EPH - International Journal of Science And Engineering
ISSN (Online): 2454 - 2016
Volume 05 Issue 04 December 2019

DOI: https://doi.org/10.53555/eijse.v5i4.34

SURVEY ON WEB SPAM AND ITS UNDERLYING PRINCIPLES

Tanvir Kaur[1]*, Varinder Kaur Attri[2]
*[*1,2]Department of Computer Science Guru Nanak Dev University Campus Jalandhar Tkaur1992@gmail.com*
[2]Email: Varinder2002@yahoo.com.

Corresponding Author:-
Email: Tkaur1992@gmail.com

Abstract:-

Search engines have become a de facto place to start information acquisition on the Internet. Although due to web spam phenomenon, search results are not always as fine as they are expected. Moreover, spam evolves that makes the problem of providing high quality search even more challenging. Over the last decade research on information retrieval has gained a lot of interest both from academics and industry. In this paper, systematic review of web spam detection techniques and underlying principles are presented. Existing algorithms are categorized into three categories based on the type of information they use: content-based methods, methods based upon links and methods based on non-traditional data such as user behavior for e.g. clicks and image spam is given. A brief survey on various spam forms is provided. Finally, the underlying principles are summarized.

Keywords: - *Click spam. Search engine. Spamming. Fradulent clicks. URL*

I. INTRODUCTION

Spam attacks any information system on internet, an email or web, social, blog or any review platform. The concept of web spam was first introduced in 1996 [1] and soon it was recognized as one of the key challenges for the search engine and internet industry. Now, all major search engine companies have identified information retrieval as a top priority because of its multiple negative effects and appearance of new challenges in this area of research. First, spam decends the quality of search results and deprives websites of revenue that they might earn in the absence of spam [2]. Secondly, it weakens trust of a user in a search engine provider, which is especially a tangible issue due to zero cost of switching from one search provider to another and cut throat competition. Third, spam websites serves as means of malware and adult content broadcast and phishing attacks. It is worth mentioning that the necessity of dealing with the malicious content is a key distinctive feature of information retrieval [3] in comparison with the traditional information retrieval methods, where algorithms operate on comparatively simpler and clean benchmark data set.

II. WEB SPAM TAXONOMY

A. Content spam

The content spam is the first and most widely spread form of web spam. Because, search engines use information retrieval models based on a content of pages to rank the web pages according to their opening and clicks. Taking a document structure of spam, into account there are 5 subtypes of content spamming.

(i) Title Spamming: Due to high importance of the title field for information retrieval, spammers have a clear incentive to overstuff it so as to achieve higher overall ranking [4].

(ii) Body Spamming: In this case, the body of a page is modified. This is the most common form of content spam because it is cheap and simultaneously allows to apply different strategies. For instance, if a spammer wants to achieve a high ranking of a page for only limited predefined set of queries, they can use the repetition strategy by overstuffing body of a page with condition that appear in the set of queries.

(iii) URL Spamming: Some search engines also consider a tokenized URL of a page as a zone of spam attack. Therefore the spammers create a URL for a page from words which should be mentioned in a targeted set of the queries. For example, if a spammer wants to be ranked high for the query, they can create a URL which lures the user.

(iv) Meta-Tags spamming: meta-tags play a specific role in a today work in the form of document description, search engines analyze them carefully. Due to this, the placement of spam content in this field of document

Might be very prospective from spammer point of view -very flexible in his strategies. He can create and Meta tags are targeted more. Because of the heavy his own link farm and carefully tune its spamming, nowadays search engines give a low priority topology to guarantee the desired properties to this field or even ignore it completely while ranking and optimality. One common ---link farm the page. Topology has been found, named as a honeypot farm. In this case a spammer creates

B. Link Spam

Two major categories of link spam are:

(i) Outgoing link spam: This is an easy method of link spam because a spammer has a direct access to his pages and therefore can add any items to them, and secondly, they can easily copy the entire web catalogue and therefore quickly create a large set of authoritative links, accumulating relevance score

(ii) Incoming link spam: In this case spammers try to raise a score of a page or simply boost a number of incoming links. One can identify the strategies depending on an access to pages

- Accessible Pages: These are the pages which and therefore can add any items to them, and secondly, spammers can modify but don't own. For they can easily copy the entire web catalogue and instance, it can be Wikipedia pages, blog with therefore quickly create a large set of authoritative links, public comments, a public discussion group, accumulating relevance score or even an open user-maintained web raise a score of a page or simply boost a number of incoming links. One can identify the strategies depending
- Own pages [7]: In this case a spammer has a both link and anchor text spamming direct control over all the pages and can be- techniques very flexible in his strategies. He can create his own link farm and carefully tune its topology to guarantee the desired properties and optimality. One common ---link farm topology has been found, named as a honey potfarm. In this case a spammer creates a page which looks absolutely innocent and may be even authoritative to the spammer's target pages

Criteria	Link Propagation	Link pruning and Reweighting	Label Refinement	Graph Regularization	Feature Based
Algorithm	PageRank TrustPage	Hits PageRank	Clustering Algorithms	PageRank	Truncated Pagerank
Working	Exploits the topological relationship between the web pages	Identify the suspicious nodes and links and their subsequent downweighting	Extacting linkbased features for each node and use various machine learning algorithms	Uses the idea of label refinement based on the web graph topology	Work based on graph regularization method
Mining Techniques	WSM	WSM,WCM	WCM	WSM	WSM
Type of information used	Topological Relationship	Downweighting of nodes & links	Link base features of each node	URL	Structural Patterns
Complexity	Internal Structure	Relationship between the nodes	Limited data set are allowed	URL Classification	_

Table 1: comparison of link based Spam Detection methods

Cloaking:
Is the way to provide different 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 with advertisements, navigational elements, and rich GUI elements. However, if exploited by spammers, cloaking takes form of an abuse. In this case, site owners serve different copies of a page to a crawler and a user with the goal to deceive the former [13]. For example, a surrogate page can be served to the crawler to manipulate ranking of previous page, *while* users are served with a user-oriented version of a page. To distinguish genuine users from crawlers, spammers analyse a user-agent field of HTTP request and keep track of IP addresses used by search engine crawlers. It is worth mentioning that JavaScript redirection spam is the most widespread and difficult to detect by crawlers.

C. Click Spam
Since search engines use click stream data as a feedback to tune in ranking functions, spammers are quick to generate fraudulent clicks with the intention to move those functions towards their websites. To achieve this target, spammers submit queries to a search engine and then click on links pointing to their target pages. To hide anomalous behavior, spammers deploy click scripts on multiple machines. The other incentive of spammers to generate fraudulent clicks that comes from online advertising [5]. In this case, in reverse, spammers click on ads of competitors in order to decrease their budgets to make them minimal and place the ads on the same spot i.e on the space of competitor's advertisement.

D. Image spam
The past few years have seen a novel approach, where the spammers embed the text message into an image. Thus, the Research of anti-spam filtering is bound to shift from textbased techniques to image-based techniques. Spam blockers [15] designed to combat it have spawned has an upsurge in creativity and innovation. Many software developers are developing new and every more effective spam filtering software.
Secure Computing Research has seen an increase of 50 percent in just the past spell. In that time, there has also been a tripling in the amount of spam that is image spam, which today accounts for 30 percent of all spam.
It consists in embedding the spam message into images which are sent through email as attachments. Its goal is to misguide the analysis of the emails' textual content performed by appointed spam filters, including automatic text classifiers. Since there is facility of displaying attached images by default by most email clients, the message is directly conveyed to the user as soon as the email is click opened. To detect image spam by conventional content filters, is very difficult. New filter techniques are needed. Often spam images are constructed by making random changes in a given template image, to make signature-based detection techniques ineffective, and are designed to prevent optical character recognition (OCR).

III. Anti-spam strategies
(i) Prevention based
This approach aims at making it tough for spam content to contribute towards social tagging system by restricting its certain access types through interfaces such as CAPTCHA (completely automated public Turing test to tell computers and humans apart) or through usage limits as tagging quota e.g Flickr introduced a limit of 75 tags per photo.[16]

(ii) Detection based

This type of approaches identify likely spam either manually or automatically by making use of machine learning or statistical analysis and then deleting the spam content or visibly mark as hidden to the user. For these methods, the corpus is treated as set of objects along with associated attributes. In e-mail spam, the messages are considered as objects and the headers are attributes. In spam of web, the web pages are objects and attributes might be in links, out links, page content and various external data.

(iii) Demotion based

This approach reduces the prominence of content likely to be spam. For instance, Rank based methods produce order of a system's content, tags or users based on the score of trust.

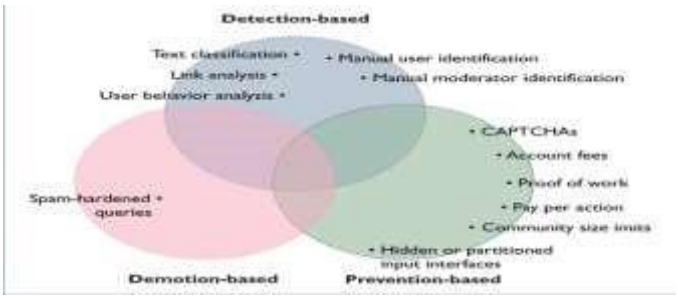

Fig 1. Anti-spam strategies [internet]

IV Key principles

After analyzing the related works devoted to the topic of web spam mining, set of underlying principles are identified which are frequently used for algorithm construction

- Due to machine-generated nature and focus on search engines manipulation[12], spam shows abnormal properties such as high level of duplicate content and links, rapid changes of content and the language models built for spam pages that deviate significantly from the models built for the normal Web.
- Spam pages deviate from the power law distributions based upon numerous web graph statistics i.e PageRank or number of in-links.
- Spammers mostly goals upon popular queries and queries which have high advertising value in market.
- According to experiments, the principle of approximate isolation of good pages (i.e. pages with genuine content) takes place [8]: good pages mostly link to good pages, while bad pages link either to good pages or a few selected pages to spam target pages. It has also come out that connected pages have some level of semantic similarity therefore label smoothing using the Web graph is a useful strategy.
- Several algorithms use the idea of trust and distrust propagation using various similarity measures such as propagation strategies and seed selection heuristics.
- Due to abundance of "neponistic" links [11] that negatively affect the performance of a link mining algorithm there is a popular idea of link removal and down weighting. Moreover, it is beneficial to analyze local subgraphs rather than the entire Web graph.
- Because, one spammer can have large number of pages under one website and use them all to boost ranking of some target pages, it will be good to analyze host graph or even perform clustering and consider clusters as a logical unit of link support.
- In addition to traditional page [9] content and links, there are plenty of other sources of information such as user behaviour or HTTP requests. More should be developed in the near future. Clever feature engineering is especially important for web spam detection.
- Despite the fact that new and sophisticated features can boost it further [14], Proper selection and training of a machine learning models is also of high importance.

V. Conclusion

In this work we surveyed existing techniques and algorithms created to fight against web spam. To draw a general picture of the web spam phenomenon, firstly provide numeric estimates of spam on the web, then discuss how spam affects users and search engine companies.

According to this work, web spam detection research has gone through a distance: starting from simple content based methods to approaches using sophisticated link mining and user behavior mining techniques. Furthermore, current anti-spam algorithms show a competitive performance in detection of spam. However, spam is constantly evolving and still negatively affects many people and businesses, so more research is needed in this area. It is believed that even more exciting and effective methods will be developed in the future. Among promising directions of research: identify click fraud for online advertising detection and construction of platforms, which don't have incentives for non-fair behavior.

For instance, pay-per-click models having this property will be very beneficial. Detection of cloaking is an open issue. There is potential and need of anti-spam methods at the intersection of Web and social media.

Acknowledgement
First of all I express my sincerest debt of gratitude to the Almighty God who always supports me in my endeavors.
I would like to thank my guide Mrs. Varinder Kaur Attri for her encouragement and support. I am thankful to all those who helped me in one way or the other at every stage of my work.

Refferences

[1]. K. Chellapilla and A. Maykov. "A taxonomy of javascript redirection spam" In Proceedings of the 3rd international workshop on Adversarial information retrieval on the web, AIRWeb'07, Canada, 2007.
[2]. J. Abernethy, O. Chapelle, C. Castillo, J. Abernethy, O. Chapelle, and C. Castillo. WITCH: A new approach to web spam detection. In Proceedings of the 4th International Workshop on Adversarial Information, 2008.
[3]. M. R. Henzinger, R. Motwani, and C. Silverstein. Challenges in web search engines. SIGIR,2002.
[4]. Z. Dou, R. Song, X. Yuan, and J.-R. Wen. Are clickthrough data adequate for learning web search rankings? Information and knowledge management, 2008.
[5]. N. Immorlica, K. Jain, M. Mahdian, and K. Talwar. Click Fraud Resistant Methods for Learning Click- Through Rates. Technical report, Microsoft Research, Redmond, 2006.
[6]. Nikita Spirin, Jiawei Han, Survey on Web Spam Detection: Principles and Algorithms Department of Computer Science, SIGKDD Explorations Volume 13, Issue 2, 2011
[7]. S. Chakrabarti. Mining the Web: Discovering Knowledge from Hypertext Data. Morgan Kaufmann, 2002.
[8]. R. Bhattacharjee and A. Goel. Algorithms and Incentives for Robust Ranking. Technical report, Stanford University, 2006.
[9]. M. Najork. Introduction to Web spam detection, 2006.
[10]. S. Nomura, S. Oyama, T. Hayamizu, and T. Ishida. Analysis and improvement of hits algorithm for detecting web communities, Japan, 35, Nov. 2004
[11]. R. Lempel and S. Moran. SALSA: "The stochastic approach for link-structure analysis". ACM Trans. Inf. System, April 2001
[12]. C. Castillo and B. D. Davison. "Adversarial web search: Found. Trends" , 4, May 2011.
[13]. B. Wu and B.D. Davison. "Detecting semantic cloaking on the web." In Proceedings of the International Conference on World Wide Web, WWW'06, Edinburgh, Scotland, 2006.
[14]. K.K. Arthi M.Sc1, Dr. V.Thiagarasu "A Study on Web Spam Classification and Algorithms" International Journal of Computer Trends and Technology (IJCTT). volume 4 , Sep 2013
[15]. Dhanraj S; Karthi keyani, V. "A study on e-mail image spam the we (at the 14th International World Wide Web Conference) chiba, filtering techniques" IEEE, Salem. ISBN- 978-1-4673-5843-9, 2013 Japan, 2005..
[16]. ZoltanGyongi; Hector Garcia-Molina.,"Web Spam Taxonomy ",.First International workshop on Adversarial Information Retrieval on the We (at the 14th International World Wide Web Conference) chiba, Japan, 2005.