Using Metasearch Engines for Chemistry-II

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 Sept. 30, 2004

YOU ARE HERE> **Alchemist's Lair** > Web Tutorial > Meta Search Engines for the WWW.

(This article appeared in the Fall 2004 issue of the Computers in Chemical Education Newsletter.)

Meta Search Engines for the WWW, Harry E. Pence, SUNY Oneonta, Oneonta, NY, pencehe@oneonta.edu

---

Introduction

The mergers and realignments of the past year are still churning the world of search engines. As if this were not enough, Amazon has recently announced that it is developing its own search engine that will make it much easier for an individual to manage his or her information resources. Until this situation becomes more stable, any attempt to compare the different engines is as fruitless as trying to nail plain jello on the wall. On the other hand, there have been some interesting developments in the world of metasearch engines, and it has been two years since this topic was discussed in these reviews (see Using Metasearch Engines for Chemistry-I). As noted in this previous article, the basic idea of a metasearch engine seems very attractive. If it is true that few search engines cover the entire searchable web (not to speak of the portions of the WWW that cannot be accessed by spiders), it seems like a very good strategy to combine the results from several different engines to obtain greater coverage and create a correspondingly greater chance of finding the best reference to the topic being searched. As is usually the case, there are several caveats. A metaengine does not search its own database of web pages, but instead simultaneously submits the search word or phrase to several different traditional search engines, compiling the results from the various engines, and presenting the results to the user. If the metaengine does a poor job of selecting which results to present or if it does not include results from the most comprehensive traditional engines, the search may be not just less comprehensive but also more deceptive than a single engine. Many individuals think that a metaengine automatically gives better results than a single engine; this may not be the case. On the other hand, some metaengines have developed very desirable ways to map of the search results.

As noted in Using Metasearch Engines for Chemistry-I, failure to include the most comprehensive engines among those being searched by the metaengine is considered to be a serious shortcoming. Thus, the engines surveyed in this presentation are divided into two groups; the first group includes those that include either Google and FAST (currently the most comprehensive search engines), and the second group which does not. In some cases the engines in the second group have some very attractive features, and the engines in this group may be helpful to searchers who do not require a truly comprehensive search.

Metasearch Engines that appear to use Google or FAST

Probably the best known of the metasearch engines that employ Goggle and/or FAST are those owned by Infospace. Many longtime searchers are probably already familiar with one or more of the engines, including Dogpile, MetaCrawler, Excite, or WebCrawler. Three of these are widely recognized to be metaengines, but some users may not realize that when Excite went bankrupt, it was purchased and redesigned for metasearch by InfoSpace. Chris Sherman reports that InfoSpace has agreed to buy search results from FAST, Google, and Inktomi as well as a number of other engines. . Although an InfoSpace Vice President suggests that Dogpile is their "destination met property", the first impression is that there seems to be little to choose among these four except in the home page features. Excite maintains the cluttered look that will be familiar to those who have used this engine in the past; the others present the relatively uncluttered query page like that which has been popularized by Google. Dogpile seems to present a more paid search results, which is probably a minus for most chemists. InfoSpace has bought the rights to Vivisimo's method for clustering and is using it on all four engines. To the left of each set of results is a column labeled "refine your search" with a topical list related to the original query. Clicking on any of these terms produces a list of the hits that would fall in this category. As will be noted below, this is one of the main assets for Vivisimo, and it is also an excellent feature of the InfoSpace engines. The clustering, however, does appear to be less complete than when one goes directly to Vivisimo. It is also possible to click on a box that will group the results by engine, providing an opportunity to compare how different engines rank sites in terms of relevance. An advanced search capacity also allows for Boolean searching as well as searching using key words or phrases All in all, this is an impressive set of meta engines, and anyone who plans to use meta search should consider these options.

[Special added note: Vivisimo has just announced the availability of a new search engine, Clusty, which is intended to compete with Google. Vivisimo says it does not plant to be as comprehensive as Google but is counting on the Vivisimo clustering technology to attract users. Searches will be based on Overture, a search engine owned by Yahoo. Past experience with Overture indicates that doesn't provide nearly as much coverage as Google, but the usefulness of this new engine will be explored in the next report on search engines.]

Another metasearch engine that covers Google results is Mamma, which also includes engines like About, Ah-ha, Business.com, LookSmart, EntireWeb, FindWhat, Kanoodle, LookSmart, MSN, Open Directory, Teoma, Gigablast, and EntireWeb. Because the listing of search results includes the number of engines that give each site a high rating, it is possible to form a good estimate of overall relevance. Mamma allows you to search for images, yellow pages, and white pages. A searcher may specify Boolean terms or search for the exact match of a phrase. It is hard to estimate how important the Boolean search capability might be, since most web surveys suggest that the number of searches performed using true Booleans is small and getting smaller. It is a matter of individual practice to decide whether the Boolean capability is significant.

One of the most unusual of the meta search engines is Ithaki, which claims to search over 500 web search engines, directories, and guides. It has several features that are interesting. For example, Ithaki claims that it searches all the engines available worldwide and chooses those which are most complete and independent. You may search "**web pages**", i.e. Wisenut, Google, Altavista, Alltheweb, Teoma, Gigablast, Hotbot, Voila, Entireweb and Cybersearch, or you may search "**web sites**", which covers web directories, such as Yahoo web sites, Open Directory , Looksmart, Galaxy, Zeal, and AskJeeves. The distinction between sites and pages is an interesting one. It appears that the latter search strategy gives more specific information, while the former search gives sites with a more general overview. Ithaki not only orders the results by relevance(as do many other engines) but also removes any duplicates. You can group the results by search engine or ask for a single, comprehensive list. Ithaki claims that it matches the best search resources on the WWW to each individual search topic to produce the most efficient search results. Searches may be either Boolean or natural language queries. Ithaki is not only available in French, Spanish, German, Portuguese, Italian, Japanese, Russian, Norwegian, Dutch, Polish, Swedish, Chinese, Greek & English, but it claims to search independent directories and search engines in these specific country rather than depending on using the international version of Google. This breadth of coverage is especially impressive, since Ithaki was created and is operated by an individual, a molecular biologist named Carla Tironi Farinati.

Metasearch Engines that do NOT appear to use Google or FAST

The table below compares several of the metasearch engines, listed in alphabetical order, that either do not include Google or FAST or else did not seem to actually use these engines. Otherwise, the list of search engines used is based on the claims of the metaengine. In some cases the metaengine claimed to search a given engine, but no hits were returned, even though that term should have produced hits on the engine in question. It may be that the failure to include results from an engine has some simple explanation; however, it seems unlikely that the omitted returns should always be from the same, major search engine. There are, of course, a number of metasearch engines that are not included in this report, but most of the popular engines are covered.

Meta-engine	Search engines used	Comments
Infonetware	This engine does represent another interesting example of clustering technology. This site submits your query to a traditional Internet search engine (not named) and then lists subtopics of this main listing in a column on the left. Clicking on any of these allows one to "drill down" to obtain more specific results.	According to a white paper that accompanies the engine, it is possible to use this technology with AOL, AltaVista, Google, FAST, and Inktomi. Unfortunately, I was unable to determine how to control this selection. Despite this, the clustering technology is a little different and it is worth considering for this reason.
Ixquick	The engines that may be selected are AOL, Ask/Teoma, EntireWeb, Find What, Gigablast, Go, LookSmart, Lycos, MSN, Netscape, Open Directory, Overture, Sprinks, Yahoo, and Wisenut. Ixquick returns a single list, with the quality of the sites indicated by awarding one star for each search engine that placed it in its top ten. The listing also includes the individual ranking from the major search engines. Paid listings are clearly identified, albeit at the top of the list of sites returned.	Ixquick allows you to search for Adobe PDF files, MP3 files, images, and news. Ixquick also translates searches that include wild cards and Booleans to match the different engines.
Profusion	The engines used are AltaVista, MSN, About, AOL, LookSmart, Lycos, Netscape, Raging Search, Teoma, WiseNut, AllTheWeb, Metacrawler, and About, but the default choices are MSN, AlltheWeb, and AltaVista. Using the advanced search feature, the selection can be customized to search the best 3, fastest 3, or all the available engines.	You may select three engines based on speed, accuracy, or personal choice. Profusion returns a single list, ranked by relevance. It is said to modify Boolean searches to work on different engines. (Boolean terms must be capitalized!)
Surfwax	The search engines it lists as being used are Yahoo, LookSmart, AOL, MSN, and Wisenut.	It does not appear to use FAST or Google, but it does return a consolidated list, with the URLs listed in order of relevance and also has several useful features, the best of which is called site snap. This allows you to click on the magnifying glass next to a site to obtain a summary of the site. This summary includes a list of "focus words", which can be used as the basis for a more focused search.
		This search engine will automatically cluster the search

Vivisimo	MSN, OpenDirectory, Overture, Looksmart, AskJeeves, Lycos, and Wisenut. This engine also provides a number of useful options, including FirstGov, PubMed@NIH and several news feeds.	results into folders. Yes! Even in chemistry it knows enough to organize enediynes into research, vita, chemistry departments, etc. A really great feature! It does not seem to include results from either Google or FAST.

Comparing metasearch Engines

The earlier report on metasearch engines suggested that the best approach would be to choose one comprehensive search engine, like Google or FAST, and learn how to use it very well. Most of these metaengines seem to perform much better than they did in the previous report, which certainly makes metasearch more attractive than suggested previously. If comprehensiveness is an important criteria, the preferred engines would be those that now include Google, that is, the four InfoSpace engines, Dogpile, MetaCrawler, Excite, or WebCrawler, plus Mamma. Among this group, my personal preference was MetaCrawler even though it is not the first choice for serious searchers according to the owner, Infospace. To complicate the issue, two real strengths of metaengines are the ability to cluster results and to include rankings from different search engines. Clustering is not simply a cute feature; going to a folder that seems to contain useful sites can be a quick substitute for spending a lot of time designing the best possible search phrase. In addition,the number of regular engines that give a high ranking to a site is a good way to evaluate relevance. Clustering seems to work much better on Vivisimo than on the InfoSpace engines. It probably should not be surprising that the technology that Vivisimo makes available to InfoSpace is not as good as the version used on its own engine. Choosing among the metasearch engines depends upon whether an individual desires comprehensiveness (in which case I would recommend MetaCrawler) or is more interested in exploring clustering, in which case Vivisimo would be my choice.

Ithaki must be placed in a special class all by itself. The claim that Ithaki selects the web resources that are best suited to the specific query is difficult to evaluate, but at the same time, potentially very powerful. Time (or lack of creativity) did not allow me to come up with an adequate method for testing this feature, and so it will be left to be explored later. The comparison of metasearch engines developed by Ithaki seems to suggest that their engine is clearly superior (are you surprised?), but despite my well-earned skepticism I must admit that the data they quote seems to support this assertion, especially since the data is supposedly based on one of my favorite search gurus, Danny Sullivan. The table indicates that Dogpile and Mamma depend much more on paid engines than does MetaCrawler, which corresponds to my comment above that Metacrawler seems to have fewer paid ads in the results. This listing also suggests that MetaCrawler actually uses a higher percentage of the engines that it claims to use that most of the other metasearch engines. Again, this agrees with my experience. Contrary to this data, Chris Sherman reports that Dogpile won the Best MetaSearch Engine Award from Search Engine Watch in 2003 and Vivisimo was second.

In summary, the two metasearch engines that I preferred were Vivisimo and MetaCrawler, even though I am not completely comfortable with these choices. In particular, I hesitate to disagree with Chris Sherman and the experts at Search Engine Watch, who obviously liked DogPile much more than I did. The wild card in the list is Ithaki. It is certainly an engine that I will be investigating in the future. Perhaps I'll have an inspiration about how to evaluate what it claims to do.

Addendum

Frequently I run across a resource that doesn't really fit into any of the topics that I discuss, but which does appear to be generally interesting (at least to me). I can resist anything but temptation, so I usually try find some way to include these sites. So far Brian Pankuch hasn't complained so I will try to be circumspect but will still include some extra tidbits. In this case the resource is ISIHighly Cited. This site recognizes researchers whose collected publications have received the highest number of citations during the past twenty years based on the Web of Science® maintained by the Thompson's Institute for Scientific Information. The names are organized into twenty-one categories, including chemistry, physics, ecology/environment, and several subdisciplines of the biological sciences. Thompson says that it updates the results each year. If you select the browse button on the home page, it is possible to review the names

by scientific category, institution, name, or country. For articles with multiple authors, each citation to that article is counted for each of the authors. Thus, an article with 20 authors may generate twenty different listings if it is cited once. The listing for each scientist includes basic biographical information and a list of recent publications.

Return to The Alchemist's Lair Web Site Home Page

Return to Web Tutorial Home Page at the Alchemist's Lair.

Return to the Fall 2004 issue of the Computers in Chemical Education Newsletter.)

You are the visitor to the Alchemist's Lair site since Jan. 10,1997.
