LOST IN CYBERSPACE
by DAVID BRAKE

Just as a library is only as good as the index that lists its books, the World Wide Web is only as useful as the search engines that service it.

On the surface, the Web is doing pretty well, thanks to remarkable search engines such as AltaVista, Infoseek and Lycos. You can search millions of Web pages simply by typing in a key word or phrase. Within a few seconds a list of pages that meet your criteria appears in front of you.

But the many users of these search engines may be surprised to learn that they cover fewer than half the pages available on the Web. And with the phenomenal growth of the Web and new methods of presenting information on its pages, the search engines are falling further and further behind. "Nobody can afford enough hardware to index the whole Web and serve it back to the entire planet," observes Louis Monier, chief technical officer at AltaVista Software.

Just how much information is out there is impossible to say, not least because anyone with a computer attached to the Internet can publish Web pages. Monier estimates that when AltaVista's search engine was launched in 1995, the Web contained around 50 million pages on 100 000 sites. Now, he says, there are between 100 million and 150 million pages on around 650 000 different sites. Meanwhile, most Web indexes have hardly grown at all over the past year.

The companies that run the search engines say that rather than making their databases bigger, they are putting their efforts into making it easier for users to find what they want on the Web. Adding more pages to a database often makes little difference to the user, they argue. "If the answer people need is in the first 50 000 pages searched, do you need the other 50 000?" asks John Nauman, vice-president of engineering at Infoseek. And as Monier points out: "There is so much duplication of information on the Web. There may be ten more versions of the answer I am looking for elsewhere unindexed, but I won't mind as long as I get my answer." He says a study has shown that, if users perform a search on more than one database, they will often believe that a small one is more effective than a larger one, because it offers them fewer "hits" and they assume that it is giving them more precisely targeted answers.

The people who complain most about the shortcomings of search engines are those who publish Web pages. In many cases, search engines are the only mechanism they have to announce themselves to the world. John Pike, webmaster for the Federation of American Scientists, was horrified when he discovered that only 600 of the 6000 pages of his site were indexed by AltaVista. "This is like buying a phone book that only has even-numbered phone numbers," he says.

Danny Sullivan, a London-based consultant studying search engines, recognises the problem that search companies are up against. "They all face an uphill battle in keeping up with the growing Web." But he feels they should be more open about their limitations. "If they are just sampling the Web, users should understand that."

AltaVista is concentrating on providing at least a sample from every Web site, says Monier. Only on the most frequently visited sites does it try to index the majority of the content. Although this works well for those favoured sites, and also for very small sites where a few pages might give a good indication of their content, it presents big problems for other larger sites such as Pike's. But Monier defends the approach. "Our current policy is based on fairness," he says. "Every site gets a chance to be represented somehow."

Infoseek adopts a similar approach. According to Nauman, its database now contains information on 25 million to 30 million pages of text. But about 90 per cent of queries are answered using the most frequently accessed 1 million pages. And more than 90 per cent of the pages in the database are never accessed as the result of a search. The company is therefore concentrating on making sure that the information it has already indexed is as up-to-date as possible, although Nauman admits that there are those inside Infoseek who think the company should be aiming to cover the whole of the known Web.

Maintaining the quality of a search engine's index is made especially difficult by the Web's volatility. A random sampling of pages on 2000 sites over three months in 1995, carried out by two US universities as part of research into a new indexing system, indicated that the average time a page of text remained unchanged on the Web was just 75 days. A substantial percentage changed every 10 days or less. Sometimes pages disappeared entirely, but more often the information they contained was simply updated, or the page was moved to a different address.

Nauman says that around 10 per cent of the pages indexed in Infoseek's database no longer exist. This, he believes, is one of the biggest frustrations for users. The company now plans to visit all its indexed pages repeatedly over a period to determine how often they change. Those that change infrequently will then be checked every two months, while those that change more regularly or are looked at more often will be checked every day or two.

All search engines work in a broadly similar way. They send out programs called "spiders" to scan and catalogue the Web automatically by following the links between documents. The pages the spiders collect are indexed by key word and stored in huge databases that can then be accessed by the public. A site will generally not be picked up by the spider unless another site links to it, or its owner registers it manually with the search engine.

In the early days of the Web, pages carried only simple text. What is more, they were open to anyone who cared to view them. Now, as technology advances and different organisations pursue varied commercial goals, the Web is beginning to fragment. Some information cannot be read with an ordinary Web browser such as Netscape or Internet Explorer. For example, the documents in the vast collection of scientific papers on a variety of subjects held by Los Alamos National Laboratory in New Mexico and known as the "xxx" e-print archive--contain numerous complex diagrams and formulae. These can only be properly displayed and printed if the user has installed additional software that can cope with the software company Adobe's PostScript or PDF file formats.

Sites that contain mainly text can be difficult to index, too. The New York Times, for example, is one of a number of sites that deliberately bar users who have not paid an entrance fee or filled out a registration form. The spider programs cannot easily register to view protected sites, although some do now index many of the more important ones.

Some sites that are officially open to all may cause other difficulties for the spiders. Paul Holbrook, director of Internet technologies at CNN, says the company used to exclude all search spiders from its site because they distort the figures that the company records on the numbers who visit it. Many Web publishers use these data to sell advertising space. The British Library's newly unveiled online catalogue is an example of another kind of site that will defeat a search engine's spider: it contains information about millions of books and periodicals, but it is visible only to users who have made a specific search through the library's online query form.

Sangam Pant, vice-president of engineering at Lycos, believes that faster software and more powerful computers may keep search engines abreast of the growth of the Web for a while. But in the long run "trying to keep up using brute force is not the answer," he says.

Instead, search engines may come to rely increasingly on "meta-data"--brief descriptions of the contents of a page or site that are embedded in the Web pages but read only by searchers.

Search engines already use meta-data to a certain extent to index pages. In the future, instead of striving to maintain a database containing the text of every page on every site, search engines may use meta-data to index entire sites. But the system will only work if Web publishers can be relied on to describe their pages accurately; no computerised catalogue would have the resources to check the validity of every description. "We could provide Web browsers with a lot more value if we could rely on the key words we are given," says Pant.

But all too often this is not the case. Already, the meta-data fields of many Web pages are stuffed with a small number of key words repeated over and over again in an attempt to ensure that search engines place the page high up in the list of hits it displays to the user. The designers of search engines are aware of this trick and try to compensate for it.

Monier is not optimistic about using meta-data to solve the accessibility problem. Instead, he foresees an increasing number of specialist search engines and indexes providing in-depth coverage of the Web by language or by subject. He suggests that the difficulties facing search engines may be overstated. And he believes that the exponential growth of the Web is bound to stop at some point. There is already evidence that the growth is slowing, he says. At the end of the day he points out, "if users felt they had serious difficulty finding what they wanted, they would stop using us".

But Sullivan thinks there is still plenty of work for the search engine companies to do. "At some point, if you are going to run these services, you have to take responsibility for maintaining standards," he says. If you know it is not good enough, you should do something about it whether your customers are complaining or not.

Related sites:

Search Engine Watch - an excellent site by Danny Sullivan with detailed information on how all of the search engines work and how they compare to one another.

ZDNet Talkback in late March featured an interesting exchange between John Pike at the Federation of Amerian Scientist's site and the Chief Technology Officer of AltaVista, Louis Monier about the limitations of his search engine.

From New Scientist, 28 June 1997