Hit or miss

Finding information on the Web requires the right search engine, the right strategy, and a bit of luck.

BY MARK S. LESNEY

Search engines are a fact of life. They have become the only effective way of dealing with the overwhelming amount of information now on the World Wide Web. But they are not necessarily intuitive devices. The number of public and free search engines is still growing, with time-honored classics jostling for position with upstart newcomers, and with older providers being absorbed into or transformed into private specialty markets in the historical shuffle. Search engines rely on one or both of two main schemes—spiders (the so-called webcrawlers, which search automatically through pages on the Web, performing indexing tasks) and the use of human-edited directories. Because these Web searchers all use different algorithms for searching and indexing, the same terms entered into a query can give different results on different engines (*1*).

Metasearch engines such as Dogpile and Copernic use their own algorithms to search through the collective gleanings of other search engines, taking what they refer to as a "best of the best" strategy. Although this can be highly effective, it can also dilute your search if you strongly prefer one search engine's algorithms, and it can result in considerable overkill if you are interested in a very common term or topic. For a summary of some of the major search engines and their websites, see "The searchers" (box, p 26).

The number and variability of search engines mean three key things for a typical user:

- you can get a broader survey of a topic by using more than one search engine;
- one search engine may match your general preferences for information better than another, depending on your topic and on your search style and needs—for example, AllTheWeb provides specialized audio and video searching capability; and
- information can be hiding in plain "site" if you do not develop the know-how, tricks, and patience to seek it out.

No free lunch

Searching and being searchable are now an industry in themselves, with organizational webpages, annual meetings, gurus and doyens and pundits—and, of course, consultants and educational software (*1*). Pay-for-use search engines designed for unique and proprietary purposes are becoming more common even as free ones still proliferate—this is especially true for rare or specialized information.

PHOTO: PHTODISC

Few end users realize the extent to which a hidden realm of finances underpins the search culture—based on the practice of paying a search engine to guarantee or prioritize visibility in its collective space of webpages examined and indexed. Often, these fees paid by companies and organizations for priority listing subsidize the "free" portions of the Web search and indexing. And, like an ad in the *Yellow Pages*, the results can benefit both "seller" and user when there is a need for a commercial matchup. For others (often most), it is the necessary if undesired consequence of having a "free" source of generalized searching in the first place. When the paid results are not desired, cutting through this commercial chaff is often best done by honing one's search and querying techniques.

Searching operands

Almost everyone knows about putting quotation marks around a phrase to ensure it being searched for exactly "as is". Most free and pay search engines have adopted several other conventional operands that help simplify and direct the standard search beyond the random jumble of keywords that most individuals initially use when they first learn about searching the Web.

According to Danny Sullivan, founder of SearchEngineWatch.com, there is a "mathematics" of search engines. For example, *+poisoning +"heavy metal"* will give you many pages related to lead and mercury toxicity, but *–poisoning+"heavy metal"* will likely give you pages on your favorite hard rock band. And if there is a mathematics, there is also a grammar, with punctuation and key "verbs" controlling the pattern of a search. An especially powerful tool is the use of the colon (:) with an appropriate keyword. For example, in AltaVista, the query *host: *.gov medicine*, where * is a wildcard, gives all of the government websites that mention medicine (*2*).

Most search engines have "advanced" search sections that help you to phrase your questions using these operands or to narrow your search to certain categories of items or site types, such as ftp, pdf, html, video, or image (*2*).

KEY TERMS: informatics, trends

© 2004 AMERICAN CHEMICAL SOCIETY

The searchers

For most, bigger is better, and that continues to be one of the first criteria considered in choosing search engines. Google (*www.google.com*) and AllTheWeb (*www.alltheweb.com*) have each indexed more than 3 billion pages and continue to battle for primacy of place, while Teoma (*www.teoma.com*) and AltaVista (*www.altavista.com*) have indexed 1.5 and 1 billion pages, respectively (*1*). Inktomi (*www.inktomi.com*)—a Yahoo (*www.yahoo.com*) subsidiary—has an index of over 3 billion pages, but it is not directly accessible, acting instead as a major Web search partner behind the scenes for user portals such as Hotbot (*www.hotbot.com*).

Sometimes, look-and-feel issues may be important to a particular user—some people, for example, find the interactive "personal" feel of a portal such as Ask Jeeves (*www.askjeeves.com*) preferable to more naked search engines that might give nearly identical results. Others prefer metasearch engines such as the free Dogpile (*www.dogpile.com*) and Look.com (*www.look.com*) and the paid Copernic (*www.copernic.com*).

One search engine that all those interested in biomedical topics should take advantage of is Entrez, the life science search engine sponsored by the National Library of Medicine (*www.ncbi.nlm.nih.gov/gquery/gquery.fcgi*).

One trick query

More is definitely not better in Web searches if that more involves thousands of websites that mention your search terms in an inappropriate, trivial, or downright mistaken fashion. Often, it is the phrasing of your query that is most easily modified to give the best results. Where possible, specify to the *n*th degree. And vary your query in both its phrasing and its vocabulary if you do not get the answer you desire.

Suppose you want to find out when the first known case of diabetes was diagnosed. Most people know enough not to search on *"diabetes"* (which gives 6,840,000 hits in Google); but even *"first case of diabetes"* (which gives a mere 14 results) does not give the first case of diabetes ever but rather the first cases in modern contexts. So, you must change your thinking to *"history of diabetes"*, and this query gives a large number of sites, one happily relating that the earliest known record of diabetes was in 1552 B.C. But here too, there are dangers, for a huge number of sites refer to modern patients, with "history of diabetes" as part of their case histories.

The take-home message is that there are no guarantees. "Try everything you can think of" is the mantra of the frustrated query-ist. And then try it again using another search engine if there are no acceptable results. Note that all of the queries above are typed within quotation marks to control the query to "as given". Eliminating the quotation marks if the initial query fails can sometimes give success, though all too often, it dilutes the query to the point of homeopathy.

Another common search technique, used to look for a specific item or piece of information, is to go into overkill and work back. For example, to see if a reference is available on the Web in citation or full text, give the full title of a paper (or as much as you know), and/or list all of the last names of authors known, the year published, and/or the journal (if known) in one great list. If no match returns, use another search engine and/or whittle down the list until a useful "hit" shows up or you can assure yourself that one is not available. Remember, the entire Web is not on any one search engine. And differences in the indexing algorithms complicate things even more.

Beyond the engines . . .

Finding specialty information on the Web can often be challenging. But in many cases, dedicated sites obviate the need for search engines, at least in narrowing the search for information in specialized fields—although you may need a search engine to find them in the first place! These are often most useful for the now nearly ubiquitous "links" page that can send you to related sites deemed the most useful or interesting by the Webmaster or her employers (see box, "Links in the chain: Biomedical Web resources").

Ultimately, the Web is what one makes of it. Search engines and linked sites are the Dewey decimal system of the modern library that the Internet is becoming. Because more and more researchers are becoming their own virtual librarians, it is ever more necessary for them to understand and make best use of these critical tools of the trade.

References

(1) http://searchenginewatch.com.
(2) www.searchenginewatch.com/webmasters/article.php/2156031. ■

Links in the chain: Biomedical Web resources

One of the most useful tools on the Internet is the dedicated subject pages where the "best of the best" links on a particular topic have been assembled by a human Webmaster. These links and information pages can be found on government, organization, corporate, university, and even individual sites. To find links pages on any topic, one of the easiest methods is to type "<topic> links" in any search engine, and generally more links pages than you could possibly want will appear.

The following are a few sources for medical and drug topics:

- ClinicalTrials.gov: *http://clinicaltrials.gov/ct/gui.*
- Medical history: *http://inventors.about.com/library/inventors/blmedical.htm.*
- Medilexicon: *www.pharma-lexicon.com.*
- Medline: *http://medlineplus.gov.*
- NIH: *www.nih.gov.*
- Protein science: *www.proteinscience.com.*
- The Biotechnology Information Directory: *www.cato.com/biotech.*
- The National Library of Medicine: *www.nlm.nih.gov.*
- The Protein Data Bank: *www.rcsb.org/pdb.*