A PARTITION AWARE ENGINE AND SPAM DETECTION FOR SEARCH ENGINES

SURAJ PAWAR
M.Tech, Dept. of CSE, Marri Laxman Reddy Institute of Technology and Management, Telangana.
Email: surajpwr7@gmail.com,

Mr. B. PRASAD
Associate Professor, Dept. of CSE, Marri Laxman Reddy Institute of Technology and Management, Telangana

ABSTRACT
Now days as additional and more people rely on the affluence of information accessible through online, enlarged interest on the World Wide Web may acquiesce considerable financial gains for individuals or organizations. Most frequently, search engines are the entryways to the Web that is why some people try to mislead search engines, so that their pages would rank high in search results, and thus, capture user attention. In the last few years, this model of reaching relevant information through the use of search engines has become pervasive. Numerous sites on the network observe an ever-increasing part of their traffic coming from search engines referrals. The purpose of a search engine is to provide high quality results by correctly identifying all web pages that are most appropriate for a specific query, and presenting the user with some of the most important of those relevant pages. Relevance is usually measured through the textual similarity between the query and a page. Pages can be given a query-specific, numeric relevance score; the higher the number, the more relevant the page is to the query.

I. INTRODUCTION
Web spamming refers to actions intentional to mislead search engines into ranking some pages higher than they deserve. In recent times, the web spam has enlarged severely, foremost to a degradation of search results. Spam suffuse any information system, be it e-mail or web, social, blog or reviews platform. The concept of web spam was first introduced in 1996 and soon was recognized as one of the key challenges for search engine industry. Recently, all major search engine organization have identified adversarial information retrieval as a top priority because of multiple negative ejects caused by spam and appearance of new challenges in this area of research. First, spams deteriorate the eminence of search results and dispossess justifiable websites of revenue that they might earn in the absence of spam. Second, it weakens trust of a user in a search engine provider which is especially tangible issue due to zero cost of switching from one search provider to another. Third, spam websites serve as means of malware and adult content dissemination and fishing attacks. For instance, ranked 100 million pages using Page Rank algorithm and found that 11 out of top 20 results were pornographic websites that achieved high ranking due to content and web link manipulation. Last, it forces a search engine company to waste a significant amount of computational and storage resources. However, many web site operators try to influence the ranking functions of search engines by using less-ethical gray-hat and black-hat SEO techniques. These include the creation of extraneous pages which link to a target page (link stuffing). Using link stuffing, web sites

can increase the desirability of target pages to search engines using link-based ranking. The content of other pages may be engineered so as to appear relevant to popular searches, a technique known as keyword-stuffing. The hope is that the target pages will rank high within the search engine results for included terms and thereby draw users to visit their web sites. The practices of crafting web pages for the sole purpose of increasing the ranking of these or some affiliated pages, without improving the utility to the viewer, are called "web spam". Figure 1 shows an example of a spam page: this page contains important keywords; however its content is, on the whole, useless to a human viewer.

MASSIVE big graphs are prevalent nowadays. Prominent examples include web graphs, social networks and other interactive networks in bioinformatics. The up to date web graph contains billions of nodes and trillions of edges.

Graph structure can represent various relationships between objects, and better models complex data scenarios. The graph-based processing can facilitate lots of important applications, such as linkage analysis community discovery, pattern matching and machine learning factorization models.

II. TYPES SEARCH ENGINE SPAM

The perspective of SEOs is spammers since their actions are intentional to make better rankings of a page without actually humanizing the quality of that page.

Web spam is detrimental to search engines in two ways because it:

- Reduces the quality of search results
- Increases the cost of each processed query due to the storage and retrieval of useless pages

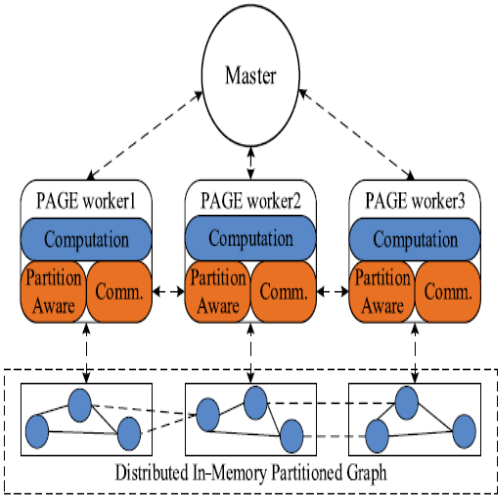
Mainly in search engines spams are categorized into following ways.

a. Term Spam
b. Content spam
c. Link spam
d. Cloaking and Redirections
e. Click spam

Term Spamming: Term spamming refers to the practice of search engine spamming. It is a form of SEO spamming. SEO is an abbreviation for Search Engine Optimization, which is the art of having your website optimized, or attractive, to the major search engines for optimal indexing. Term spamming is the practice of creating websites that will be dishonestly indexed with a high situation in the search engines. Sometimes, Term Spamming is used to try and manipulate a search engine's understanding of a category. The objective of a web designer is to create a web page that will find constructive rankings in the search engines, and they create their pages according the standards that they believe will help – unfortunately, some of them resort to spamdexing, unbeknown to the person who hired them.

Content Spamming: Once popular, but not particularly effective anymore, is hiding content using background and foreground colors that match. Hiding links is only slightly harder and can be achieved with 1×1 pixel images. CSS brought with it a few new tricks such as setting page elements to be not

visible along with other tricks like negative indents.



(a) Architecture

Link Spamming: There are two major categories of link spam: outgoing link spam and incoming link spam.

- Outgoing link spam
- Incoming link spam

Click Spam: Since search engines use click stream data as an implicit feedback to tune ranking functions, spammers are eager to generate fraudulent clicks with the intention to bias those functions towards their websites. To achieve this goal spammers submit queries to a search engine and then click on links pointing to their target pages [92; 37]. To hide anomalous behavior they deploy click scripts on multiple machines or even in large bot nets [34; 88]. The other incentive of spammers to generate fraudulent clicks comes from online advertising. In this case, in reverse, spammers click on ads of competitors in order to decrease their budgets, make them zero, and place the ads on the same spot.

Cloaking and Redirection: Cloaking is the way to provide di↵erent versions of a page to crawlers and users based on information contained in a request. If used with good motivation, it can even help search engine companies because in this case they don't need to parse a page in order to separate the core content from a noisy one (advertisements, navigational elements, rich GUI elements). However, if exploited by spammers, cloaking takes an abusive form. In this case adversary site owners serve di↵erent copies of a page to a crawler and a user with the goal to deceive the former [28; 108; 110; 75]. For example, a surrogate page can be served to the crawler to manipulate ranking, while users are served with a user-oriented version of a page. To distinguish users from crawlers spammers analyze a user-agent field of HTTP request and keep track of IP addresses used by search engine crawlers. The other strategy is to redirect users to malicious pages by executing JavaScript activated by page onLoad() event or timer. It is worth mentioning that JavaScript redirection spam is the most widespread and difficult to detect by crawlers, since mostly crawlers are script-agnostic.

III. CONCLUSION

In this paper we presented a variety of commonly web spamming and we have studied various aspects of search engine spam on the web. It is also possible to address the problem of spamming as a whole, despite the differences

Among individual spamming techniques. This paper aim identification of some common features of spam pages. For instance, the spam detection methods presented in [5] take advantage of the approximate isolation of reputable, non-spam pages: reputable web pages seldom point to spam. Thus, adequate link analysis algorithms can be used to separate reputable pages from any form of spam, without dealing with each spamming technique individually.

REFERENCES:

1. P. Metaxas and J. DeStefano. Web Spam, Propaganda and Trust. In 1st International Workshop on Adversarial Information Retrieval on the Web, May 2005.
2. Z. Gy¨ongyi, H. Garcia-Molina and J. Pedersen. Combating Web Spam with TrustRank. In 30th International Conference on Very Large Data Bases, Aug. 2004.
3. D. Fetterly, M. Manasse and M. Najork. Detecting Phrase-Level Duplication on the World Wide Web. In 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Aug. 2005.
4. D. Fogaras and B. Racz. Towards scaling fully personalized pagerank. In Proceedings of the 3rd Workshop on Algorithms and Models for the Web-Graph, WAW'04, 2004.
5. 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, AIRWeb'07, Ban↵, Alberta, 2007.
6. J. Piskorski, M. Sydow, and D. Weiss. Exploring linguistic features for web spam detection: a preliminary study. In Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web, AIRWeb'08, Beijing, China.
7. B. Poblete, C. Castillo, and A. Gionis. Dr. searcher and mr. browser: a unified hyperlink-click graph. In Proceedings of the 17th ACM conference on Information and knowledge management, CIKM'08, 2008.
8. G. Mishne, D. Carmel and R. Lempel. Blocking Blog Spam with Language Model Disagreement. In 1st International Workshop on Adversarial Information Retrieval on the Web, May 2005.
9. Dennis Fetterly, Mark Manasse, and Marc Najork. Spam, damn spam, and statistics. In Proceedings of the Seventh International Workshop on the Web and Databases (WebDB), 2004.
10. Zolt´an Gy¨ongyi, Hector Garcia-Molina, and Jan Pedersen. Combating web spam with TrustRank. In Proceedings of the 30th International Conference on Very Large Databases (VLDB), 2004.