SEARCH AND SEARCHABILITY
By Kevin Werbach

‘Twas the best of searching times, ‘twas the worst of searching times....

The Internet has collected massive quantities of information and opened it up to the general public. Thousands of concurrent users can perform full-text searches across libraries of more than 100 million documents... all in a matter of seconds. Sites built around search engines are among the most popular destinations on the Web.

Yet how many of those millions of page views are the “wrong” documents users must sift through to find what they were really looking for? Internet search engines often produce long lists of pages wholly unrelated to the desired topic. Even then, they index only a minority of the Web. Entire businesses do nothing but manipulate search services to rank their clients' sites higher. Search engines today are clearly more powerful than in the past, but are they really more useful?

Better search engines will require more than bigger indexes and faster processing of Boolean queries. Search is fundamentally linked to meaning, that deepest and most slippery of concepts. To uncover meaning, search engines must go beyond document text to glean knowledge about structure and context. In this issue of Release 1.0, we explore recent developments and consider the future of search technology.

Search engines narrow vast rivers of information into usable channels. There is no greater information resource than the Web, and it is through the Web that search services have achieved mainstream recognition.

A search engine maps between two items: a query and a response. The query is typically a few keywords and the response is usually a list of documents, but both are proxies for fuzzier concepts. Keywords and documents are the denotation, but what users really want resides in penumbras of connotation. When I ask for information about world leaders and cigars, I may be thinking about Fidel Castro or Bill Clinton. A query about Gates and Windows may mean:

THE FORUM IS FILLING UP FAST!!!
concern Microsoft or home remodeling. The documents in the response list are also surrounded by clouds of associations: An editorial advocating marijuana legalization carries different implications depending on whether it runs in the New York Times or High Times.

Keyword matching elides these distinctions. Keywords are woefully inadequate approximations of the semantic structures of the mind, but they are the only explicit information traditional search engines have to work with. The new companies we discuss below offer a variety of solutions, including placement auctions (GoTo.com), popularity tracking (Direct Hit), link structure analysis (Google and IBM's CLEVER project), natural language processing (Ask Jeeves) and lexical object creation (Lexeme).

**Portal combat**

If portals are the desktops of the Web, search services are the file systems. For many users, search services take the place of structure beyond a few regularly visited sites. Consequently, the evolution of search technology will influence the development of the Web on both technological and economic levels (but see page 14 for Esther Dyson's analysis of why search services will become less relevant over time).

Portals are always looking for ways to distinguish themselves from competitors and gain a greater share of advertising or other revenue. In the past two years, the major services have built out content beyond their core search franchises, spending money and time adding community features, e-mail, e-commerce, e-t cetera. The portals have had to work hard enough just to scale the capacity of their search engines to meet demand. Consequently, there have been few significant improvements in search quality.

The pendulum is gradually swinging the other way. Danny Sullivan, editor of Search Engine Watch, points out that AltaVista has added features such as spell-checking, language translation and a query refinement tool. Infoseek has incorporated new technology to improve result relevance, although its index remains relatively small. With every portal offering a similar suite of destination features, search once again becomes a point of differentiation. Users may stay on portals for the added features, but they still come in the door primarily for searching. If competing sites can offer better search results, users may switch and find other places to get their free e-mail or stock quotes.

**I STILL HAVEN'T FOUND WHAT I'M LOOKING FOR**

Why is it so hard to find anything on the Internet?

First off, the Web is big. Lee Giles and Steve Lawrence of NEC Research estimated early in 1998 that there were at least 320 million searchable Web pages, and the number has probably doubled since then. That doesn't include documents behind firewalls, dynamically generated or otherwise inaccessible to search engines, which may still contain information users are looking for. Today the most comprehensive search service indexes roughly 140 million pages, a fraction of the total.
Quantity isn’t the only problem. The Web changes constantly, with thousands of pages being added and disappearing every day. Even if you could fit the Web on a disk array, you would simply have a snapshot at a particular point in time (something not without value, and the aim of Brewster Kahle’s Internet Archive. See **Release 1.0, 5-98**). In fact, it’s worse than that. The only way to pull together a large quantity of Web pages is to use robots or spiders to crawl in search of new sites (see page 7). By the time you’re done taking your snapshot the Web will have moved, like those old-fashioned photos where people have walked through the background as the plate was being exposed. Some of the information is already out of date the moment it becomes searchable.

The demand on search engines is increasing as well. The World Wide Web Worm, one of the first search services, handled 1,500 queries per day in April 1994; AltaVista now handles over 40 million queries per day. In addition, the community of searchers is broadening with the Internet itself. Users may look for information in many different languages, for example.

**No structure no find**

The Net operates on a least-common-denominator basis. Web browsers are effective “thin clients” because they do not require data to be formatted in a specific manner. Client-server systems rely on a tight coupling between the two ends of the connection; data on the server side is formatted in a specific manner that the client understands. With IP and HTML, the Web browser need only understand a simple set of markup commands. As a result, content creators don’t have to worry about the specific characteristics of every client system. The tradeoff is that the browser generally sees information on the server as an undifferentiated mass of text, graphics and other objects. Only a simple set of formatting tags suggest internal structure, with no consistency from site to site.

Commercial database providers such as Lexis-Nexis and Dow Jones spend significant time and effort organizing content that goes into their systems so that it can be retrieved more easily. Even if tools for such markup were available, many Websites wouldn’t use them. The Internet Engineering Task Force and World Wide Web Consortium (W3C) can promulgate standards, but beyond the basics needed for connectivity it’s up to users whether to follow them. (A good example is HTML 3.0, which added many new features but was never fully implemented by browser vendors. In HTML 3.2, W3C eliminated most of the proposed additions and simply tried to codify existing practice.)

Efforts are underway to give Web content — and by extension the Web itself — greater structure. The extensible markup language (XML) transcends the limited scope of HTML and allows information to be organized under any schema embodied in a document type definition (see **Release 1.0, 5-98 and 9-98**). The W3C’s Resource Definition Format (RDF) standard provides an XML-based structure for categorizing and rating Web content (See **Release 1.0, 5-98**).

XML and RDF will be important for specialized communities and for content aggregators. For example, a site could provide links and search capability limited to kid-safe pages, or to documents accessible to the sight-impaired (or, we must acknowledge, to content that a government has ap-
proved). Chemical engineers could tag their research papers so that colleagues could search using categories familiar to that discipline. Italian speakers could locate only materials available in their native language. XML can also open up the proprietary data structures of commercial databases to browser-based access, assuming the business models can be worked out.

XML is not a panacea. Someone still has to tag all that content, and many sites simply won't go to the trouble. And structure alone is not sufficient, because people search for different things. Without agreement on terms and categories, people will still retrieve results that fail to correspond to their interests.

Dealing with information overload

Editors have been the traditional way of dealing with information overload. Good human editors guide users through the thicket of information to find nuggets of knowledge, and also create them by ordering the world and the information in it. That's how newspapers and magazines work: They channel a vast array of data into a usable package.

The Web also has edited directories of information, Yahoo! being the most prominent. But Yahoo! captures only a tiny fraction of the Net, and Yahoo!’s categories may not match up with your own. The Mining Co. (see Release 1.0, 7/8-98) offers more detailed information, using editors not only to categorize information but also to put it in context and recommend the best sites for certain subjects. Yet this results in even narrower coverage: The Mining Co. has about 600 topic headings compared to roughly 20,000 on Yahoo! Northern Light attempts a hybrid approach, dividing search results on the fly into topic categories, but that doesn’t necessarily get users to the documents they are looking for.

The problem is that human editors can’t keep up with the Web. People (and their salaries) simply aren't sufficiently scalable to organize the entire Web through any explicit process. (As we describe below, however, implicit preferences can be harnessed to improve search. See pages 11-20.) There is also an inevitable tradeoff between depth and breadth. Companies such as Content Advisor (see Release 1.0, 5-98) can categorize pages to filter material deemed inappropriate for children or the workplace, but the subject categories are too broad to be of much use for searching.

Information overload has secondary consequences. One is that search services, limited as they are, have become increasingly important and valuable resources. In the proprietary database world, value lies in the information itself, or perhaps in the consistent manner that information has been tagged for retrieval. On the Web, information is free, but the value is in the information-gathering and search technology necessary to sift through that mass of data. Search services (or their VCs) have recognized the value of this asset and have marketed themselves into portal sites with dizzying market capitalization.

A further consequence is that placement in search results on major portals becomes a valuable commodity for online businesses. People may never find out about your site if it doesn't show up when they query a search engine. One way to gain visibility on a search site is to buy an advertising ban-
ner linked to a keyword. That costs money, though, and doesn't place the site directly within the search results listing. Consequently, many people try to trick the search services into ranking their pages highly in response to common queries (see page 6).

In addition to the breadth/depth dilemma, there are other tradeoffs any search site must contend with. Even if you can design an effective algorithm to categorize pages, it has to run quickly enough to answer large numbers of queries. AltaVista, for example, reports that at peak times its service processes over 1,500 queries per second.

THE ECONOMICS OF SEARCH

Search services are driven by more than technical considerations; they must also consider business realities. Because search is now big business, companies take financial considerations into account when designing their services.

Specifically, commercial Web search sites don't necessarily want to take you directly to the site you're looking for. Unlike proprietary databases, the search service has no control over the underlying content. Instead of charging for the content, it must extract value from the search process itself. One way to do so would be to charge a subscription or usage fee for searching the Web, but that seems unrealistic given the tradition of free Web-based services and the number of free competing search sites. (For analysis of the economics of online content, see Release 1.0, 12-94, 1-96.) The primary alternative is to charge for advertising, which is how all the current search services make the bulk of their money.

For an advertising-driven site, the more pages a user views the better. If the search engine takes you immediately to your desired site, you won't stay long. But if you have to click through several pages and perhaps enter multiple queries, you'll generate more revenue for the search service even as you have a less-satisfying experience. The search services compete with one another, so they need to offer good enough results to keep users around, but that's all.

As search-oriented sites have become portals, moreover, they have gained other goals beyond slow but accurate search results. Where originally search services merely returned a list of links with a banner ad, now they provide a range of other options dynamically generated from the search query. For example, the search service might offer a link (advertiser-supported) to books available through Barnesandnoble.com on the requested topic, or a football-related query might return a link to the portal's own ad-filled football section in addition to the external sites in the search results. This is in addition to the topic-specific ad banner, which the portal can sell at a higher cost per thousand (CPM) because it is more likely to be relevant to the user. As the portals become more and more inclusive, their incentives shift ever further away from offering good search results.

The recent launch of Disney's Go.com shows this trend to its full extent. Unlike most portals, which grew out of narrower services, Go emerged fully formed from the head of Michael Eisner. Go provides a consistent inter-
face of links to Disney-owned content such as ESPN sports information, ABC news and Disney children's content, along with a Web-wide search service powered by Infoseek. According to Barak Berkowitz, senior vp of Infoseek and general manager of the Go network, this offers a better experience for users because navigation commands and content are always in consistent locations. However, the more portals become their own self-contained worlds, the less the Web differs from the closed models of traditional media such as magazines.

Because placement in search results is so valuable, the search services also have an incentive to sell their listings to the highest bidder. To users, search engines are black boxes, so it can be almost impossible to tell if a site received a favorable ranking because it paid for it. In 1996, Open Text launched a search service that sold preferred placement in search results. The Open Text Index generated significant controversy because many people felt selling placement compromised editorial integrity. The site has since been folded into a business-oriented product without the preferred placement option. The idea of selling premium placement in search results has since been resurrected by GoTo.com (see page 7).

Search engine persuasion

An entire sub-industry has emerged to improve sites' placement in search results. Companies promise (often in spammed promotional e-mails) that they can deliver higher rankings, greater traffic and greater revenues. Some techniques are benign, such as making sure the site has been submitted for inclusion in all the major services and that page titles provide good information about their content. In other cases, though, people deliberately try to fool the search services. Massimo Marchiori, a researcher at the World Wide Web Consortium (W3C), calls such manipulations search engine persuasion (SEP). Search engines give each page a score in response to a particular query and then return results in ranked order. Sprinkling repetitive invisible keywords on a page is the simplest way to enhance those scores. Sites that want more traffic can fill their pages with common keywords even if they do not correspond in any way to the subject area of the page. This is particularly common for pornography sites.

Search services try to prevent sites from artificially obtaining high rankings. If the quality of search results degrades too far as a result of such manipulations, people will be less likely to use search services. More to the point, search services don't get any revenue when sites artificially boost their rankings; they would prefer that companies pay them for banner ads tied to keywords instead. A final, less obvious negative consequence is what Marchiori calls flattening. If several sites all receive the highest possible score, the search engine has no good way to order them. Thus, when many sites use artificial techniques to improve their ranking, the rankings themselves become less useful.

The major search services keep their exact ranking algorithms secret, and also tweak the algorithms frequently to foil manipulators. The manipulators, of course, keep trying to reverse engineer the search algorithms by analyzing sites that score highly in response to common queries. It's a never-ending arms race, much like the battles over computer viruses.
As long as search result placement is a valuable commodity (that is to say, as long as search services are widely used), people will do anything they can to fool the search services. Fortunately, the new search technologies we describe below have the side benefit of making it harder to do so. One reason it's relatively easy to fool search services is that they rely almost exclusively on the page text, which the owner can control. Other methods, such as popularity tracking (see Direct Hit, page 13) or link structure analysis (see Google and CLEVER, pages 16, 17) determine relevance from the activities of others. It's possible to create artificial links and traffic, but this is more difficult and less effective than manipulating the page itself.

How a search engine works

Because the Web doesn't exist in any finite location, a search engine must do more than simply scan through a defined corpus. Web search engines perform three primary functions: acquiring information about pages, organizing it and responding to queries.

The first step for any search engine is to create an index. There is no central map of the Web, so search engines must find all the pages to search themselves. They do so by operating distributed networks of crawlers or spiders that follow links through the Web to identify new pages. When a crawler finds a new or updated page, the search engine processes the page to extract the useful information (such as the unique words on the page and their frequency). Most search engines don't actually store the full text of pages themselves (Google is an exception) because of the disk space and other overhead required.

This information collected by the crawlers is assembled into an index, optimized for rapid queries. When a user types in a request, the search engine extracts the most relevant documents from the index through a scoring algorithm. The completeness of the index, and the sophistication of the algorithm, are what distinguish one search service from another.

Directories such as Yahoo! are not search engines, because they rely on human ontologists to categorize information. Directories allow users to follow structured taxonomies to locate information on a particular topic. As IBM's CLEVER project demonstrates, however, the demarcation between search services and directories is not as sharp as it may seem (see page 18). As search engines improve, they will obviate some of the need for human-assembled directories, but the directories can add more detailed descriptions in response.

GoTo.com: If you can't beat 'em, join 'em

An alternative solution to search engine persuasion is to put listings up for sale. After all, the yellow pages gives priority (through bold-facing, color or display ads) to companies willing to pay for more prominent listings. No one seems to mind that when you turn to a given category, the first thing you see is the company that paid the most. In fact, the size of the ad itself provides some relevance feedback. Larger compa-
nies can afford more prominent ads, but they may not need them if they have strong brands that customers will look for directly. Smaller companies will pay for bigger ads if they strongly want your business... or if they have backers who believe in them and will fund such marketing.

GoTo.com, based in Pasadena, CA, wants to bring this yellow-pages model to the Web. Founded in late 1997 by Bill Gross' idealab!, with additional funding from Draper Fisher Jurvetson, GoTo.com offers traditional search results generated by Inktomi and supplemented with category-specific rankings assembled by human editors. What makes the site unique, however, is that it auctions top listings to the highest bidder. The amount an advertiser pays is disclosed right on the search results page, and GoTo charges the advertiser only when a user clicks through to a site. Ceo Jeffrey Brewer joined from CitySearch in February 1998, and the site launched in June. The company now has over 70 employees.

Brewer asserts that GoTo's approach benefits both consumers and advertisers. Advertisers can target searchers directly, rather than rely on banners, and can calibrate their spending by bidding only what they are willing to pay. Over 5,000 have signed up so far. Consumers see which advertisers will pay the most for their business. In a user survey conducted in October by the NPD Group, GoTo ranked second behind Ask Jeeves (see page 21) in overall effectiveness and tied for first in search success.

GoTo is a free market, and markets are complex systems that evolve in unpredictable ways (just ask Long-Term Capital Management!). On the other hand, this is a market with fairly detailed rules. For example, GoTo limits advertisers to search terms that actually relate to their site, to prevent companies from tricking users into visiting. An advertiser recently accused GoTo of favoring another idealab! company, but Brewer convincingly denies that anyone receives preferential treatment. He acknowledges that GoTo has to make the market as transparent and efficient as possible, so that advertisers understand exactly where they stand vis-a-vis competitors.

Biological evolution not only produces the survival of the fittest, but also generates increasing diversity and new species. If GoTo were to follow a similar pattern, advertisers over time would purchase more-specific category listings, targeting their marketing to narrower consumer interests. Users would then get smaller numbers of results that more closely matched their interests. Since GoTo launched, the percentage of clicks on revenue-generating listings has increased from 1 percent to 10 percent, and the average price per click has increased from 1 cent to 10 cents.

GoTo's model is most effective when many advertisers compete. On average, 20 percent of users click on the first site listed, 9 percent on the second and five percent on the third, meaning that top placement makes a real difference. Bids have reached as high as $2.80 per click in the Web hosting category, with more than 80 competing advertisers. Paid listings are not limited to e-commerce categories. A search on "cancer," for example, brought back sponsored links to a drug company, community sites, information publishers and a topic guide from The Mining Co. (see page 4).

In addition to its own site, GoTo partners with other sites that want to add search functionality. Brewer notes that because GoTo only does
search, it does not compete with community-oriented sites that are potential partners. Because GoTo gets paid for search listings, it can give its partners all the banner advertising revenues and still make money.

So far, GoTo has avoided the controversy that sank the Open Text Index. Perhaps the explanation is that the Web has gotten far more commercial since 1996, and people have become accustomed to the notion of search services as advertising vehicles. GoTo has gone out of its way to be open about its methodology and why it believes this system benefits users as well as advertisers. With proliferation of Web content and the rise of search engine persuasion, customers may be more willing to accept an alternative approach.

What doesn't get indexed

The economics of search depend on more than search engines themselves. The search services don't index a great deal of the content accessible through the Web. Some of the content (such as dynamically generated pages or non-HTML files) is hard to index directly, and in many cases sites specifically direct search engine crawlers not to index them. The de facto robot exclusion standard provides a mechanism for sites to declare themselves off limits to search engines. Most news-oriented sites do so.

The problem is that these sites can't capture the full benefits from search-service traffic. Ad-supported news sites want users to come in through their main home page, rather than zip directly to a specific story. The problem is even more acute for proprietary databases that make revenue from subscription fees. Because there is no universal system for micropayments and copyright protection on the Web, such sites don't want users to click to their precious content. If and when payment and protection technology becomes commonplace, these providers will want to release their links to the search services.

BUILDING A BETTER MOUSE TRAP

Computer search technology has evolved over a period of decades. Most pre-Internet work in information retrieval involved central databases that could be structured and searched directly. Companies such as Lexis-Nexis developed proprietary, vertically integrated systems down to the terminals, and sold them to professionals in fields such as law and finance that could justify the relatively high expense.

As search technology, processing power and computer networking advanced, it became possible to search many different content databases, and to query the full text of documents rather than just the headers. The problem, in the age of client-server, was that every database had its own query language and client software (for a detailed discussion of this problem and some responses, see Release 1.0, 4-91). The Web solved many of those problems with its universal browser client, but at the expense of the sophisticated index management and query structures the proprietary systems offered.

Most Internet search services began as university or corporate research
projects. Lycos was a project of Carnegie Mellon university. AltaVista was designed to show off the performance of Digital's Alpha processor chip that powered the service. Inktomi grew out of distributed computing technology developed by Berkeley computer science professors. Excite was developed by Stanford students. Venture capitalists quickly saw the potential of search and navigation sites as commercial ventures, however, especially after Yahoo!'s IPO. As the search services shifted from technology projects to the corporate world, they began to focus less on the technology of searching itself.

Inktomi: e pluribus unum

Inktomi was the first technology provider to target Internet search services. Its technology, developed by Eric Brewer at the University of California at Berkeley, allows tasks to be efficiently broken up among many computers. A room full of workstations can therefore provide better performance than a more expensive supercomputer at a computationally intensive task like crawling the Web and answering search queries. Inktomi doesn't offer search services to end users; it is a technology wholesaler to companies such as HotBot and Yahoo! Inktomi has moved into other segments that can benefit from its approach, including caching (see Release 1.0, 6-98) and comparison shopping engines.

There are interesting parallels between Inktomi's role and another pioneering technology company, Thinking Machines. Several years ago, we described how Dow Jones had purchased a $2.5-million Thinking Machines massively parallel supercomputer to power its full-text search service (see Release 1.0, 1-88). Thinking Machines wasn't a search company; it was a technology company with a product that had many possible applications. Despite the sophistication of its technology, however, Thinking Machines eventually failed. Inktomi has so far succeeded in part because its distributed technology is not only better but cheaper than the alternatives. Moreover, the market for high-performance search (and the text available to search) is much greater today.

Meta-search engines

Try as they might, none of the existing search services covers more than a fraction of the Web. Because each search service crawls the Web independently, though, their databases do not completely overlap. Using more than one search service therefore provides greater coverage. Meta-search engines such as Metacrawler, Savvy Search and Dogpile automatically submit queries to several search services and aggregate the results with duplicates removed. NEC's Lee Giles and Steve Lawrence concluded that searching five engines at once returned roughly three times as many documents as a search on a single engine.

Some meta-search engines are accessible directly from the desktop, including the Sherlock search tool in Apple's new System 8.5 for the Macintosh. Sherlock integrates remote Web searching with search functions on the local computer. A tabbed interface allows users to find terms in local file names, the full text of local files (if the user has indexed his or her hard drive beforehand) or Web pages using a combination of several major search services.
Meta-search engines may provide better coverage, but they don't necessarily improve results. In fact, by increasing the number of sites that come back, they may even make it harder to find the desired information.

Giles and Lawrence have developed a meta-search engine at NEC that seeks to rectify these problems. Inquirus downloads and analyzes the full pages suggested by the initial search engines, rather than simply returning a combined list of URLs. Downloading the full pages provides two significant benefits. First, the system displays the search term in context instead of the title or first few lines of the page. Proprietary databases such as Lexis-Nexis have long offered “keyword in context” views. Studies have shown that this approach gives users a better sense of whether a page is relevant. Second, Inquirus re-ranks the pages based on the proximity of search terms within the document. Pages where the search terms are in the same sentence will likely be more relevant than pages where they are far apart, but without the full text of a page there’s no way to do this kind of analysis for each query.

NEW APPROACHES: FINDING NEEDLES BY ANALYZING THE HAY

The problem with traditional search engines is insufficient information. Just analyzing the text on a page provides only a crude indication of what that page means. Search engines can use algorithms to estimate relevance based on frequency and proximity of search terms in documents, but those algorithms are inherently limited, especially in unstructured data. Even when search engines are able to match up pages with queries, the sheer size of the Web can make the information they return less than useful. Try typing “Microsoft and monopoly” into AltaVista and sifting through the 402,010 pages that come back. The same algorithms might have generated a manageable number of results when searching smaller proprietary databases, but on the Web additional ranking is essential.

Consistent and universal use of labels (XML-based or otherwise) would provide some additional information outside the documents themselves. However, as we described above (see pages 3-4), many content creators won’t do so. Moreover, because of the propensity of sites to mislead search services (see page 6), any schema for categorizing Web pages wouldn’t necessarily work even if widely adopted.

To be reliable, the supplemental information about Web pages must not be generated consciously by the page creator. This sounds counterintuitive, but a Web page author may not necessarily be the best one to describe it honestly. Reasonable minds may disagree about how a certain page should be categorized. It doesn't matter who's right; if the person doing the searching has a different term in mind he or she may never find the desired page. Many times searchers don't form their queries as precisely as they should, because they don't know the syntax of a given service or Boolean searching in general.

Data-mining technologies may help. Search engines have traditionally been considered a form of information retrieval, modeled on libraries. The focus has been on retrieving documents based on their content and classification. Data mining developed as a means to identify hidden associations
in relational databases. In other words, information retrieval looks at explicit information, while data mining extracts implicit information. Fortunately, these approaches are complementary. Companies are now exploiting data-mining techniques to derive additional implicit information from Web pages, in order to enhance the quality of search results.

What people search for

One of the challenges for search services is users. The more information the users puts into the query, the easier it is for the search engine to provide a good result. The basic Boolean interfaces of the major search services allow users to narrow the scope of a search significantly. The average query, however, is only 2.6 words long.

So what do people actually search for? Perhaps it should not be surprising that the most common thing people look for is dirty pictures. After all, sex sells. The following are currently the top 25 search terms used on the MetaCrawler meta-search service:

1. free
2. nude
3. sex
4. mp3
5. pictures
6. download
7. pics
8. new
9. music
10. de*
11. games
12. warez
13. university
14. women
15. girls
16. stories
17. software
18. chat
19. video
20. school
21. world
22. history
23. cheats
24. computer
25. art

* Germans apparently use this query frequently to identify German-language content under the .de top-level domain. This term ranks so high because Germans are probably the largest population of non-English speakers on the Web today.
Direct Hit: Popularity counts

The search process doesn't end when a computer returns a list of relevance-ranked pages. The user must still decide whether any of those pages answer the question. If most users who type in a query for "long underwear" select the ninth page in the list returned, there is reason to believe that that page is most relevant for that query. An engine that tracks user preferences in this manner won't tell you which page is most relevant for a new topic, but it can indicate that a much-selected page deserves a higher ranking than the initial Boolean algorithm suggests.

Direct Hit uses this popularity-based method to enhance search results. The company, located in Wellesley Hills, MA, was founded in April 1998 by Mike Cassidy and Gary Culliss. Culliss frequently searched online databases in his work as a patent lawyer. He realized the Net could automate the process of sharing effective strategies with fellow searchers, harnessing "the efforts and human judgements of the millions of people performing searches every day" to improve results. He hooked up with Cassidy, who had previously founded and sold computer-telephony software vendor Stylus Innovation, and developed a prototype. The Direct Hit technology won the grand prize at MIT's 1998 annual entrepreneurship competition in May. The company has closed two funding rounds from Draper Fisher Jurvetson and Mosaic Venture Partners totaling $3.4 million, and currently has nearly 30 employees.

At first glance Direct Hit's technology looks like collaborative filtering, but there are significant differences. Companies such as Firefly (now owned by Microsoft), LikeMinds (now owned by Andromedia), Net Perceptions and Alexa Internet employ collaborative filtering to make recommendations based on correlated user responses (see Release 1.0, 11-96 on collaborative filtering; Release 1.0, 5-98 on Alexa). However, where collaborative filtering identifies clusters of associations within groups, Direct Hit passively aggregates implicit user relevance judgments around a particular topic. WiseWire (now owned by Lycos) uses content analysis and collaborative filtering to categorize documents by topic. However, the topics themselves are generally human-defined, and users must actively rate pages they view. Direct Hit provides the most popular documents for any topic, rather than continuously monitoring defined subject areas.

Direct Hit partners with existing search services. As users move through those sites, Direct Hit tracks their queries and the pages they select from result lists, capturing behavior but not identities. It associates queries with more-general topics, developing a database of the sites most commonly selected for each topic. Direct Hit also considers factors such as the time searchers spend on a site once they've chosen it and the location of the site on the original results list.

A Direct Hit icon appears above the initial page of results. If the user clicks on the icon the search service displays a new list of the most popular sites for that topic. Direct Hit makes money by splitting the revenue from ad banners when users invoke the service. Its service is currently available on HotBot, and the company has also signed deals with Apple, America Online and two unannounced partners on similar non-exclusive terms.
On beyond search

By Esther Dyson

Long ago, I worked as a securities analyst on Wall Street, following growth companies for investors. One of the stocks I covered was Federal Express (whose chief operating officer was none other than Jim Barksdale). Like most people, I kept files but rarely used them. One day, as I poked into the FedEx file for some reason or other, I found my FedEx bills for some packages I had sent.

Now, whenever I tell that story to people, they laugh. But to a computer (or a less than clueful secretary), it would seem perfectly logical, wouldn't it? Keep that in mind....

Second story, more recent: some friends of mine were looking for an SGML. One of them had done a Web search and had not come up with a useful answer. I fired off an e-mail to Dave Winer, and got an answer back in half an hour, with an e-mail address for the supplier. Yes, I was lucky, but it was an altogether satisfactory experience. He copied another old friend of ours... and a good time was had by all.

Third story: Another friend sent me (and a few hundred other people) an e-mail about "developing an information infrastructure and... services for using people's assessments of online documents for improved navigation, and apply them to Usenet messages." (See Resources under Sasha Chislenko.)

These tales all concern search. As Kevin notes, portals are interested in making search results good enough to please their users, but not so good that customers pass through the portal right away.... There's a fundamental conflict here, and I think it's bad news for the generic portals. Even as Disney is creating the newest brand-name generic portal, Go, hundreds of other perfectly respectable sites focused on, say, medicine or landscape gardening want to become the "medicine portal" or the landscape gardening portal, with information organized in relevant ways rather than in massive alphabetical indexes.

While the search engine companies are focusing on providing better flashlights (query tools and flat alphabetical indexes) to poke around in the dark, a more interesting approach is to build better floodlights — to light up areas of the Net as a whole rather than pinpoint single items in it. Along with those floodlights, they are developing maps, signposts, topographical maps, building directories and other cues to help us pick out what we want — in context.

When you look for a store or restaurant or a place to leave the kids, you consider the neighborhood: Is it expensive? Is it a shopping district or a public park? Are the office buildings old or new? And when you make your choice and walk into a restaurant, you aren't led to your table blindfolded. You look around: Is it crowded? Noisy or subdued? How old are the people? Well-dressed or comfortable? Stiff, or had a few?
When you look at content, you want to know, similarly, what's the neighborhood? Is the site well-visited? Frequently updated? Do a lot of other sites point to this site — a sign that it is authoritative or at least important. Do visitors come from the financial district, or do the just want to send a package? (Sites have different areas and entry points: "Federal Express" as a simple and "accurate" search term doesn't make that distinction....)

The point is not to focus not on faster queries. And making them more accurate is tough because the real problem is that people don't describe what they want. The point is to do a better job describing the Web, so that people can navigate for themselves, starting in the right neighborhood and following the right cues, and see what they want.

Now the task of developing all the semiotics (which word does not appear in my MS Word thesaurus, but which means the symbol systems that describe or indicate things) for the Web is a huge one, well beyond the capabilities of any portal. You can buy (or rent) search engine technology, but how can you buy or even manage a catalogue for the entire Web? Even Yahoo!, the major portal with a catalogue of the Web rather than a search engine, doesn't cover most of the Net's territory. (Interestingly, it is also the most profitable of the portals, unless you include AOL in that category.)

But relax. The good news is that the Web is starting to describe itself. Sasha's project is more formal than most. Another company, Realize, wants to get people to rate one another's postings to improve the quality of discourse in online communities.

Everywhere, people are putting up signposts, pointing to other sites, building and sharing bookmark lists, and e-mailing links to one another. All the cross-references and hyperlinks you see on the typical Website are just parts of the human-built structure of the Web that is slowly accreting over time as people read and point and lay trails all over the Web. Companies such as IBM and Google (pages 16, 17) are building tools to detect and follow those links and aggregate them, and then let people pick sites by whether they are hubs — with lots of outward links — or authorities — with lots of inward links.

Meanwhile, other people are building different kinds of catalogues and directories more consciously: systems that classify goods by price (or by some specific metric such as tube size for piping, skin tone for cosmetics, disease in medicine, chemical structure in proteins). Even the booming classifieds and auction services are a way of ordering the world of available goods and services.

Kevin worries that there's no standard language for describing everything on the Web, but that's because there's no standard language for describing everything in the world. Right now we have a Web that's opaque and constructed artificially. The major way to find things is by brute force. But in a few years, most of the content on the Web will have become much better at describing itself, through a range of
methods ranging form formal catalogues to the kinds of trails people leave by their behavior. And at that point the Web will be like the real world: comprehensible up close, and visible as clearly as it needs to be from a distance.

Direct Hit has developed two additional offerings. The first, available now through HotBot, displays other topics related to the request. Direct Hit analyzes variants of the original search terms (both broader and narrower), and displays the 10 most popular alternatives. This helps users refine their searches based on the paths others have followed.

The second new service, Personalized Search, returns different search results based on the user's gender, age, geographic location or other demographic factors. Direct Hit is in discussions with potential partners to implement this technology, which considers the pages most often selected by a given group rather than the overall population.

Personalized Search raises some interesting questions. British users searching for "football" sites probably have something different in mind from Americans. (Danny Sullivan of Search Engine Watch, an American living in the UK, provided this example.) But what exactly does it mean when women prefer a different page from men? Will Third Age Media want to offer a senior citizen's search engine? What about a search engine for Republicans? Or open-source programmers? Will it be long before personalized search engines become a standard element of online community sites, along with free e-mail and home pages?

**Gaga for Google**

Google is the work of two Stanford graduate students, Larry Page and Sergey Brin. It's perhaps too obvious to point out that industrious Stanford students also founded Yahoo! and Excite. Where Yahoo! began as a largely manual directory and gradually expanded and evolved as the creators realized its potential, Google has been designed from the ground up as a highly efficient search service.

Despite the remarkable success of portals built around search engines, there has been little published research on improving Internet search results. Page and Brin stepped into this breach with Google. As Page describes it, he was looking for a dissertation topic three years ago and decided to analyze link structures on the Web. He was originally interested in the analogy between links and academic citations, but quickly realized that link structures could be used to rank pages for relevance to search queries.

Google uses a metric called PageRank (named for Larry Page) to determine the relevance of a page to a given query. As Brin and Page explain, PageRank "corresponds to the principal eigenvector of the normalized link matrix of the Web." In layman's terms, PageRank for a given page A is derived from the PageRanks of all pages linking to A, adjusted for the number of links on each of those pages. In other words, PageRank represents the possibility that a user clicking on links at random from one page to the next, will come upon a given page. Pages that many other

---

1 Market research is another possible application. Wouldn't Coke want to know that 60 percent of 18-to-24 year old men choose the Pepsi site when they search for "cola?"
pages point to get higher rankings, but not all links are given equal weight. A page may come up more frequently if many sites link to it, or if a smaller number of more influential sites do so, because the user is more likely to reach those intermediate sites.2

Google assigns a PageRank to every document in the index. When a user types in a query, the search engine extracts pages that contain the search terms, and uses the PageRank data to help order the results delivered to the user. In ordering search results, Google also considers factors such as the proximity of search terms within a document and whether text is in boldface or larger point sizes. Because Google caches the full text of every page in its index, it is able to display the search term in context to help users select the most relevant document.

Google associates the highlighted words in a hyperlink not only with the page on which they reside, but also with the pages the links point to. This allows the search engine to match keywords not only on the target page, but also based on descriptions of that page by others. In many cases outside links provide a better summary of page content than the text of the page itself, especially for a computer program incapable of directly understanding that content. Google can also find pages that its crawlers don't reach, so long as there are links to them from pages the system does analyze. The tradeoff is that Google has to index many more page references this way; its initial universe of 24 million pages included 259 million links. Fortunately, the falling price of processing makes this a manageable number.

For an effective Web search service, generating results from the database is only part of the challenge. The system must have high-performance crawlers to find Web pages to include in its database. AltaVista claims its crawlers can visit 6 million pages per day; Google's architects claim that at peak rates their system can exceed 200 pages crawled per second, or 17.2 million per day (assuming they are willing to pay for the bandwidth). Google's public site currently has about 60 million pages indexed, and the company plans to release a much bigger index soon.

In September, Google's creators formed a company to bring Google to market. Page serves as CEO and Brin as president, and the company received seed funding from angel investors led by Andy Bechtolsheim. Page says that setting up a company was the best way to get Google's technology out into the world, although like most Internet entrepreneurs he identifies an IPO as the company’s goal. Google is hiring engineers and planning for a commercial launch “pretty soon;” an alpha test version of the search service is available at www.google.com. Page emphasizes that search technology is a particularly rich opportunity for innovation, and that “it will be a long time before search is solved.”

Boy, are those IBM guys CLEVER

Client-Side Eigenvector Enhanced Retrieval (CLEVER), a project of IBM's Almaden research center in San Jose, also uses link structure to improve search results. The technology grew out of the Hypertext-Induced Topic Search (HITS) algorithm developed by Cornell computer science professor Jon Kleinberg, at the time a visiting scholar at Almaden. According to Prabhakar Raghavan, senior manager of computer science at Almaden, IBM is
refining the technology and discussing commercial possibilities with portals and others.

CLEVER uses an iterative process. The system begins with a group of pages related to a topic, usually obtained through an existing search engine. It then collects all the pages linked to those pages and all the other pages it can find that link to the initial group. CLEVER initially ranks each page based on the number of pages that link to it, then repeatedly recalculates the scores by giving links from pages with many links greater weight. CLEVER uses a variety of other techniques to improve relevance, for example giving greater weighting to links that contain the search term in their highlighted text and giving lower weighting to links within the same site.

CLEVER and Google both use link structure to determine the most relevant pages, but they go about this process differently. As Kleinberg explains, Google first ranks and then searches, whereas CLEVER searches and then ranks. Google assigns PageRanks to everything and then uses the query terms to extract the most relevant pages. CLEVER starts with a more limited set of pages based on the query, and then generates relevance scores for those pages. Because CLEVER must query a search engine and compute rankings each time, it generally takes much longer than Google to generate results. However, CLEVER's results are more sensitive to whether a page is an authority or hub for the particular topic under consideration.

CLEVER is designed to find what the IBM researchers call authority and hub sites. Authority sites are those that contain the best information on a given topic, and hub sites provide links to many authority sites. To put it another way, hubs have many good links out, and authorities have many good links in. CLEVER gives each page both a hub score and an authority score, in contrast to Google which looks only for authorities.

According to Kleinberg, "in a lot of situations the hubs are as valuable if not more than the authorities." A good jumping-off point for a particular topic may give you more information than a specific document, no matter how relevant. In real life, for example, many people decide what movie to see by following reviewers they respect, rather than reading all the reviews of a particular movie. Websites that offer good resources or good links similarly develop reputations through word of mouth, e-mail or published recommendations (see Release 1.0, 1-98). Mechanical search engines can't understand reputations in the same way as humans, but CLEVER can leverage the human activity implicit in webs of reciprocal hyperlinks.

Because CLEVER excels at finding groups of related pages around hubs and authority sites, it is a perfect tool to uncover subject-oriented Web communities. A group of users and site owners may not even realize that they are congregating around the same set of sites, but CLEVER can find the forest of link relationships that tie those sites together. A practical use for this capability is building Web category directories such as Yahoo! The real Yahoo! employs dozens of human ontologists to sift
through and categorize links. People are expensive, and even Yahoo! indexes only some 1 million pages. In tests, IBM researchers found that 81 percent of the time CLEVER assembled groups of category links that users found more accurate than those Yahoo! offered.

Based on work with CLEVER, Raghavan estimates that there are more than 100,000 thematically unified virtual Web communities. Members of these communities may not even realize that others share their interests, but the IBM research suggests than 96 percent of the time pages with overlapping link structures have a concrete thematic relationship.

Kleinberg acknowledges that CLEVER is attuned to communities of page creators, rather than page browsers, because the former create the hyper-links that the system analyzes. Because different types of information on the Web have different styles of authorship, CLEVER will be more effective on some than on others.

Learning from links

Google and CLEVER do more than just find the most relevant pages. Link structure analysis suggests not only what a page means, but also what others think about it. As IBM’s Raghavan points out, “relevance is not the same as authority,” and link structure is an excellent way to unearth authority patterns on the Web. The Net has no central hierarchy or formal constitution, but authority does matter. People and companies follow IETF standard because the IETF has authority, not because anyone has officially delegated power to it.

A link isn’t always a positive recommendation. Positive links will generally receive higher scores in CLEVER or Google because the sites they point to are more likely to link back. However, a negative review of a software product may be at least as important to a searcher as a positive one, even though the software vendor neglected to point to it. If enough other sites link to the critical information, it will show up as either a hub or an authority.

Links can also tell you more than sites themselves. As Raghavan points out, IBM’s own Website doesn’t talk much about IBM as a mainframe company, because of the connotations of the bad old pre-Gerstner days. However, IBM is still a leading player in the mainframe market. CLEVER would make the association from outside links even though IBM wouldn’t.

Links also can tell us something about the relationship of the Web to society as a whole. According to Raghavan, most of the links on English literature are situated not in England but at American universities, because of the over-representation of those sites on the Web compared to the real world. There are far more communities around English literature than German literature for the same reason, although as the Web grows we can anticipate that such imbalances will begin to disappear. Similarly, Page says that the more wired universities (such as Stanford) tend to show up more often on Google than other campuses, because they have richer link structures.

A final benefit of link structure analysis is that it is difficult to artificially manipulate such algorithms. Both Google and CLEVER consider
not simply the number of links to a given page, but how important those links are. Someone who wants an artificially high ranking can establish dummy pages linking to his or her site, but unless authoritative sites point to those dummy pages the ploy won't be effective. Search engines such as Google and CLEVER can also use techniques based on linear algebra to identify and further devalue such artificial links.

ANSWERS, NOT DOCUMENTS

What's the point of a search engine, anyway?

Users are ultimately interested not in documents but in what those documents contain. They are looking for answers to questions, but those questions don't always map to keyword queries. Someone who uses the keyword "Ford" may be looking for the Ford Motor Company home page, prices at local Ford dealers, a biography of President Gerald Ford, an image of George Washington fording the Delaware River or something completely different. Because Boolean search engines typically treat documents as nothing more than streams of characters, they can't differentiate among these questions. They may find documents, but not the desired answer.

The problem is that we ask search engines to do things they aren't good at. In the physical world, we use the yellow pages to find the address of John's Pizza, but the Zagat's guide to find a good pizza place on the Upper West Side. Similarly, if you want the American Airlines home page, a directory service such as Centraal's RealNames works better than a search service. But RealNames won't be as useful if you're researching a history of the airline industry and looking for other papers on the topic. As Esther explains (see page 14), search services are popular partly because humans haven't yet mapped the Web themselves.

The companies we've discussed up to now take the role of search engines for granted and use new techniques to improve the results. The alternative is to change search services to focus more on answers than documents.

There's something about queries

What comes out of a search engine depends on what goes in. Keywords, even supplemented with Boolean connectors, provide limited information. And mainstream users rarely use even the Boolean tools available today. Link structure and popularity analysis to some degree compensate for the limited information in the query. In some cases, however, a richer query language, analogous to the structured query language (SQL) standard for relational databases, would be a better solution.

Query languages work best on structured data. Interest in Web query languages has therefore paralleled work on XML, RDF and other standards to add structure to Web documents (see page 3). In early December, the W3C held a query languages workshop to begin thinking about requirements and solutions. More than 90 W3C members participated, and so far 66 position papers have been posted on the W3C Website.
Others such as Ask Jeeves and Lexeme are working on new query systems that go in the opposite direction. Natural language processing allows questions to be expressed directly rather than as keywords, making it easier for mainstream users to express what they're looking for.

Ask not what Jeeves can do for you...

Ask Jeeves is designed to make it simple for users to ask questions and receive direct answers. It offers a natural language interface tied to a custom knowledgebase of common question and answer types. The company was founded in Berkeley, CA, in 1996. President and CEO Rob Wrubel came from educational software publisher Knowledge Adventure in June 1998. He says that “most people don’t want everything,” when they query a search service; they prefer high-quality results pre-selected by trusted editors.

Ask Jeeves responds to queries with a list of related questions that it can answer. For example, “Who won the World Series in 1980?” brings back “What happened in the World Series of the year 1980?” along with “Who won Tony Awards in 1980?” “What was unique about the 1905 World Series?” and several other choices. The user selects the most relevant question, which links to a Website that has the answer. Ask Jeeves also includes a meta-search engine that provides more traditional results if the knowledgebase is insufficient. The recent NPD survey ranked Ask Jeeves first in overall effectiveness out of 12 major search engines. The results, however, show just how far everyone has to go. Only 24 percent of Ask Jeeves users found information they were looking for “every time,” yet that was the best score in the survey (tied with GoTo).

Wrubel argues there’s more to search than answering the narrow question in the user’s head. When Ask Jeeves doesn’t have a good response, it says so, rather than providing a long list of weakly-related results. This helps put users at ease, giving the Net “a more humanized face.” Wrubel believes even answers that seem off-target, like the one above about the Tony Awards, help illuminate possibilities the user may find interesting.

Ask Jeeves’ natural language processing engine considers both semantic and syntactic factors to extract the essence of a question. The system’s knowledgebase contains alternate forms of common questions, mapped to one or more templates that link to Web pages with the relevant answer. Ask Jeeves uses human researchers to find useful Websites that are incorporated into the answer templates. The researchers monitor the stream of questions from users and develop new question and answer templates for common queries.

Ask Jeeves is available on a standalone site and through AltaVista. The company also offers a separate site designed for kids. In December, Ask Jeeves launched a tailored version of its service on the Dell Website called Ask Dudley. Much of the knowledgebase is specific to Dell, but some of the question and answer templates will contribute to the general Ask Jeeves knowledgebase.

Wrubel says Ask Jeeves could become a sort of “portal’s portal,” helping
users navigate the sprawling content networks of the major portals. Google and Direct Hit aren’t threats, he says, because Ask Jeeves sits higher in the value chain. Libraries offer both comprehensive card catalogs and reference librarians who can recommend good resources. By analogy, most users will look to editorially-selected sites first and then use open-ended search services as a backstop.

Search from the inside out

Search engines are like mainframes: they put all the intelligence at the center. In theory, pushing processing out to the endpoints would be more efficient, for the same reasons the Internet has triumphed over centralized networking models (see Release 1.0, 6-98). But search depends on universal coverage, and the only way to assure that something gets indexed is to do it yourself.

Distributed indexing systems such as Harvest and WordCruncher have been around for several years. Sites generate local indexes in a common format, and the central search engine need only tap into those indexes rather than crawling the raw pages. These systems can be effective for small communities, but on the Internet as a whole there’s no way to guarantee that all sites will generate and update their indexes.

WordCruncher, based on technology developed in the 1980s at Brigham Young University, is used in universities to search local documents and databases. In late 1996, James Johnston and Daniel Lunt formed WordCruncher Internet Technologies to bring WordCruncher to the Web. The company has licensed the technology from Brigham Young and plans to launch in the first quarter of 1999.

Lexeme: Lexical memes

John Clippinger, ceo of Lexeme, believes that improvements in processing power and algorithms have finally made it possible to extract knowledge automatically out of unstructured text. Clippinger developed one of the first corporate intranets for knowledge management while at Coopers & Lybrand. He founded Lexeme with cto James Pustejovsky, a professor of computational linguistics at Brandeis University, and vp of business development Jim Keller, formerly with the Harvard Information Infrastructure Project. The company, based in Cambridge, MA, will make a company presentation at PC Forum.

Lexeme’s engine parses text to extract entities, relations and concepts. It can distinguish objects from their characteristics, and can identify related or identical concepts even if they use different words. The engine in effect creates its own categories and populates a relational database with lexical objects, which can then be searched through a query interface. Traditional Boolean search engines look at files as merely strings of ASCII characters. Clippinger says that “English is the mother of all protocols,” and therefore search engines must understand language to extract the full meaning from documents. Lexeme processes actual linguistic structures as objects, which gives it a rich understanding of concepts and attributes. Where Ask Jeeves uses human editors to hard-
code question and answer templates, Lexeme organizes information automatically.

As Pustejovsky puts it, "We're in the business of providing answers, not hits." Lexeme uses a conversational natural language interface, rather than the more rigid structure of Boolean queries. The system is designed to bring back specific information that addresses users' needs, instead of a list of documents containing related material. By making both queries and content representations richer, Lexeme brings to bear the penumbras of meaning surrounding words (see page 1). Pustejovsky draws an analogy to the human genome. Identifying the building blocks isn't enough; you need to understand something about structure and have a model of how the components interact.

Lexeme's technology has broad application, although the startup is initially targeting vertical markets where it can generate high margins. Medstract.org, a site funded by the National Institutes of Health, is using Lexeme's technology to develop a database of functional characteristics of genes and proteins from scientific abstracts available through the Medline service. New research arrives daily, which is why Lexeme's ability to automatically organize information is so valuable.

Another area Lexeme plans to address is customer service for electronic commerce sites. The Lexeme engine will allow users to find the customer support information most relevant to their problems (for more on this market, see Release 1.0, 9-98). Because of the processing involved, Lexeme's system doesn't make sense for searching or "understanding" the entire Web, but it could serve as an adjunct to existing search services. Lexeme is talking to several portals about licensing deals.

one size doesn't fit all

No search engine will ever be flawless, because people have a variety of needs. Do they want a specific answer, or an opportunity to rummage through related materials? Are they willing to trust others' judgments about what's interesting, or do they want to make their own determinations? Different tools may work better in each of these situations. (This is similar to the point we made in a previous issue about content labels. See Release 1.0, 5-98.)

The technologies we describe can improve search results, but they can't replace the human brain. Users don't always know what they're really looking for, and sometimes it changes depending on what they find along the way. The concept of a universal search engine has seductive appeal. As the artificial intelligence community has learned, however, some challenges are more difficult than they seem.

Search engines are like the map in the Borges story: They can only achieve perfection at a scale of 1:1, at which point they save no time at all. Yet there is real value in making search engines less imperfect. The Net isn't getting any smaller; people will always need tools to find what they're looking for. The companies discussed in this issue deliver significant enhancements in usability and result quality. But there's plenty of work left to be done. Because it invokes deep con-
cepts such as meaning, search will remain a challenge for a long time to come.

COMING SOON

- Cable and the future of the Net.
- How big companies innovate.
- Portals vs. portholes.
- Wireless and embedded networking.
- The Net swallows the phone network.
- Living on the Web.
- And much more... (If you know of any good examples of the categories listed above, please let us know.)
RESOURCES & PHONE NUMBERS

Rob Wrubel, Ask Jeeves, (510) 649-3550; fax, (510) 649-8633; rob@aj.com; www.ask.com

John Kleinberg, Cornell University, (607) 255-3600; fax, (607) 255-9555 kleinber@cs.cornell.edu; www.cs.cornell.edu/home/kleinber

Gary Culliss, Direct Hit, (781) 235-7570; fax, (781) 239-0196; gculliss@dirhit.com; www.dirhit.com

Harold Kester, Encyclopedia Britannica, (619) 622-4700; fax, (619) 622-4709; hkester@eb.com; www.eb.com

Barak Berkowitz, Infoseek/Go.com, (408) 543 6000; fax, (408) 734-9350

Lawrence Page, Sergey Brin, Google, (650) 330-0100; fax, (650) 618-1499; larry@google.com, sergey@google.com

Jeff Brewer, GoTo.com, (626) 535-2733; fax, (626) 535-2701; jeffrey@goto.com

Prabhakar Raghavan, IBM Almaden Research Center, (408) 927-1804; pragh@almaden.ibm.com

David Peterschmidt, Inktomi, (650) 653-2800; fax, (650) 653-2801

John Clippinger, Jim Keller, James Pustejovsky, Lexeme, (617) 492-7377; john@lexeme.com, jamesp@lexeme.com, keller@lexeme.com

Jim King, Lexis-Nexis, (937) 865-1182; fax, (937) 865-1786; james.king@lexis-nexis.com

Dan Pliske, Lexis-Nexis, (937) 865-6800, ext 4956; fax, (937) 865-1655; daniel.pliske@lexis-nexis.com

Jeff Levy, Media Metrix, (404) 224-3301; fax, (404) 224-2110; jeff@rkinc.com; www.mediametrix.com

Mark Peterson, Go2Net/MetaCrawler, (206) 447-1595 x299; fax, (206) 447-1625; mark@go2net.com

Tom Barrett, Namestake.com (Thomson & Thomson), (617) 368-3938; tom.barrett@namestake.com

Steve Lawrence, NEC Research, (609) 951-2676; fax, (609) 951-2488; lawrence@research.nj.nec.com; www.neci.nj.nec.com/~lawrence

Lee Giles, NEC Research, (609) 951-2642; fax, (609) 951-2488; giles@research.nj.nec.com; www.neci.nj.nec.com/homepages/giles

Sasha Chislenko, The Newsfilter Project, sasha1@netcom.com, www.lucifer.com/~sasha/newsfilter_plan.html

Danny Sullivan, Search Engine Watch, +44 (171) 446-0443, danny@calafia.com; www.searchenginewatch.com

Massimo Marchiori, World Wide Web Consortium, (617) 253-2442; fax, (617) 258-5999; massimo@w3.org; www.w3c.org

Except as noted otherwise, all companies’ Websites are at the likely address, http://www.domain_name.com.

For further reading:


World Wide Web Consortium Query Languages Workshop — www.w3.org/TandS/QL/QL98/Overview.html

Release 1.0 15 January 1999
2 On the other hand, some sites such as Yahoo! have so many links in both directions that each link isn't worth much.

3 Infoseek also analyzes link structure to improve search results.
Release 1.0 Calendar


February 7-10    #Demo 99 - Indian Wells, CA. Chris Shipley picks the hot startups. Call Alexa Hanes (650) 286-2730; e-mail alexa@demo.com; www.demo.com.


February 9-12    Milia '99 - Cannes, France. The international content market for interactive media. Contact Barney Bernhard, (212) 689-4220; fax, (212) 689-4348; e-mail infomilia-us@compuserve.com; www.milia.com.


March 1-3       Jupiter Consumer Online Forum - New York, NY. Focuses exclusively on the intersection between the consumer Internet and traditional media, entertainment and communications companies. To register, call (888) 780-5010 x103; fax, (212) 780-6075; e-mail jon@jup.com; www.jup.com/events/forums/cof/.


March 6-9       SPA/IIA Spring Symposium - Los Angeles, CA. Issues critical to the future of software and information providers. Contact Anika Valentine, (202) 452-1600 x339; fax, (202) 785-3649; avalentine@spa.org.

March 21-24     *#PC Forum - Scottsdale, AZ. Sponsored by EDventure Holdings. You read the newsletter; now meet the players. Call Daphne Kis, (212) 924-8800; fax, (212) 924-0240; daphne@edventure.com; www.edventure.com.


April 6-8       Computers, Freedom, and Privacy 99 - Washington, DC. The ninth annual conference on technology and public policy. This year's theme is "The Global Internet." E-mail info@cfp99.org; www.cpr99.org for more info.

April 12-16     Spring Internet World - Los Angeles, CA. For information call (800) 500-1959; e-mail siwprogram@mecklermedia.com; events.internet.com/spring99/.

April 26-29  ISPCON Spring 99 - Baltimore, MD.  The largest ISP trade show, now owned by Mecklermedia, er make that Penton Media.  For information call (800) 632-5537; ispcon.internet.com/spring99.


June 22-24  PC Expo - New York, NY.  Over 100,000 corporate technology buyers in search of new toys.  Sponsored by Miller Freeman; keynote speakers include Bob Herbold and Chuck Geschke.  For information, call (800) 829-3976; www.pcexpo.com.

June 22-25  INET '99 - San Jose, CA.  The Internet Society's annual conference.  For information e-mail inet99-register@isoc.org; www.isoc.org/inet99/.

* Events Esther plans to attend.
# Events Kevin plans to attend.

Lack of a symbol is no indication of lack of merit.
The full, current calendar is available on our Website, www.edventure.com.
Please let us know about other events we should include.  - Mari Katsunuma
SUBSCRIPTION FORM

Please enter my subscription to Release 1.0 at the rate of $695 per year in the U.S. and Canada. Overseas subscriptions are $750, airmail postage included. Payment must be enclosed. Satisfaction guaranteed or your money back.

Name

Title

Company

Address

City State Zip

Country

Telephone Fax

E-mail URL

☐ Check enclosed

☐ Charge my

☐ American Express  ☐ Master Card  ☐ Visa

Card Number Expiration Date

Name and Billing Address

Signature

☐ Please send me information on your multiple copy rate.

Please fill in the information above and send to:

EDventure Holdings Inc.
104 Fifth Avenue, 20th Floor
New York, NY 10011

If you have any questions, please contact us at 1 (212) 924-8800; fax 1 (212) 924-0240; e-mail us@edventure.com; www.edventure.com.

Daphne Kis
Publisher