Blown to Bits

Your Life, Liberty, and Happiness After the Digital Explosion

Hal Abelson
Ken Ledeen
Harry Lewis

Addison-Wesley

Upper Saddle River, NJ • Boston • Indianapolis • San Francisco
New York • Toronto • Montreal • London • Munich • Paris • Madrid
Cape Town • Sydney • Tokyo • Singapore • Mexico City

Many of the designations used by manufacturers and sellers to distinguish their products are claimed as trademarks. Where those designations appear in this book, and the publisher was aware of a trademark claim, the designations have been printed with initial capital letters or in all capitals.

The authors and publisher have taken care in the preparation of this book, but make no expressed or implied warranty of any kind and assume no responsibility for errors or omissions. No liability is assumed for incidental or consequential damages in connection with or arising out of the use of the information or programs contained herein.

The publisher offers excellent discounts on this book when ordered in quantity for bulk purchases or special sales, which may include electronic versions and/or custom covers and content particular to your business, training goals, marketing focus, and branding interests. For more information, please contact:

U.S. Corporate and Government Sales
(800) 382-3419
corpsales@pearsontechgroup.com

For sales outside the United States, please contact:

International Sales
international@pearson.com

Visit us on the Web: www.informit.com/aw

Library of Congress Cataloging-in-Publication Data:

Abelson, Harold.
Blown to bits : your life, liberty, and happiness after the digital explosion / Hal Abelson, Ken Ledeen, Harry Lewis.
p. cm.
ISBN 0-13-713559-9 (hardback : alk. paper) 1. Computers and civilization. 2. Information technology—Technological innovations. 3. Digital media. I. Ledeen, Ken, 1946- II. Lewis, Harry R. III. Title.
QA76.9.C66A245 2008
303.48'33—dc22

2008005910

For information regarding permissions, write to:

Pearson Education, Inc.
Rights and Contracts Department
501 Boylston Street, Suite 900
Boston, MA 02116
Fax (617) 671 3447

ISBN-13: 978-0-13-713559-2
ISBN-10: 0-13-713559-9
Text printed in the United States on recycled paper at RR Donnelley in Crawfordsville, Indiana.
Third printing December 2008

This Book Is Safari Enabled

The Safari® Enabled icon on the cover of your favorite technology book means the book is available through Safari Bookshelf. When you buy this book, you get free access to the online edition for 45 days.

Safari Bookshelf is an electronic reference library that lets you easily search thousands of technical books, find code samples, download chapters, and access technical information whenever and wherever you need it.

To gain 45-day Safari Enabled access to this book:

- Go to http://www.informit.com/onlineedition
- Complete the brief registration form
- Enter the coupon code 9SD6-IQLD-ZDNI-AGEC-AG6L

If you have difficulty registering on Safari Bookshelf or accessing the online edition, please e-mail customer-service@safaribooksonline.com.

Editor in Chief
Mark Taub

Acquisitions Editor
Greg Doench

Development Editor
Michael Thurston

Managing Editor
Gina Kanouse

Senior Project Editor
Kristy Hart

Copy Editor
Water Crest Publishing, Inc.

Indexer
Erika Millen

Proofreader
Williams Woods Publishing Services

Publishing Coordinator
Michelle Housley

Interior Designer and Composition
Nonie Ratcliff

Cover Designer
Chuti Prasertsith

CHAPTER 4

Needles in the Haystack

Google and Other Brokers in the Bits Bazaar

Found After Seventy Years

Rosalie Polotsky was 10 years old when she waved goodbye to her cousins, Sophia and Ossie, at the Moscow train station in 1937. The two sisters were fleeing the oppression of Soviet Russia to start a new life. Rosalie's family stayed behind. She grew up in Moscow, taught French, married Nariman Berkovich, and raised a family. In 1990, she emigrated to the U.S. and settled near her son, Sasha, in Massachusetts.

Rosalie, Nariman, and Sasha always wondered about the fate of Sophia and Ossie. The Iron Curtain had utterly severed communication among Jewish relatives. By the time Rosalie left for the U.S., her ties to Sophia and Ossie had been broken for so long that she had little hope of reconnecting with them—and, as the years wore on, less reason for optimism that her cousins were still alive. Although his grandfather dreamed of finding them, Sasha's search of immigrant records at Ellis Island and the International Red Cross provided no clues. Perhaps, traveling across wartime Europe, the little girls had never even made it to the U.S.

Then one day, Sasha's cousin typed "Polotsky" into Google's search window and found a clue. An entry on a genealogical web site mentioned "Minacker," the name of Sophia's and Ossie's father. In short order, Rosalie, Sophia, and Ossie were reunited in Florida, after 70 years apart. "All the time when he was alive, he asked me to do something to find them," said Sasha, recalling his grandfather's wish. "It's something magic."

The digital explosion has produced vast quantities of informative data, the Internet has scattered that data across the globe, and the World Wide Web has

put it within reach of millions of ordinary people. But you can't reach for something if you don't know where it is. Most of that vast store of digital information might as well not exist without a way to find it. For most of us, the way to find things on the Web is with web search engines. Search is a wondrous, transformative technology, which both fulfills dreams and shapes human knowledge. The search tools that help us find needles in the digital haystack have become the lenses through which we view the digital landscape. Businesses and governments use them to distort our picture of reality.

The Library and the Bazaar

In the beginning, the Web was a library. Information providers—mostly businesses and universities, which could afford to create web pages—posted information for others to see. Information consumers—mostly others in business and academia—found out where to get the information and downloaded it. They might know where to look because someone sent them the URL (the "Uniform Resource Locator"), such as `mit.edu` (the URL for MIT). Ordinary people didn't use the Web. Instead, they used services such as CompuServe for organized access to databases of various kinds of information.

As the Web went commercial, directories began to appear, including printed "Yellow Pages." These directories listed places to go on the Web for various products and services. If you wanted to buy a car, you looked in one place, and you looked in another place to find a job. These lists resembled the categories AOL and CompuServe provided in the days before consumers could connect directly to the Internet. Human beings constructed these lists—editors decided what went in each category, and what got left out entirely.

The Web has changed drastically since the mid-1990s. First, it is no

WEB 1.0 VS. WEB 2.0

In contemporary jargon, the newer, more participatory web sites to which users can contribute are dubbed "Web 2.0." The older, more passive web sites are now called "Web 1.0." These look like software release numbers, but "Web 2.0" describes something subtler and more complex. Web 2.0 sites—Facebook and Wikipedia, for example—exploit what economists call "network effects." Because users are contributing information as well as utilizing information others supply, these sites become more valuable the more people are using them. See `http://www.oreillynet.com/lpt/a/6228` for a fuller explanation of Web 2.0.

longer a passive information resource. Blogs, Wikipedia, and Facebook are contributory structures, where peer involvement makes the information useful. Web sites are cheap and easy to create; ordinary individuals and even the smallest of organizations can now have them. As a result, the content and connectedness of the Web are changing all the time.

Second, the Web has gotten so big and so unstructured that it is not humanly possible to split it up into neat categories. Web pages simply don't lend themselves to organization in a nice structure, like an outline. There is no master plan for the Web—vast numbers of new pages are added daily in an utterly unstructured way. You certainly can't tell what a web page contains by looking at its URL.

Moreover, hierarchical organization is useless in helping you find information if you can't tell where in the hierarchy it might belong. You don't usually go to the Web to look for a web page. You go to look for *information*, and are glad to get it wherever you can find it. Often, you can't even guess where to look for what you want to know, and a nice, structured organization of knowledge would do you no good. For example, any sensible organization of human knowledge, such as an encyclopedia, would have a section on cows and a section on the moon. But if you didn't know that there was a nursery rhyme about the cow jumping over the moon, neither the "cow" nor the "moon" entry would help you figure out what the cow supposedly did to the moon. If you typed both words into a search engine, however, you would find out in the blink of an eye.

Search is the new paradigm for finding information—and not just on the Web as a whole. If you go to Wal-Mart's web site, you can trace through its hierarchical organization. At the top level, you get to choose between "accessories," "baby," "boys," "girls," and so on. If you click "baby," your next click takes you to "infant boys," "toddler girls," and so on. There is also a search window at the top. Type whatever you want, and you may be taken directly to what you are looking for—but only on Wal-Mart's site. Such limited search engines help us share photos, read newspapers, buy books online from Amazon or Barnes and Noble, and even find old email on our own laptops.

Search makes it possible to find things in vast digital repositories. But search is more than a quick form of look-up in a digital library. *Search is a new form of control over information.*

Search is a new form of control over information.

Information retrieval tools such as Google are extraordinarily democratizing—Rosalie and Sasha Berkovich did not need to hire a professional people-finder. But the power that has been vested in individuals is not the only kind

that search has created. We have given search engines control over where we get reliable information—the same control we used to assign to authoritative sources, such as encyclopedias and "newspapers of record." If we place absolute trust in a search engine to find things for us, we are giving the search engine the power to make it hard or impossible for us to know things. Use Google in China, and your searches will "find" very different information about democracy than they will "find" if you use Google in the United States. Search for "401(K)" on John Hancock's web site, and Fidelity's 401(K) plans will seem not to exist.

For the user, search is the power to find things, and for whoever controls the engine, search is the power to shape what you see. Search is also power of a third kind. Because the search company records all our search queries, we are giving the search company the power that comes with knowing what we want to know. In its annual "Zeitgeist" report, Google takes the pulse of the population by revealing the questions its search engine is most often asked. It was amusing to know that of the most popular "Who is ...?" searches of 2007, "God" was #1 and "Satan" was #10, with "Buckethead" beating "Satan" at #6. Search engines also gather similar information about each one of us individually. For example, as discussed in Chapter 2, Amazon uses the information to suggest books you might like to read once you have used its web site for a bit.

Here are some interesting Google Zeitgeist results from 2007: among "What is" questions, "love" was #1 and "gout" was #10; among "How to" queries, "kiss" was #1 and "skateboard" was #10.

The Web is no longer a library. It is a chaotic marketplace of the billions of ideas and facts cast up by the bits explosion. Information consumers and information producers constantly seek out each other and morph into each other's roles. In this shadowy bits bazaar, with all its whispers and its couriers running to and fro, search engines are brokers. Their job is not to supply the undisputed truth, nor even to judge the accuracy of material that others provide. Search engines connect willing producers of information to willing consumers. They succeed or fail not on the quality of the information they provide, because they do not produce content at all. They only make connections. Search engines succeed or fail depending on whether we are happy with the connections they make, and nothing more. In the bazaar, it is not always the knowledgeable broker who makes the most deals. To stay in business, a broker just has to give most people what they want, consistently over time.

Search does more than find things for us. Search helps us discover things we did not know existed. By searching, we can all be armchair bits detectives,

finding surprises in the book *next* to the one we were pulling off the digital bookshelf, and sniffing out curious information fragments cast far and wide by the digital explosion.

Forbidden Knowledge Is Only a Click Away

Schizophrenia is a terrible brain disease, afflicting millions of people. If you wanted to know about the latest treatment options, you might try to find some web sites and read the information they contain.

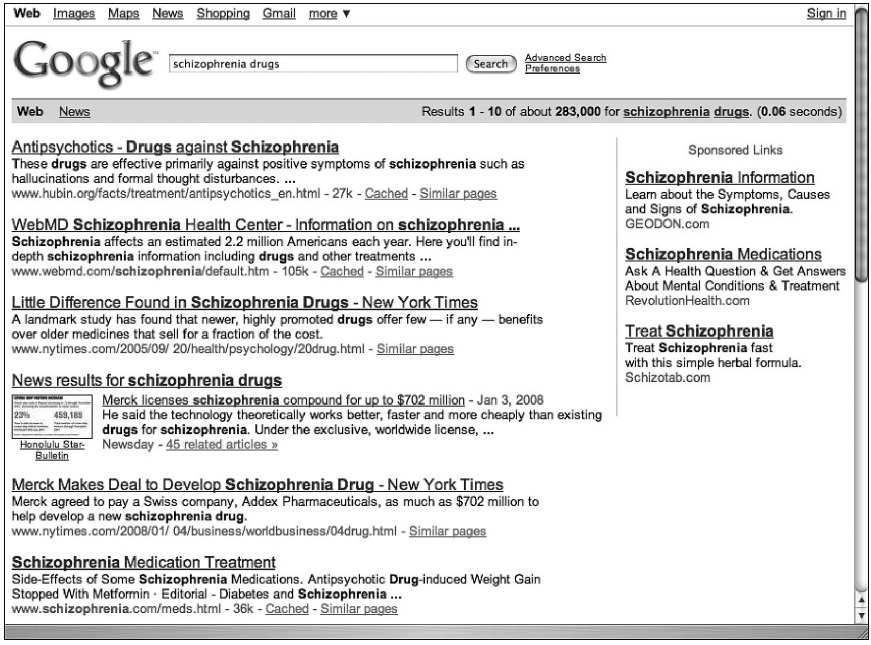
Some people already know where they think they can good find medical information—they have bookmarked a site they trust, such as `WebMD.com` or `DrKoop.com`. If you were like us, however, you'd use a search engine—`Google.com`, `Yahoo.com`, or `Ask.com`, for example. You'd type in a description of what you were looking for and start to click links and read. Of course, you should *not* believe uncritically anything you read from a source you don't know anything about—or act on the medical information you got through your browsing, without checking with a physician.

Britney in the Bits Bazaar

Providing what most people want creates a tyranny of the majority and a bias against minority interests. When we searched for "spears," for example, we got back three pages of results about Britney Spears and her sister, with only three exceptions: a link to Spears Manufacturing, which produces PVC piping; one to comedian Aries Spears; and one to Prof. William M. Spears of the University of Wyoming. Ironically, Prof. Spears's web page ranked far below "Britney Spears' Guide to Semiconductor Physics," a site maintained by some light-hearted physicists at the University of Essex in the UK. That site has a distinctive URL, `britneyspears.ac`—where "`.ac`" stands not for "academic" but for "Ascension Island" (which gets a few pennies for use of the `.ac` URL, wherever in the world the site may be hosted). Whatever the precise reason for this site's high ranking, the association with Britney probably didn't hurt!

When we tried searching for "schizophrenia drugs" using Google, we got the results shown in Figure 4.1. The top line tells us that if we don't like these results, there are a quarter-million more that Google would be glad to show us. It also says that it took six-hundredths of a second to get these results for us—we didn't sense that it took even that long. Three "Sponsored Links" appear to the right. A link is "sponsored" if someone has paid Google to have it put there—in other words, it's an advertisement. To the left is a variety of ordinary links that Google's information retrieval algorithms decided were

most likely to be useful to someone wanting information about "schizophrenia drugs." Those ordinary links are called the search engine's *organic* results, as opposed to the sponsored results.



Google ™ is a registered trademark of Google, Inc. Reprinted by permission.

FIGURE 4.1 Google's results from a search for "schizophrenia drugs."

THOSE FUNNY NAMES

Yahoo! is an acronym—it stands for "Yet Another Hierarchical Officious Oracle" (`docs.yahoo.com/info/misc/history.html`). "Google" comes from "googol," which is the number represented by a 1 followed by 100 zeroes. The Google founders were evidently thinking big!

Just looking at this window raises a series of important questions:

- The Web is enormous. How can a search engine find those results so fast? Is it finding every appropriate link?
- How did Google decide what is search result number 1 and what is number 283,000?
- If you try another search engine instead of Google, you'll get different results. Which is right? Which is better? Which is more authoritative?

- Are the sponsored links supposed to be better links than the organic links, or worse? Is the advertising really necessary?
- How much of this does the government oversee? If a TV station kept reporting lies as the truth, the government would get after them. Does it do anything with search engines?

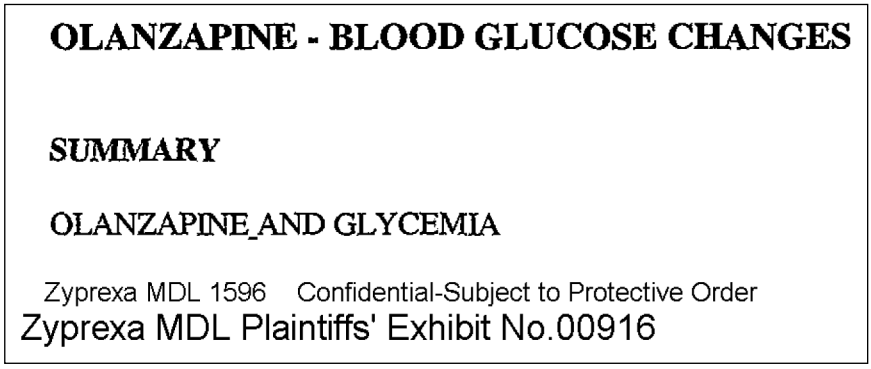
We shall take up each of these questions in due course, but for the time being, let's just pursue our medical adventure.

When we clicked on the first organic link, it took us to a page from the web site of a distinguished Swedish university. That page contained some information about the different kinds of schizophrenia drugs. One of the drugs it mentioned was "olanzapin (Zyprexa)." The trade name rang a bell for some reason, so we started over and searched for "Zyprexa."

The first of the organic links we got back was to `www.zyprexa.com`, which described itself as "The Official ZYPREXA Olanzapine Site." The page was clearly marked as maintained by Eli Lilly and Company, the drug's manufacturer. It provided a great deal of information about the drug, as well as photographs of smiling people—satisfied patients, presumably—and slogans such as "There is Hope" and "Opening the Door to Possibility." The next few links on our page of search results were to the medical information sites drugs.com, rxlist.com, webmd.com, and askapatient.com.

Just below these was a link that took us in a different direction: "ZyprexaKills wiki." The drug was associated with some serious side effects, it seems, and Lilly allegedly kept these side effects secret for a long time. At the very top of that page of search results, as the only sponsored link, was the following: "Prescription Drug Lawsuit. Zyprexa-olanzapine-lawyer.com. Pancreatitis & diabetes caused by this drug? Get legal help today." That link took us to a web form where a Houston attorney offered to represent us against Lilly.

It took only a few more mouse clicks before a document appeared that was entitled "Olanzapine—Blood glucose changes" (see Figure 4.2). It was an internal Lilly memorandum, never meant to be seen outside the company, and marked as a confidential exhibit in a court case. Some patients who had developed diabetes while using Zyprexa had sued Lilly, claiming that the drug had caused the disease. In the course of that lawsuit, this memo and other confidential materials were shared with the plaintiffs' lawyers under a standard discovery protocol. Through a series of improper actions by several lawyers, a *New York Times* reporter procured these documents. The reporter then published an exposé of Lilly's slowness to acknowledge the drug's side effects. The documents themselves appeared on a variety of web sites.



Source: www.furiousseasons.com/zyprexa%20documents/ZY1%20%20%2000008758.pdf.

FIGURE 4.2 Top and bottom lines of a document filed in a court case. It was supposed to be kept secret, but once on the Web, anyone searching for "Zyprexa documents" finds it easily.

Lilly demanded that the documents be returned, that all copies be destroyed, and that the web sites that had posted them be required to take them down. A legal battle ensued. On February 13, 2007, Judge Jack B. Weinstein of the U.S. District Court in New York issued his judgment, order, and injunction. Yes, what had been done with the documents was grievously wrong and contrary to earlier court orders. The lawyers and the journalist had cooked up a scam on the legal system, involving collusion with an Alaska lawyer who had nothing to do with the case, in order to spring the documents. The lawyers who conspired to get the documents had to give them back and not keep any copies. They were enjoined against giving any copies to anyone else.

But, concluded Judge Weinstein, the web sites were another matter. The judge would not order the web sites to take down their copies. Lilly was entitled to the paper documents, but the bits had escaped and could not be recaptured. As of this writing, the documents are still viewable. We quickly found them directly by searching for "zyprexa documents."

The world is a different place from a time when the judge could have ordered the return of *all* copies of offending materials. Even if there were hundreds of copies in file cabinets and desk drawers, he might have been able to insist on their return, under threat of harsh penalties. But the Web is not a file cabinet or a desk drawer. "Web sites," wrote Judge Weinstein, "are primarily fora for speech." Lilly had asked for an injunction against five web sites that had posted the documents, but millions of others could post them in the future. "Limiting the fora available to would-be disseminators by such an

infinitesimal percentage would be a fruitless exercise," the judge concluded. It probably would not be effective to issue a broader injunction, and even if it were, "the risk of unlimited inhibitions of free speech should be avoided when practicable."

The judge understood the gravity of the issue he was deciding. Fundamentally, he was reluctant to use the authority of the government in a futile attempt to prevent people from saying what they wanted to say and finding out what they wanted to know. Even if the documents had been visible only for a short time period, unknown numbers of copies might be circulating privately among interested parties. Grasping for an analogy, the judge suggested that God Himself had failed in His attempt to enjoin Adam and Eve from their pursuit of the truth!

Two sponsored links appeared when we did the search for "zyprexa documents." One was for another lawyer offering his services for Zyprexa-related lawsuits against Lilly. The other, triggered by the word "documents" in our search term, was for Google itself: "Online Documents. Easily share & edit documents online for free. Learn more today. docs.google.com." This was an ironic reminder that the bits are out there, and the tools to spread them are there too, for anyone to use. Thanks to search engines, anyone can find the information they want. Information has exploded out of the shells that used to contain it.

In fact, the architecture of human knowledge has changed as a result of search. In a single decade, we have been liberated from information straightjackets that have been with us since the dawn of recorded history. And many who should understand what has happened, do not. In February 2008, a San Francisco judge tried to shut down the Wikileaks web site, which posts leaked confidential documents anonymously as an aid to whistleblowers. The judge ordered the name "Wikileaks" removed from DNS servers, so the URL "`Wikileaks.org`" would no longer correspond to the correct IP address. (In the guts of the Internet, DNS servers provide the service of translating URLs into IP addresses. See the Appendix.) The publicity that resulted from this censorship attempt made it easy to find various "mirrors"—identical twins, located elsewhere on the Web—by searching for "Wikileaks."

The Fall of Hierarchy

For a very long time, people have been organizing things by putting them into categories and dividing those categories into subcategories. Aristotle tried to classify everything. Living things, for example, were either plants or animals. Animals either had red blood or did not; red-blooded animals were

either live-bearers or egg-bearers; live-bearers were either humans or other mammals; egg-bearers either swam or flew; and so on. Sponges, bats, and whales all presented classification enigmas, on which Aristotle did not think he had the last word. At the dawn of the Enlightenment, Linnaeus provided a more useful way of classifying living things, using an approach that gained intrinsic scientific validity once it reflected evolutionary lines of descent.

Our traditions of hierarchical classification are evident everywhere. We just love outline structures. The law against cracking copyright protection (discussed in Chapter 6, "Balance Toppled") is Title 17, Section 1201, paragraph (a), part (1), subpart (A). In the Library of Congress system, every book is in one of 26 major categories, designated by a Roman letter, and these major categories are internally divided in a similar way—B is philosophy, for example, and BQ is Buddhism.

If the categories are clear, it may be possible to use the *organizing* hierarchy to *locate* what you are looking for. That requires that the person doing the searching not only know the classification system, but be skilled at making all the necessary decisions. For example, if knowledge about living things was organized as Aristotle had it, anyone wanting to know about whales would have to know *already* whether a whale was a fish or a mammal in order to go down the proper branch of the classification tree. As more and more knowledge has to be stuffed into the tree, the tree grows and sprouts twigs, which over time become branches sprouting more twigs. The classification problem becomes unwieldy, and the retrieval problem becomes practically impossible.

The system of Web URLs started out as such a classification tree. The site `www.physics.harvard.edu` is a web server, of the physics department, within Harvard University, which is an educational institution. But with the profusion of the Web, this system of domain names is now useless as a way of finding anything whose URL you do not already know.

In 1991, when the Internet was barely known outside academic and government circles, some academic researchers offered a program called "Gopher." This program provided a hierarchical directory of many web sites, by organizing the directories provided by the individual sites into one big outline. Finding things using Gopher was tedious by today's standards, and was dependent on the organizational skills of the contributors. Yahoo! was founded in 1994 as an online Internet directory, with human editors placing products and services in categories,

"Gopher" was a pun—it was software you could use to "go for" information on the Web. It was also the mascot of the University of Minnesota, where the software was first developed.

making recommendations, and generally trying to make the Internet accessible to non-techies. Although Yahoo! has long since added a search window, it retains its basic directory function to the present day.

The practical limitations of hierarchical organization trees were foreseen sixty years ago. During World War II, President Franklin Roosevelt appointed Vannevar Bush of MIT to serve as Director of the Office of Strategic Research and Development (OSRD). The OSRD coordinated scientific research in support of the war effort. It was a large effort—30,000 people and hundreds of projects covered the spectrum of science and engineering. The Manhattan Project, which produced the atomic bomb, was just a small piece of it.

From this vantage point, Bush saw a major obstacle to continued scientific progress. We were producing information faster than it could be consumed, or even classified. Decades before computers became commonplace, he wrote about this problem in a visionary article, "As We May Think." It appeared in the *Atlantic Monthly*—a popular magazine, not a technical journal. As Bush saw it,

> The difficulty seems to be, not so much that we publish unduly ... but rather that publication has been extended far beyond our present ability to make real use of the record. The summation of human experience is being expanded at a prodigious rate, and the means we use for threading through the consequent maze to the momentarily important item is the same as was used in the days of square-rigged ships. ... Our ineptitude in getting at the record is largely caused by the artificiality of systems of indexing.

The dawn of the digital era was at this time barely a glimmer on the horizon. But Bush imagined a machine, which he called a "memex," that would augment human memory by storing and retrieving all the information needed. It would be an "enlarged intimate supplement" to human memory, which can be "consulted with exceeding speed and flexibility."

Bush clearly perceived the problem, but the technologies available at the time, microfilm and vacuum tubes, could not solve it. He understood that the problem of finding information would eventually overwhelm the progress of science in creating and recording knowledge. Bush was intensely aware that civilization itself had been imperiled in the war, but thought we must proceed with optimism about what the record of our vast knowledge might bring us. Man "may perish in conflict before he learns to wield that record for his true good. Yet, in the application of science to the needs and desires of man, it would seem to be a singularly unfortunate stage at which to terminate the process, or to lose hope as to the outcome."

A Futurist Precedent

In 1937, H. G. Wells anticipated Vannevar Bush's 1945 vision of a "memex." Wells wrote even more clearly about the possibility of indexing everything, and what that would mean for civilization:

> *There is no practical obstacle whatever now to the creation of an efficient index to all human knowledge, ideas and achievements, to the creation, that is, of a complete planetary memory for all mankind. And not simply an index; the direct reproduction of the thing itself can be summoned to any properly prepared spot. ... This in itself is a fact of tremendous significance. It foreshadows a real intellectual unification of our race. The whole human memory can be, and probably in a short time will be, made accessible to every individual. ... This is no remote dream, no fantasy.*

Capabilities that were inconceivable then are commonplace now. Digital computers, vast storage, and high-speed networks make information search and retrieval necessary. They also make it possible. The Web is a realization of Bush's memex, and search is key to making it useful.

It Matters How It Works

How can Google or Yahoo! possibly take a question it may never have been asked before and, in a split second, deliver results from machines around the world? The search engine doesn't "search" the entire World Wide Web in response to your question. That couldn't possibly work quickly enough—it would take more than a tenth of a second just for bits to move around the earth at the speed of light. Instead, the search engine has *already* built up an index of web sites. The search engine does the best it can to find an answer to your query using its index, and then sends its answer right back to you.

To avoid suggesting that there is anything unique about Google or Yahoo!, let's name our generic search engine Jen. Jen integrates several different processes to create the illusion that you simply ask her a question and she gives back good answers. The first three steps have nothing to do with your particular query. They are going on repeatedly and all the time, whether anyone is posing any queries or not. In computer speak, these steps are happening in the *background*:

1. **Gather information.** Jen explores the Web, visiting many sites on a regular basis to learn what they contain. Jen revisits old pages because their contents may have changed, and they may contain links to new pages that have never been visited.
2. **Keep copies.** Jen retains copies of many of the web pages she visits. Jen actually has a duplicate copy of a large part of the Web stored on her computers.
3. **Build an index.** Jen constructs a huge index that shows, at a minimum, which words appear on which web pages.

When you make a query, Jen goes through four more steps, in the *foreground*:

4. **Understand the query.** English has lots of ambiguities. A query like "red sox pitchers" is fairly challenging if you haven't grown up with baseball!
5. **Determine the relevance of each possible result to the query.** Does the web page contain information the query asks about?
6. **Determine the ranking of the relevant results.** Of all the relevant answers, which are the "best"?
7. **Present the results.** The results need not only to be "good"; they have to be shown to you in a form you find useful, and perhaps also in a form that serves some of Jen's other purposes—selling more advertising, for example.

Each of these seven steps involves technical challenges that computer scientists love to solve. Jen's financial backers hope that her engineers solve them better than the engineers of competing search engines.

We'll go through each step in more detail, as it is important to understand what is going on—at every step, more than technology is involved. Each step also presents opportunities for Jen to use her information-gathering and editorial powers in ways you may not have expected—ways that shape your view of the world through the lens of Jen's search results.

The background processing is like the set-building and rehearsals for a theatrical production. You couldn't have a show without it, but none of it happens while the audience is watching, and it doesn't even need to happen on any particular schedule.

Step 1: Gather Information

Search engines don't index everything. The ones we think of as general utilities, such as Google, Yahoo!, and Ask, find information rather indiscriminately throughout the Web. Other search engines are domain-specific. For example, Medline searches only through medical literature. ArtCylopedia indexes 2,600 art sites. The FindLaw LawCrawler searches only legal web sites. Right from the start, with any search engine, some things are in the index and some are out, because some sites are visited during the gathering step and others are not. Someone decides what is worth remembering and what isn't. If something is left out in Step 1, there is no possibility that you will see it in Step 7.

Speaking to the Association of National Advertisers in October 2005, Eric Schmidt, Google's CEO, observed that of the 5,000 terabytes of information in the world, only 170 terabytes had been indexed. (A *terabyte* is about a trillion bytes.) That's just a bit more than 3%, so 97% was not included. Another estimate puts the amount of indexed information at only .02% of the size of the databases and documents reachable via the Web. Even in the limited context of the World Wide Web, Jen needs to decide what to look at, and how frequently. These decisions implicitly define what is important and what is not, and will limit what Jen's users can find.

How *often* Jen visits web pages to index them is one of her precious trade secrets. She probably pays daily visits to news sites such as `CNN.com`, so that if you ask tonight about something that happened this morning, Jen may point you to CNN's story. In fact, there is most likely a master list of sites to be visited frequently, such as `whitehouse.gov`—sites that change regularly and are the object of much public interest. On the other hand, Jen probably has learned from her repeated visits that some sites don't change at all. For example, the Web version of a paper published ten years ago doesn't change. After a few visits, Jen may decide to revisit it once a year, just in case. Other pages may not be posted long enough to get indexed at all. If you post a futon for sale on `Craigslist.com`, the ad will become accessible to potential buyers in just a few minutes. If it sells quickly, however, Jen may never see it. Even if the ad stays up for a while, you probably won't be able to find it with most search engines for several days.

Jen is clever about how often she revisits pages—but her cleverness also codifies some judgments, some priorities—some *control*. The more important Jen judges your page to be, the less time it will take for your new content to show up as responses to queries to Jen's search engine.

Jen roams the Web to gather information by following links from the pages she visits. Software that crawls around the Web is (in typical geek

How a Spider Explores the Web

Search engines gather information by wandering through the World Wide Web. For example, when a spider visits the main URL of the publisher of this book, `www.pearson.com`, it retrieves a page of text, of which this is a fragment:

```
<div id="subsidiary">
<h2 class="hide">Subsidiary sites links</h2>
<label for="subsidiarySites" class="hide">Available
sites</label>
<select name="subsidiarySites" id="subsidiarySites" size="1">
<option value="">Browse sites</option>
<optgroup label="FT Group">
<option value="http://www.ftchinese.com/sc/index.jsp">
  Chinese.FT.com</option>
<option value="http://ftd.de/">FT Deutschland</option>
```

This text is actually a computer program written in a special programming language called HTML ("HyperText Markup Language"). Your web browser renders the web page by executing this little program. But the spider is retrieving this text not to render it, but to index the information it contains. "FT Deutschland" is text that appears on the screen when the page is rendered; such terms should go into the index. The spider recognizes other links, such as `www.ftchinese.com` or `ftd.de`, as URLs of pages it needs to visit in turn. In the process of visiting those pages, it indexes them and identifies yet more links to visit, and so on!

A spider, or web crawler, is a particular kind of *bot*. A bot (as in "robot") is a program that endlessly performs some intrinsically repetitive task, often an information-gathering task.

irony) called a "spider." Because the spidering process takes days or even weeks, Jen will not know immediately if a web page is taken down—she will find out only when her spider next visits the place where it used to be. At that point, she will remove it from her index, but in the meantime, she may respond to queries with links to pages that no longer exist. Click on such a link, and you will get a message such as "Page not found" or "Can't find the server."

Because the Web is unstructured, there is no inherently "correct" order in which to visit the pages, and no obvious way to know when to stop. Page A may contain references to page B, and also page B to page A, so the spider has to be careful not to go around in circles. Jen must organize her crawl of

the Web to visit as much as she chooses without wasting time revisiting sections she has already seen.

A web site may stipulate that it does not want spiders to visit it too frequently or to index certain kinds of information. The site's designer simply puts that information in a file named robots.txt, and virtually all web-crawling software will respect what it says. Of course, pages that are inaccessible without a login cannot be crawled at all. So, the results from Step 7 may be influenced by what the sites want Jen to know about them, as well as by what Jen thinks is worth knowing. For example, Sasha Berkovich was fortunate that the Polotsky family tree had been posted to part of the genealogy.com web site that was open to the public—otherwise, Google's spider could not have indexed it.

Finally, spidering is not cost free. Jen's "visits" are really requests to web sites that they send their pages back to her. Spidering creates Internet traffic and also imposes a load on the web server. This part of search engines' background processing, in other words, has unintended effects on the experience of the entire Internet. Spiders consume network bandwidth, and they may tie up servers, which are busy responding to spider requests while their ordinary users are trying to view their pages. Commercial search engines attempt to schedule their web crawling in ways that won't overload the servers they visit.

Step 2: Keep Copies

Jen downloads a copy of every web page her spider visits—this is what it means to "visit" a page. Instead of rendering the page on the screen as a web browser would, Jen indexes it. If she wishes, she can retain the copy after she has finished indexing it, storing it on her own disks. Such a copy is said to be "cached," after the French word for "hidden." Ordinarily Jen would not do anything with her cached copy; it may quickly become out of date. But caching web pages makes it possible for Jen to have a page that no longer exists at its original source, or a version of a page older than the current one. This is the flip side of Jen never knowing about certain pages because their owners took them down before she had a chance to index them. With a cached page, Jen knows what used to be on the page even after the owner intended it to disappear.

Caching is another blow to the Web-as-library metaphor, because removing information from the bookshelf doesn't necessarily get rid of it. Efforts to scrub even dangerous information are beyond the capability of those who posted it. For example, after 9/11, a lot of information that was once available on the Web was pulled. Among the pages that disappeared overnight

were reports on government vulnerabilities, sensitive security information, and even a Center for Disease Control chemical terrorism report that revealed industry shortcomings. Because the pages had been cached, however, the bits lived on at Google and other search engine companies.

Not only did those pages of dangerous information survive, but anyone could find them. Anytime you do a search with one of the major search engines, you are offered access to the cached copy, as well as the link to where the page came from, whether or not it still exists. Click on the link for the "Cached" page, and you see something that looks very much like what you might see if you clicked on the main link instead. The cached copy is identified as such (see Figure 4.3).

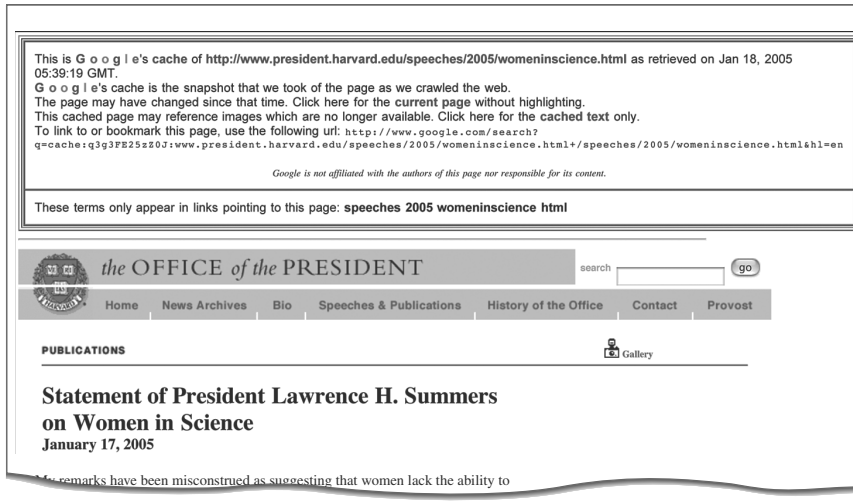


FIGURE 4.3 Part of a cached web page, Google's copy of an official statement made by Harvard's president and replaced two days later after negative public reaction. This copy was retrieved from Google after the statement disappeared from the university's web site. Harvard, which holds the copyright on this once-public statement, refused to allow it to be printed in this book (see Conclusion).

This is an actual example; it was the statement Lawrence Summers released on January 17, 2005, after word of his remarks about women in science became public. As reported in *Harvard Magazine* in March–April 2005, the statement began, "My remarks have been misconstrued as suggesting that women lack the ability to succeed at the highest levels of math and science. I did not say that, nor do I believe it." This unapologetic denial stayed on the

The digital explosion grants the power of both instant communication and instant retraction—but almost every digital action leaves digital fingerprints.

Harvard web site for only a few days. In the face of a national firestorm of protest, Summers issued a new statement on January 19, 2005, reading, in part, "I deeply regret the impact of my comments and apologize for not having weighed them more carefully." Those searching for the President's statement were then led to the contrite new statement—but for a time, the original, defiant version remained visible to those who clicked on the link to Google's cached copy.

Finding Deleted Pages

An easy experiment on finding deleted pages is to search using Google for an item that was sold on craigslist. You can use the "site" modifier in the Google search box to limit your search to the craigslist web site, by including a "modifier":

```
futon site:craigslist.com
```

The results will likely return pages for items that are no longer available, but for which the cached pages will still exist.

The digital explosion grants the power of both instant communication and instant retraction—but almost every digital action leaves digital fingerprints. Bits do not die easily, and digital words, once said, are hard to retract.

If Jen caches web pages, it may be possible for you to get information that was retracted after it was discovered to be in error or embarrassing. Something about this doesn't feel quite right, though—is the information on those pages really Jen's to do with as she wishes? If the material is copyrighted—a published paper from ten years ago, for example—what right does Jen have to show you her cached copy? For that matter, what right did she have to keep a copy in the first place? If you have copyrighted something, don't you have some authority over who can make copies of it?

This enigma is an early introduction to the confused state of copyright law in the digital era, to which we return in Chapter 6. Jen cannot index my web page without receiving a copy of it. In the most literal sense, any time you "view" or "visit" a web page, you are actually copying it, and then your web browser renders the copy on the screen. A metaphorical failure once again: The Web is *not a library*. Viewing is an exchange of bits, not a passive activity, as far as the web site is concerned. If "copying" copyrighted materials was totally prohibited, neither search engines nor the Web itself could work, so some sort of copying must be permissible. On the other hand, when Jen caches the material she indexes—perhaps an entire book, in the case of the

Google Books project—the legal controversies become more highly contested. Indeed, as we discuss in Chapter 6, the Association of American Publishers and Google are locked in a lawsuit over what Google is and is not allowed to do with the digital images of books that Google has scanned.

Step 3: Build an Index

When we searched the Web for "Zyprexa," Jen consulted her index, which has the same basic structure as the index of a book: a list of terms followed by the places they occur. Just as a book's index lists page numbers, Jen's index lists URLs of web pages. To help the search engine give the most useful responses to queries, the index may record other information as well: the size of the font in which the term appears, for example, and where on the page it appears.

INDEXES AND CONCORDANCES

The information structure used by search engines is technically known as an *inverted index*—that is, an index of the words in a document or a set of documents, and the places where those words appear. Inverted indexes are not a new idea; the biblical concordances laboriously constructed by medieval monks were inverted indexes. Constructing concordances was one of the earliest applications of computer technology to a nonmathematical problem.

Indexes are critical because having the index in order—like the index of a book, which is in alphabetical order—makes it possible to find things much faster than with sequential searching. This is where Jen's computer scientists really earn their salaries, by devising clever ways of storing indexed information so it can be retrieved quickly. Moore's Law also played a big role in the creation of web indexes—until computer memories got fast enough, cheap enough, and big enough, even the cleverest computer scientists could not program machines to respond instantly to arbitrary English queries.

When Jen wants to find a term in her index, she does not start at the beginning and go through it one entry at a time until she finds what she is looking for. That is not the way you would look up something in the index of a book; you would use the fact that the index is in order alphabetically. A very simple strategy to look up something in a big ordered index, such as a phone book, is just to open the book in the middle and see if the item you are looking for belongs in the first half or the second. Then you can ignore half the phone book and use the same strategy to subdivide the remaining half. The number of steps it takes to get down to a single page in a phone book with n pages using this method is the number of times you have to divide n by 2 to get down to 1. So if n is 1000, it takes only 10 of these probing steps to find any item using *binary search*, as this method is known.

In general, the number of steps needed to search an index of *n* things using binary search is proportional, not to *n*, but to the number of digits in *n*. That means that binary search is exponentially faster than linear search—searching through a million items would take only 20 steps, and through a billion items would take 30 steps. And binary search is fairly dumb by comparison with what people actually do—if you were looking for "Ledeen" in the phone book, you might open it in the middle, but if you were looking for "Abelson," you'd open it near the front. That strategy can be reduced to an even better computer algorithm, exponentially faster than binary search.

How big is Jen's index, in fact? To begin with, how many terms does Jen index? That is another of her trade secrets. Jen's index could be useful with a few tens of millions of entries. There are fewer than half a million words in the English language, but Jen probably wants to index some numbers too (try searching for a number such as 327 using your search engine). Proper names and at least some words in foreign languages are also important. The list of web pages associated with a term is probably on disk in most cases, with only the information about *where* on the disk kept with the term itself in main memory. Even if storing the term and the location on disk of the list of associated URLs takes 100 bytes per entry, with 25 million entries, the table of index entries would occupy 2.5 gigabytes (about 2.5 billion bytes) of main memory. A few years ago, that amount of memory was unimaginable; today, you get that on a laptop from Wal-Mart. The index can be searched quickly—using binary search, for example—although retrieving the list of URLs might require going to disk. If Jen has Google's resources, she can speed up her query response by keeping URLs in main memory too, and she can split the search process across multiple computers to make it even faster.

Now that the preparations have been made, we can watch the performance itself—what happens when you give Jen a query.

Step 4: Understand the Query

When we asked Google the query *Yankees beat Red Sox*, only one of the top five results was about the Yankees beating the Red Sox (see Figure 4.4). The others reported instead on the Red Sox beating the Yankees. Because English is hard for computers to understand and is often ambiguous, the simplest form of query analysis ignores syntax, and treats the query as simply a list of keywords. Just looking up a series of words in an index is computationally easy, even if it often misses the intended meaning of the query.

To help users reduce the ambiguity of their keyword queries, search engines support "advanced queries" with more powerful features. Even the simplest, putting a phrase in quotes, is used by fewer than 10% of search

engine users. Typing the quotation marks in the query "Red Sox beat Yankees" produces more appropriate results. You can use "~" to tell Google to find synonyms, "-" to exclude certain terms, or cryptic commands such as "allinurl:" or "inanchor:" to limit the part of the Web to search. Arguably we didn't ask our question the right way, but most of us don't bother; in general, people just type in the words they want and take the answers they get.

Often they get back quite a lot. Ask Yahoo! for the words "allergy" and "treatment," and you find more than 20,000,000 references. If you ask for "allergy treatment"—that is, if you just put quotes around the two words—you get 628,000 entries, and quite different top choices. If you ask for "treating allergies," the list shrinks to 95,000. The difference between these queries may have been unintentional, but the search engine thought they were drastically different. It's remarkable that human-computer communication through the lens of the search engine is so useful, given its obvious imperfections!

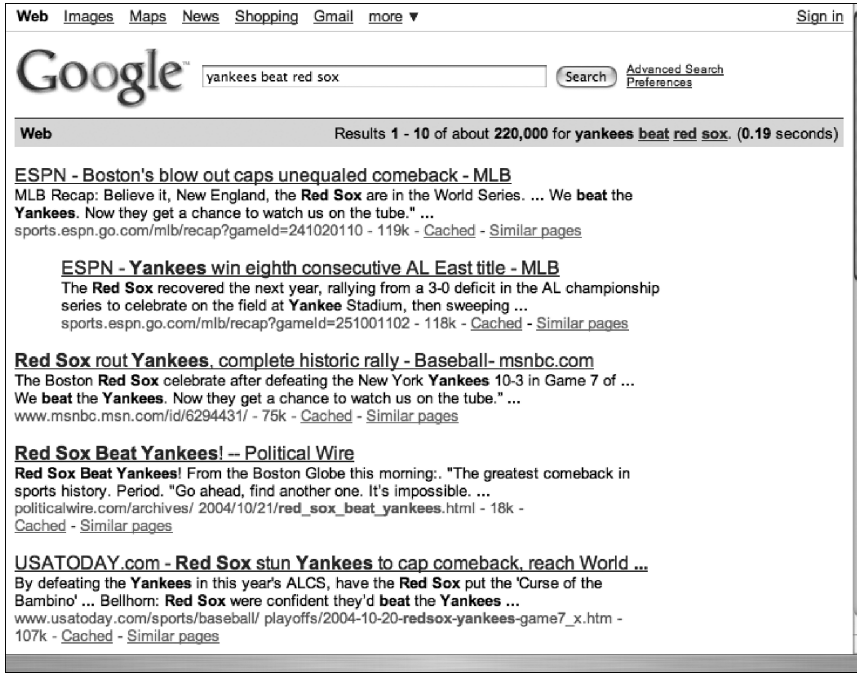


FIGURE 4.4 Keyword search misses the meaning of English-language query. Most of the results for the query "Yankees beat Red Sox" are about the Red Sox beating the Yankees.

NATURAL LANGUAGE QUERIES

Query-understanding technology is improving. The experimental site `www.digger.com`, for example, tells you when your query is ambiguous and helps you clarify what you are asking. If you ask Digger for information about "java," it realizes that you might mean the beverage, the island, or the programming language, and helps get the right interpretation if it guessed wrong the first time.

Powerset (`www.powerset.com`) uses natural language software to disambiguate queries based on their English syntax, and answers based on what web pages actually say. That would resolve the misunderstanding of "Yankees beat Red Sox."

Ongoing research promises to transfer the burden of disambiguating queries to the software, where it belongs, rather than forcing users to twist their brains around computerese. Natural language understanding seems to be on its way, but not in the immediate future. We may need a hundred-fold increase in computing power to make semantic analysis of web pages accurate enough so that search engines no longer give boneheaded answers to simple English queries.

Today, users tend to be tolerant when search engines misunderstand their meaning. They blame themselves and revise their queries to produce better results. This may be because we are still amazed that search engines work at all. In part, we may be tolerant of error because in web search, the cost to the user of an inappropriate answer is very low. As the technology improves, users will expect more, and will become less tolerant of wasting their time sorting through useless answers.

Step 5: Determine Relevance

A search engine's job is to provide results that match the intent of the query. In technical jargon, this criterion is called "relevance." Relevance has an objective component—a story about the Red Sox beating the Yankees is only marginally responsive to a query about the Yankees beating the Red Sox. But relevance is also inherently subjective. Only the person who posed the query can be the final judge of the relevance of the answers returned. In typing my query, I probably meant the New York Yankees beating the Boston Red Sox of Major League Baseball, but I didn't say that—maybe I meant the Flagstaff Yankees and the Continental Red Sox of Arizona Little League Baseball.

Finding all the relevant documents is referred to as "recall." Because the World Wide Web is so vast, there is no reasonable way to determine if the search engine is finding everything that is relevant. Total recall is unachievable—but it is also unimportant. Jen could give us thousands or even millions more responses that she judges to be relevant, but we are unlikely to look beyond the first page or two. Degree of relevance always trumps level of recall. Users want to find a few good results, not all possible results.

The science of measuring relevance is much older than the Web; it goes back to work by Gerald Salton in the 1960s, first at Harvard and later at Cornell. The trick is to automate a task when what counts as success has such a large subjective component. We want the computer to scan the document, look at the query, do a few calculations, and come up with a number suggesting how relevant the document is to the query.

As a very simple example of how we might calculate the relevance of a document to a query, suppose there are 500,000 words in the English language. Construct two lists of 500,000 numbers: one for the document and one for the query. Each position in the lists corresponds to one of the 500,000 words—for example, position #3682 might be for the word "drugs." For the document, each position contains a count of the number of times the corresponding word occurs in the document. Do the same thing for the query—unless it contains repeated words, each position will be 1 or 0. Multiply the lists for the document and the query, position by position, and add up the 500,000 results. If no word in the query appears in the document, you'll get a result of 0; otherwise, you will get a result greater than 0. The more frequently words from the query appear in the document, the larger the results will be.

Search Engines and Information Retrieval

Three articles offer interesting insights into how search engines and information retrieval work:

"The Anatomy of a Large-Scale Hypertextual Web Search Engine" by Sergey Brin and Larry Page was written in 2000 and gives a clear description of how the original Google worked, what the goal was, and how it was differentiated from earlier search engines.

"Modern Information Retrieval: A Brief Overview" by Amit Singhal was written in 2001 and surveys the IR scene. Singhal was a student of Gerry Salton and is now a Google Fellow.

"The Most Influential Paper Gerald Salton Never Wrote" by David Dubin presents an interesting look at some of the origins of the science.

Figure 4.5 shows how the relevance calculation might proceed for the query "Yankees beat Red Sox" and the visible part of the third document of Figure 4.4, which begins, "Red Sox rout Yankees ...." (The others probably contain more of the keywords later in the full document.) The positions in the two lists correspond to words in a dictionary in alphabetical order, from "ant" to "zebra." The words "red" and "sox" appear two times each in the snippet of the story, and the word "Yankees" appears three times.

Lexicon:	ant, …,	beat, …,	defeating, …,	new, …,	patriots, …,	red, …,	sox, …,	Yankees, …,	zebra, …
Doc:	0, …,	1, …,	2, …,	1, …,	0, …,	2, …,	2, …,	3, …,	0, …
Query:	0, …,	1, …,	0, …,	0, …,	0, …,	1, …,	1, …,	1, …,	0, …
Doc × Query	0, …,	1, …,	0, …,	0, …,	0, …,	2, …,	2, …,	3, …,	0, …

Sum of elements of Doc × Query = 1+2+2+3 = 8 = "relevance" of document to query

FIGURE 4.5 Document and query lists for relevance calculation.

That is a very crude relevance calculation—problems with it are easy to spot. Long documents tend to be measured as more relevant than short documents, because they have more word repetitions. Uninteresting words such as "from" add as much to the relevance score as more significant terms such as "Yankees." Web search engines such as Google, Yahoo!, MSN, and Ask.com consider many other factors in addition to which words occur and how often. In the list for the document, perhaps the entries are not word counts, but another number, adjusted so words in the title of the page get greater weight. Words in a larger font might also count more heavily. In a query, users tend to type more important terms first, so maybe the weights should depend on where words appear in the query.

Step 6: Determine Ranking

Once Jen has selected the relevant documents—perhaps she's chosen all the documents whose relevance score is above a certain threshold—she "ranks" the search results (that is, puts them in order). Ranking is critical in making the search useful. A search may return thousands of relevant results, and users want to see only a few of them. The simplest ranking is by relevance—putting the page with the highest relevance score first. That doesn't work well, however. For one thing, with short queries, many of the results will have approximately the same relevance.

More fundamentally, the documents Jen returns should be considered "good results" not just because they have high relevance to the query, but also because the documents themselves have high quality. Alas, it is hard to say what "quality" means in the search context, when the ultimate test of success is providing what people want. In the example of the earlier sidebar, who is to judge whether the many links to material about Britney Spears are really "better" answers to the "spears" query than the link to Professor Spears? And whatever "quality" may be, the ranking process for the major web search engines takes place automatically, without human intervention. There is no way to include protocols for checking professional licenses and past convictions for criminal fraud—not in the current state of the Web, at least.

Even though quality can't be measured automatically, something like "importance" or "reputation" can be extracted from the structure of linkages that holds the Web together. To take a crude analogy, if you think of web pages as scientific publications, the reputations of scientists tend to rise if their work is widely cited in the work of other scientists. That's far from a

What Makes a Page Searchable

No search provider discloses the full details of its relevance and ranking algorithm. The formulas remain secret because they offer competitive advantages, and because knowing what gives a page high rank makes abuse easier. But here are some of the factors that might be taken into account:

- Whether a keyword is used in the title of the web page, a major heading, or a second-level heading
- Whether it appears only in the body text, and if so, how "prominently"
- Whether the web site is considered "trustworthy"
- Whether the pages linked to from within the page are themselves relevant
- Whether the pages that link to this page are relevant
- Whether the page is old or young
- Whether the pages it links to are old or young
- Whether it passes some objective quality metric—for example, not containing any misspellings

Once you go to the trouble of crawling the Web, there is plenty to analyze, if you have the computing power to do it!

perfect system for judging the importance of scientific work—junk science journals do exist, and sometimes small groups of marginal scientists form mutual admiration societies. But for the Web, looking at the linkage structure is a place to start to measure the significance of pages.

One of Google's innovations was to enhance the relevance metric with another numerical value called "PageRank." PageRank is a measure of the "importance" of each a page that takes into account the external references to it—a World Wide Web popularity contest. If more web pages link to a particular page, goes the logic, it must be more important. In fact, a page should be judged more important if a lot of *important* pages link to it than if the same number of unimportant pages link to it. That seems to create a circular definition of importance, but the circularity can be resolved—with a bit of mathematics and a lot of computing power.

This way of ranking the search results seems to reward reputation and to be devoid of judgment—it is a mechanized way of aggregating mutual opinions. For example, when we searched using Google for "schizophrenia drugs," the top result was part of the site of a Swedish university. Relevance was certainly part of the reason that page came up first; the page was specifically about drugs used to treat schizophrenia, and the words "schizophrenia" and "drugs" both appeared in the title of the page. Our choice of words affected the relevance of the page—had we gone to the trouble to type "medicines" instead of "drugs," this link wouldn't even have made it to the first page of search results. Word order matters, too—Google returns different results for "drugs schizophrenia" than for "schizophrenia drugs."

Sergey Brin and Larry Page, Google's founders, were graduate students at Stanford when they developed the company's early technologies. The "Page" in "PageRank" refers not to web pages, but to Larry Page.

This page may also have been ranked high because many other web pages contained references to it, particularly if many of those pages were themselves judged to be important. Other pages about schizophrenia drugs may have used better English prose style, may have been written by more respected scientific authorities, and may have contained more up-to-date information and fewer factual errors. The ranking algorithm has no way to judge any of that, and no one at Google reads every page to make such judgments.

Google, and other search engines that rank pages automatically, use a secret recipe for ranking—a pinch of this and a dash of that. Like the formula

for Coca-Cola, only a few people know the details of commercial ranking algorithms. Google's algorithm is patented, so anyone can read a description. Figure 4.6 is an illustration from that patent, showing several pages with links to each other. This illustration suggests that both the documents themselves and the links between them might be assigned varying numbers as measures of their importance. But the description omits many details and, as actually implemented, has been adjusted countless times to improve its performance. A company's only real claim for the validity of its ranking formula is that people like the results it delivers—if they did not, they would shift to one of the competing search engines.

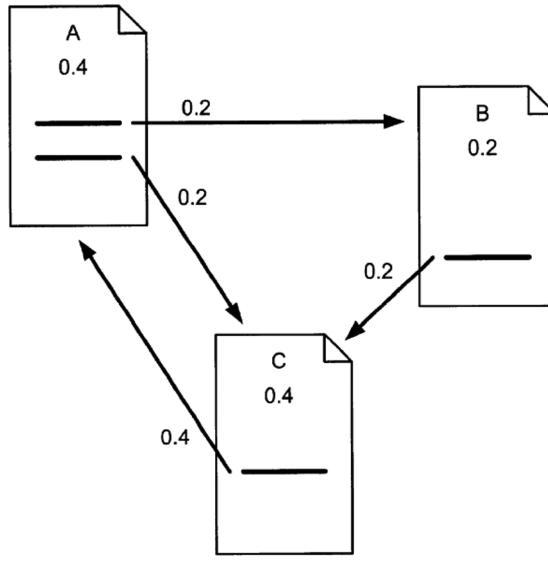


FIGURE 4.6 A figure from the PageRank patent (U.S. Patent #6285999), showing how links between documents might receive different weights.

It may be that one of the things people like about their favored search engine is consistently getting what they believe to be unbiased, useful, and even truthful information. But "telling the truth" in search results is ultimately only a means to an end—the end being greater profits for the search company.

Ranking is a matter of opinion. But a lot hangs on those opinions. For a user, it usually does not matter very much which answer comes up first or whether any result presented is even appropriate to the query. But for a

company offering a product, where it appears in the search engine results *can* be a matter of life and death.

KinderStart (www.kinderstart.com) runs a web site that includes a directory and search engine focused on products and services for young children. On March 19, 2005, visits to its site declined by 70% when Google lowered its PageRank to zero (on a scale of 0 to 10). Google may have deemed KinderStart's page to be low quality because its ranking algorithm found the page to consist mostly of links to other sites. Google's public description of its criteria warns about pages with "little or no original content." KinderStart saw matters differently and mounted a class action lawsuit against Google, claiming, among other things, that Google had violated its rights to free speech under the First Amendment by making its web site effectively invisible. Google countered that KinderStart's low PageRank was just Google's opinion, and opinions were not matters to be settled in court:

> Google, like every other search engine operator, has made that determination for its users, exercising its judgment and expressing its opinion about the relative significance of web sites in a manner that has made it the search engine of choice for millions. Plaintiff KinderStart contends that the judiciary should have the final say over that editorial process.

No fair, countered KinderStart to Google's claim to be just expressing an opinion. "PageRank," claimed KinderStart, "is not a mere statement of opinion of the innate value or human appeal of a given web site and its web pages," but instead is "a mathematically-generated product of measuring and assessing the quantity and depth of all the hyperlinks on the Web that tie into PageRanked web site, under programmatic determination by Defendant Google."

SEEING A PAGE'S PAGERANK

Google has a toolbar you can add to certain browsers, so you can see PageRanks of web pages. It is downloadable from toolbar.google.com. You can also use the site www.iwebtool.com/pagerank_checker to enter a URL in a window and check its PageRank.

The judge rejected every one of KinderStart's contentions—and not just the claim that KinderStart had a free speech right to be more visible in Google searches. The judge also rejected claims that Google was a monopoly guilty of antitrust violations, and that KinderStart's PageRank of zero amounted to a defamatory statement about the company.

Whether it's a matter of opinion or manipulation, KinderStart is certainly much easier to find using Yahoo! than Google. Using Yahoo!, `kinderstart.com` is the top item returned when searching for "kinderstart." When we used Google, however, it did not appear until the twelfth page of results.

A similar fate befell `bmw.de`, the German web page of automaker BMW. The page Google indexed was straight text, containing the words "gebrauchtwagen" and "neuwagen" ("used car" and "new car") dozens of times. But a coding trick caused viewers instead to see a more conventional page with few words and many pictures. The effect was to raise BMW's position in searches for "new car" and "used car," but the means violated Google's clear instructions to web site designers: "Make pages for users, not for search engines. Don't deceive your users or present different content to search engines than you display to users, which is commonly referred to as 'cloaking.'" Google responded with a "death penalty"—removing `bmw.de` from its index. For a time, the page simply ceased to exist in Google's universe. The punitive measure showed that Google was prepared to act harshly against sites attempting to gain rank in ways it deemed consumers would not find helpful—and at the same time, it also made clear that Google was prepared to take *ad hoc* actions against individual sites.

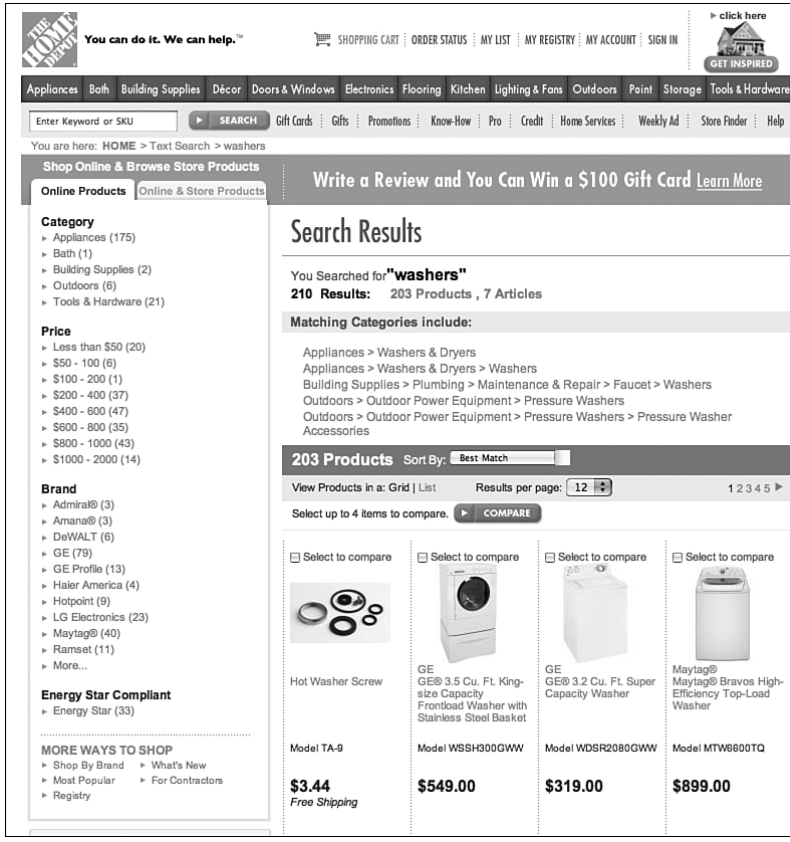
Step 7: Presenting Results

After all the marvelous hard work of Steps 1–6, search engines typically provide the results in a format that is older than Aristotle—the simple, top-to-bottom list. There are less primitive ways of displaying the information.

If you search for something ambiguous like "washer" with a major web search engine, you will be presented with a million results, ranging from clothes washers to software packages that remove viruses. If you search Home Depot's web site for "washer," you will get a set of automatically generated choices to assist you in narrowing the search: a set of categories, price ranges, brand names, and more, complete with pictures (see Figure 4.7).

Alternatives to the simple rank-ordered list for presenting results better utilize the visual system. Introducing these new forms of navigation may shift the balance of power in the search equation. Being at the top of the list may no longer have the same economic value, but something else may replace the currently all-important rank of results—quality of the graphics, for example.

No matter how the results are presented, something else appears alongside them, and probably always will. It is time to talk about those words from the sponsors.



Source: Home Depot.

FIGURE 4.7 Results page from a search for "washers" on the Home Depot web site.

Who Pays, and for What?

Web search is one of the most widely used functions of computers. More than 90% of online adults use search engines, and more than 40% use them on a typical day. The popularity of search engines is not hard to explain. Search engines are generally free for anyone to use. There are no logins, no fine print to agree to, no connection speed parameters to set up, and no personal information to be supplied that you'd rather not give away. If you have an Internet connection, then you almost certainly have a web browser, and it probably comes with a web search engine on its startup screen. There are no directions

to read, at least to get started. Just type some words and answers come back. You can't do anyone any harm by typing random queries and seeing what happens. It's even fun.

Perhaps because search is so useful and easy, we are likely to think of our search engine as something like a public utility—a combination of an encyclopedia and a streetlamp, a single source supplying endless amounts of information to anyone. In economic terms, that is a poor analogy. Utilities charge for whatever they provide—water, gas, or electricity—and search firms don't. Utilities typically don't have much competition, and search firms do. Yet we trust search engines as though they were public utilities because their results just flow to us, and because the results seem consistent with our expectations. If we ask for American Airlines, we find its web site, and if we ask for "the price of tea in China," we find both the actual price ($1.84 for 25 tea bags) and an explanation of the phrase. And perhaps we trust them because we assume that machines are neutral and not making value judgments. The fact that our expectations are rarely disappointed does not, however, mean that our intuitions are correct.

Who pays for all this? There are four possibilities:

- The users could pay, perhaps as subscribers to a service.
- Web sites could pay for the privilege of being discovered.
- The government or some nonprofit entity could pay.
- Advertisers could pay.

All four business models have all been tried.

Commercial-Free Search

In the very beginning, universities and the government paid, as a great deal of information retrieval research was conducted in universities under federal grants and contracts. WebCrawler, one of the first efforts to crawl the Web in order to produce an index of terms found on web pages, was Brian Pinkerton's research project at the University of Washington. He published a paper about it in 1994, at an early conference on the World Wide Web. The 1997 academic research paper by Google's founders, explaining PageRank, acknowledges support by the National Science Foundation, the Defense Advanced Research Projects Agency, and the National Aeronautics and Space Administration, as well as several industrial supporters of Stanford's computer science research programs. To this day, Stanford University owns the patent on the PageRank algorithm—Google is merely the exclusive licensee.

Academia and government were the wellsprings of search technology, but that was before the Web became big business. Search needed money to grow. Some subscription service web sites, such as AOL, offered search engines. Banner ads appeared on web sites even before search engines became the way to find things, so it was natural to offer advertising to pay for search engine sites. Banner ads are the equivalent of billboards or displayed ads in newspapers. The advertiser buys some space on a page thought promising to bring in some business for the advertiser, and displays an eye-catching come-on.

With the advent of search, it was possible to sell advertising space depending on what was searched for—"targeted advertising" that would be seen only by viewers who might have an interest in the product. To advertise cell phones, for example, ads might be posted only on the result pages of searches involving the term "phone." Like billboards, banner ads bring in revenue. And also like billboards, posting too many of them, with too much distracting imagery, can annoy the viewer!

There was a presumed, generally acknowledged ethical line. Payola was a no-no.

Whichever business model was in use, there was a presumed, generally acknowledged ethical line. If you were providing a search engine, you were not supposed to accept payments to alter the presentation of your results. If you asked for information, you expected the results to be impartial, even if they were subjective. Payola was a no-no. But there was a very fine line between partiality and subjectivity, and the line was drawn in largely unexplored territory. That territory was expanding rapidly, as the Web moved out of the academic and research setting and entered the world of retail stores, real estate brokers, and impotence cures.

Holding a line against commercialism posed a dilemma—what Brin and Page, in their original paper, termed the "mixed motives" of advertising-based search engines. How would advertisers respond if the engine provided highly ranked pages that were unfriendly to their product? Brin and Page noted that a search for "cell phones" on their prototype search engine returned an article about the dangers of talking on cell phones while driving. Would cell phone companies really pay to appear on the same page with information that might discourage people from buying cell phones? Because of such conflicts, Google's founders predicted "that advertising funded search engines will be inherently biased toward the advertisers and away from the needs of the consumers." They noted that one search engine, Open Text, had already gotten out of the search engine business after it was reported to be selling rank for money.

Placements, Clicks, and Auctions

Only a year later, the world had changed. Starting in 1998, Overture (originally named GoTo.com) made a healthy living by leaping with gusto over the presumed ethical line. That line turned out to have been a chasm mainly in the minds of academics. Overture simply charged advertisers to be searchable, and charged them more for higher rankings in the search results. The argument in favor of this simple commercialism was that if you could afford to pay to be seen, then your capacity to spend money on advertising probably reflected the usefulness of your web page. It mattered not whether this was logical, nor whether it offended purists. It seemed to make people happy. Overture's CEO explained the company's rationale in simple terms. Sounding every bit like a broker in the bazaar arguing with the authorities, Jeffrey Brewer explained, "Quite frankly, there's no understanding of how any service provides results. If consumers are satisfied, they really are not interested in the mechanism."

Customers were indeed satisfied. In the heady Internet bubble of the late 1990s, commercial sites were eager to make themselves visible, and users were eager to find products and services. Overture introduced a second innovation, one that expanded its market beyond the sites able to pay the substantial up-front fees that AOL and Yahoo! charged for banner ads. Overture charged advertisers nothing to have their links posted—it assessed fees only if users clicked on those links from Overture's search results page. A click was only a penny to start, making it easy for small-budget Web companies to buy advertising. Advertisers were eager to sign up for this "pay-per-click" (PPC) service. They might not get a sale on every click, but at least they were paying only for viewers who took the trouble to learn a little bit more than what was in the advertisement.

As a search term became popular, the price for links under that term went up. The method of setting prices was Overture's third innovation. If several advertisers competed for the limited real estate on a search results page, Overture held an auction among them and charged as much as a dollar a click. The cost per click adjusted up and down, depending on how many other customers were competing for use of the same keyword. If a lot of advertisers wanted links to their sites to appear when you searched for "camera," the price per click would rise. Real estate on the screen was a finite resource, and the market would determine the going rates. Auctioning keywords was simple, sensible, and hugely profitable.

Ironically, the bursting of the Internet bubble in 2000 only made Overture's pay-for-ranking, pay-per-click, keyword auction model more attractive. As profits and capital dried up, Internet businesses could no longer afford up-front capital to buy banner ads, some of which seemed to yield meager results. As a result, many companies shifted their advertising budgets to Overture and other services that adopted some of Overture's innovations. The bursting bubble affected the hundreds of early search companies as well. As competition took its toll, Yahoo! and AOL both started accepting payment for search listings.

Uncle Sam Takes Note

Different search engines offered different levels of disclosure about the pay-for-placement practice. Yahoo! labeled the paid results with the word "Sponsored," the term today generally accepted as the correct euphemism for "paid advertisement." Others used vaguer terms such as "partner results" or "featured listings." Microsoft's MSN offered a creative justification for its use of the term "featured" with no other explanation: MSN's surveys showed that consumers already assumed that search results were for sale—so there was no need to tell them! With the information superhighway becoming littered with roadkill, business was less fun, and business tactics became less grounded in the utopian spirit that had given birth to the Internet. "We can't afford to have ideological debates anymore," said Evan Thornley, CEO of one startup. "We're a public company."

At first, the government stayed out of all this, but in 2001, Ralph Nader's watchdog organization, Consumer Alert, got involved. Consumer Alert filed a complaint with the Federal Trade Commission alleging that eight search engine vendors were deceiving consumers by intermingling "paid inclusion" and "paid placement" results along with those that were found by the search engine algorithm. Consumer Alert's Executive Director, Gary Ruskin, was direct in his accusation: "These search engines have chosen crass commercialism over editorial integrity. We are asking the FTC to make sure that no one is tricked by the search engines' descent into commercial deception. If they are going to stuff ads into search results, they should be required to say that the ads are ads."

The FTC agreed, and requested search engines to clarify the distinction between organic results and sponsored results. At the same time, the FTC issued a consumer alert to advise and inform consumers of the practice (see Figure 4.8). Google shows its "sponsored links" to the right, as in Figure 4.1, or slightly indented. Yahoo! shows its "sponsor results" on a colored background.

Source: Federal Trade Commission.

FIGURE 4.8 FTC Consumer Alert about paid ranking of search results.

Google Finds Balance Without Compromise

As the search engine industry was struggling with its ethical and fiscal problems in 2000, Google hit a vein of gold.

Google already had the PageRank algorithm, which produced results widely considered superior to those of other search engines. Google was fast, in part because its engineers had figured out how to split both background and foreground processing across many machines operating in parallel. Google's vast data storage was so redundant that you could pull out a disk drive anywhere and the engine didn't miss a beat. Google was not suspected of taking payments for rankings. And Google's interface was not annoying—no flashy banner ads (no banner ads at all, in fact) on either the home page or the search results page. Google's home page was a model of understatement. There was almost nothing on it except for the word "Google," the search window, and the option of getting a page of search results or of "feeling lucky" and going directly to the top hit (an option that was more valuable when many users had slow dialup Internet connections).

There were two other important facts about Google in early 2000: Google was expanding, and Google was not making much money. Its technology was successful, and lots of people were using its search engine. It just didn't have a viable business model—until AdWords.

Google's AdWords allows advertisers to participate in an auction of keywords, like Overture's auction for search result placement. But when you win an AdWords auction, you simply get the privilege of posting a small text

advertisement on Google's search results pages under certain circumstances—not the right to have your web site come up as an organic search result. The beauty of the system was that it didn't interfere with the search results, was relatively unobtrusive, was keyed to the specific search, and did not mess up the screen with irritating banner ads.

At first, Google charged by the "impression"—that is, the price of your AdWords advertisement simply paid for having it shown, whether or not anyone clicked on it. AdWords switched to Overture's pay-per-click business model in 2002. Initially, the advertisements were sold one at a time, through a human agent at Google. AdWords took off when the process of placing an advertisement was automated. To place an ad today, you simply fill out a web form with information about what search terms you want to target, what few words you want as the text of your ad—and what credit card number Google can use to charge its fee.

Google's technology was brilliant, but none of the elements of its business model was original. With the combination, Google took off and became a giant. The advertising had no effect on the search results, so confidence in the quality of Google's search results was undiminished. AdWords enabled Google to achieve the balance Brin and Page had predicted would be impossible: commercial sponsorship without distorted results. Google emerged—from this dilemma, at least—with its pocketbooks overflowing and its principles intact.

Banned Ads

Targeted ads, such as Google's AdWords, are changing the advertising industry. Online ads are more cost-effective because the advertiser can control who sees them. The Internet makes it possible to target advertisements not just by search term, but geographically—to show different ads in California than in Massachusetts, for example. The success of web advertising has blown to bits a major revenue source for newspapers and television. The media and communications industries have not yet caught up with the sudden reallocation of money and power.

The success of web advertising has blown to bits a major revenue source for newspapers and television.

As search companies accumulate vast advertising portfolios, they control what products, legal or illegal, may be promoted. Their lists result from a combination of legal requirements, market demands, and corporate philosophy. The combined effect of these decisions represents a kind of soft censorship—with which newspapers have long been familiar, but which acquires new significance as search sites become a

dominant advertising engine. Among the items and services for which Google will not accept advertisements are fake designer goods, child pornography (some adult material is permitted in the U.S., but not if the models *might* be underage), term paper writing services, illegal drugs and some legal herbal substances, drug paraphernalia, fireworks, online gambling, miracle cures, political attack ads (although political advertising is allowed in general), prostitution, traffic radar jammers, guns, and brass knuckles. The list paints a striking portrait of what Joe and Mary Ordinary want to see, should see, or will tolerate seeing—and perhaps also how Google prudentially restrains the use of its powerfully liberating product for illegal activities.

Search Is Power

At every step of the search process, individuals and institutions are working hard to control what we see and what we find—not to do us ill, but to help us. Helpful as search engines are, they don't have panels of neutral experts deciding what is true or false, or what is important or irrelevant. Instead, there are powerful economic and social motivations to present information that is to our liking. And because the inner workings of the search engines are not visible, those controlling what we see are themselves subject to few controls.

Algorithmic Does Not Mean Unbiased

Because search engines compute relevance and ranking, because they are "algorithmic" in their choices, we often assume that they, unlike human researchers, are immune to bias. But bias can be coded into a computer program, introduced by small changes in the weights of the various factors that go into the ranking recipe or the spidering selection algorithm. And even what *counts* as bias is a matter of human judgment.

Having a lot of money will not buy you a high rank by paying that money to Google. Google's PageRank algorithm nonetheless incorporates something of a bias in favor of the already rich and powerful. If your business has become successful, a lot of other web pages are likely to point to yours, and that increases your PageRank. This makes sense and tends to produce the results that most people feel are correct. But the degree to which power should beget more power is a matter over which powerful and marginal businesses might have different views. Whether the results "seem right," or the search algorithm's parameters need adjusting, is a matter only humans can judge.

For a time, Amazon customers searching for books about abortion would get back results including the question, "Did you mean adoption?" When a pro-choice group complained, Amazon responded that the suggestion was automatically generated, a consequence of the similarity of the words. The search engine had noticed, over time, that many people who searched for "abortion" also searched for "adoption." But Amazon agreed to make the *ad hoc* change to its search algorithm to treat the term "abortion" as a special case. In so doing, the company unintentionally confirmed that its algorithms sometimes incorporate elements of human bias.

Market forces are likely to drive commercially viable search engines toward the bias of the majority, and also to respond to minority interests only in proportion to their political power. Search engines are likely to favor fresh items over older and perhaps more comprehensive sources, because their users go to the Internet to get the latest information. If you rely on a search engine to discover information, you need to remember that others are making judgment calls for you about what you are being shown.

Not All Search Engines Are Equal

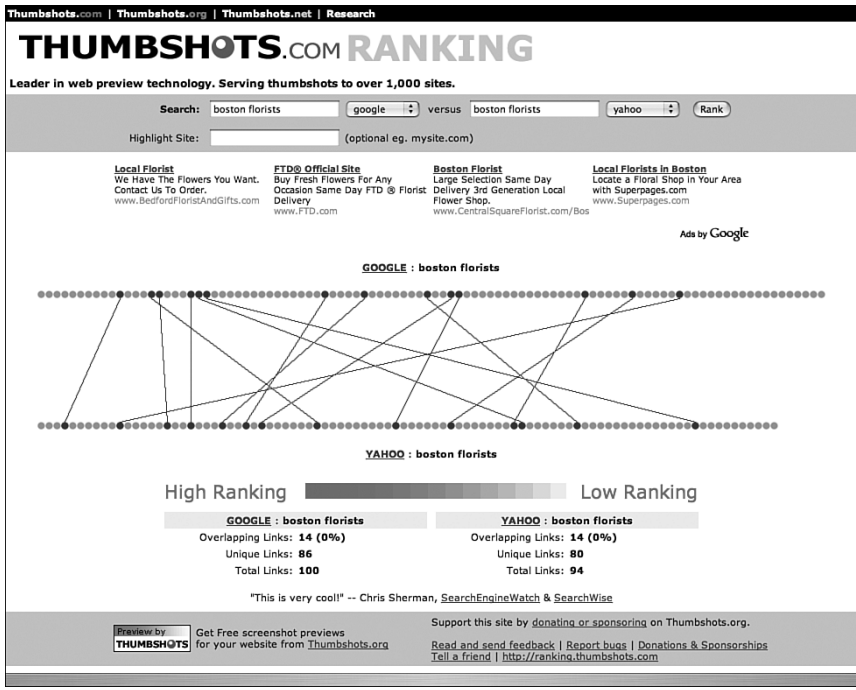
When we use a search engine, we may think that what we are getting is a representative sample of what's available. If so, what we get from one search engine should be pretty close to what we get from another. This is very far from reality.

A study comparing queries to Google, Yahoo!, ASK, and MSN showed that the results returned on the first page were unique 88% percent of the time. Only 12% of the first-page results were in common to even two of these four search engines. If you stick with one search engine, you could be missing what you're looking for. The tool ranking.thumbshots.com provides vivid graphic representations of the level of overlap between the results of different search engines, or different searches using the same search engine. For example, Figure 4.9 shows how little overlap exists between Google and Yahoo! search results for "boston florist."

Each of the hundred dots in the top row represents a result of the Google search, with the highest-ranked result at the left. The bottom row represents Yahoo!'s results. A line connects each pair of identical search results—in this case, only 11% of the results were in common. Boston Rose Florist, which is Yahoo's number-one response, doesn't turn up in Google's search at all—not in the top 100, or even in the first 30 pages Google returns.

Ranking determines visibility. An industry research study found that 62% of search users click on a result from the first page, and 90% click on a result within the first three pages. If they don't find what they are looking for, more

than 80% start the search over with the same search engine, changing the keywords—as though confident that the search engine "knows" the right answer, but they haven't asked the right question. A study of queries to the Excite search engine found that more than 90% of queries were resolved in the first three pages. Google's experience is even more concentrated on the first page.



Reprinted with permission of SmartDevil, Inc.

FIGURE 4.9 Thumbshots comparison of Google and Yahoo! search results for "boston florists."

Search engine users have great confidence that they are being given results that are not only useful but authoritative. 36% of users thought seeing a company listed among the top search results indicated that it was a top company in its field; only 25% said that seeing a company ranked high in search results would not lead them to think that it was a leader in its field. There is, in general, no reason for such confidence that search ranking corresponds to corporate quality.

Cat and Mouse with Blog Spammers

You may see comments on a blog consisting of nothing but random words and a URL. A malicious bot is posting these messages in the hope that Google's spider will index the blog page, including the spam URL. With more pages linking to the URL, perhaps its PageRank will increase and it will turn up in searches. Blogs counter by forcing you to type some distorted letters—a so-called *captcha* ("Completely Automated Public Turing test to tell Computers and Humans Apart"), a test to determine if the party posting the comment is really a person and not a bot. Spammers counter by having their bot take a copy of the captcha and show it to human volunteers. The spam bot then takes what the volunteers type and uses it to gain entry to the blog site. The volunteers are recruited by being given access to free pornography if they type the captcha's text correctly! Here is a sample captcha:

This image has been released into the public domain by its author, Kruglov at the wikipedia project. This applies worldwide.

Search Results Can Be Manipulated

Search is a remarkable business. Internet users put a lot of confidence in the results they get back from commercial search engines. Buyers tend to click on the first link, or at least a link on the first page, even though those links may depend heavily on the search engine they happen to be using, based on complex technical details that hardly anyone understands. For many students, for example, the library is an information source of last resort, if that. They do research as though whatever their search engine turns up must be a link to the truth. If people don't get helpful answers, they tend to blame themselves and change the question, rather than try a different search engine—even though the answers they get can be inexplicable and capricious, as anyone googling "kinderstart" to find kinderstart.com will discover.

Under these circumstances, anyone putting up a web site to get a message out to the world would draw an obvious conclusion. Coming out near the top of the search list is too important to leave to chance. Because ranking is algorithmic, a set of rules followed with diligence and precision, it must be possible to manipulate the results. The Search Engine Optimization industry (SEO) is based on that demand.

Search Engine Optimization is an activity that seeks to improve how particular web pages rank within major search engines, with the intent of

increasing the traffic that will come to those web sites. Legitimate businesses try to optimize their sites so they will rank higher than their competitors. Pranksters and pornographers try to optimize their sites, too, by fooling the search engine algorithms into including them as legitimate results, even though their trappings of legitimacy are mere disguises. The search engine companies tweak their algorithms in order to see through the disguises, but their tweaks sometimes have unintended effects on legitimate businesses. And the tweaking is largely done in secret, to avoid giving the manipulators any ideas about countermeasures. The result is a chaotic battle, with innocent bystanders, who have become reliant on high search engine rankings, sometimes injured as the rules of engagement keep changing.

Google proclaims of its PageRank algorithm that "Democracy on the web works," comparing the ranking-by-inbound-links to a public election. But the analogy is limited—there are many ways to manipulate the "election," and the voting rules are not fully disclosed.

The key to search engine optimization is to understand how particular engines do their ranking—what factors are considered, and what weights they are given—and then to change your web site to improve your score. For example, if a search engine gives greater weight to key words that appear in the title, and you want your web page to rank more highly when someone searches for "cameras," you should put the word "cameras" in the title. The weighting factors may be complex and depend on factors external to your own web page—for example, external links that point to your page, the age of the link, or the prestige of the site from which it is linked. So significant time, effort, and cost must be expended in order to have a meaningful impact on results.

Then there are techniques that are sneaky at best—and "dirty tricks" at worst. Suppose, for example, that you are the web site designer for Abelson's, a new store that wants to compete with Bloomingdale's. How would you entice people to visit Abelson's site when they would ordinarily go to Bloomingdale's? If you put "We're better than Bloomingdale's!" on your web page, Abelson's page might appear in the search results for "Bloomingdale's." But you might not be willing to pay the price of mentioning the competition on Abelson's page. On the other hand, if you just put the word "Bloomingdale's" *in white text on a white background* on Abelson's page, a human viewer wouldn't see it—but the indexing software might index it anyway. The indexer is working with the HTML code that generates the page, not the visible page itself. The software might not be clever enough to realize that the word "Bloomingdale's" in the HTML code for Abelson's web page would not actually appear on the screen.

A huge industry has developed around SEO, rather like the business that has arisen around getting high school students packaged for application to college. A Google search for "search engine optimization" returned 11 sponsored links, including some with ads reading "Page 1 Rankings Guarantee" and "Get Top Rankings Today."

Is the search world more ethical because the commercial rank-improving transactions are indirect, hidden from the public, and going to the optimization firms rather than to the search firms? After all, it is only logical that if you have an important message to get out, you would optimize your site to do so. And you probably wouldn't have a web site at all if you thought you had nothing important to say. Search engine companies tend to advise their web site designers just to create better, more substantive web pages, in much the same way that college admissions officials urge high school students just to learn more in school. Neither of the dependent third-party "optimization" industries is likely to disappear anytime soon because of such principled advice.

And what's "best"—for society in general, not just for the profits of the search companies or the companies that rely on them—can be very hard to say. In his book, *Ambient Findability*, Peter Morville describes the impact of search engine optimization on the National Cancer Institute's site, `www.cancer.gov`. The goal of the National Cancer Institute is to provide the most reliable and the highest-quality information to people who need it the most,

Google Bombing

A "Google bomb" is a prank that causes a particular search to return mischievous results, often with political content. For example, if you searched for "miserable failure" after the 2000 U.S. presidential election, you got taken to the White House biography of George Bush. The libertarian Liberty Round Table mounted an effort against the Center for Science in the Public Interest, among others. In early 2008, `www.libertyroundtable.org` read, "Have you joined the Google-bombing fun yet? Lob your volleys at the food nazis and organized crime. Your participation can really make the difference with this one—read on and join the fun! Current Target: Verizon Communications, for civil rights violations." The site explains what HTML code to include in your web page, supposedly to trick Google's algorithms.

Marek W., a 23-year-old programmer from Cieszyn, Poland, "Google bombed" the country's president, Lech Kaczyński. Searches for "kutas" using Google (it's the Polish word for "penis") returned the president's web site as the first choice. Mr. Kaczyński was not pleased, and insulting the president is a crime in Poland. Marek is now facing three years in prison.

often cancer sufferers and their families. Search for "cancer," and the NCI site was "findable" because it appeared near the topic of the search page results. That wasn't the case, though, when you looked for specific cancers, yet that's exactly what the majority of the intended users did. NCI called in search engine optimization experts, and all that is now changed. If we search for "colon cancer," the specific page on the NCI site about this particular form of cancer appears among the top search results.

Is this good? Perhaps—if you can't trust the National Cancer Institute, who *can* you trust? But WebMD and other commercial sites fighting for the top position might not agree. And a legitimate coalition, the National Colorectal Cancer Roundtable, doesn't appear until page 7, too deep to be noticed by almost any user.

Optimization is a constant game of cat and mouse. The optimizers look for better ways to optimize, and the search engine folks look for ways to produce more reliable results. The game occasionally claims collateral victims. Neil Montcrief, an online seller of large-sized shoes, prospered for a while because searches for "big feet" brought his store, `2bigfeet.com`, to the top of the list. One day, Google tweaked its algorithm to combat manipulation. Montcrief's innocent site fell to the twenty-fifth page, with disastrous consequences for his economically marginal and totally web-dependent business.

Manipulating the ranking of search results is one battleground where the power struggle is played out. Because search is the portal to web-based information, controlling the search results allows you, perhaps, to control what people think. So even governments get involved.

Search Engines Don't See Everything

Standard search engines fail to index a great deal of information that is accessible via the Web. Spiders may not penetrate into databases, read the contents of PDF or other document formats, or search useful sites that require a simple, free registration. With a little more effort than just typing into the search window of Google or Yahoo!, you may be able to find exactly what you are looking for. It is a serious failure to assume that something is unimportant or nonexistent simply because a search engine does not return it. A good overview of resources for finding things in the "deep web" is at Robert Lackie's web site, `www.robertlackie.com`.

Search Control and Mind Control

To make a book disappear from a library, you don't have to remove it from the bookshelf. All you need to do is to remove its entry from the library

catalog—if there is no record of where to find it, it does not matter if the book actually still exists.

When we search for something, we have an unconfirmed confidence that what the search engine returns is what exists. A search tool is a lens through which we view information. We count on the lens not to distort the scene, although we know it can't show us the entire landscape at once. Like the book gone from the catalog, information that cannot be found may as well not exist. So removing information in the digital world does not require removing the documents themselves. You can make things disappear by banishing them into the un-indexed darkness.

You can make things disappear by banishing them into the un-indexed darkness.

By controlling "findability," search tools can be used to hide as well as to reveal. They have become a tool of governments seeking to control what their people know about the world, a theme to which we return in Chapter 7, "You Can't Say That on the Internet." When the Internet came to China, previously unavailable information began pouring into the country. The government responded by starting to erect "the great firewall of China," which filtered out information the government did not want seen. But bits poured in more quickly than offending web sites could be blocked. One of the government's counter-measures, in advance of a Communist Party congress in 2002, was simply to close down certain search engines. "Obviously there is some harmful information on the Internet," said a Chinese spokesman by way of explanation. "Not everyone should have access to this harmful information." Google in particular was unavailable—it may have been targeted because people could sometimes use it to access a cached copy of a site to which the government had blocked direct access.

Search was already too important to the Chinese economy to leave the ban in place for very long. The firewall builders got better, and it became harder to reach banned sites. But such a site might still turn up in Google's search results. You could not access it when you clicked on the link, but you could see what you were missing.

In 2004, under another threat of being cut off from China, Google agreed to censor its news service, which provides access to online newspapers. The company reluctantly decided not to provide any information at all about those stories, reasoning that "simply showing these headlines would likely result in Google News being blocked altogether in China." But the government was not done yet.

The really hard choice came a year later. Google's search engine was available inside China, but because Google's servers were located outside the

country, responses were sluggish. And because many of the links that were returned did not work, Google's search engine was, if not useless, at least uncompetitive. A Chinese search engine, Baidu, was getting most of the business.

Google had a yes-or-no decision: to cooperate with the government's web site censorship or to lose the Chinese market. How would it balance its responsibilities to its shareholders to grow internationally with its corporate mission: "to organize the world's information and make it universally accessible and useful"? Would the company co-founded by an émigré from the Soviet Union make peace with Chinese censorship?

GOOGLE U.S. VS. GOOGLE CHINA

You can try some searches yourself:

- `www.google.com` is the version available in the United States.
- `www.google.cn` is the version available in China.

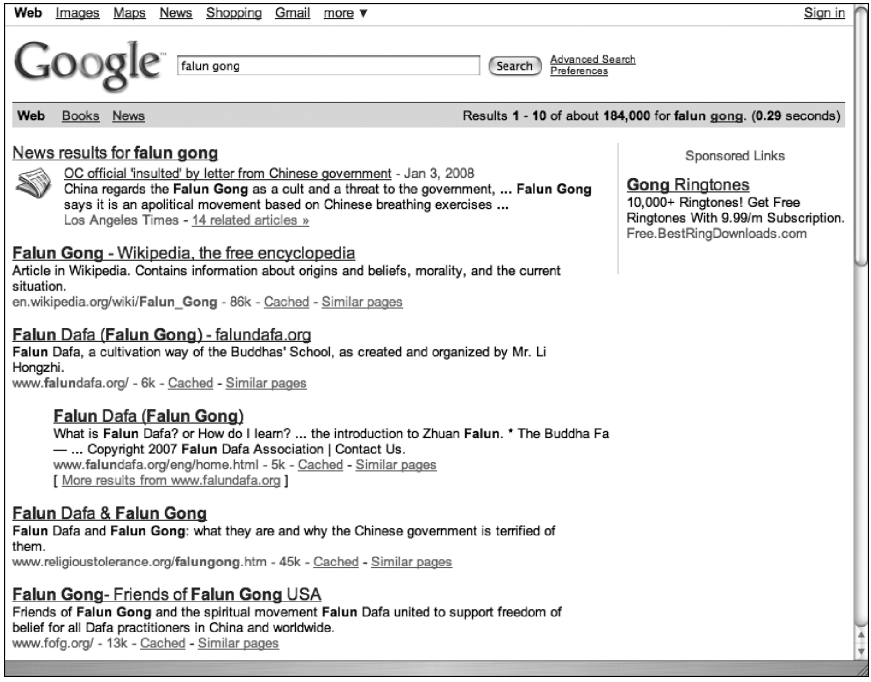
Completely universal accessibility was already more than Google could lawfully accomplish, even in the U.S. If a copyright holder complained that Google was making copyrighted material improperly accessible, Google would respond by removing the link to it from search results. And there were other U.S. laws about web content, such as the Communications Decency Act, which we discuss in Chapter 7.

Google's accommodation to Chinese authorities was, in a sense, nothing more than the normal practice of any company: You have to obey the local laws anywhere you are doing business. China threw U.S. laws back at U.S. critics. "After studying internet legislation in the West, I've found we basically have identical legislative objectives and principles," said Mr. Liu Zhengrong, deputy chief of the Internet Affairs Bureau of the State Council Information Office. "It is unfair and smacks of double standards when (foreigners) criticize China for deleting illegal and harmful messages, while it is legal for U.S. web sites to do so."

And so, when Google agreed in early 2006 to censor its Chinese search results, some were awakened from their dreams of a global information utopia. "While removing search results is inconsistent with Google's mission, providing no information (or a heavily degraded user experience that amounts to no information) is more inconsistent with our mission," a Google statement read. That excuse seemed weak-kneed to some. A disappointed libertarian commentator countered, "The evil of the world is made possible by the sanction that you give it." (This is apparently an allusion to another Google maxim, "Don't be evil"—now revised to read, "You can make money without doing evil.") The U.S. Congress called Google and other search companies on the carpet. "Your abhorrent activities in China are a disgrace," said

California Representative Tom Lantos. “I cannot understand how your corporate executives sleep at night.”

The results of Google’s humiliating compromise are striking, and anyone can see them. Figure 4.10 shows the top search results returned by the U.S. version of Google in response to the query “falun gong.”



Google ™ is a registered trademark of Google, Inc. Reprinted by permission.

FIGURE 4.10 Search results for “falun gong” provided by Google U.S.

By contrast, Figure 4.11 shows the first few results in response to the same query if the Chinese version of Google is used instead. All the results are negative information about the practice, or reports of actions taken against its followers.

Most of the time, whether you use the U.S. or Chinese version of Google, you will get similar results. In particular, if you search for “shoes,” you get sponsored links to online shoe stores so Google can pay its bills.

But there are many exceptions. One researcher tested the Chinese version of Google for 10,000 English words and found that roughly 9% resulted in censored responses. Various versions of the list of blocked words exist, and the specifics are certainly subject to change without notice. Recent versions

contained such entries as “crime against humanity,” “oppression,” and “genocide,” as well as lists of dissidents and politicians.

FIGURE 4.11 Results of “falun gong” search returned by Google China.

The search engine lens is not impartial. At this scale, search can be an effective tool of thought control. A Google executive told Congress, “In an imperfect world, we had to make an imperfect choice”—which is surely the truth. But business is business. As Google CEO Eric Schmidt said of the company’s practices, “There are

The home page of the OpenNet Initiative at the Berkman Center for Internet and Society, `opennet.net`, has a tool with which you can check which countries block access to your favorite (or least favorite) web site. A summary of findings appears as the book *Access Denied* (MIT Press, 2008).

many, many ways to run the world, run your company ... If you don't like it, *don't participate*. You're here as a volunteer; we didn't force you to come."

You Searched for WHAT? Tracking Searches

IMAGE SEARCH

There are search engines for pictures, and searching for faces presents a different kind of privacy threat. Face recognition by computer has recently become quick and reliable. Computers are now better than people at figuring out which photos are of the same person. With millions of photographs publicly accessible on the Web, all that's needed is a single photo tagged with your name to find others in which you appear. Similar technology makes it possible to find products online using images of similar items. Public image-matching services include `riya.com`, `polarrose.com`, and `like.com`.

Search engine companies can store everything you look for, and everything you click on. In the world of limitless storage capacity, it pays for search companies to keep that data—it might come in handy some day, and it is an important part of the search process. But holding search histories also raises legal and ethical questions. The capacity to retain and analyze query history is another power point—only now the power comes from knowledge about what interests you as an individual, and what interests the population as a whole.

But why would search companies bother to keep every keystroke and click? There are good reasons not to—personal privacy is endangered when such data is retained, as we discuss in Chapter 2. For example, under the USA PATRIOT Act, the federal government could, under certain circumstances, require your search company to reveal what you've been searching for, without ever informing you that it is getting that data. Similar conditions are even easier to imagine in more oppressive countries. Chinese dissidents were imprisoned when Yahoo! turned over their email to the government—in compliance with local laws. Representative Chris Smith asked, "If the secret police a half century ago asked where Anne Frank was hiding, would the correct answer be to hand over the information in order to comply with local laws?" What if the data was not email, but search queries?

From the point of view of the search company, it is easy to understand the reason for retaining your every click. Google founder Sergey Brin says it all

on the company's "Philosophy" page: "The perfect search engine would understand exactly what you mean and give back exactly what you want." Your search history is revealing—and Jen can read your mind much better if she knows what you have been thinking about in the past.

Search quality can improve if search histories are retained. We may prefer, for privacy reasons, that search engines forget everything that has happened, but there would be a price to pay for that—a price in performance to us, and a consequent price in competitiveness to the search company. There is no free lunch, and whatever we may think in theory about Jen keeping track of our search queries, in practice we don't worry about it very much, even when we know.

Even without tying search data to our personal identity, the aggregated search results over time provide valuable data for marketing and economic analysis. Figure 4.12 shows the pattern of Google searches for "iPhone" alongside the identity of certain news stories. The graph shows the number of news stories (among those Google indexes) that mentioned Apple's iPhone. Search has created a new asset: billions of bits of information about *what* people want to know.

You can track trends yourself at `www.google.com/trends`.

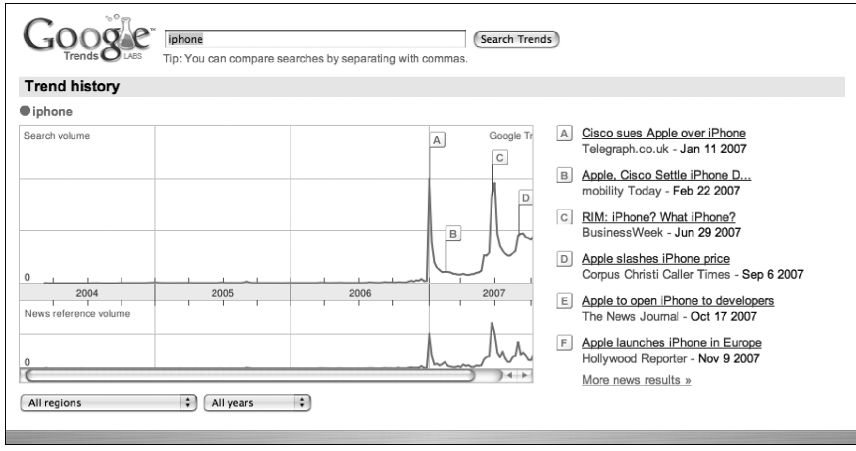


FIGURE 4.12 The top line shows the number of Google searches for "iphone," and the bottom line shows the number of times the iPhone was mentioned in the news sources Google indexes.

Regulating or Replacing the Brokers

Search engines have become a central point of control in a digital world once imagined as a centerless, utopian universe of free-flowing information. The important part of the search story is not about technology or money, although there is plenty of both. It is about power—the power to make things visible, to cause them to exist or to make them disappear, and to control information and access to information.

Search engines have become a central point of control in a digital world once imagined as a centerless, utopian universe of free-flowing information.

Search engines create commercial value not by creating information, but by helping people find it, by understanding what people are interested in finding, and by targeting advertising based on that understanding. Some critics unfairly label this activity "freeloading," as though they themselves could have created a Google had they not preferred to do something more creative (see Chapter 6). It is a remarkable phenomenon: *Information access has greater market value than information creation*. The market capitalization of Google ($157 billion) is more than 50% larger than the combined capitalization of the *New York Times* ($3 billion), Pearson Publishing ($13 billion), eBay ($45 billion), and Macy's ($15 billion). A company providing access to information it did not create has greater market value than those that did the creating. In the bits bazaar, more money is going to the brokers than to the booths.

Open Alternatives

There are hundreds of open source search projects. Because the source of these engines is open, anyone can look at the code and see how it works. Most do not index the whole Web, just a limited piece, because the infrastructure needed for indexing the Web as a whole is too vast. Nutch (`lucene.apache.org/nutch`, `wiki.apache.org/nutch`) is still under development, but already in use for a variety of specialized information domains. Wikia Search, an evolving project of Wikipedia founder Jimmy Wales (`search.wikia.com/wiki/Search_Wikia`), uses Nutch as an engine and promises to draw on community involvement to improve search quality. Moreover, privacy is a founding principle—no identifying data is retained.

The creation and redistribution of power is an unexpected side effect of the search industry. Should any controls be in place, and should anyone (other than services such as `searchenginewatch.com`) watch over the industry? There have been a few proposals for required disclosure of search engine selection and ranking algorithms, but as long as competition remains in the market, such regulation is unlikely to gain traction in the U.S. And competition there is—although Microsoft pled to the FTC that Google was close to "controlling a virtual monopoly share" of Internet advertising. That charge, rejected by the FTC, brought much merriment to some who recalled Microsoft's stout resistance a few years earlier to charges that it had gained monopoly status in desktop software. Things change quickly in the digital world.

METASEARCH

Tools such as `copernic.com`, `surfwax.com`, and `dogpile.com` are *metasearch engines*—they query various search engines and report results back to the user on the basis of their own ranking algorithms. On the freeloading theory of search, they would be freeloading on the freeloaders!

We rely on search engines. But we don't know what they are doing, and there are no easy answers to the question of what to do about it.

French President Jacques Chirac was horrified that the whole world might rely on American search engines as information brokers. To counter the American hegemony, France and Germany announced plans for a state-sponsored search engine in early 2006. As Chirac put it, "We must take up the challenge posed by the American giants Google and Yahoo. For that, we will launch a European search engine, Quaero." The European governments, he explained, would enter this hitherto private-industry sphere "in the image of the magnificent success of Airbus. ... Culture is not merchandise and cannot be left to blind market forces." A year later, Germany dropped out of the alliance, because, according to one industry source, the "Germans apparently got tired of French America-bashing and the idea of developing an alternative to Google."

So for the time being at least, the search engine market rules, and the buyer must beware. And probably that is as it should be. Too often, well-intentioned efforts to regulate technology are far worse than the imagined evils they were intended to prevent. We shall see several examples in the coming chapters.

Search technology, combined with the World Wide Web, has had an astonishing effect on global access to information. The opportunities it presents for limiting information do not overshadow its capacity to enlighten. Things unimaginable barely a decade ago are simple today. We can all find our lost relatives. We can all find new support groups and the latest medical information for our ailments, no matter how obscure. We can even find facts in books we have never held in our hands. Search shines the light of the digital explosion on things we want to make visible.

Encryption technology has the opposite purpose: to make information secret, even though it is communicated over open, public networks. That paradoxical story of politics and mathematics is the subject of the next chapter.