Improving information retrieval effectiveness by using domain knowledge stored in ontologies*

Gábor Nagypál

FZI Research Center for Information Technologies at the University of Karlsruhe
Haid-und-Neu-Str. 10–14
D-76131 Karlsruhe, Germany
nagypal@fzi.de

Abstract. The huge number of available documents on the Web makes finding relevant ones a challenging task. The quality of results that traditional full-text search engines provide is still not optimal for many types of user queries. Especially the vagueness of natural languages, abstract concepts, semantic relations and temporal issues are handled inadequately by full-text search. Ontologies and semantic metadata can provide a solution for these problems. This work examines how ontologies can be optimally exploited during the information retrieval process, and proposes a general framework which is based on ontology-supported semantic metadata generation and ontology-based query expansion. The framework can handle imperfect ontologies and metadata by combining results of simple heuristics, instead of relying on a "perfect" ontology. This allows integrating results from traditional full-text engines, and thus supports a gradual transition from classical full-text search engines to ontology-based ones.

1 Introduction

The huge number of available documents on the Web makes finding relevant ones a challenging task. Full-text search that is still the most popular form of search provided by the most used services such as Google, is very useful to retrieve documents which we have already seen (and therefore we know the exact keywords to search for), but it is normally not suitable to find not yet seen relevant documents for a specific topic.

The major reasons why purely text-based search fails to find some of the relevant documents are the following:

– *Vagueness of natural language*: synonyms, homographs and inflection of words can all fool algorithms which see search terms only as a sequence of characters.

* This work was partially funded by the VICODI (EU-IST-2001-37534) and DIP (no. FP6 - 507483) EU IST projects

- *High-level, vague concepts*: High-level, vaguely defined abstract concepts like the "Kosovo conflict", "Industrial Revolution" or the "Iraq War" are often not mentioned explicitly in relevant documents, therefore present search engines cannot find those documents.
- *Semantic relations*, like the *partOf* relation, cannot be exploited. For example, if users search for the European Union, they will not find relevant documents mentioning only Berlin or Germany.
- *Time* dimension: for handling time specifications, keyword matching is not adequate. If we search documents about the "XX. century" using exactly this phrase, relevant resources containing the character sequences like "1945" or "1956" will not be found by simple keyword matching.

Although most of the present systems can successfully handle various inflection forms of words using stemming algorithms, it seems that the lots of heuristics and ranking formulas using text-based statistics that were developed during classical IR research in the last decades [1] cannot master the other mentioned issues. One of the reasons is that term co-occurrence that is used by most statistical methods to measure the strength of the semantic relation between words, is not valid from a linguistic-semantical point of view [2].

Besides term co-occurrence-based statistics another way to improve search effectiveness is to incorporate background knowledge into the search process. The IR community concentrated so far on using background knowledge expressed in the form of thesauri. Thesauri define a set of standard terms that can be used to index and search a document collection (controlled vocabulary) and a set of linguistic relations between those terms, thus promise a solution for the vagueness of natural language, and partially for the problem of high-level concepts.

Unfortunately, while intuitively one would expect to see significant gains in retrieval effectiveness with the use of thesauri, experience shows that this is usually not true [3]. One of the major cause is the "noise" of thesaurus relations between thesaurus terms. Linguistic relations, such as synonyms are normally valid only between a specific meaning of two words, but thesauri represent those relations on a syntactic level, which usually results in false positives in the search result. Another big problem is that the manual creation of thesauri and the annotation of documents with thesaurus terms is very expensive. As a result, annotations often incomplete or erroneous, resulting in decreased search performance.

Ontologies form the basic infrastructure of the Semantic Web [4]. As "ontology" we consider any formalism with a well-defined mathematical interpretation which is capable at least to represent a subconcept taxonomy, concept instances and user-defined relations between concepts. Such formalisms allow a much more sophisticated representation of background knowledge than classical thesauri. They represent knowledge on the semantic level, i.e., they contain semantic entities (concepts, relations and instances) instead of simple words, which eliminates the mentioned "noise" from the relations. Moreover, they allow specifying custom semantic relations between entities, and also to store well-known

facts and axioms about a knowledge domain (including temporal information). This additional expression power allows the identification of the validity context of specific relations. E.g., while in the context of the "Napoleon invades Russia" event the "Napoleon – Russia" relation is valid, it does not hold in general.

Based on that, ontologies theoretically solve all of the mentioned problems of full-text search. Unfortunately, ontologies and semantic annotations using them are hardly ever perfect for the same reasons that were described at thesauri. Indeed, presently good quality ontologies and semantic annotations are a very scarce resource. This claim is based on both personal experiences during the VICODI project [5], and on our analysis of available ontologies and metadata on the present Web[1].

During the VICODI project an ontology-based web portal was developed which contained documents about European history[2]. A comprehensive ontology of European history was also developed. Although VICODI showed some of the potentials of an ontology-based information portal, the quality of the results were plagued by the lack of proper ontological information, due to the prohibitive cost of developing an ontology fully covering such a wide domain. The main lesson learned from the project is that it is very hard to switch from present full-text based information systems to semantic based ones in a one big step, but rather a gradual approach is needed, which combines the merits of statistical and ontological approaches, and thus provides a smooth transition between the two worlds.

In addition to the costs of ontology creation, another cause for imperfection is the limited expression power of ontology formalisms. Although they are much more powerful than thesauri, there are still many important aspects that cannot be modeled in present-day ontology languages. Therefore, imperfection in ontologies and metadata should be considered probably even in the long run, as the expression power of ontologies cannot be significantly raised without losing decidability of ontology reasoning.

The goal of this thesis is to examine and validate whether and how ontologies can help improving retrieval effectiveness in information systems, considering the inherent imperfection of ontology-based domain models and annotations.

This research builds on the results of VICODI, and therefore its major domain is also history. While history is a very interesting application domain from a theoretical point of view, it has also a strong practical relevance. After all, what is news today, will be history tomorrow. Therefore it is likely that the developed techniques will be directly exploitable in news portals on the Web. Indeed, we also plan to make some experiments with the IT-News domain.

The main contribution of this work is the demonstration of the utility of semantical information in an application domain of practical relevance without making unrealistic assumptions about the quality of ontologies and semantic

[1] E.g. http://www.daml.org/ontologies/, http://ontolingua.nici.kun.nl:5915/ and http://protege.cim3.net/cgi-bin/wiki.pl?ProtegeOntologiesLibrary

[2] Accessible from http://eurohistory.net

metadata. This can provide a strong motivation for the creation of new ontologies which is a crucial step toward the Semantic Web vision.

2 Approach

This research evaluates the following hypotheses:

1. Ontologies allow to store domain knowledge in a much more sophisticated form than thesauri. We therefore assume that by using ontologies in IR systems a significant gain in retrieval effectiveness can be measured.
2. The better (more precise) an ontology models the application domain, the more gain is achieved in retrieval effectiveness.
3. It is possible to diminish the negative effect of ontology imperfection on search results by combining different ontology-based heuristics during the search process which are immune against different kinds of ontology errors.
4. It is a well-known fact that there is a trade-off between algorithm complexity and performance. This insight is also true for ontologies: most of the ontology formalisms do not have tractable reasoning procedures. Still, our assumption is that by combining ontologies with traditional IR methods, it is possible to provide results with acceptable performance for real-world size document repositories.

These hypotheses are evaluated by implementing a prototype ontology-based IR system, and running experiments on a test collection. Gains in retrieval effectiveness in terms of classical IR measures such as precision and recall [1] are expected.

2.1 IR process and architecture

A schematic description of a usual IR process is shown on Fig. 1. Background knowledge stored in the form of ontologies can be used at practically every step of the process. For performance reasons, however, it does not seem to be feasible to use background information in the similarity measure used during matching and ranking, as it would be prohibitively expensive. Although, e.g., case-based reasoning systems apply domain-specific heuristics in their similarity measure [6], they operate on document collections which contain only several hundreds or thousand cases. We rather believe that it is possible to extend the query (and/or the document representation) syntactically based on the information stored in ontologies so that a simple, syntax-based similarity measure will yield semantically correct results (see also Hypothesis 4).

In this work, solutions are therefore provided for the issues of ontology-based query extension, ontology-supported query formulation and ontology-supported metadata generation (indexing). This leads to a conceptual system architecture (see Fig. 2) where the Ontology Manager component has a central role, and it is extensively used by the Indexer, Search Engine and GUI components[3].

[3] The GUI component is responsible for supporting the user in query formulation.

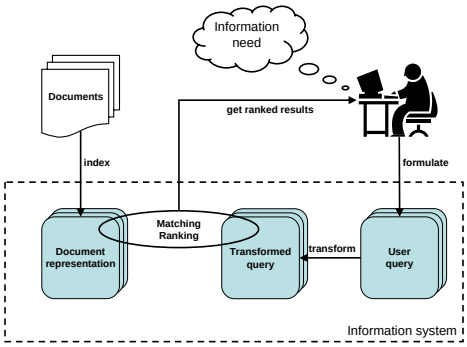

Fig. 1. IR process

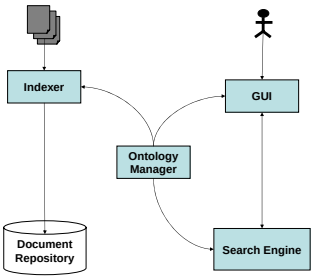

Fig. 2. Architecture

2.2 Information model

The information model defines how documents and the user query are represented in the system. The model used in this work is based on the model we developed during the VICODI project, and represents the content of a resource as a weighted set of instances from a suitable domain ontology (the conceptual part) together with a weighted set temporal intervals (the temporal part). The representation of the conceptual part is practically identical with the information model used by classical IR engines built on the vector space model [1], with the difference that vector terms are ontology instances instead of words in a natural language. This has the advantage that retrieval algorithms, index structures, or even complete IR engine implementations can be reused with our model.

Time, as a continuous phenomenon has different characteristics than the discrete conceptual part of the information model. The first question according time is how to define similarity among weighted sets of time intervals. A possible solution which is being considered, is to use the "temporal vector space model", described in [7]. The main idea of the model is that if we choose a discrete time representation, the lowest level of granules can be viewed as "terms" and the vector space model is applicable also for the time dimension.

Another time-related problem is caused by some special properties of history as an application domain. Generally, it turned out that traditional time intervals are not suitable to represent historical time specifications because of uncertainty, vagueness and subjectivity. E.g., it is impossible to exactly specify the birth date of Stalin (there are two possible dates) or to define the precise starting date of the "Russian Revolution". To address these issues, a fuzzy temporal model together with a temporal algebra was proposed in [8]. Interestingly, recently a user study in the business news domain [7] further validates the claim that for some temporal specifications fuzzy time intervals are better suited than traditional time intervals (or a set of time intervals).

The previously mentioned "temporal vector space" model can be naturally extended to incorporate fuzzy temporal intervals. In this case the set of lowest

level granules form the universe and fuzzy temporal intervals define fuzzy subsets of this universe.

A problem with the "temporal vector space" approach is the potentially huge number of time granules which are generated for big time intervals. E.g. to represent the existence time of concepts such as the "Middle Ages", potentially many tens of thousand terms are needed if we use days as granules. This problem can be diminished by "granule switching", i.e. using bigger granules for queries and/or documents which define a wide time range as relevant. The intuition is that in those cases the information loss caused by the transformation will not distort relevance scores significantly.

In addition to the original VICODI model, our information model also contains a traditional bag of words representation of the document content (and query terms), as we see the ontology-supported IR heuristics as an extension of the traditional IR approaches and not as a replacement.

2.3 Query formulation

VICODI experience showed that navigation in a full-fledged ontology is too complicated for most of the users. Therefore during query formulation we use the ontology only to disambiguate queries specified in textual form. E.g., if the users type "Napoleon" we provide them a list of Napoleons stored in the ontology (by running classical full-text search on ontology labels), and users only have to choose the proper term interpretation. This is much easier than finding the proper Napoleon instance starting from the ontology root. We also plan to make experiments with completely automatic disambiguation techniques like in [9] and in [10].

2.4 Ontology-supported query expansion and indexing

Recent research shows [11] that a combination of ranking results of simpler queries can yield a significantly better result than a monolithic query extension. This is probably because every algorithm has its own strengths and weaknesses and it is not possible to find "the optimal" method. Therefore, it is better to combine the results of different algorithms than just using only one (see also Hypothesis 3).

Motivated by these results our query process applies various ontology-based heuristics one-by-one to create separate queries which are executed independently using a traditional full-text engine. The ranked results are then combined together to form the final ranked result list (see Fig. 3). The combination of results is based on the belief network model [12] which allows the combination of various evidences using Bayesian inference. New types of ontology-based or statistical heuristics can be easily added to the system. If no ontological information is available, the system simply uses the "bag of words" part of the information model. Therefore the proposed search process supports a gradual transition between full-text and ontology-based IR systems.

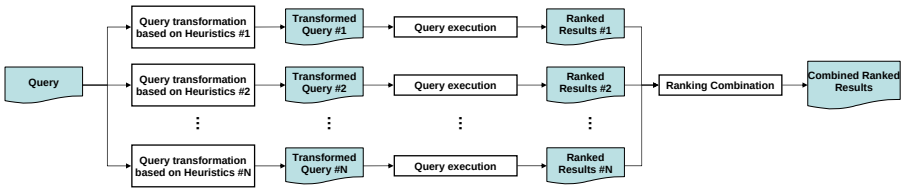

Fig. 3. The search process

2.5 Test collection and evaluation strategy

A new test collection has to be built because unfortunately, presently no test collections are available that incorporate ontologies. The approach for building the test collection is the following: Wikipedia is reused as the basis[4], which is refined by adding documents describing specific miniworlds, such as "World War II" or the "Iraq War". For these miniworlds (specific topics) facts are also added to the ontology and metadata is generated.

In other words, we follow the gradual approach also during test collection creation, as it is clearly not feasible to semantically annotate the whole Wikipedia for these experiments. After this process there should be miniworlds that are only partially or erroneously covered by semantic information, and there should be other topics which are not covered by the ontology at all. This is needed to test the robustness of the approach.

One advantage of this approach is that recall can be estimated well, as most of the relevant documents for the miniworlds are added manually to the test collection. Wikipedia can also contain relevant documents, of course. Those will be explored using the pooling method, which is also used on the TREC conferences (see e.g. [13]) to estimate the number of relevant documents in big test collections. Further, since Wikipedia contains more than 700000 documents, the performance of the system can be tested in a real-world situation.

Classical IR laboratory experiments will be conducted to get recall and precision figures. Various ontology-based heuristics will be switched on one-by-one, and a gradual gain in effectiveness is expected, which will hopefully validate Hypothesis 1 and partially Hypothesis 2. To further validate Hypothesis 2, fuzzy time information will be switched on and off to see if the simplification caused by using traditional crisp temporal intervals degrade retrieval effectiveness. To validate Hypothesis 3, the results of various heuristics will be integrated into one query extension step, and the retrieval figures will be compared with the other approach which combines the ranking results. Hypothesis 4 will be validated by comparing the response time of the new system with the response time of traditional full-text search.

[4] Available for download at http://download.wikimedia.org/

2.6 Research status and implementation

The fuzzy time model for modeling temporal specifications in history is complete, presently we are working on tool support for defining such intervals in a user-friendly way. Tool support for ontology development is in place and it is improved continuously. The test collection for the evaluation is presently being created.

During the VICODI project a first prototype of the described IR system was developed [14], using crisp temporal intervals and a one-step query expansion. This system used the KAON1 system [15] for reasoning, which unfortunately did not scale well. Based on the experiences with the VICODI system a new prototype is presently under implementation using the more powerful KAON2 system[5], fuzzy intervals and the ranking combination approach. The KAON2 system is used as an efficient disjunctive Datalog engine. Although KAON2 also supports OWL-DL reasoning, Datalog is used directly as the ontology implementation language because OWL-DL (and generally description logic) is not suitable for the application domain of history because some required reasoning patterns, such as temporally dependent transitive partOf relations on locations, are not supported. KAON2 also supports user-defined datatypes, which is crucial for implementing fuzzy time interval-based reasoning in the ontology.

As full-text retrieval engine, the Apache Lucene library[6] is used.

For the ontology-supported automatic semantic metadata generation, we use the GATE system[7], and develop GATE components which can exploit background information stored in our ontology. During the indexing process we also exploit traditional GATE components which are not ontology-aware, i.e. we follow the same gradual approach as during the query process.

3 State of the Art

It is well known fact in the field of IR that simple syntactical matching of the document and query representations do not always yield optimal results. A big body of literature exists where approaches using thesauri is described (see e.g. [1], Chapter 5 for an overview). This includes both automatically constructed thesauri based on statistical properties of the document collection or hand-crafted thesauri. As ontologies also codify background knowledge of a domain, lessons learned for thesauri are also relevant for ontologies.

Another related area is Information Extraction (IE) [16] that provides methods for extracting pieces of information from textual documents. Results of this research are useful for the indexing task in the IR process. Although usually a general claim is made by the IE community that semantic indexing (extracting semantic metadata from documents) provides better retrieval effectiveness than traditional full-text search, the emphasis of these systems is not retrieval but indexing. Probably therefore these claims are not yet validated, although because of the imperfection issues this claim is not trivial.

[5] Available from http://kaon2.semanticweb.org
[6] http://lucene.apache.org
[7] http://gate.ac.uk/

Recently, ontology-based information retrieval attracted a considerable interest. Most of the systems, however, concentrate on retrieving ontology instances, rather than documents [10, 17]. The approaches of these works can be used, however, as part of our ontology-based heuristics to extend the IR query.

Probably the most relevant works to this research are the KIM system [18] and the work reported in [9]. They also define a general framework for ontology-supported document retrieval, and integrate full-text search with ontology-based methods. These systems, however, start with a semantic query in an ontology query language and use the resulting instances to retrieve relevant documents. This is different from our approach where the ontology is rather used to extend a non-ontological query, which also contains semantic elements. This has the advantage, that our system can be also used for information filtering because the representation of documents and queries coincide.

KIM uses the retrieved ontology instances to initiate a traditional full-text search on documents where ontology annotations are embedded into the document text as terms. We also use a traditional search engine on the syntactic representation of the ontology elements, but we do not embed annotations to documents.

In the system of Vallet et al. documents are connected with ontology instances via weighted annotations, which is practically the same solution as our "bag of ontology instances" model. As they also include documents and annotation to the ontology, they can directly use the annotation weights to calculate semantic document relevance using the classical tf-idf formula, after executing the ontology-based query. They also execute a full-text search separately and combine its returned relevance weight with the result of the semantic query to diminish the effect of ontology imperfection. This is a similar, but simpler idea that we use when we combine the results of ontology-based heuristics.

None of the solutions handle the temporal dimension separately, and correspondingly they do not provide any support for fuzzy temporal intervals which is needed in some application domains such as history or business news.

4 Conclusion

This PhD work examines how background knowledge stored in ontologies and semantic metadata can be optimally exploited for the task of information retrieval. A special emphasis is placed on the issue of imperfect ontologies and metadata which is the reality on the present Web. History is used as application domain, which is very challenging from the temporal modeling perspective and allows evaluating the effect of ontology modeling simplifications on retrieval effectiveness.

During further work we validate that the proposed solution significantly improves retrieval effectiveness of information systems and thus provides a strong motivation for developing ontologies and semantic metadata. The gradual approach described allows a smooth transition from classical text-based systems to ontology-based ones.

References

1. Baeza-Yates, R.A., Ribeiro-Neto, B.A.: Modern Information Retrieval. ACM Press / Addison-Wesley (1999)
2. Kuropka, D.: Uselessness of simple co-occurrence measures for IF&IR – a linguistic point of view. In: Proceedings of the 8th International Conference on Business Information Systems, Poznan, Poland (2005)
3. Salton, G.: Another look at automatic text-retrieval systems. Commun. ACM **29** (1986) 648–656
4. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Scientific American **284** (2001) 34–43
5. Nagypál, G., Deswarte, R., Oosthoek, J.: Applying the Semantic Web – the VICODI experience in creating visual contextualization for history. Literary and Linguistic Computing (2005) to appear.
6. Lenz, M., Bartsch-Spörl, B., Burkhard, H.D., Wess, S., eds.: Case-Based Reasoning Technology: From Foundations to Applications. Volume 1400 of Lecture Notes in Computer Science. Springer (1998)
7. Kalczynski, P.J., Chou, A.: Temporal document retrieval model for business news archives. Information Processing & Management **41** (2005) 635–650
8. Nagypál, G., Motik, B.: A fuzzy model for representing uncertain, subjective, and vague temporal knowledge in ontologies. In: On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE, Springer (2003) 906 – 923
9. Vallet, D., Fernández, M., Castells, P.: An ontology-based information retrieval model. In: The Semantic Web: Research and Applications: Second European Semantic Web Conference, ESWC 2005, Springer (2005) 455–470
10. Rocha, C., Schwabe, D., Aragao, M.P.: A hybrid approach for searching in the semantic web. In: WWW '04: Proceedings of the 13th international conference on World Wide Web, ACM Press (2004) 374–383
11. Silveira, M.L., Ribeiro-Neto, B.: Concept-based ranking: a case study in the juridical domain. Information Processing & Management **40** (2004) 791–805
12. Ribeiro, B.A.N., Muntz, R.: A belief network model for IR. In: SIGIR '96: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, ACM Press (1996) 253–260
13. Voorhees, E.M., Buckland, L.P., eds.: NIST Special Publication 500-261: The Thirteenth Text REtrieval Conference Proceedings (TREC 2004). (2004)
14. Surányi, G.M., Nagypál, G., Schmidt, A.: Intelligent retrieval of digital resources by exploiting their semantic context. In: On The Move to Meaningful Internet Systems 2004: CoopIS, DOA, and ODBASE, Springer (2004) 705–723
15. Motik, B., Maedche, A., Volz, R.: A conceptual modeling approach for semantics-driven enterprise applications. In: Proc. 1st Int'l Conf. on Ontologies, Databases and Application of Semantics (ODBASE-2002). (2002)
16. Vlach, R., Kazakos, W.: Using common schemas for information extraction from heterogeneous web catalogs. In: Proceedings of the 7th East-European Conference on Advances in Databases and Informations System (ADBIS), Springer (2003) 118–132
17. Stojanovic, N., Studer, R., Stojanovic, L.: An approach for the ranking of query results in the semantic web. In: ISWC 2003. (2003) 500–516
18. Kiryakov, A., Popov, B., Terziev, I., Manov, D., Ognyanoff, D.: Semantic annotation, indexing, and retrieval. Journal of Web Semantics **2** (2005)