Evaluating Search Engines for Chemistry - 2005

Part of The Alchemist's Lair Web Site
Maintained by Harry E. Pence, Professor of Chemistry, SUNY Oneonta, for the use of his students. Any opinions are totally coincidental and have no official endorsement, including the people who sign my pay checks. Comments and suggestions are welcome (pencehe@oneonta.edu).

Last Revised October 8, 2005

Evaluating WWW Search Engines for Chemistry 2005, Harry E. Pence, SUNY Oneonta, Oneonta, NY

INTRODUCTION

It seems that the world of search engines has begun to simplify to three main competitors, each supported by a company with deep pockets. Google continues to maintain a very strong position; Yahoo, having purchased several of the major engines, used the human talent that it acquired to develop a new engine as the basis for its searches; and Microsoft, has rolled out its own search engine. (Previously, Yahoo! had used a combination of Google results and results from a human generated directory. Chris Sherman suggests that combining so much search expertise with Yahoo!'s extensive e-mail experience may make the new engine more resistant to spam.) Some chemists may still choose to use one of the engines that serve a specific niche or a meta engine, but the number of individuals in this category seems to be decreasing. The main exception would be those at institutions that subscribe to a number of Elsevier science journals, for whom Scirus may be an attractive choice for several reasons, not the least of which is the fact that it is specifically science oriented and includes on-line access to resources like MEDLINE, Beilstein, BioMed Central and several preprint servers. The focus in this article is only on the three main engines. As noted below, developments in the engine industry may make a science-focused engine even more attractive in the near future, since the main engines are increasingly focused on serving general interest clientele. This may make it harder for chemists to find the specialized pages that are of interest to them.

As noted in a previous article in this series, there are at least three important criteria that should be used to evaluate search engines: comprehensiveness, currency, and relevance. *Comprehensiveness*, the measure of what fraction of the total web the search engine index actually includes, is particularly important for chemists, because they are often looking for unusual information that may not be included in smaller search engine indices. *Currency* measures how often the search engine revisits sites to determine whether or not there have been any changes. This is important to all web searches, since failure to revisit sites allows dead links to be included in the index. The final important criterion is *relevancy.* Are the most useful sites listed early in the search results? This is probably the most difficult criterion to evaluate quantitatively, since even among chemists what is important to an inorganic chemist may be of little interest to an organic chemist and visa versa. In fact, the needs of a single searcher may vary from day to day. Danny Sullivan has an excellent (and highly relevant) summary of both the issues and problems involved in a web page titled, "In Search Of The Relevancy Figure."

The basic measure of comprehensiveness is the size of the index, and all three of the major engines have

been claiming large increases in index size. In November, MSN announced that its new engine had an index of 5 billion pages. Google immediately bumped its claimed index size to 8 billion pages. Then Yahoo claimed an index that provided access to over 20 billion items, including 19.2 billion web sites, 1.9 billion images, and more than 50 million audio and video files. Google responded that their "scientists are not seeing the increase claimed in the Yahoo! index." Matthew Cheney and Mike Perry, former students of Professor Vernon Burton at the University of Illinois, reported a study of the two engines to attempt to clarify the controversy. They ran a random sample of 10,012 queries and concluded that for a given search, Google would produce 65% more web sites than using Yahoo! search. In fact, Google produced more results in 83.7% of the cases. Based on this research, they concluded that the Yahoo! claim was "suspicious." To further confuse the issue, Jean V�ronis has questioned the methodology of the Cheney-Perry study. On another level, one must ask, "Do all these claims and counter claims actually mean anything?" The web continues to grow rapidly in size and various estimates place the total size of the web at from 12 to 45 billion publicly available pages, plus as many as 200 billion more pages that are not readily accessible to engine spiders. Danny Sullivan probably speaks for many in his column entitled, "Screw Size! I dare Google and Yahoo! to report on relevance."

Relevance is still the major concern, does an engine provide the sites you want early in the list of returned sites? The Rusty Engine search engine reports a survey of search engine user opinions on relevance that puts Yahoo slightly in the lead, followed closely by Google, with MSN Search being outscored by Ask Jeeves. The interesting observation about these results is that many those taking the survey tended to rank all the engines either very relevant or not very relevant, with relatively fewer votes for the intermediate evaluations that one might expect to be most popular. Apparently it is a case of love or hate, with few emotions in between. Yahoo received significantly more votes of 5 (very relevant) than Google , but Google made up some of the difference in the votes of 4. It is hard to tell what these figures mean for chemists, since the group surveyed are more likely to be looking for information about Britney Spears than about homogeneous catalysis. The results do suggest, however, that neither Yahoo and Google have thus far established a commanding lead in the relevance for general searches. [Added Note: Google has announced that it will no longer list how many pages it has in its index. This may signal a basic change or it may just be that Google is trying to say that it is above such mundane arguments.]

There have been several recent efforts to shine some light on the issue of relevance. The first of these was a new search engine, called jux2, which compared of the results when the three major engine, Google, Yahoo!, and MSN, searched the same term. This was given high marks by many reviewers, but these favorable reviews attracted so much traffic that the available resources were overwhelmed, and this site was soon removed by the owners. A similar site, called Thumbshots.com ranking also allowed such a comparison, but this reviewer found attempts to search were very slow, probably indicating that it suffers from the same lack of resources that caused jux2 to go of business.

A site of this type that is available and is equally interesting is missingpieces, available through Dogpile. Missing pieces compares the top-rated results from three engines, Google, Yahoo!, and Ask Jeeves (MSN Search is expected to be added soon). Click on the "missing pieces" text (on the lower right corner) on the missingpieces URL produces a search box on the left and a series of concentric circles, which represent the results for the three engines. The outer rink represents results found in the top returns of only one engine (indicated by a yellow dot); the inner ring represents results from two engines (blue dots); and the center circle is the results returned by all three engines (green dots). Rolling the mouse over each dot gives the corresponding URL. A paw print on the dot indicates that the Dogpile algorithm has selected this as one of the most relevant sites. [Note: In the short time between my writing this article and it appearing in the Newsletter, the missing pieces site seems to have gone missing. Hopefully, it will return soon, but in the interim, you may wish to try DoubleTrust, or GrabAll. Each of these does a similar job, but only allows for the comparison of two engines. I like missing pieces better, but I guess you have to be satisfied with what you have now, not what you had yesterday.]

The agreement did not seem much different for popular terms compared with scientific terms. For example, a search on Madonna gave six sites that were listed on two or three of the engines; a search on Jessica Simpson returned nine dual or triple listings; and Sean Penn gave five. On the other hand, superconductor gave five; dendrimer gave two; and paramagnetism gave six. An informal examination of the various returns for several search terms seemed to indicate to this reviewer that the Google hits were more

relevant. If you are like most searchers and only look at the first few page or two of returns, it may come as a shock to realize how much can be missed by using only one engine. Of course, this is the message that Dogpile would like to send; using a meta-engine will allow you to pick up sites that you might miss if you just used one of the major engines. [Important Note: Some Dogpile publicity is quoting a statistic that in general only 3% of the results are shared by all three engines. This is deceptive, because it includes both paid and algorithmic results. It is fair game to compare algorithmic results, but paid results are not usually important for chemical searches and they are expected to differ significantly from one engine to another, depending on who pays whom.

WHAT IS GOING ON WITH SEARCH ENGINES?

Even if the choice of search engines can be narrowed down to only three, there is still clearly something fundamental changing. Yahoo! and Microsoft have both created new search engines, after years of depending on external services; Google recently made a public offering of its stock, which currently is selling for around $300 per share; Microsoft and Google are in court, with the former claiming that the latter has illegally hired a web engineer; and the intensity of agreements over index size seems to be escalating (as described above). Perhaps most surprising is the report that Microsoft is in discussions to sell its web portal (*New York Times*, page C1, 9/16/05). This last observation represents a major shift. For several years, the major web companies have focused heavily on developing a strong portal that would attract large numbers of users. If this latter piece of news indicates that Microsoft plans to focus most of its attention on search rather than attracting the public to a web portal, it is a major strategic change for a company that is not always first to catch new trends, but rarely has made major mistakes. What is happening?

The answer may lie not with the engines themselves but with broader changes that are happening in the media. Despite the best (or worst) efforts of the major media companies, the business plan that has worked well for so many years seems to be running into trouble. The basic question is, "Who will control access to information and entertainment?" The control of content is slipping away from media conglomerates and shifting to individual viewers. Napster may have been tamed, but other peer-to-peer sharing networks are more than taking up the slack. Many users are ignoring the threat of law suits in order to make their own selection of what music they wish to hear. And it is not just music. Podcasting, blogging, exchange of cell phone photos, and many other activities are already empowering individuals to be not just a consumer but also a producer of content. As network capacity continues to increase, sharing of video and movie files is also becoming more common and it is safe to say that new capabilities are one the way. It may be a portent of things to come that this July the Live 8 concerts were broadcast on VH1, MTV, and ABC, but far more people watched online on AOL. Some are even talking about the equivalent of a million channel network. There is surely some hyperbole in this claim, but it does appear clear that entertainment choices are multiplying rapidly and the Internet is a major component in this increase. Video clips are already a widely traded commodity on the Internet, and options like BitTorrent are making this increasingly easy. Journalist A.J. Liebling is credited with the quote that, "Freedom of the press is limited to those who own one." The problem is that the more popular your creation becomes, the more it costs for distribution (as jux2 and thumbshots founds out! see above.) With BitTorrent, each new user also becomes a new supplier; the old problem of providing resources for expansion is eliminated.

The fly in the ointment for this expanding universe is the problem of finding what you desire among the ever-increasing options. Many individuals have trouble finding what want in the current 50-100 channel network. (Of course, it does not help when major TV networks shift program schedules at the last minute to counter popular programs on other networks!) Even TV Guide is having trouble keeping up with the small slice of the bandwidth that it chooses to cover. How will we navigate the network of the future? The smart money is betting that the "million channel network" will be organized by search engines. An article entitled "The Super Network" in the September, 2005 issue of Wired magazine (which is not always right, but is always thought provoking) suggests that new and improved search engines will be the universal point of access. This is by no means going to be easy. Cataloguing text is much easier than cataloging images or videos, but the progress made thus far gives grounds for optimism. To some this is welcome news that Vannevar Bush's original idea for an individualized information source is coming closer to reality. The major providers of movies, music, TV, and other entertainment greet this possibility with consternation, since it is a threat to the business plan that has provided their revenue stream for decades

Both Google and Yahoo! are trying to position themselves to be this universal access provider, but they are following significantly different strategies. John Battelle provides some interesting comments on the different cultures that prevail at these two search engines, and how this is reflected in the search experience they provide. In essence, Yahoo! is trying to become an entertainment provider, hiring executives with TV experience who will make Yahoo! the prime destination for both entertainment and search. Google is pursuing a more technology-based strategy. Meanwhile, Microsoft is looming in the background, not usually mentioned in the competition, but never to be counted out. As this competition drives the main players to rapidly expand, the stakes required to buy into the game will continue to increase and it will become ever more expensive for a brand new engine to become competitive. It should be very interesting to watch this game develop in the next few years.

CONCLUSION

If the stakes are high for the engines, their investors, and the entertainment industry, the outcome is also important to the humble chemist who only wishes for efficient web searches. As the amount of material in the engine index increases, the question of relevance becomes of paramount importance. Will we reach a point where a chemist must sort through so much extraneous material that all patience will be exhausted before one gets to what is relevant? Or will personalized search finally become a reality, so that the engine keeps track of each user's preferences so well that an inorganic chemist will obtain different results than an organic chemist? Like any good soap opera, the search engine saga continues to move from crisis to crisis, continually balancing on the edge of disaster. For the time being, Google appears to be the engine of choice for chemists, but stay tuned for the next thrilling episode, where everything may change. We can only hope that it changes for the better.
