Erschienen in: Proceedings of the 21st ACM Conference on Hypertext and Hypermedia : June 13 - 16, 2010, Toronto, ON, Canada / Chignell, Mark (Hrsg.). - New York : ACM, 2010. - S. 297-298. - ISBN 978-1-4503-0041-4

On the Robustness of Google Scholar against Spam

Jöran Beel
UC Berkeley / OvGU
beel@berkeley.edu

Bela Gipp
UC Berkeley / OvGU
gipp@berkeley.edu

ABSTRACT
In this research-in-progress paper we present the current results of several experiments in which we analyzed whether spamming Google Scholar is possible. Our results show, it is possible: We 'improved' the ranking of articles by manipulating their citation counts and we made articles appear in searchers for keywords the articles did not originally contained by placing invisible text in modified versions of the article.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *search process, information filtering*

General Terms
Algorithms, Measurement, Reliability, Experimentation, Human Factors, Legal Aspects

Keywords
search engines, academic search engines, citation spam, spamdexing, spam

1. INTRODUCTION
Researchers should have an interest in having their articles indexed by Google Scholar and other academic search engines such as CiteSeer(X). The inclusion of their articles in the index improves the ability to make their articles available to the academic community. In addition, authors should not only be concerned about the fact *that* their articles are indexed, but also *where* they are displayed in the result list. As with all ranked search results, articles displayed in top positions are more likely to be read.

In recent studies we researched the ranking algorithm of Google Scholar [1-3] and gave advice to researchers on how to optimize their scholarly literature for Google Scholar [4]. However, there are provisos in the academic community against what we called "Academic Search Engine Optimization" [4]. There is the concern that some researchers might use the knowledge about ranking algorithms to 'over optimize' their papers in order to push their articles' rankings in non-legitimate ways.

We conducted some experiments to find out how robust Google Scholar is against spamming. The experiments are not all completed yet but those that are completed show interesting results which are presented in this paper.

2. RELATED WORK
Most web search engines rank web pages based on two factors, namely the web page content and the amount and quality of links that point to the web page. Accordingly, spammers try to manipulate one or both of these factors to improve the ranking of their web sites for a specific set of keywords. This practice is commonly known as 'link spam' and 'content spam'.

Link spammers have different options to create fraudulent links. They can create dummy web sites which link to the website they want to push (link farms), exchange links with other webmasters, buy links on third party web pages, and post links to their websites, for instance, in blogs. To detect link spam, much research has been performed, among others [5-12].

Content spammers try to make their website appear more relevant for a certain keyword search than it is. This can be accomplished by taking content of other websites and combining different (stolen) texts as 'new content', or by stuffing many keywords in a web page's meta tags[1], title, ALT-tags of images, the body text, creating doorway pages, and placing invisible text on a web page. Invisible text usually means text in the same color as the background or text in layers which are behind the normal text or which are invisible. Again, much research has been performed to identify content spam, among others [13-15].

Although spammers are continuously adjusting their methods and developing new techniques (e.g. scraper sites, page hijacking, social media spam, Wikipedia spam, and gadget spam), overall, search engines are capable to fight web spam quite well.

No studies are available, to our knowledge, on the existence of spam in *academic* search engines or whether it can be recognized and prevented.

3. METHODOLOGY
We modified already published academic articles by adding references and (invisible) text. These modified articles were then uploaded as PDF to the Web to see whether Google Scholar was indexing them. Currently, not all experiments are completed. The current results are presented in the following. A more detailed analysis and explanation of the methodology shall be provided in another paper as soon as all experiments are completed.

4. RESULTS
Google Scholar did index the PDFs with invisible text and grouped these PDFs with the original article. That means, a researcher could add invisible keywords to his article after its publication and upload this PDF to the web. This way a researcher could make the article appear for keyword searches the article originally was not relevant for.

[1] Meta tags usually are not used by spammers any more since most search engines ignore meta tags due to spam issues

Moreover, Google Scholar did count references that were added to a modified version of an already published article. Citation counts and rankings of the cited articles increased. That means, authors could add additional references in their articles after official publication. If these altered articles were uploaded to the Web, Google would index them. This way, researchers could increase citation counts and rankings of the additionally cited articles. They could also arouse more attention to their articles because the cited authors might investigate who has cited them. It is to assume that researchers could also modify articles from other authors and add references to their own articles. This way, scholars could create the impression that their articles were cited by an authority in their field and increase citation counts as well.

5. DISCUSSION

Our study on the robustness of Google Scholar delivers surprising results: It seems that Google Scholar is far easier to spam than the classic Google Search for web pages. Apparently, Google Scholar applies no or only very rudimentary mechanisms to detect and prevent spam. With comparatively little effort we could manipulate articles' citation counts and hence their rankings and make Google Scholar indexing invisible text.

6. OUTLOOK

This poster is work-in-progress. We are currently conducting more experiments. For instance, we created nonsensical text with the random paper generator SciGen to see if this text, when published on the Web, is indexed by Google Scholar. We are also analyzing whether we can make one article appear as several search results to spam result sets and whether PDFs containing advertisement is indexed by Google Scholar.

We would like to note that the intention of this paper was not to expose Google Scholar. The intention was to stimulate a discussion about academic search engine optimization. We chose Google Scholar as the subject of our study because Google Scholar probably is the best and largest academic search engine, indexing PDFs from the Web. Currently, we are developing our own academic search engine SciPlore (www.sciplore.org), however as yet, SciPlore has not any precautions against spam either.

7. REFERENCES

[1] Jöran Beel and Bela Gipp. Google Scholar's Ranking Algorithm: The Impact of Citation Counts (An Empirical Study). In André Flory and Martine Collard, editors, *Proceedings of the 3rd IEEE International Conference on Research Challenges in Information Science (RCIS'09)*, pages 439–446, Fez (Morocco), April 2009. IEEE. doi: 10.1109/RCIS.2009.5089308. ISBN 978-1-4244-2865-6. Available on http://www.sciplore.org.

[2] Jöran Beel and Bela Gipp. Google Scholar's Ranking Algorithm: An Introductory Overview. In Birger Larsen and Jacqueline Leta, editors, *Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI'09)*, volume 1, pages 230–241, Rio de Janeiro (Brazil), July 2009. International Society for Scientometrics and Informetrics. ISSN 2175-1935. Available on http://www.sciplore.org.

[3] Jöran Beel and Bela Gipp. Google Scholar's Ranking Algorithm: The Impact of Articles' Age (An Empirical Study). In Shahram Latifi, editor, *Proceedings of the 6th International Conference on Information Technology: New Generations (ITNG'09)*, pages 160–164, Las Vegas (USA), April 2009. IEEE. doi: 10.1109/ITNG.2009.317. ISBN 978-1424437702. Available on http://www.sciplore.org.

[4] Jöran Beel, Bela Gipp, and Erik Wilde. Academic Search Engine Optimization (ASEO): Optimizing Scholarly Literature for Google Scholar and Co. *Journal of Scholarly Publishing*, 41 (2): 176–190, January 2010. doi: 10.3138/jsp.41.2.176. University of Toronto Press. Available on http://www.sciplore.org.

[5] Z. Gyöngyi and H. Garcia-Molina. Link spam alliances. In *Proceedings of the 31st international conference on Very large data bases*, page 528. VLDB Endowment, 2005.

[6] AA Benczur, K Csalogány, T Sarlós, and M Uher. SpamRank – Fully Automatic Link Spam Detection. In *Adversarial Information Retrieval on the Web (AiRWEB'05)*, 2005.

[7] I. Drost and T. Scheffer. Thwarting the nigritude ultramarine: Learning to identify link spam. *Lecture Notes in Computer Science*, 3720: 96, 2005.

[8] D. Fetterly, M. Manasse, and M. Najork. Spam, damn spam, and statistics: Using statistical analysis to locate spam web pages. pages 1–6, 2004.

[9] A. Benczúr, K. Csalogány, and T. Sarlós. Link-based similarity search to fight web spam. *Adversarial Information Retrieval on the Web (AIRWEB), Seattle, Washington, USA*, 2006.

[10] H. Saito, M. Toyoda, M. Kitsuregawa, and K. Aihara. A large-scale study of link spam detection by graph algorithms. In *Proceedings of the 3rd international workshop on Adversarial information retrieval on the web*, page 48. ACM, 2007.

[11] B. Wu and K. Chellapilla. Extracting link spam using biased random walks from spam seed sets. In *Proceedings of the 3rd international workshop on Adversarial information retrieval on the web*, page 44. ACM, 2007.

[12] Q. Gan and T. Suel. Improving web spam classifiers using link structure. In *Proceedings of the 3rd international workshop on Adversarial information retrieval on the web*, page 20. ACM, 2007.

[13] T. Urvoy, T. Lavergne, and P. Filoche. Tracking web spam with hidden style similarity. In *AIRWeb 2006*, 2006.

[14] I.S. Nathenson. Internet infoglut and invisible ink: Spamdexing search engines with meta tags. *Harv. J. Law & Tec*, 12: 43–683, 1998.

[15] C. Castillo, D. Donato, A. Gionis, V. Murdock, and F. Silvestri. Know your neighbors: Web spam detection using the web topology. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, page 430. ACM, 2007.