When Everything Is Searchable

ERIC A. BREWER

After communication, search is the most visible and important aspect of the Internet. The first major sites were search engines; although they have evolved into portals, search remains fundamental to the vision, purpose, and performance of the network. The Internet's technical (and social) elegance is its anarchy and freedom of information sharing; everyone can add information and express his or her views. Search is the way even the most casual user confronts this chaos and tries to make sense of the billion documents online today and the trillion that are on the way. Thus, instead of trying to organize information—essentially hopeless for humans at such vast scales—all of us manage to use technology to find information, despite its lack of universal organization.

Search is also one of the aspects of the Internet that makes it fundamentally different from previous media; interactivity is the other obvious one. Consider the difference between a mail-order catalog and an e-commerce site. I use the catalog by scanning every page and folding the corners of the pages with the items I might buy; on the Web site, I use search. Both types of exploration can be serendipitous as well, helping find things I didn't know I wanted, but search provides a level of productivity traditional media can't match. In fact, there is a plausible argument that search capabilities in the broadest sense have led to increased overall productivity by millions of workers, and thus to our recent global economic expansion. Anecdotal examples include FedEx tracking numbers, procurement systems, and intranets. Today, I can find and learn things in seconds that I would not have attempted to find five years ago.

In looking at the future of search, I focus on four areas that will affect our lives in the coming decades: integration of textual search and database technologies; distributed repositories; context; and integration with the physical world.

Information retrieval vs. databases. Today's search technology, which derives from research in information retrieval and statistics, primarily involves the ranking of documents by relevance. Many factors determine relevance, but this fundamentally fuzzy game is essentially always the same: statistically combine different facets of relevance information into one number that effectively predicts the relevance of the document to the query.

Databases focus on manipulation of sets; primary operations are joining, sorting, and pruning. Although a search engine can be built using database technology, it is difficult to do well. The main problem is that databases are not good at fuzzy unstructured things, as they are designed for very structured data, such as bank accounts and employee records. Still, most of the interesting data in the world is in databases.

Thus, search must evolve to include this data and, more generally, knowledge of the structure of the data. Researchers must combine the probabilistic techniques of the current generation with the structured queries that make databases so powerful. Creating this combined method will not be easy for a huge set of reasons, the most difficult that the structures vary widely and evolve over time and the lack of agreement on the meaning of even simple terms like "state" and "salary." Many researchers expect the Extensible Markup Language (XML) to solve these issues, even though this metalanguage provides only a common way to describe structure and can't directly help interpret what the structure means.

Distributed repositories. The current generation of search technology barely deals with distributed sources of information. The basic process is to "crawl" the Web and bring all the data from the remote sources into one centralized database from which the queries are handled. It works, mostly, but only for Web pages, not for the large repositories holding most of the interesting data. At some point in the grand scheme of information access, it is better to move the queries to the data in real time than to move all the data to one place. Examples of mov-

ing the queries to the data include "function shipping" in distributed databases, the Federated Facts and Figures site at the University of California, Berkeley (see fff.cs.berkeley.edu), and the search facility for Gnutella (the search facility of the better-known Napster is centralized). However, in practice, getting this approach to work involves solving a number of very difficult issues.

First is the availability and access time of the information. In today's centralized approach, there is enough control to ensure that all the information is available all the time and that it can be searched in the fraction of a second we have (unreasonably) come to expect. With distributed sources, we cannot ensure that the sources are available or even reachable, and even when they are, we can't depend on their performance. Moreover, there is no way to hide these issues from the end user. Users of Napster experience this phenomenon during most hoped-for music downloads; it is fundamental to peer-to-peer systems, which are an extreme form of distributed sources.

Much more difficult to resolve is the issue of trust. Distributed search requires a level of trust that is in general not worthy of that trust. Today, we see this in spam; sites lie about their content, often prompted by economic incentive to do so (such as the need to drive more traffic). Even if the user trusts a set of distributed databases, there is still the fundamental problem of which ones to include for a given query and the relative importance of their responses.

In the end, this is a social problem requiring social solutions, such as branding and reputation. Search engines today use both techniques in small ways. They are designed to blacklist sites or pages that prove untrustworthy and give more coverage and weight to known high-quality (branded) sites, such as the *Wall Street Journal* and the Centers for Disease Control and Prevention. To reflect reputation, we look at the number of pages that link to a given page, the assumption being that if someone takes the time to link to your page, you must have a pretty good page. Hence, these links act as a measurable form of reputation. In the future, reputation will have to be a more direct automated aspect of data sources. Ultimately, it must be in the eye of the beholder, thus personal and not global. The short term will focus on reliable identification of sites (authentication); reputation is farther off.

The power of context. The personal nature of reputation highlights the fact that relevance is itself personal and not the same for everyone. This lack of a singular definition of "relevant" exemplifies a larger problem: that the meaning of a query depends on its context—who, when, and where. The task, today, for even the simplest search is to find the "right" document out of a billion, given only about two words in the query. It's remarkable that we find anything. We expect something, because daily life is filled with an implicit context for every interaction. Search context means knowing a great deal more about the issuer of the query and even simple things like the time of day or location of the issuer. For example, "good restaurants" is a query that depends on all three. The same person in different roles, such as at work or at home, implies different context and thus different criteria for relevance.

> I would love to have a search engine for my office, for the same reasons—anarchy and entropy—that make search critical for the Internet.

We see some use of context in the form of vertical portals, such as health sites, and mobile users, for whom search systems can find local restaurants and stores. But, in general, search engines lack useful information about users or their tasks, have little experience discerning relevance based on contextual information, and, over time, need to encode implicit assumptions in a useful form, as in the CYC project, an early attempt to encode the hundreds of millions of facts and heuristics comprising common sense. Making direct use of context in search systems will take decades but will yield incredible power; different users or even the same user in different situations will get different results, each better than the generic "right" answer today's systems provide.

The physical world. Longer term, researchers will strive to apply the power of search to the physical world. I would love to have a search engine for my

office, for the same reasons—anarchy and entropy—that make search critical for the Internet. Today, we might search for our missing packages, identify the location of an airplane, and find our misplaced cordless phone. But as we deepen the integration of the virtual and physical worlds, we'll extend the power of search. Research at Berkeley and Xerox PARC reveals that we can create interactive labels at low cost using essentially a fancy inkjet printer that prints circuits and that we can build tiny networked sensors and actuators ("smart dust") using integrated-circuit technology. Simple uses include tracking physical objects, labels that refer you to a more detailed Web site, and price/feature comparisons. Books and files would know if they are in the right place, and, in general, inventory management would be simpler and more powerful. I would finally get that search engine for my office.

Fundamental Solution to Chaos

Search is the fundamental solution to anarchy and information chaos. It is search that enables anyone to be a publisher, in turn promoting freedom of information and expression. We are only part way to dealing with this anarchy and its fundamental issues of trust and relative value. Similarly, we have just started to understand how to combine traditional search with the large important repositories of structured data, such as patent and census data. In the longer term, the full expression of the power of search depends on the use of context and on exploiting the continuing integration of the Internet into the physical world. It really is the most exciting and profound time for information sharing in many centuries and promises to be so for a long time to come. C

Eric A. Brewer (brewer@eecs.berkeley.edu) is a professor of computer science at the University of California, Berkeley, and a cofounder and chief scientist of Inktomi Corp., Foster City, CA.

Bandwidth and the Creation of Awareness

MARTIN COOPER

Modern telecommunications is the transmission of electromagnetic signals using technology that ultimately delivers useful information to people or machines in the form of voice, data, and video. But the essence of telecommunications is *content*, not technology or signals. Until now, the properties of this content have been limited by available bandwidth. But wired (or fiber) bandwidth is now abundant and becoming more so. Likewise, wireless bandwidth has effectively doubled every 2.5 years since Guglielmo Marconi received his first wireless telegraphy patent 105 years ago. The same pace of innovation (more than a trillion times) will continue for the next 100 years (an observation I call Cooper's Law). This inexorable growth and the resulting abundance of fixed and untethered bandwidth will set the stage for innovations that will profoundly affect human society for the next 1,000 years.

Thus, when I speak of "the future" in general terms, I'm referring to the changes that will characterize both the coming century and the coming millennium. Other changes—those reflecting the effects of specific technologies—are defined accordingly.

My approach to viewing the coming centuries is to first examine the networks that carry the content and the terminals that transfer it between people and machines in the context of the abundance of bandwidth and the continuing advancement of device technology. That's the easy part. More difficult is predicting the effect of this combination on people, but since neither you nor I will survive its realization in the projected time frame, there's little risk in doing so (other than the ridicule of dis-