DOCUMENT RESUME

ED 411 835 IR 056 662

AUTHOR Weiner, Michael L.; Rusch, Peter F.
TITLE New Searching Technologies and Interfaces.
PUB DATE 1996-00-00
NOTE 5p.; In: Online Information 96. Proceedings of the International Online Information Meeting (20th, Olympia 2, London, England, United Kingdom, December 3-5, 1996); see IR 056 631.
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Computer Interfaces; *Information Retrieval; Information Seeking; Information Systems; *Information Technology; Librarians; Navigation (Information Systems); *Online Searching; *Psycholinguistics; Reference Services; Research Libraries; Search Strategies; User Needs (Information)
IDENTIFIERS Natural Language; *Query Processing

ABSTRACT

The DR-LINK (Document Retrieval through Linguist Knowledge) search system was created to help automate the process that research librarians use to convert information needs, as stated by users, into the information retrieval process. DR-LINK resulted from participation in the United States Government initiative called Tipster that was sponsored by the Advance Research Project Agency. DR-LINK contains several independent retrieval technologies that are blended to create a new information retrieval process; central to this new process are the concepts of psycholinguistics. The benefits of the DR-LINK search system permit natural language queries to be input by trained information specialists or novice users. For the novice, the query can be formulated as an expression of interest to the research librarian who would convert the information needs into the information retrieval process. For the experienced searcher, the process allows more time to be given to analysis of results. Searching goes beyond the traditional search term-based system, allowing for automatic term expansion and the removal of ambiguities of search terms. Consequences of events and temporal references may be part of the query. Results are presented in a ranked order presenting the most relevant items to be scanned first. (Contains 18 references.) (Author/AEF)

New searching technologies and interfaces

Michael L. Weiner and Peter F. Rusch
Manning & Napier Information Services, USA

Abstract: *The increasing demand for information has resulted in an increasing number of duties for the corporate library staff. With the advent of the library without walls, librarians and information specialists are tasked increasingly to provide a variety of digital library services beyond the traditional intermediated search. Searches of high economic value remain the purview of the skilled information specialist who has both subject and information skills to apply to the searches required. Preliminary screening searches, competitive intelligence searches and searches of lower economic value are now performed by the end-user of the information satisfying his or her own needs. These end-users may search infrequently and they usually have fewer formal information skills than their professional information colleagues. A new range of databases and search engines with natural language interfaces are available to assist the occasional user perform high-quality searches. The processes embedded in these systems provide an ease of use more closely related to the research skills of the end-user. The techniques of database preparation, query formulation and search execution using the DR-LINK interface will be described.*

1. Introduction

For many years research in information retrieval has followed several paths leading to development of both experimental and commercial systems. Historically, the straightforward word- and phrase-indexed search systems were commercially implemented earliest. These systems took advantage of a confluence of events in technology and database development. Improvements in computer technology through the 1960s brought both faster central processors and larger random-access disk storage, each of which benefited both information retrieval and database development.

The improving computer technologies were applied to database collection, processing and production usually as an adjunct to some printed publication process. Although initially small, many of these databases today contain millions of records and billions of bytes. Some organisations followed dual development paths, working on both database development and information retrieval development. By the early 1970s the commercial database industry was generally divided into database development, practised by a group of producers (usually publishers), and commercial information retrieval from these databases, offered by the growing online services of the day.

Databases adopted a variety of old and new techniques in creating their content. Standards for bibliographic data evolved and were used. Subject specific indexing schemes were applied and augmented as the databases grew. Information retrieval depended solely upon the words and phrases entered by the database producer. Typically, database production was an extension of manual production processes that were assisted by computer technologies in the capture and preparation of magnetic distribution media. Accordingly, database content was usually brief and was (and still is) subject to human error. Thus, concept searching was performed using word, or phrase, searching that required allowance for any number of misspellings and word variants. Inference and nuance were lacking. Controlled vocabulary indexing relieved some of the problems encountered but led to the need for information specialists to search effectively the specialised vocabularies that developed. Further, these controlled vocabularies evolved as the subject matter itself evolved and changed. As time passed and databases grew larger, even controlled vocabularies lost some of their value since they were not consistently applied throughout the time span of a database.

Commercial, online information retrieval systems of today are often characterised as 'Boolean' or statistical/probabilistic retrieval systems. Perhaps a more apt designation is 'search term based' (or STB) systems. The systems are designed to retrieve search terms, whether they are words or phrases. Databases are prepared for search by deriving search terms and presenting these to the user. Such information retrieval systems are truly WYSIWYG systems in the sense that even misspellings appearing in the databases become search terms and retrieving these items depends on anticipating the misspelling. Users are generally required to specify the databases where they believe the information resides. They give a detailed query composed of search terms that are words and phrases hoped to describe the query in the context of the databases' coverage. Results are presented as a group of 'hits' that must be sorted manually for relevance to the query posed.

2. Improvements to information retrieval

As the number and size of databases grew so did the information retrieval problems associated with them. New kinds of databases appeared. Commercial information retrieval services grew, added more databases and adapted to the changing needs of both databases and users. More full-text was available as the result of so much text being captured from word processing. No longer were secondary (abstracting and indexing) databases the only means of information retrieval. Search and delivery of the complete document became expected.

At the same time, computer technology and telecommunications were becoming ubiquitous so that the number of people equipped for information retrieval grew enormously. With relatively few information specialists to handle an increasing number of requests, new computer-based methods were needed to assist this growing number of information seekers. Results from the information retrieval research community were refined and implemented to advance information retrieval. Many of these techniques are incorporated into DR-LINK (Document Retrieval through Linguist Knowledge) that was created to help automate the process that research librarians use to convert information needs, as stated by users, into the information retrieval process.

3. The DR-LINK search system

DR-LINK resulted from participation in the US Government initiative called Tipster that was sponsored by the Advanced Research Project Agency (ARPA). This was an effort of significant proportion to stretch the limits of information retrieval. The test queries used to evaluate the responses of competing information retrieval systems included temporal and consequential aspects requiring linguistic knowledge. Indeed, the idea was to expand the capabilities of information retrieval to process not only queries formulated in natural language but to search and retrieve natural language 'documents'. The results have been impressive, and substantial improvements in precision and recall can be demonstrated.

To illustrate the expectations of the project look at an example Tipster query on the topic, 'Black Monday' (Ref 1):

> A relevant document will contain at least one reason why US stock markets experienced a huge drop on 19 October 1987, losses of equity so large that markets were said to have crashed (the Dow, for example, lost 508 points on that day alone); the date of the crash has become known as 'Black Monday.' A preferable document would contain a detailed analysis of the crash. The best document would link analysis of events to actions taken or recommendations made by federal authorities or the stock markets to prevent further crashes. *Not* relevant are reports which simply reference, without analysis, 'Black Monday,' such as anniversary stories generated by the press around every October 19th.

The DR-LINK search system was designed and developed to handle queries of the kind given above. It has been demonstrated to do so successfully. DR-LINK actually contains several independent retrieval technologies that are blended to create a new information retrieval process. Central to this new process are the concepts of psycholinguistics. Scientists in the field of psycholinguistics have determined that humans process text on at least six known levels: morphological; lexical; semantic; syntactic; discourse; and pragmatic. The **discourse** level refers to the all of the meaning we derive beyond the word and sentence level such as reference to information provided previously in an article. It also refers to the difference between definite information related to the past, and prediction and expectations in the future. The **pragmatic** level refers to that which we know outside of the text of an article. References to general nouns imply the specifics: for example, 'third world' implies the countries that are known to comprise that grouping.

Working with this model of six levels of text processing, the value of applying them to information retrieval can be illustrated. For example, if we encounter 'White House' in a document about the US Government or the US President, we know at a pragmatic level in this context it refers to a branch of the United States government, physically located in Washington, DC in the USA in North America. We also know that it is *not* a European nation, nor a Third World country, nor is in it South America.

Natural language processing (or NLP) systems such as DR-LINK use this kind of ancillary information in a way that permits them to become part of the search process. Both the databases at the time of loading and the search query at the time of execution are processed using these relationships to build a richer environment for information retrieval.

A search query about future, evaluative comments on the outcome of a specific pending event is a very different request from one about historical results. Many queries contain this temporal and consequential component. Statistical/probabilistic online information retrieval systems rarely deal with these degrees of subtlety. To the extent it is available, it may be part of a controlled vocabulary indexing system. Thus, effective use of statistical/probabilistic systems may require extra efforts by trained information retrieval specialists to formulate the query iteratively and review the results in detail.

A search system that interprets and expands on a user's natural language query automatically can expand the access to information without the help of a trained information retrieval specialist (Ref 13). This works to the benefit of the trained information specialist in two ways. Ideally, it frees the trained searcher to focus on the most difficult and strategically vital searches. Even for the information specialist, the improved results of a search permit more time for analysis and interpretation. Overall usage of internal and external information in the organisation increases.

4. Specific applications of DR-LINK

In the area of searching for prior art for patents, this type of linguistically informed search capability could change fundamentally the manner in which patents are prosecuted. If the patent offices accessed such a powerful search system, one that could overcome some of the complexities and obfuscations of language, what patent applicant could afford to be without a comparable, comprehensive search?

With a complete database collection it would be possible to search not only newspapers, magazines, abstracts and bibliographic data but also entire reference books, specialised encyclopedias and complete doctoral dissertations. All of these represent significant amounts of non-patent prior art that has been impractical to search previously.

DR-LINK, with its large collection of databases containing computer software prior art, is able to locate significant prior art that could be used to challenge patent claims. An example query of this nature is:

> 'I would like information about searching for textual, graphical, audio or video information in a multi-media software environment that combines these attributes into a single, interactive, related program for entertainment or education.'

The top ten articles retrieved from the collection of databases containing computer software prior art are on point and should be reviewed seriously with respect to the disclosures made in them (Refs 2–11).

Competitive intelligence is similar in the desire to obtain *all* prior information. In competitive intelligence the objective is to gather all information about a competitor and its products and services. Using a natural language profile that searches a variety of document types from many localities it is possible to alert an organisation to any competitor's new or rumoured products. Such a system can return results of reasonably relevant information, interpreted, expanded and filtered to match what was needed rather than results that merely matched lexically.

Likewise, a market survey search is best done with a variety of sources to get a broad view of some market opportunity. An example of such a query is:

> 'I would like information about point of sale devices, linked to computer systems, that scan bar code displays and then compare the product code to certain data to generate discount coupons for the customer automatically at the point of sale such as a supermarket checkout counter cash register.'

The top five ranked articles to appear from this query have titles that are relevant to the topic of point of sale coupons in supermarkets (Refs 14–17). The rank order of the documents is that given among the references cited and illustrates the benefits of retrieval over a broad collection of diverse databases. More than a dozen databases were searched, ranging from full-text, primary sources such as newspapers and research reports to secondary abstract and index.

5. Conclusion

The benefits of the DR-LINK search system permit natural language queries to be input by trained information specialists or novice users. For the novice, the query can be formulated as an expression of interest to a research librarian who would convert the information need into the information retrieval process. For the experienced searcher, the process allows more time to be given to analysis of results. Searching goes beyond the tradition search term-based system, allowing for automatic term expansion and the removal of ambiguities of search terms. Consequences of events and temporal references may be part of the query. Results are presented in a ranked order presenting the most relevant items to be scanned first.

Acknowledgements

The authors want to thank the staff of TextWise Co. who helped with some of the examples presented.

Peter F. Rusch
355 Verano Drive
Los Altos
CA 94022
USA
Tel/fax: +1 (415) 941 8120

References

[1] Liddy, E. (1994) Information retrieval via natural language processing or an intelligent digital librarian, *ASIS Mid-Year Conference*.

[2] Mackay, W. and G. Davenport (1989) Virtual video editing in interactive multimedia applications, *Comm. of the ACM*, pp. 13.

[3] Stevens, S. (1989) Intelligent interactive video simulation of a code inspection, *Comm. of the ACM*, pp. 15.

[4] Halasz, F. (1988) Reflections on NoteCards: seven issues for the next generation of hypermedia systems, *Comm. of the ACM*, pp. 29.

[5] Barbour, B. (1990) Enhancing the BBC Domesday Videodisc as a learning environment for New Zealand secondary schools, *Brit. J. Educ. Technology*, 4.

[6] Warwick, M. (1993) Survey of telecommunications in Business (10), *Financial Times*, 12 July 1993.

[7] Sonsino, S. (1993) Survey of information and communications technology, *Financial Times*, 18 April 1993.

[8] Schnase, J.L., J.J. Leggett, D.L. Hicks and R.L. Szabo (1993) Semantic data modelling of hypermedia associations, *ACM Transactions on Info. Systems*, pp. 1

[9] Gibbs, ., D. Tsichritzis, E. Casais, O. Nierstrasz and X. Pintado (1990) Class management for software communities, *Object-Oriented Programming, Comm. of the ACM*, pp. 25.

[10] Davis, J.I. (1992) File 1 — a computer & information technologies platform, *Computer Underground Digest*, pp. 18.

[11] Chen, C.C. (1990) DATAPLEX: an access to heterogeneous distributed databases, *Comm. of the ACM*, pp. 15.

[12] Weiner, M.L. and E.D. Liddy (1995) Intelligent text processing and intelligent tradecraft, *The Journal of AGSI*, July, 60–67.

[13] Feldman, S. (1996) *DR-LINK, DIALOG and TARGET: A Comparative Study* (in publication) (a complimentary copy of this article may be obtained by writing the authors).

[14] Gellene, D. (1994) Check it out supermarkets track top patrons, reward them with custom deals, *LA Times*, 4 August.

[15] Anonymous (1988) Interactive POS video yields instant results, *Chain Store Age Executive*, pp. 1.

[16] Coleman, L. (1988) 'Smart Card', coupon eater targeted to grocery retailers, *Marketing News*, pp. 1.

[17] Bradshaw, D. (1994) Technology: the end of the queue — automated supermarket checkouts will cut waiting time, *Financial Times*, 13 July 1994.

[18] Tomkins, R. (1994) Management (marketing and advertising): time to cut it out — US manufacturers are questioning the value of money-off coupons, *Financial Times*, 18 May 1994.

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

NOTICE

REPRODUCTION BASIS

[X] This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

[ ] This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").