11-1-1993

Natural Language Searching with WIN

Carol Tenopir
University of Tennessee - Knoxville

ONLINE DATABASES

BY CAROL TENOPIR

Natural Language Searching with WIN

IMAGINE LOGGING ONTO an online system and entering a search statement exactly like this: "What is the government's obligation to warn military personnel of the dangers of past exposure to radiation?" In almost every online system you would get zero postings or a string of puzzled error messages. Standard online systems force users to conform their search statements to a structured and often convoluted string of commands, Boolean operators, and proximity operators that the *system understands.*

Depending on the online system, the question above might have to look like this instead:

DIALOG: SS government(s)warn???(s) (soldier or sailor or service(w)member or serviceman or military) (s)radiation;

WESTLAW: government /p warn*** /p soldier sailor "service member" serviceman military /p radiation.

These statements use command input, as opposed to natural language, coupled with the standard Boolean search engine, which is an unforgiving, exact-match system. Only those documents that contain all four of the concepts—linked with (s) or AND on DIALOG and with /p or AND on WESTLAW—will be retrieved.

This may be fine for trained and frequent searchers, but it is difficult, at best, for novices or infrequent users. Every online system requires a different set of commands, a different way to express proximity operators, and a different syntax. This can be confusing even for experienced searchers.

Natural language/partial match

Natural-language input, coupled with partial-match search techniques such as word frequency ranking, is just beginning to compete with command input/Boolean logic searching in commercial online systems. The natural language query above can be entered as it is stated above, in WESTLAW's WIN ("WESTLAW Is Natural") system. The search process involves a complex formula that examines such things as how many of the concepts occur, how many times each word occurs, and how important the terms are in the document. WIN will display first those documents deemed to be most relevant, but those that contain some of the terms may be retained as well; this is what is meant by partial match.

WESTLAW is one of the first major commercial online systems to embrace both natural-language input and partial-match searching. (See Online Databases, October 1, p. 67–68, for a description of others.)

What is WESTLAW?

If you don't work in a law library, you may not be familiar with the WESTLAW online system. WESTLAW is the computer-assisted legal research arm of West Publishing Company, the longtime publisher of printed legal materials. WESTLAW, along with Mead's LEXIS, its major competitor, is perennially one of the most heavily used online systems. Law librarians, lawyers in law firms, and law school students are all frequent WESTLAW searchers, needing comprehensive and current full-text retrieval of statutes, case law, and other related legal materials.

West Publishing was founded in Minnesota more than 120 years ago, and its many innovative print publications have helped shape the way law is taught in this country. Notable among the company's many innovations are West's Reporters, casebooks, and the American Digest System.

WESTLAW first went online in 1975 as a bibliographic and synopsis legal research service. Full text was added soon after, with major enhancements to the system in the late 1970s and early 1980s. Today, WESTLAW includes the complete texts of federal and state case law, statutes, and regulations. It also has full texts of legal newsletters, reference books, law reviews, and bar journals.

The West Publishing editorial staff adds intellectual value to WESTLAW through its system of synopses, headnotes, and key numbers, which provide enhanced access to the full text of court decisions. Synopses are narrative summaries of each case; headnotes are brief summaries of each point of law in an opinion and the numerical citations to statutes that have been interpreted in a case; and key numbers provide a classification system that arranges point of law by main topics and subtopics.

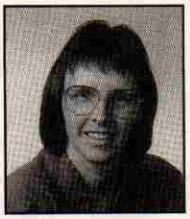
Westlaw goes natural

Like almost all other online systems, WESTLAW based its retrieval on commands and Boolean logic. To make searching easier for novices or infrequent users, a menu-driven interface (called EZ ACCESS) was added in 1990. EZ ACCESS provides help at each step of the search process, including choosing a database, but the EZ menus rely on Boolean searching. Last year, after three years of development, West made the search process even easier.

West recommends choosing the WIN interface "if you are not an experienced WESTLAW user and you know concepts related to your research issue rather than specific information, such as citation or title." The company is adamant that WIN is an additional search method that will continue to coexist with traditional methods. The traditional "terms and connectors" command system is recommended "if you are experienced with Boolean searching and want to search specific facts, or a title or a citation, rather than concepts."

Stairway to WIN

When a natural language statement is entered in WIN, what happens next? Though the process is transparent to the searcher, most online searchers understand what happens in relatively simple exact-match Boolean systems—words representing concepts are either all found in each docu-



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

ment or they are not. WIN's search engine is much more complex and sophisticated.

WIN undergoes five steps in processing a search query: In the first step, "stop phrases" are removed (i.e., common, trivial phrases such as "find articles about" or "I'm interested in"). Next, WIN identifies legal phrases by matching them to a dictionary of common legal phrases built by West editors. These phrases will be searched as intact phrases, a process the user can instigate by inputting known phrases in quotation marks. Stopwords (e.g., to, the, is, for, and so on) are removed after legal phrases are identified.

Thirdly, the system's stemming program generates common word form variations for all the remaining words. Singulars, plurals, past tense, and gerunds are all generated, as are common legal abbreviations (e.g., regs for regulation) and standard equivalencies (4 and four). In the search above, for example, obligations, obligated, obligate, and obligates would all be generated, as would warn, warning, warns, warned, warnings, etc. (Uncommon word forms, such as obligatory, are not a part of the stemming, however, and must be entered by the searcher).

Narrowing the search

In the fourth step, WIN finally begins the search process. Words, word stems, and phrases are searched throughout the database, with any document that contains any of the words or phrases retrieved for further statistical analysis. Statistical methods compare how many times the terms appear in each record with how many times they appear in the database as a whole. Each term is not treated equally. According to West, "the more often a concept appears in the database, the less weight it is given. The more often a concept appears in the document, the greater weight it is given." This helps the system predict relevance of any document to the search query. Legal phrases are given more weight than other words.

The final step involves ranking the documents according to the likelihood of relevance. The document most likely to be most relevant is displayed first, with display progressing in decreasing likelihood of relevance. Up to 20 documents are displayed; the searcher can choose to display fewer or more: up to 100 may be displayed.

Other optional features improve the WIN search and display. Users can choose to view only those portions of full-text documents that are most related to their query by using the BEST mode. An online thesaurus suggests alternative terms or synonyms to improve a search statement.

Does it work?

After a year of WIN searching, reactions are mostly positive, and many are enthusiastic. The ranking by likelihood of relevance seems usually to match the intuition of the user. Some searchers get drastically different results when searching the same topic on the same database with commands/Boolean and with WIN, with unique relevant documents found in each. A comprehensive search should probably use both approaches.

There are some cautionary notes. Some law librarians fear that novices mistakenly think WIN is doing more than it really does. WIN does not analyze concepts, for example, nor does it interpret a query for meaning. Although it goes beyond straight pattern-matching with its automatic stemming and online thesaurus, it does not automatically add synonyms.

WIN does not eliminate the need for thinking online and for formulating a good query. Even though users don't have to memorize Boolean or proximity connectors or commands, all natural-language statements will not yield equal results. The WESTLAW manual offers hints for entering a WIN query to get the best results. (The best natural-language input isn't completely natural after all!)

Query statements should be clear and concise with a minimum of extraneous words. For example, "give me all the information addressing the issue of interpreting the statute of limitations for personal injury actions" could better be stated as "What is the statute of limitations in a personal injury action?" All of the extra words that are not stopwords in the first example would be weighted and treated as concepts.

Searchers still need to think about alternative ways to describe each concept. Synonyms, antonyms, or alternative terms should be input in parentheses after each concept or located in a search of the WESTLAW thesaurus. Thesaurus choices for the concept "creditors," for example, are "mortgage holder," "secured party," "bank," "debtee," "lender," and "mortgagee." Users can choose from this list after entering the thesaurus command or putting the desired alternatives in their initial statement. For example, rather than checking the thesaurus, a search might be directly input: "Is the wife (spouse) of a debtor a necessary party in a creditor's (or lender's, mortgage holder's, mortgagee's) action?"

The precision of legal terminology may contribute to WIN's positive search results. The dictionary of legal phrases and the thesaurus, along with relevance ranking, are a powerful combination. In a full-text magazine or mixed topics, the development of a thesaurus and term dictionary are more complicated, as is relevance ranking.

Exactly what we've needed

Commands, menus, or function keys are all more familiar interfaces for online or CD-ROM database searching. Natural-language searching has been relegated to the research laboratory or to software for small, in-house databases. (One notable exception is Personal Librarian, described in Online Databases, *LJ*, October 1).

Exact-match, Boolean logic search engines are also a familiar standard. Boolean searching is called exact-match because all concepts linked with an AND operator *must* be present in a document for that document to be retrieved. Documents that have three out of four ANDed terms are just as lost as documents that contain none of the terms. Although laboratory tests and software for in-house databases have proven the value of partial-match, non-Boolean engines, most of the commercial online world has been slow to embrace such major changes.

West is to be congratulated for its innovation and for offering choices. WIN's success is causing other company's to rethink their systems, and some are beginning to follow suit. Look for many more natural-language and non-Boolean retrieval systems in the commercial online world in 1994.

For more information about WIN

Pritchard-Schoch, Teresa, "Natural Language Comes of Ages," *Online*, May 1993, p. 33–43.

Pritchard-Schoch, Teresa, "WIN—WESTLAW Goes Natural," *Online*, January 1993, p. 101–103.

Quint, Barbara, "Easy Does It," *Wilson Library Bulletin*, June 1993, p. 86–91.

West Publishing Company, 610 Opperman Dr., Eagan, MN 55123-1308; 800-688-6363.