Web Search With Document Space Adapted Ontologies

Stein L. Tomassen
IDI, NTNU
Sem Saelandsvei 7-9,
NO-7491 Trondheim, Norway
+ 47 735 94218

stein.l.tomassen@idi.ntnu.no

ABSTRACT

An increasing number of recent information retrieval systems make use of ontologies to help the users clarify their information needs and come up with semantic representations of documents. In this paper, we present an approach that utilizes ontologies to enhance the effectiveness of large-scale search systems for the Web. The ontology concepts are adapted to the domain terminology by computing a feature vector for each concept. We explain how these feature vectors are constructed and finally present some results.

Categories and Subject Descriptors

H.3.3 [**Information Storage And Retrieval**]: Information Search and Retrieval - *information filtering, query formulation, search process, selection process.*

General Terms

Experimentation, Management.

Keywords

Ontology, Web Search, Feature Vector Construction.

1. INTRODUCTION

Typical IR systems work in word-space while we humans deal with information in concept-space [1]. Consequently, concept-based search is a promising approach since the burden of knowing how the documents are written is taken off the user and hence the user can focus on searching on a conceptual level instead. However, a problem with this approach is finding good concepts.

Ontologies can define concepts and the relationships among them for any domain of interest. It is assumed that the success of the Semantic Web strongly depends on ontologies [2]. However, creating and maintaining ontologies is both time-consuming and costly. Therefore, it is important to use the ontologies for many different tasks to increase return on investment (ROI). The focus of this paper is an approach for reuse of ontologies for traditional vector-space information retrieval (IR) systems. The idea is to associate every concept (classes and instances) of the ontology with a feature vector (*fv*) to tailor these concepts to the specific terminology used in the document collection, terms that tend to be used in connection with the concept and to provide a contextual definition of it.

2. RELATED WORK

The related work to our approach comes mainly from two areas. Ontology based IR, in general, and approaches to conceptual query expansion, in particular. General approaches to ontology based IR can further be sub-divided into Knowledge Base (KB) and vector space model driven approaches.

There are approaches combining both ontology based IR and vector space model. For instance, some start with semantic querying using ontology query languages and use resulting instances to retrieve relevant documents [3, 4]. Vallet et al. [4] use weighted annotation when associating documents with ontology instances. The weights are based on the frequency of occurrence of the instances in each document. Nagypal [5] combines ontology usage with vector-space model by extending a non-ontological query. There, ontology is used to disambiguate queries. Simple text search is run on the concepts' labels and users are asked to choose the proper term interpretation.

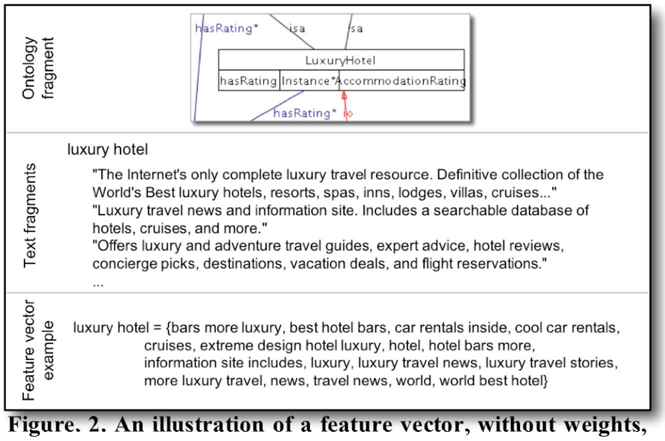
3. FEATURE VECTOR CONSTRUCTION

The process of constructing feature vectors constitutes three main steps. The aim of the first step is to do some preparation to optimize the construction of the feature vectors for each of the concepts of the selected ontology. The concepts are ranked according to assumed relevancy to the ontology. This list is the input to the next step.

The main aim of the second step is to extract and group sets of candidate terms being relevant to each concept. For each concept a query is formulated based on how the concepts relate to other concepts of the current ontology. The queries are submitted to the underlying search engine that retrieves a set of documents for each concept. Then the individual documents of each concept are clustered to group those documents having high similarity. For each cluster a set of candidate terms are extracted based on the documents of each cluster. These candidate terms are the input to the next and final step, which is to identify the relevant documents and construct a corresponding feature vector for each of the ontology concepts.

At this stage the grouped candidate terms are not necessarily relevant to the domain defined by the ontology currently used. Consequently, the aim of the third and last step is to identify those candidate terms being relevant to both the corresponding concepts and the current ontology. To find those domain relevant candidate terms we calculate the similarity between each of the grouped candidate terms of the current concept and the grouped candidate terms of the neighboring concepts. Weighting is used to differentiate on the importance of the ontology relation types when calculating these scores. Finally, the set of grouped candidate terms with the highest similarity to the neighboring concepts is selected and next used when creating the *fv* of the corresponding concept.

Figure 2 depicts an illustrative example.



luxury hotel

"The Internet's only complete luxury travel resource. Definitive collection of the World's Best luxury hotels, resorts, spas, inns, lodges, villas, cruises..."
"Luxury travel news and information site. Includes a searchable database of hotels, cruises, and more."
"Offers luxury and adventure travel guides, expert advice, hotel reviews, concierge picks, destinations, vacation deals, and flight reservations."
...

luxury hotel = {bars more luxury, best hotel bars, car rentals inside, cool car rentals, cruises, extreme design hotel luxury, hotel, hotel bars more, information site includes, luxury, luxury travel news, luxury travel stories, more luxury travel, news, travel news, world, world best hotel}

Figure. 2. An illustration of a feature vector, without weights, created for the concept LuxuryHotel. A fragment of the Travel ontology is shown at the top. While in the middle some of the summary snippets of the potentially related pages are shown. At the bottom, the resulting feature vector for the LuxuryHotel concept is shown.

4. ONTOLOGY-DRIVEN SEARCH

WebOdIR[1] is an ontology-driven information retrieval system for the Web that uses ontologies to bring the query closer to the real intention of the user's query. In our approach, the user can specify one or more concepts related to a domain of interest when formulating a query. In addition, the user can specify a set of keywords to narrow the search even further. By differentiating on concepts and keywords the real intention of the user's query can better be understood by the underlying machinery and thus present more relevant results.

WebOdIR is ontology-driven, which means that ontologies are used extensively in the search process. Firstly, an ontology is used to help a user formulating a query. Then the specified query concepts are used by the system to formulate one or more new queries, which are sent to the underlying search engine (e.g. Yahoo!). The new query is based on how the current concept relates to other neighboring concepts. Finally, the concepts are used to re-rank and filter out those documents retrieved by the underlying search engine that are considered to be of no or little relevance to the current domain.

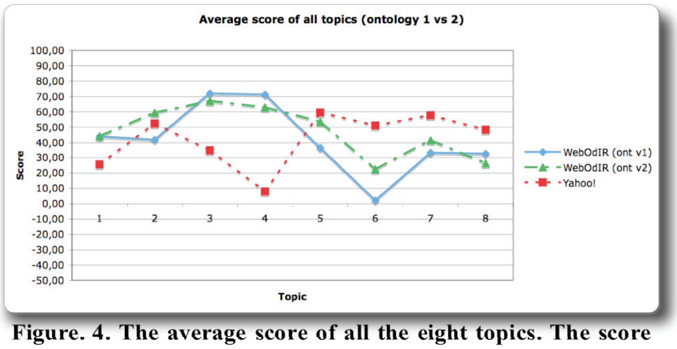


Figure. 4. The average score of all the eight topics. The score is in the range of -50 to 100 (see [6] for further details). The graph shows effectiveness of ontology 1 versus ontology 2 but also how they perform versus a similar search done with Yahoo! Web search.

5. Results

Figure 4 depicts a graph showing how the different ontologies versions influence on the search result relevance score and how they perform versus the baseline (in this case Yahoo! Web search). The graph shows that WebOdIR performs in general well especially for topic 3 and 4 (travel). The graph also shows that in general a more advanced ontology in the sense of having more relations, properties and individuals does perform better than a similar simpler ontology. A reason for this can be that for the more advanced ontologies more knowledge is available in the process of creating the concept *fv*s and hence will contain less noise compared to those of a simpler ontology. This observation matches well with one of our hypothesis regarding these issues.

6. CONCLUSION

Preliminary analysis of the experiment shows that the approach performs well, especially for shorter queries. In a survey the participants were asked to rate the quality of the results compared to the baseline in a scale from 1 (very bad) to 5 (very good), and the average score was 3.5. This score indicates that the approach for automatic construction of concept feature vectors based on an ontology works quite well. However, the evaluation results are preliminary and more analysis is needed to figure out when and why this approach performs better or worse than traditional keyword-based search. More analysis is also needed to state what aspects of an ontology is important for the process of generating good concept feature vectors.

7. ACKNOWLEDGMENTS

This research work is funded by the Integrated Information Platform for reservoir and subsea production systems (IIP) project, which is supported by the Norwegian Research Council (NFR). NFR project number 163457/S30. In addition, I would like to thank Darijus Strasunskas (NTNU), Jon Atle Gulla (NTNU), Per Gunnar Auran (Yahoo!), and Robert Engels (ESIS) for their support and help.

8. REFERENCES

1. Ozcan, R., Aslangdogan, Y.A.: *Concept Based Information Access Using Ontologies and Latent Semantic Analysis*. Technical Report CSE-2004-8. University of Texas at Arlington (2004) 16
2. Maedche, A., Staab, S.: *Ontology learning for the Semantic Web*. IEEE Intelligent Systems and Their Applications 16 (2001) 72-79
3. Kiryakov, A., Popov, B, Terziev, I., Manov, D., and Ognyanoff, D.: *Semantic Annotation, Indexing, and Retrieval*. Journal of Web Semantics 2(1), Elsevier, (2005)
4. Vallet, D, Fernández, M., Castells, P.: *An Ontology-Based Information Retrieval Model*. Gómez-Pérez, A., Euzenat, J. (Eds.): Proceedings of ESWC 2005, LNCS 3532, Springer-Verlag. (2005) 455-470.
5. Nagypal, G.: *Improving Information Retrieval Effectiveness by Using Domain Knowledge Stored in Ontologies*. OTM Workshops 2005, LNCS 3762, Springer-Verlag, (2005) 780-789
6. Brasethvik, T.: *Conceptual modelling for domain specific document description and retrieval - An approach to semantic document modelling*. IDI, Vol. PhD. NTNU, Trondheim (2004) 257

[1] WebOdIR prototype (23.07.2008), http://129.241.110.220