, disease diagnosing. Symptoms, causes and consequences of many diseases are uncertain which complicates conceptualization of such domains in formal ontologies and thus restricts machine understanding.

In this demonstration we present Pronto - a probabilistic OWL reasoner that has been developed to address this issue [1]. Pronto implements reasoning services of P-$\mathcal{SHIQ}$(D) - a very expressive formalism which is a probabilistic generalization of OWL with the exception of nominals [4]. Pronto can represent and reason about probabilistic facts as well as generic probabilistic relationships that typically arise from statistical experiments. The demonstration aims to illustrate that important problems can be reduced to probabilistic reasoning in P-$\mathcal{SHIQ}$(D).

Although the demonstration will describe the features and capabilities of Pronto, it will be geared to its usage in practical applications. The goal is to clearly show what can be done using Pronto, what the typical patterns of probabilistic modeling are and what can be expected from probabilistic reasoning. Most of these aspects will be illustrated on the Breast Cancer Risk Assessment (BRCA) ontology[1] that has been developed specifically for that purpose. The ontology will serve as an example of the ontological alternative to the traditional Bayesian modeling of such uncertain problems.

2. PRONTO: FUNCTIONAL OVERVIEW

As a P-$\mathcal{SHIQ}$(D) reasoner, Pronto is capable of representing and reasoning about uncertainty in both, generic background knowledge and individual facts. It complements the OWL syntax with *conditional constraints* - constructs that express probabilistic relationships between OWL classes or an OWL class and an individual. The following are the examples of conditional constraints from the BRCA ontology:

- (WomenUnderAbsoluteBRCRisk|Women)[0,0.123] meaning that an average woman's risk to develop breast cancer is under 12.3%
- Ann:(WomenWithHighLevelOfEstrogen|$\top$)[0.9,1.0] meaning that the degree of belief that Ann's level of estrogen is high is over 90%

The first is an example of a generic, statistical relationship (TBox constraint) whether the second expresses a rather subjective degree of belief about a specific individual (ABox constraint). Combining these two sources of probabilistic knowledge is one of the fundamental features of Pronto.

These extra constructs do not interfere with OWL in any way. Pronto is built on top of Pellet [5] so all the classical OWL representation and reasoning services are retained. Conditional constraints are added in the form of OWL 2.0 annotations which are semantic-free for all other applications. Importantly this allows modelers to reuse existing classical ontologies in probabilistic models.

Pronto provides reasoning services in the form of generic and individual entailments. Both types of entailment are non-monotonic, i.e. new knowledge (classical or probabilistic) can affect previously made entailments. Entailed constraints always have the tightest possible, and thus the most informative, probability intervals.

Finally, Pronto helps to understand the reasoning results by providing *explanations* - the minimal subsets of probabilistic statements that support the inferred constraint. This function is essential because result of non-monotonic probabilistic reasoning can be very obscure and counterintuitive. A number of such simple yet confusing entailments will be demonstrated.

[1] http://www2.cs.man.ac.uk/klinovp/pronto/brc /cancer_cc.owl

3. PROBABILISTIC MODELING AND BREAST CANCER RISK ASSESSMENT ONTOLOGY

Probabilistic modeling will be demonstrated for the BRCA domain. The ontology will illustrate how a statistical background knowledge can be incorporated into an OWL ontology and then used for the risk assessment. The latter can be formulated as probabilistic entailment. That reduction is simplified by distinguishing evidence and conclusion categories among OWL classes.

As usual in P-$\mathcal{SHIQ}$(D), the ontology is split on the classical and probabilistic parts. The classical part is the OWL ontology that contains classes describing different categories of women. It is partitioned on evidence and conclusion subtaxonomies:

- Evidence subtaxonomy contains subclasses of $WomanWithRiskFactors$. They are used to represent risk factors such as age, ethnicity, etc.
- Conclusion subtaxonomy contains subclasses of $WomanUnderBRCRisk$. They are used to model breast cancer risks, such as lifetime risk, short term risk or relative risk.

Such separation might turn out to be typical for probabilistic ontologies, especially if applications are concerned with uncertain classification. In that case evidence classes represents probabilistic characteristics of objects whereas conclusion classes - classification categories. Classification can then be reduced to computing probabilities that objects fall into certain conclusion categories given their membership in evidence classes. This approach is followed for thr BRCA problem.

The ontology heavily uses the *overriding* feature of P-$\mathcal{SHIQ}$(D) which allows to override the effect of more generic probabilistic statements by more specific ones. For example, if no information is available about Ann, her lifetime risk would be determined by the statement (WomenUnderAbsoluteBRCRisk|Women)[0,0.123] which applies to all women. But this constraint will be overriden by a new one if more specific information becomes available, such as age, family history, etc.

It will be shown how to express various dependencies between risk factors. One possibility is to represent how the presence of one risk factor allows to guess on the presence of others. This is the principal way to use *inferred* risk factors, i.e., those unknown to a woman. For example, it is known that Ashkenazi Jews are more likely to develop BRCA gene mutation [3].

Finally, the ontology contains a number of ABox axioms that represent risk factors for specific individuals. The motivation is that while the generic probabilistic model that provides all the necessary statistics can be developed and maintained by a central cancer research institute, individual women can supply the knowledge about the risk factors that are known to them, e.g., age. It will also be shown how to express uncertainty in having some particular risk factor.

The modeling described above is necessary to reduce the problem of assessing breast cancer risks to the standard lexicographic entailment implemented in Pronto. Risk assessment for a particular woman corresponds to the entailment of an ABox constraint. For example, $(WomanWithBRCInLongTerm|\top)[0.6, 0.8]$ implies that some woman's risk of developing cancer in life time is 60%-80%. The reasoning will be demonstrated on a number of test probabilistic individuals.

It will also be presented how Pronto justifies the results of the risk assessment by generating the *explanations* for the entailments. In particular, it can retrieve exactly those risk factors and generic statistical axioms that caused the inference for a particular woman and filter our all the irrelevant risk factors. In addition to being useful for end users, this capability can aid the model developers in testing the accuracy and adequacy of their model.

Finally, the demo will reveal some pitfalls of default probabilistic reasoning by presenting seemingly unobvious, yet sound entailments. This will provide a better insight into the nature of probabilistic reasoning and also demonstrate the need of explanations.

4. SUMMARY

The demo does not pretend to cover all the aspects of default probabilistic reasoning in the Semantic Web. However, it is expected that the attendees will learn the advantages of modeling uncertain knowledge *inside* OWL ontologies. One of the goals is to present it as an alternative to more traditional Bayesian approaches.

The demo will also serve as an addition to the research paper that explains probabilistic reasoning and its evaluation in in more detail [2]. It is expected that the demo will be useful as the up-to-date presentation of the ongoing work focused on the scalability and performance improvements.

5. REFERENCES

[1] P. Klinov. Pronto: a non-monotonic probabilistic description logic reasoner. In *System Demo at European Semantic Web Conference*, 2008.

[2] P. Klinov and B. Parsia. Optimization and evaluation of reasoning in probabilistic description logic: Towards a systematic approach. Accepted to International Semantic Web Conference, 2008.

[3] S. G. Komen. Breast cancer risk factors table, 2007. http://cms.komen.org/Komen/AboutBreastCancer/.

[4] T. Lukasiewicz. Expressive probabilistic description logics. *Artificial Intelligence*, 172(6-7):852–883, 2008.

[5] S. Sirin, B. Parsia, B. C. Grau, A. Kalyanpur, and Y. Katz. Pellet: A practical OWL-DL reasoner. Technical Report CS 4766, University of Maryland, College Park, MD, 2005.