PACIS 1993 Proceedings

Pacific Asia Conference on Information Systems (PACIS)

December 1993

Assertive Retrieval System for Textual Databases

Ernst Schuegraf
St. Francis Xavier University Antigonish

ASSERTIVE RETRIEVAL SYSTEMS FOR TEXTUAL DATABASES

ERNST J. SCHUEGRAF, Department of Mathematics and Computing Sciences,
St. Francis Xavier University Antigonish, Nova Scotia. B2G 1C0. Canada.

ABSTRACT

A brief survey of the use of some artificial intelligence techniques in document retrieval systems is given. The paper explores the addition of deductive capability during the search process in a retrieval system, so that an answer to a query can be inferred from information found in several records.

Information systems that provide scientists with fast access to the bibliographic references and abstracts of the latest publications in their field are an essential tool for working in highly competitive areas such as computer or bio technology. These so-called document retrieval systems may respond to standing query of a user commonly called a user profile. In this case the profile remains constant and the database changes periodically over time. These SDI systems (Selective Dissemination of Information) alert the user to new publications in the literature in his field of interest. Researchers that want to explore a new area and want to retrieve the relevant literature of the last several years submit retrospective queries to a retrieval system. These one-shot queries search a database only once [1]. In this case the database remains constant and the queries change. Many information systems that provide that service are available at reasonable rates, and the databases available cover almost any field.

Current document retrieval systems such as DIALOG use Boolean operators and search terms (with or without truncation) to formulate a query. The system searches the database for matching terms and if the logical connections between the terms is satisfied, it submits to the user a list of relevant documents. Relevance feedback [2] and other statistical techniques are frequently available to improve precision and recall of such searches. However, researchers in information retrieval believe that the statistical techniques have reached their performance limit in terms of precision and recall [3], and that IR systems could benefit from incorporating artificial intelligence techniques.

The three most promising areas of AI were deemed to be expert systems, knowledge representation, and natural language processing[4]. Many applications of intelligent information retrieval have found their way into libraries and their use covers a wide spectrum. A survey of these applications is given by Schuegraf [5].

The first adaptation of expert systems to document retrieval systems was through the development of expert systems as intermediaries [6] between users and databases.

Automatic indexing of documents is another area in which attempts were made to incorporate expert systems. Most noteworthy are the systems described by Humphreys [7], Driscoll et al. [8] and Schuegraf [9].

Other research experimented with replacing the standard text representations of documents with rule- or logic-based representations [10,11,12]. These experiments have shown that the difficulty is not in being able to answer queries from text, but getting the large amounts of text into a suitable representation by automatic means.

Natural language processing offered syntactic and semantic analysis for manipulating of natural language queries, and derivation of "concepts" that take the role of search terms in Boolean queries [13]. Many of these systems that showed promise were using legal text [14]. This promise is due to the fact that legal text has many standard phrases and deals with a narrow domain.

These previously described intelligent components of a retrieval system are not part of the search engine. Croft [15] has made an attempt to add intelligence to the search engine by selecting different search strategies based on user requirements.

The basic premise of current document retrieval systems is that terms or concepts of a query are matched against those of a single document and if the match is successful the document is relevant. The match must occur in one document.

Most retrieval system implementations are done by sequential searching in SDI systems, or use of an inverted index for retrospective searches. The search engine treats a document as being self-contained and no information is carried from the search of one document to another.

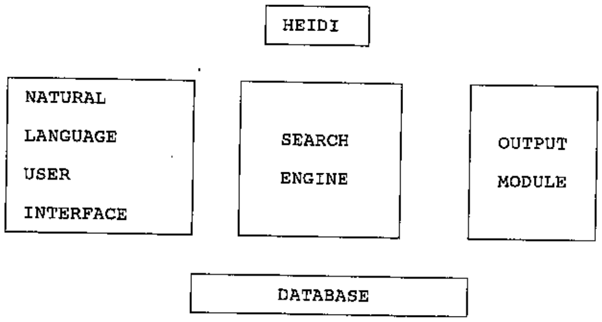

Figure 1. Architecture of an Assertive Document Retrieval System

An assertive document retrieval system adds an extra component to a system as seen in Figure 1. This module is responsible for attempting to prove assertions that involve more than one document and may be called HEIDI (Heuristic Engine for Inter Document Inference). HEIDI may answer queries that need inference between two or more documents. What effect that module can have on a query is illustrated by an example.

EXAMPLE: *Which famous Polish pianist was a signatory to the Treaty of Versailles?*

The skeleton Boolean query that may be extracted from that natural language query can be represented as follows:

QUERY:
Pianist
& Polish
& Treaty of Versailles
END

•Query expansion may include such terms as "signatory" or "famous", but unless some synonyms of these terms are included the success of a match is very small, because the same language may not be used in the database. There is a good chance that the database has no single record that satisfies the query, because the query involves two "unrelated" areas. However, there may be some documents dealing only with the Treaty of Versailles, or famous pianists, in the database. Simple inference using information in two records, can produce the answer, provided there is some link.

Such situations are common in deductive question-answering systems, where normally inference is produced from a whole set of assertions. Hypertext systems let the user answer such questions too. It is the users responsibility to establish the necessary inference through browsing, and can be done only if links between the concepts exist. Database management systems can handle such questions through a join operation provided the attribute on which the join is based is known.

An assertive document retrieval system only needs the module HEIDI and modifications to the output module. The operation of HEIDI is transparent to the user and does not affect the query formulation. However, users must be alerted if the answer is contained in several documents and has been obtained by inference. There may be some degradation in performance under certain circumstances.

The implementation of HEIDI presents some interesting questions.The designers of any document retrieval system that attempts to answer queries such as the one posed in the example have to decide whether all documents that satisfy the query must be found or wether one answer to the query will suffice. In the latter case there may be no need for HEIDI at all if a satisfactory answer is found within one document. Otherwise if no matching document is found HEIDI can wait until the search of the database is completed before it is trying to establish an answer by inference from several records.

When examining records in the database for a match with the query four different cases may arise.

I. All search terms have been found and the query logic is satisfied. This is the desired answer and the document must be output to the user.

II. None of the search terms was found in the query. The document may be completely ignored.

III. All search terms are found but the logic of the query is not satisfied.

IV. Not all of the search terms are found in the document, there is at least one search term missing.

Cases I and II are straightforward, Case III is more intriguing. The document obviously has some relevance though the Boolean logic is not satisfied. In this case the document and the natural language query are subject to a detailed semantic and syntactic analysis to see if they can satisfy the query. Some reasoning about some of the assertions contained in the query and the document will take place. The reason for this apparent backward step is that the skeleton boolean query was constructed to permit the use of a sequential search or an inverted index. However, there is some loss of information when generating the query with regard to word order, word type or syntactic component of the sentence. After the analysis and inference it can be decided if the document should be output to the user.

Case IV is obviously the most challenging in the implementation of HEIDI. To be able to draw inferences later on HEIDI must store the document identifier, the search terms that were found in set {A}, and the ones that were missing in set {B}. This has to be done for all documents that fall into the case IV category.

To answer the query using two documents D1 and D2 the following conditions must be satisfied. Assume set {Q} is the set of all search terms in the query.

$$\{A\}_{D1} \cup \{B\}_{D2} = \{Q\}$$
$$\{A\}_{D1} \cap \{B\}_{D2} \neq \phi$$

The first condition guarantees that all query terms are found in the two documents, and the second one guarantees the existence of at least one common search term between the two documents. If these conditions are satisfied the semantic analysis and inference part of HEIDI will be employed to see if the query can be answered successfully.

A detailed description of the internal algorithms of HEIDI is tedious and beyond the scope of this paper.

To return to the example, one document tells us that the treaty of Versailles was signed in 1919 and that it was signed for Poland by Ignace Paderewski. Another document about famous pianists around the turn of the century also lists Ignace Paderewski. Thus our query can be answered and the link between the documents is the actual answer, the name of the person.

The addition of an inference facility to a document retrieval system provides human associative capability for large textual databases. It can retrieve information that cannot be found with the standard Boolean search engines. It thus expands the capability of document retrieval systems to allow human reasoning and deductions about the information found in documents.

References

[1] Salton, G.A. *Automatic Text Processing*, Addison Wesley, 1989.

[2] Harman, D. "Relevance Feedback Revisited," Proceed 15 Intl. ACM Conference on Research and Development in Information Retrieval, *ACM Press*, pp. 1-10, 1992.

[3] Croft, B.W. "Approaches to Intelligent Retrieval," *Information Processing and Management*, Vol. 23, pp. 247-254, 1987.

[4] Brooks, H.M. "Expert Systems and Intelligent Retrieval", *Information Processing and Management*, Vol 23(4), pp. 367-382, 1987.

[5] Schuegraf, E.J. "A Survey of Expert Systems in Library and Information Science", *Canadian Journal of Information Science*, Vol 15(3), pp. 42-57, 1990.

[6] Pollitt, A.J. "Can Search: An Expert Systems Approach to Document Retrieval, "*Information Processing and Management*, Vol. 23, pp. 119-138, 1987.

[7] Humphrey, S. "MedIndEx System: Medical Indexing Expert System", *Information Processing and Management*, Vol 25(1), pp. 73-78, 1989.

[8] Driscoll, J.D., Rajala, D., Shaffer, W., Thomas, D. "The operation and performance of an artificially intelligent keywording system", *Information Processing and Management*, Vol 27(1), pp. 43-54, 1991.

[9] Schuegraf, E.J., van Bommel, M.F. "An Automatic Document Indexing System Based on Cooperating Expert Systems: Design and Development." *Canadian Journal of Information and Library Science*, (In Press).

[10] Croft, B.W. "I^3R: A New Approach to the Design of Document Retrieval Systems," *Journal of ASIS*, Vol. 38, pp. 389-404, 1987.

[11] Simmons, R. "A Text Knowledge Base from the AI Handbook," *Information Processing and Management*, Vol. 23, pp. 321-340, 1987.

[12] McCune,B.P.,Tong,R.,Dean,J., Shapiro,D.G. "RUBRIC: A system for rule-based Information Retrieval", *IEEE Transactions on Software Engineering*, SE-11, pp. 939-944, 1985.

[13] Mauldin, M.L. "Performance in FERRET: A Conceptual Information Retrieval System," Proceed. 14th Intl. ACM Conference on Research and Development in Inform. Retrieval, *ACM Press*, pp. 347-355, 1991.

[14] Gelbart, D., Smith, J.C. "Towards a comprehensive legal Information Retrieval System", *Proceedings 1990 Conference on Database and Expert Systems Applications*, Springer Verlag Vienna, pp. 121-125.

[15] Croft, B.W. "The Use of adaptive mechanisms for selection of Search Strategies", *Proceedings of the third ACM-BCS Conference on Research and Development in Information Retrieval*, C.J.Rijsbergen ed. Cambridge England: Cambridge University Press, July 1984, pp 95-110.