Probabilistic Information Retrieval Based on Conceptual Overlap in Semantic Web Ontologies

Markus Holi and Eero Hyvönen

University of Helsinki
Department of Computer Science
http://www.cs.helsinki.fi/group/seco
firstname.lastname@cs.helsinki.fi

The Problem

- Ontologies are used in information retrieval to represent background knowledge about a domain in a machine understandable format.
- Ontologies are based on crisp logic and do not provide means for expressing subtleties in relations between concepts. Hinders the usability of ontologies.
- A Crisp partOf meronymy can not express the situation of the adjacent Venn diagram correctly. No way to quantify the concepts and the overlap between them.

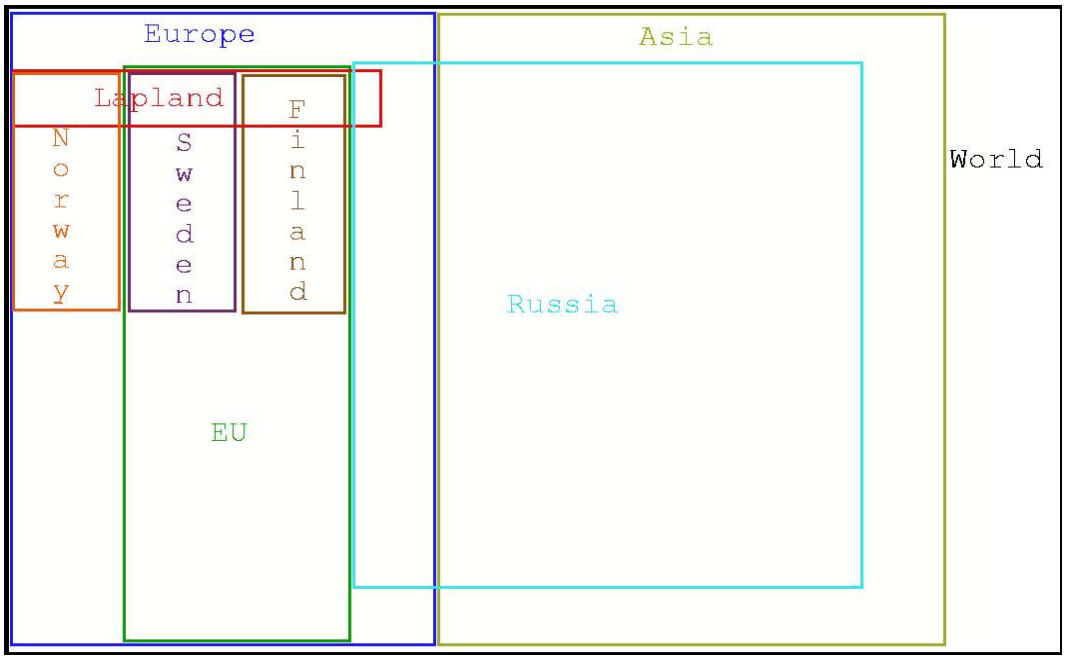


Figure 1: A Venn diagram illustrating countries, areas their overlap and size in the world.

Solution Approach

- The goal is to create a method for computing the overlaps between a selected concept and the other concepts of a taxonomy.
  - o = |Referred ∩ Selected| / |Referred|
  - An overlap table for each concept
- Each value depicts the relevance of the selected concept to the referred one
  - Hits can be sorted accordingly in a search engine
- The goal is reached in two steps:

1. Create a notation to represent concepts and degrees of subsumption in quantified form as an overlap graph
2. Compute the overlap values by transforming the graph into a Bayesian network.

Selected	Referred	Value (o)
Lapland	World	0.0306
	Europe	0.0754
	Asia	0
	EU	0.0953
	Norway	0.2222
	Sweden	0.2222
	Finland	0.2222
	Russia	0.0059

Table 1: An overlap table created for Lapland, based on the Venn diagram of figure 1.

The Process

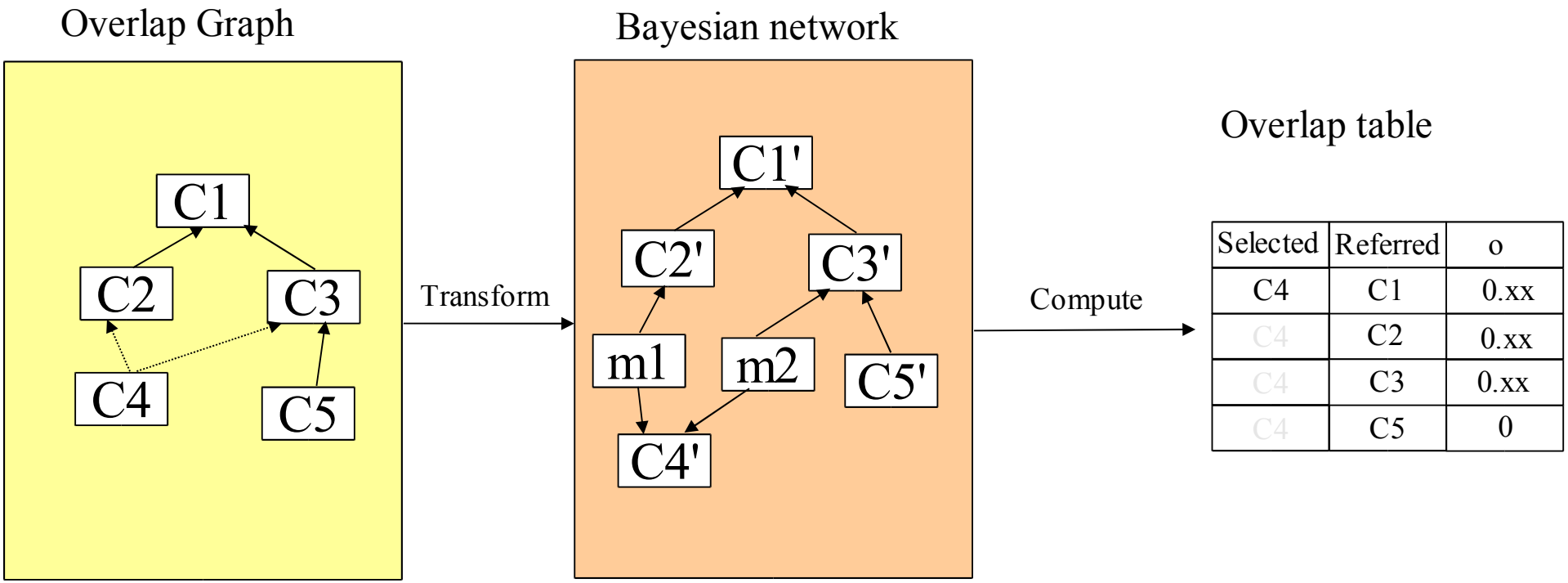


Figure 2: The overlap table computation process.

Representing Overlap

- In RDFS and OWL concept (class) refers to a set of individuals. Subsumption reduces into the subset relationship between the respective sets
  - A taxonomy can be represented as a Venn diagram
- In the graph (DAG), concepts are nodes, black arrows denote crisp subsumption, dashed arrows disjointness, and the dotted arrows quantified partial subsumption.
- Can be easily represented in RDF(S)

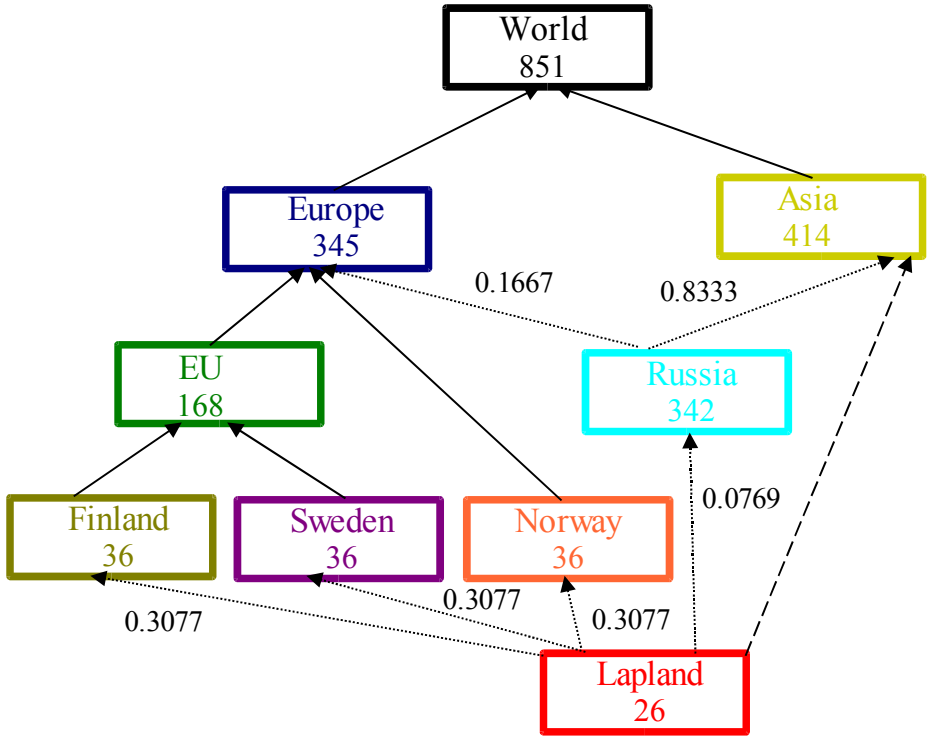


Figure 3: An overlap graph representing the Venn diagram of figure 1.

Computing the Overlaps

- An overlap value between concept A (selected) and B (referred) can be written as the conditional probability P(B'=true |A'=true).

- A' is a binary variable, s.t, A'=true means that the user is interested in concept A. P(B'=true|A'=true) is the probability that the user is interested in documents about concept B when we know that she is interested in concept A.

- For computing the overlaps we transform the overlap graph into a Bayesian network.

Solid Path Structure

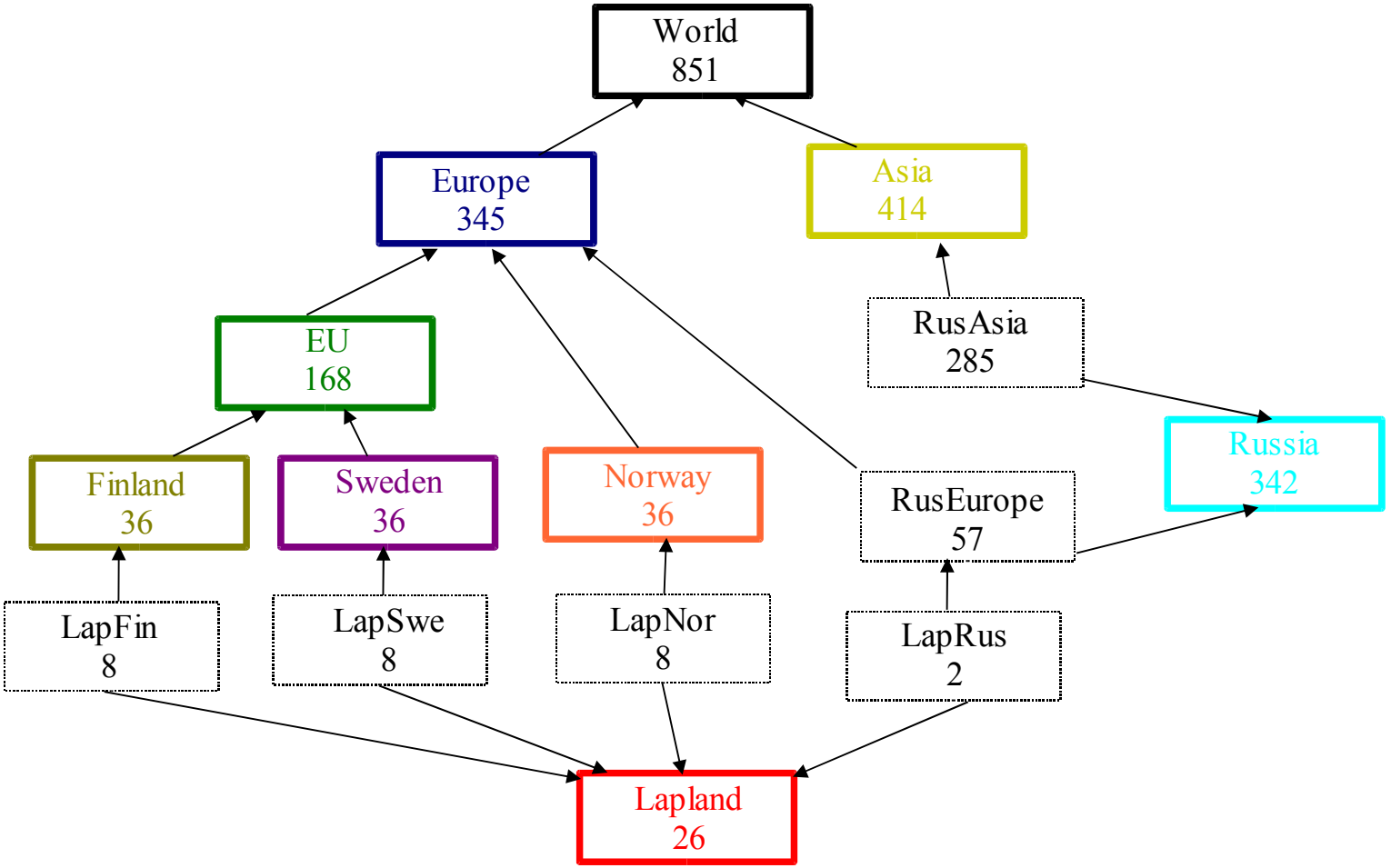


Figure 4: The graph of figure 3 transformed into a solid path structure. The dashed boxes indicate middle concepts.

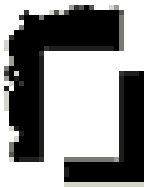
Discussion

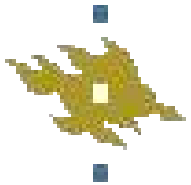
- The problem of representing uncertainty or vagueness in ontologies has been tackled also by using fuzzy logic and rough sets
  - Crisp set theory and Bayesian networks were chosen because of the sound mathematical foundations they offer
  - The calculations are simple but still enable the representation of overlap and degrees of subsumption
  - The Bayesian representation of a taxonomy can be used also in other tasks than in sorting hits in a search engine.
- Open questions and future work
  - How to quantify concept that do not have natural extensions?
  - The method makes ontologies more complicated. What are the consequences of this?
  - Is conceptual overlap a sensible basis for sorting hits in a search engine?
  - The method should be tested in realistic application situations.

Research Consortium

National Board
of Antiquities

Suomen valokuvataiteen museo
Finlands fotografiska museum
The Finnish Museum of Photography





UNIVERSITY
OF HELSINKI

UNIVERSITY
OF TAMPERE

Finnish National
Gallery