
June 3, 2007

Google Keeps Tweaking Its Search Engine

By [SAUL HANSELL](#)

Mountain View, Calif.

THESE days, [Google](#) seems to be doing everything, everywhere. It takes pictures of your house from outer space, copies rare Sanskrit books in India, charms its way onto Madison Avenue, picks fights with Hollywood and tries to undercut [Microsoft](#)'s software dominance.

But at its core, Google remains a search engine. And its search pages, blue hyperlinks set against a bland, white background, have made it the most visited, most profitable and arguably the most powerful company on the Internet. Google is the homework helper, navigator and yellow pages for half a billion users, able to find the most improbable needles in the world's largest haystack of information in just the blink of an eye.

Yet however easy it is to wax poetic about the modern-day miracle of Google, the site is also among the world's biggest teases. Millions of times a day, users click away from Google, disappointed that they couldn't find the hotel, the recipe or the background of that hot guy. Google often finds what users want, but it doesn't always.

That's why Amit Singhal and hundreds of other Google engineers are constantly tweaking the company's search engine in an elusive quest to close the gap between often and always.

Mr. Singhal is the master of what Google calls its "ranking algorithm" — the formulas that decide which Web pages best answer each user's question. It is a crucial part of Google's inner sanctum, a department called "search quality" that the company treats like a state secret. Google rarely allows outsiders to visit the unit, and it has been cautious about allowing Mr. Singhal to speak with the news media about the magical, mathematical brew inside the millions of black boxes that power its search engine.

Google values Mr. Singhal and his team so highly for the most basic of competitive reasons. It believes that its ability to decrease the number of times it leaves searchers disappointed is crucial to fending off ever fiercer attacks from the likes of [Yahoo](#) and Microsoft and preserving the tidy advertising gold mine that search represents.

"The fundamental value created by Google is the ranking," says John Battelle, the chief executive of Federated Media, a blog ad network, and author of "The Search," a book about Google.

Online stores, he notes, find that a quarter to a half of their visitors, and most of their new customers,

come from search engines. And media sites are discovering that many people are ignoring their home pages — where ad rates are typically highest — and using Google to jump to the specific pages they want.

“Google has become the lifeblood of the Internet,” Mr. Battelle says. “You have to be in it.”

Users, of course, don’t see the science and the artistry that makes Google’s black boxes hum, but the search-quality team makes about a half-dozen major and minor changes a week to the vast nest of mathematical formulas that power the search engine.

These formulas have grown better at reading the minds of users to interpret a very short query. Are the users looking for a job, a purchase or a fact? The formulas can tell that people who type “apples” are likely to be thinking about fruit, while those who type “Apple” are mulling computers or iPods. They can even compensate for vaguely worded queries or outright mistakes.

“Search over the last few years has moved from ‘Give me what I typed’ to ‘Give me what I want,’ ” says Mr. Singhal, a 39-year-old native of India who joined Google in 2000 and is now a Google Fellow, the designation the company reserves for its elite engineers.

Google recently allowed a reporter from The New York Times to spend a day with Mr. Singhal and others in the search-quality team, observing some internal meetings and talking to several top engineers. There were many questions that Google wouldn’t answer. But the engineers still explained more than they ever have before in the news media about how their search system works.

As Google constantly fine-tunes its search engine, one challenge it faces is sheer scale. It is now the most popular Web site in the world, offering its services in 112 languages, indexing tens of billions of Web pages and handling hundreds of millions of queries a day.

Even more daunting, many of those pages are shams created by hucksters trying to lure Web surfers to their sites filled with ads, pornography or financial scams. At the same time, users have come to expect that Google can sift through all that data and find what they are seeking, with just a few words as clues.

“Expectations are higher now,” said Udi Manber, who oversees Google’s entire search-quality group. “When search first started, if you searched for something and you found it, it was a miracle. Now, if you don’t get exactly what you want in the first three results, something is wrong.”

Google’s approach to search reflects its unconventional management practices. It has hundreds of engineers, including leading experts in search lured from academia, loosely organized and working on projects that interest them. But when it comes to the search engine — which has many thousands of interlocking equations — it has to double-check the engineers’ independent work with objective, quantitative rigor to ensure that new formulas don’t do more harm than good.

As always, tweaking and quality control involve a balancing act. “You make a change, and it affects some queries positively and others negatively,” Mr. Manber says. “You can’t only launch things that are 100

percent positive.”

THE epicenter of Google’s frantic quest for perfect links is Building 43 in the heart of the company’s headquarters here, known as the Googleplex. In a nod to the space-travel fascination of Larry Page, the Google co-founder, a full-scale replica of SpaceShipOne, the first privately financed spacecraft, dominates the building’s lobby. The spaceship is also a tangible reminder that despite its pedestrian uses — finding the dry cleaner’s address or checking out a prospective boyfriend — what Google does is akin to rocket science.

At the top of a bright chartreuse staircase in Building 43 is the office that Mr. Singhal shares with three other top engineers. It is littered with plastic light sabers, foam swords and Nerf guns. A big white board near Mr. Singhal’s desk is scrawled with graphs, queries and bits of multicolored mathematical algorithms. Complaints from users about searches gone awry are also scrawled on the board.

Any of Google’s 10,000 employees can use its “Buganizer” system to report a search problem, and about 100 times a day they do — listing Mr. Singhal as the person responsible to squash them.

“Someone brings a query that is broken to Amit, and he treasures it and cherishes it and tries to figure out how to fix the algorithm,” says Matt Cutts, one of Mr. Singhal’s officemates and the head of Google’s efforts to fight Web spam, the term for advertising-filled pages that somehow keep maneuvering to the top of search listings.

Some complaints involve simple flaws that need to be fixed right away. Recently, a search for “French Revolution” returned too many sites about the recent French presidential election campaign — in which candidates opined on various policy revolutions — rather than the ouster of King Louis XVI. A search-engine tweak gave more weight to pages with phrases like “French Revolution” rather than pages that simply had both words.

At other times, complaints highlight more complex problems. In 2005, Bill Brougher, a Google product manager, complained that typing the phrase “teak patio Palo Alto” didn’t return a local store called the Teak Patio.

So Mr. Singhal fired up one of Google’s prized and closely guarded internal programs, called Debug, which shows how its computers evaluate each query and each Web page. He discovered that [Theteakpatio.com](http://theteakpatio.com) did not show up because Google’s formulas were not giving enough importance to links from other sites about Palo Alto.

It was also a clue to a bigger problem. Finding local businesses is important to users, but Google often has to rely on only a handful of sites for clues about which businesses are best. Within two months of Mr. Brougher’s complaint, Mr. Singhal’s group had written a new mathematical formula to handle queries for hometown shops.

But Mr. Singhal often doesn’t rush to fix everything he hears about, because each change can affect the

rankings of many sites. “You can’t just react on the first complaint,” he says. “You let things simmer.”

So he monitors complaints on his white board, prioritizing them if they keep coming back. For much of the second half of last year, one of the recurring items was “freshness.”

Freshness, which describes how many recently created or changed pages are included in a search result, is at the center of a constant debate in search: Is it better to provide new information or to display pages that have stood the test of time and are more likely to be of higher quality? Until now, Google has preferred pages old enough to attract others to link to them.

But last year, Mr. Singhal started to worry that Google’s balance was off. When the company introduced its new stock quotation service, a search for “Google Finance” couldn’t find it. After monitoring similar problems, he assembled a team of three engineers to figure out what to do about them.

Earlier this spring, he brought his squad’s findings to Mr. Manber’s weekly gathering of top search-quality engineers who review major projects. At the meeting, a dozen people sat around a large table, another dozen sprawled on red couches, and two more beamed in from New York via video conference, their images projected on a large screen. Most were men, and many were tapping away on laptops. One of the New Yorkers munched on cake.

Mr. Singhal introduced the freshness problem, explaining that simply changing formulas to display more new pages results in lower-quality searches much of the time. He then unveiled his team’s solution: a mathematical model that tries to determine when users want new information and when they don’t. (And yes, like all Google initiatives, it had a name: QDF, for “query deserves freshness.”)

Mr. Manber’s group questioned QDF’s formula and how it could be deployed. At the end of the meeting, Mr. Singhal said he expected to begin testing it on Google users in one of the company’s data centers within two weeks. An engineer wondered whether that was too ambitious.

“What do you take us for, slackers?” Mr. Singhal responded with a rebellious smile.

THE QDF solution revolves around determining whether a topic is “hot.” If news sites or blog posts are actively writing about a topic, the model figures that it is one for which users are more likely to want current information. The model also examines Google’s own stream of billions of search queries, which Mr. Singhal believes is an even better monitor of global enthusiasm about a particular subject.

As an example, he points out what happens when cities suffer power failures. “When there is a blackout in New York, the first articles appear in 15 minutes; we get queries in two seconds,” he says.

Mr. Singhal says he tested QDF for a simple application: deciding whether to include a few news headlines among regular results when people do searches for topics with high QDF scores. Although Google already has a different system for including headlines on some search pages, QDF offered more sophisticated results, putting the headlines at the top of the page for some queries, and putting them in

the middle or at the bottom for others.

GOOGLE'S breakneck pace contrasts with the more leisurely style of the universities and corporate research labs from which many of its leaders hail. Google recruited Mr. Singhal from [AT&T Labs](#). Mr. Manber, a native of Israel, was an early examiner of Internet searches while teaching computer science at the [University of Arizona](#). He jumped into the corporate fray early, first as Yahoo's chief scientist and then running an [Amazon.com](#) search unit.

Google lured Mr. Manber from Amazon last year. When he arrived and began to look inside the company's black boxes, he says, he was surprised that Google's methods were so far ahead of those of academic researchers and corporate rivals.

"I spent the first three months saying, 'I have an idea,' " he recalls. "And they'd say, 'We've thought of that and it's already in there,' or 'It doesn't work.' "

The reticent Mr. Manber (he declines to give his age), would discuss his search-quality group only in the vaguest of terms. It operates in small teams of engineers. Some, like Mr. Singhal's, focus on systems that process queries after users type them in. Others work on features that improve the display of results, like extracting snippets — the short, descriptive text that gives users a hint about a site's content.

Other members of Mr. Manber's team work on what happens before users can even start a search: maintaining a giant index of all the world's Web pages. Google has hundreds of thousands of customized computers scouring the Web to serve that purpose. In its early years, Google built a new index every six to eight weeks. Now it rechecks many pages every few days.

And Google does more than simply build an outsized, digital table of contents for the Web. Instead, it actually makes a copy of the entire Internet — every word on every page — that it stores in each of its huge customized data centers so it can comb through the information faster. Google recently developed a new system that can hold far more data and search through it far faster than the company could before.

As Google compiles its index, it calculates a number it calls PageRank for each page it finds. This was the key invention of Google's founders, Mr. Page and Sergey Brin. PageRank tallies how many times other sites link to a given page. Sites that are more popular, especially with sites that have high PageRanks themselves, are considered likely to be of higher quality.

Mr. Singhal has developed a far more elaborate system for ranking pages, which involves more than 200 types of information, or what Google calls "signals." PageRank is but one signal. Some signals are on Web pages — like words, links, images and so on. Some are drawn from the history of how pages have changed over time. Some signals are data patterns uncovered in the trillions of searches that Google has handled over the years.

"The data we have is pushing the state of the art," Mr. Singhal says. "We see all the links going to a page, how the content is changing on the page over time."

Increasingly, Google is using signals that come from its history of what individual users have searched for in the past, in order to offer results that reflect each person's interests. For example, a search for "dolphins" will return different results for a user who is a Miami football fan than for a user who is a marine biologist. This works only for users who sign into one of Google's services, like Gmail.

(Google says it goes out of its way to prevent access to its growing store of individual user preferences and patterns. But the vast breadth and detail of such records is prompting lust among the nosey and fears among privacy advocates.)

Once Google corrals its myriad signals, it feeds them into formulas it calls classifiers that try to infer useful information about the type of search, in order to send the user to the most helpful pages. Classifiers can tell, for example, whether someone is searching for a product to buy, or for information about a place, a company or a person. Google recently developed a new classifier to identify names of people who aren't famous. Another identifies brand names.

These signals and classifiers calculate several key measures of a page's relevance, including one it calls "topicality" — a measure of how the topic of a page relates to the broad category of the user's query. A page about President Bush's speech about Darfur last week at the White House, for example, would rank high in topicality for "Darfur," less so for "George Bush" and even less for "White House." Google combines all these measures into a final relevancy score.

The sites with the 10 highest scores win the coveted spots on the first search page, unless a final check shows that there is not enough "diversity" in the results. "If you have a lot of different perspectives on one page, often that is more helpful than if the page is dominated by one perspective," Mr. Cutts says. "If someone types a product, for example, maybe you want a blog review of it, a manufacturer's page, a place to buy it or a comparison shopping site."

If this wasn't excruciating enough, Google's engineers must compensate for users who are not only fickle, but are also vague about what they want; often, they type in ambiguous phrases or misspelled words.

Long ago, Google figured out that users who type "Brittany Speers," for example, are really searching for "[Britney Spears](#)." To tackle such a problem, it built a system that understands variations of words. So elegant and powerful is that model that it can look for pages when only an abbreviation or synonym is typed in.

Mr. Singhal boasts that the query "Brenda Lee bio" returns the official home page of the singer, even though the home page itself uses the term "biography" — not "bio."

But words that seem related sometimes are not related. "We know 'bio' is the same as 'biography,'" Mr. Singhal says. "My grandmother says: 'Oh, come on. Isn't that obvious?' It's hard to explain to her that bio means the same as biography, but 'apples' doesn't mean the same as '[Apple](#).'"

In the end, it's hard to gauge exactly how advanced Google's techniques are, because so much of what it

and its search rivals do is veiled in secrecy. In a look at the results, the differences between the leading search engines are subtle, although Danny Sullivan, a veteran search specialist and blogger who runs Searchengineland.com, says Google continues to outpace its competitors.

Yahoo is now developing special search formulas for specific areas of knowledge, like health. Microsoft has bet on using a mathematical technique to rank pages known as neural networks that try to mimic the way human brains learn information.

Google's use of signals and classifiers, by contrast, is more rooted in current academic literature, in part because its leaders come from academia and research labs. Still, Google has been able to refine and advance those ideas by using computer and programming resources that no university can afford.

“People still think that Google is the gold standard of search,” Mr. Battelle says. “Their secret sauce is how these guys are doing it all in aggregate. There are 1,000 little tunings they do.”

[Copyright 2007 The New York Times Company](#)

[Privacy Policy](#) | [Search](#) | [Corrections](#) | [RSS](#) | [First Look](#) | [Help](#) | [Contact Us](#) | [Work for Us](#) | [Site Map](#)
