BARRAGE of Internet information will drop to a focused trickle with new search engines that can take context—such as a user's long-term interests, location or other factors—into account.

By Javed Mostafa

Illustrations by Neil Brennan

Seeking Better Web Searches

Deluged with superfluous responses to online queries, users will soon benefit from improved search engines that deliver customized results

In less than a decade, Internet search engines have completely changed how people gather information. No longer must we run to a library to look up something; rather we can pull up relevant documents with just a few clicks on a keyboard. Now that "Googling" has become synonymous with doing research, online search engines are poised for a series of upgrades that promise to further enhance how we find what we need.

New search engines are improving the quality of results by delving deeper into the storehouse of materials available online, by sorting and presenting those results better, and by tracking your long-term interests so that they can refine their handling of new information requests. In the future, search engines will broaden content horizons as well, doing more than simply processing keyword queries typed into a text box. They will be able to automatically take into account your location—letting your wireless PDA, for instance, pinpoint the nearest restaurant when you are traveling. New systems will also find just the right picture faster by matching your sketches to similar shapes. They will even be able to name that half-remembered tune if you hum a few bars.

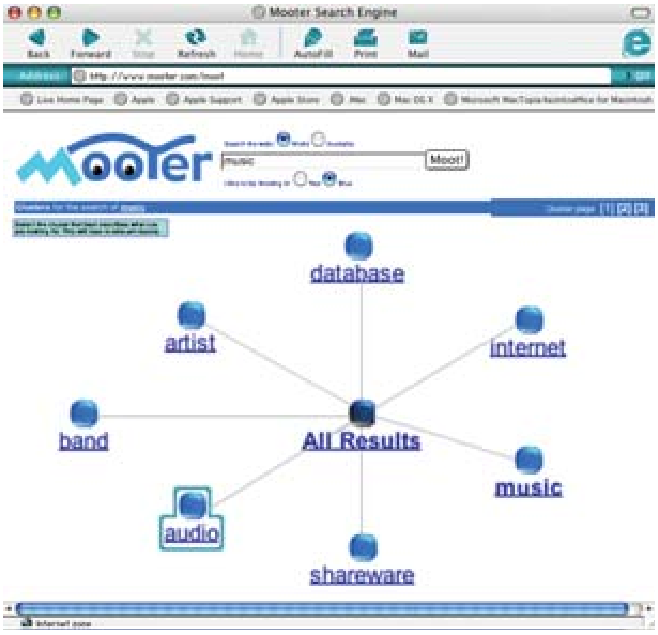
Today's search engines have their roots in a research field called information retrieval, a computing topic tracing back nearly 50 years. In a September 1966 *Scientific American* article, "Information Storage and Retrieval," Ben Ami Lipetz described how the most advanced information technologies of the day could handle only routine or clerical tasks. He then concluded perceptively that breakthroughs in information retrieval would come when researchers gained a deeper understanding of how humans process information and then endowed machines with analogous capabilities. Clearly, computers have not yet reached that level of sophistication, but they are certainly taking users' personal interests, habits and needs into greater account when completing tasks.

Prescreened Pages

BEFORE DISCUSSING new developments in this field, it helps to review how current search engines operate. What happens when a computer user reads on a screen that Google has sifted through billions of documents in, say, 0.32 second? Because matching a user's keyword query with a single Web page at a time would take too long, the systems carry out several key steps long before a user conducts a search.

First, prospective content is identified and collected on an ongoing basis. Special software code called a crawler is used to probe pages published on the Web, retrieve these and linked pages, and aggregate pages in a single location. In the second step, the system counts relevant words and establishes their importance using various statistical techniques. Third, a highly efficient data structure, or tree, is generated from the relevant terms, which associates those terms with specific Web pages. When a user submits a query, it is the completed tree, also known as an index, that is searched and not individual Web pages. The search starts at the root of the index tree, and at every step a branch of the tree (representing many terms and related Web pages) is either followed or eliminated from consideration, reducing the time to search in an exponential fashion.

To place relevant records (or links) at or near the top of the retrieved list, the search algorithm applies various ranking strategies. A common ranking method—term frequency/inverse document frequency—considers the distribution of words and their frequencies, then generates numerical weights for words that signify their importance in individual documents. Words that are frequent (such as "or," "to" or "with") or that appear in many documents are given substantially less weight than words that are more relevant semantically or appear in comparatively few documents.



MOOTER, a new search engine, simplifies the user's assessment of results by categorizing the collected information and clustering related sites under on-screen buttons. Subcategory buttons surround the central general topic cluster. Clicking on a cluster button retrieves lists and new, associated clusters.

WWW.MOOTER.COM

Overview/Beyond Google

- As Web sites continue to multiply rapidly, Internet users need more precise search engines that find what they are looking for more quickly and efficiently.
- The next search engines will improve results by digging deeper through online materials, by better classifying and displaying the catch, and by monitoring users' interests to respond more intelligently to future searches. New software will track a user's location and will handle graphics and music as well as text.
- New business models will eventually open up nearly all published digital information—text, audio and video resources that are currently unavailable on the Web—to smart search functions.

In addition to term weighting, Web pages can be ranked using other strategies. Link analysis, for example, considers the nature of each page in terms of its association with other pages—namely, if it is an authority (by the number of other pages that point to it) or a hub (by the number of pages it points to). Google uses link analysis to improve the ranking of its search results.

Superior Engines

DURING THE SIX YEARS in which Google rose to dominance, it offered two critical advantages over competitors. One, it could handle extremely large-scale Web crawling tasks. Two, its indexing and weighting methods produced superior ranking results. Recently, however, search engine builders have developed several new, similarly capable schemes, some of which are even better in certain ways.

Much of the digital content today remains inaccessible because many systems hosting (holding and handling) that material do not store Web pages as users normally view them. These resources generate Web pages on demand as users interact with them. Typical crawlers are stumped by these resources and fail to retrieve any content. This keeps a huge amount of information—approximately 500 times the size of the conventional Web, according to some estimates—concealed from users. Efforts are under way to make it as easy to search the "hidden Web" as the visible one.

To this end, programmers have developed a class of soft-

Efforts are under way to make it as easy to **search the "hidden Web"** as the visible one.

ware, referred to as wrappers, that takes advantage of the fact that online information tends to be presented using standardized "grammatical" structures. Wrappers accomplish their task in various ways. Some exploit the customary syntax of search queries and the standard formats of online resources to gain access to hidden content. Other systems take advantage of application programming interfaces, which enable software to interact via a standard set of operations and commands. An example of a program that provides access to the hidden Web is Deep Query Manager from BrightPlanet. This wrapper-based query manager can provide customized portals and search interfaces to more than 70,000 hidden Web resources.

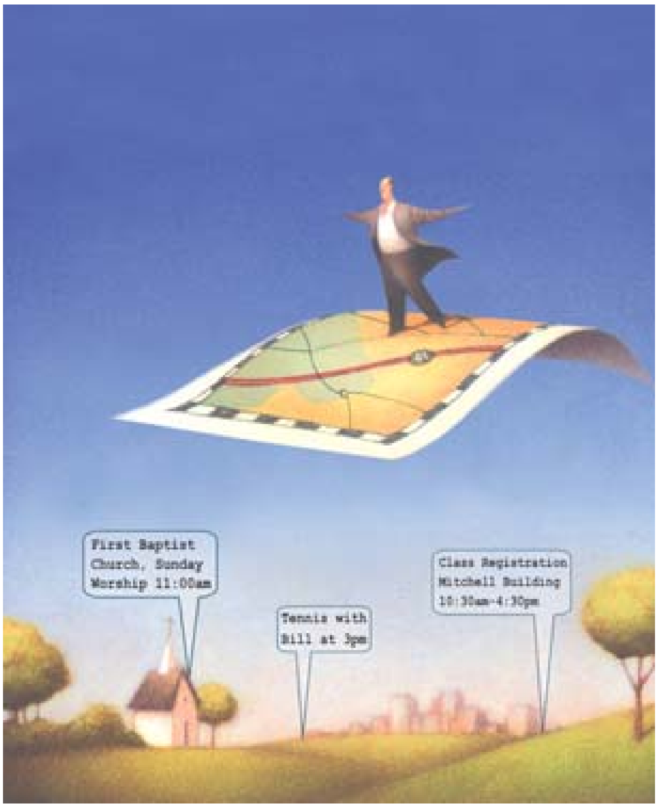
Relying solely on links or words to establish ranking, without placing any constraint on the types of pages that are being compared, opens up possibilities for spoofing or gaming the ranking system to misdirect queries. When the query "miserable failure," for example, is executed on the three top search engines—Google, Yahoo and MSN—a page from the whitehouse.gov site appears as the top item in the resulting set of retrieved links.

Rather than providing the user with a list of ranked items (which can be spoofed relatively easily), certain search engines attempt to identify patterns among those pages that most closely match the query and group the results into smaller sets. These patterns may include common words, synonyms, related words or even high-level conceptual themes that are identified using special rules. These systems label each set of links with its relevant term. A user can then refine a search further by selecting a particular set of results. Northern Light (which pioneered this technique) and Clusty are search engines that present clustered results.

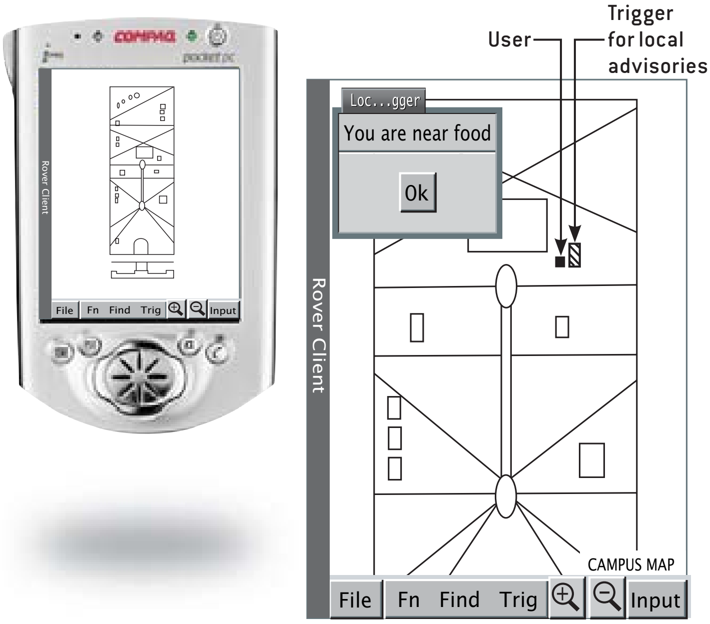
Mooter, an innovative search engine that also employs clustering techniques, provides researchers with several additional advantages by presenting its clusters visually [*see illustration on opposite page*]. It arrays the subcategory buttons around a central button representing all the results, like the spokes of a wheel. Clicking on a cluster button retrieves lists of relevant links and new, associated clusters. Mooter remembers the chosen clusters. By clicking on the "refine" option, which combines previously retrieved search clusters with the current query, a user can obtain even more precise results.

A similar search engine that also employs visualization is Kartoo. It is a so-called metasearch engine that submits the user's query to other search engines and provides aggregated results in a visual form. Along with a list of key terms associated with various sites, Kartoo displays a "map" that depicts important sites as icons and relations among the sites as labeled paths. Each label can be used to further refine the search.

Another way computer tools will simplify searches is by looking through your hard drive as well as the Web. Currently


First Baptist Church, Sunday Worship 11:00am
Tennis with Bill at 3pm
Class Registration Mitchell Building 10:30am-4:30pm

I KNOW WHERE WE ARE: A computing environment that is aware of its location, such as the University of Maryland's Rover technology, makes it possible for a wireless handheld device to know its position on the map at all times. This feature allows Rover to provide customized information about the local surroundings to the user on the go.


Loc...gger

You are near food

Ok

MIND LAB, UNIVERSITY OF MARYLAND, COLLEGE PARK

Recently Amazon, Ask Jeeves and Google announced initiatives that would **allow users to personalize their searches.**

searches for a file on a computer user's desktop require a separate software application. Google, for example, recently announced Desktop Search, which combines the two functions, allowing a user to specify a hard disk or the Web, or both, for a given search. The next release of Microsoft's operating system, code-named Longhorn, is expected to supply similar capabilities. Using techniques developed in another Microsoft project called Stuff I've Seen, Longhorn may offer "implicit search" capabilities that can retrieve relevant information without the user having to specify queries. The implicit search feature reportedly harvests keywords from textual information recently manipulated by the user, such as e-mail or Word documents, to locate and present related content from files stored on a user's hard drive. Microsoft may extend the search function to Web content and enable users to transform any text content displayed on screens into queries more conveniently.

Search Me

RECENTLY AMAZON, Ask Jeeves and Google announced initiatives that attempt to improve search results by allowing users to personalize their searches. The Amazon search engine, A9.com, and the Ask Jeeves search engine, MyJeeves.ask.com, can track both queries and retrieved pages as well as allow users to save them permanently in bookmark fashion. In MyJeeves, saved searches can be reviewed and reexecuted, providing a way to develop a personally organized subset of the Web. Amazon's A9 can support similar functions and also employs personal search histories to suggest additional pages. This advisory function resembles Amazon's well-known book recommendation feature, which takes advantage of search and purchasing patterns of communities of users—a process sometimes called collaborative filtering.

The search histories in both A9 and MyJeeves are saved not on users' machines but on search engine servers so that they can be secured and later retrieved on any machine that is used for subsequent searches.

In personalized Google, users can specify subjects that are of particular interest to them by selecting from a pregenerated hierarchy of topics. It also lets users specify the degree to which they are interested in various themes or fields. The system then employs the chosen topics, the indicated level of interest, and the original query to retrieve and rank results.

Although these search systems offer significant new features, they represent only incremental enhancements. If search engines could take the broader task context of a person's query into account—that is, a user's recent search subjects, personal behavior, work topics, and so forth—their utility would be greatly augmented. Determining user context will require software designers to surmount serious engineering hurdles, however. Developers must first build systems that monitor a user's interests and habits automatically so that search engines can ascertain the context in which a person is conducting a search for information, the type of computing platform a user is running, and his or her general pattern of use. With these points established beforehand and placed in what is called a user profile, the software could then deliver appropriately customized information. Acquiring and maintaining accurate information about users may prove difficult. After all, most people are unlikely to put up with the bother of entering personal data other than that required for their standard search activities.

THE AUTHOR

JAVED MOSTAFA is Victor H. Yngve Associate Professor of Information Science at Indiana University. He is an associate editor of *ACM Transactions on Information Systems* and directs the Laboratory of Applied Informatics Research at Indiana. The author dedicates this article to one of his favorite teachers, Ms. Shaila of BWA School in Chittagong, Bangladesh.

Good sources of information on personal interests are the records of a user's Web browsing behavior and other interactions with common applications in their systems. As a person opens, reads, plays, views, prints or shares documents, engines could track his or her activities and employ them to guide searches of particular subjects. This process resembles the implicit search function developed by Microsoft. PowerScout and Watson are the first systems introduced capable of integrating searches with user-interest profiles generated from indirect sources. PowerScout has remained an unreleased laboratory system, but Watson seems to be nearing commercialization. Programmers are now developing more sophisticated software that will collect interaction data over time and then generate and maintain a user profile to predict future interests.

The user-profile-based techniques in these systems have not been widely adopted, however. Various factors may be responsible: one issue may be the problems associated with maintaining profile accuracy across different tasks and over extended periods. Repeated evaluation is necessary to establish robust profiles. A user's focus can change in unpredictable and subtle ways, which can affect retrieval results dramatically.

Another factor is privacy protection. Trails of Web navigation, saved searches and patterns of interactions with applications can reveal a significant amount of secret personal information (even to the point of revealing a user's identity). A handful of available software systems permit a user to obtain content from Web sites anonymously. The primary means used by these tools are intermediate or proxy servers through which a user's transactions are transmitted and processed so that the site hosting the data or service is only aware of the proxy systems and cannot trace a request back to an individual user. One instance of this technology is the anonymizer.com site, which permits a user to browse the Web incognito. An additional example is the Freedom WebSecure software, which employs multiple proxies and many layers of encryption. Although these tools offer reasonable security, search services do not yet exist that enable both user personalization and strong privacy protection. Balancing the maintenance of privacy with the benefits of profiles remains a crucial challenge.

On the Road

ANOTHER CLASS of context-aware search systems would take into account a person's location. If a vacationer, for example, is carrying a PDA that can receive and interpret signals from the Global Positioning System (GPS) or using a radio-frequency technique to establish and continuously update position, systems could take advantage of that capability. One example of such a technology is being developed by researchers at the University of Maryland. Called Rover, it is a system that makes use of text, audio or video services across a wide geographic area [*see illustration on page 69*]. Rover can present maps of the region in a user's vicinity that highlight appropriate points of interest. It is able to identify these spots automatically by applying various subject-specific "filters" to the map.

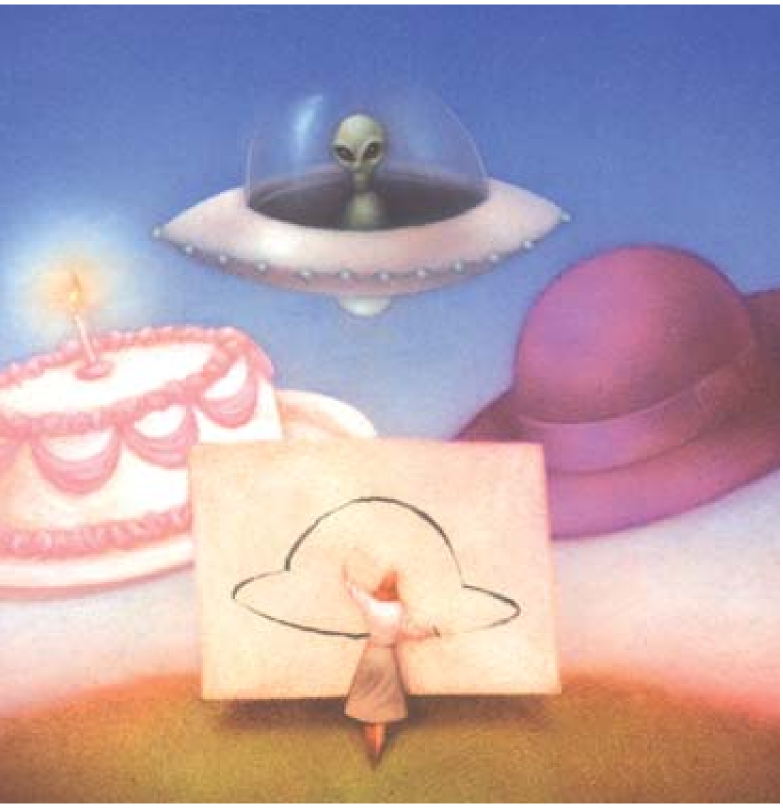
The system can provide additional information as well. If a Rover client were visiting a museum, for example, the handheld device would show the institution's floor plan and nearby displays. If the user stepped outside, the PDA display would change to an area map marking locations of potential interest. Rover would also permit an operator to enter his or her position directly and retrieve customized information from the networked database. In 2003 the group that created Rover and KoolSpan, a private network company, received funding from the Maryland state government to develop jointly applications for secure wireless data delivery and user authentication. This collaboration should result in a more secure and commercially acceptable version of Rover.

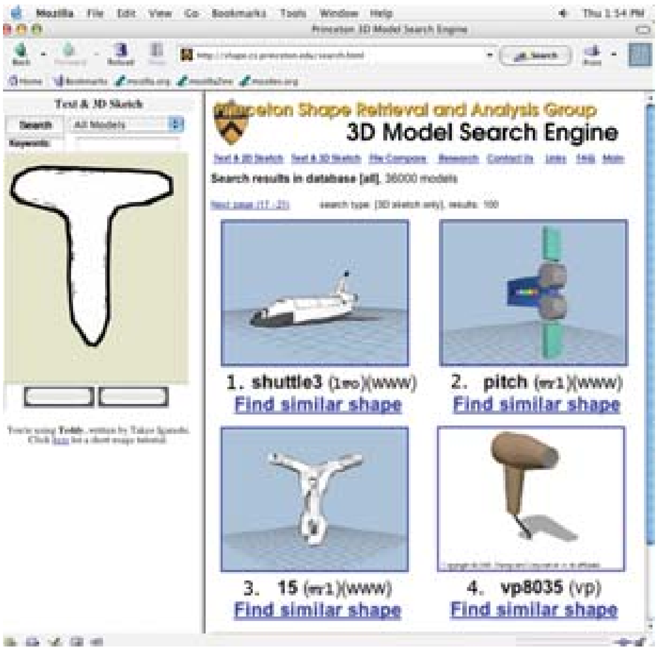
Unfortunately, the positional error of GPS-based systems (from three to four meters) is still rather large. Even though this resolution can be enhanced by indoor sensor and outdoor beacon systems, these technologies are relatively expensive to implement. Further, the distribution of nontext information, especially images, audio and video, would require higher bandwidth capacities than those currently available from handheld devices or provided by today's wireless networks. The IEEE 802.11b wireless local-area network protocol, which offers bandwidths of up to 11 megabits per second, has been tested successfully in providing location-aware search services but is not yet widely available.

Picture This

CONTEXT CAN MEAN MORE than just a user's personal interests or location. Search engines are also going beyond text queries to find graphical material. Many three-dimensional images are now available on the Web, but artists, illustrators and designers cannot effectively search through these drawings or shapes using keywords. The Princeton Shape Retrieval and Analysis Group's 3-D Model Search Engine supports three methods to generate such a query [*see illustration below*]. The first approach uses a sketchpad utility called Teddy, which allows a person to draw basic two-dimensional shapes. The soft-

THOMAS FUNKHOUSER, PRINCETON SHAPE RETRIEVAL AND ANALYSIS GROUP





Princeton Shape Retrieval and Analysis Group
3D Model Search Engine

Search results in database [all], 36000 models

1. shuttle3 (1mo)(www) Find similar shape
2. pitch (mr1)(www) Find similar shape
3. 15 (mr1)(www) Find similar shape
4. vp8035 (vp) Find similar shape

FIND THAT SHAPE: The 3-D Model Search Engine from the Princeton Shape Retrieval and Analysis Group matches a desired shape with multiple images of similar forms available on the Internet. Designers, engineers and architects can locate analogous three-dimensional objects much more rapidly than before.

Web searchers will steer through voluminous data repositories using interfaces that **establish broad patterns in information.**

ware then produces a virtual solid extrusion (by dragging 2-D images through space) from those shapes. The second lets a user draw multiple two-dimensional shapes (approximating different projections of an image), and the search engine then matches the flat sketches to 13 precomputed projections of each three-dimensional object in its database. Theoretically, this function can be generalized to support retrieval from any 2-D image data set. The third way a person can find an image is to upload a file containing a three-dimensional model.

The system, still in development, matches queries to shapes by first describing each shape in terms of a series of mathematical functions—harmonic functions for three-dimensional images and trigonometric ones for two-dimensional representations. The system then produces certain "fingerprinting" values from each function that are characteristic for each associated shape. These fingerprints are called spherical or circular signatures. Two benefits arise from using these descriptors: they can be matched no matter how the original and search shapes are oriented, and the descriptors may be computed and matched rapidly.

What's That Song?

MUSIC HAS ALSO ENTERED the search engine landscape. A key problem in finding a specific tune is how to best formulate the search query. One type of solution is to use musical notation or a musical transcription-based query language that permits a user to specify a tune by keying in alphanumeric characters to represent musical notes. Most users, however, find it difficult to transform the song they have in mind to musical notation.

The Meldex system, designed by the New Zealand Digital Library Project, solves the problem by offering a couple of ways to find music [*see illustration on opposite page*]. First, a user can record a query by playing notes on the system's virtual keyboard. Or he or she can hum the song into a computer microphone. Last, users can specify song lyrics as a text query or combine a lyrics search with a tune-based search.

To make the Meldex system work, the New Zealand researchers had to overcome several obstacles: how to convert the musical query to a form that could be readily computed; how to store and search song scores digitally; and how to match those queries with the stored musical data. In the system, a process called quantization identifies the notes and pitches in a query. Meldex then detects the pitches as a function of time automatically by analyzing the structure of the waveforms and maps them to digital notes. The system stores both notes and complete works in a database of musical scores. Using data string-matching algorithms, Meldex finds musical queries converted into notes that correspond with notes from the scores database. Because the queries may contain errors, the string-matching function must accommodate a certain amount of "noise."

Searching the Future

FUTURE SEARCH SERVICES will not be restricted to conventional computing platforms. Engineers have already integrated them into some automotive mobile data communications (telematics) systems, and it is likely they will also embed

PLACING ALL MEDIA ON THE NET

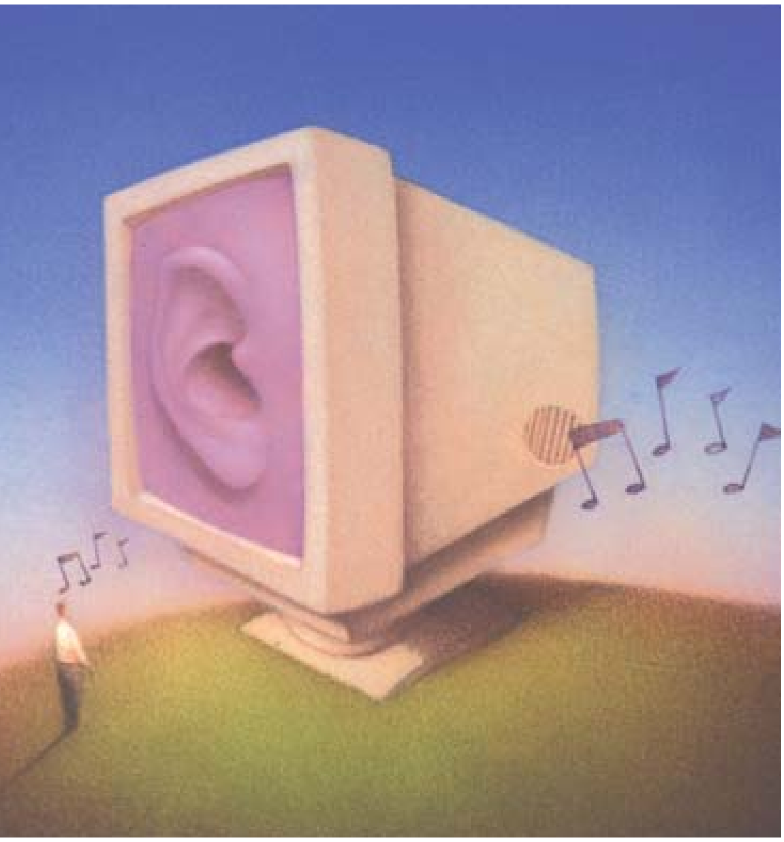
Although the Internet covers a tremendous amount of information, much of what is published today—text, audio and video—is not available online. Content is costly, and its producers want to exercise maximum control over what they generate, so they limit access severely. That situation is changing, however, as collaboration grows among content producers (such as Time-Warner, Sony, Hearst, Elsevier, and so forth) and "brand-name" search engine players (particularly the big three, Yahoo, Google and MSN). The challenge is to forge business relations that are beneficial to both sides.

If suitable contractual agreements were established between media publishers and search engine companies, arranging for the content producers' sites to be crawled and indexed by search engines would be relatively straightforward. Say a search engine user were to find a citation to a specific content producer's item; the link could then route the user to the appropriate site for access. There the user could be offered various options for obtaining the complete content.

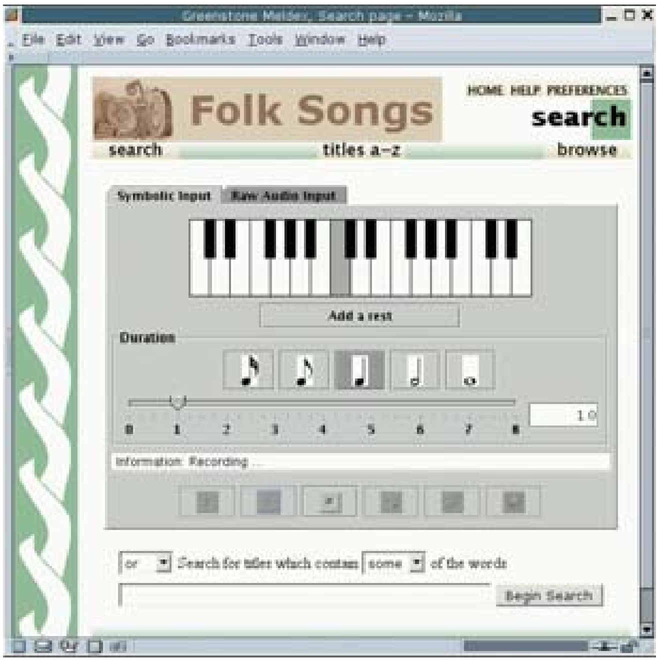
In some pilot projects, content providers are allowing indexing of their raw product. Amazon, for instance, has instituted an experimental project through which customers can read the full texts of books. Google recently introduced a service for publishers and large libraries to submit their books for indexing so they can be included in the same indexes as Web content.

Related concerns exist in the audio and video fields. Production houses and studios are reluctant to adopt new avenues of distribution. Here, too, however, alternative marketing models are appearing. Apple has promoted its iTunes music store aggressively, and both Dell and Hewlett-Packard have announced music delivery services.

Eventually, industry observers say, search engines will most likely serve as "hubs" or gateways to all types of content. They will generate and maintain indexes as well as provide search services for diverse classes of published media. Content providers will meanwhile concentrate on their core creative businesses. *—J.M.*



NAME THAT TUNE: Formulating a query to find a particular song or melody in an Internet-based musical database is not easy for people without formal music training. The Meldex system from the New Zealand Digital Library Project lets a user hum a half-remembered tune into a PC's microphone or type in some of its lyrics, and the software identifies the matching song or melody fast.



NEW ZEALAND DIGITAL LIBRARY PROJECT, UNIVERSITY OF WAIKATO

search capabilities into entertainment equipment such as game stations, televisions and high-end stereo systems. Thus, search technologies will play unseen ancillary roles, often via intelligent Web services, in activities such as driving vehicles, listening to music and designing products.

Another big change in Web searching will revolve around new business deals that greatly expand the online coverage of the huge amount of published materials, including text, video and audio, that computer users cannot currently access [*see box on opposite page*].

Ironically, next-generation search technologies will become both more and less visible as they perform their increasingly sophisticated jobs. The visible role will be represented by more powerful tools that combine search functions with data-mining operations—specialized systems that look for trends or anomalies in databases without actually knowing the meaning of the data. The unseen role will involve developing myriad intelligent search operations as back-end services for diverse applications and platforms. Advances in both data-mining and user-interface technologies will make it possible for a single system to provide a continuum of sophisticated search services automatically that are integrated seamlessly with interactive visual functions.

By leveraging advances in machine learning and classification techniques that will be able to better understand and categorize Web content, programmers will develop easy-to-use visual mining functions that will add a highly visible and interactive dimension to the search function. Industry analysts expect that a variety of mining capabilities will be available, each tuned to search content from a specialized domain or format (say, music or biological data). Software engineers will design these functions to respond to users' needs quickly and conveniently despite the fact they will manipulate vast quantities of information. Web searchers will steer through voluminous data repositories using visually rich interfaces that focus on establishing broad patterns in information rather than picking out individual records. Eventually it will be difficult for computer users to determine where searching starts and understanding begins. SA

MORE TO EXPLORE

Information Storage and Retrieval. Ben Ami Lipetz in *Scientific American*, Vol. 215, No. 3, pages 224–242; September 1966.

Exploring the Web with Reconnaissance Agents. H. Lieberman, C. Fry and L. Weitzman in *Communications of the ACM*, Vol. 44, No. 8, pages 69–75; August 2001.

Web Search—Your Way. E. Glover et al. in *Communications of the ACM*, Vol. 44, No. 12, pages 97–102; December 2001.

Rover: Scalable Location-Aware Computing. S. Banerjee et al. in *Computer*, Vol. 35, No. 10, pages 46–53; October 2002.

A Search Engine for 3D Models. T. Funkhouser et al. in *ACM Transactions on Graphics*, Vol. 22, No. 1, pages 83–105; January 2003.

Simulation Studies of Different Dimensions of Users' Interests and Their Impact on User Modeling and Information Filtering. Javed Mostafa, S. Mukhopadhyay and M. Palakal in *Information Retrieval*, Vol. 6, No. 2, pages 199–223; April 2003.

For the URLs of the Web sites referred to in the article, see **www.sciam.com/ontheweb**