Modeling Degrees of Conceptual Overlap in Semantic Web Ontologies

Markus Holi and Eero Hyvönen

Helsinki Institute for Information Technology (HIIT),
Helsinki University of Technology, Media Technology 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. Information retrieval systems have to deal with uncertain knowledge and query results should reflect this uncertainty in some manner. However, 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. Degrees of overlap indicate how well an individual data item matches the query concept, which can be used as a well-defined measure of relevance in information retrieval tasks.

1 Ontologies and Information Retrieval

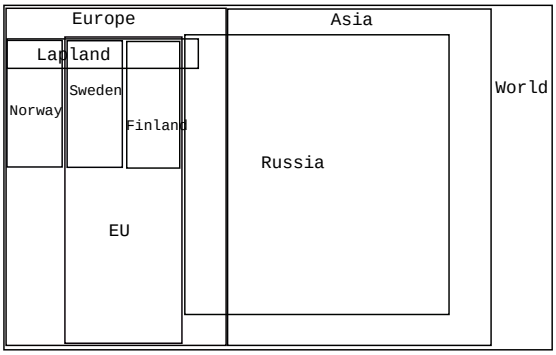
A key reason for using ontologies in information retrieval systems, is that they enable the representation of background knowledge about a domain in a machine understandable format. Humans use background knowledge heavily in information retrieval tasks [7]. For example, if a person is searching for documents about Europe she will use her background knowledge about European countries in the task. She will find a document about Germany relevant even if the word 'Europe' is not mentioned in it. With the help of an appropriate geographical ontology also an information retrieval system could easily make the above inference. Ontologies have in fact been used in a number of information retrieval system in recent years [15, 10, 11].

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. For example, consider the case of representing the relationships between geohgraphical areas in the world with a partonomy where each concept represents an area in the world. We will run into difficulties, because RDFS [2] and OWL [1] do not provide standard ways to express the facts that Germany covers a much larger part of the area of Europe than Andorra, or that Russia is a part of both Asia and Europe, for example. In addition, the information system itself can be a source of uncertainty, as the indexing of the documents is often inexact. These representational shortcomings could hinder the performance of the information retrieval system and produce wrong search results.

This paper presents a new method to approach the above problems. The method is based on the modeling of 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. Then our implementation of the method is discussed, and finally conclusions are drawn and related work discussed. These paper is an extended version of [9]. For a more detail presentation of the method see [8].

2 Modeling Uncertainty in Ontologies




```
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
```

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

The Venn diagram of figure 1 illustrates some countries and areas in the world. A crisp partonomy cannot represent the partial overlap between the geographical area Lapland and the countries Finland, Sweden, Norway, and Russia, for example. Furthermore, it is not possible to quantify the coverage and the overlap of the areas involved.

According to figure 1, the size of Lapland is 26 units, and the size of Finland is 36 units. The size of the overlapping area between Finland and Lapland is 8 units. Thus, $8/26$ of Lapland belongs to Finland, and $8/36$ of Finland belongs to Lapland. On the other hand, Lapland and Asia do not have any overlapping area, thus no part (0) of Laplan is part of Asia, and no part of Asia is part of Lapland. If we want a partonomy to be an accurate representation of the 'map' of figure 1, there should be a way to make this kind of inferences based on the partonomy.

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:

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

An *overlap table* is created for the selected concept. The overlap table can be created for every concept of a hierarchy.

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.

The overlap value between the selected concept (e.g. Lapland) and the referred concept (e.g. Finland) can in fact be written as the conditional probability $P(Finland'|Lapland')$ whose interpretation is the following: If a person is interested in data records about Lapland, what is the probability that the annotation "Finland" matches her query? X' is a binary random variable such that $X' = true$ means that the annotation "X" matches the query, and $X' = false$ means that "X" is not a match. This conditional probability interpretation of overlap values will be used in section 4 of this paper.

It is mathematically easy to compute the overlap tables, if a Venn diagram (the sets) is known. In practice, the Venn diagram may be difficult to create from the modeling view point, and computing with explicit sets is computationally complicated and inefficient. For these reasons our method calculates the overlap values from a hierarchical representation of the Venn diagram.

Our method consists of two parts. First, a graphical notation by which partial subsumption and concepts can be represented in a quantified form. The notation can be represented easily in RDF(S). Second, a method for computing degrees of overlap between the concepts of a hierarchy. Overlap is quantified by transforming the concept hierarchy first into a Bayesian network [4].

3 Representing Overlap

As shown in the above section, a partonomy can be represented as a Venn diagram. Also class hierarchies, i.e. taxonomies can be reprsented in the same manner. This follows from the fact that in RDFS and OWL a class refers to a set of individuals. Subsumption reduces essentially into the subset relationship between the sets corresponding to classes [1]. A taxonomy is therefore a set of sets and can be represented, e.g., 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 uncertainty and 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 mutally 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.

For example, figure 2 depicts the partonomy of figure 1 as an overlap graph. The geographic sizes of the areas are used as masses and the partial overlap values are determined based on the Venn diagram. This graph notation is complete in the sense that any Venn diagram can be represented by it. However, sometimes the accurate representation of a Venn diagram requires the use of auxiliary concepts, which represent results of set operations over named sets, for example $s(A) \setminus s(B)$, where A and B are ordinary concepts.

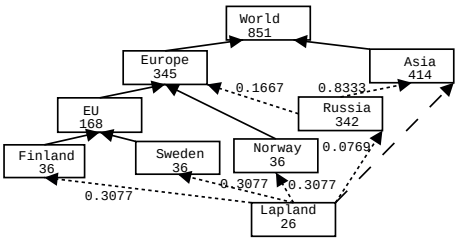


Fig. 2. The taxonomy corresponding to the Venn diagram of figure 1.

4 Solid Path Structure

Our method creates an overlap table for each concept in the concept hierarchy. Computing the overlaps 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 the solid path is directed from B to A, then $o = \frac{|s(A) \cap s(B)|}{|s(B)|} = \frac{m(B)}{m(B)} = 1$. If there is not a directed path between A and B, then $o = \frac{|s(A) \cap s(B)|}{|s(B)|} = \frac{|\emptyset|}{m(B)} = 0$.

If there is a mixed path of solid and dotted arcs between A and B, then the calculation is not as simple. Consider, for example, the relation between $Lapland$ and EU in figure 2. To compute the overlap, we have to follow all the paths emerging from $Lapland$, take into account the disjoint relation between $Lapland$ and $Asia$, and sum up the partial subsumption values somehow.

To exploit the simple solid arc case, a hierarchy with partial overlaps is first transformed into a *solid path structure*, in which crisp subsumption is the only relation between the concepts. The transformation is done 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 $o \cdot m(A)$.*

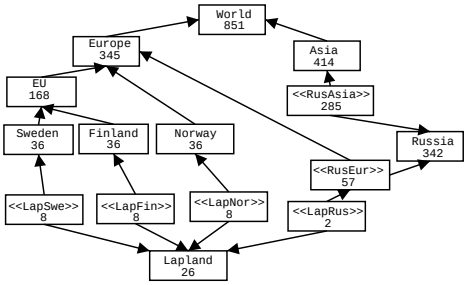


Fig. 3. The taxonomy of figure 2 as a solid path structure.

For example, the partonomy of figure 2 is transformed into the solid path structure of figure 3. The original partial overlaps of Lapland and Russia are transformed into crisp subsumption by using middle concepts. The transformation algorithm processes the overlap graph in a breadth-first manner starting from the root concept. A concept is processed only after all of its super concepts (partial or complete) are processed. Because the graph is acyclic, all the concepts will eventually be processed. For a more detailed description of the transformation algorithm see [8].

5 Computing the Overlaps

Recall from section 2 that if A is the selected concept and B is the referred one, then the overlap value o can be interpreted as the conditional probability

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

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 effi cient evidence propagation algorithms developed for Bayesian networks [4] to take care of the overlap computations. Furthermore, we saw a Bayesian representation of the taxonomy valuable as such.

The joint probability distribution of the Bayesian network is defi ned 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:

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

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. The above formula is based on the above defi nition of conditional probability. The intuition behind the formula is the following. If a user is interested in Sweden and in Finland, in the Bayesian network both Finland and Sweden will be set "true". Thus, the bigger the number of European countries that the user is interested in, the bigger the probability that the annotation "Europe" matches her query, i.e., $P(Europe' = true|Sweden' = true, Finland' = true) > P(Europe' = true|Finland' = true)$.

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 conditional probabilities only, i.e., from the overlap information.

The whole overlap table of a concept can now be determined effi ciently by using the Bayesian network with its conditional and prior probabilities. 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.

First, to be able to easily use the the solid path structure as the topology of the Bayesian network. The CPTs can be calculated directly based on the masses of the concepts. Second, with this defi nition the Bayesian evidence propagation algorithm returns the overlap values readily as posterior propabilities. We experimented with various ways to construct a Bayesian network according to probabilistic interpretations of the Venn diagram. However, none of these constructions answered to our needs as well as the construction described above.

Third, in the solid path structure d-separation indicates disjointness between concepts. We see this as a useful characteristic, because it makes the simultaneous selection of two or more disjointed concepts possible.

6 Implementation

The presented method has been implemented as a proof-of-concept. In the implementation overlap graphs are represented as RDF(S) ontologies in the following way. Concepts are represented as RDFS classes[1] The concept masses are represented using a special *Mass* class. It has two properties, subject and mass that tell the concept resource in question and mass as a numeric value, respectively. The subsumption relation can be implemented with a property of the users choice.

Partial subsumption is implemented by a special *PartialSubsumption* class with three properties: subject, object and overlap. The subject property points to the direct partial subclass, the object to the direct partial superclass, and overlap is the partial overlap value. The disjointness arc is implemented by the disjointFrom property used in OWL.

The input to the system is an RDF(S) ontology, the URI of the root node of the overlap graph, and the URI of the subsumption property used in the ontology. The output is the overlap tables for every concept in the taxonomy extracted from the input RDF(S) ontology. Next, each submodule in the system is discussed briefly.

The *preprocessing* module transforms the taxonomy into a predefined standard form. The *transformation* module implements the transformation algorithm, and defines the CPTs of the resulting Bayesian network. In addition to the Bayesian network, it creates an RDF graph with an identical topology, where nodes are classes and the arcs are represented by the *rdf:subClassOf* property. This graph will be used by the *selection* module that expands the selection to include the concepts subsumed by the selected one, when using the Bayesian network. The *Bayesian reasoner* does the evidence propagation based on the selection and the Bayesian network. The selection and Bayesian reasoner modules are operated in a loop, where each concept in the taxonomy is selected one after the other, and the overlap table is created.

The *preprocessing*, *transformation*, and selection modules are implemented with SWI-Prolog[2]. The Semantic Web package is used. The *Bayesian reaoner* module is implemented in Java, and it uses the Hugin Lite 6.3[3] through its Java API.

7 Discussion

7.1 Lessons Learned

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

[1] Actually, any resources including instances could be used to represent concepts.
[2] http://www.swi-prolog.org/
[3] http://www.hugin.com/

notation enables the representation of any Venn diagram, but there are set structures, which lead to complicated representations.

Such a situation arises, for example, when three or more concepts mutually partially overlap each other. In these situations some auxiliary concepts have to be used. However, we have not met such situations in geospatial ontologies.

The Bayesian network structure that is created with the presented method is only one of the many possibilities. This one was chosen, because it can be used for computing the overlap tables in a most direct manner.

7.2 Related Work

The problem of representing uncertain or vague inclusion in ontologies and taxonomies has been tackled also by using methods of fuzzy logic [21] and rough sets [19, 17]. With the rough sets approach only a rough, egg-yolk representation of the concepts can be created [19]. Fuzzy logic, allows for a more realistic representation of the world.

Straccia [18] presents a fuzzy extension to the description logic SHOIN(D)corresponding to the ontology description language OWL DL. It enables the representation of fuzzy subsumption for example.

Widyantoro and Yen [20] have created a domain-specific search engine called PASS. The system includes an interactive query refinement mechanism to help to find the most appropriate query terms. The system uses a fuzzy ontology of term associations as one of the sources of its knowledge to suggest alternative query terms. The ontology is organized according to narrower-term relations. The ontology is automatically built using information obtained from the system's document collections.

The fuzzy ontology of Widyantoro and Yen is based on a set of documents, and works on that document set. However, our focus is on building taxonomies that can be used, in principle, with any data record set. The automatic creation of ontologies is an interesting issue by itself, but it is not considered in this paper. At the moment, better and richer ontologies can be built by domain specialists than by automated methods.

The fuzzy logic approach is critisized because of the arbitrariness in finding the numeric values needed and mathematical indefiniteness [19]. In addition, the representation of disjointness between concepts of a taxonomy seems to be difficult with the tools of fuzzy logic. For example, the relationships between Lapland, Russia, Europe, and Asia are very easily handled probabilistically, but in a fuzzy logic based taxonomy, this situation seems complicated. There is not a readily available fuzzy logic operation that could determine that if Lapland partly overlaps Russia, and is disjoint from Asia, then the fuzzy inclusion value between Europe and $Lapland \cap Russia$ is 1 even though Russia is only a fuzzy part of Europe.

We chose to use crisp set theory and Bayesian networks, because of the sound mathematical foundations they offer. The set theoretic approach also gives us means to overcome to a large degree the problem of arbtrariness. The calculations are simple, but still enable the representation of overlap and vague subsumption between concepts. The Bayesian network representation of a taxonomy is useful not only for the matching problem we discussed, but can also be used for other reasoning tasks [14].

Ding and Peng [3] present principles and methods to convert an OWL ontology into a Bayesian network. Their methods are based on probabilistic extensions to de-

scription logics [13, 5]. The approach has some differences to ours. First, their aim is to create a method to transform any OWL ontology into a Bayesian network. Our goal is not to transform existing ontologies into Bayesian networks, but to create a method by which overlap between concepts could be represented and computed from a taxonomical structure. However, we designed the overlap graph and its RDF(S) implementation so, that it is possible, quite easily, to convert an existing crisp taxonomy to our extended notation. Second, in the approach of Ding and Peng, probabilistic information must be added to the ontology by the human modeler that needs to know probability theory. In our approach, the taxonomies can be constructed without virtually any knowledge of probability theory or Bayesian networks.

Also other approaches for combining Bayesian networks and ontologies exist. Gu [6] present a Bayesian approach for dealing with uncertain contexts. In this approach probabilistic information is represented using OWL. Probabilities and conditional probabilities are represented using classes constructed for these purposes. Mitra [16] presents a probabilistic ontology mapping tool. In this approach the nodes of the Bayesian network represents matches between pairs of classes in the two ontologies to be mapped. The arrows of the BN are dependencies between matches.

Kauppinen and Hyvönen [12] present a method for modeling partial overlap between versions of a concept that changes over long periods of time. The approach differs from ours in that we are interested in modelling degrees of overlap between different concepts in a single point of time.

8 ACKNOWLEDGEMENTS

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

References

1. *OWL Web Ontology Language Guide*. http://www.w3.org/TR/2003/CR-owl-guide-20030818/.
2. *RDF Vocabulary Description Language 1.0: RDF Schema*. http://www.w3.org/TR/rdf-schema/.
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. R. Giugno and T. Lukasiewicz. P-shoq(d): A probabilistic extension of shoq(d) for probabilistic ontologies in the semantic web. INFSYS Research Report 1843-02-06, Technische Universität Wien, 2002.
6. T. Gu and D.Q. Zhang H.K. Pung. A bayesian approach for dealing with uncertain contexts. In *Advances in Pervasive Computing*, 2004.
7. N. Guarino. Formal ontology in information systems. In *Proceedings of FOIS'98*. IOS Press, 1998.
8. 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/.
9. M. Holi and E. Hyvönen. A method for modeling uncertainty in semantic web taxonomies. In *Proceedings of WWW2004, Alternate Track Papers and Posters*, 2004.

10. E. Hyvönen, M. Junnila, S. Kettula, E. Mäkelä, S. Saarela, M. Salminen, A. Syreeni, A. Valo, and K. Viljanen. Finnish Museums on the Semantic Web. User's perspective on museumfinland. In *Proceedings of Museums and the Web 2004 (MW2004), Arlington, Virginia, USA*, 2004. http://www.archimuse.com/mw2004/papers/hyvonen/ hyvonen.html.
11. E. Hyvönen, A. Valo, K. Viljanen, and M. Holi. Publishing semantic web content as semantically linked HTML pages. In *Proceedings of XML Finland 2003, Kuopio, Finland*, 2003. http://www.cs.helsinki.fi/u/eahyvone/publications/ xmlfinland2003/swehg_article_xmlfi2003.pdf.
12. T. Kauppinen and E. Hyvönen. Geo-spatial reasoning over ontology changes in time. In *Proceedings of IJCAI-2005 Workshop on Spatial and Temporal Reasoning*, 2005.
13. D. Koller, A. Levy, and A. Pfeffer. P-classic: A tractable probabilistic description logic. In *Proceedings of AAAI-97*, 1997.
14. A. Kuenzer, C. Schlick, F. Ohmann, L. Schmidt, and H. Luczak. An empirical study of dynamic bayesian networks for user modeling. In R. Schafer, M.E. Muller, and S.A. Macskassy, editors, *Proc. of the UM'2001 Workshop on Machine Learning for User Modeling*, 2001.
15. K. Mahalingam and M.N. Huhns. Ontology tools for semantic reconciliation in distributed heterogeneous information environments. *Intelligent Automation and Soft Computing*, 1999.
16. P. Mitra, N. Noy, and A.R. Jaiswal. Omen: A probabilistic ontology mapping tool. In *Working Notes of the ISCW-04 Workshop on Meaning Coordination and Negotiation*, 2004.
17. J. Pawlak. Rough sets. *International Journal of Information and Computers*, 1982.
18. U. Straccia. Towards a fuzzy description logic for the semantic web. In *Proceedings of the Second European Semantic Web Conference, ESWC 2005*, 2005.
19. H. Stuckenschmidt and U. Visser. Semantic translation based on approximate reclassification. In *Proceedings of the 'Semantic Approximation, Granularity and Vagueness' Workshop*, 2000.
20. D.H. Widyantoro and J. Yen. A fuzzy ontology-based abstract seawrch engine and its user studies. In *The Proceedings of the 10th IEEE International Conference on Fuzzy Systems*, 2002.
21. L. Zadeh. Fuzzy sets. *Information and Control*, 1965.