6-1-1992

Innovations in Text Retrieval Software

Carol Tenopir
University of Tennessee - Knoxville

ONLINE DATABASES

BY CAROL TENOPIR

Innovations in Text Retrieval Software

SOFTWARE FOR commercial online systems was mostly developed 20 years ago. The systems are comfortable and familiar to longtime searchers. It is better than many databases or OPAC software and certainly much better than manual ways to retrieve information, but the search features and programming conventions represent a past era in information retrieval software.

New methods and changes in search features are more common in CD-ROM systems, although many of these just emulate online systems. Innovations are easier to accomplish and are the most common in software for in-house databases. A new generation of software provides many enhanced search and display features for textual databases you build and maintain.

Natural language queries

The traditional way to search a text database is to learn the command language of the system or follow menus until told to enter search terms or phrases. In most systems if a user enters a statement like "find articles about the relationship between the rate of inflation and election years," he or she will get either an error message or the system will try to search the entire string as a phrase and report back zero postings.

Natural language input is allowed by some text retrieval software for in-house databases. If the above statement was entered in Personal Librarian, the system would first take out all the stop words (as defined at setup by the database creator) and then search individually for the rest of the words in the input string. It would be reasonable to define "find," "about," "the," "between," "of," and "and" as stop words and perhaps also "articles." Personal Librarian would then search for "relationship," "rate," "inflation," "election," and "years." The database creator specifies at setup what relationship between these words is used—either the default Boolean OR or Boolean AND or adjacency. Natural language input with words joined in a Boolean OR relationship may seem like it would retrieve too many false drops (e.g, all the articles with the word "years" in them). It works on Personal Librarian because documents are ranked and displayed according to a word frequency formula.

Carol Tenopir is Associate Professor at the School of Library and Information Studies, University of Hawaii at Manoa, Honolulu

Word frequency ranking

Traditional document displays present the most current documents first. All items retrieved are displayed in reverse chronological order so you know the first items in a set will be the newest, and the last items will be the oldest. This is great for current events searching, when you want to find the latest developments on a topic, or for updating a search done earlier. It is not so useful for most subject searching when any item in a retrieved set may be useful regardless of publication date. For full-text databases, such "last in-first out" displays mix articles that focus on a topic with those that just mention it in passing.

Word frequency ranking displays first those documents that contain the most occurrences of the search terms or phrases. The documents likely to be most relevant thus come out first, so a user can just stop looking at items once the displayed documents begin to seem less relevant.

Software packages use different formulas for determining word frequency ranking. A common method looks at how many of the individual words occur (a document that contains all of the terms "relationship," "rate," "inflation," "election," and "years" would rank higher than one that had only four or fewer of the five) and then how many times each of the terms occur. A document with all of the terms occurring many times will be displayed first, followed by others in order by decreasing number of occurrences.

Lotus Magellan, Search Express, and Personal Librarian offer word frequency ranking. Personal Librarian also offers the option of displaying a graph that depicts word occurrences in the documents before displaying each document in turn in order. It allows users to change the display order by sorting when date or some other order of display is desired.

Another version of ranking is offered by topic. It ranks and displays documents in order of likely relevance based on the database builder's defined weights for terms. Database creators or users define "topics," hierarchies of words and phrases that define a topic with each word assigned a weight depending on how important it is to that topic. For example, the topic "computer security" may have children of "data security or computer crime"; may have children of "break in," "hackers," "unauthorized access," "virus," etc. The word "hackers" may be assigned a higher weight for this topic than the less specific term "break in." Output ranks are calculated based on the relative weights assigned to those words that occur in each document.

Thesaurus/word equivalencies

Topic's topic definition feature is one kind of user-defined word equivalency feature. In building topics, the database designer sets up all of the terms or phrases that represent that topic; users can browse topic lists and select topics of interest to be searched. All terms in the topic list then will be searched according to the relationships defined by the builder (for example, Boolean OR or Boolean AND between terms or within the same sentence or within the same paragraph).

The problem with this kind of feature is that topics and queries likely to be of interest to users must be anticipated in advance by the database designer. In most textual databases, this is rarely possible. (Alternatively, users can build topics, but this requires time and knowledge.) Word equivalency features in other systems just build synonym lists for automatic searching. ZyINDEX, Search Express, Personal Librarian, and Concept Finder all allow the database builder to specify such automatic synonym equivalen-

cies. For example, "waterways" can be defined as "stream" or "brook" or "river" or "lake." When a searcher enters any of the terms, all will be searched. Some systems will allow phrases to be named as equivalents, e.g., F.B.I.=FBI=Federal Bureau of Investigation=federal agents.

Automatic equivalencies for singulars/plurals, British/American spelling, some abbreviations or acronyms/spelled-out versions, forms of dates, etc., may be labeled a "thesaurus" by the database software. Rarely true hierarchical thesauri—with a formal NT, BT, RT structure—these are more commonly of the *Roget's* type of synonym structure. In most online systems, if a database's hierarchical thesaurus is loaded, it is available for viewing only. (Wilsonline is an exception by supporting automatic invocation of its "use" cross references, so a "use" term is automatically searched if a "used for" is entered.) CAIRS and STAR in-house software include modules for true thesaurus building.

Hypertext

Word assistance features enhance the traditional methods of searching for words in a text. Other search methods go beyond straight word searching. Hypertext was the buzzword of a few years ago. Although the term and concept were coined many years ago by Ted Nelson, it did not catch on until development of the Hypercard program for the Macintosh. Entire conferences were devoted to explaining the hypertext concept and to showcasing software or applications that use hypertext.

In text retrieval, hypertext is an alternative way of providing relationships among related documents or parts of documents. For example, an article about Antarctica may have a section or paragraph about the fauna of the region. This paragraph could be linked to another document all about penguins. The penguin article could, in turn, be linked to articles about flightless birds or relevant paragraphs in larger articles about cold-weather adaptations. Links can also be made between the text and related pictures or other graphics.

Many text retrieval packages have some sort of hypertext, but they range from quite simple to fairly complex. At the simplest level, what is called hypertext just searches throughout the documents in a database to find a term that a user highlights in a document. No "links" were established by the database builder; the system just uses the term as a search input term. A more complex hypertext feature allows the database builder to establish conceptual links or multimedia links.

askSam will do the simplest word matches within one file or from one file to another while supporting more complex links. Topic, Personal Librarian, and Sonar require building the links at database creation.

For any kind of link more complex than just simple word links (finding all documents that have the word "penguin"), hypertext links must be built by the database creator or database "indexer"—it is an intellectual process that requires examining all of the documents in a database and determining where links should be created. It works well with static texts like an encyclopedia; with continually updated document databases, appropriate links must be added at each update.

A similar but more sophisticated way to use a relevant retrieved document as a search query is "like document" retrieval. If a user locates a document that is particularly relevant and wants others "like" it, Personal Librarian will search and find all documents that have the same word occurrence pattern as the seed document. The user doesn't have to specify what it is about the first one that makes it relevant; the system will analyze all of the words and word occurrences to find those that best match.

The price, capabilities, and complexity of software packages vary, but extra features of many provide enhanced text searching. If you are considering building your own database, the new generation of text retrieval software packages are worth a look.

Software

askSam Version 5.0
askSam Systems
PO Box 1428
Perry, FL 32347
800-327-5726; 904-584-5690
$395; educators, $99.95

BRS/Search (mini,micro)
BRS
1200 Rte. 7
Latham, NY 12110
800-235-1209; 518-783-1161
DOS, $2000
multiuser, $5000–$180,000

Concept Finder
MMIM Inc.
566A S. York Rd.
Elmhurst, IL 60126
312-941-0090
286 PC version, $1595
386 PC version, $1995

Concordance
Dataflight Software
10573 W. Pico Blvd., Suite 68
Los Angeles, CA 90064
213-785-0623
standard ed., $495
professional ed., $995
network (up to 10 users), $3000

Inmagic Plus Version 7.2
Inmagic, Inc.
2067 Massachusetts Ave.
Cambridge, MA 02140
617-661-8124
PC version, $1250
multiuser (2 & up), $1950

Library Master
Balboa Software
61 Lorraine Dr.
Willowdale, Ontario
CANADA M2N 2E3
416-730-1896
$179.95

Notebook II Version 4.1
Pro/Tem Software Inc.
3790 El Camino Real
Palo Alto, CA 94306
800-533-6922; 415-323-4083
$189; w/bibliog./convert., $299

Papyrus 7.0
Research Software Design
2718 SW Kelly St., Suite 181
Portland, OR 97201
503-796-1368
PC, $99; Mac version due 1993

Personal Librarian
Personal Library Software
15215 Shady Grove Rd.
Rockville, MD 20850
301-926-1402
PC & Mac versions, $995
XENIX/UNIX version, $1295
network/multiuser, $4975+

Pro-Cite Version 1.4 (PC), 1.3 (Mac)
Personal Bibliographic Software
PO Box 4250
Ann Arbor, MI 48106
313-996-1580
PC & Mac, $395

Reference Manager Version 5.0 (PC), 2.0 (Mac)
Research Information Systems
2355 Camino Vida Roble
Carlsbad, CA 92009
800-722-1227; 619-438-5526
w/capture & formats, $499

Search Express Version 2.51
Executive Technologies, Inc.
2120 16th Ave. S
Birmingham, AL 35205
205-933-5494
single user, $3000
network (5 users), $5000+

STAR
Cuadra Associates, Inc.
11835 W. Olympic Blvd.
Los Angeles, CA 90064
800-366-1390; 213-478-0066

Topic 3.1
Verity, Inc.
1550 Plymouth St.
Mt. View, CA 94043-7600
415-960-7600
Servers, $15,600–150,000
DOS/Mac client software, $795/station

ZyINDEX
ZyLAB Corp.
3105-T N. Wilke Rd.
Arlington Heights, IL 60004
800-544-6339; 708-632-1100
DOS, Windows, $395