Research on Ontology-Driven Information Retrieval

Stein L. Tomassen

Department of Computer and Information Science,
Norwegian University of Technology and Science,
NO-7491 Trondheim, Norway
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. A particular concern here is the integration of these semantic approaches with traditional search technology. The research presented in this paper examines how ontologies can be efficiently applied to large-scale search systems for the web. We describe how these systems can be enriched with adapted ontologies to provide both an in-depth understanding of the user's needs as well as an easy integration with standard vector-space retrieval systems. The ontology concepts are adapted to the domain terminology by computing a feature vector for each concept. Later, the feature vectors are used to enrich a provided query. The whole retrieval system is under development as part of a larger Semantic Web standardization project for the Norwegian oil & gas sector.

1 Introduction

A problem with traditional information retrieval (IR) systems is that they typically retrieve information without an explicitly defined domain of interest to the user. Consequently, the system presents a lot of information that is of no relevance to the user. The research presented in this paper examines how ontologies can be efficiently utilized for traditional vector-space IR systems. The ontologies are adapted to the document space within multi-disciplinary domains where different terminology is used. The objective is to enhance the user-experience by improvement of search result quality for large-scale search systems.

One of the reasons for why IR systems do not have an explicitly defined domain of interest to the user is that most users tend to use very few terms (3 or less) in their search queries [1, 2]. As a result, the systems cannot *understand* the context of the user's query, which results in lower precision. By adding more relevant terms to the query, the domain of interest can, to some extent, be identified. However, adding both *correct* and *distinctive* terms is not always trivial, since the user needs knowledge about the terminology used in that particular domain to find those *correct* terms.

A novel and promising approach is concept-based search [3, 4, 5]. With this approach, 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. One problem with this approach is to find good concepts. The approach described in [3, 5] finds concepts based on the result set of the search, which then are used to refine the search. However, the relationships between the concepts are neglected.

Concepts and, in particular, relations between them can be specified in ontologies. Ontologies define concepts and the relationships among them [6]; therefore, they are often used to capture knowledge about domains. A growing number of IR systems make use of ontologies to help clarifying the information needs of the users, further described in section 3. However, a concern with these semantic approaches is the integration with traditional commercial search technologies.

In our approach [7], we propose a query enrichment approach that uses contextually enriched ontologies to bring the queries closer to the user's preferences and the characteristics of the document collection. 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 document collection and terminology used. The structure of the ontology is taken into account during the construction of the feature vectors. The ontology and its associated feature vectors are later used for post-processing of the results provided by the search engine.

This paper is organized as follows. In section 2, we describe the context of this research. In section 3, related work is discussed. In section 4, we describe the approach including some research questions and the methodology used. Where in section 5, we present the current status of this research. Finally, section 6 concludes the paper.

2 Research Context

The context of this research is information retrieval utilizing ontologies. Furthermore, the work of this PhD is part of the Integrated Information Platform for reservoir and subsea production systems (IIP) project. The IIP project is funded by the Norwegian Research Council (NFR)[1]. The project started in 2004 and will end in 2007. The project employs two PhD students and one research scientist.

The IIP project is creating an ontology for all subsea equipment used by oil and gas industry. Unlike other initiatives, this project endeavors to integrate life-cycle data spanning several standards and disciplines. A goal of this project is to define an unambiguous terminology of the domain and build an ontology that will ease integration of systems between disciplines. A common terminology is assumed to reduce risks and improve the decision making process in the industry. The project will also make this ontology publicly available and standardized by the International Organization for Standardization (ISO)[2].

[1] NFR project number 163457/S30

[2] http://www.iso.org/

3 State-of-the-Art

Traditional information retrieval techniques (i.e., vector-space model) have an advantage of being fast and give a fair result. However, it is difficult to represent the content of the documents meaningfully using these techniques. That is, after the documents are indexed, they become a “bag of terms” and hence the semantics is partly lost in this process.

In order to increase quality of IR much effort has been put into annotating documents with semantic information [8, 9, 10, 11]. That is a tedious and labor-intensive task. Furthermore, hardly any search engines are using metadata when indexing the documents. AltaVista[3] is one of the last major search engines which dropped its support in 2002 [12]. The main reason for this is that the meta information can be and has been misused by the content providers in the purpose of giving the documents a misleading higher ranking than it should have had [12]. However, there is still a vision that for ontology based IR systems on Semantic Web, “it is necessary to annotate the web’s content with terms defined in ontology” [13].

The related work to our approach comes from two main areas. Ontology based IR, in general, and approaches to 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. KB approaches use reasoning mechanism and ontological query languages to retrieve instances. Documents are treated either as instances or are annotated using ontology instances [13, 14, 15, 16]. These approaches focus on retrieving instances rather than documents. Some approaches are often combined with ontological filtering [17, 18, 19].

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 [16, 20]. [20] use weighted annotation when associating documents with ontology instances. The weights are based on the frequency of occurrence of the instances in each document. [21] 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. A similar approach is described in [22] where documents are associated with concepts in the ontology. The concepts in the query are matched to the concepts of the ontology in order to retrieve terms and then used for calculation of document similarity.

[17] is using ontologies for retrieval and filtering of domain information across multiple domains. There each ontology concept is defined as a domain feature with detailed information relevant to the domain including relationships with other features. The relationships used are hypernyms (super class), hyponyms (sub class), and synonyms. Unfortunately, there are no details in [17] provided on how a domain feature is created.

Most query enrichment approaches are not using ontologies like [3, 4, 5]. Query expansion is typically done by extending provided query terms with synonyms or hyponyms (cf. [23]). Some approaches are focusing on using ontologies in the process of enriching queries [15, 17, 22]. However, ontology in such case typically serves as

[3] AltaVista, http://www.altavista.com/

thesaurus containing synonyms, hypernyms/hyponyms, and do not consider the context of each term, i.e. every term is equally weighted.

[4] is using query expansion based on similarity thesaurus. Weighting of terms is used to reflect the domain knowledge. The query expansion is done by similarity measures. Similarly, [3] describes a conceptual query expansion. There, the query concepts are created from a result set. Both approaches show an improvement compared to simple term based queries, especially for short queries.

The approaches presented in [5, 24] are most similar to ours. However, [5] is not using ontologies but is reliant on query concepts. Two techniques are used to create the feature vectors of the query concepts, i.e. based on document set and result set of a user query. While the approach presented in [24] is using ontologies for the representation of concepts. The concepts are extended with similar words using a combination of Latent Semantic Analysis (LSA) and WordNet[4]. Both approaches get promising results for short or poorly formulated queries.

To show the difference from the related work discussed above we emphasize on the main features of our approach as follows. Our approach relies on domain knowledge represented in ontology when constructing feature vectors, then traditional vector-space retrieval model is used for the information retrieval task, where feature vectors are used to enrich provided queries. The main advantage of our approach is that the concepts of an ontology is tailored to the terminology of the document collection, which can vary a lot even within the same domain.

4 Research Approach

The overall objective of this research is to enhance the user-experience by improving search result quality for large-scale search systems. This objective contains the following sub goals:

- Explore and analyze the usage of ontologies for large-scale search systems for the web.
- Contribute with a method for applying ontologies efficiently to large-scale search systems for the web.

[4] WordNet, http://wordnet.princeton.edu/

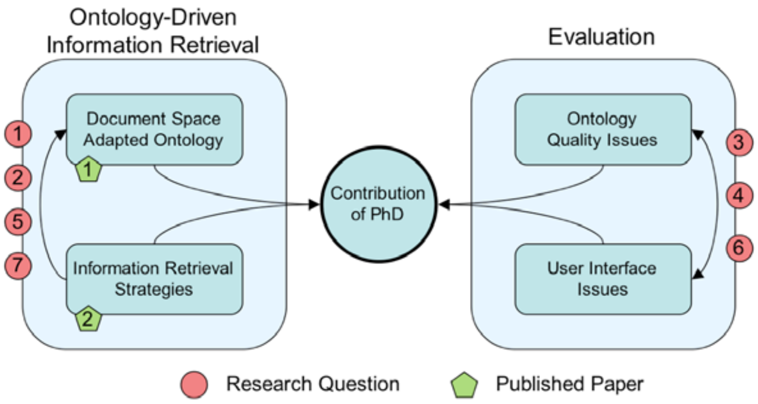

Fig. 1. Overview of research approach with the different main components and the relationship between them.

Fig. 1 depicts the overall approach of this research. The research is divided into two main areas, which are *Ontology-Driven Information Retrieval* and *Evaluation*. The former deals with aspects regarding ontologies in IR systems and different strategies of using ontologies and the latter deals with aspects regarding evaluation of ontology-driven IR systems, quality of IR ontologies, and user interface of ontology-driven IR systems. Fig. 1 also shows how the research questions, described in section 4.1, and relevant published papers, described in section 5.2, relates.

4.1 Research Questions

Research questions relevant to this work are as follows:

Q1: Can the retrieval effectiveness of large-scale search systems be improved by utilizing ontologies?

Q2: What components of large-scale search systems will benefit of using ontologies?

Q3: How can ontologies be used to help the user to improve the user experience of search systems?

Q4: What features of an ontology influence on the search quality?

Q5: How to provide the search system with more information of the user's intention of a query?

Q6: How to evaluate search systems where the user experience is taken into consideration?

Q7: How can ontologies be used to enhance search systems for web?

4.2 Research Method

This research will consist of several tasks being part of a cycle illustrated in Fig. 2. This cycle will be used for all the areas of research illustrated in Fig. 1 and will be an iterative process. The tasks are as follows:

Theoretical Framework:

- This task mainly consists of doing literature studies and establishing the state-of-the-art within the relevant areas of this research. A new theory will be created being inspired by the literature survey and the results from preliminary evaluations.

Implementation:

- This task consists of implementing the theories created in the previous task for testing.

Testing:

- The testing will be done using both quantitative and qualitative methods depending on what is being tested. For laboratory testing, typically precision and recall [25] measures will be used. Questionnaires will also be used since precision and recall does not take into account i.e. the user experience. In addition, it might be necessary to use observations and/or semi-structured interviews to gather all knowledge about the user experience of using the prototypes.
- Different test collections will be used depending on the ontologies. As part of the IIP project both some ontologies and text collections within the oil and gas domain will be available. Wikipedia[5] will also be used as a text collection for testing usage of smaller ontologies both manually created and found on the Web. In addition, the API from Yahoo will be used for large-scale search testing.

Analysis:

- The results of the testing will be analyzed and compared with previously gathered results. Based on this analysis the theoretical framework will be revised or a new one will be created, which next will be implemented and tested, etc.

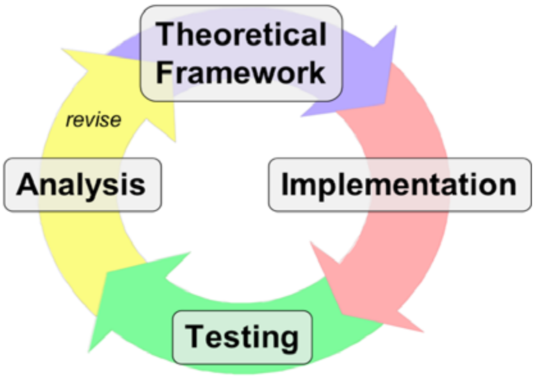

Fig. 2. Research tasks cycle. The cycle illustrates the tasks that are used for each individual aspect of the overall approach depicted in Fig. 1.

[5] Available for download from: http://download.wikimedia.org

5 Approach and Research Status

In this section the proposed architecture and some preliminary results are presented.

5.1 Proposed Architecture

Fig. 3 illustrates the overall architecture of the ontology-driven information retrieval system. Next the individual components of the system are briefly described.

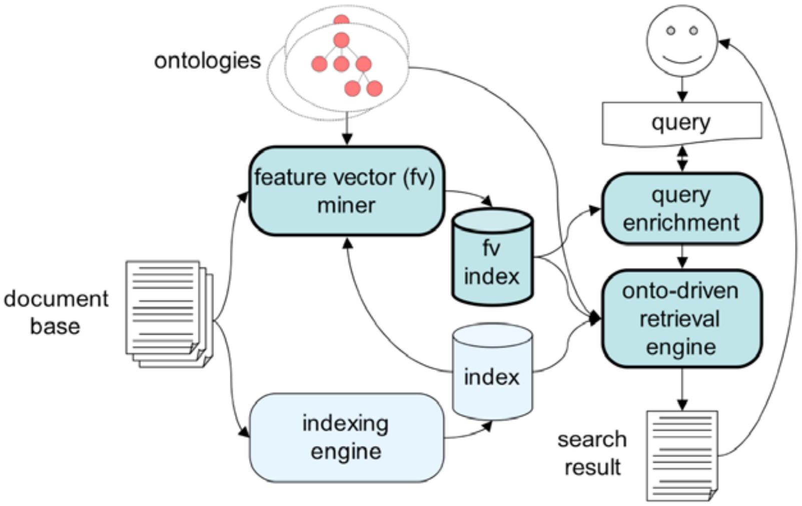

Fig. 3. The overall architecture of the ontology-driven information retrieval system. The no-transparent objects illustrate the components of the system. The outlined components illustrate those components being the contribution of this work to typically existing IR systems.

Feature vector miner: This component extracts the terms from the document collection and associates them with relevant concept(s) from the ontologies. The *fv index* is created offline equal to the *index* of the search engine.

Indexing engine: The main task of this component is to index the document collection. The indexing system is built on top of Lucene[6], which is a freely available and fully featured text search engine from Apache[7]. We will also do experiments using the index provided by Yahoo.

Query enrichment: This component handles the query specified by the user. The query can initially consist of concepts and/or ordinary terms (keywords). The concepts will be replaced by corresponding *fv*s. Each concept or term can be individually weighted. This component is further described in [7].

Onto-driven retrieval engine: This component performs the search and post-processing of the retrieved results. The ontologies and their corresponding *fv*s are used when post-processing the retrieved documents before presented to the user.

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

[7] http://www.apache.org/

5.2 Preliminary Results

Some components of the proposed architecture shown in Fig. 3 are implemented and individually tested. These components are all related to the *Query enrichment* component that was presented in a paper [7] at the NLDB 2006 conference[8], depicted as paper 1 in Fig. 1. Main architectural components and techniques constituting the method were presented in that paper. The components implemented are built upon the full-text retrieval engine Lucene from Apache. As research reported here is still in progress, we have not been able to formally evaluate the approach. However, preliminary results indicate that the quality of the feature vectors is very important for the quality of the search result. Further, we have proposed that concepts and ordinary terms or keywords of the query should be handled differently since they have different roles identified by the user. This proposal is described in paper [26], depicted as paper 2 in Fig. 1.

6 Conclusion

In this PhD work we will explore and analyze methods for utilizing ontologies to improve the retrieval quality. The concepts in the ontology are associated with contextual definitions in terms of weighted feature vectors tailoring the ontology to the content of the document collection. Further, the feature vectors are used to enrich a provided query. Query enrichment by feature vectors provides means to bridge the gap between query terms and terminology used in a document set, and still employing the knowledge encoded in the ontology.

Acknowledgements. 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 Jon Atle Gulla and Darijus Strasunskas for their support and help.

References

1. Gulla, J.A., Auran, P.G., Risvik, K.M.: *Linguistic Techniques in Large-Scale Search Engines*. Fast Search & Transfer (2002) 15
2. Spink, A., Wolfram, D., Jansen, M.B.J., Saracevic, T.: *Searching the Web: the public and their queries*. J. Am. Soc. Inf. Sci. Technol. 52 (2001) 226-234
3. Grootjen, F.A., van der Weide, T.P.: *Conceptual query expansion*. Data & Knowledge Engineering 56 (2006) 174-193
4. Qiu, Y., Frei, H.-P.: *Concept based query expansion*. Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval. ACM Press, Pittsburgh, Pennsylvania, USA (1993) 160-169

[8] http://www.nldb.org

5. Chang, Y., Ounis, I., Kim, M.: *Query reformulation using automatically generated query concepts from a document space*. Information Processing and Management 42 (2006) 453-468
6. Gruber, T.R.: *A translation approach to portable ontology specifications*. Knowledge Acquisition 5 (1993) 199-220
7. Tomassen, S.L., Gulla, J.A., Strasunskas, D.: *Document Space Adapted Ontology: Application in Query Enrichment*. 11th International Conference on Applications of Natural Language to Information Systems. Springer, Klagenfurt, Austria (2006)
8. Desmontils, E., Jacquin, C.: *Indexing a Web Site with a Terminology Oriented Ontology*. In I.F. Cruz, S. Decker, J. Euzenat and D.L. McGuinness (eds.) The Emerging Semantic Web. IOS Press, (2002) 181- 198
9. Motta, E., Shum, S.B., Domingue, J.: *Case Studies in Ontology-Driven Document Enrichment: Principles, Tools and Applications*. International Journal of Human-Computer Studies 6 (2000) 1071-1109
10. Popov, B., Kiryakov, A., Kirilov, A., Manov, D., Ognyanoff, D., Goranov, M.: *KIM - Semantic Annotation Platform*. In: Fensel, D., Sycara, K.P., Mylopoulos, J. (eds.): The Semantic Web - ISWC 2003, Second International Semantic Web Conference, Sanibel Island, FL, USA, October 20-23, 2003, Proceedings, Vol. 2870. Springer (2003) 834-849
11. Fensel, D., Harmelen, F.v., Klein, M., Akkermans, H., Broekstra, J., Fluit, C., Meer, J.v.d., Schnurr, H.-P., Studer, R., Hughes, J., Krohn, U., Davies, J., Engels, R., Bremdal, B., Ygge, F., Lau, T., Novotny, B., Reimer, U., Horrocks, I.: *On-To-Knowledge: Ontology-based Tools for Knowledge Management*. In Proceedings of the eBusiness and eWork 2000 (EMMSEC 2000) Conference, Madrid, Spain (2000)
12. Sullivan, D.: *Death of a Meta Tag*. Search Engine Watch (2002)
13. Song, J-F., Zhang, W-M., Xiao, W., Li, G-H., Xu, Z-N.: *Ontology-Based Information Retrieval Model for the Semantic Web*. Proceedings of EEE 2005. IEEE Computer Society (2005) 152-155
14. Rocha, C., Schwabe, D., de Aragao, M.P.: *A hybrid approach for searching in the semantic web*. Proceeding of WWW 2004, ACM (2004) 374-383
15. Ciorăscu, C., Ciorăscu, I., Stoffel, K.: *knOWLer - Ontological Support for Information Retrieval Systems*. In Proceedings of Sigir 2003 Conference, Workshop on Semantic Web, Toronto, Canada (2003)
16. Kiryakov, A., Popov, B, Terziev, I., Manov, D., and Ognyanoff, D.: *Semantic Annotation, Indexing, and Retrieval*. Journal of Web Semantics 2(1), Elsevier, (2005)
17. Braga, R.M.M., Werner, C.M.L., Mattoso, M.: *Using Ontologies for Domain Information Retrieval*. Proceedings of the 11th International Workshop on Database and Expert Systems Applications. IEEE Computer Society (2000) 836-840
18. Borghoff, U.M., Pareschi, R.: *Information Technology for Knowledge Management*. Journal of Universal Computer Science 3 (1997) 835-842
19. Shah, U., Finin, T., Joshi, A., Cost, R.S., Mayfield, J.: *Information Retrieval On The Semantic Web*. Proceedings of Conference on Information and Knowledge Management. ACM Press, McLean, Virginia, USA (2002) 461-468
20. 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.
21. Nagypal, G.: *Improving Information Retrieval Effectiveness by Using Domain Knowledge Stored in Ontologies*. OTM Workshops 2005, LNCS 3762, Springer-Verlag, (2005) 780-789
22. Paralic, J., Kostial, I.: *Ontology-based Information Retrieval*. Information and Intelligent Systems, Croatia (2003) 23-28
23. Chenggang, W., Wenpin, J., Qijia, T. et al.: *An information retrieval server based on ontology and multiagent*. Journal of computer research & development 38(6) (2001) 641-647.

24. 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

25. Baeza-Yates, R., Ribeiro-Neto, B.: *Modern information retrieval*. ACM Press, New York (1999)

26. Tomassen, S.L., Strasunskas, D.: *Query Terms Abstraction Layers*. Submitted to Web Semantics (SWWS'06) in conjunction with OnTheMove Federated Conferences (OTM'06), Montpellier, France (2006)