Architecture of a knowledge based interactive Information Retrieval System

Melanie Gnasa
Institute of Computer Science III
University of Bonn
gnasa@bonn.edu

Jens Woch
Department of Computer Science
University of Koblenz-Landau
woch@uni-koblenz.de

Abstract

The dramatic explosion of internet information sources that become available to an exponentially growing number of users necessitates new strategies for navigating through the content space. This paper proposes an innovative architecture for an interactive information retrieval system which differs from common architectures for its unique knowledge representation which integrates syntactical and domain knowledge in such a way that it allows for the implementation of natural language driven document analysis processes, user input analysis processes, and feedback generation processes without having different knowledge bases and formalisms for each task.

1 Introduction

The dramatic explosion of internet information sources that become available to an exponentially growing number of users necessitates new strategies for navigating through the content space. The major web search engines have significant limitations in terms of user support by query formulation and interactivity to substantiate the individual information need. As such, information retrieval systems should (1) bridge the semantic gap between the user's and the author's vocabulary and (2) give the user feedback to his search result in order to refine the search request.

(Gnasa and Harbusch, 2002) have shown that the understanding of artificial query languages commonly is quite imperfect. Misconceptions range from syntactical to perceptional issues and produce the full bandwidth of unintended results, e.g. too few or too much or simply wrong results and the user is left alone to figure out whether this is a consequence of syntactical or other matters regarding the query language, the data or the user's perception of both of them. For example, *"Give me all authors who write essays"* is much more intuitive than "$X \leftarrow \exists y : essay(y), author\_of(X, y)$" and a simple keyword match for *"author AND essay"* obviously has different semantics.

In this paper, we give a detailed view of an interactive information retrieval system supporting natural language processing of documents and queries.

This work is partly funded by the German Research Foundation.

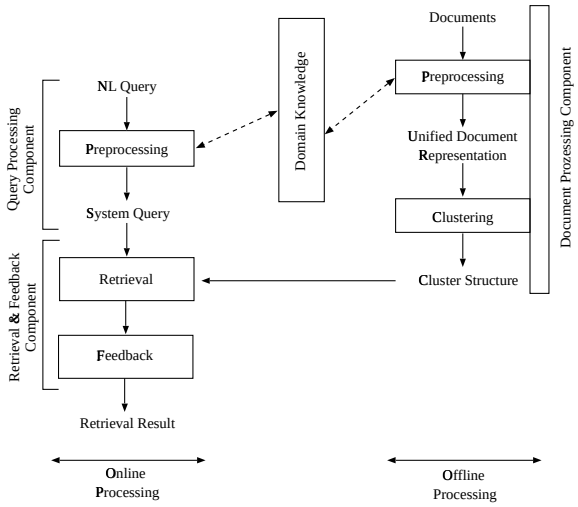

Figure 1: Architecture of the proposed information retrieval system

2 Architecture

Fig. 1 gives an overview of the architecture of the proposed retrieval system. Three main components could be distinguished: document processing component, query processing component and retrieval & feedback component. First, the document processing component fulfills two tasks. The documents are preprocessed to build a unified document representation, assisted by the ontology and supertagging which in this case help to generate natural language representations of ontological content

(and therefore document content). Getting a common representation of all different document formats is the pre-requisite of the final clustering such that the computed cluster structure is the basis for the search. Second, the query processing component performs a mapping from the information need (in our case natural language formulated by speech or text) to a system query with the help of a domain ontology with associated supertags. Third, the result of the document processing, i.e. the retrieved clusters build the input for the retrieval and feedback component. An iterative process between query modification and comparison with the cluster structure is started, whenever the result set is too large or the user is not satisfied. The feedback process ends, once the user does not want to restrict his or her query any more or the bottom of the cluster hierarchy is reached. In the following the three main components of our proposed system architecture are presented in detail.

3 Document Processing Component

3.1 Document Analysis

In order to provide a natural language motivated retrieval system the document collection needs to be preprocessed and classified. Our approach uses supertags for the analysis of user input as well as syntactical analysis of the documents. Situated in an ontology, supertags play a key role of our design.

An essential feature of a retrieval system is the contribution of information from different resources. Hence, a set of methods and tools that analyse the diversity of information sources is necessary.

The basis for this aim builds the supertags. In contrast to the small complexity of formulated queries, the textual information of the data collection must be more elaborated. The text, of course, is more than just a sequence of words. Taking this into account, a complex structure for each document is developed. Our general approach to this problem consists of three phases and a final mapping of all collected information to a unified representation.

In order to analyse all documents, the corpus is first segmented into sentences. Subsequent to this process each sentence needs to be parsed. For the syntactical offline analysis of the documents, the grammatical structures of the ontology are used as if being supertags, i.e., it is tried to assert the supertags to the documents' sentences. This does not perform a thorough analysis, because the grammatical structures of the ontology are by far not complete in order to analyze arbitrary sentences. Thus, we suspect, that the analysis' quality can be raised significantly by either applying a common parsing strategy (which may be impossible because of the sheer size of the domain space, even if processed offline), or by applying a well-trained supertagger as of (Chandrasekar and Srinivas, 1997), if a tagged domain specific corpus is available. However, further investigations have to reveal if it suffices to simply enrich the grammatical structures of the ontology to gain a sufficient precision. This should be the case, because the descriptions are supposed to be abstract enough for the analysis to focus on the essentials and concrete enough for the what-to-say component of the sentence generation. The results of that analysis (mainly the assertion of supertags and the corresponding instantiation of the ontology) represent the syntactical and conceptual level. The structural analysis takes structural information of the document (i.e. sections, subsections, headers etc.) into account.

Finally, the results of all three analyses are stored into an XML-file for further processing without to have to worry about the different formats of the documents. This file is the source for later retrieval processes and feedback generation.

3.2 Clustering

For the representation of documents within a retrieval system as one possible organization a classification or cluster structure is conceivable. Further research has to reveal whether the potential of clustering based on our unified document representation and thus the inclusion of syntactical and structural information for computing the document similarity produces better results than common clustering techniques, but this seems to be plausible.

4 Query Processing Component

The classical IR technique for analyzing a search query is *parsing*. Combined with the Boolean retrieval model, logical operators could be used to specify the search query. The popularity of the Boolean retrieval model results from the easy implementation and the time efficiency in processing a query. However, this model requires the user to be experienced in formulating Boolean logical terms. The evaluation of the information seeking behaviour of IT-professionals (cf. (Gnasa and Harbusch, 2002)) showed that empty result sets are produced in 50% of the cases by syntactic errors. The

other half of empty result sets is produced in 78% of the cases by one single search term, i.e. it apparently does not result from too complex queries as was expected by, e.g., (Cooper, 1988). This seems to be a surprising result for IT-professionals.

For this reason, we propose a robust parsing strategy for individualized search input strings. However, in case of spoken input (which should be the common case in our scenario), a user independent recognizer (e.g. Nuance, www.nuance.com) tries to extract a list of words from the input. The bottleneck is the restricted size of the vocabulary and the insufficient reliability of the recognition. Such problems must be considered by the analysis, i.e. the stock of words of a speech recognizer for spoken language is presently not arbitrary large and the input analysis of in part incorrectly recognized speech could only ensue rudimentary.

As introduced before the syntactical knowledge is represented by supertags. If no token of the input sequence matches with the anchors of the supertags, the supertagging fails, and the user is requested to paraphrase his statement. Otherwise, the anchor is analyzed in its structural context given by the supertag combination, i.e., the input is partially and only "almost" parsed (Srinivas, 1997). This is no real parsing, since only the combination is checked, whether the input can be derived from it.

5 Retrieval & Feedback Generation Component

The feedback generation performs several heuristics on a meta-level in order to decide whether the feedback should present query results or whether it should ask for query refinements. For example, if the search reveals that there are too many clusters, a conceptual refinement would be advisable, e.g., *There are too many meanings of X. You should try to be more precise on X"*. Or there are manageable many results, but they are somehow contradictory: *Did you mean JAVA being a programming language, or being an island?"* Finally, simply presenting the results as exemplified earlier, is obviously a task of the feedback generation, too.

For the automated feedback generation we adopted the integrated natural language system described in detail in (Harbusch and Woch, 2002). The how-to-say component remains the same, namely an into Schema-TAGs ((Harbusch and Woch, 2000)) transformed XTAG, which is a large English grammar (cf. (Doran et al., 1994)). However, domain specific expressions and vocabulary has to be added. The what-to-say component is extracted from the grammar structures associated with the ontology. Additional discourse related grammatical descriptions are required to perform dialog structures, as, e.g., asking for user input, presenting retrieval results and so forth. Since they are not domain dependent, they form an artificial third component, which is due to its nature not associated with the ontology and therefore can be reused in other domains. The advantage, besides those for the generation process itself, are the relinquishment of an explicit dialogue-graph, whose poor flexibility in case of extending or modifying the domain knowledge is a well known problem.

The sequential text planning is warranted by defining discourse related rules as being most general rules, the grammar structures of the ontology as being micro-planning rules, and XTAG as being the most fine-grained ones.

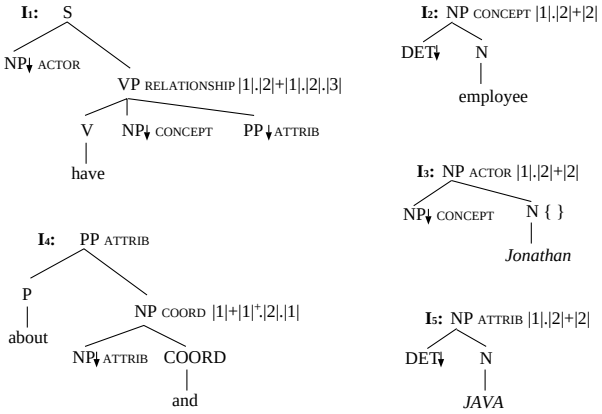

Figure 2: Some associated grammar structures for the instantiated ontology

The snippet depicted in Fig. 2 gives a sketch on what is being instantiated during feedback generation, and what is known directly from the uninstantiated ontology. Every tree, whose leaves are not written in italics, is member of the grammatical structure as is, without further modification[1]. Those trees, whose leaves are in italics, are automatically instantiated from uninstantiated ones, after the document processing (Fig. 2 does not show uninstantiated trees). Actually, the document processing does instantiate the ontology by trying to

[1]The creation, however, can be partly automated, since the information attached to the leaves are found in the ontology itself.

match the structures with the document by instantiating the underspecified leaves with the document's content and if that does not fail, it instantiates the ontology according to the matching values. The example above allows to generate sentences from *Jonathan has knowledge about JAVA and C++*, *The employee Jonathan has knowledge about JAVA* and even *Jonathan and Martha have knowledge about JAVA*, if there would be an I_3' for $Martha$, and an appropriate ontology instantiation for her knowledge. This flexibility is a direct result of the grammar formalism's flexibility, for example, I_4 is responsible for an arbitrary large enumeration of attributes.

6 Conclusion

This paper proposes an innovative architecture for an interactive information retrieval system which differs from common architectures for its unique knowledge representation which integrates syntactical and domain knowledge in such a way that it allows for the implementation of natural language driven document analysis processes, user input analysis processes, and feedback generation processes without having different knowledge bases and formalisms for each task. This aim is tried to be achieved by a unique combination of syntactical knowledge and domain knowledge which allows for a certain situation of the two different knowledge sources. The result is a tight knowledge representation which is abstract enough to be maintenable and concrete enough to produce natural language feedback.

For the future much work remains to be done. First, we need to investigate whether the combination of ontologies with supertags is suitable and the automatic instantiation builds a sufficient large knowledge domain in order to improve retrieval precision and recall. Second, the intended natural language query formulation is dependent of the training quality of the domain specific corpus. Further studies are needed to evaluate the usability of supertags to cover common queries. Finally, and probably the most intricate, we want to approach to a fully automated building of a domain specific information retrieval system: starting by the automated modelling of a portal (i.e. gathering and structuring documents to a special content) in order to build a document collection which can be preprocessed and domain knowledge can be extracted to a common representation and ending with the automatic generation of a retrieval interface supporting NL queries and automated feedback generation . Only the domain of the portal which should be build and searched through must be described at the beginning. An automated support for this task is also indented for further developments.

References

R. Chandrasekar and Bangalore Srinivas. 1997. Using supertags in document filtering: The effect of increased context on information retrieval effectiveness. In Ruslan Mitkov and Nicolas Nicolov, editors, *Procs. of Recent Advances in Natural Language Processing*, Tzigov Chark, Bulgaria, September.

W. S. Cooper. 1988. Geeting byond boole. *Information Processing and Management*, 24:243–248.

Christine Doran, Dania Egedi, Beth Ann Hockey, Bangalore Srinivas, and Martin Zaidel. 1994. XTAG system — a wide coverage grammar for english. In Nagao (Nagao, 1994), pages 922–928.

Melanie Gnasa and Karin Harbusch. 2002. Evaluation of search tactics of it-professionals in the framework of a boolean retrieval model. Technical report, University Koblenz-Landau, Computer Science Department.

Karin Harbusch and Jens Woch. 2000. Reuse of plan–based knowledge sources in a uniform tag–based generation system. In Anne Abeillé and Giorgio Satta, editors, *Procs. of the 5th International Workshop on Tree Adjoining Grammars and Related Formalisms (TAG+5)*, pages 245–248, Paris, France, May 25–27. University of Paris 7.

Karin Harbusch and Jens Woch. 2002. Integrated natural language generation with schema-tree adjoining grammars. In *Procs. of the 3rd International Conferences on Intelligent Text Processing and Computational Linguistics*, Mexico, Mexico City. Springer-Verlag. LNCS.

Makoto Nagao, editor. 1994. *Procs. of the 15th International Conference on Computational Linguistics (COLING)*, volume 2, Kyoto, Japan, August 23–28.

Bangalore Srinivas. 1997. Performance evaluation of supertagging for partial parsing. In *Procs. of the 5th International Workshop on Parsing Technologies (IWPT)*, Boston/USA, September.