INTERNATIONAL
Herald Tribune
THE GLOBAL EDITION OF THE NEW YORK TIMES

Technology & Media

Exploring a 'Deep Web' that Google can't grasp

By Alex Wright

Published: February 23, 2009

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.

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

In France, Rue89 brings readers into the newsroom

Reaching tech folks on their turf

Barnes & Noble buys an e-book retailer

"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 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.

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," 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."