search insider

David Green returns to the Search Insider arena to review developments since his last visit, and to describe the likely future of searching on the internet.

Web searching has come a long way. Each year the number of search providers and portals continues to grow, while the variety of technologies and approaches used to deliver the results to the intrepid searcher continues to expand.

Just as well: researchers at IBM published findings in *Scientific American* in which they estimated the volume of information being published onto the web was a daily deluge of over one million pages (Members of the Clever Team, 'Hypersearching the Web', *Scientific American*, June 1999). So what have been the major developments and where are we heading?

Static attraction

First it is important to clarify what we mean by 'the web'. Despite its uniform interface and seamless linked integration, the web is not a single coherent element. Essentially, there are two types of web page: static (manually produced) and dynamic (database driven). The major differences are given in the following table:

Static web pages	**Dynamic web pages**
manually produced	computer generated
generic information	customised information
most are indexable	not indexable

Available for indexing to all search engines, static web pages together constitute the 'visible' web. The 'invisible web' largely refers to database driven, dynamically generated web pages. For example, most publishers distribute their data on the web by integrating huge databases with a front-end search interface. By virtue of its professionally published origin, such information is typically high value and more highly structured and indexed than the 'visible' web. The user's search enquiry will generate customised, as opposed to generic, results. Nonetheless, the 'visible' web constitutes a significant contribution to the dissemination of human knowledge, and as the NEC studies (see 'Search Insider', *IWR* Oct 99) have acknowledged, 'much of (this) material is not available in traditional databases'. A very attractive resource.

Static development

Two approaches sprang up to address the demand for finding information – automated search engines and manually compiled directories. The early providers of these services (such as Yahoo!, AltaVista, Excite and Lycos) began to focus on realising ecommerce opportunities and generating advertising revenue rather than building their core service. Portalisation also began to blur the boundaries between the two types of services as the search engines licensed web directories and vice versa, while other portal players licensed both types of service.

From late 1996 until September 1997 the growth of the main search engine indexes and web directories was negligible (Sullivan, Danny, 'Search Engine Sizes', September 1999, Search Engine Watch: www.searchengine.com/reports sizes.html), despite the relentless growth of the web. This period was also marked by a distinct lack of search technology innovation. Although meta search engines such as Mamma, Dogpile and Metacrawler first rose to prominence during this period, their search functionality was essentially based on the keywords approach was developed by the main search engines. It was not until the arrival of a 'second-generation' search engine providers in 1998 that new approaches became available.

Moving with the next generation

'Second generation' refers to a more recent family of search providers who share some common characteristics. Many of these firms have eschewed the portal model adopted by their earlier predecessors. Several cannot be accessed directly, but are licensed to the major portal sites as refining technologies. The variety of technological innovations that these firms have developed include results determined by web ➤p 26

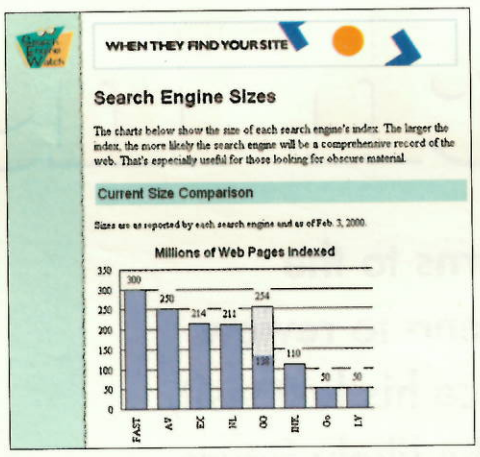
➠p25 site popularity, natural language searching and links-based analysis.

Launched in April 1998, Direct Hit represented a radical departure from the traditional search engine approach based on location and frequency of terms. The system claimed to be 'user-controlled', as the ranking of results is based on web sites that users have visited.

First generation search engines also failed to consider the context of the search terms – the syntactical relationships between the search terms and other vocabulary within their index. By looking for literal, exact matches they also failed to consider semantics or use thesauri. June 1998 represented a landmark in addressing these limitations. Two new search engines were launched, both offering natural language searching. The Electric Monk uses Artificial Intelligence to analyse user's questions and convert these to complex Boolean search statements which were then submitted to the AltaVista index. Ask Jeeves attempts to match users queries to its database of template questions whilst also conducting meta-searches across the major first-generation indexes and directories.

Linking it all up

The first-generation search engines focused on the content of each page in their indexes with little consideration of how those pages inter-relate. Links-based analysis attempts to overcome this limitation by examining the relationships between pages – the one billion or so hyperlinks that weave the web together (Sullivan, 1999). In this way links-based analysis offers methodologies for identifying authoritative sources of topic-specific information, eliciting highly relevant results. While this approach does factor in the relevance algorithms of several first-generation engines, Google is the only search engine that is exclusively focused on links-based searching currently publicly available for web-wide searching. IBM has developed two interesting technologies called Clever and Focused Crawler that focus on links-based analysis. However, neither is publicly available yet for web-wide searching.



Search Utilities

Meta-search sites such as Dogpile and Mamma grew in popularity, as they allow to users to search across different search indexes simultaneously with results de-duplicated and re-ranked. Search utilities represent the logical evolution of this functionality. Unlike meta-search engines, where the processing power to refine results still remains on the server that the user is interrogating, search utilities are programs that are installed onto the user's hard drive. By shifting processing power away from the server, and onto the user's own desktop, search utilities offer a much greater range of search and results analysis functionality. Popular search utilities include BullsEye, Copernic and Mata Hari.

Intelligent Agents

Like several of the second-generation search technologies (Electric Monk, Google) many of these search utilities incorporate intelligent agents (or bots). Indeed, many of the powerful features offered by search utilities, such as language-independent searching, filtering, automatic refinement of results and document summaries, active hyperlinking of query words and live highlighting of search terms are possible because of the nature of intelligent agents.

A common function of agents is that they allow the user to specify a high-level goal instead of issuing explicit instructions – leaving the 'how' and 'when' decisions to the agent. This – combined with their abilities to search across data in unstructured format, to automatically learn and adapt to user preferences and to identify patterns – is giving agent technology an ever-increasing role in web searching.

The Future

So where are these trends taking us? What could the future of web searching look like? Much of the search technology innovation over the last eighteen months has come from second-generation search companies.

By focusing on portalisation and ecommerce the first generation of search firms have ceded control of technological innovation to their younger cousins. However a symbiotic relationship appears to be emerging – second generation firms often need access to the large indexes that have been developed by the first generation firms. The cost of developing such a large index was deemed prohibitive, acting as an effective barrier to entry. One public exception to this are Norwegian upstarts FAST.

They re-wrote the business model by adopting PCs and parallel processing in favour of mainframes. Despite quickly building the world's largest index of the web (300 million plus and heading for 400 million by mid-2000) this younger firm still needs access to the more developed experience and expertise that older first generation firm Lycos can offer. Such strategic alignments are certainly in line with current business process outsourcing (BPO) trends and should prove popular with these firm's institutional investors.

Will we be charged micro-payments for web searching? Unlikely – at least in the near future. Just look at the current bloodbath over free unlimited internet access in the UK, with AltaVista leading the way and Excite announcing its intention to join the fray.

XML will become very important. Unlike HTML, it allows the manipulation and exchange of information across different platforms. I discussed the impact of XML on search engine development in my last Search Insider feature (*IWR* Dec 99). Together with wireless application protocol (WAP), the future of information retrieval is anything anywhere you want.

Nobody does it better

Finally, despite all this technological innovation, human categorisation is also becoming increasing important. Web directories such as About.com, LookSmart and Yahoo! have all been joined by Lycos, MSN Search, AOL Search and Netscape who all now use human editors to improve and refine the categorisation of information in their indexes. Between November '98 and October '99, the volunteer based Open Directory project grew from just under 5,000 contributors who had catalogued 84,000 web sites to over 16,500 contributors and about 1 million web sites. No wonder their slogan is 'Humans do it better'.

David Green is a council member of the Institute of Information Scientists. He writes here in a personal capacity. His web site is www.clickmedia.freeserve.co.uk.