Webology, Volume 3, Number 1, March, 2006

Home	Table of Contents	Titles & Subject Index	Authors Index

Editorial

Alireza Noruzi

Link Spam and Search Engines

The growing number of blogs has caused problems for search engines, problems such as the highly frequent blog spam. Spammers use blogs to promote their websites. Spammers are trying to win the attention of search engines, not of bloggers or their readers.

Spam in blogs (also called simply *blog spam* or *comment spam*) is a form of *search engine spamming* done manually or automatically by posting random comments, promoting commercial services, to blogs, wikis, guestbooks, or other publicly-accessible online discussion boards. Any web application that accepts and displays hyperlinks submitted by visitors may be a target of '*Link Spam*' (Wikipedia, 2006b). This is the placing or solicitation of links randomly on other sites, placing a desired keyword into the hyperlinked text of the backlink. Blogs, guest books, forums and any site that accepts visitors' comments are particular targets and are often victims of drive-by spamming, where automated software creates nonsense posts with links that are usually irrelevant and unwanted (Wikipedia, 2006a).

Link spam dishonestly and deliberately manipulates link-based ranking algorithms of search engines like Google's PageRank to increase the rank of a web site or page so that it is placed as close to the top of search results as possible. A link-based ranking algorithm gives a higher ranking to a site that has many backlinks, especially from highly-ranked sites/pages.

The link spammers' underlying assumption is that link spam within the comments of blogs increases traffic to a site from search engines and optimizes site backlinks to help search engines index the site better. The big advantage of this to spammers and marketers/advertisers is that it probably gives their site a great *PageRank* (Google's ranking algorithm). The more links the spammers can propagate across the Web, the better their rankings in the search engine results.

According to the *PageRank* methodology explanation, Google interprets a link from page 'A' to page 'B' as a vote, by page 'A', for page 'B'. But Google considers more than the sheer volume of votes, or [back]links a page receives; it also analyzes the page that casts the vote. Votes cast by pages that are themselves "*important*" weigh more heavily and help to make other pages "*important*" (Google, 2006).

If site 'A' links to site 'B', Google calculates this as a *vote* for site 'B'. The higher the number of votes, the higher the overall value for site 'B'. In a perfect web society, this would be true. However, some bloggers and authors abuse the system, implementing '*link spam*' and Google bombing: linking to web sites that have little or nothing to contribute. It is obvious that web bloggers and authors have been able to "bomb" Google and are still playing with Google and other search engines.

Spammers may use different methods to spam, for example they may praise other blogs, linking to their own site/page. However, in some cases it is not clear whether it is spam or not. Therefore, not all blog comments are spam. But now the question arises, "Do search engines use backlinks in blog comments to crawl and rank web sites?" When a search engine crawls a blog comment, its crawler reads what the anchor text says about the page that it is linked to, and then follows each link to index the target page/site and the topic or theme of the page. From a search engine's point of view, *anchor text* determines the topic of the page the link points to.

A simple search on Google shows that it displays blog comments and thus it presents site backlinks in blog comments. For example,

"Comments +on" site:blogspot.com

It seems that backlinks in comments can increase the visibility, popularity and PageRank of backlinked sites. However, Google does not allow a link search to be restricted to a special site. Therefore, it is not possible to perform a combination search with link command (i.e. link:). For example,

link:webology.org AND site:blogspot.com

But it is possible to perform a combination search with link command on Yahoo. A link search shows that Yahoo counts links in certain blog comments for the Webology site and blog. For example,

linkdomain:webology.persianblog.com AND site:netbib.de
linkdomain:webology.org AND site:netbib.de
linkdomain:webology.org AND site:blogspot.com

Blog Spam Solutions

Most blog software and services now have spam prevention options and anti-spam solutions. General solutions include:

1. Deleting every spam in blog comments manually;
2. Changing the comments template to display the comment's associated URL as text only, instead of a hyperlinked address, converting all URLs to text, to make them useless in increasing the URL's PageRank.
3. Using the Authimage plugin in the server. This plugin asks the commenter to insert a randomly generated code displayed in an image before accepting the comment.
4. Adding the *rel="nofollow"* attribute to links in blog comments so that search engine robots do not follow the link when crawling and ranking pages. Google, MSN and Yahoo support a tag called "nofollow" to exclude links in blog comments from search-engine crawlers and to prevent spam posts from influencing search rankings (Hicks, 2005; MSN, 2005).
5. Blocking spam in blog comments, or banning the IP address of recognized spammers;
6. Removing the comments part of the blog;

It can be concluded that because of spamming, search engines should ignore links in blog comments.

Articles in This Issue

This issue includes three articles and a book review. The first article concerns with aspects of *stemming and root-based approaches to the retrieval of Arabic documents*, the second article discusses *unsolicited commercial e-mail and legal solutions* and the third article examines *environmental knowledge and marginalized communities*. The final section of this issue is a short review of the book '*Digital Libraries: Principles and Practice in a Global Environment*' written by Lucy A. Tedd and Andrew Large.

Haidar Moukdad: *Stemming and root-based approaches to the retrieval of Arabic documents on the Web*. The Web is a growing multilingual information resource. Users generally prefer to access web pages in their native language. Search engines are the interface between users and the billions of web pages. English-language search engines have their own limitations in indexing and retrieval of non-English pages, especially Asian alphabets (e.g. Arabic, Persian, Chinese, Japanese, and Korean). The author argues that "rules, theories, algorithms, and retrieval methods designed and developed for English and other morphologically similar languages may or may not apply in the linguistic environments of other languages. The problem is particularly acute in languages that differ radically from English on account of morphological rules. This paper compares the effects stemming and root retrieval on information retrieval in Arabic through an exploratory study of the handling of Arabic words by an English-language search engine."

Li Xingan: *E-marketing, Unsolicited Commercial E-mail, and Legal Solutions*. This paper explores the legal solutions to spam, doing a comparison between traditional and e-mail advertisements. It discusses the costs and benefits of the spammer and the spammed. The spam, which is unsolicited commercial e-mail, poses challenges for all e-mail users. The author argues that "spam messages are annoying in that the users have to spend time and money dealing with them" and concludes that "comprehensive mechanisms must be established to protect the spammed and to discourage the spammer."

A. Neelameghan & Greg Chester: *Environmental Knowledge and Marginalized Communities: The Last Mile Connectivity*. It is noteworthy to mention that in obtaining and transferring the knowledge received from a foreign country, all factors and elements should be carefully considered, because otherwise "a missing apparently minor detail made a great difference." Thus, adaptation, experimentation, and flexibility to local or national circumstances are the best ways to assure that progress in knowledge transfer will be secure, rapid, and long-lived. Consider how information collected should be disseminated in local languages. "The contents should be carefully selected and planned and should relate to the needs of the target audience." It is concluded that "the knowledge and technology are essential for the well-being of not only the Indigenous and rural peoples but also the economically and technologically advanced peoples. Both can gain from the exchange of ideas. On the other hand if the interchange is handled inappropriately it can be either of little benefit for the Indigenous people at best or very harmful at worst. Therefore, much planning and cross cultural communication must occur before sharing begins."

References

- Google (2006). Our search: Google technology, PageRank explained. Retrieved Mach 25, 2006 from http://www.google.com/technology/index.html
- Hicks, Matthew (2005, January 18). Search Engines, Bloggers Team to Fight Spam. eWeek. Retrieved Mach 25, 2006 from http://www.eweek.com/article2/0,1759,1752331,00.asp
- MSN (2005, January 18). Working Together Against Blog Spam. *MSN Search's WebLog*. Retrieved Mach 25, 2006 from http://blogs.msdn.com/msnsearch/archive/2005/01/18/nofollow_tags.aspx
- Wikipedia (2006a). Spam in blogs. *Wikipedia*, the free encyclopedia. Retrieved Mach 25, 2006 from http://en.wikipedia.org/wiki/Spam_in_blogs
- Wikipedia (2006b). Spamdexing. *Wikipedia*, the free encyclopedia. Retrieved Mach 25, 2006 from http://en.wikipedia.org/wiki/Spamdexing

Bibliographic information of this paper for citing:

Noruzi, A. (2006). "Editorial: Link Spam and Search Engines." *Webology*, **3**(1), editorial 7. Available at: http://www.webology.org/2006/v3n1/editorial7.html