search insider

David Green considers the search engines' claims about Web coverage

The second NEC study into the *Accessibility of Information on the Web* was published in the 8 July issue of the science journal *Nature*. Its impact has been dramatic. It found that, of the 800 million publicly indexable pages on the Web (as of Feb 1999), search engine coverage has decreased substantially since the first NEC study, published over a year ago (which I covered in *IWR* 141, November 1998). The main search engines' rankings are given in the table to the right. Since its publication the report has had a number of interesting repercussions.

Rank	Search Engine	% Coverage
1	Northern Light	16.0
2	AltaVista	15.5
3	SNAP*	15.5
4	HotBot*	11.3
5	MSN Search*	8.5
6	Infoseek	8.0
7	Google	7.8
8	Yahoo!*	7.4
9	Excite	5.6
10	Lycos	2.5

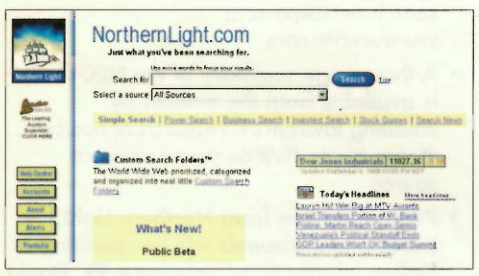
* Powered by Inktomi. Not all companies access Inktomi's full 110M index. Note also that HotBot is part of the Lycos network.

Combined, the search engines only index 42 per cent of the available Web. The study also uncovered a bias in search engine indexing:

- Search engines are more likely to index popular sites (popularity defined as number of links to a site).
- They are more likely to index US sites, and commercial rather than educational sites.
- Indexing of new and modified pages can take months.

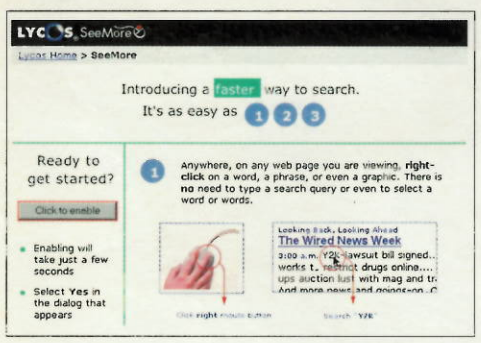
Users see the (northern) light

Coming in at pole position has benefited Northern Light enormously. Its traffic increased three-fold after the study's publication. The immediate strain that this placed on their infrastructure forced the company to reduce the number of hits presented in each search results page from 25 to 10.



Lycos lurks near the bottom

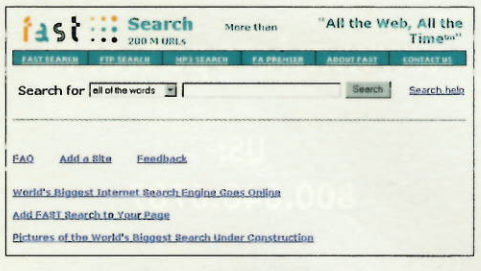
When questioned about the study, in an interview on www.silicon.com, the UK Managing Director of Lycos, Charles Walker, commented: "It's actually a good thing if you have a search engine that doesn't cover 100 per cent of the Web pages out there, because a lot o f them, quite frankly, are not worth looking at." The NEC



study showed that Lycos had one of the smallest indexes and contained the largest percentage of dead links. Frankly, users wouldn't have been able to view many Web pages using the Lycos search engine index. *The Sunday Times* rubbished Mr Walker's comments as "a gobsmacking piece of arrogant balderdash". Right on.

Going everywhere fast

This is not the mentality at new upstart FAST, the Norwegian company behind the Alltheweb search engine. As their URL suggests, the company is seeking to create a search engine index that will encompass everything. If successful, users will be able to go everywhere on the Web.



Such information democracy would deliver some of the Web's potential to offer equal accessibility of information and to overcome the bias in indexing that the NEC study highlighted.

FAST's approach is blindingly simple. Rather than use the monolith mainframes ➳p32

p31 adopted by all other main search engines, the company has linked together a few hundred Dell PCs (Dell has a 5 per cent stake in the company) and uses parallel processing to accomplish its goal much more cost effectively than the mainframe model. By August, FAST had established a new milestone when it announced a 200 million page index – making it the largest ever search engine index. The company aims to increase this to 300 million by December. Not surprisingly, the growth of FAST's share price has been dramatic.

Lucky link for Google

The Web is so-called because Web site links connect a constellation of computers in an integrated universal structure – like a spider's web. Link

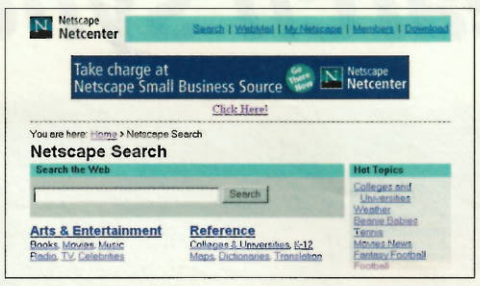
analysis is a factor in the relevance-ranking algorithms of several search engines such as HotBot and Excite. However, by focusing exclusively on this method, Google has emerged as the clear leader of the link-based method of searching the Web. Like Northern Light, Google has benefited enormously from its favourable results in the NEC study. It recently announced an agreement with AOL subsidiary Netscape to be the main search provider on the Netcenter portal. This link-up has resulted in a huge surge in users accessing the Google index. Although Google may only index about 70 million Web pages, through its link analysis, it claims to allow users to reach up to 300 million Web pages – a greater reach than anyone else – for the time being.

Like FAST, the company has stated that it is not interested in becoming a portal but will focus on co-branding its search technology. Like FAST, Google intends its Web site primarily to act as a showcase for its technology – after all it's been in beta mode for over a year. In its press material, Google describes itself as offering 'co-branded Web search and site search solutions for information content providers'. Considering the Netscape deal is the company's first, this could be regarded as a very interesting statement of intent. Watch carefully.

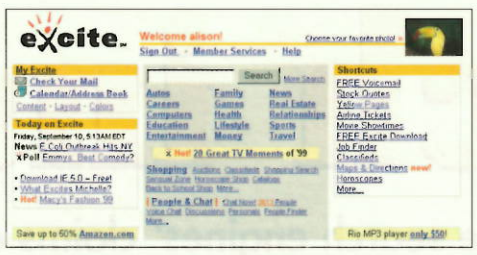
No deep thinking

Storm clouds lurk on the horizon. Link-based searching could be endangered. A small, but growing band of companies is trying to argue that certain kinds of links are illegal. Culprits include Ticketmaster (who started all this) and Universal Studios. Concerned about 'deep linking' – that is, links to pages buried deep within their sites – they've used their lawyers to force other Web site owners, such as Movie-List.com, to remove such 'deep links'. Instead, they request the other Web site owners to redirect all links to their home page.

These companies argue that 'deep links' bypass their home pages on which they earn revenue by selling ad space. This is despite the fact that the 'deep link' could be taking users directly to a page where they can book tickets for a particular concert or movie – on which of course Ticketmaster and Universal Studios will earn income. Not to mention the ads that will probably be on that particular 'deep-linked' page too. Perhaps we should ask *The Sunday Times* for their opinion.



All this is in spite of the fact that, in investigating the "probability of indexing random sites versus 'popularity' or connectedness of sites", the NEC study proved that sites with few links to them have a low probability of being indexed. Low index presence equates with lower hit rates,



which is a perfect example of the failings of short-term corporate greed. Bear in mind also that IBM does not currently make its links-based Clever indexing technology available for Web-wide searching due to 'legal liabilities'. This goes to prove that old joke – that three lawyers up to their necks in sand is just not enough sand.

Old dogs, new clicks

The first wave of search engines, which have become a 'Big Five', transformed themselves into portals. To increase 'stickiness' so that people spend longer on the site, they offered a variety of other services such as free emails, Web hosting and personalised news headlines. As a result, work on improvements to the core search technology was neglected.

This created the opportunity for a second generation of search engines such as Google, FAST, Direct Hit and Inktomi. Search functionality has proved resiliently core to Internet users' needs – and these companies have been offering better technology to fulfil this need.

According to data from PriceWaterhouseCoopers and research firm IPO Monitor, in the last year search engine companies have raised more than $274.7 million in private funds and another $282 million in public offerings – almost all of which is going to this second generation of search firms.

What is interesting is that none of the 'second generation' search companies have adopted the portal model so enthused about by their predecessors, the 'Big Five'. Perhaps this could prove to be the smartest trick to learn from.

Time will tell but the NEC study, which acted as both a benchmark and an audit, together with the threat posed by these new upstarts, is forcing the original search engines to improve.

Size Wars

The 'Big Five' search engines have all have joined the size war and have announced new features. Northern Light is increasing its index size. Lycos now also derives its primary results from the Open Directory, rather than its own index (funny that). It has also launched a search utility called 'See More' and has signed up Direct Hit to provide popularity-orientated results. As we went to press, Excite was due to unveil a 250 million page index that also offered 'twice as powerful' a search functionality.

Competition from the 'second generation' has alarmed the oldies. The NEC study helped to focus the public mind on search engine performance. Together these two factors have unleashed a burst of innovation and improvement. Smashing!

To the authors of the NEC study – your impact has been very real and very positive. Would you be kind enough to do this on an annual basis? Everyone would love you for it – except perhaps, Web search engine companies.

Further information

- www.northernlight.com
  www.lycos.com/seemore
  www.alltheweb.com
  www.google.com
  search.netscape.com
  www.excite.com
- A three-page summary of the NEC study is available from the authors by emailing lawrence@research.nj.nec.com. Please quote *IWR* as your reference source.
- Representatives from Northern Light, Excite and Alltheweb will be speaking at Online Information Conference 99. David Green will also be speaking on 'Making the Web work as a source of sales and marketing data' and on 'The evolution of Web searching'.
  www.online-information.co.uk