Proceedings of "National Conference on Emerging Trends in Science and Advances in Engineering"
Organized by Phaltan Education Society's College of Engineering Phaltan,
www.coephaltan.edu.in
International Journal of Innovations in Engineering Research and Technology [IJIERT]
ISSN: 2394-3696, Website: www.ijiert.org, July, 2023

PARALLEL INDEXING ENGINE AND SPAM FILTERING FOR SEARCH ENGINES

Suraj Pawar*
Department of Computer Engineering, DBATU University,
PES's COE, Phaltan.
Email: surajpwr7@gmail.com

Abstract-
The growing reliance on online information and the potential financial benefits associated with it. Search engines serve as gateways to the vast amount of information available on the World Wide Web, and as a result, some individuals or organizations attempt to deceive search engines to achieve higher rankings in search results, aiming to attract user attention. This practice has become increasingly common in recent years.

Many websites now receive a significant portion of their traffic from search engine referrals. The main objective of a search engine is to provide high-quality search results by accurately identifying web pages that are most suitable for a specific query and presenting them to the user. Relevance is typically assessed based on the textual similarity between the query and a web page. Pages are assigned a query-specific numeric relevance score, where a higher score indicates greater relevance to the query.

I. INTRODUCTION:-

Web spamming refers to intentional actions taken to deceive search engines in order to obtain higher rankings for certain web pages. This practice has become increasingly prevalent and has led to a degradation of search results. Web spam can be found in various information systems, including email, social media, blogs, and review platforms.

The concept of web spam was first introduced in 1996 and has since been recognized as a significant challenge for the search engine industry. Major search engine organizations have identified adversarial information retrieval as a top priority due to the negative effects caused by spam and the emergence of new challenges in this research area.

There are several negative effects of web spam. Firstly, it diminishes the quality of search results and deprives legitimate websites of the revenue they could earn in the absence of spam. Secondly, it undermines users' trust in search engine providers, which is particularly concerning given that users can easily, switch to alternative search providers. Thirdly, spam websites can serve as a means of spreading malware, adult content, and engaging in phishing attacks.

Web spam also imposes a significant burden on search engine companies in terms of computational and storage resources. Additionally, many website operators attempt to manipulate search engine rankings using unethical techniques known as gray-hat and black-hat SEO. These techniques include link stuffing, where extraneous pages are created to link to a target page, and keyword stuffing, where the content of pages is engineered to appear relevant to popular searches.

Crafting web pages solely for the purpose of increasing rankings, without improving the utility for users, is known as "web spam." These pages often contain important keywords but lack meaningful content for human viewers.

On a different topic, massive big graphs have become prevalent in various domains, such as web graphs, social networks, and bioinformatics. These graphs can consist of billions of nodes and trillions of edges. Graph-based processing is valuable for analyzing relationships between objects and enables applications like linkage analysis, community discovery, pattern matching, and machine learning factorization 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.

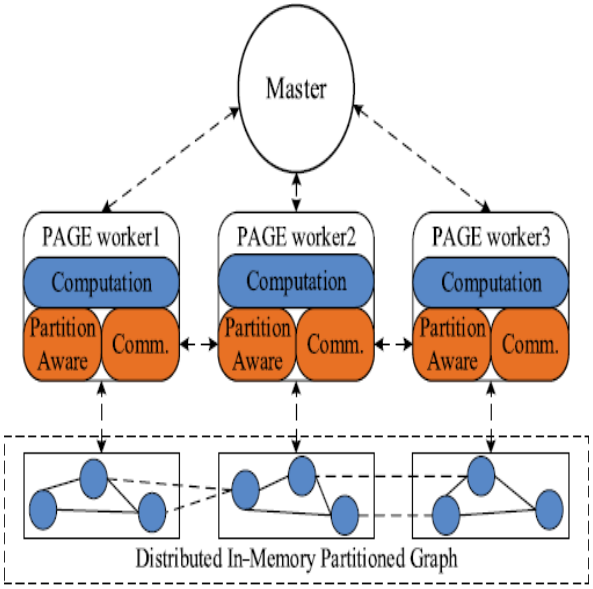
- Increases the cost of each processed query due to the storage and retrieval of useless pages

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 deferent 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 deferent 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 on Load () 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.

I. 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 Trust Rank. In 30th International Conference on Very Large DataBases, 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 theWeb-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 featuresfor web spam detection: apreliminary study. In Proceedingsof the 4th International Workshop on Adversarial InformationRetrieval 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.