LIBER QUARTERLY, ISSN 1435-5205

K.G. Saur, Munich. Printed in Germany

Information Retrieval and Search Engines in Full-text Databases

by HARTMUT ZILLMANN

ABSTRACT

OSIRIS and *ELIB* are two cooperative projects at the Osnabrück University supported by the German Research Society (DFG) and the Ministry of Science and Culture (MWK) Niedersachsen (Germany). They deal with natural language retrieval systems and with indexing techniques for full-text databases using natural language processing. Very complex processes in the context of syntactical and semantical analysis of textual phrases including sophisticated valuation criteria could be implemented in large relational databases with online realtime indexing requirements.

1. PROBLEMS IN ORDINARY RETRIEVAL SYSTEMS

Thematic queries in databases of any kind (bibliographical databases, fact databases, full-text databases etc.) presently require the application of Boolean operators („and", „or" „not") to structure complex queries. Questions precisely defined in natural language, e.g.

- Auslandsberichterstattung in den Massenmedien
- cognition in animals
- Fremdsprachen in der Grundschule
- der Wald im Unterricht

must be fragmented by Boolean operators into structures that have little in common with their original syntactical-semantical construction. The modification of a subject, explicitly formulated by the user with an expression like „in den Massenmedien", „zur Zeit der Stauffer", „in the North-East of France", „by means of spectral analysis", „im Unterricht", and so on is not recognized as such, but must be treated in the query as a separate element:

- auslandsberichterstattung *and* massenmedien
- cognition *and* animal
- fremdsprache *and* Grundschule
- wald *and* unterricht

The meaning of the original question cannot be reconstructed. The query „africa *and* television" may originate from the two completely different phrases

- africa on television *or*
- television in africa

Accordingly, the presentation of search results is incidental and vague, and very often shows as a result „no hit" or „too many hits". - Extended retrieval techniques (truncation, additional discriminating information, proximity etc.) as a rule require the user's detailled knowledge of the data structures. In library databases, for instance, this generally results in the user first of all learning to put queries like the librarian who is familiar with the data structures. Usual relevance-ranking-techniques cannot solve these problems either.

Incalculable search results arise in large full-text systems. The fact, that a document contains two search terms (e.g. „women" and „politics") - ignoring the syntactical-semantical context - means really nothing for the potential relevance of the document.

2. The OSIRIS Project in Osnabrück

OSIRIS is a cooperative project of the *University Library* and the *Institute for Semantic Information Processing (ISIV)* at the University of Osnabrück, supported by the German Research Society (DFG) and the Ministry of Science and Culture (MWK) Niedersachsen (Germany) in the years 1996 - 1999.

The Institute for Semantic Information Processing at the University of Osnabrück is a cooperative institution of the Faculty of Language and Literature and the Faculty of Computer Science and Mathematics. It deals with the basics of the theoretical analysis, technical design and realisation of complex information processing systems. This is seen as an interdisciplinary task of computer science, computational linguistics, linguistics, and cognitive science.

The University Library of Osnabrück, founded in 1975, is an academic library serving for 12,000 students. The library holds 1,000,000 volumes and subscribes to 6,200 periodicals. It has an annual budget of ca. 3 mio. DM and a staff of 80 people.

It is the objective of the OSIRIS project to significantly improve formal and subject queries in information retrieval systems by applying a knowledge base. The OSIRIS knowledge base is fully automatically generated from the indexing elements already contained in the database.

The use of the OSIRIS developments is not restricted to library applications. A simple, natural language approach for databases, which guarantees qualified search results is able to solve a lot of problems in retrieval systems in general.

The OSIRIS developments were transferred to other applications. The best known systems using the OSIRIS technology too are *BREWIS* (BrEmer WirtschaftsInformationsSystem, State- and University Library Bremen) and an open system architecture called *CAMBase* (Complementary and Alternative Medicine, University of Witten-Herdecke).

3. THE OSIRIS SYSTEM

The OSIRIS system accepts natural language input for thematic queries. The user's input is interpreted syntactically and semantically by a compiled declarative grammar and is processed in an adequate way in the OSIRIS knowledge base, not losing the recognized syntactical-semantical connections by a reduction to Boolean operators. The OSIRIS application at the University Library of Osnabrück considerably improves the users' catalogue searches, first of all because of the possibility of natural language input.

The input for a thematic query in an OSIRIS system is the completion of the phrase „Searching for information in the field of ..." presented at the user interface. Thus complexity and ambiguity of input are reduced in a way that seems very natural to the user while the parser module profits from the input's comparatively poor syntax and semantics.

The OSIRIS components for natural language processing are optimized for the german and english language today (phonetics, morphological reduction, analysis of compound words).

The development of the OSIRIS system started in September 1996. A prototype completed in March 1997 was successfully presented at the CeBIT '97 in Hannover as version 1.0 of the OSIRIS system. Version 1.1 has been released in August 1997 for users of the University Library of Osnabrück. The actual version running in the University Library of Osnabrück is V. 4.0.

The most important scientific and technical development in the OSIRIS-System is, that very complex processes in the context of syntactical and semantical analysis of textual phrases including sophisticated valuation criteria could be implemented in large relational databases with realtime indexing requirements.

4. THE OSIRIS SEARCH ENGINE IN FULL-TEXT DATABASES

4.1 The ELIB Project: Electronic Library

The ELIB project started as a common initiative of the *Faculty of Mathematics and Computer Science* and the *University Library of the Osnabrück University*. Its objective is, to build up an ordered and well structured collection ('a library') of electronic scientific resources accessible via the WWW.

The project was supported by the German Research Society (DFG) and the Ministry of Science and Culture (MWK) 1997 - 1999.

The most important result of these developments is an appropriate infrastructure for the use of electronic scientific information at Osnabrück University.

The acquisition of electronic scientific information was a cooperative task carried out by scientists and librarians. The materials represented in the Electronic Library have a large variety:

- Electronic Versions of Scientific Journals
- Preprints
- Multimedia Courses
- Archives (Midi-Server, News Agencies, ...)
- Databases (interactive Java Applications)
- Software

- Scientific Societies
- ...

All the information the Electronic Library needs for representing these materials in an appropriate way are gathered by a WEB-Crawler. Indexing for the underlying database is done using the OSIRIS indexing tools including natural language processing.

4.2 Specific Problems in Full-text Databases

In databases of this kind - e.g. full-text databases for scientific information in the internet - several other problems arise.

The first problem is, that these databases usually grow very fast. The WEB-Crawler of the Electronic Library indexed about 8 GB of data in a few weeks with an average of 700 MB per day. At the moment we hope, that the database will reach its saturation point in some sense with 20 GB of indexed data. Compared with that the OSIRIS database at the University Library contains 2,5 GB of data (1 mio. records) only.

In this context of a full-text database the OSIRIS searching and indexing algorithms proved to be very efficient.

Very essential is, that a large full-text database of this kind has to deal with an unlimited size of interim results in the matching algorithms. Usually interim results are limited to a certain size (e.g. 15,000 document addresses). A database user, searching with terms like „Geschichte" and „Deutschland" in a „german speaking" database will get certain error messages in those limited databases. Very often this problem could only partially be solved only with very expensive hardware investions.

The Crawler-database of the Electronic Library needs matching algorithms for 40,000 - 90,000 document addresses as a rule, e.g. for queries like „women in mathematics". These circumstances asked for a special development we called „learning inverted file".

4.3 Learning Inverted File

Online realtime indexing and simultaneous fast retrieval requirements based on (well known) inverted file structures for indexes are in some sense contradictory for relational databases. Very often the only solution was to invest in the hardware configuration.

The generating process for the learning inverted file is based on the users „click stream". Terms, an inverted file index is necessary and usefull for, are learned from the users input. Updating processes for the inverted file are running periodically in the background. So, the system avoids the usual overhead for global inverted file structures and simultaneous online realtime indexing is possible too. The learning rate of the inverted file structure for the OSIRIS-database in the University Library was very high and the initial phase of this index was very short in time.

The special effect is, that these techniques allow to generate the OSIRIS-database of the University Library and the Crawler-database of the Electronic Library on a PC-architecture (costs: DM 5,000 - 7,000). Both databases are driven without the usual limits for interim results in the matching algorithms.

5. Indexing Terra Bytes of Data: Multidimensional Inverted File Structures

The latest developments in the OSIRIS-project deal with indexing techniques in terra byte areas.

In the context of these developments the database plays the role of a „record container" only. Especially developed, multidimensional index structures - realized as database blobs - guarantee very fast indexing processes. The search algorithms on the other side may be considered as a process in which searching and matching is the same task. These techniques are very promising and show an interesting behaviour in time for queries with several search terms.

Searching with several search terms leads to precise result sets in the semantical and syntactical context. Searching with four or more terms may be faster than searching with three or two terms only. The generating process for this index structure needs some hypothesis about the word statistics of the underlying document domain.

Dr. Hartmut Zillmann
Universitätsbibliothek Osnabrück
Postfach 4469
49034 Osnabrück, Germany
Tel: +49 541 9694359
Hartmut.zillmann@ub.uni-osnabrueck.de