search insider

In the second of three articles on Internet searching (see also *IWR* 141 November 1998), David Green continues his exploration of ways to get what you want — or, sometimes perhaps, what you need — from the World Wide Web.

We all know the Web is exploding in size (a billion Web pages predicted by the end of 1999). Meanwhile the percentage of Web sites indexed by search engines is falling. However, this is not to say that the number of Web pages indexed is decreasing, and more available pages means more hits.

The way search engines work is to index every single word on a Web page, to search the full text and then apply mathematical algorithms that will rank the results. Unfortunately, the algorithms in use are only partially successful and the increasing number of hits often corresponds to decreasing relevancy.

Clearly the Web in general, and search engines in particular, need to improve, if users are to obtain the valuable information they seek. Developments such as XML (see *IWR* 144, February 1999 for a fuller discussion of XML) herald radical improvements in how information will be published on the Web in the near future. Alongside this, other technological developments to improve searching on the Web include:

- Proprietary software tools that work alongside search engines.
- Search engine technology that analyses the link structure of the Web.
- Intelligent agents that can search 'the invisible Web'.

Search utilities and plug-ins

NPD Online Research has produced statistics which reveal that 77 per cent of Web users access multiple search sites simultaneously when they are searching the Web. Consequently 'meta-search' sites such as Dogpile and Mama are becoming increasingly popular. There is also a growing range of commercially available software tools such as Bullseye and Mata Hari that perform meta searches while offering greater search and results analysis functionality. Such programs are referred to as search utilities.

In the face of such commercial utilities, some search engine providers have released a variety of free basic search utility programs as 'plug ins'. As the name suggests, once installed, they are incorporated within the user's Web browser. By shifting processing power away from their own servers, and onto the user's own desktop, such plug-ins enable the search engine provider to offer more features. Two of the most popular tools currently available are Infoseek Express and AltaVista Discovery.

Infoseek Express

Infoseek Express is designed for people who search the Web frequently. It is a meta search tool in that it allows the user to search multiple search engines and Web directories simultaneously (for example, you might be searching using Infoseek, Yahoo! and Excite all at the same time). In total, Infoseek Express can search over 300 sites, including a selection of 'favourite sites' that can be added by the user.

Search strategies, which can be saved for future use, can be narrowed by searching within predefined categories, such as current news, stock quotes and so on. These categories can be customised according to personal preferences. As Infoseek Express supports multiple users, individual search strategies can be password protected.

When the search is complete, users can pre-select the aggregated, de-duplicated results they wish to view, and in what order. While you are reading a particular page from your list of results, Express will download other Web pages in the background, thus reducing the overall time required to scan through the results.

AltaVista Discovery

Unlike Infoseek Express, AltaVista Discovery will search not only the AltaVista database of indexed Web sites but it will also search across the user's own computer and intranet. While it doesn't offer some of the advanced features that can be found on Infoseek Express, it does offer other useful features. These include a 'more like this' search strategy, called 'find similar pages', and 'summarise' which will, not surprisingly, summarise long documents according to the format and length that you have specified.

If the user looks at a particular Web page from their list of results, Discovery will find all other pages on that site and then provide a description of them. Discovery also enables the searcher to view Web sites that make reference to the information in their search results.

Future developments

Of course, humans can only process a certain amount of information at any given time. Consequently, search engine technology is focusing not on increasing the size of the database index but on ways of improving search capabilities. Until quite recently, search engines only examined the location and frequency of words and terms within Web pages to build their database indexes. However, a Web site is also important if it is popular, has lots of hyperlinks elsewhere referring to or connecting to it, and if it itself refers to other related Web sites.

➳p20

➻p19 Three key technologies have emerged that seek to exploit these ideas. Two – Google and Clever – are based on analysing the links structure of the Web, while the third – called Direct Hit – is based on the popularity of a Web site.

Google

Google was developed by students at Stanford University (don't be surprised, so was Yahoo!). This technology uses a methodology known as PageRank (named after Larry Page, one of its creators) to crawl the Web and analyse how Web sites link to each other. The results are ranked on 'importance' – ie how many other Web sites link to them. If you, as a Web site author, have included hyperlinks to other sites that you deem important, then you have exercised some editorial judgement. In the same way that Web directories such as Yahoo! are compiled by editors on a manual basis, Google seeks to capitalise on the editorial judgement of the millions of Web site authors, but on an automatic basis.

As a result, of course, it can analyse far more Web sites than the humans who build directories such as Yahoo! and Mining Co. In fact, unlike search engines, which in reality become less useful the larger their index of Web sites, Google claims to return even better results from a bigger index. Google also seeks to capitalise on the accompanying editorial commentary by processing the text around each hyperlink.

Clever

A team of IBM researchers examining search engine effectiveness developed a system which they referred to internally as HITS – which stood for Hyperlink-Induced Topic Search. Following some further tweaks, the marketing staff became involved and the projected was branded 'Clever'. Yes, very.

Based on the citation index – which is widely used in academia – Clever examines the hypertext context of a keyword search. Like Google, Clever examines hyperlinks and the surrounding commentary. Unlike Google, which crawls the Web, Clever first submits the query to a search engine such as AltaVista, and then conducts its links analysis on a core set of pages from the results produced by that search engine. So, in other words, its results would not be as comprehensive as those provided by Google's larger data set for each enquiry.

Clever's next step is to use the information from its analysis to sort and rank the results, which are divided into two categories: authorities and hubs. Authorities are Web pages about a particular topic that have lots of links to them, ie they are authoritative sources of information. Hubs are Web pages which are a guide to, or list of, authoritative sources, ie they do the most citing. Hubs are similar to portals in that they act as a starting point for anyone interested in the topic that they cover. IBM has been experimenting with using Clever automatically to develop Yahoo!-style directories.

Unsurprisingly, Clever's ability to identify and target a huge range of different types of audience and community has attracted the large portal sites. IBM is seeking to licence its new technology not only to the larger portal sites but also to organisations with large intranets that are seeking to create their own internal directories.

Direct Hit

Search engines such as Excite match search terms with the content on indexed Web pages. Web directories identify, review and index Web sites on an individual basis.

Based on the concept of 'popularity', Direct Hit adopts a third approach. As the ranking of results is based on those Web sites that users have visited, Direct Hit claims to be 'user-controlled'. It isn't a separate search engine, instead being incorporated within existing search engines, one being HotBot.

Before licensing Direct Hit, HotBot would return a list of results which were based on the standard methodology of matching search terms with content on the Web sites in its index. Provided their search term is a popular one, searchers can now run a second-level ranking on their results by simply clicking 'Top 10 most visited sites for [search term]'. Direct Hit will analyse the list of results, identify those which are 'popular' and then re-rank the search results accordingly, with the most popular Web sites matching your search term presented first. Direct Hit is also available within Netscape Communicator 4.5 and in Apple's Sherlock search utility.

Of course, the popularity of a Web site can be largely determined by its search engine rankings and there are all sorts of ways to manipulate those if you have a good understanding of how search engines work. The sending of electronic junk mail and junk newsgroup postings – 'spamming' – is one of the most crude ways of doing this. Direct Hit tries to compensate for such activities by boosting the ranking of what it calls 'hidden gems'.

A Web site might provide lots of valuable information about a particular topic but, for a variety of reasons, it might feature quite a long way down the list of results produced by standard search engines. If a previous searcher has been tenacious enough to scroll down as far as result number 100 on the list (they'd probably be an information professional!) and have clicked on it, Direct Hit's algorithms will give this site a big boost up the list of results when it next appears in other searchers' lists. If subsequent users don't click on this 'hidden gem' – in a sense, if they don't feel it deserves its high ranking – it will again drop down the list of results for subsequent searchers.

Search engine plug-ins and improved interrogation technology are fine; but they only query the search engine database index. In my previous article (*IWR* 141, November 1998), I discussed how, as a result of the trend towards integrating databases to Web sites, information was increasingly accessed via the Web, not on it. The information that resides in these databases is referred to as 'the invisible Web'. As this information doesn't exist in Web page format, it's not publicly available for view or for access by search engines to search in. Therefore, it's necessary to visit and interrogate each and every database that you are interested in, separately. In the third, and final, article, I'll be looking at the role of intelligent agents in interrogating databases on the Web.

David Green can be contacted via his Web site at www.clickmedia.freeserve.co.uk.

Useful links:

• Infoseek Express	www.express.infoseek.com
• AltaVista Discovery	www.discovery.altavista.com
• Dogpile	www.dogpile.com
• Mama	www.mama.com
• Bullseye	www.intelliseek.com
• Mata Hari	www.thewebtools.com
• Direct Hit	www.directhit.com
• Clever	www.almaden.ibm.com/cs/k53/clever.html
• Google	www.google.com