October 1991

Survey, Analysis and Evaluation Criteria of Full Text Systems

Carol Tenopir
University of Tennessee - Knoxville, ctenopir@utk.edu

Gerald W. Lundeen

SIG/SRT, SIG/ALP AND SIG/HCI—FULL TEXT: FROM TUTORIAL TO INNOVATIONS—PART 1. INTRODUCTION

Moderator: Emil Levine

SURVEY, ANALYSIS AND EVALUATION CRITERIA OF FULL TEXT SYSTEMS

Carol Tenopir and Gerald Lundeen
University of Hawaii at Manoa
Honolulu, HI

ABSTRACT

Textual databases, including bibliographic, referral, and full text, have many unique characteristics that impact software selection for text retrieval. Full text databases, in particular, contain records that vary greatly in length and structure. Software for full text retrieval must have powerful retrieval capabilities, whether the software deals with structured text, unstructured text, or a hybrid combination. Besides the standard textual retrieval features, full text software should offer features such as the ability to recognize the grammatical structure of sentences and paragraphs, keyword in context display, and word occurrence and ranking output.

INTRODUCTION

Software for text searching and retrieval has been an area of active development. Many packages are available for the creation of textual databases on microcomputers as well as minicomputers and mainframes. In this paper we will survey the types of text retrieval software available and look at the unique characteristics and requirements for full text databases that impact full text software evaluation criteria.

Textual databases can be bibliographic, referral (directory), or full text or a combination of all three. Bibliographic databases, once the most common type of textual database, contain surrogates of documents, including citations and further descriptive information. Referral databases typically provide names, addresses, phone numbers, and information about organizations or individuals.

Full text databases provide the complete text of articles, books, reports, or other items. They may also include some bibliographic-type information such as subject headings and a structured citation. Full text databases are becoming common these days for several reasons: most organizations now use word processing software to generate textual documents; disk storage costs are coming down and capacities are going up; optical scanning and optical character recognition technology is now a practical method of data entry; and online or CD-ROM full text databases are widely available for downloading.

CHARACTERISTICS OF TEXTUAL DATABASES

Search and retrieval capabilities for textual databases are necessarily different from those of software suited for other types of databases, due to some unique characteristics of typical text databases. Textual databases, whether they are bibliographic, referral, or full text differ from the typical numeric or DBMS application in several ways:

- the data is primarily alphabetic and when numbers are included they are more likely to be treated as text.
- the databases are often very large
- records often have many fields (e.g author, title, source, date, abstract, subjects, location, etc.)
- the fields tend to be variable length and may be very lengthy
- fields often contain repeating values, such as multiple authors, but whether they repeat and how often they repeat varies from record to record
- most applications require searchable access to all or most of the fields in the records
- search capabilities are important. Information is represented in various ways using natural language and sophisticated retrieval techniques are needed. Retrieval is based on searching rather than selecting.

Full text databases can be further differentiated from other types of textual databases. Because of the wide variation in textual documents there is more variation in full text than in other types of textual databases.

- The size of the texts can vary from short memos to multi-volume encyclopedias.
- The database may consist of a single text (e.g. an encyclopedia) or may be a collection of many discrete and independent texts.

- Collections of texts may be relatively uniform in format and size or they may vary (memos, letters, reports, contracts, articles, books).

- Full text databases may exhibit varying degrees of structure
    - Fielded records
    - Amorphous text with appended fixed fields
    - Text with inherent structure (format) e.g. letters with date, addressee, body, sender, etc.
    - Amorphous text with its inherent structure (sentences, paragraphs, other grammatical structure).

SOFTWARE CATEGORIES

Software for full text databases is not always labeled as such, and there is a wide variety in capabilities, costs, and complexities. Because different packages are better suited for different things, it is useful to categorize full text systems by their general characteristics. (These categories are indicative, rather than exact, because they describe typical features and general strengths. Some packages don't neatly fit into only one category.) We have divided full text software into 3 categories: 1) structured text retrieval, 2) unstructured text retrieval, and 3) hybrid text retrieval.

Software in all of these categories are good for handling text; which means they may not be the best for handling numbers. Typically they are poor at calculations, mathematical manipulation, or forecasting. Output is geared to the special needs of information that is structured into words, sentences, and paragraphs and may not have the capability to generate reports structured in rows and columns such as with a spreadsheet or generic DBMS package.

Structured text retrieval packages demand that the creator specify fields and field characteristics before building the database. Text must be structured into these fields either before it is transferred into the program or at input. Because of the field structure searching may be more precise and retrieval may be very fast. The structure allows more control of output formatting. Examples of structured text retrieval packages include BRS/Search, STAR, Personal Librarian, and InMagic. These packages are also often used for bibliographic databases.

Unstructured text retrieval packages accept text files without fielded structure. This is particularly good for existing word processing files or files downloaded from a variety of systems. Although these packages do not allow searching of specified fields, they often recognize grammatical structure of texts such as sentences and paragraphs. These can be used both for searching and output. Some examples of unstructured text retrieval packages include: ZyINDEX, Lotus Magellan, and Gofer.

Hybrid (or combination) text retrieval packages support fielded information as well as unstructured text. The fielded information is often forced into fixed length fields. These fields are typically used for bibliographic data, while the unstructured feature is used for complete texts. The software packages in this category tend to offer the most powerful search capabilities. Examples include: Concept Finder, Concordance, askSam, and Search Express.

RETRIEVAL CAPABILITIES

Retrieval capabilities for full text must be powerful. Certain capabilities, that are by now de facto standard for bibliographic systems, must also be a part of full text systems. These can be considered minimum level retrieval capabilities; that is they must be present in all good text retrieval software. Minimum level capabilities include:

- Boolean logic (AND, OR, AND NOT with the ability to nest)

- comparison operators (greater than, less than, equal to)

- truncation (right hand user-specified as a minimum)

- inverted index display

- free text searching

- word adjacency searching

- field specification

- set building

In addition, there are search features called for by full text that may go beyond what is needed for bibliographic software. These can be considered essential for full text and should be included in all good text retrieval packages for full text. These include:

- Searching within a specified number of words

- Key word in context displays

- Searching making use of grammatical structure, including within a sentence or specified number of sentences and within a paragraph or specified number of paragraphs

- Left hand and internal truncation

- Thesaurus features -- synonym expansion, word profiles, etc.

- Automatic stemming for singular/plural and other word form variations

- Automatic language enhancement, including automatic abbreviation expansion, British/American spelling, etc.

Additional search and retrieval capabilities are available in some full text retrieval packages. These additional features are not standard, so all are rarely found in one package. Still they increase the power and capability of full text databases. These features include:

- Hypertext capabilities
- word occurrence information
- ranked output
- fuzzy sets
- relevance feedback
- sound-alike retrieval and/or other partial match algorithms
- the ability to include images

STORAGE OPTIONS

There are two basic types of full text software packages from the perspective of storage options. Archival text retrieval software requires that a copy of the texts be loaded into the system before it is searchable. If the text is from active files, for example active word processing files, then this means that two copies of the files are needed -- one in the retrieval system and the other in the word processing subdirectory. The structured type of text databases are typically archival. These generally offer the most powerful retrieval and output capabilities.

The second type of full text retrieval package is the text file indexer. These packages search files in place and can generally deal with most popular word processing formats. Here only one copy of the textual data is needed and the software maintains an index which provides rapid access to the content of the files. ZyINDEX, and Lotus Magellan are examples of text file indexer software.

CONCLUSION

These are some of the major features and considerations when evaluating full text retrieval software. What is most important depends on each application and situation. Evaluation requires a thorough analysis of your needs so that the best choice from the wide variations of types and capabilities can be made. The kind of data, its method of creation or capture, its volatility, and uses will determine the specific features needed. There are many excellent packages to choose from. There should be at least one that will meet your needs very well.

FURTHER READING

Carnahan, Ron. "PC Text Management." DBMS (January 1991): 52-61.

Perez, Ernest. "Managing Text." Databased Advisor 8 (June 1990): 83-

Tenopir, Carol. Full Text Databases. Westport, CT: Greenwood Press, 1990.

Tenopir, Carol and Lundeen, Gerald W. Managing Your Information: How to Design and Create a Textual Database on Your Microcomputer. NY: Neal Schuman, 1988.

Tenopir, Carol and Lundeen, Gerald W. "Software Choices for In-House Databases." Database 12 (June 1988): 34-42.