Web alert

Searching the web

Chemistry & Biology September 1999, 6:R261–R262

1074-5521/99/$ – see front matter

The world wide web is a huge repository of information. Netcraft's July 1999 survey (http://www.netcraft.co.uk/survey/) indicated that there are at least 6.5 million servers hosting web pages. A survey of 2500 random servers estimated that each server hosts nearly 300 pages on average, and that ~6% of servers have scientific or educational content (http://www.wwwmetrics.com/). Taken together, these surveys suggest that there are over 100 million pages on the web that might be of interest to academics (and this is undoubtedly a gross underestimate). The aim of this Web alert is to give you some guidance on how best to search the web.

Directories and search engines

There are two major tools used to locate information on the web — directories and search engines. The difference between them is not obvious and is becoming increasingly blurred, but understanding the difference is important because what you are searching for will influence the tool you use.

A directory is a human-generated list of websites, complete with descriptions and some type of classification system. Because people decide which websites are listed, there should be more signal and less noise. The first directory was Yahoo! (http://www.yahoo.com/) and it is still the most popular, and others include Looksmart (http://www.looksmart.com/) and the Open Directory Project (http://www.dmoz.org/).

A search engine is a computer-generated index of websites. A robot or spider crawls the web and automatically generates an index of all the words that it finds. This index can then be searched by users. Although this means a search engine should be comprehensive, it also means the quality of the search results may vary. You can choose between a number of search engines, but the major ones are AltaVista (http://www.altavista.com/), Excite (http://www.excite.com/), HotBot (http://www.hotbot.com/), GO (http://www.go.com/), Lycos (http://www.lycos.com/) and Northern Light (http://www.northernlight.com/).

If you are searching for a popular or very general term then it is best to use a directory. For example, searching for "biology" using Excite will find 250,000 pages, most of which will be irrelevant. Searching for "biology" on Yahoo!, however, produces a list of 121 biological categories that allows you to drill down quickly to your area of interest.

If you are searching for a more esoteric subject you will need to use a search engine. For example, searching for "auxin-binding protein" using Yahoo! does not produce any matches from its directory, but the same search using Excite finds 22 web pages.

Coverage

The distinction between directories and search engines is not as important as it once was because many search engines now have their own directories (for example, AltaVista uses Looksmart's directory) and many directories default to a search engine if no matches are found (for example, Yahoo! uses a search engine powered by Inktomi, a company that specialises in developing search engines). Search engines do not cover the entire web, however. A recent experiment (http://www.wwwmetrics.com/) indicated that the largest search engine was Northern Light, but that it only indexed 16% of the web. When combined, the 11 search engines that were surveyed were estimated to index only 42% of web pages. Since these results were published, a relatively new search engine, called FAST Search (http://www.alltheweb.com/), has claimed to index 25%, but, nevertheless, no search engine covers the entire web.

One way to improve coverage is to use a metacrawler — a service that queries many search engines simultaneously and consolidates the results. This effectively means that more of the web is searched. For example, searching for "auxin-binding protein" on Go2Net (http://www.go2net.com/) produces 42 matches, which is nearly twice as many as Excite alone. Other metacrawlers include Dogpile (http://www.dogpile.com/) and SavvySearch (http://www.savvysearch.com/). Importantly, metacrawlers only combine results from search engines, so it may be more appropriate to use a directory for a very general search.

Search engines are, however, trying to improve the results they produce for general search terms. One method, used by Google (http://www.google.com/), is to give a higher ranking to sites that are heavily linked to from other sites (because the best sites will be linked more often).

Subject gateways

All the search engines and directories mentioned above attempt to cover the complete spectrum of web-based information. Another method, which is becoming increasingly more common, is to produce subject-specific search engines and directories. These services are generally known as subject gateways. For example, SciCentral (http://www.scicentral.com/) is a directory of scientific websites. More specific examples are Chemie.DE (a chemistry subject gateway; http://www.chemie.de/) and Biocrawler.com (a biology subject

gateway; http://www.biocrawler.com/). Often more than one gateway covers the same subject area, for example, ChemDex (http://www.chemdex.org/) is another chemistry subject gateway. As yet, there is no comprehensive directory of subject gateways, but Search Engine Watch (http://www.searchenginewatch.com/links/) and PINAKES (http://www.hw.ac.uk/libWWW/irn/pinakes/pinakes.html) are good starting points.

There are many options available for searching the web, but no single one is suitable for every possible search. This Web alert has attempted to guide you in choosing which tool to use for a particular search. It is possible, however, to further optimise your searching by constructing your search terms carefully, which will be the subject of a future Web alert.

Ian M Sealy, Institute for Learning and Research Technology, University of Bristol, Bristol, BS8 1HH, UK; ian.sealy@bristol.ac.uk