

SEARCH SMARTS

NEW TECHNOLOGY COULD LEAVE WEB 'CRAWLERS' IN THE DUST





One day in the not-too-distant future, you'll be able to type a query into an online search engine and have it deliver not Web pages that may contain an answer, but just the answer itself.

User: "Who starred in the film *Casablanca*?"

Search Engine: "Humphrey Bogart and Ingrid Bergman."

Not impressed?

Imagine asking a more nuanced question, such as "What do Americans think of offshore drilling?" A search engine will be able to respond with a report indicating trends in opinion based on what has been posted to the Web.

Search engines may eventually be used to conduct polling and even help sort fact from fiction, said Weiyi Meng, a professor of computer science at Binghamton University. He's helping to make such futuristic possibilities a reality, both through his research and as president of a company called Webscalers.

The way Meng sees it, big search engines such as Google and Yahoo are fundamentally flawed.

You see, the Web has two parts: The surface Web and the deep Web. The surface Web is made up of perhaps 60 billion pages. The deep Web, at some 900 billion pages, is about 15 times larger.

NOT ONLY CAN A METASEARCH ENGINE PROBE DEEPER, IT CAN ALSO OFFER THE LATEST INFORMATION.



Google, which relies on a “crawler” to examine pages and catalog them for future searches, can search about 20 billion pages, just a small fraction of the entire Web. Web crawlers follow links to reach pages and often miss content that isn’t linked to any other page or is in some other way “hidden.”

Meng, along with colleagues at the University of Illinois at Chicago and the University of Louisiana at Lafayette, has helped pioneer large-scale metasearch-engine technology that harnesses the power of numerous small search engines to come up with results that are more accurate and more complete.

“Most of the pages on the deep Web aren’t directly ‘crawlable.’ We want to connect to small search engines and reach the deep Web,” he said. “That’s the idea. Many people have the misconception that Google can search everything, and if it’s not there it doesn’t exist. But we should be able to retrieve many times more than what Google can search.”

Not only can a metasearch engine probe deeper, it can also offer the latest information.

“In principle,” Meng noted, “small guys are much better able to maintain the freshness of their data. Google has a program to ‘crawl’ all over the world. Depending on when the crawler has last visited your server, there’s a delay of days or weeks before a new page will show up in that search. We can get fresher results.”

The concept is not new. In fact, the first metasearch engine was built in 1994.

“The big difference between our technology and the ones pursued by other people is that most of the other technologies do the metasearching on top of a small number of general-purpose search engines, such as Yahoo, Google or MSN,” Meng explained. “We have a completely different perspective. We want to build large-scale metasearch engines on top of many small search engines.”

The Web has about 1 million search engines. Most universities have search engines, most newspapers have search engines and many companies and organizations have search engines. Since 1997, and with the support of five grants from the National Science Foundation, Meng and his collaborators have found innovative ways to run queries across multiple search engines and sort through the results.

Webscalers, founded in 2002, is now based in the Start-Up Suite at Binghamton University’s Innovative Technologies Complex, which is home to several young companies that have their roots in faculty inventions.

“If the Web keeps on growing, a company like Google may run out of resources to crawl all of those pages,” said Vijay V. Raghavan, a vice president of Webscalers and a distinguished professor of computer science at the University of Louisiana at Lafayette. “We won’t have that problem. We will scale much better.”

The firm has already launched several metasearch products.

The first is a news metasearch engine called AllinOneNews. Available at www.allinonenews.com, it connects to 1,800 news sources in 200 countries. That’s the largest metasearch engine in the world.

Webscalers also offers MySearchView, a customized metasearch-engine generation system that allows any user to create his or her own metasearch engine just by checking off a few options at v2.mysearchview.com.

That site is a showcase for the company’s attempts to develop automated solutions to link multiple search engines.

This kind of technology could be useful for large organizations with many branches or divisions. If each one has its own search engine, but the organization as a whole does not, a metasearch engine can connect all of the parts to the whole.



Wei Yi Meng

For example, Webscalers has developed a prototype that would allow a search of all 64 campuses in the State University of New York system as well as SUNY's central administration.

"People can use it to find collaborators," Meng said. "It could also help prospective students find programs they're interested in."

The technology could be adapted to large companies or even the government, Meng said.

Webscalers has incorporated another unusual feature in its AllinOneNews metasearch engine called a "semantic match." A search engine that's capable of making such a match will find results for words with the same meaning, even if they're not part of the original query. It will include pages with the word "ballerina" if you search for "ballet dancer," for example, and "hypertension" if you search for "high blood pressure."

Challenges for large-scale metasearch engines include determining which search engines are the best for a given query, automating the interaction with search engines as well as organizing the search results. Meng and his colleagues have done extensive and pioneering research on these topics, publishing about 50 papers so far.

Meng hopes to one day build a grand metasearch engine that would integrate all of the 1 million small search engines into a single system. "There are still a lot of significant challenges in creating a system of such magnitude," he said, "but I am optimistic that such a metasearch engine can be built." ■

— Rachel Coker