- Follow Wired
- Twitter
- Facebook
- RSS

Exclusive: How Google's Algorithm Rules the Web

- By Steven Levy
- 02.22.10 |
- 12:00 pm |
- Permalink

- Share on Facebook
  0
- Tweet 1
- 10
- Share
-

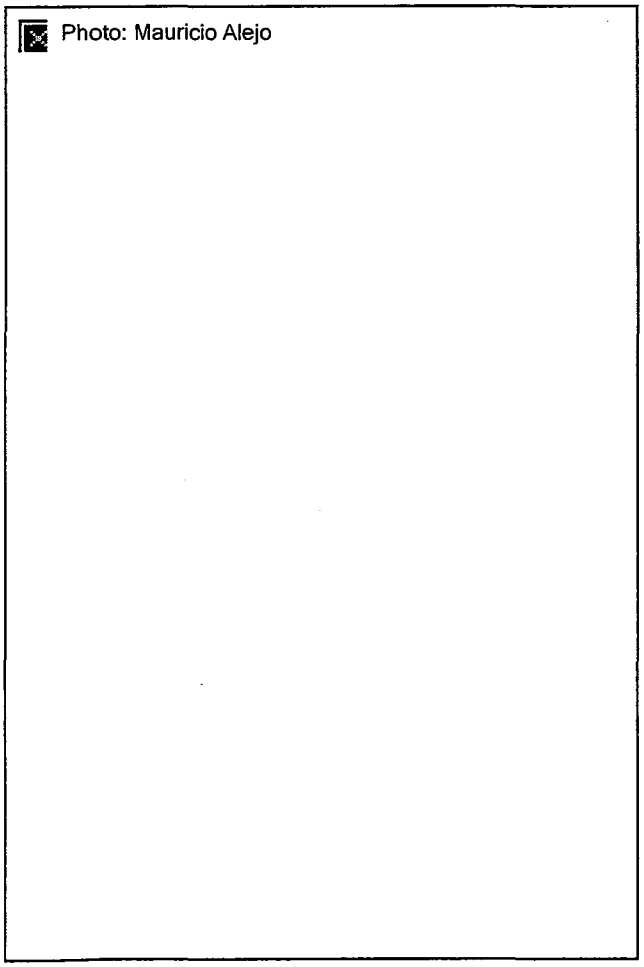
Photo: Mauricio Alejo

When it comes to finding stuff, there's Google — and there's everyone else.
Photo: Mauricio Alejo

Want to know how Google is about to change your life? Stop by the Ouagadougou conference room on a Thursday morning. It is here, at the Mountain View, California, headquarters of the world's most powerful Internet company, that a room filled with three dozen engineers, product managers, and executives figure out how to make their search engine even smarter. This year, Google will introduce 550 or so improvements to its fabled algorithm, and each will be determined at a gathering just like this one. The decisions made at the weekly Search Quality Launch Meeting will wind up affecting the results you get when you use Google's search engine to look for anything — "Samsung SF-755p printer," "Ed Hardy MySpace layouts," or maybe even "capital Burkina Faso," which just happens to share its name with this conference room. Udi Manber, Google's head of search since 2006, leads the proceedings. One by one, potential modifications are introduced, along with the results of months of testing in various countries and multiple languages. A screen displays side-by-side results of sample queries before and after the change. Following one example — a search for "guitar center wah-wah" — Manber cries out, "I did that search!"

You might think that after a solid decade of search-market dominance, Google could relax. After all, it holds a commanding 65 percent market share and is still the only company whose name is synonymous with the verb *search*. But just as Google isn't ready to rest on its laurels, its competitors aren't ready to concede defeat. For years, the Silicon Valley monolith has used its mysterious, seemingly omniscient algorithm to, as its mission statement puts it, "organize the world's information." But over the past five years, a slew of companies have challenged Google's central premise: that a single search engine, through technological wizardry and constant refinement, can satisfy any possible query. Facebook launched an early attack with its implication that some people would rather get information from their friends than from an anonymous formula. Twitter's ability to parse its constant stream of updates introduced the concept of real-time search, a way of tapping into the latest chatter and conversation as it unfolds. Yelp helps people find restaurants, dry cleaners, and babysitters by crowdsourcing the ratings. None of these upstarts individually presents much of a threat, but together they hint at a wide-open, messier future of search — one that isn't dominated by a single engine but rather incorporates a grab bag of services.

Still, the biggest threat to Google can be found 850 miles to the north: Bing. Microsoft's revamped and rebranded search engine — with a name that evokes discovery, a famous crooner, or Tony Soprano's strip joint — launched last June to surprisingly upbeat reviews. (*The Wall Street Journal* called it "more inviting than Google.") The new look, along with a $100 million ad campaign, helped boost Microsoft's share of the US search market from 8 percent to about 11 — a number that will more than double once regulators approve a deal to make Bing the search provider for Yahoo.

Team Bing has been focusing on unique instances where Google's algorithms don't always satisfy. For example, while Google does a great job of searching the public Web, it doesn't have real-time access to the byzantine and constantly changing array of flight schedules and fares. So Microsoft purchased Farecast — a Web site that tracks airline fares over time and uses the data to predict when ticket prices will rise or fall — and incorporated its findings into Bing's results. Microsoft made similar acquisitions in the health, reference, and shopping sectors, areas where it felt Google's algorithm fell short.

Even the Bingers confess that, when it comes to the simple task of taking a search term and returning relevant results, Google is still miles ahead. But they also think that if they can come up with a few areas where Bing excels, people will get used to tapping a different search engine for some kinds of queries. "The algorithm is extremely important in search, but it's not the only thing," says Brian MacDonald,

Microsoft's VP of core search. "You buy a car for reasons beyond just the engine."

Google's response can be summed up in four words: *mike siwek lawyer mi*.

Amit Singhal types that koan into his company's search box. Singhal, a gentle man in his forties, is a Google Fellow, an honorific bestowed upon him four years ago to reward his rewrite of the search engine in 2001. He jabs the Enter key. In a time span best measured in a hummingbird's wing-flaps, a page of links appears. The top result connects to a listing for an attorney named Michael Siwek in Grand Rapids, Michigan. It's a fairly innocuous search — the kind that Google's servers handle billions of times a day — but it is deceptively complicated. Type those same words into Bing, for instance, and the first result is a page about the NFL draft that includes safety Lawyer Milloy. Several pages into the results, there's no direct referral to Siwek.

The comparison demonstrates the power, even intelligence, of Google's algorithm, honed over countless iterations. It possesses the seemingly magical ability to interpret searchers' requests — no matter how awkward or misspelled. Google refers to that ability as search quality, and for years the company has closely guarded the process by which it delivers such accurate results. But now I am sitting with Singhal in the search giant's Building 43, where the core search team works, because Google has offered to give me an unprecedented look at just how it attains search quality. The subtext is clear: You may think the algorithm is little more than an engine, but wait until you get under the hood and see what this baby can really do.

Key Advances in Google Search

Google's search algorithm is a work in progress — constantly tweaked and refined to return higher-quality results. Here are some of the most significant additions and adaptations since the dawn of PageRank. — Steven Levy

Backrub
[September 1997]

This search engine, which had run on Stanford's servers for almost two years, is renamed Google. Its breakthrough innovation: ranking searches based on the number and quality of incoming links.

New algorithm
[August 2001]

The search algorithm is completely revamped to incorporate additional ranking criteria more easily.

Local connectivity analysis
[February 2003]

Google's first patent is granted for this feature, which gives more weight to links from authoritative sites.

Fritz
[Summer 2003]

This initiative allows Google to update its index constantly, instead of in big batches.

Personalized results
[June 2005]

Users can choose to let Google mine their own search behavior to provide individualized results.

Bigdaddy
[December 2005]

Engine update allows for more-comprehensive Web crawling.

Universal search
[May 2007]

Building on Image Search, Google News, and Book Search, the new Universal Search allows users to get links to any medium on the same results page.

Real-Time Search
[December 2009]

Displays results from Twitter and blogs as they are published.

The story of Google's algorithm begins with PageRank, the system invented in 1997 by cofounder Larry Page while he was a grad student at Stanford. Page's now legendary insight was to rate pages based on the number and importance of links that pointed to them — to use the collective intelligence of the Web itself to determine which sites were most relevant. It was a simple and powerful concept, and — as Google quickly became the most successful search engine on the Web — Page and cofounder Sergey Brin credited PageRank as their company's fundamental innovation.

But that wasn't the whole story. "People hold on to PageRank because it's recognizable," Manber says. "But there were many other things that improved the relevancy." These involve the exploitation of certain signals, contextual clues that help the search engine rank the millions of possible results to any query, ensuring that the most useful ones float to the top.

Web search is a multipart process. First, Google crawls the Web to collect the contents of every accessible site. This data is broken down into an *index* (organized by word, just like the index of a textbook), a way of finding any page based on its content. Every time a user types a query, the index is

combed for relevant pages, returning a list that commonly numbers in the hundreds of thousands, or millions. The trickiest part, though, is the *ranking* process — determining which of those pages belong at the top of the list.

That's where the contextual signals come in. All search engines incorporate them, but none has added as many or made use of them as skillfully as Google has. PageRank itself is a signal, an attribute of a Web page (in this case, its importance relative to the rest of the Web) that can be used to help determine relevance. Some of the signals now seem obvious. Early on, Google's algorithm gave special consideration to the title on a Web page — clearly an important signal for determining relevance. Another key technique exploited anchor text, the words that make up the actual hyperlink connecting one page to another. As a result, "when you did a search, the right page would come up, even if the page didn't include the actual words you were searching for," says Scott Hassan, an early Google architect who worked with Page and Brin at Stanford. "That was pretty cool." Later signals included attributes like freshness (for certain queries, pages created more recently may be more valuable than older ones) and location (Google knows the rough geographic coordinates of searchers and favors local results). The search engine currently uses more than 200 signals to help rank its results.

Google's engineers have discovered that some of the most important signals can come from Google itself. PageRank has been celebrated as instituting a measure of populism into search engines: the democracy of millions of people deciding what to link to on the Web. But Singhal notes that the engineers in Building 43 are exploiting another democracy — the hundreds of millions who search on Google. The data people generate when they search — what results they click on, what words they replace in the query when they're unsatisfied, how their queries match with their physical locations — turns out to be an invaluable resource in discovering new signals and improving the relevance of results. The most direct example of this process is what Google calls personalized search — a feature that uses someone's search history and location as signals to determine what kind of results they'll find useful.[1] But more generally, Google has used its huge mass of collected data to bolster its algorithm with an amazingly deep knowledge base that helps interpret the complex intent of cryptic queries.

Take, for instance, the way Google's engine learns which words are synonyms. "We discovered a nifty thing very early on," Singhal says. "People change words in their queries. So someone would say, 'pictures of dogs,' and then they'd say, 'pictures of puppies.' So that told us that maybe 'dogs' and 'puppies' were interchangeable. We also learned that when you boil water, it's hot water. We were relearning semantics from humans, and that was a great advance."

But there were obstacles. Google's synonym system understood that a dog was similar to a puppy and that boiling water was hot. But it also concluded that a hot dog was the same as a boiling puppy. The problem was fixed in late 2002 by a breakthrough based on philosopher Ludwig Wittgenstein's theories about how words are defined by context. As Google crawled and archived billions of documents and Web pages, it analyzed what words were close to each other. "Hot dog" would be found in searches that also contained "bread" and "mustard" and "baseball games" — not poached pooches. That helped the algorithm understand what "hot dog" — and millions of other terms — meant. "Today, if you type 'Gandhi bio,' we know that bio means biography," Singhal says. "And if you type 'bio warfare,' it means biological."

Throughout its history, Google has devised ways of adding more signals, all without disrupting its users' core experience. Every couple of years there's a major change in the system — sort of equivalent to a new version of Windows — that's a big deal in Mountain View but not discussed publicly. "Our job is to basically change the engines on a plane that is flying at 1,000 kilometers an hour, 30,000 feet above Earth," Singhal says. In 2001, to accommodate the rapid growth of the Web, Singhal essentially revised Page and Brin's original algorithm completely, enabling the system to incorporate new signals quickly

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.