UC Berkeley >

University of California

Campus News > Media Relations

Berkeley

NEWS SEARCH

Go

NEWS HOME

ARCHIVES

EXTRAS

MEDIA RELATIONS

- Press Releases
- Image Downloads
- Contacts

PRESS RELEASE

Election 2000 Web site at UC Berkeley demonstrates power of new Internet technology for mining the "deep Web"

25 Oct 2000

By Robert Sanders, Media Relations

Berkeley - Want to find out which Hollywood stars donated to Vice President Al Gore's Presidential campaign? How about the home prices of the donors to Texas Governor George Bush's campaign? Or the crime rates in the neighborhoods of donors to either candidate?

As this year's Presidential campaign climaxes, a University of California, Berkeley, professor has created a Web site that makes such searches easy, and demonstrates the power of new Internet technology he has developed to mine the "deep Web."

"This is more powerful than search engines on the Web," said Joseph Hellerstein, associate professor of computer sciences in the College of Engineering at UC Berkeley, who created the site with fellow computer sciences associate professor Michael Franklin and the help of five graduate students and one undergraduate. "With this you can do real data analysis, not just find a neat new Web page."

The software that Hellerstein and Franklin developed is called Telegraph, after the street near the UC Berkeley campus famous for its street vendors and street people.

"Like the Berkeley main street after which it is named, Telegraph is the natural thoroughfare for a volatile, eclectic mix coming from all over the world," Hellerstein wrote on his Web site.

The "deep Web" refers to information on the Internet that is not available by simply following hyperlinks, and thus not accessible through search engines like Google or Inktomi. Some people estimate the deep Web contains 500 times as much information as the rest of the Web, most of it in free databases that require a person to fill out a form in order to submit a query.

The database searches and cross-referencing that Hellerstein, Franklin and their students make available can be done by anyone willing to delve into publicly available databases compiled and run by the Federal Election Commission, the APBNews.com Crime Statistics site, the Yahoo Real Estate database, the Yahoo Actor and Actress List, the U.S. Census, and others. But such painstaking digging is laborious and time consuming.

Hellerstein's Web site makes it easy by automating the form searches, so that information in one database can be brought up for comparison and correlation with information in other databases. The computer does the tedious data searching - "screen scraping" in computer jargon - while the new UC Berkeley technology choreographs the search in the most efficient way.

"This is about the facts and figures on the Web. Web crawlers can't get to this information," he said. "We call this data the 'Facts and Figures Federation.'"

For example, journalists today can trace money spent by political action committees (PACs) by laboriously pouring over lists of donors and tracking the money back through numerous other PACs to its corporate or private source. The UC Berkeley team has set up a way to connect the money easily by automatically "crawling" the donations back to the source.

"You can track the six degrees of separation of PACs - which PACs give to other PACs," Hellerstein said. Philip Morris, for example, doesn't give directly to Bush, but through its PAC gives to other PACs, like the Fund For a Responsible Future, that in turn give to Bush. Similarly, the AFL-CIO doesn't give directly to Gore, but does give to other PACs, like the Evergreen Fund, that give to Gore.

Alternatively, a comparison of crime rates in the neighborhoods of donors to the two candidates show that Bush's donors live predominantly in low crime areas, while Gore's donors are spread out among low and medium crime areas.

And what do Gwyneth Paltrow, Jack Nicholson, Candice Bergen and Jerry Seinfeld have in common? They've each given to the Gore campaign, along with a slew of other actors and actresses. It's hard to identify any Hollywood donors to the Bush campaign.

"With software like this, the publicly available databases are more powerful than people thought," Hellerstein said.

And, he added, potentially more scary. With all the databases on the Web, it also is possible to correlate names with addresses for the entire U.S. population, and cross-reference those with individual home prices, neighborhood crime statistics and even AIDS infection rates. Marketing firms with their own private data could mine veritable gold by combining their data with these deep Web databases.

"This software enables new things, and those things have consequences," he said. "Obviously it's good to have some of these databases publicly available, but they can lead to serious invasions of privacy. This technology may cause people to rethink the balance between freedom of information and privacy."

Telegraph merges two technologies that make the Internet work: the technology that makes data flow smoothly through the Web, and database query technology. He calls Telegraph an adaptive data flow system because it adapts to the fact that data doesn't flow at the same rate from all sources. One database may be slower than another, or the Internet may slow down and then speed up. Correlating data in real time requires a system that can adapt to such unpredictable behavior.

Using the metaphor of flowing water, he described the technique as creating an eddy of data analysis within a river of information streaming across the Internet. He employs a kind of lottery to determine which database is queried at each step, and is able to optimize and speed up the process of data collection. Telegraph also is designed to harness streams of live data coming from networks of sensors on the Internet, or even from smart devices.

The adaptive dataflow technique is well suited to searching databases that are updated frequently, such as lists of campaign donors, and also databases that are impractical to download in their entirely and cache for later analysis. Telegraph downloads the information "live" from the source. The software also makes it easy to handle new databases that appear on the Internet.

"This will be useful for anyone who wants to look at trends or the big picture using lots of data. That means marketers, pollsters, businesses and researchers of all kinds," Hellerstein said.

For the moment, Hellerstein and his students have scripted specific database searches and made them available on their Web site, such as a correlation of crime rate with a campaign donor's Zip code. More general comparisons could be allowed, but a user would have to learn how to write a proper query. Many users would not go to that trouble, Hellerstein said, so making that process easier is an area of future research. But he and his students can script database queries quickly, and they plan to provide further examples on his Web site.

"Once the election is over, we'll draw on new sources of information on the deep Web and elsewhere," he said. "It will be interesting."

###

The Federated Facts and Figures Website, complete with access to Election 2000 information, is at http://fff.cs.berkeley.edu/.

For more information on the adaptive dataflow system, Telegraph, check out http://telegraph.cs.berkeley.edu/.

UC Berkeley | News | Archives | Extras | Media Relations

Comments? E-mail newscenter@pa.urel.berkeley.edu.