Searching for Web sites is one of the most common tasks performed on the Web. It is also one of the most frustrating. In fact, the situation has become a notorious symbol of the Web’s growing size and lack of structure, as well as the inadequacy of Web search technologies.

However, a number of Web companies and research organizations are taking a variety of approaches to try to solve this problem.

Traditional search technology (see the sidebar “Traditional Web Search Technology”) is based on users typing in keywords for the information they want to receive. Search services then scan Web pages for those keywords. This approach consistently causes a number of well-known problems.

Users must try to come up with the correct keywords for their search. If the keywords are too general or have multiple meanings, users may receive too many results or too many irrelevant results to find the information they want. For example, a search for “history of rock” could yield results related to popular music, geology, or history classes at a university. Meanwhile, the wrong choice of keywords may lead to useless results or no results at all.

“In general, what’s happening is that Boolean search techniques from the 60s and 70s are running out of gas,” said Kevin Werbach, managing editor of *Release 1.0*, a newsletter on emerging communications and computing technologies.

There are technical, commercial incentives for improving search technology.

Search services are a key to attracting users because they are an important reason people use portals in the first place. Moreover, about 71 percent of Internet users utilize search services to find Web sites, according to Nielsen Media Research, a company that measures computer and Internet usage (as well as television audience levels).

Currently, though, said Werbach, “For consumers, most of the search engines are pretty comparable.” Users thus frequently choose a portal for reasons other than its search service.

However, a company might attract more users to its portal if it could offer an improved search technology. Researchers are thus looking at a variety of new technologies and techniques.

**CHALLENGES**

The sheer size of the Web is a challenge to improving search technology. There are more than 350 million Web pages, and AltaVista contains only about 140 million of them, one of the largest totals of any search service.

Meanwhile, the Web is constantly changing, with new URLs added and old pages discarded. NEC’s Research Institute estimated that in 1998, more than 5 percent of search results in one prominent search service were invalid or “dead” links.

**NEW APPROACHES**

Researchers are taking a variety of approaches to improving Web search technology. For example, some search services are making their Web indexes bigger, in an effort to make their results more comprehensive.

**Human annotation**

The human-annotation approach bases search results on the behavior of and the results obtained by previous Web searchers, rather than just on keywords. Proponents say the results of previous searches, as well as Webmasters’ decisions about which pages their sites should link to, better indicate which sites will satisfy new searches. They also say this technique reduces the ability of a Web site to use keywords to manipulate search services.

Search services are a key to attracting users because they are an important reason people use portals in the first place. Moreover, about 71 percent of Internet users utilize search services to find Web sites, according to Nielsen Media Research, a company that measures computer and Internet usage (as well as television audience levels).

Currently, though, said Werbach, “For consumers, most of the search engines are pretty comparable.” Users thus frequently choose a portal for reasons other than its search service.

However, a company might attract more users to its portal if it could offer an improved search technology. Researchers are thus looking at a variety of new technologies and techniques.

**CHALLENGES**

The sheer size of the Web is a challenge to improving search technology. There are more than 350 million Web pages, and AltaVista contains only about 140 million of them, one of the largest totals of any search service.

Meanwhile, the Web is constantly changing, with new URLs added and old pages discarded. NEC’s Research Institute estimated that in 1998, more than 5 percent of search results in one prominent search service were invalid or “dead” links.

**NEW APPROACHES**

Researchers are taking a variety of approaches to improving Web search technology. For example, some search services are making their Web indexes bigger, in an effort to make their results more comprehensive.

**Human annotation**

The human-annotation approach bases search results on the behavior of and the results obtained by previous Web searchers, rather than just on keywords. Proponents say the results of previous searches, as well as Webmasters’ decisions about which pages their sites should link to, better indicate which sites will satisfy new searches. They also say this technique reduces the ability of a Web site to use keywords to manipulate search services.

**CHALLENGES**

The sheer size of the Web is a challenge to improving search technology. There are more than 350 million Web pages, and AltaVista contains only about 140 million of them, one of the largest totals of any search service.

Meanwhile, the Web is constantly changing, with new URLs added and old pages discarded. NEC’s Research Institute estimated that in 1998, more than 5 percent of search results in one prominent search service were invalid or “dead” links.

**NEW APPROACHES**

Researchers are taking a variety of approaches to improving Web search technology. For example, some search services are making their Web indexes bigger, in an effort to make their results more comprehensive.

**Human annotation**

The human-annotation approach bases search results on the behavior of and the results obtained by previous Web searchers, rather than just on keywords. Proponents say the results of previous searches, as well as Webmasters’ decisions about which pages their sites should link to, better indicate which sites will satisfy new searches. They also say this technique reduces the ability of a Web site to use keywords to manipulate search services.

However, *Release 1.0*’s Werbach and ...
Industry Trends

According to Media Metrix, a company that provides Internet and digital-media measurement services, the Web-search sites with the greatest number of unique visitors as of May 1999 were Yahoo, Infoseek, Excite, Lycos, and AltaVista.
The company credits its search service's performance in part to fast indexing algorithms; large arrays of off-the-shelf servers, storage systems, and interconnects; and software that efficiently utilizes server capabilities.

Fast Search & Transfer says this scalable architecture will be able to effectively handle a growing number of search queries and search an increasingly large index of Web pages.

The search service's index includes 80 million pages and is slated to grow to 200 million in the near future, which would make it one of the biggest in use.

Filtering query responses

Search services such as iAtlas and Northern Light use filtering technology. Filtering narrows the scope of queries to yield results that are more relevant. When submitting a keyword search, users can fill out electronic forms to specify that, for example, they want only information relating to certain industries or certain geographic locations.

"The roles and goals of the user have to be taken into account. This takes you down the road to context sensitivity, which is crucial," said Curt Monash, CEO and cofounder of Elucidate Technologies, a software company working on a variety of products, including some that are search related. "That gives you a chance for accurate searches. The results can be a near-perfect search."

However, if users filter their searches too broadly, they could screen out potentially useful results.

Natural language

Search services are beginning to work with natural-language queries, designed to make them easier to use. For example, with Ask Jeeves, instead of typing in one or more keywords, users who are considering selling their automobile could type, "How much is my used car worth?" The service would then refer them to a site that provides the market value of used cars.

Natural-language engines analyze a query's grammatical structure for meaning and then use the analysis to conduct keyword searches.

"There hasn't been great success in using natural language," said John Lafferty, associate professor at Carnegie Mellon University's Computer Science Department and Language Technologies Institute. He said natural-language search technology's ability to effectively parse queries is limited because the algorithms are still quite immature.

Elucidate's Monash, on the other hand, said "I think natural-language patterns, à la Ask Jeeves, are clearly productive..."

Continued on page 11

Traditional Web Search Technology

In 1990, researchers at McGill University in Montreal developed Archie, the first Internet search engine. Archie searches the files of Internet FTP servers. Two other early engines search gopher servers: Veronica, developed in 1992 at the University of Nevada; and Jughead, developed in 1993 at the University of Utah.

SEARCH ENGINES AND DIRECTORIES

Search services such as iAtlas and Northern Light use filtering technology. Filtering narrows the scope of queries to yield results that are more relevant. When submitting a keyword search, users can fill out electronic forms to specify that, for example, they want only information relating to certain industries or certain geographic locations.

"The roles and goals of the user have to be taken into account. This takes you down the road to context sensitivity, which is crucial," said Curt Monash, CEO and cofounder of Elucidate Technologies, a software company working on a variety of products, including some that are search related. "That gives you a chance for accurate searches. The results can be a near-perfect search."

However, if users filter their searches too broadly, they could screen out potentially useful results.

Natural language

Search services are beginning to work with natural-language queries, designed to make them easier to use. For example, with Ask Jeeves, instead of typing in one or more keywords, users who are considering selling their automobile could type, "How much is my used car worth?" The service would then refer them to a site that provides the market value of used cars.

Natural-language engines analyze a query's grammatical structure for meaning and then use the analysis to conduct keyword searches.

"There hasn't been great success in using natural language," said John Lafferty, associate professor at Carnegie Mellon University's Computer Science Department and Language Technologies Institute. He said natural-language search technology's ability to effectively parse queries is limited because the algorithms are still quite immature.

Elucidate's Monash, on the other hand, said "I think natural-language patterns, à la Ask Jeeves, are clearly productive..."

Continued on page 11

Traditional Web Search Technology

In 1990, researchers at McGill University in Montreal developed Archie, the first Internet search engine. Archie searches the files of Internet FTP servers. Two other early engines search gopher servers: Veronica, developed in 1992 at the University of Nevada; and Jughead, developed in 1993 at the University of Utah.

SEARCH ENGINES AND DIRECTORIES

Current search services can be divided into search engines and directories.

Search engines

Search engines (such as AltaVista and HotBot) traditionally consist of three components: the crawler, the index, and the search software. Crawlers, also called spiders, are programs that automatically scan various Web sites and create indexes of URLs, keywords, links, and text. Crawlers also follow the links on a site to find other relevant pages. They return to sites periodically to look for changes.

When a user submits a search query, the engine's software goes through the index to find Web pages with keyword matches and ranks the pages in terms of relevance.

Directories

Instead of working with indexes, directories (such as LookSmart and Yahoo) work with descriptions of Web pages submitted by either Webmasters or editors who have reviewed the pages. Directories respond to queries by searching through these descriptions. Some search engines, such as Microsoft's MSN and Netscape Search, take a hybrid approach by also using directories.

Because they don't use crawlers, directories don't automatically find changes in Web pages. But proponents say that human-generated descriptions can produce more relevant responses to some search queries.

However, said Kevin Werbach, managing editor of Release 1.0, a newsletter on emerging communications and computing technologies, it's difficult for manually produced directories to keep up with the rapid growth of the Internet. He said, "It's really important to automate some of the processes that they do."

KEYWORD SEARCHES

Traditionally, a user enters a keyword (or keywords along with Boolean modifiers, such as "and," "or," "not") into a search engine, which then scans indexed Web pages for the keywords. To determine in which order to display pages to the user, the engine uses an algorithm to rank sites that contain the keyword.

For example, the engine may count the number of times the keyword appears on a page. The engine also may look for keywords in metatags. A metatag is an HTML tag that provides information about a Web page. Unlike most HTML tags, metatags don't affect a document's appearance. Instead, they include such information as a Web page's contents and some relevant keywords.

In the past, users have subverted keyword-based techniques by stuffing their Web pages with keywords or loading their metatags with keywords that don't relate to their site's content. However, search services have taken steps to counteract this. For example, some don't scan metatags anymore, and some lower the relevance rankings of sites that use keywords unrelated to their content.
be ready to scrap their expensive infrastructure, particularly when they still have network capacity.

The transition may also be complicated because there's more competition between 2G standards in North America than in Japan or Europe, Potiency said.

Several North American groups, all working under the auspices of the Telecommunications Industry Association (TIA), have proposed competing standards. The CDMA Development Group and Qualcomm favor cdma2000/Wideband cdmaOne. The Universal Wireless Communications Consortium (UWCC) is working on UWC-136 to provide an upgrade path for TDMA- and GSM-based carriers.

However, progress is being made toward compatibility, noted GTE’s Levitan, who said the big issue is providing the investment many operators have made in 2G technologies.

Levitan expects that North America will primarily migrate to W-CDMA and cdma2000, which will probably be the most widely used standards worldwide, but will also make some use of TD-CDMA.

OBSTACLES IN 3G’S PATH

It appears that worldwide implementation of IMT-2000 technology won’t be easy, fast, or inexpensive.

Although the ITU has worked hard to make proposed 3G standards less divergent, more work is needed.

Also, in some cases, new 3G technologies will require carriers to obtain new radio frequencies and undertake expensive upgrades to software and hardware, including transmitting and receiving equipment, noted Dataquest’s Bruederle.

On the other hand, increasingly cheaper chipsets and increased competition will help keep prices down for 3G consumer products.


However, the ITU subsequently determined that the 230-MHz allocation would not provide enough capacity for the projected level of usage or provide enough frequencies in the same range that could be used throughout the world to permit seamless global roaming. The ITU thus has been recalculating spectrum needs and considering other ways to provide sufficient capacity.

If any killer app will drive 3G wireless technology, said Bruederle, it will be the Internet. That’s because 3G will enhance important existing and future Internet capabilities, including Web access, voice applications, and e-commerce.

Wireless expert Seybold doesn’t see such a bright future. He said that there is not a significant demand for 3G technology’s capabilities now and that technology vendors can no longer push carriers into adopting new technologies without sound business reasons.

Even in Japan, he said, carriers are rolling out 3G technology only because they’re running out of 2G voice capacity, not because consumers are demanding new capabilities.

Bruederle said the key to 3G’s future success will be wireless technology’s progression from a voice-centric to a data-centric medium. According to many industry analysts, the more that users want to use mobile devices to access data, the more they will need a faster wireless technology.

GTE’s Levitan said that 3G technology will be driven by market demand for the ability to use a single wireless device worldwide. However, he said, “It’ll be an evolution, not just a flip of the switch. As systems evolve, you’ll see a lot of the newer technology leveraging off previous generations.”

Directories

Some search sites are considered directories (as seen in the sidebar “Traditional Web Search Technology” on page 7). When Netscape Communications acquired the NewHoo Community Directory Project recently, it also acquired the Open Directory project (http://dmoz.org). The project uses volunteer experts in various subject areas to produce and maintain comprehensive directories of Web sites that contain information in their areas of expertise. The Open Directory has been licensed to a number of search services, including HotBot and Lycos.

Proponents claim the use of volunteers will help the project scale as the Internet grows. However, some question whether volunteers will do as effective a job as experts paid and trained by search services to work on directories.

Even the military wants better search technology. The US Defense Advanced Research Projects Agency’s (DARPA’s) Space and Naval Warfare Systems Center has invested $2 million in a classified search-technology project at Mississippi State University.

Meanwhile, Release 1.0’s Werbach said, commercial search services are likely to move into niche areas. Niche services would focus only on certain topics and/or index only certain sets of documents, he said.

For Web portals, the future of searching will likely entail incorporating not one but several of the approaches now spearheaded by the vanguard of new search services. “People think search is a monolithic activity,” said Werbach. “But it isn’t. The reality is that sometimes you will want to use different techniques for different results.”

Ilan Greenberg is a freelance technology writer based in San Francisco. Contact him at ilang@ix.netcom.com.

David Clark is a freelance technology writer based in Torrance, California. Contact him at dwclark@earthlink.net.

Lee Garber is Computer’s news editor. Contact him at l.garber@computer.org.