POSTCARD FROM SILICON VALLEY

SEARCH AND DEPLOY

The race to build a better search engine.

BY MICHAEL SPECTER

It's not easy to impress the people who fly into Scottsdale, Arizona, each spring to attend the annual PC Forum. The event, organized by the Internet impresario Esther Dyson, is held at a resort near the foot of the McDowell Mountains, and it has become a sort of digital Renaissance Weekend. This year, the conference was so heavily stocked with the fatted calves of the "new" economy—most of them dressed in the casual E-commerce outfit of khaki pants and blue oxford shirt—that controllers at Scottsdale's tiny airport struggled to accommodate all the corporate jets.

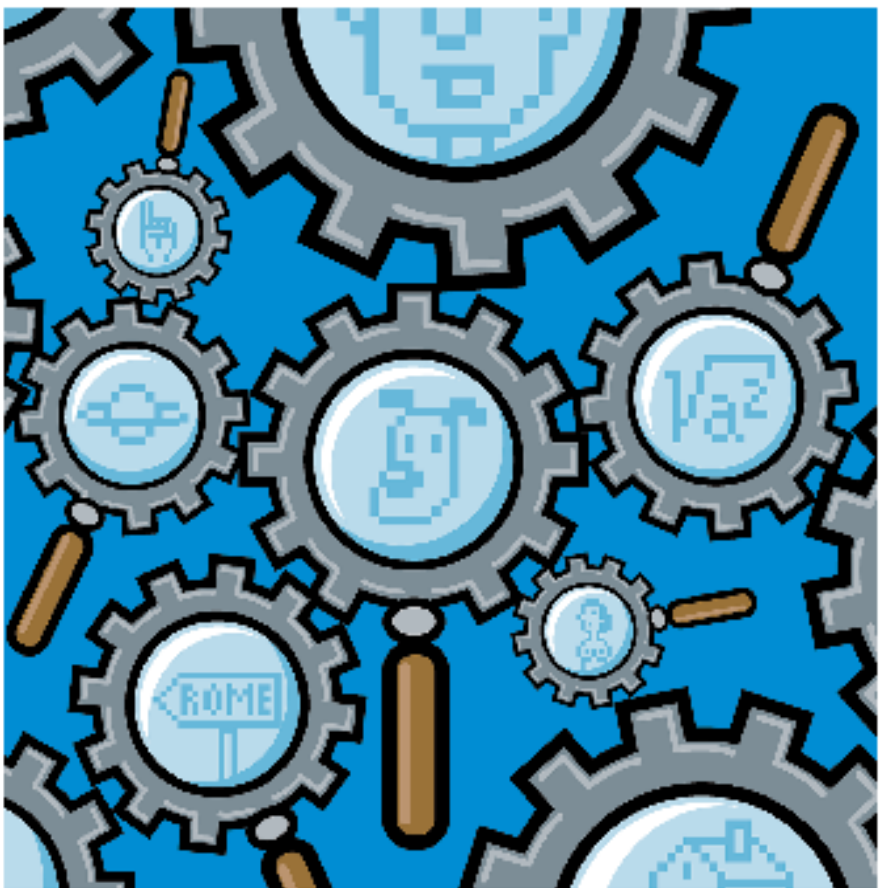
The Dyson conference began as a specialized gathering twenty-three years ago, when the Web was mostly a military secret. Like the Internet itself, however, the PC Forum has spread far beyond its initial boundaries. (It retains its quaint name as a reminder of what it was; these days, PCs are beside the point.) "We try hard to keep the meetings from becoming just a survey of what's going on," Ms. Dyson told me, when I asked her how she decided what to focus on each year.

Order is on Dyson's mind at the moment because the Internet has become so resistant to discipline. There are now more than a billion pages on the World Wide Web, all loosely tied together by seven billion annotated links, called hyperlinks, which is at least one link for every person on the planet. Each day, more than a million pages are added, and a page can appear in any language, written by any person, for any reason; it can be three lines long or the length of the Bible. For the first time in history, people everywhere have access to the thoughts, products, and writing of a large—and growing—percentage of the earth's population.

This much democracy can be daunting. As more information fills the Web, and more people become dependent upon it, search engines—programs that hunt for Web pages in response to specific words or phrases—have become overwhelmed. "People are beginning to feel a little lost in all this opportunity," Dyson said. "For the Internet to work, and to be liberating, it has to be easy to use."

Too often, however, it is not. When I type "How do you skin a mule?" into most search engines, for example, I get thousands of answers—and they refer to everything from drug dealers to shoes to skin color and radiation treatment for a variety of cancers. Most answers are useless. If I want to know the current population of Rome and type the phrase "population of Rome" into Infoseek, a well-known search engine, approximately twenty-nine thousand replies come back, and among the first are pages on "world depopulation and slavery" and "the history of the white race," and a page titled "The Cure of the Neurobiological Sickness of Religion, Part 2."

The reason for the muddle is simple: most search engines are programmed to unleash software called "spiders," which systematically crawl through the Web sucking up every link on every page. When they have digested what they have found, the spiders generate indexes of the words and the links. So even if the word "population" appears in a sentence about ancient Greece, and the word "Rome" appears far away on the same page, perhaps in an advertisement for a hotel in upstate New York, most search engines would consider the page relevant. It would have been easier to track down Rome's chief demographer. "It's ironic, but, the bigger the Internet gets, the more difficult it is to find a simple, accurate answer to your questions," Lawrence Page told me before the first major presentation,



Google's PageRank system made it the default search engine for the digital in-crowd.

CHRISTOPH NIEMANN

TNY—5/29/00—PAGE 88—LIVE OPI ART—NIEMANNSEARCHMACHINE—133SC.

on navigating the World Wide Web, at this year's PC Forum.

At the age of twenty-seven, Page runs—with a fellow former Stanford graduate student, Sergey Brin—a small company, based in Mountain View, California, named Google, which has become the default search engine of the digital in-crowd. "The more information there is out there, the more likely you are to get junk or lies for an answer," Page told me. "You want relevant information, but you are fighting with chaos."

The moderator of the presentation, Kevin Werbach, was having trouble getting the audience to focus, because everyone was distracted by a series of seemingly unrelated phrases scrolling by on a giant screen: "Fishing boats. Lesson plans format. ICRA. Woodpecker control. Origin of God." (Google, which derives from the word "googol"—the numeral one followed by a hundred zeros—had set up a live feed of the thirteen million queries that it gets each day.) "Unholy dancers. Drug testing in high schools. Compulsive hoarding. Free wife-swapping stories. Bald. Shaved."

"Let's go to the panel," Werbach said as the scroll continued. "Hopefully, it will be more interesting than seeing the queries." That produced a chorus of boos, because it's hard to imagine a computer conference generating anything more exciting than the thrill of watching what the world is trying to find out.

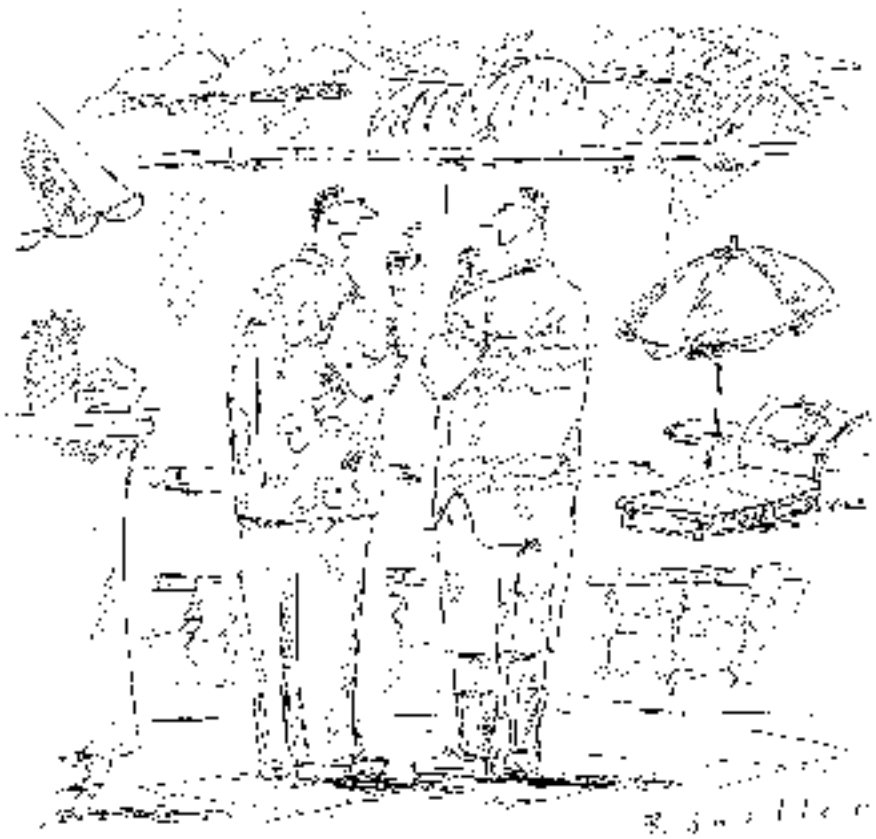
A few days after the conference ended, I walked into the Gates Computer Science Building at Stanford University. It is a gaudy place on a campus that works hard at being sedate, and it is where Page and Brin were working toward Ph.D.s when they thought up the idea of Google. I had come to see Rajeev Motwani, an associate professor in the computer-science department and the author of a standard work on computational algorithms—the mathematical recipes that make software work. Motwani, a cheerful thirty-seven-year-old man with short black hair, a mustache, and eyes the color of wet coal, has spent a lot of his recent career trying to figure out a better way to search.

Before the Internet, there were electronic information services—like Lexis-Nexis—but they have always been narrowly focussed, expensive, and, for most people, difficult to deploy. When the World Wide Web came into popular use, people realized that search engines were a powerful tool. Most people, though, never understood that the searches were limited and that the quality of the results varied greatly. (This is still true; even the largest of the search engines, Inktomi, has indexed only about half the Web. So far, the rest is dark matter; if the page you want is trapped there, it doesn't make any difference which search engine you use.)

It is common knowledge that if a search fails to retrieve relevant information within a couple of seconds, most surfers will click away and try someplace else. In those few seconds, as the engine crawls through millions of links, many problems need to be solved—the biggest of which is called "the verbal-disagreement problem." Verbal disagreement means that if you have a certain concept in mind and you ask two people to describe it, they will give you two completely different, but entirely correct, words. Conversely, two people using the same word could be talking about entirely different concepts. The Internet magnifies that problem immensely. If you search for the word "automobile" on the Web, for instance, you are likely to miss many pages that use the word "car" instead. "Search engines are far better than even five years ago," Motwani told me. "But most of them are still like primitive buzz saws cutting down giant forests to look for a single tree. If you ask me if they are delivering the way I think they should, I would say we are at Step One in a ten-step process."

Internet search has evolved rapidly since 1993, when a program called WebCrawler became the first widely used search engine. These days, there are about two dozen major search engines, most of which rank Web sites based on their contents. Yahoo!, which is probably the most popular, isn't really a search engine at all. It employs a team of editors to index the Internet; if you want a page to show up in a Yahoo! search, you must submit a form with information about the site.

Some people will do almost anything to receive a top ranking from a heavily used search engine, and it's easy to understand why: the first response in a search will bring more viewers, more



"Meritocracy worked for my grandfather, it worked for my father, and it's working for me."

TNY—5/29/00—PAGE 90—LIVE OPI ART #A5446—120 LS.

business—and the sort of prominence that gets a site ranked more highly by other search engines. The ploys people use to get there are often deceptive. Pages can repeat words many times in invisible type (masked in a color that is the same as the color of the page) so that the search engine picks them up and ranks them as more relevant than it otherwise would. For example, some automobile Web sites have stooped to writing "BUY THIS CAR" dozens of times in hidden fonts. That way, a search engine will count the words "buy" and "car" and rate it highly—a subliminal version of listing AAAA Autos in the Yellow Pages.

The most direct way to get your Web site to the top of a search—and the most pernicious—is to pay for it. At GoTo, a popular search engine, payment is routine. As the Internet newsletter *Search Engine Watch* has pointed out, "A company might bid on the word 'travel,' agreeing to pay twenty-five cents per click. If no one agrees to pay more than this, then your company would occupy the top spot—and every time someone clicked on your link, you'd owe GoTo twenty-five cents." That's your "cost per click," and for a much frequented travel site it's a bargain. (At other engines, you can pay for how many times somebody sees your ad, rather than just for clicks.)

As a result, if you type "Harvard" into GoTo, you won't get to the Harvard University home page until you have seen links for Gradschools.com and Harvard Hotels, among others. The people who run search engines say you need to deliver the goods within the first ten entries, but at GoTo the Harvard home page is No. 14. At Infoseek and Google, neither of which takes money for placement, the Harvard home page comes up first.

Motwani knew that for a search to be more effective it would have to move beyond lists and pay for placement. "The Web is a network of hyperlinks, and this network is sometimes called a graph," he said. "If someone goes to the effort of introducing a hyperlink into a Web page, you ought to be able to make judgments about it."

What Motwani and several other researchers recognized was that one could look at surfing around the Web as similar to taking a random walk on a giant

TNY—5/29/00—PAGE 91—#2 PAGE!!!

grid, sort of like wandering aimlessly around Manhattan. If you pick a starting point at random, click on a series of random hyperlinks, and watch long enough as people surf around, you can make statistical statements about how likely it is that a person will end up at any particular site.

"I understood all this, and so did many other people," Motwani said, smiling sheepishly. "But I didn't see the implications." Lawrence Page and Sergey Brin did. "They had this idea, a new way to look at the links on the Web. Other people had thought about link structure, of course. But they took it further. All of a sudden, we were no longer talking about Web pages. We were talking about a giant community, and each link was a relationship between members of that community."

The system, which Page called PageRank, permitted Brin and Page to improve on the standard practice of counting how often a key word appears on a Web site. They realized that if a page is linked to many other pages it's like a vote—the collective voice of the Web has decided that the page has a certain value. If millions of people link to a page, it's a good endorsement. It doesn't mean that the link is accurate, but it's likely to be a more useful authority than a page nobody points to. Page and Brin realized that it was possible to map the Web and rate pages primarily by analyzing links instead of words. (In fact, they are so confident of Google's accuracy that they put an "I'm Feeling Lucky" button on their page. Click on it, and you go directly to the highest-ranked site for your search.)

Such searches can require millions of computations, but essentially the rating you get is based on who "voted" for you by establishing links to your site. (The engine also looks at how many votes were cast for the pages that were linked to those pages. If the home page of the *Times* links to your page, you will be ranked more highly than if, say, just your cousin Harvey links to your home page. That's because many other pages link to the *Times*, so it brings in lots of votes.) "Before this, people were just looking at the content," Motwani told me. "They were completely ignoring the fact that people were going to the effort of putting a link from one page to another and that there must be a meaning to that."

Google is not the first search engine to look at the links on the page; Excite and Lycos have also done it. But Page and Brin's Google has raised the bar. "Their system just works much better than anybody else's does," Danny Sullivan, the editor of *Search Engine Watch,* told me. "Now every major search engine will have to use it. Nobody can afford to do anything less."

I tried it out. I typed "population of Rome" into Google. The program did a quick calculation of the value of all the pages with those words on it, assessed the links that connected them, and figured out the relative value of each page on which the words appeared. It then looked at the position of the words on the page, the size of the fonts, and the likelihood that the words were related to each other. That took 0.38 seconds. By then, I had a list of eighty-four thousand possible responses. It wasn't a perfect search; Google had no way of knowing whether I meant ancient or modern Rome. Unlike any other search engine I tried, however, Google did address my query about the population of Rome, Italy. (My first ten responses in Yahoo!, on the other hand, included two entries for Rome, Maine, and one for Rome, New York. The first mention of the Rome I had in mind was on a page entitled "Xiphoid's Rise of Rome Conpluvium." AltaVista wasn't much better. It had nothing of direct value in its first ten responses, one of which was the home page of a Baltimore real-estate agent whose last name is Rome.)

Google can be fooled, of course. Anybody who takes the trouble to set up a group of pages with links to each other can force his way into the rankings, with some rather odd results. Brin told me to type in the phrase "more evil than Satan itself," and the first response was the Microsoft home page. (The response just shows that there are many people on the Web who seem to use the words "evil" and "Satan" when referring to Micro-

TNY—5/29/00—PAGE 92

"Whose level do you want to stoop to tonight, mine or yours?"

• •

soft—and that they tend to link to each other.) "It's not a trick," Brin said one evening. "But if you want to just say it's possible to get bad information on Google, I'll understand. It's possible to get bad information anywhere."

Google's offices are spread through a sort of dot-com strip mall not far from Palo Alto. It's a graduate-student Disneyland, filled with Rollerblades and assorted hockey paraphernalia for twice-weekly company hockey games. The offices are stocked with enough free M&M's, PowerBars, barrels of granola, urns of coffee, and coolers of fruit juice to drive anybody through to 4 A.M.—which is not an unusual time to find people in the office. Not everything is in place yet, though. When I visited, a baby-grand piano and a new espresso bar were both on order, so the lobby looked a little bare.

A gym is on the lower floor, next to a sauna and a room for massages. There is a massage therapist on site every day—and all employees are encouraged to make use of her services. The biggest perk, however, is the cafeteria. Page and Brin have hired an accomplished chef—he formerly worked for the Grateful Dead—to cook organic meals. The food is free, and all employees are fed lunch and dinner (and so are friends and family members who wish to join them, as long as the chef is given advance notice).

I had lunch one day with a few of the company's researchers, including Jim Reese, who told me that he was employee No. 19. His business card de-

TNY—5/29/00—PAGE 97—LIVE OPI—A5540—120 LS

scribes him as chief operations engineer and head neurosurgeon. That's because, before coming to Google, he was a neurosurgeon, at Stanford. Reese spends a lot of his time at Google's "server farms," warehouses filled with computers that have the fastest connections to the Internet. One of Google's facilities is run by a company called Exodus, in nearby Santa Clara, and Google stores some of its network of nearly four thousand Linux computers there, each with eighty gigabytes of hard-drive space, on which it keeps constant downloads of the Web. (Many other companies, including Hotmail and eBay, use Exodus as their electronic storage vault.)

Reese told me that he has been too busy lately to bother with newspapers or television. He gets his news by watching the questions people ask Google in search queries. "Usually, the most popular queries are sex and MP3," he said. "One day it will be sex, the next MP3. But you can sort of gauge important events by looking at the queries. The day after the Grammys, for instance, we were getting tons of hits that involved the words 'Jennifer Lopez' and 'dress' and 'naked.' " I asked him if there were many requests for information with the words 'Gore' or 'Bush' or 'campaign' in them. "Nope," he said, and laughed. "It's a sad comment to make, but nobody seems interested."

About ten per cent of Google queries are for pornography. The figure is lower than that of most other search engines. This reflects the demographics of the people who use the search engine, but perhaps it also demonstrates one of Google's obvious failings: porn sites are sought out by millions of Internet users but are rarely linked to prominent Web pages. Without links, even the most popular page is invisible.

If you add up the ages of Google's founders, it comes to fifty-three—younger than the average age of a C.E.O. of a major company that doesn't have "dot" or "com" in its name. Page and Brin are pleasantly dishevelled workaholics who find it amazing that they don't have to subsist on burritos. The company has not yet gone public; Brin and Page each take eighty thousand dollars a year in salary, which, as Brin pointed out, is more than eighty times what he was making while he was in graduate school.

Brin's family came to America from Russia when he was six. His father teaches math at the University of Maryland, and his mother works at the Goddard Space Flight Center at NASA. Page's father, who died a few years ago, was a computer-science professor at Michigan State. Page was one of those kids who spend their youth taking everything in the house apart. When he and Brin met, at Stanford, they had complementary interests in computers. "I was working on the link structure of the Web," Page said. "A sort of mathematical problem about which pages pointed to which other pages. That's all I was doing. Sergey was working on data mining. He was looking at how useful information could be extracted from large quantities of information."

It didn't take long for them to attract backers. Stanford has put money into Google, as have the venture-capital firms Sequoia Capital and Kleiner Perkins Caufield & Byers. The Sun Microsystems co-founder Andy Bechtolsheim is an investor, too. Still, one of Google's draws is also its biggest liability: all it does is search. There usually isn't much money in that, which is why so many search engines—like AltaVista, Infoseek, Excite, HotBot, and, above all, Yahoo!—have become Web portals where you are encouraged to chat with friends, use E-mail, and look at news wires or stock prices.

Page plans to sell his service to portals like Yahoo! and Microsoft, which would pay Google a fee based on how many of their searches Google manages to complete. It already has an arrangement with such partners as Netscape and the Washington *Post.* Advertising has increased sharply this year—largely because users have, too. So far, Google permits ads to appear only in text form, since text loads faster than graphics, and the company allows no more than two to appear on any page. "We want to be the fastest search engine," Brin told me. "The fastest and the best."

There seems to be a generation of people for whom the Internet is the principal source of information about the world. When they need to solve a problem or answer a question, they go to the Web, and that is where they find

TNY—5/29/00—PAGE 98—LIVE OPI ART #N17290—PLEASE INSPECT.

"reality"—even though the Web often confuses what is "true" with what is "popular." (In "The Economic Analysis of Law," Richard Posner observed, "The true utterance is like the brand of beer that commands ninety-five per cent of the market and the false like the brand with only five per cent.") If you ask most search engines how many home runs Mickey Mantle hit in 1958, you will get some answers that are right and some that are wrong; on the Web, where fantasy-baseball sites are at least as popular as Yankee statistics, it is hard to distinguish what is popular from what is true. "That is the greatest challenge," Andrew Tomkins told me. "Making the truth shine through."

Tomkins, who recently received a Ph.D. from Carnegie Mellon, is a researcher on the I.B.M. Clever Project, a search engine that so far is used only at the I.B.M. Almaden Research Center, in San Jose, California. Clever is similar in approach to Google—it looks at links and not just at key words—but it may yet produce a more finely tuned way to find information and assess it. Where Google essentially assigns a fixed value to all links, based on how highly other links value them, Clever's rating allows the links to shift in value depending on the search request.

Clever's analysis follows from this sociological observation: the Web contains many pages filled with useful pointers to specific information. Someone interested in fishing can find plenty of pages with titles like "My Fishing Links." In a traditional search engine, you would type "fishing" and get back a considerable amount of useless information. "Eventually, though, you would probably find a valuable page," Tomkins told me. "Call it 'Joe's fishing links.' Joe is not the guy who won the bass master classic, but he is a grad student in some place and he loves to fish. He is enthusiastic and he has the perseverance to keep his page up to date, and he is really versed in the on-line fishing community. So he created this page with a bunch of links. About fishing. So it's a familiar experience to find a page filled with these useful links. And when you see it you say, 'Ah, finally,' and maybe you bookmark it. This we call a 'hub page.'

"Just through the evolution of the Web, these pages are all over the place," Tomkins continued. "And they are there on every conceivable topic. We found really good hubs on oil spills off the coast of Japan. And on Australian fire brigades—and on people who go off into the woods on the weekend and wear inflatable sumo costumes and wrestle. Clever tests each link, analyzes the text on the pages, and looks for key words."

Then, unlike Google, it analyzes the hubs to discover "authorities"—pages that on-line fishing experts regard as the most useful and interesting—and uses the authorities to help judge the quality of the hubs. Emerging from all that is what Tomkins describes as "the footprint of a community," and he goes on, "The surprising thing is that as the number of pages grows—the billions, zillions, trillions—the number of these communities that emerge from random association shrinks. I decide that it's really important to me to find out wherever fish turn up in stained-glass windows. I find a picture of some stained-glass windows and create a Web page. This is my page with the links to stained-glass windows on it. Nobody cares. Then in Siberia there is some guy who happens to have the same interest, and he creates a page that also links to that stuff. And some other similar stuff. And as soon as that happens we find it. Because I link to these pages. And he links to these pages as well. Even though neither of us knows there is a community on this topic, we can find it and use it in any way we want. This is a way to understand the emergence of low-level grassroots sort of things. We can see patterns as they are developing, trends, ideas, communities. That really could be powerful. It could be beyond search. It could give people what they are looking for."

Over at Google, Page and Brin also wonder whether Clever will be what people are looking for. "It's a good approach," Page told me. The two systems "were conceived in similar ways. But Clever uses additional information that is very prone to manipulation—or spam—by people trying to mislead the search engine for commercial gain." Page went on, "The great thing about search is that we are not going to solve it any time soon. There are so many problems and failings. I see no end to what we need to do. If we aren't a lot better next year, we will already be forgotten." ♦

TNY—5/29/00—PAGE 100