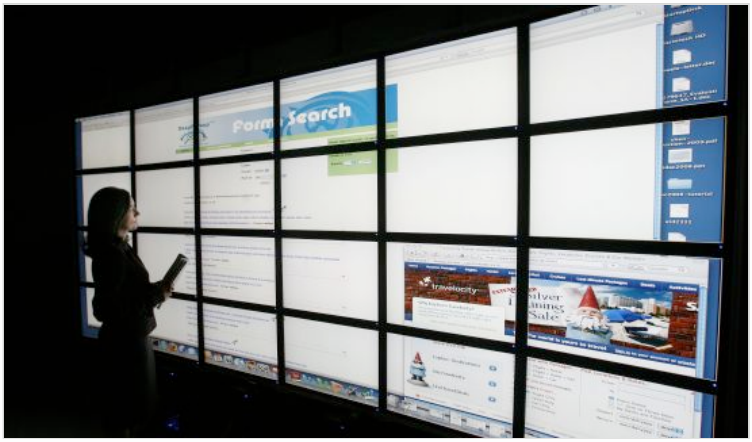
INTERNATIONAL
Herald Tribune
THE GLOBAL EDITION OF THE NEW YORK TIMES

Technology & Media



Juliana Freire's DeepPeep project is meant to index every database on the public Web.
(Jeffrey D. Allred for The New York Times)

Emerging search technologies aim for Web's hidden depths

By Alex Wright Published: February 23, 2009

NEW YORK: One day last summer, Google's search engine trundled quietly past a milestone. It added the 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 cannot 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.

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.

Today in Technology & Media

Ads not merely commercial in France

Playboy says it is open to sale of business

New York Times Co. suspends dividends to shareholders

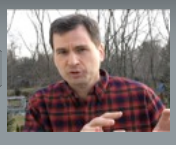
"The crawlable Web is the tip of the iceberg," says Anand Rajaraman, co-founder of Kosmix (www.kosmix.com), a Deep Web search startup whose investors include Jeffrey Bezos, chief executive of Amazon.com. Kosmix has developed software that matches searches with the databases most likely to yield relevant 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," 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.

Video See all videos »



Windows 7 Beta
David Pogue looks at Windows 7 Beta.

Most E-Mailed

24 Hours | 7 Days | 30 Days

1. The diamond 'overhang'
2. Mystery endures in Brazilian town of twins
3. Paul Krugman: Banking on the brink
4. India celebrates Oscar victories by 'Slumdog'
5. China keeps wary eye on displaced migrant workers
6. Bagging bargains in Cape Town, South Africa
7. At 44, a running career again in ascent
8. U.S. to give $900 million in aid to Gaza
9. Statement from Binyam Mohamed
10. A soft spot for print is costing Murdoch