Warning: Blogs Can Be Infectious

By Amit Asaravala

Story location: http://www.wired.com/news/culture/0,1284,62537,00.html

02:00 AM Mar. 05, 2004 PT

The most-read webloggers aren't necessarily the ones with the most original ideas, say researchers at Hewlett-Packard Labs.

Using newly developed techniques for graphing the flow of information between blogs, the researchers have discovered that authors of popular blog sites regularly borrow topics from lesser-known bloggers -- and they often do so without attribution.

These findings are important to sociologists who are interested in learning how ideas grow from isolated topics into full-blown epidemics that "infect" large populations. Such an understanding is also important to marketers, who hope to be able to pitch products and ideas directly to the most influential people in a given group.

"There is a lot of speculation that really important people are highly connected, but really, we wonder if the highly connected people just listen to the important people," said Lada Adamic, one of the four researchers working on the project.

To satisfy their curiosity, the researchers began analyzing data from Intelliseek's BlogPulse Web crawler, which regularly mines thousands of blogs for references to people, places and events.

When they plotted the links and topics shared by various sites, they discovered that topics would often appear on a few relatively unknown blogs days before they appeared on more popular sites.

"What we're finding is that the important people on the Web are not necessarily the people with the most explicit links (back to their sites), but the people who cause epidemics in blog networks," said researcher Eytan Adar.

These infectious people can be hard to find because they do not always receive attribution for being the first to point to an interesting idea or news item.

Indeed, the team at HP Labs found that when an idea infected at least 10 blogs, 70 percent of the blogs did not provide links back to another blog that had previously mentioned the idea.

To get past this obstacle, the researchers developed techniques to infer where information might have come from, based on the similarities in text, links and infection rates.

For instance, if Blog A used the words "furry germs" to link to an infectious topic like Giantmicrobes just days after Blog B in the same social circle used the exact same words and link, that would be a good sign that Blog A copied Blog B.

The researchers have incorporated their techniques into a search algorithm they call iRank. Unlike Google's PageRank algorithm, which ranks websites based on overall popularity, the iRank algorithm ranks sites based on how good they are at injecting ideas into the mainstream.

"A lot of sites that get listed by search engines as most relevant are not always the most relevant," said Adar. "For instance, Slashdot often gets listed at the top, but it's just an aggregator. I may want to go to the source."

Adar and Adamic say it's too soon to tell if iRank will be incorporated into popular search engines.

For one thing, they plan to refine the algorithm after seeing how it works on more data. They would also like to modify the algorithm to resist manipulation from Google-bomb-type attacks, where collaborators link to each other's sites to boost themselves in Google's ranking mechanism.

In the meantime, the team has made some of its research available online in the form of the Blog Epidemic Analyzer, a Java program that reveals the implicit and inferred links between blogs in an interactive, visual form.

"Blogs are helping us get a better understanding of how things happen on the Internet," said Adar. "We're hopeful that in being able to do this research, we can apply the technology to other information, like e-mail, to improve productivity."

Wired News: Staff | Contact Us | Advertising | RSS | SUBSCRIBE
We are translated daily into Spanish, Portuguese, and Japanese

© Copyright 2004, Lycos, Inc. All Rights Reserved.
Your use of this website constitutes acceptance of the Lycos **Privacy Policy** and Terms & Conditions