Evaluating WWW Search Engines for Chemistry

Harry E. Pence, Richard Bachelder, Michael Branciforti, Susan Donadio, Brian John, Joo M Jung, Matthew Glidden, Melanie Krom, Kelly Modoo, and Todd Morris,
SUNY Oneonta, Oneonta, NY
pencehe@oneonta.edu

INTRODUCTION

In a relatively short time, the World Wide Web has become a widely used source of chemical information for both college students and faculty. At its best, the Web is a rapid and convenient method to search for information; at its worst, it can be frustrating and time wasting. Frequently, the difference between success and failure at using the Web depends on the search engine chosen. Thus, search engine selection can be a critical decision for chemists who use the WWW.

For many Web users, including some chemists, the choice of a search engine is based more on Web location or external advertising than any other criteria. Engines that are well placed on popular Web portals or are widely advertised in the media seem to do much better, regardless of how useful they are. This selection process may be adequate for casual surfers, but chemists need to be more selective if they expect to find the specialized information they need.

Many popular computer journals review search engines, but these evaluations are intended for the general user. A good summary of such reviews is available on the Web.(1) One effort to focus specifically on chemistry is "Best Search Engines for Finding Scientific Information on the Web." (2), which was developed by Alexander Lebedev of Moscow University. On August 3, 1996 and on February 10, 1997 Lebedev compared the number of hits recorded by eleven different search engines for eight different keywords important in physics and/or chemistry. He discovered that the number of hits could differ by several orders of magnitude from one search engine to another. Unfortunately this work has not been updated since May 17, 1997.

The rate of change on the Web is so rapid that Lebedev's results need to be reevaluated. Even in two years there have been important changes, including the elimination or modification of some search engines in the original study. In May of 1999, the senior chemistry seminar class at SUNY Oneonta set out to update Lebedev's results.

CRITERIA FOR SELECTING A SEARCH ENGINE

There are at least three important criteria that should be used to evaluate search engines, comprehensiveness, currency, and efficiency. Comprehensiveness is a measure of what fraction of the total web sites the search engine actually reviews. This is particularly important for chemists. An article by Lawrence and Giles (3) reported that not all Web sites can be accessed by search engines and even the best of the search engines misses over a third of these accessible sites. Since engines are more likely to identify popular sites, that is, those with many links to other pages, this partially explains why chemists cannot find the specialized pages they need.

Currency measures how often the search engine revisits its sites to determine whether or not there have been any changes. Not only are new web sites constantly being created, but also many sites are vanishing. Failure to keep up to date can produce useless links that no longer exist. In September, 1998, a further study (4) by Lawrence and Giles concluded that the Web is growing faster than the increase in the search engine coverage, and engines are returning a greater percentage of dead links. The situation is getting worse, not better.

The final concern is efficiency. Are the most useful sites not just included but listed early in the search results? This is probably the most difficult to evaluate quantitatively.

Lebedev argues that the number of documents is most important when looking for scientific information and so focused mainly on comprehensiveness. He argues that the number of scientific publications is only 10-20% of the total number of documents found by search engines, and so listing more documents increases the probability that nothing useful will be missed. This approach ignores two other important criteria mentioned above, currency and efficiency, but it does provide a helpful perspective for chemists.

Lebedev chose a short list of scientific terms and recorded the number of hits for each term on each search engine. He found that the number of hits changed by several orders of magnitude from one search engine to another. Based on his results, he recommended AltaVista as the most comprehensive search engine.

SEARCH ENGINE RESULTS

During May of this year, students in the senior seminar at SUNY Oneonta repeated the survey that had previously been done by Lebedev, with several changes. The list of Search engines used was modified by eliminating those that gave very few hits with scientific search terms, as well as those that had changed format in such a way that they no longer could be compared. Yahoo, which is highly rated for general use, consistently returns very small numbers of hits for these scientific terms, and so was eliminated. Two search engines, Northern Light and Microsoft Network, were added, since these are reputed to give good results. A slightly shorter list of search terms was used. The results are shown in Table I below. The 1996 and 1997 results are from Lebedev's study and the 1999 results are the current project.

DISCUSSION

Several of the trends reported by Lebedev are continued with the most recent data. During the period from 1996 to 1997, two search engines, Inktomi and NlightN, terminated or became inaccessible. Since then, Magellen has changed to focus mainly on forming chat groups and Lycos no longer displlays the number of hits. The number of hits recorded with Excite decreased in each case from 1996 to 1997, and these values were, in turn, even less in 1999. AltaVista usually returned the greatest number of hits. Although Lebedev reported

TABLE I

	AV	HB	EX	IS	NL	MN
crystallography						
96	31186		24975	1464		
97	28597	25232	15360	11513		
99 1	20365	28530	12260	21123	42108	18794
catalysis						
96	27431		18061	550		
97	21841	18521	12308	9471		
99	73020	23340	7902	17163	37397	13988
benzene						
96	27533		17304	374		
97	24764	19879	12372	8875		
99	51488	27745	9351	17538	51217	15315
luminescence						
96	9731		7231	206		
97	7597	8103	5733	4104		
99	11228	10490	4144	8397	21604	6539
ferroelectric						
96	8354		4362	166		
97	5622	4579	2983	2439		
99	7091	4810	1846	3820	14402	3652
EXAFS						
96	3144		2167	64		
97	2677	2225	1639	1005		
99	3501	2450	963	1446	4595	1713

AV=AltaVista, HB=HotBot, EX=Excite, IS=Infoseek,
NL=Northern Light, and MN=Microsoft Network.

that the number from AltaVista declined from 1996 to 1997, and the latest data generally shows these values have increased, often quite substantially. Northern Light, which was not among the engines in Lebedev's study, competes best with AltaVista, and in some cases even provides more hits.

There are several alternative sources of evaluations that tend to confirm these results. One of the most useful sources of information about search engines is the Search Engine Watch site edited by Danny Sullivan. This site compares search engines in several ways, including the size of each search engine's index (5). The most recent results from that site (May 1, 1999) indicate that AltaVista has the largest index, followed by Northern Light, then Inktomi (used by several engines, including HotBot and MSN). A larger index indicates a greater chance of finding unusual information, which would presumably include chemical terms.

The article by Lawrence and Giles (3) reports that the most comprehensive engines are HotBot (which is powered by Inktomi), AltaVista, and Northern Light (in that order). In their later report (4), they note that in comparison to their previous study, Northern Light has significantly increased its coverage relative to the other engines, and the difference between the largest and smallest coverage of the engines is not as great. All of these results are in agreement with the results obtained in this paper. Finally, it should be noted that Lawrence and Giles found that Northern Light, Microsoft Network, and Lycos returned about twice as high a percentage of invalid links as the other engines.

CONCLUSIONS

Perhaps the most important conclusion from the available data is that no single search engine covers the entire WWW, and so a really through search would require the use of more than one engine. Even though a number of engines now have roughly equivalent indexes, AltaVista still seems to be slightly better for use by scientists, with Northern Light giving results that are almost as good and sometimes may be a little better. It is possible to search multiple WWW engines by using a metasearch engine, like Dogpile (6), but these are usually limited in the number of hits that are returned. Instead of crawling the web to build an index, metacrawlers send search terms to several search engines, then combine the results. Finally, it is clear that the WWW is still in a state of rapid development, and even these conclusions must be considered to be tentative, until the next new development. A more extensive repokrt of this project is available on the WWW (7).

REFERENCES

1. http://www.searchenginewatch.com/reports/reviewchart.html
2. http://www.chem.msu.su/eng/comparison.html
3. Science, 280, April 3, 1998, pgs. 98-100
4. http://www.neci.nj.nec.com/homepages/lawrence/websize98.html
5. http://www.searchenginewatch.com/reports/sizes.html
6. www.dogpile.com/
7. www.oneonta.edu/~pencehe/engineselect.html