Modeling Degrees of Conceptual Overlap in Semantic Web Ontologies

Markus Holi and Eero Hyvönen
Helsinki University of Technology, Media Technology,
Helsinki Institute for Information Technology (HIIT), and University of Helsinki
P.O. Box 5500, FI-02015 TKK, FINLAND
http://www.cs.helsinki.fi/group/seco/
email: firstname.lastname@tkk.fi

Abstract

Semantic Web ontologies are based on crisp logic and do not provide well-defined means for expressing uncertainty. We present a new probabilistic method to approach the problem. In our method, degrees of subsumption, i.e., overlap between concepts can be modeled and computed efficiently using Bayesian networks based on RDF(S) ontologies.

1 Introduction

Ontologies are based on crisp logic. In the real world, however, relations between entities often include subtleties that are difficult to express in crisp ontologies. RDFS [rdf, 2004] and OWL [owl, 2003] do not provide standard ways to express partial overlap and degrees of overlap in general.

This paper presents a method for modeling degrees of overlap between concepts. In the following we first introduce the principles of our method. Then a notation that enables the representation of degrees of overlap between concepts in an ontology is presented after which a method for doing inferences based on the notation will be described. For a more detailed presentation of the method see [Holi, 2004].

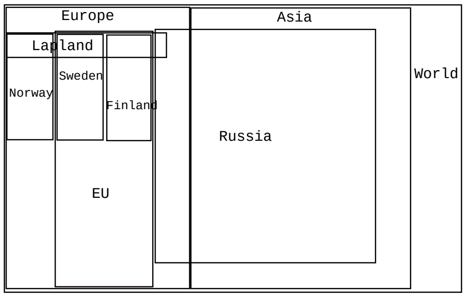
2 Modeling Uncertainty in Ontologies

Figure 1 illustrates various countries and areas in the world. There are important properties in the figure, that are not modeled in a crisp partonomy. For example, EU is a bigger part of Europe than Lapland, and Russia partly overlaps Europe and Asia.

Our method enables the representation of overlap in concept hierarchies, including class hierarchies and partonomies, and the computation of overlap between a *selected* concept and every other, i.e. *referred* concept in the hierarchy. The overlap value is defined as follows:

$$Overlap = \frac{|Selected \cap Referred|}{|Referred|} \in [0, 1].$$

Intuitively, the overlap value has the following meaning: The value is 0 for disjoint concepts (e.g., Lapland and Asia) and 1, if the referred concept is subsumed by the selected one. High values lesser than one imply, that the meaning of the selected concept approaches the meaning of the referred one.



```
World 37*23 = 851
Europe 15*23 = 345
Asia 18*23 = 414
EU  8*21 = 168
Sweden 4*9 = 36
Finland 4*9 = 36
Norway 4*9 = 36
Lapland 13*2 = 26  Lapland&(Finland | Sweden | Norway) = 8
Lapland&EU = 16 Lapland&Russia = 2
Russia 18*19 = 342  Russia&Europe = 57  Russia&Asia = 285
```

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

3 Representing Overlap

A concept hierarchy can be viewed as a set of sets and can be represented by a Venn diagram.

If A and B are sets, then A must be in one of the following relationships to B.

1. A is a subset of B, i.e. $A \subseteq B$.
2. A partially overlaps B, i.e. $\exists x, y : (x \in A \wedge x \in B) \wedge (y \in A \wedge y \notin B)$.
3. A is disjoint from B, i.e. $A \cap B = \emptyset$.

Based on these relations, we have developed a simple graph notation for representing overlap in a concept hierarchy as an acyclic *overlap graph*. Here concepts are nodes, and a number called *mass* is attached to each node. The mass of concept A is a measure of the size of the set corresponding to A, i.e. $m(A) = |s(A)|$, where $s(A)$ is the set corresponding to A. A solid directed arc from concept A to B denotes crisp subsumption $s(A) \subseteq s(B)$, a dashed arrow denotes disjointness $s(A) \cap s(B) = \emptyset$, and a dotted arrow represents quantified partial subsumption between concepts, which means that the concepts partially overlap in the Venn diagram. The amount of overlap is represented by the *partial overlap value* $p = \frac{|s(A) \cap s(B)|}{|s(A)|}$.

In addition to the quantities attached to the dotted arrows, also the other arrow types have implicit overlap values. The overlap value of a solid arc is 1 (crisp subsumption) and the value of a dashed arc is 0 (disjointness). The quantities of the arcs emerging from a concept must sum up to 1. This means that either only one solid arc can emerge from a node or several dotted arcs (partial overlap). In both cases, additional dashed arcs can be used (disjointness). Intuitively, the outgoing arcs constitute a quantified partition of the concept. Thus, the dotted arrows emerging from a concept must always point to concepts that are mutually disjoint with each other.

Notice that if two concepts overlap, there must be a directed (solid or dotted) path between them. If the path includes dotted arrows, then (possible) disjointness between the concepts must be expressed explicitly using the disjointness relation. If the directed path is solid, then the concepts necessarily overlap.

4 Computing the Overlaps

Computing the overlap is easiest when there are only solid arcs, i.e. complete subsumption relation between concepts. If there is a directed solid path from A (selected) to B (referred), then overlap $o = \frac{|s(A) \cap s(B)|}{|s(B)|} = \frac{m(A)}{m(B)}$. If there is a mixed path then the computation is not as simple. To exploit the simple case we transform the graph into a solid path structure according to the following principle:

Transformation Principle 1 *Let A be the direct partial subconcept of B with overlap value o. In the solid path structure the partial subsumption is replaced by an additional* middle concept, *that represents $s(A) \cap s(B)$. It is marked to be the complete subconcept of both A and B, and its mass is $o \cdot m(A)$.*

If A is the selected concept and B is the referred one, then the overlap value o can be interpreted as the conditional probability

$$P(B' = true | A' = true) = \frac{|s(A) \cap s(B)|}{|s(B)|} = o, \qquad (1)$$

where $s(A)$ and $s(B)$ are the sets corresponding to the concepts A and B. A' and B' are boolean random variables such that the value $true$ means that the corresponding concept is a match to the query, i.e, the concept in question is of interest to the user.

Based on the above, we chose to use the solid path structure as a Bayesian network topology. In the Bayesian network the boolean random variable X' replaces the concept X of the solid path structure. The efficient evidence propagation algorithms developed for Bayesian networks [Finin and Finin, 2001] to take care of the overlap computations.

The joint probability distribution of the Bayesian network is defined by conditional probability tables (CPT) $P(A'|B'_1, B'_2, \ldots B'_n)$ for nodes with parents $B'_i, i = 1 \ldots n$, and by prior marginal probabilities set for nodes without parents. The CPT $P(A'|B'_1, B'_2, \ldots B'_n)$ for a node A' can be constructed by enumerating the value combinations (true/false) of the parents $B'_i, i = 1 \ldots n$, and by assigning:

$$P(A' = true | B'_1 = b_1, \ldots B'_n = b_n) = \frac{\sum_{i \in \{i : b_i = true\}} m(B_i)}{m(A)} \qquad (2)$$

The value for the complementary case $P(A' = false | B'_1 = b_1, \ldots B'_n = b_n)$ is obtained simply by subtracting from 1.

By instantiating the nodes corresponding to the selected concept and the concepts subsumed by it as evidence (their values are set "true"), the propagation algorithm returns the overlap values as posterior probabilities of nodes. The query results can then be ranked according to these posterior probabilities.

5 Discussion

Overlap graphs are simple and can be represented in RDF(S) easily. Using the notation does not require knowledge of probability theory. The concepts can be quantified automatically, based on data records annotated according to the ontology, for example.

The problem of representing uncertainty in ontologies has been tackled previously by using methods of fuzzy logic, rough sets [Stuckenschmidt and Visser, 2000] and Bayesian networks [Ding and Peng, 2004; Gu and H.K. Pung, 2004].

Acknowledgments

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

References

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

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

[Gu and H.K. Pung, 2004] T. Gu and D.Q. Zhang H.K. Pung. A bayesian approach for dealing with uncertain contexts. In *Advances in Pervasive Computing*, 2004.

[Holi, 2004] M. Holi. Modeling uncertainty in semantic web taxonomies, 2004. Master of Science Thesis. Department of Computer Science, University of Helsinki, http://ethesis.helsinki.fi/julkaisut/mat/tieto/pg/holi/.

[owl, 2003] *OWL Web Ontology Language Guide*, 2003. http://www.w3.org/TR/2003/CR-owl-guide-20030818/.

[rdf, 2004] *RDF Vocabulary Description Language 1.0: RDF Schema*, 2004. http://www.w3.org/TR/rdf-schema/.

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