A Method for Modeling Uncertainty in Semantic Web Taxonomies

Markus Holi and Eero Hyvönen
University of Helsinki, Helsinki Institute for Information Technology (HIIT)
P.O. Box 26, 00014 UNIVERSITY OF HELSINKI, FINLAND
http://www.cs.helsinki.fi/group/seco/
firstname.lastname@cs.helsinki.fi

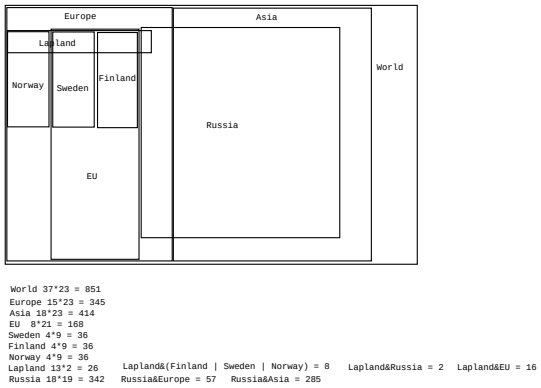

Figure 1: A Venn diagram illustrating countries, areas, their overlap, and size in the world.

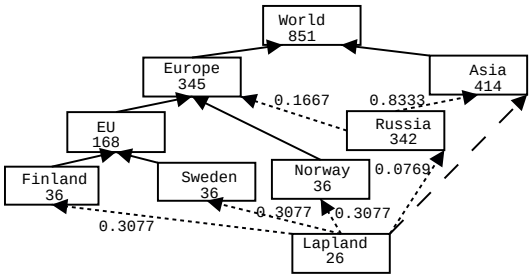

Figure 2: The taxonomy corresponding to the Venn diagram of figure 1.

ABSTRACT

We present a method for representing and reasoning with uncertainty in RDF(S) and OWL ontologies based on Bayesian networks.

Categories and Subject Descriptors: I.2.4 Artificial Intelligence: Knowledge Representation Formalisms and Methods

General Terms: Design

Keywords: Semantic Web, ontology, uncertainty

1. UNCERTAINTY IN ONTOLOGIES

Taxonomical concept hierarchies constitute an important part of the RDF(S)[1] and OWL[2] ontologies used on the semantic web. For example, subsumption hierarchies based on the *subClassOf* or *partOf* properties are widely used. In the real world, concepts are not always subsumed by each other, and cannot always be organized in crisp subsumption hierarchies. Many concepts only partly overlap each other. See, for example, the Venn diagram of figure 1 illustrating various countries and areas in the world. A crisp *partOf* meronymy cannot express the simple fact that Lapland partially overlaps Finland, Sweden, Norway, and Russia, nor quantify the overlap and the coverage of the areas involved.

[1]http://www.w3.org/TR/rdf-schema/
[2]http://www.w3.org/TR/2003/CR-owl-guide-20030818/

WWW2004, May 17–22, 2004, New York, New York, USA.
ACM 1-58113-912-8/04/0005.

Semantic web ontologies are based on crisp logic and do not usually provide well-defined means for expressing degrees of subsumption. To address this foundational problem, this paper presents a new probabilistic method to model conceptual overlap in taxonomies, and an algorithm to compute the overlap between a selected concept and the other concepts of a taxonomy. Our approach can be applied, for example, to sorting hits in an ontology based search engine. The degree of overlap can also be used as a measure of semantic distance between concepts.

In the following, a graphical notation is first presented by which partial subsumption and concepts can be represented in a quantified form. The notation can be represented easily in RDF(S). Then a method for computing degrees of overlap between the concepts of a taxonomy is presented. Overlap is quantified by transforming the taxonomy first into a Bayesian network [4].

2. REPRESENTING OVERLAP

In RDFS and OWL a concept class refers to a set of individuals. Subsumption reduces essentially into the subset relationship between the sets corresponding to classes. A taxonomy is therefore a set of sets. It can be represented, e.g., as a Venn diagram.

We have developed a simple graph notation for representing uncertainty and overlap in taxonomies. Here concepts are nodes, solid directed arcs denote crisp subsumption, dashed arrows disjointness between concepts, and dotted arrows quantified partial subsumption. The values attached to dotted arcs emerging from a concept node of must sum up to 1. Intuitively, the arcs constitute a partition of the concept. For example, figure 2 depicts the meronymy of figure 1. The graph notation is complete in the sense that any Venn diagram can be represented by it.

This graph notation is transformed into an RDF(S) ontology easily in the following way: Concepts are represented either as classes

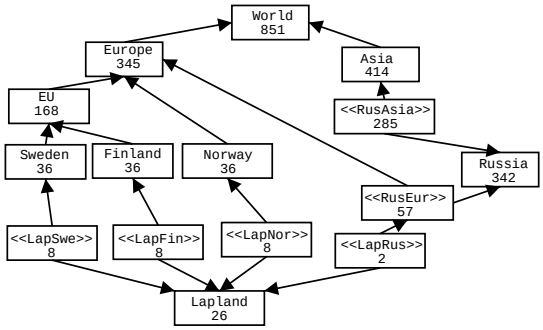

Figure 3: The taxonomy of figure 2 transformed into the *solid path structure* (Bayesian network). The original partial inclusions of Lapland and Russia is transformed into crisp subsumption by using middle concepts. Note that disjoint concepts are d-separated.

or as instances. Disjointness (dashed arc) is represented by a special property. Partial inclusion, i.e., a dotted arc, is represented as an instance with three properties: subject (meronym), object (holonym), and overlap, where the value of overlap quantifies the amount of overlap.

3. COMPUTING OVERLAPS

Given a taxonomy we want to know how much the concepts have in common, i.e., overlap with each other. For example, assume that one is interested in a concept A. We want a method to evaluate, how much the other concepts in the taxonomy have in common with A. A is called the *selected* concept and the evaluated concepts are called *referred* concepts. Let B be a *referred* concept. The question of how much B has in common with A can be quantified in well-defined sense in terms of the conditional probability $P(B|A = true)$. This probability is based on the set theoretic structure of the taxonomy.

In theory, the conditional probability can be computed directly from the Venn diagram. In practice, this is complicated, inefficient, and the Venn diagram may not be available. To solve the problem, we have developed an algorithm for transforming the RDF(S) graph into a Bayesian network. After this, the efficient evidence propagation algorithms developed for Bayesian networks can be used for computing the needed probabilities. We briefly describe next how this can be done.

The overlap value o between concepts A (selected) and B (referred) is $o = \frac{n(A \cap B)}{n(B)}$, where $n(.)$ denotes the mass of a concept. In figure 1, the sizes of the geographical areas are used as the mass values. Computing overlaps is easiest when there are only solid arcs, i.e., complete subsumption relation, between concepts. To exploit this simple case, the taxonomy is first transformed into a *solid path structure*, in which subsumption is the only relation between concepts. This is done according to the following principle:

TRANSFORMATION PRINCIPLE 1. *Let A be the direct partial meronym of B with the overlap value o. In the transformed structure the partial subsumption is replaced by an additional* middle concept, *that represents $A \cap B$. It is marked to be the meronym of both A and B, and it gets the mass $n(A \cap B)$.*

For example, the taxonomy of figure 2 is transformed into the solid path structure of figure 3. Now the overlap between two concepts can be calculated according to the following principle:

OVERLAP CALCULATION PRINCIPLE 1. *Let A (selected) and B (referred) be concepts in a* solid path structure. *If B is subsumed by A, then overlap $o = 1$. If B is not subsumed by A, then all the concepts subsumed by A are marked as selected, as constituents of A. If C is the collection of the selected concepts that are also subsumed by B, then $o = \frac{n(\bigcup C)}{n(B)}$. If $C = \emptyset$, then $o = 0$.*

As can be seen, the topology of the solid path structure is well-suited to be used as a Bayesian network. Let A (selected) and B (referred) be concepts with the overlap value o. Probabilistically A and B are boolean random variables, and $P(B|A = true) = o$.

The conditional probability table (CPT) for each node A can be constructed in the following way: 1) Go through all the value combinations of the parents of A. 2) The *true* value in the CPT for a given entry is $\frac{n(\bigcup TrueStateVariables)}{n(A)}$. If A has no parents, then $P(A = true) = \lambda$, where λ is a very small non-zero probability, because we want the posterior probabilities to result from only conditional probabilities (overlap). When we give to the Bayesian evidence propagation algorithm the selected concept and all the concepts subsumed by it as evidence, the algorithm returns the overlap values as posterior probabilities.

To validate and evaluate the method, we have implemented the transformation algorithm from RDF(S) to a Bayesian net format using SWI-Prolog[3] and its RDF parser. Hugin Lite 6.3[4] was then used as the Bayesian reasoner through its Java API.

4. DISCUSSION

We chose to use crisp set theory and Bayesian networks, because of the sound mathematical foundations they offer. The calculations are simple, but still enable the representation of overlap and vague subsumption between concepts. The Bayesian representation of a taxonomy is useful not only for the matching problem we discussed, but also to other reasoning tasks.

The problem of representing uncertain or vague inclusion in ontologies and taxonomies has been tackled by using methods of fuzzy logic [1, 2], roughs sets [5]. The work that is closest to ours is that of Ding et al. [3]. They present principles and methods to convert an OWL ontology into a Bayesian network. Their transformation method is, however, quite different from ours, and the semantics of the transformation is not explicitly specified.

Acknowledgments

Our research was funded mainly by the National Technology Agency Tekes.

5. REFERENCES

[1] G. Akrivas, M. Wallace, G. Andreou, G. Stamou, and S. Kollias. Context - sensitive semantic query expansion. In *Proceedings of the IEEE International Conferrence on Artificial Intelligence Systems (ICAIS)*, 2002.

[2] R.A. Angryk and F.E. Petry. Consistent fuzzy concept hierarchies for attribute generalization. In *Proceeding of the IASTED International Conference on Information and Knowledge Sharing (IKS' 03)*, 2003.

[3] Z. Ding and Y. Peng. A probabilistic extension to ontology language owl. In *Proceedings of the Hawai'i Internationa Conference on System Sciences*, 2004.

[4] F. V. Finin and F. B. Finin. *Bayesian Networks and Decision Graphs*. Springer-Verlag, 2001.

[5] H. Stuckenschmidt and U. Visser. Semantic translation based on approximate re-classifi cation. In *Proceedings of the 'Semantic Approximation, Granularity and Vagueness' Workshop*, 2000.

[3] http://www.swi-prolog.org/

[4] http://www.hugin.com/