Welcome to TimesPeople
What's this?
TimesPeople Lets You Share and Discover the Best of NYTimes.com
8:34 PM
Get Started
No, thanks

HOME PAGE | TODAY'S PAPER | VIDEO | MOST POPULAR | TIMES TOPICS
Get Home Delivery | Log In | Register Now

The New York Times

Internet

Search All NYTimes.com
Go

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION | ARTS | STYLE | TRAVEL | JOBS | REAL ESTATE | AUTOS

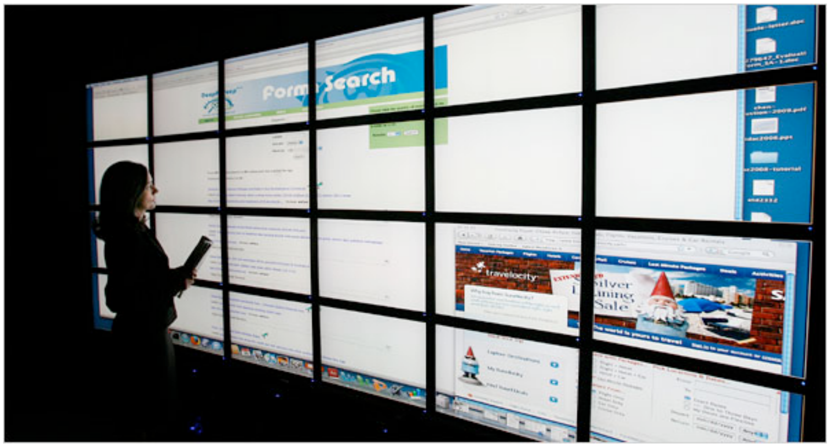
Search Technology
Go

Inside Technology
Internet | Start-Ups | Business Computing | Companies

Bits Blog »

Personal Tech »
Cellphones, Cameras, Computers and more

Exploring a 'Deep Web' That Google Can't Grasp



Jeffrey D. Allred for The New York Times

At the University of Utah, Prof. Juliana Freire is working on DeepPeep, an ambitious effort to index every public database online.

By ALEX WRIGHT
Published: February 22, 2009

SIGN IN TO E-MAIL
PRINT
REPRINTS
SHARE

One day last summer, Google's search engine trundled quietly past a milestone. It added the one trillionth address to the list of Web pages it knows about. But as impossibly big as that number may seem, it represents only a fraction of the entire Web.

Beyond those trillion pages lies an even vaster Web of hidden data: financial information, shopping catalogs, flight schedules, medical research and all kinds of other material stored in databases that remain largely invisible to search engines.

The challenges that the major search engines face in penetrating this so-called Deep Web go a long way toward explaining why they still can't provide satisfying answers to questions like "What's the best fare from New York to London next Thursday?" or "When will the Yankees play the Red Sox this year?" The answers are readily available — if only the search engines knew how to find them.

Now a new breed of technologies is taking shape that will extend the reach of search engines into the Web's hidden corners. When that happens, it will do more than just improve the quality of search results — it may ultimately reshape the way many companies do business online.

Next Article in Technology (1 of 13) »

News for Education Professionals
What's This?
FROM NYTIMES.COM
- The Big Test Before College? The Financial Aid Form
- Class Size in New York City Schools Rises, but the Impact is Debated
- Alfred J. Kahn, Specialist in Child Welfare Issues, Dies at 90
- Endowment Director Is on Harvard's Hot Seat
- 18 Students Are Suspended as Protest at N.Y.U. Ends

Powered by LinkedIn

Breaking News Alerts by E-Mail
Sign up to be notified when important news breaks.
Sign Up
Privacy Policy

Subscribe to Technology RSS Feeds
Technology News
Internet
Start-Ups
Business Computing
Companies
Bits Blog
Personal Tech
Pogue's Posts

MOST POPULAR - TECHNOLOGY
E-MAILED | BLOGGED

Search engines rely on programs known as crawlers (or spiders) that gather information by following the trails of hyperlinks that tie the Web together. While that approach works well for the pages that make up the surface Web, these programs have a harder time penetrating databases that are set up to respond to typed queries.

“The crawlable Web is the tip of the iceberg,” says Anand Rajaraman, co-founder of Kosmix (www.kosmix.com), a Deep Web search start-up whose investors include Jeffrey P. Bezos, chief executive of Amazon.com. Kosmix has developed software that matches searches with the databases most likely to yield relevant information, then returns an overview of the topic drawn from multiple sources.

“Most search engines try to help you find a needle in a haystack,” Mr. Rajaraman said, “but what we’re trying to do is help you explore the haystack.”

That haystack is infinitely large. With millions of databases connected to the Web, and endless possible permutations of search terms, there is simply no way for any search engine — no matter how powerful — to sift through every possible combination of data on the fly.

To extract meaningful data from the Deep Web, search engines have to analyze users’ search terms and figure out how to broker those queries to particular databases. For example, if a user types in “Rembrandt,” the search engine needs to know which databases are most likely to contain information about fine art (like, say, museum catalogs or auction houses), and what kinds of queries those databases will accept.

That approach may sound straightforward in theory, but in practice the vast variety of database structures and possible search terms poses a thorny computational challenge.

“This is the most interesting data integration problem imaginable,” says Alon Halevy, a former computer science professor at the University of Washington who is now leading a team at Google that is trying to solve the Deep Web conundrum.

Google’s Deep Web search strategy involves sending out a program to analyze the contents of every database it encounters. For example, if the search engine finds a page with a form related to fine art, it starts guessing likely search terms — “Rembrandt,” “Picasso,” “Vermeer” and so on — until one of those terms returns a match. The search engine then analyzes the results and develops a predictive model of what the database contains.

In a similar vein, Prof. Juliana Freire at the University of Utah is working on an ambitious project called DeepPeep (www.deeppeep.org) that eventually aims to crawl and index every database on the public Web. Extracting the contents of so many far-flung data sets requires a sophisticated kind of computational guessing game.

“The naïve way would be to query all the words in the dictionary,” Ms. Freire said. Instead, DeepPeep starts by posing a small number of sample queries, “so we can then use that to build up our understanding of the databases and choose which words to search.”

Based on that analysis, the program then fires off automated search terms in an effort to dislodge as much data as possible. Ms. Freire claims that her approach retrieves better than 90 percent of the content stored in any given database. Ms. Freire’s work has recently attracted overtures from one of the major search engine companies.

As the major search engines start to experiment with incorporating Deep Web content into their search results, they must figure out how to present different kinds of data without overcomplicating their pages. This poses a particular quandary for Google, which has long resisted the temptation to make significant changes to its tried-and-true search results format.

1. Basics: Low-Tech Fixes for High-Tech Problems
2. More Customers Give Up the Cellphone Contract
3. How to Friend Mom, Dad, and the Boss on Facebook...Safely
4. Don’t Worry, Be Athletic: These Headphones Fit Comfortably
5. Q & A: Tip of the Week: A Way to Clean Up Your Bookmarks
6. Pocket Projector Stands Out, Even in a Small Crowd
7. Bits: Fabled City of Atlantis Spotted on Google Earth?
8. Bits: I.B.M. Delivers Rural Broadband Over Power Lines
9. Bits: Baby Boomers, Luddites? Not So Fast
10. Facebook Withdraws Changes in Data Use

Go to Complete List »

“Google faces a real challenge,” said Chris Sherman, executive editor of the Web site Search Engine Land. “They want to make the experience better, but they have to be supercautious with making changes for fear of alienating their users.”

Beyond the realm of consumer searches, Deep Web technologies may eventually let businesses use data in new ways. For example, a health site could cross-reference data from pharmaceutical companies with the latest findings from medical researchers, or a local news site could extend its coverage by letting users tap into public records stored in government databases.

This level of data integration could eventually point the way toward something like the Semantic Web, the much-promoted — but so far unrealized — vision of a Web of interconnected data. Deep Web technologies hold the promise of achieving similar benefits at a much lower cost, by automating the process of analyzing database structures and cross-referencing the results.

“The huge thing is the ability to connect disparate data sources,” said Mike Bergman, a computer scientist and consultant who is credited with coining the term Deep Web. Mr. Bergman said the long-term impact of Deep Web search had more to do with transforming business than with satisfying the whims of Web surfers.

A version of this article appeared in print on February 23, 2009, on page B1 of the New York edition.

Next Article in Technology (1 of 13) »

Related Articles

FROM THE NEW YORK TIMES

Lawsuit Says Google Was Unfair to Rival Site
(February 18, 2009)

Search Service By Google Briefly Fails
(February 1, 2009)

Google Beats Forecast Even as Its Profit Tapers
(January 23, 2009)

PING; At First, Funny Videos. Now, a Reference Tool.
(January 18, 2009)

INSIDE NYTIMES.COM

OPINION »

Op-Ed: The Gay Marriage Reconciliation

THEATER »

A Radical Vixen Retakes the Stage

REAL ESTATE »

The Financial District Has Babies

OPINION »

Room for Debate

Experts discuss what China expects from the Obama administration.

THE CITY »

Queens Ice Cream Hut at Center of Ethnic Divide

FASHION & STYLE »

Weddings and Celebrations