'metasearch' engines, such as SherlockHound, 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.

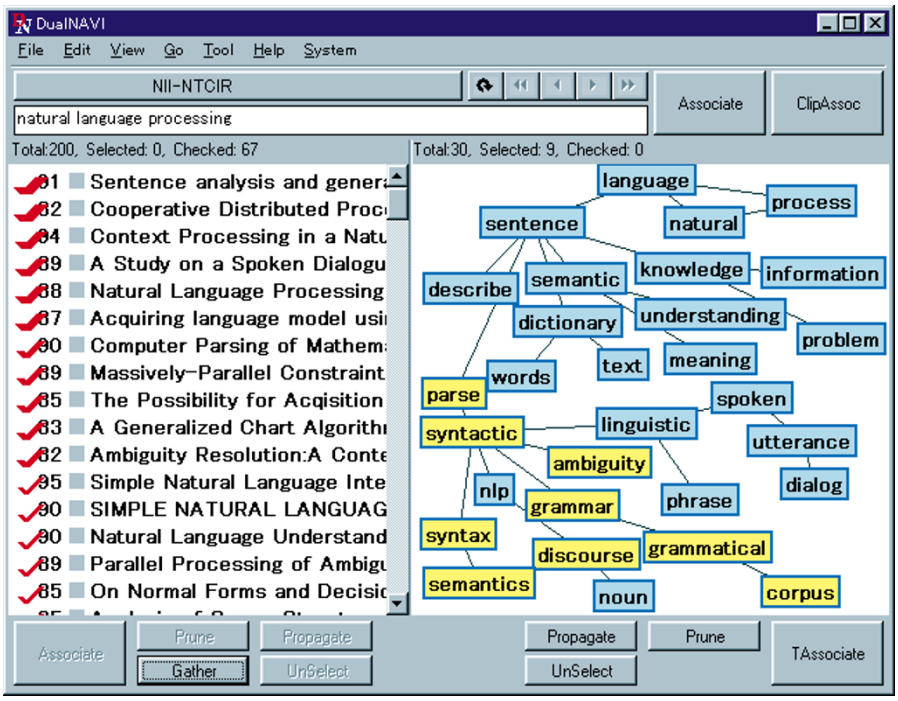
Searches for researchers

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 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."



HITACHI

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

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 Hatoyama, 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.