5-1-1988

Searching Full-Text Databases

Carol Tenopir
University of Tennessee - Knoxville

BY CAROL TENOPIR

Searching Full-Text Databases

FULL-TEXT DATABASES are being added almost every month to many of the major commercial online systems. In the last five to six years, the complete texts of many magazines, newspapers, and books have joined full-text legal documents and newswires online. Although these materials are not complete replacements of printed works (they do not yet include graphics and certain parts of printed works, e.g., letters to the editor, advertisements, short news items), full-text databases are attractive alternatives to bibliographic databases or, in some cases, to printed publications.

The full-text systems

Some of the major commercial online systems that offer a large number of full-text databases and that are commonly used in U.S. libraries include LEXIS and NEXIS, BRS, DIALOG, STN International (Chemical Abstracts Service Online), VU/TEXT, NEWSNET, and Dow Jones News Retrieval. (Systems less often used in libraries such as CompuServe and The Source have more limited search features and are excluded from this discussion.)

Each of these systems uses some standard search and display features that are also used for searching bibliographic databases and are familiar to all searchers. The major online systems allow every word in the texts to be searched (except stop words) and use inverted index file structures. Standard search techniques include: Boolean operations (typically AND, OR, and NOT), proximity searching (usually adjacent, sometimes within a certain number of words or within the same field), and truncation (most commonly right-hand truncation of either a specified or unspecified number of characters). Typical display features allow the user to specify what fields or combination of fields they wish to see displayed.

Some search and display powers on these systems are especially useful for full-text searching: proximity operators that allow words to be searched in the same grammatical sentence (BRS); proximity operators that allow words to be searched in the same grammatical paragraph (BRS, DIALOG, STN, Dow Jones); proximity operators that allow the user to specify any number of intervening words (DIALOG, LEXIS/NEXIS, STN, NEWSNET, Dow Jones); automatic searching of plurals and singulars (LEXIS/NEXIS, VU/TEXT, BRS); automatic searching of equivalent words such as British/American spelling or abbreviations (LEXIS/NEXIS, BRS); word frequency counts to sort display output or to help with relevance judging (BRS, STN, VU/TEXT); highlighting of search words (LEXIS/NEXIS, VU/TEXT, STN, DIALOG); and display of only those portions of texts that contain the search terms (LEXIS/NEXIS, BRS, VU/TEXT, STN, DIALOG).

The searcher's arsenal

These features need to be part of a searcher's arsenal to search full text most effectively. Experienced full-text searchers recommend replacing the Boolean AND to link concepts with the "same paragraph" operator or within approximately 20 words. A study I am now doing on search techniques for full texts of popular magazines suggests that searching within the same paragraph retrieves on the average the best combination of relevant articles without an unwieldy number of false drops. The Boolean AND operator retrieves more documents (both relevant and not) and is useful when there are several concepts in a search, not all of which could be expected to be mentioned in a paragraph or when one or more concepts is imprecise. For example, a search using AND between the concepts morals or ethics AND televangelists retrieved documents that did not use the terms morals or ethics in the same paragraph with the term televangelists, but that discussed specific types of ethical behavior.

Other studies have shown that the more often search terms occur in a document, the more likely that document will be relevant. Systems that provide word occurrence tables (BRS and STN) or allow sorting by number of times words occur (VU/TEXT) offer a good way to deal with the large numbers of documents that are sometimes retrieved in a full-text search. Highlighting and displaying only those portions of the text that contain the search terms further facilitates relevance judging and makes full-text viewing more cost effective.

Types of full-text databases

The materials that are available in full-text form on each of these systems varies considerably. Speaking of full-text databases as a single entity may be as fallacious as lumping together all people who search databases for their own use as "end users." The amount of information and type of information available in full-text form varies as much as individual end users' experience, expertise, and needs do. Effective search strategies for each type can be expected to vary also. Types of full-text databases readily available include:

- scholarly or technical journals (the American Chemical Society journals on STN International and BRS);
- popular magazines (widely read, nonscholarly magazines such as those available in Magazine ASAP on BRS, DIALOG, and NEXIS. These can be further subdivided into categories: news [*Time* and *Newsweek*]; business [*Forbes* and *Money*]; hobby [*Popular Photography* and *Popular Mechanics*]; political/commentary [*New Republic* and *Nation*]; "women's" [*Ladies' Home Journal* and *Redbook*]; entertainment [*Rolling Stone*, *Sports Illustrated*, and *Teen*]; and popular science [*Science* and *Psychology Today*]);
- newsletters (mostly highly specialized, industry-oriented, available on NEWSNET and NEXIS);
- newspapers (dailies or weeklies ranging from the *New York Times* to the *Allentown Morning Call*. Mostly on

Carol Tenopir is Assistant Professor at the Graduate School of Library Studies, University of Hawaii, Manoa

NEXIS, VU/TEXT, and Dow Jones News Retrieval);

- newswire services (national and international, such as AP, UPI, Reuters, and even Tass, are on NEXIS, NEWSNET, Dow Jones, and DIALOG as well as on services such as CompuServe and The Source);

- reference books (encyclopedias are available on most of these systems. In addition, standard technical reference books such as the *Merck Index*, *Kirk-Othmer*, *Mental Measurements Yearbook*, and several drug handbooks are available on DIALOG, BRS, NEXIS, STN, and Dow Jones);

- directories (the most common type of reference book online is the directory, including *Books in Print*, company directories, *Marquis Who's Who*, the *Official Airlines Guide*, "Peterson's Guides," on various systems including BRS, DIALOG, NEXIS, Dow Jones);

- government documents (including periodicals such as *Code of Federal Regulations*, *Department of State Bulletin*, and *Federal Register* on NEXIS and *Commerce Business Daily* on DIALOG); and

- statutes and court decisions (on LEXIS, WESTLAW, and JURIS).

Some full-text databases are made up of a single publication. Harvard Business Review Online on DIALOG and BRS includes several years of articles from that one journal. *The Academic American Encyclopedia* and most directories contain a single title per file. NEXIS and LEXIS always allow the user to select a single title for searching, forming an ad hoc single title file. Other databases and the grouped library option on NEXIS and LEXIS put several different publications together: McGraw-Hill Publications Online file on DIALOG has over 30 major journals published by McGraw and the Magazine ASAP database includes over 100 popular magazines.

A variety of strategies

Search strategy and uses should be expected to vary with the type and amount of information in full text. Searching within the same sentence or paragraph should be a more successful strategy with chemistry journals. Paragraph searching and display on popular-style periodicals can easily give a reader a false picture. For example, a search of "marriage contract" in the same paragraph as "Charles and Diana" might retrieve a recent article in *McCall's* that contained a paragraph that gave details of their marriage contract. Only by reading the next paragraph does the reader learn that in fact there is no contract; the "relevant" paragraph was reporting nonsense.

According to my study of Magazine ASAP, the mixture of types of journals sometimes makes good search strategy difficult. The news magazines cause many false drops even when searching for concepts within five or ten words of each other because articles about political campaigns just list many unrelated concepts discussed in a speech. These strategies are often too restrictive to retrieve more factual, lengthy articles found in magazines such as *Psychology Today* or *Science* where paragraph retrieval works better.

Uses of full text

Like bibliographic databases, full-text databases can be used to compile a list of documents on a subject. For this use searching for words in titles or on controlled vocabulary descriptors if available will yield a cost-effective, high-precision search. Searching for concepts within the same paragraph will retrieve many more documents, including some that treat the desired concepts peripherally to the main focus of the article. Searchers need to review titles and preferably KWIC portions of all documents before printing or displaying full texts.

Studies conducted by the American Chemical Society show that if the printed copy of a journal is readily available, users prefer to print out citations or relevant paragraphs and citations and get the complete articles in original form. As document locators full-text databases are not used to replace printed journals; instead they are used to allow enhanced bibliographic-style searching.

On the other hand, full-text databases as document delivery aids allow a known article to be located online and printed out or downloaded on demand. Here the emphasis is not so much on the search capabilities but on the database as a substitute for hard copy.

Just browsing

Full texts as browsing devices have not been cost-effective on systems that have connect-time pricing, but it can be valuable if price does not enter into the picture. Browsing through certain journals whenever they are updated or browsing through articles after a broad subject search will allow users to serendipitously find material of interest just as they do with printed journals.

Search strategies should be broad for browsing—the latest update of a particular journal or a simple subject search. Searching for concepts linked with the Boolean AND operator may work here because studies have shown that although precision is poor with AND searching in full text, many relevant documents are retrieved that are not retrieved by paragraph or word proximity searching. Other studies have shown that people who are searching full-text databases for their own use have a higher tolerance for irrelevant materials than they do when an intermediary does a search for them.

A final, unique use of full-text databases is to retrieve isolated facts or paragraphs in a document. For this purpose searching for concepts within a certain number of words is often the best strategy. This is the real power of full text and has some interesting implications for the way people read texts and perhaps for the way authors write.

A couple of years ago I was giving a speech on full-text databases to a group of authors and publishers. As I explained the advantages of being able to search on every word in the article and then to display only those portions of the article that contained the search terms they stopped me. "You mean that someone can read only the paragraphs or lines that contain the terms searched and then go onto the next document and read only a paragraph or a few lines in that one also?" The idea of paragraph retrieval—of reading isolated portions of text out of context—appalled them. They saw it as a threat to the integrity of writing and to an author's meaning.

Hypertext links between related portions of documents will further facilitate this new kind of reading. Full text online will become even more paragraph retrieval—not document or article retrieval. The searcher and database designer will have powers over reading that most authors have never considered.