Study shows Google favored over other search engines by webmasters

November 15 2007

Web site policy makers who use robots.txt files as gatekeepers to specify what is open and what is off limits to Web crawlers have a bias that favors Google over other search engines, say Penn State researchers whose study of more than 7,500 Web sites revealed Google's advantage.

That finding was surprising, said C. Lee Giles, the David Reese Professor of Information Sciences and Technology who led the research team which developed a new search engine—BotSeer—for the study.

"We expected that robots.txt files would treat all search engines equally or maybe disfavor certain obnoxious bots, so we were surprised to discover a strong correlation between the robots favored and the search engines' market share," said Giles of Penn State's College of Information Sciences and Technology (IST).

Robots.txt files are not an official standard, but by informal agreement, they regulate Web crawlers—also known as "spiders" and "bots"—which mine the Web 24/7 for everything from the latest news to e-mail addresses. Web policy makers use the files found in a Web site's directory to restrict crawler access to non-public information. Robots.txt files also are used to reduce server load which can result in denial of service and shut down Web sites. But some Web policy makers and administrators are writing robots.txt files which are not uniformly blocking access.

Instead, those robots.txt files give access to Google, Yahoo and MSN

while restricting other search engines, the researchers learned.

As an example, some U.S. government sites favor Google's bot—Googlebot—followed by Yahoo and MSN, according to the researchers.

While the study doesn't include explanations for why Web policy makers have opted to favor Google, the researchers know the choice was made consciously. Not using a robots.txt file gives all robots equal access to a Web site.

"Robots.txt files are written by Web policy makers and administrators who have to intentionally specify Google as the favored search engine," Giles said.

That finding is described in a paper, "Determining Bias to Search Engines from Robots.txt," given at the recent 2007 IEEE/WIC/ACM International Conference on Web Intelligence in Silicon Valley. Besides Giles, the authors include Yang Sun and Ziming Zhuang, IST graduate students, and Isaac Councill, an IST post-doctoral scholar.

Not every site has a robots.txt file although the number is growing. Of the 7,500 sites analyzed by the researchers, about four in 10 had a robots.txt file—up from less than 1 in 10 in 1996.

That growth, which the researchers anticipate will continue, was one reason for the study.

The researchers didn't know what they would find when they set BotSeer on the loose to look at and index the content of the robots.txt files of the Web sites which spanned several market segments including government, newspaper, university and Fortune 1000 companies.

“Our intent was exploratory—to see if there was anything interesting,” Councill said. Consumers with a soft spot for Google aren’t affected by the bias. But consumers who prefer other search engines may be at a disadvantage.

“With the preference, Google can index some information which other search engines can’t,” Giles said.

Source: Penn State