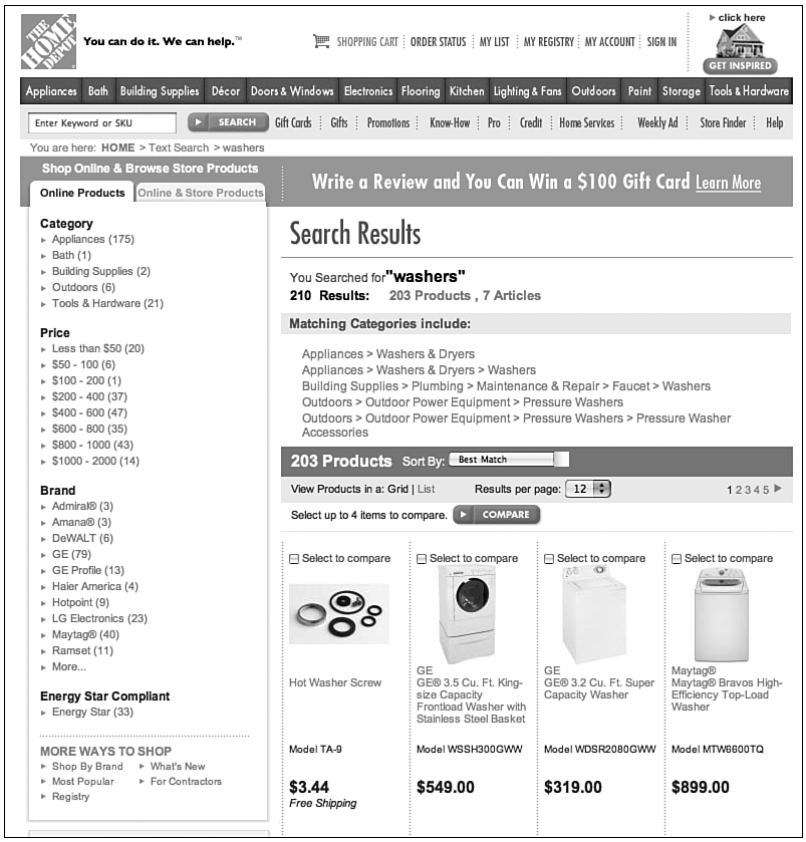


Source: Home Depot.

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.

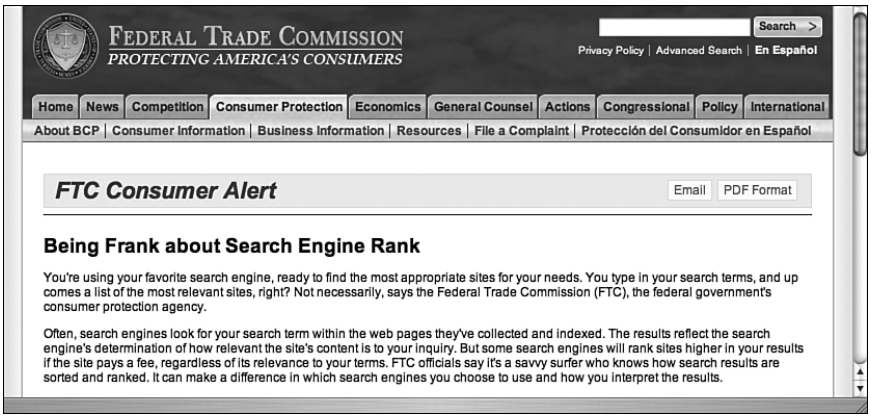
Ironically, the bursting of the Internet bubble in 2000 only made Overture's pay-for-ranking, pay-per-click, keyword auction model more attractive. As profits and capital dried up, Internet businesses could no longer afford up-front capital to buy banner ads, some of which seemed to yield meager results. 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

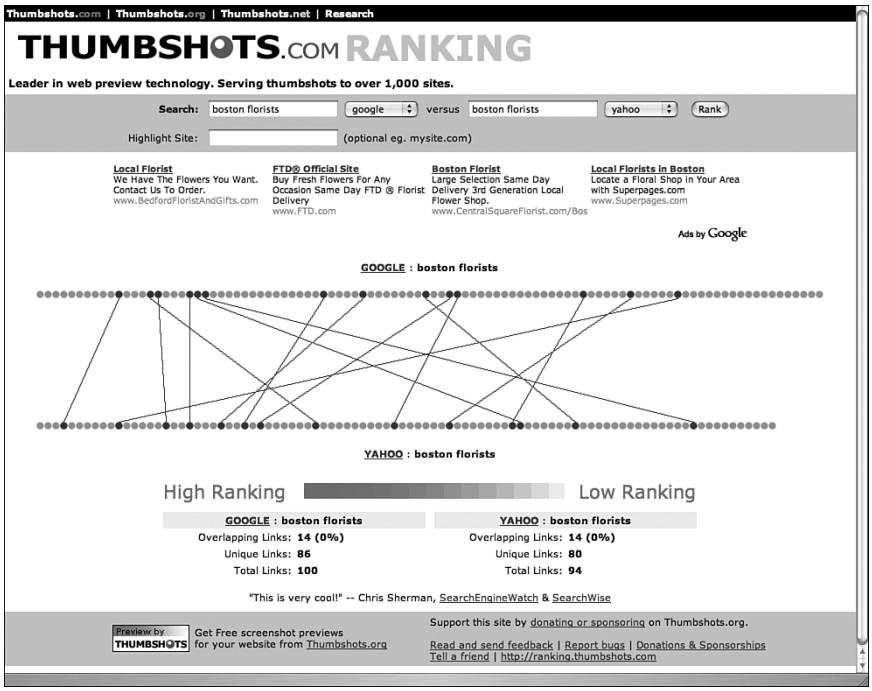
When we use a search engine, we may think that what we are getting is a representative sample of what's available. If so, what we get from one search engine should be pretty close to what we get from another. This is very far from reality.

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.

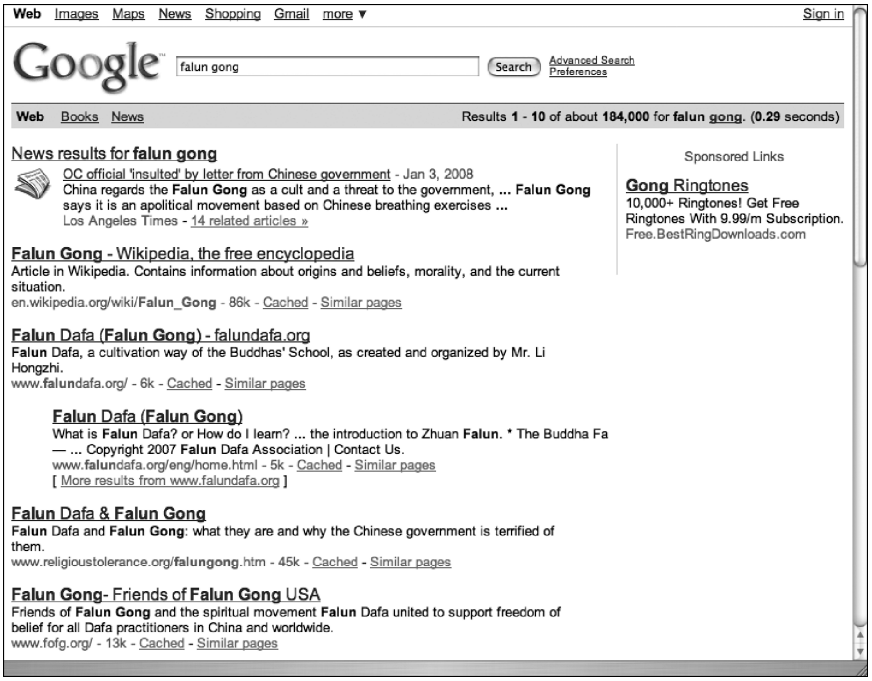
Completely universal accessibility was already more than Google could lawfully accomplish, even in the U.S. If a copyright holder complained that Google was making copyrighted material improperly accessible, Google would respond by removing the link to it from search results. And there were other U.S. 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.

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

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`.

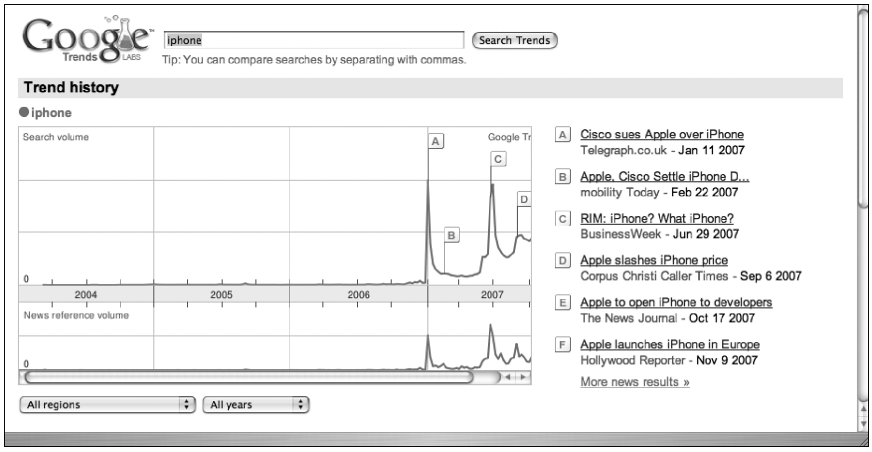


FIGURE 4.12 The top line shows the number of Google searches for "iphone," and the bottom line shows the number of times the iPhone was mentioned in the news sources Google indexes.

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.