Effects of Using a Research Context Ontology for Query Expansion

Carola Carstens

German Institute for International Educational Research, Information Center for Education
Schloßstraße 29, 60486 Frankfurt, Germany
carstens@dipf.de

Abstract. This thesis investigates the question whether and how ontologies such as the ones currently evolving in the Semantic Web can serve as knowledge structures for the generation of query expansion terms in information retrieval systems. This issue is examined using a specific example domain, namely educational research. Initial results support the already well-researched finding that query expansion can increase recall. Subsequent experiments will focus on comparing the effectiveness of thesaurus and ontology-based query expansion, on identifying ontological relationships which are especially useful for the generation of query expansion terms in the domain of educational research, and on evaluating the usefulness of different ontological relationships as expansion terms for different types of queries, as well as for different query expansion modes.

Keywords: Information Retrieval, Ontologies, Semantic Web, Query Expansion

1 Research Interest

Following the Semantic Web vision [1], more and more ontologically organized Semantic Web data is currently being produced. With regard to the retrieval of this data, the Semantic Web has a strong focus on data retrieval, allowing to query the data by formal query languages. In addition to being directly queryable for certain ontology objects, ontological knowledge bases can also support information retrieval scenarios. Ontologically structured Semantic Web data constitutes a new abundance of corpus-independent knowledge that may be exploited for the implementation of query expansion mechanisms.

While the existing research on query expansion focuses on the use of thesauri, this research aims at investigating the use of ontologies for that purpose, their semantic relationships usually being more numerous than those typically comprised in thesauri. As research context related data is widely available on the Semantic Web, the use of a research context ontology for query expansion is examined. More specifically, the domain of educational research serves as an example application area.

The main research interest of this work is to investigate if ontologies such as the ones currently evolving in the Semantic Web are promising knowledge structures for

the implementation of query expansion mechanisms in information retrieval systems. Furthermore, the following more-detailed research questions are addressed:

- Can ontologies generate added value in query expansion mechanisms, as compared to thesauri?
- Which ontological relationships are most useful as query expansion terms for the field of educational research?
- Which ontological query expansion terms are most suitable for which type of query terms (concept, project, person, organization queries)?
- Which ontological relationships are suitable for automatic query expansion; which for interactive query expansion?

2 State of the Art

In the Semantic Web, a large proportion of the data that is published is research context related data. This is for example illustrated by the datasets of the ISWC and ESWC conferences[1] and the data of the AIFB semantic portal of the University of Karlsruhe[2]. For the purpose of publishing such data, the vocabulary of evolving standard ontologies such as SWRC [2], FOAF[3] and SKOS[4] can be used.

While this semantically interpretable data can be easily retrieved with the help of formal query languages (data retrieval), the role that full-text search and information retrieval play on the Semantic Web is still an open research issue [3, 4]. One step towards the integration of data and information retrieval is the augmentation of traditional search engine results with structured Semantic Web data. For example, this is achieved in the Semantic Search application by Guha et al. [5] and is also supported by Yahoo!'s SearchMonkey initiative[5].

Another possible integration of ontologies and information retrieval is the use of ontologically structured data as background knowledge for the implementation of query expansion mechanisms in information retrieval systems. Much research has already been conducted on the use of corpus-independent thesauri for this purpose [6, 7, 8, 9]. This research has shown that thesaurus-based query expansion often induces an increase in recall, usually accompanied by a significant loss in precision.

Taking a more detailed look at the effect of certain thesaurus relationships on the effectiveness of query expansion, Greenberg determined that synonyms and narrower terms are well suited for automatic query expansion, because they "increased relative recall with a decline in precision that was not statistically significant" [6]. This finding was further reinforced in her follow-up study focusing on the differences between automatic query expansion and interactive query expansion [7]. A more recent study by Navigli and Velardi examined the use of expansion terms derived from WordNet [10], coming to the conclusion that the use of gloss words for query expansion achieved top scores for the precision@10 measure, outmatching query expansion by synsets and hyperonyms, for example.

[1] http://data.semanticweb.org/dumps/
[2] http://www.aifb.uni-karlsruhe.de/english
[3] http://www.foaf-project.org/
[4] http://www.w3.org/2004/02/skos/
[5] http://developer.yahoo.com/searchmonkey/

As in the aforementioned study, several other authors also refer to thesauri as ontologies in the context of query expansion, although these lightweight ontologies are mostly restricted to typical thesaurus relationships, namely synonyms, broader terms, narrower terms and related terms [11]. Query expansion based on ontologies with additional semantic relationships, by contrast, has yet to be studied extensively.

3 Research Design

In order to simulate the existence of Semantic Web data for the domain of educational research, a domain-specific ontological knowledge base is created in the first step. The research context ontology (rescon) schema is modeled around the basic concepts *person*, *organization*, *project* and *concept*. With the aim of representing commonly used Semantic Web relationships, the schema partly integrates vocabulary from the FOAF, SWRC and SKOS ontologies, as illustrated in figure 1. This schema is instantiated with data from several domain-specific, structured or semi-structured data sources, resulting in a total of 27,082 concepts, 5,526 persons, 7,665 organizations and 540 projects. As shown in figure 1, the objects are interrelated by semantic relationships such as *swrc:employs*, *skos:related* and *foaf:topic_interest*.

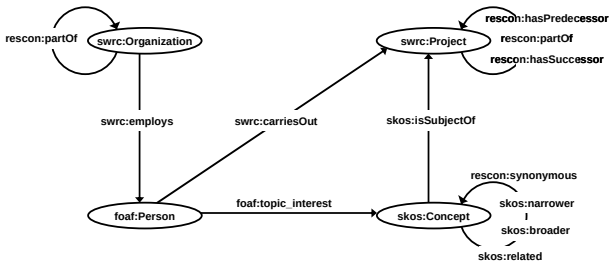

Fig. 1: Research context ontology schema

In the second step, a prototypical retrieval system based on Lucene[6] is implemented, incorporating both an automatic and an interactive mode for query expansion. Its configuration determines which ontology relationships are used for the generation of query expansion terms.

In the automatic query expansion mode, the expansion terms are added directly to each of the original query terms with the Boolean OR operator, before the query is sent to the Lucene index. This is known as the building-block strategy [12]. In the interactive mode, by contrast, the system generates expansion terms as suggestions that are presented to the user, who can then decide whether or not to make use of them to start a new search. Moreover, the user has the option to either broaden the original query or to narrow it down with the Boolean AND and OR operators.

The system's effectiveness is then measured in information retrieval experiments. This evaluation phase consists of two parts: the first part focuses on the automatic query expansion mode, while the second part investigates the interactive mode.

In the automatic mode, different ontology relationships are used sequentially as

[6] http://lucene.apache.org/

query expansion terms to determine how they affect information retrieval effectiveness. This can be evaluated in terms of the standard measures recall and precision [13]. Depending on the type of query – be it a *concept*, a *person*, an *organization* or a *project* – the use of each semantic relationship as a query expansion term is analyzed sequentially. Accordingly, the concept query *Musikunterricht* (*music lessons*) can be expanded by the synonym *Musikerziehung* (*music education*), the related term *Tanzunterricht* (*dance lessons)*, the broader term *Unterricht* (*lessons)* and the narrower term *Privatmusikunterricht* (*private music lessons*).

Compared to the automatic mode, the effectiveness of interactive retrieval systems is more difficult to evaluate, because the system's effect is difficult to isolate from other variables, such as user influences [14:15]. To circumvent this problem, a pre-questionnaire is used to record data about the search experience and the topic-related knowledge of the participants. In a between-groups design, one group of users works with a baseline retrieval system while another group works with the ontology enhanced system. As in the final TREC Interactive Track experiments [14], users are given realistic tasks to accomplish, such as to identify documents that give evidence of the cooperation of two organizations, for example. Comparable topics are defined for each of the different types of queries. Finally, it is possible to analyze which group accomplishes the tasks most effectively based on the number of result documents gathered in a certain time frame and the number of steps necessary for collecting the result documents.

This research setup makes it possible to identify which ontological relations are most suitable for which query expansion mode, which relations are most effective as query expansion terms for the domain under consideration, which relations are most effective for certain query types, and if ontological relations can generate added value compared to thesaurus relations. Although the experiments apply to one specific domain, it can be expected that the results on the use of relationships from a general research context ontology may be at least partly transferable to other domains.

4 Preliminary Results and Outlook

At the current stage, a pre-test for the automatic mode has been conducted, which already has some implications for the design of further experiments. For this pre-test, an excerpt from the German Education Index[7], a bibliographic database for the domain of educational research, served as a test corpus. Its 215,948 documents were all in German and contained full-text abstracts. A Lucene index was built using the document fields *abstract*, *title* and *subtitle*. The intellectually assigned keywords of the documents served as relevancy judgments. As these keywords were restricted to terms of the type *concept*, only the relations *synonymous (SYN)*, *narrower (NT)*, *broader (BT)* and *related (RT)* have been used sequentially as query expansion terms in this experiment. The average effects of these relations on information retrieval effectiveness are listed in table 1[8]. On the whole, this experiment took into account 3,067 SYN, 1,404 NT, 1,625 BT and 1,013 RT expansion terms.

[7] http://www.fachportal-paedagogik.de/start_e.html

[8] Expansion terms that did not occur in the test corpus were excluded from this calculation.

Table 1: Average influence of expansion terms on retrieval effectiveness (rounded)

	SYN	NT	BT	RT
Average recall deviation from baseline query	+ 8 %	+ 1%	+ 15%	+ 5%
Average precision deviation from baseline query	- 5%	- 3%	- 38%	- 15%

These preliminary results are in line with the results attained by Greenberg [6] insofar as they show that query expansion with semantic relationships can induce an increase in recall. Nevertheless, Greenberg reports a much higher increase in recall, up to nearly one quarter. This difference may be due to the fact that certain expansion terms only scarcely occur in the test corpus, thereby not permitting the expanded query to detect a high number of additional relevant documents. Therefore, the corpus frequency of the expansion terms will be taken into account in further calculations.

These pre-test experiences will inspire the more comprehensive upcoming experiments. By applying further evaluation measures, such as the precision at different ranks, and by studying user influences in the interactive mode, they can be expected to produce even more meaningful results in practice.

References

1. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Scientific American May 2001.
2. Sure, Y., Bloehdorn, S., Haase, P., Hartmann, J., Oberle, D.: The SWRC Ontology - Semantic Web for Research Communities. In: Bento, C., Cardoso, A., Dias, G. (eds.) EPIA 2005. LNCS, vol. 3803, pp. 218-231. Springer, Covilha, Portugal (2005).
3. Scheir, P., Pammer, V., Lindstaedt, S.N.: Information Retrieval on the Semantic Web - Does it exist? In: LWA 2007, pp. 252-257. Martin-Luther-University Halle-Wittenberg (2007).
4. Finin, T., Mayfield, J., Joshi, A., Cost, R.S., Fink, C.: Information Retrieval and the Semantic Web. In: 38th Annual Hawaii International Conference on System Sciences (2005).
5. Guha, R., McCool, R., Miller, E.: Semantic Search. In: 12th International Conference on World Wide Web, pp. 700-709 (2003).
6. Greenberg, J.: Automatic Query Expansion via Lexical-Semantic Relationships. Journal of the American Society for Information Science and Technology 52(5), 402-415 (2001).
7. Greenberg, J.: Optimal Query Expansion (QE) Processing Methods with Semantically Encoded Structured Thesauri Terminology. Journal of the American Society for Information Science and Technology 52(6), 487-498 (2001).
8. Kristensen, J.: Expanding End-Users' Query Statements for Free Text Searching with a Search-Aid Thesaurus. Information Processing and Management 29(6), 733-744 (1993).
9. Voorhees, E.M.: Query Expansion using Lexical-Semantic Relations. In: Annual ACM Conference on Research and Development in Information Retrieval, pp.61-69 (1994).
10. Navigli, R., Velardi, P.: An Analysis of Ontology-based Query Expansion strategies. In: 14th European Conference on Machine Learning, pp. 42-49 (2003).
11. ISO 2788: Documentation - Guidelines for the Establishment and Development of Monolingual Thesauri. International Organization for Standardization (1986).
12. Harter, S.P.: Search Strategies and Heuristics. In: Online Information Retrieval. Concepts, Principles and Techniques, pp. 170-204. Academic Press, Orlando, Florida (1986).
13. Salton, G., McGill, M.: Introduction to Modern Information Retrieval. McGraw-Hill, New York (1983).
14. Voorhees, E.M., Harman, D.K.: TREC: Experiment and Evaluation in Information Retrieval. MIT Press, Cambridge, MA (2005).