Pinning down the fleeting Internet: Web crawler archives historical data for easy searching

November 17 2008

Zoetrope scans BBC headline articles for past stories about the Ukraine. Image: University of Washington

(PHysOrg.com) -- The Internet contains vast amounts of information, much of it unorganized. But what you see online at any given moment is just a snapshot of the Web as a whole -- many pages change rapidly or disappear completely, and the old data gets lost forever.

"Your browser is really just a window into the Web as it exists today," said Eytan Adar, University of Washington computer science and engineering doctoral student. "When you search for something online, you're only getting today's results."

Now, Adar and his colleagues at UW and Adobe Systems Inc. are grabbing hold of the fleeting Web and storing historical sites that users

can easily search using an intuitive application called Zoetrope.

"There are so many ways of finding and manipulating and visualizing data on what we call 'the today Web' that it's kind of amazing that there's no way to do anything similar to the ephemeral Web," said Dan Weld, a UW computer science and engineering professor who also worked on the application. One service, the Internet Archive, has been capturing old versions of Web sites for years, but the records for the stored sites are inconsistent, Weld said. More importantly, there's no easy way to search the archive.

With Zoetrope, anyone will be able to use easy keyword searches to find archived Web information or look for patterns over time. The research was presented Oct. 22 by Mira Dontcheva, the system's co-creator and a recently graduated UW computer science and engineering doctoral student now at Adobe Systems Inc., at the ACM Symposium on User Interface Software and Technology in Monterey, Calif.

There are a variety of ways people might want to search the historical Internet. For example, to find a history of traffic patterns in the Seattle area, you'd have to sort through lengthy PDF files from the state Department of Transportation, Adar said. With Zoetrope, you could easily view past versions of any traffic Web site, and getting more specific, search for drive-times on Interstate 90 at 6 p.m. on rainy Fridays. Zoetrope can also capture and help analyze information that might otherwise not be available anywhere.

Sports fanatics could use the program to check historical rankings of their favorite teams or players, information that currently may not be easy to find. The application can do more than just simple keyword searches, Adar said. It also can be used to analyze historical data or link information from different sites. For example, Adar wondered whether air pollution conditions could affect the performance of Olympic

athletes, so he used Zoetrope to find daily records of pollution levels in Beijing and the number of world records broken in the 2008 Olympics on each day, and looked to see whether fewer records were broken on days with high pollution levels.

"Zoetrope is aimed at the casual researcher," Weld said. "It's really for anyone who has a question."

Zoetrope could eventually be built in to any other Web browser, Adar said. If you just want to browse the past versions of a given site, you drag a slider backwards to see older and older versions. Alternatively, you can draw a box around just one part of the site, if you're interested in, say, the lead story on CNN.com but don't care about the rest of the page. These boxes can be filtered by keyword searches or date, so you could look only for lead stories featuring Hollywood actors or stories that ran on Fridays.

Users can view historical data by moving the slider, but more sophisticated analyses are available as well. If you're looking at something numerical, such as gas prices over time, the program can draw graphs for you. Or you can pull out images from specific times, such as traffic pictures, and compare them all side by side. These kinds of visualizations can be further organized in a timeline or by clustering -- Zoetrope can make an image comparing traffic patterns on sunny days versus cloudy days, for example.

Right now, Zoetrope saves a new version of approximately 1,000 different sites every hour, Adar said. It's been running for four months, so records go no further than that, but Adar hopes to eventually incorporate information from the Internet Archive's nearly 14 years of records into the program.

He wants to figure out how to scale the program up from 1,000 Web

pages to all pages in existence, and has run studies to figure how often each page would need to be captured. For example, a traffic site or stock-watching page would need versions saved much more often than every hour, but there are many unchanging pages that could be archived less frequently. Eventually, Zoetrope could automatically figure out how often to capture a page based on how frequently it changes, Adar said.

"This is really a new way to think about storing information on the Web," he said.

The researchers hope to release Zoetrope free, and say it may be available as early as next summer.

Provided by University of Washington