

Souped-up search engines

For scientists, finding the information they want on the World-Wide Web is a hit-and-miss affair. But, as Declan Butler reports, more sophisticated and specialized search technologies are promising to change all that.

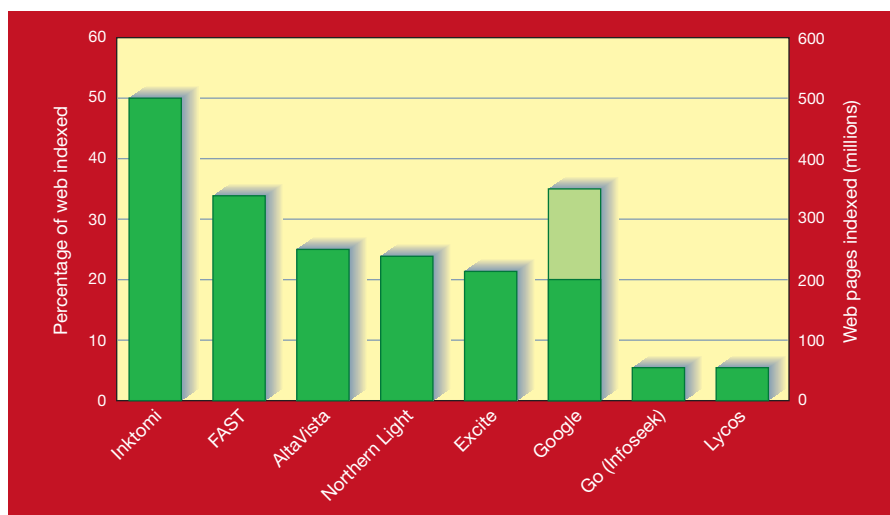
“What is nature?” would seem to be about as vague a question as you can get. But put it to Ask Jeeves, a popular Internet search engine, and its preferred response is clear: “*Nature*, international weekly journal of science”. Search with other leading engines using the single keyword ‘nature’, and *Nature* or its sister journals appear in the top ten returns.

Sadly, many scientific resources on the web are much harder to find. Search engines either miss them entirely, or make you scroll down dozens of pages of hits, your eyes propped open in the hope that pertinent links will leap out from the screen.

Today’s mainstream search engines are simply not run with scientists in mind. Leaving aside their inability to interrogate the vast amount of scientific information held in databases — be they of spectral lines of stars, genome data or events from particle colliders — scientists are poorly served even when they search for text on the web. “There are substantial limitations to search engines and they have bigger implications for scientists than for regular consumers,” observes Steve Lawrence of the NEC Research Institute in Princeton, New Jersey, co-author of a seminal paper on the accessibility of online information (see *Nature* 400, 107; 1999).

Crawling towards the answer

Most search engines rely on ‘crawler’ programs that index a web page, jump to others that this links to, index these, and so on. The starting points bias the end results, and tend to be pages on topics of mass interest — so sport is covered more extensively than quantum computing, for instance. Web pages purposely submitted to search



Falling short: most search engines only cover a fraction of the web. Google’s light green bar shows pages it indexes without crawling to. With permission from <http://searchenginewatch.com>.

engines — perhaps by their authors — can also gain prominence. Commercial websites, meanwhile, use a variety of tricks to boost their search rankings, filling their pages with strings of repeated keywords in a colour that makes them invisible to the viewer, or embedding keywords in the HTML (HyperText Markup Language) code that underlies a page.

If search technologies were to stand still, the phenomenal growth of the web would soon render them useless. There are already more than a billion web pages, and even the most wide-reaching search engines cover barely half of these (see figure, above). Within two years, the web may grow to 100 billion pages, and search engines face huge difficulties keeping pace.

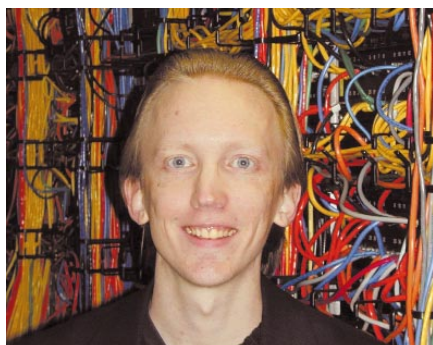
A new study of the structure of the web provides little comfort (see ‘The web is a bow tie’, page 113). This contradicts earlier suggestions that any two pages on the web are connected by a relatively small number of hyperlinks (see *Nature* 401, 131; 1999). The implication is that search engines must crawl from a greater diversity of starting points if they are to have

Rajagopalan: predicts a wave of specialist search engines.

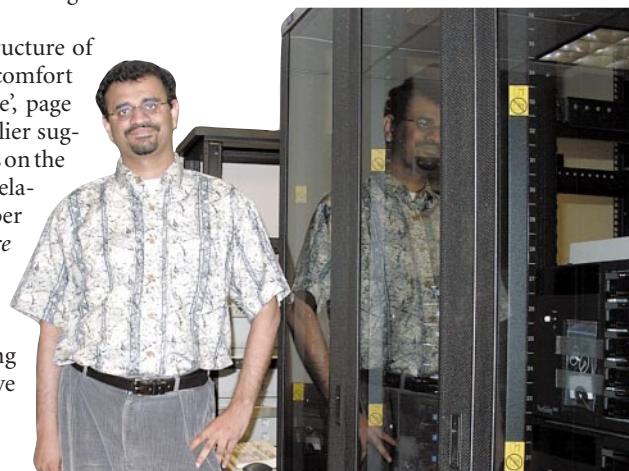
any hope of giving a reasonable breadth of coverage.

Thankfully, new search technologies promise to increase vastly the precision of web searches — and scientists are poised to reap the benefits. The introduction of XML (eXtensible Markup Language), seen as the successor to HTML for coding web pages, should make it possible to restrict a search — for instance to scientific papers, or even to papers that report work using specific biochemical reagents (see ‘The sweet XML of success’, page 114).

Experts predict that within five years, searching the entire web by keywords will be a thing of the past for most researchers. Your



Lawrence: scientific searchers suffer.



IBM

NEC

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



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

Nature, which again suggests that *Nature* is a more important authority."

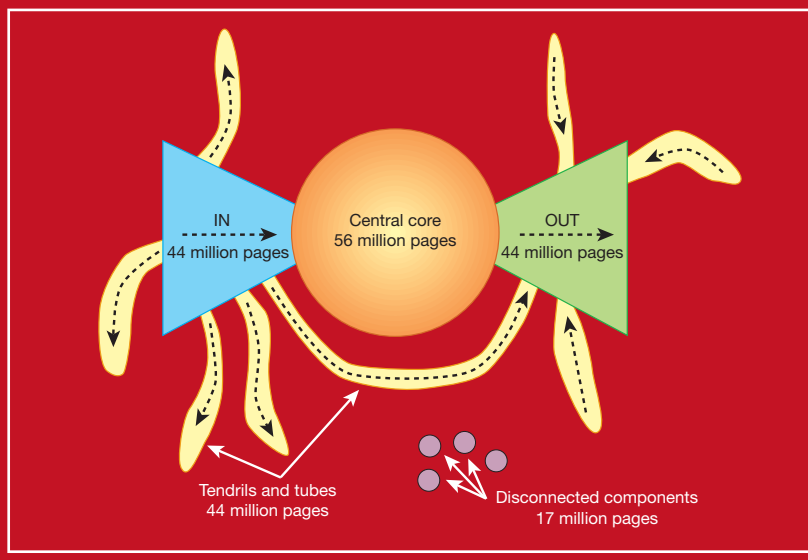
Whereas most search engines only associate the text of a link with the page the link is on, Google also associates it with the page the link points to. This allows it to cover many more pages than it actually crawls, even yielding links to sites that bar search engines' crawler programs.

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.



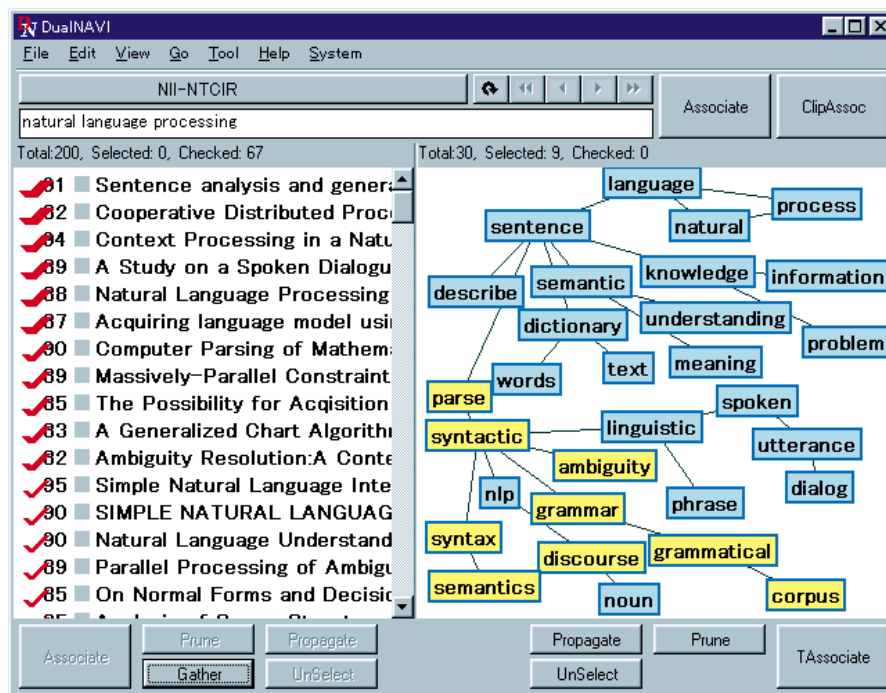
► ‘metasearch’ engines, such as Sherlock-Hound, which allow users to query multiple search engines simultaneously. The drawback of metasearch engines is that they can give enormous lists of hits. But NEC is working on a next-generation metasearch engine, called Inquirus. Rather than simply spewing out the results from other search engines, it reindexes this initial list to provide a better ranking. Inquirus also checks for broken links, and weeds out duplicate entries.

Searches for researchers

But what about search engines designed specifically for scientists? NEC’s prototype, called ResearchIndex, gathers fragmented scientific resources from around the web, and automatically organizes them within a citation index. And unlike most search engines, ResearchIndex retrieves PDF (portable document format) and postscript files, widely used by scientists to format manuscripts. It starts by querying dozens of popular search engines for a series of terms likely to be associated with scientific pages, such as ‘PDF’, or ‘proceedings’. Hundreds of thousands of scientific papers can be located quickly in this way, says Lawrence.

ResearchIndex uses simple rules based on the formatting of a document to extract the title, abstract, author and references of any research paper it finds. It recognizes the various forms of presenting bibliographies, and by comparing these with its database of other articles can conduct automatic citation analyses for all the papers it indexes. This information can also be used to quickly identify articles related to any indexed paper.

The prototype form of ResearchIndex is being applied to the computer sciences. Its archive of papers in this subject alone, at 270,000 articles, is bigger than leading online scientific archives such as the Highwire Press, which has almost 150,000 articles, and the Los Alamos archive of physics preprints, which contains about 130,000 papers. The



Helpful suggestions: DualNAVI generates a graph of potential keywords with which to refine searches.

engine already has an enthusiastic following among computer scientists. Stevan Harnad of the University of Southampton, who has tested the system on his CogPrints archive of preprints in the cognitive sciences, is another convert. “For the literature it covers, it is a gold mine,” he says.

NEC is giving the software free to non-commercial users, and Lawrence hopes it will be applied across many disciplines: “Our goal is to not just create another digital library of scientific literature, but to provide algorithms, techniques and software that can be widely used to help improve communication and progress in science.”

Concept albums

Other prototype search engines boast features that could make trawling the scientific

literature more efficient. A team at Hitachi’s Advanced Research Laboratories, in Hato-yama, Japan, is developing an engine called DualNAVI which could improve the efficiency of searches on collections of scientific literature such as Medline. Hits for a keyword are listed on the left-hand side of the screen. But DualNAVI also generates a set of related keywords by analysing the retrieved documents, and displays these on the right of the screen as a ‘topic word graph’ (illustrated above). Click any topic, and related documents are highlighted in the left-hand window. This often yields articles that the initial query missed. And in a further twist, groups of documents can be selected, indexed for keywords, and run against other literature databases to find related papers.

Collexis, a Dutch firm based in The

The sweet XML of success

Scientific information would be easier to find on the web if it were clearly marked as such. This is the promise of XML, soon to become the language of choice for web pages.

Current HTML coding tells browser programs little more than how a page should look. XML allows web pages to specify data and what they are, allowing browsers not just to read pages, but to process data referred to in the pages by machine readable tags, or ‘metadata’. Using XML, one could, for example, state that a page is a scientific paper, and provide information such as author,

address and keywords. Tags can also represent fields in a database, allowing browsers to interface directly with datasets on the web.

“It would be possible to label a page as being about, say, the Viking missions to Mars, and have specific metadata attached to images that could identify them as being linked to the names of the features they depict,” says Robert Miner of Geometry Technologies, a company in St Paul, Minnesota, specializing in web sites for scientific applications.

Some experts are sceptical of any strategy that relies on the entire

web community agreeing formats for tagging information. But in well-organized scientific circles, it should work better. Some disciplines have already drafted their own metadata standards, such as MathML, agreed by the mathematics working group of the World Wide Web Consortium (W3C). At present, mathematical notation is usually represented on web pages using image files, but with MathML it can be described precisely. This would allow researchers to search for pages containing particular symbols. Some software developers are already

developing tools that will generate the metadata automatically.

The humble hypertext link is also set for a facelift. The W3C is developing XLink and XPointer, which will make hyperlinks much more sophisticated. Xlink will let users append their own links to pages on the web, for example, with a single link offering multiple destinations. Unlike today’s hyperlinks, XPointer allows links to point to precise paragraphs or sentences, so search engines will be able to return the precise part of a document that seems relevant.

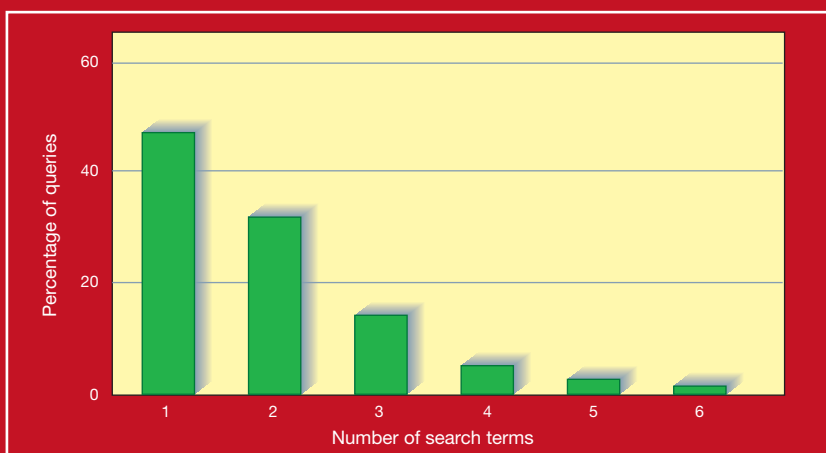
Never trust a human

As search engines become ever more sophisticated, human inadequacies are becoming the biggest limiting factor. A survey by the NEC Research Institute in Princeton, New Jersey, reveals that up to 70 per cent of web users typically type in only one keyword or search term. This is a recipe for obtaining long lists of irrelevant hits. Yet even among the staff of the NEC Research Institute, who should be web-savvy, almost half fail to define their searches more precisely (see figure).

Statistics such as these have convinced

search engine developers that there is little point trying to get users to read the 'search tips' or use the advanced search options provided by most search engines.

To address this problem, more sophisticated engines are now building classifications hidden behind the search interface, which allow them to prompt for more information when presented with ambiguous search terms. Asked to search for 'green', for example, the engine might ask whether the user wants to search for green beans, green parties, and so on.



Hague, has taken this approach further by developing algorithms that can analyse documents not just for keywords but for 'concepts'. Taking its guide from thesauri such as the US National Library of Medicine's Unified Medical Language System, which recognizes more than 600,000 different biomedical concepts, Collexis's engine looks for defined patterns of terms and analyses their contextual relationships.

Users can paste any text, such as a scientific paper, into a search box. This is automatically analysed to identify its major concepts, creating a profile that is then used to search for similar texts. Collexis also allows users to fine-tune their searches by adjusting the weighting given to each individual concept.

Because the Collexis engine's concept matching does not depend on a document's precise grammatical structure, it can generate concepts for texts in many languages, based on a rough machine translation of a document's keywords. Indeed, the software was developed as part of a European Union-funded project to facilitate international cooperation between health researchers. In keeping with these origins, 20 per cent of Collexis's profits will go to promoting collaboration between medical researchers in developed and developing countries.

Autonomy, a San Francisco-based company, is also using concepts. It generates these using neural networks and pattern-

matching technologies developed at the University of Cambridge. Rather than simply responding to specific queries, however, Autonomy is developing programs that automatically analyse any text in an active window on a user's screen to suggest related web resources. The company has released a free, pared-down version of its software, called Kenjin. As you type, Kenjin continuously proposes web resources in a small window at the bottom of the screen. This is still rudimentary, but gives a pointer to the future.

Google's Brin predicts that in five years the search engine as we know it will no longer exist, or be marginal. In its place will come 'intelligent' programs that search by using their experience of the needs and interests of their users. Rajagopalan agrees: "In future there will be automatic feedback loops based on what search results you have selected in the past in relation to this or that query, or how long you stayed at particular web pages."

Learning from previous search sessions, such programs may, in the manner of Autonomy's program, continuously conduct background searches and suggest web resources — with these being tailored to the interests of individual users. Such automated technologies should also help solve one of the biggest limitations to efficient searching: the fact that most people don't frame their queries to maximize the relevance of search returns (see 'Never trust a human', above).

As search technologies improve, journal publishers and the operators of electronic preprint archives are also taking steps that should make it much easier to search for scientific papers on the web. Crossref, a scheme agreed to last November by leading commercial scientific publishers, will link references in the articles they publish to the source papers in their respective publications (see *Nature* 402, 226; 1999). More than three million articles across thousands of journals will become searchable this year, with half a million being added every year thereafter. And in February, the operators of the world's leading archives of electronic preprints, or 'e-prints', agreed standard formats that should allow scientists to search across all of them simultaneously.

This Santa Fe Convention will allow e-prints to be tagged as 'refereed' or 'unrefereed', along with other information such as author and keywords. Using software that will become available this month, any scientist could, in principle, set up an e-print or refereed website, or a site of conference proceedings, in the knowledge that it would be compatible with this global system.

If these efforts succeed in weaving a seamless web from the scientific literature, researchers should find that the hours spent trawling through pages of irrelevant search returns are consigned to history.

Declan Butler is Nature's European correspondent.

An enhanced version of this feature is available online

Search engines

AltaVista ▶ <http://www.altavista.com>

Ask Jeeves ▶ <http://www.askjeeves.com>

Direct Hit ▶ <http://www.directhit.com>

Excite ▶ <http://www.excite.com>

FAST Search ▶ <http://www.alltheweb.com>

Go (Infoseek) ▶ <http://www.go.com>

Lycos ▶ <http://www.lycos.com>

HotBot ▶ <http://www.hotbot.com>

Inktomi ▶ <http://www.inktomi.com/products/portal/search/tryit.html>

Northern Light ▶ <http://www.northernlight.com>

Advanced search engines

Google ▶ <http://www.google.com>

IBM Clever project ▶ <http://www.almaden.ibm.com/cs/k53/clever.html>

ResearchIndex ▶ <http://www.researchindex.com>

Metasearch engine

SherlockHound ▶ <http://www.sherlockhound.com>

Automated search tools

Autonomy ▶ <http://www.autonomy.com>

Kenjin ▶ <http://www.kenjin.com>

Standards

The Santa Fe Convention ▶ <http://www.dlib.org/dlib/february00/vandesompel-oai/02vandesompel-oai.html>

W3C XML page ▶ <http://www.w3.org/XML/Activity>

MathML ▶ <http://www.w3.org/Math>

XLink ▶ <http://www.w3.org/TR/xlink>

XPointer ▶ <http://www.w3.org/TR/xptr>