personalized search needs may be met by dedicated science search portals. These webs within the web will concentrate many of the online resources you need within an easily navigable environment.

The various online repositories of the scientific literature may by then have adopted common standards to allow seamless searching across them. And the parallel development of 'intelligent' search software could mean that, as you type e-mails or word processing documents, your computer automatically delivers suggestions about relevant web resources, tailored to your particular interests and expertise.

Quality, not quantity

Portals are a hot topic on the web at the moment. The idea is to organize related content so that it can be searched in isolation from the web as a whole. This approach trades off the sheer scale of the available content against quality and ease of navigation. It also allows search engines that are overwhelmed by the public web to perform excellently.

Popular search engines have cottoned on to the trend. The Hotbot engine, for example, lets users search only academic sites with a domain name ending in '.edu'. "Soon you will see a whole slew of search engines specializing in particular sectors," predicts Sridhar Rajagopalan of IBM's Almaden Research Center in San Jose, California.

For the present, however, the biggest innovation in search engine technology takes its inspiration from the citation analyses used on the scientific literature. Conventional search engines use algorithms and simple rules of thumb to rank pages based on the frequency of the keywords specified in a query. But a new breed of engines is also exploiting the structure of the myriad links between web pages. Pages with many links pointing to them — akin to highly cited papers — are considered as 'authorities', and are ranked highest in search returns.

This approach has been pioneered by Sergey Brin and Lawrence Page, two graduate students in computer science at Stanford University in California. In less than a year, their Google search engine has become the most popular on the web, yielding more precise results for most queries than conventional engines — and transforming the lives of its developers. "I haven't finished my PhD," says Brin. "I'm afraid to say I've been too busy with Google."

Google's algorithms rank web pages by analysing their hyperlinks in a series of iterative cycles. "We don't just look at the number of links, but where they come from," explains Brin. "A link from the *Nature* home page will be given more weight than a link from my home page; more things point to *Nature*, therefore it is likely to be more important, and more important things tend to point to *Nature*, which again suggests that *Nature* is a more important authority."

Boy wonders: Brin (right) and Page's Google search engine is the most popular on the web.

Whereas most search engines only associate the text of a link with the page the link is on, Google also associates it with the page the link points to. This allows it to cover many more pages than it actually crawls, even yielding links to sites that bar search engines' crawler programs.

Clever, a prototype search engine being developed at IBM Almaden, takes the citation analogy further. Like Google, it produces a ranking of authorities, but it also generates a list of 'hubs', pages that have many links to authorities. Hubs are akin to review articles, which cite many top papers. Those that link to many of the most highly cited authorities are given higher rankings than those that link to less well-respected sites. "Not all links are equal," says Rajagopalan, the driving force behind Clever. Users get not only the top hits, but the hubs provide a good starting point for browsing.

However, search strategies that rely on analysing hyperlinks are no panacea. New pages will have few links to them, and may be missed. And such strategies may be of little use to a scientist seeking highly specific information found on specialist sites with few incoming links. "All search engines have their weaknesses for certain queries," says Brin.

Effective searching requires a mix of techniques. If you want to trawl for background information before beginning a research project, you might use an engine like Google to identify key sites, in combination with

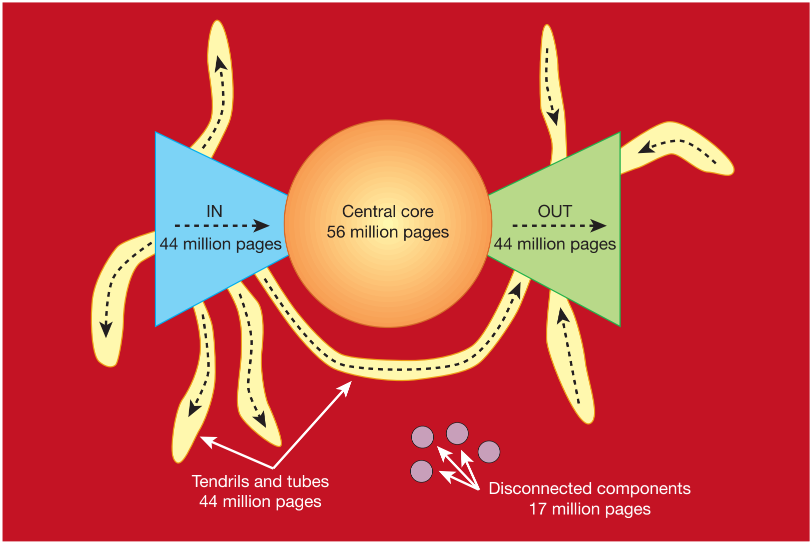
▶

The web is a bow tie

A study of the web's structure, five times larger than any attempted previously, reveals that it isn't the fully interconnected network that we've been led to believe. The study suggests that the chance of being able to surf between two randomly chosen pages is less than one in four.

Researchers from three Californian groups — at IBM's Almaden Research Center in San Jose, the Altavista search engine in San Mateo and Compaq Systems Research Center in Palo Alto — have analysed 200 million web pages and 1.5 billion hyperlinks. Their results, which will be presented next week at the World Wide Web 9 Conference in Amsterdam, indicate that the web is made up of four distinct components.

A central core contains pages between which users can surf easily. Another large cluster, labelled 'in', contains pages that link to the core but cannot be reached from it. These are often new pages that have not yet been linked to. A separate 'out' cluster consists of pages that can be reached from the core but do not link to it, such as corporate websites containing only internal links. Other groups of pages, called 'tendrils' and 'tubes', connect to either the in or out clusters, or both, but not to the core, whereas some pages are completely unconnected. To illustrate this structure, the researchers picture the web as a plot shaped like a bow tie with finger-like projections.


IN
44 million pages
Central core
56 million pages
OUT
44 million pages
Tendrils and tubes
44 million pages
Disconnected components
17 million pages