Statistical Reasoning – A Foundation for Semantic Web Reasoning

Shashi Kant • Evangelos Mamas
Massachusetts Institute of Technology
Cambridge, MA 02139
skant@sloan.mit.edu • emamas@sloan.mit.edu

Abstract

There has been considerable debate as to the merits and the applicability of probabilistic or statistical reasoning to Semantic Web. Much of this debate seems to have centered on the applicability of statistical methods in a supposedly deterministic setting. In this paper, we argue that statistical reasoning ("reasoning with uncertainty") need not be a substitute for traditional Description Logic (DL) / First-Order Logic (FOL) reasoning, instead statistical methods can serve as a complement to logic-based reasoning systems in two ways: (i) Offer a meta-reasoning (or audit) mechanism to validate logical reasoning, and (ii) Act as a "filler" where Ontological information either does not exist, or is insufficient to reason conclusively.

1 Introduction

Much of the Semantic Web effort has focused on the design and development of Ontologies and related technologies. This approach presupposes that a critical mass of Ontologies will exist that can sufficiently and accurately respond to reasoning queries. As Sir Tim Berners-Lee puts it [Berners-Lee, 1998]: "*The choice of classical logic for the Semantic web is not an arbitrary choice among equals. Classical logic is the only way that inference can scale across the web*."

However, a pure logic-based approach looks increasingly implausible given the paucity of Ontologies and the difficulty in constructing and maintaining Ontologies. Just like the World Wide Web (WWW) had a ready and mature platform to run on i.e. the Internet - which had been in existence for a long time prior to the emergence of the WWW, we feel that the Semantic Web needs an underlying platform, upon which Ontologies can function and interoperate.

We argue that this platform should be a web of statistical "metadata" – which expresses semantic relations in probabilistic terms. Such systems (e.g. Bayesian Networks, Probabilistic Relational Models) have also been in existence for a while and are used in various Machine Learning and AI applications such as Machine Vision, Speech Recognition, and Robotics etc. The Semantic Web would do well to re-use some of these efforts in building this underlying framework.

2 Ontologies and Probabilistic Models

We introduce the notion that Probabilistic Graph Models (PGM) or Bayesian Networks can be viewed as *fuzzy* Ontologies; conversely an Ontology can be viewed as a *crisper* Bayesian Networks. In our proposed architecture, there may not be a clear dividing line between them. A good way of visualizing this relation would be to view Ontologies and Bayesian Networks as ships floating in a sea of statistical "metadata". We use this metaphor to describe the notion that the sea of statistical metadata fills-in the gaps between the islands of Ontologies. Lately there have been some efforts to develop Probabilistic Ontologies by annotating OWL or RDF Ontologies with probabilistic information e.g. BayesOWL [Ding,Peng 2004]. We argue against this approach, and suggest that probabilistic and logic-based reasoning approaches should be viewed as orthogonal to each other. It makes most sense to keep the Ontological information separate from the statistical data, along the lines of how the WWW operates - in which an HTML page *links* to a "FTP" site or a "mailto" to an email hyperlink and the necessary protocols invoked only when clicked.

Figure 1 illustrates a hierarchical mechanism of aligned Ontologies and Bayesian Networks. At the very top are the top-level Ontologies on which there is general agreement and acceptance, at the bottom are the fuzzier, grayer-scale Bayesian Networks which represent relations between resources using probabilistic mechanisms.

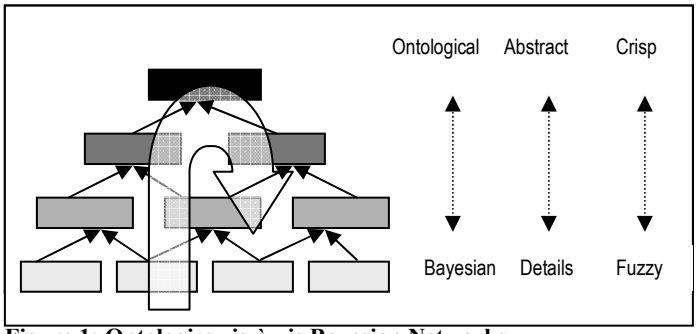

Figure 1: Ontologies vis-à-vis Bayesian Networks

We suggest, that probabilistic (or statistical) information be encoded using any of the widely accepted Bayesian Interchange Formats such as XML-BIF [Cozman, 1998], or Microsoft Research's XBN [Microsoft, 1998] or Hugin.net [Jensen, 2004] format. We propose that the Ontological model encapsulate what it is designed for - expressing logical relations between resources, and the probabilistic model express the statistical relation between them. We do not see a need to mix-and-match as they offer very different views on the same information-set and are perceptually orthogonal.

3 A Hybrid Reasoning Model

Reasoning using Ontologies is based on predicate logic and belongs in the classical tradition of monotonic deductive reasoning i.e. propositions are either true or false. But this proposed framework provides a mechanism for handling fuzzier, incomplete and inaccurate inputs. In this model, reasoning can be performed using a "bottom-up" approach where a query unanswered by a pure Ontological match is extended further up the hierarchy (Fig 1.) until all required information is found. An adjunct application might be to validate traditional reasoning with a mathematical confidence level (meta-reasoning).

Some examples of the reasoning activities possible using this system are:

1. **Deductive Reasoning**: Deductive reasoning allows a system to deduce information given a set of (possibly incomplete and erroneous) information. For example, it can *deduce* that the best course to learn "Machine Vision", "Genomics" and "Political Science" at MIT is *most probably* "6.804J Computational Cognitive Science" even though the course does not directly teach Political Science. It is making a best-guess fit for the requirements [OCW, 2005].
2. **Abductive Reasoning**: Abductive reasoning allows a system to infer the possible causes for a certain effect. For example, the possible courses for learning Artificial Intelligence at MIT are 6.803, 6.825 etc. This is the equivalent of diagnostic reasoning in Bayesian Networks [OCW, 2005].
3. **Monotonic reasoning, non-monotonic reasoning and default values**: Traditional DL-based Ontologies can represent information for monotonic reasoning. For example, one might declare that Universities in the US have a GPA scale of 4.0, but MIT uses a 5.0 GPA scale – so the system *monotonically* cannot reason with that information unless it has been explicitly encoded.

This kind of *non-monotonic* reasoning is possible with the proposed approach.

4 Conclusion

"Reasoning with Uncertainty" is probably a misnomer to describe the efforts required in this area - a more appropriate phraseology would be "reasoning without certainty". While the difference may seem pedantic, the underlying notion is that "uncertainty" is not a state unto itself, but merely the absence of certainty. In a Semantic Web sense, it is a state where Ontological information is non-existent, incomplete or inconclusive. Statistical reasoning could therefore be the bedrock upon which DL/FOL based querying and reasoning can be performed.

This means that the semantic web can operate in areas currently out-of-bounds because of a lack of Ontological information. We therefore hypothesize that statistical "metadata" could be the building-block of the Semantic Web leading to better and more accurate reasoning mechanisms.

5 Acknowledgments

We would like to acknowledge the gracious help and support of the members of the World Wide Web Consortium (W3C) and faculty of MIT-CSAIL for their help and support. We would especially like to thank Ralph Swick and Sir Tim Berners-Lee for their critique and feedback.

6 References

[Berners-Lee, 1998] Tim Berners-Lee. Axioms of Web Architecture: n, 1998, Available at: http://www.w3.org/DesignIssues/Rules.html, Accessed on June 20, 2005.

[Cozman, 1998] Fabio Gagliardi Cozman, The Interchange Format for Bayesian Networks, 1998, Available at: http://www-2.cs.cmu.edu/afs/cs/user/fgcozman/www/Research/InterchangeFormat/, Accessed on June 23, 2005.

[Microsoft, 1998] Microsoft Research, XML Belief Network File Format: Main Page, 1998, Available at: http://research.microsoft.com/dtas/bnformat/ Accessed on June 23, 2005.

[Jensen, 2004] Finn V. Jensen, A Brief Overview of the Three Main Paradigms of Expert Systems, Available at: http://developer.hugin.com/Getting_Started/Paradigms/, Accessed on June 23, 2005.

[OCW, 2005] MIT Open Courseware, Massachusetts Institute of Technology, Available at: http://ocw.mit.edu, Accessed on June 23, 2005.

[Ding, Peng, 2004] Zhongli Ding and Yun Peng. A Probabilistic Extension to Ontology Language OWL, in *Proceedings of the 37th Hawaii International Conference on System Sciences* - 2004