6-1-1994

Overcoming the Black Box Syndrome

Carol Tenopir
University of Tennessee - Knoxville

BY CAROL TENOPIR

Overcoming the "Black Box" Syndrome

A MAGICIAN'S BLACK box is the place where things happen mysteriously and out of sight of the audience. When a magician waves his hand over the box and says the magic words, the rabbit (or whatever) appears or disappears. The audience is supposed to appreciate the final result, not worry about how it came about. To understand too much may even spoil the magic.

To make online or CD-ROM search software friendlier, developers often favor the "black box" approach. In the electronics world, wires go into one side of a black box, wires come out at the other side, and an unseen process in the middle makes something happen. In the early days of radar, Royal Air Force pilots referred to the slightly magical radar box as the black box.

Online like the black box

Black box advocates believe searchers should not have to worry about what's going on behind the scenes—they should just appreciate the results. A search that retrieves relevant items is a satisfying achievement, not to be spoiled by analyzing why it occurred.

Search software should be straightforward, logical, and easy to follow. Search features should take care of routine, repetitive things and facilitate the seachers' success without bogging them down with the mechanics of how it is being done.

But good online searching is not a trick, nor is it a purely mechanical process. It is a more complex process that benefits from interaction, intuition, and an understanding of how the computer is processing the information by whomever is doing the searching. Even if end users are doing most of the searching, there must be knowledgeable experts who understand what is going on inside the box to help when something doesn't work right or when a user doesn't understand what is happening.

To look inside the black box, knowledgeable searchers need to ask several important questions about each online or CD-ROM system they use. Knowing these inside secrets can help explain strange results, help improve search strategies, and help with troubleshooting.

Boolean or statistical searching?

It used to be safe to assume that all major online or CD-ROM systems relied on Boolean logic searching. In the last few years, with the introduction of statistical search engines in the commercial sector, this can no longer be assumed. Westlaw (WIN), DIALOG (Target), Mead (FreeStyle), America Online, Compton's MultiMedia Encyclopedia on CD-ROM, and others now offer statistical retrieval and relevance ranking. These search engines calculate how often each term entered in a search appears in each record, compare occurrences within each record with the database as a whole, and display first those articles with the greatest likelihood of being relevant.

Good search strategies are different for Boolean and statistical search engines. In a Boolean system, the more terms a searcher adds linked with a Boolean AND, the fewer documents will be retrieved. Boolean ORs increase retrieval, so the experienced Boolean searcher can manipulate the number of records in a search by adding or deleting concepts and terms linked with ANDs or ORs as desired.

In a statistical relevance system, the more terms a searcher adds, the more documents will be retrieved (usually up to a system-imposed maximum), because each term is searched in an "or-like" relationship. However, the more terms that are entered, the more likely the first documents displayed will be relevant, because they will most likely include more of the search terms.

Troubleshooting also differs between Boolean and statistical search engines. In a statistical system, hard-to-pinpoint false drops are likely to occur if the system doesn't match for common phrases. For example, if a searcher inputs *New York,* Compton's MultiMedia Encyclopedia and Westlaw's WIN will recognize that as a phrase. DIALOG TARGET will treat each term independently (unless the searcher explicitly enters the phrase with quotes *"New York"*) and find some articles that only have "new" and others that only have "york."

How does it treat a blank?

When you input the terms *reference services* in DIALOG's Boolean system, you will get zero hits or almost 2000, depending on whether you are searching in Library and Information Science Abstracts (LISA) or ERIC. The reason is not because LISA doesn't cover the topic, it's because of the way DIALOG treats a blank space.

There are at least five different ways that systems handle a blank: DIALOG online, as in the example above, treats a blank as a bound-phrase indicator. Only valid descriptors or identifiers are treated as bound phrases in the subject indexes for these databases on DIALOG, so a blank between words automatically defaults the search to the descriptor or identifier fields. *Reference services* is a valid ERIC descriptor; the zero postings for LISA only means LISA doesn't use the descriptor *reference services* (it uses *reference work* instead).

DIALOG Ondisc, on the other hand, defaults a blank to an adjacency operator [a (W) in DIALOGese] and will search for the entered phrase wherever it occurs in titles, abstracts, or texts, as well as in descriptor or identifier fields. This is the friendliest way to deal with a blank and is used by many Boolean systems that cater to end users (including Knowledge Index, STN, LEXIS/NEXIS, SilverPlatter, and others).

A third way systems treat a blank is to interpret it as a Boolean AND. Online public access catalog (OPAC) software, such as the CARL system, often do it this way. Entering the phrase *reference services* in a CARL database will search for the term *reference* any-

Carol Tenopir is Professor at the School of Library and Information Science, University of Tennessee at Knoxville

where in the database AND the term *services* anywhere in the same records. This works reasonably well for short records but can lead to false drops if abstracts or text are present.

Blanks can also be treated as a Boolean OR. Both Data-Star and BRS (as of mid-1994) default a blank to OR, so entering *reference services* will cause the system to search for *reference OR services*. This is useful mostly to experienced searchers who are trying to save a few keystrokes, because it can lead to many false drops in Boolean logic systems.

A fifth way to treat a blank is to have it change depending on what a searcher has entered prior to the blank. Some systems will interpret a blank the same as the last previously entered operator, for example, Data-Star and BRS will extend the adjacency operator to all blanks in the search statement *online adj database searching*.

What does it search by default?

If a search term is entered without a field designation on DIALOG or STN, the system will search only in those fields designated as subject-related. The subject-related fields (usually title, descriptors, abstract, and text) are searched together in a default "basic index." Searchers must explicitly ask for non-subject fields such as author, corporate source, and the like with the appropriate field tags (e.g., AU or CS). This means false drops are sometimes avoided (you won't retrieve articles by John Bacon when you want articles about pork), but you must check system documentation for the appropriate codes.

Menu modes of several CD-ROM systems, including InfoTrac and WILSONDISC, default to searching just descriptors or subject headings. This eliminates unwitting false drops but restricts the power of the search system.

Other systems default to searching every field in each record. NEXIS, for example, puts authors, corporate sources, and all other fields in with all of the subject fields. This simplifies the search process but may at times cause false drops.

How does it define a word?

Free-text searching is the process of searching for words as they are extracted from database records by the loading software for entry into each database's index. A good searcher anticipates what words are likely to be in a database index, partly by knowing the subject matter and partly by knowing each system's rules for defining a word when the index is created. DIALOG's rules are the easiest to understand because they are the most simpleminded.

DIALOG defines a word as any character string surrounded by blanks or punctuation marks. This means that for the sentence *My cat has brown fur*, five words are placed into the index. Since punctuation is treated as the end or beginning of words, DIALOG will also put five words in the index for the sentence *Rock-bottom prices encourage sales.*

DIALOG is consistent in its treatment of punctuation and makes no exceptions to the rules. This makes it predictable, but it sometimes leads to absurdities. *Children's literature*, for example, has three words in DIALOG database indexes: *children*, *s*, and *literature.* To free-text search it, you have to enter *children(w)s(w)literature* or it can be searched *children(1w)literature*.

Other systems, such as LEXIS/NEXIS, BRS, and Data-Star make some exceptions in their treatment of punctuation. Periods are discarded, but hyphens are retained (*rock-bottom* is just one word), and apostrophes are excised (*children's* becomes *childrens*). Friendly software will automatically search all variations with or without punctuation.

What stop words are there?

Not all words are searchable in any system. A good searcher must know each system's stop words and whether they vary from database to database. Most systems make trivial words stop words, but how many they include varies quite a bit.

DIALOG online uses only nine stop words (an, and, by, for, from, or, the, to, with); most others include dozens or even hundreds. LEXIS and Data-Star, which include databases in languages other than English, have different lists depending on the language of the database. DIALOG and STN count the stop words for searching, making *gone* two words removed from *wind* in *gone with the wind.* In most systems *gone* is considered right next to *wind.*

Good software lessens the need to memorize stop word lists. NEXIS and CARL, among many others, will discard stop words that are entered in a search, inform the user, and proceed with the search without the stop words.

What is truncated automatically?

Automatic singulars/plurals are considered essential for systems that cater to end users, especially CD-ROM systems, but the system should inform the searcher of what variations are being searched. Unless singulars/plurals can be turned off, you will retrieve both forms whether you want them or not (electronics journals as well as electronic journals, for example). NEXIS, BRS, and Data-Star all have automatic singulars/plurals (you can turn it off in the latter two).

Some systems, such as Westlaw's WIN, do automatic stemming of other common word-form variations as well. Past tense, gerunds, and endings such as -tion will be stemmed for some words. For common word variations this obviates the need for explicit truncation, but automatic stems usually do not work for irregular verb forms or unusual words.

What automatic substitutes are made?

Mead (LEXIS/NEXIS) has led the way for years in other automatic word substitute features. It substitutes common acronyms and abbreviations—fifth/5th, FBI/Federal Bureau of Investigation, IBM/International Business Machines—as well as British/American and Chinese romanization spelling variations. Other online systems, including Westlaw, Data-Star, and BRS, do some substitution. The problem with relying on automatic substitutions is that they are not always consistent. NEXIS substitutes tumor for tumour but not vice versa. "IBM" will pick up International Business Machines, but "I.B.M." will not. Informing users of what substitutes are being made would help.

How much should end users know?

Should end users be expected to know all of these details for every system they search? Of course not. The software should make many of these features as easy and transparent as possible à la the little black box. But for those of us who do troubleshooting, instruction, and point-of-use reference, a deeper understanding is essential. Someone has to know what's going on inside the black box when the rabbit fails to appear.

In June, Carol Tenopir will leave the University of Hawaii at Manoa to become a Professor at the Graduate School of Library and Information Science, University of Tennessee at Knoxville. Her new address is: Graduate School of Library and Information Science, University of Tennessee at Knoxville, 804 Volunteer Blvd., Knoxville, TN 37996-4330; 615-974-7908; Internet: tenopir@utkvs.utk.edu.