Technology Review
PUBLISHED BY MIT
July 1998

Untangling Web Searches

A more discerning way to find information

By Herb Brody

Web searches often evoke a two-part reaction. First: Wow, that was fast! Followed sadly by: But none of this is what I want. Lightning-quick online searches typically lead Web users into piles of documents that are, to be kind, of dubious reliability. Unlike the carefully catalogued stacks in a library, the Web often appears to be untouched by human judgment.

This chaos has been the price Web users pay for an open system to which anyone can contribute. But it is an unnecessary price, says Jon M. Kleinberg, a professor of computer science at Cornell University. Kleinberg has devised an approach for sifting the contents of the Web that could go a long way toward solving what he calls the Web's "abundance problem."

Kleinberg's technique relies on the premise that despite the jumbled appearance of the Web, critical thinking is in fact woven throughout it. Every time a page's creator includes a link to another site, that is a vote of confidence in the linked-to page. Thus a rough measure of a site's value can be derived by counting how many other sites are linked to it. "The Web is explicitly annotated with precisely the type of human judgment that we need in order to formulate a notion of authority," says Kleinberg.

But this measure needs to be refined, because if it were used alone, the Yahoo search directory and the Netscape homepage would come out near the top every time. "We need a way to throw those pages out," Kleinberg explains. The solution? Kleinberg applies a second level of filtering that assigns higher value to pages that include lots of links to other sites that are themselves relevant to the search.

By viewing the Web through its linkages and not merely by key words, Kleinberg's search algorithm solves another common search problem. A conventional Web search on the word "jaguar," for example, generates an unsorted roster of sites-most related to the sports car or to an obsolete computer with the same name. Information on the jungle cat that inspired these brands, however, is harder to come by. Kleinberg's system automatically groups hit lists into "communities" of sites that reference one another, in this case providing a list subdivided by those related to cars, computers and cats.

Kleinberg developed the algorithm while at IBM's Almaden Research Center in San Jose, Calif., which still owns it. For now, the enhanced searching tool remains experimental, but IBM researchers are shopping it around to companies that run online search services, including Alta Vista operator Digital Equipment Corp. Widespread availability is "inevitable," says Prabhakar Raghavah, manager of computer science principles at IBM Almaden. "This is a great idea whose time will surely come."