ISSN:0975-9646

Mining Search Log for Privacy Definitions and Phish Safe

Janani. V.D

Teaching Fellow, Dept. of Computer Science and Engineering,
Anna University Guindy, Chennai

***Abstract*— Search engine spamming is a practice of misleading the search engine and increasing the page rank of undeserving websites. The black hat search engine optimization (SEO) techniques leads to untrustworthy results for search engines. Some commonly used black hat techniques has been characterized and proposed a new way to counter those techniques using link based spam detection technique. This technique enhances discovering of the target page and it investigates the problem of protecting privacy for publishing search engine logs. With the exponential growth of the available information on the WWW, current search engines do not just store and index web pages, they also store and mine information about their users. Hence the attackers can actively influence the search log and therefore the privacy of the user is lost. A novel algorithm called ZEALOUS and SLICING has been introduced to publish privacy guaranteed search logs. The result of this paper filter out the malicious sites from the search results and enables the search logs available without disclosing their user's sensitive information.**

***Keywords*— Search Engine Optimization, black hat techniques, link based spam detection technique, ZEALOUS, SLICING, privacy preserving.**

I. INTRODUCTION

Search engines play a crucial role in the navigation through the vastness of the Web. With increasing dependencies on World Wide Web (WWW), we rely on search engines to provide us right information at the right time. Search Engine Optimization (SEO) is the practice of building a web site search engine friendly, so that it can be found easily on the search engine with its relevant keywords. There are many free lancing SEO companies which provide such facilities. The main role of these companies is to list the websites on search engine. Some of those small SEO providers or other people also use some automated tools and/or other unethical techniques commonly known as black hat techniques to increase the traffic on website. These kinds of people, which use black hat techniques to increase the page rank of an undeserving websites, are known as spammers giving rise to spam. Due to their regular spamming, the search engine result becomes unreliable, untrustworthy and annoying. Present search engines do not just collect and index web pages, they also collect and mine information about their users. They store the queries, clicks, IP-addresses, and other information about the interactions with users in what is called a search log. Search logs contain valuable information that search engines use to tailor their services better to their users' needs. They enable the discovery of trends, patterns, and anomalies in the search behavior of users, and they can be used in the development and testing of new algorithms to improve search performance and quality.

Search engines such as Bing, Google, or Yahoo log interacts with their users. When a user submits a query and clicks on one or more results, a new entry is added to the search log. Without loss of generality, we assume that a search log has the following schema:

(USER-ID, QUERY, TIME, CLICKS),

where a USER-ID identifies a user, a QUERY is a set of keywords, and CLICKS is a list of url's that the user clicked on. The user-id can be determined in various ways; for example, through cookies, IP addresses or user accounts.

A. Spamming Techniques

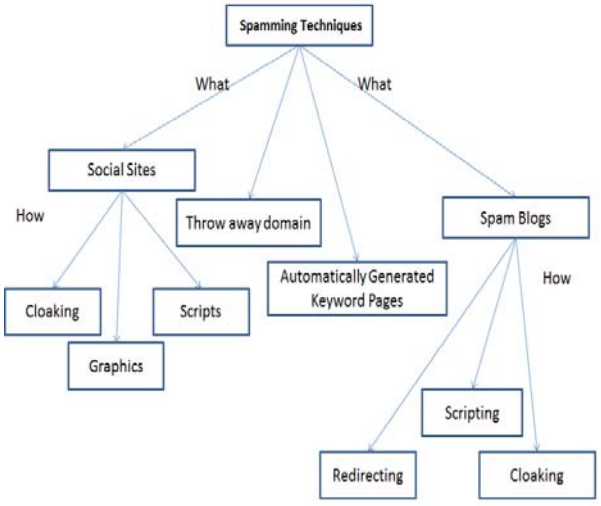

Fig. 1 Evolving Spamming Techniques

The spam technique is mainly classified into two basic categories, namely; boosting technique and the hiding technique. The boosting technique is the technique, which is used to make the page look more relevant to the search engine. One of the boosting techniques is keyword stuffing. Keyword stuffing is also known as on page technique. During the on page, the target keyword is stuffed into the web pages i.e. the HTML page, PHP page or any other available source page on the web server. These keywords are stuffed into HTML tag i.e. META tag, H1 tag, HEADER tag etc. Each tag is rated explicitly by the search engine and the summation of all the ratings provides the total keyword density for the particular page. Even the sub directory, URLs and contents are rated and included in

calculating the keyword density for the website. Hiding technique are the techniques that are used to hide the boosting techniques. These techniques are responsible for generating traffic from the user and misguiding the crawlers. Through hidden links a lot of anonymous traffic is generated. This can also be done by redirecting the domain to other domain. These techniques use the attractive platform, where it's easy to conceal and generate some heavy traffic by camouflaging with the website.

II. COMMONLY IDENTIFIED SPAMMING TECHNIQUES

With increasing age of search engine spammers, a variety of new techniques have been evolved, some of the commonly identified techniques are given below:

- *Social Networking sites:* Many social networking sites such as the twitter and facebook are targeted by the spammers. According to the spam analytical report in 2011, the 7% of the message on the twitter are spam. The links are scattered so as to increase the traffic on the webpage. This technique is the finest way to increase the traffic on the website.
- *Throwaway Domains*: Create new domains with short term that are very popular keywords such as gamesforchildrens.com, i.e. games for children's. These domains are used to practice some black hat techniques such as redirecting or clocking and are short life domain used to redirect to the target page.
- *Cloaking with the flash*: Cloaking with the scripts, that are not read by the search engines such as flash. When an flash embedded website is visited by the crawler, the crawler view its texts field only, like the header tag, Meta tag etc. but unable to predict, where it's flash will redirect the user. So such websites may lead user to some other domain whereas the crawler may not pursue the target. Some of the flash advertisements are also embedded into sites, which leads to the target site.
- *Automatically Generated Keyword Pages:* It has been found that many product selling websites create pages automatically according to the keywords applied on the search engines.
- *Blog Spam:* They use content management system (CMS) and install it in their websites with the spam protection on. Now, all they need to wait for someone to spam on them, then they parse the keyword on it, remove the link, and then they append their own links on it. It has proven as one of the most legitimate looking spamming technique.
- *Nofollow:* This tag is default for the blogs and forum sites, when a link is pasted on the websites. The human browsers only can follow it but not the search engine crawlers. For example:

your solution

In this way, the search engines can stop creating low valued backlinks. The backlinks are the inbound links, which are considered relevant according to the search engine. At present, automated tools are used to comment and blog links, which is not followed by the crawlers but generates a lot of traffic for the target page.

III. DISCLOSURE LIMITATIONS FOR PUBLISHING SEARCH LOGS

The problem of publishing keywords, clicks and other items of a search log are evaluated.

A. K-anonymity

A simple type of disclosure is the identification of a particular user's search history in the published search log. The concept of k-anonymity has been introduced to avoid such identifications. There are several proposals in the literature to achieve different variants of k-anonymity for search logs. One way to propose k-anonymity is to partition the search log into sessions and then to discard queries that are associated with fewer than k different user-ids. In each session the user-id is then replaced by a random number. The output of this algorithm is called k-query anonymous search log.

1) Insufficiency of Anonymity: K-anonymity and its variants prevent an attacker from uniquely identifying the user that corresponds to a search history in the sanitized search log. Nevertheless, even without unique identification of a user, an attacker can infer the keywords or queries used by the user. K-anonymity does not protect against this severe information disclosure.

B. Differential Privacy

An algorithm is differentially private if for all search logs S and S' differing in the search history of a single user and for all output search logs. Search logs that only differ in the search history of a single user as neighboring search logs are considered. From this the infrequent used keywords, clicks and other items can be hided. The Search Log S can be deleted by the user, whereas differential privacy ensures that the output of the algorithm is insensitive to changing/omitting the complete search history S' of a single user.

1) Impossibility of Differential Privacy: Differential privacy guarantees that an attacker learns roughly the same information about a user whether or not the search history of that user was included in the search log.

IV. LINK BASED SPAM DETECTION TECHNIQUE

Many websites use the nofollow, cloaking, spam blogs, social sites etc. methods to spam, where the crawler fails to detect the spamming. So, it is not possible to detect the spamming through the crawlers alone. Spammed pages are linked with non-spammed pages. Therefore, human interference is required. A spam detection technique is proposed, which uses identified links from the social networking websites or other spam detecting data sources.

A. Spam Detection

1) Starting with the root link: Since the crawlers are cloaked to the different results, a real browser is used to search and manipulate the result. First, the link is used to visit the target URL from the spam link with a real browser. Then, the target URL and the root URL is used to find different guest books, blogs, and forums etc. containing link.

2) Refining the Result set: The generated list may contain the false positive results too that can either be the good links posted on the same website or the spam links. Since the spam result is cloaked, automated programs that will use a real browser to visit all the primary provided URL and then record the secondary URL. Now, the software result can be associated with crawlers result. If both ends to the same page then it may be a good page else it may be the spammed page. Further, the result is transferred to the human expert i.e. the manual investigation to confirm its spamming URL.

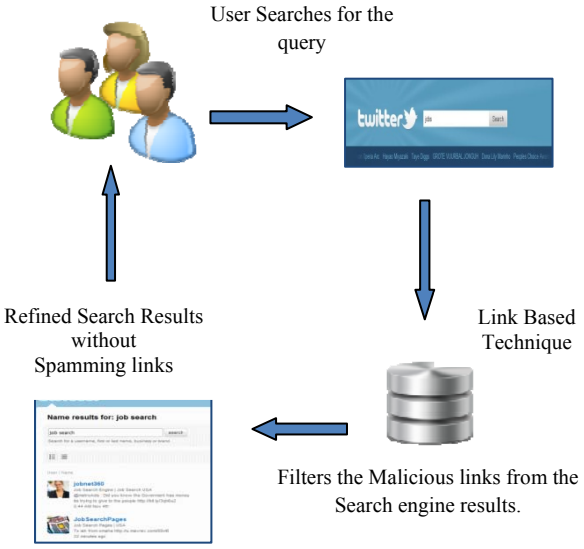

Fig.2 Spam Detection

V. ZEALOUS ALGORITHM FOR ACHIEVING PRIVACY IN SEARCH LOGS

A search log publishing algorithm called ZEALOUS has been introduced. ZEALOUS ensures probabilistic differential privacy, and it follows a simple two-phase framework. In the first phase, ZEALOUS generates a histogram of items in the input search log, and then hides from the histogram the items with frequencies below a threshold. In the second phase, ZEALOUS adds noise to the histogram counts, and eliminates the items whose noisy frequencies are smaller than another threshold. The resulting histogram (referred to as the sanitized histogram) is then returned as the output.

Algorithm ZEALOUS for Hiding Infrequent Items of a Search Log.

Input: Search log S, positive numbers m, τ, τ'

1. For each user u select a set of m distinct items from user's search history in S_u.
2. Based on the selected item's, create a histogram consisting of pairs (k, c_k), where k denotes an item and c_k denotes the number of users u that have k in their search history S_u. We call this histogram the original histogram.
3. Delete from the histogram the pairs (k, c_k) with count c_k smaller than the threshold value τ.
4. For each pair (k, c_k) in the histogram, sample a random number from the Laplace distribution Lap(λ)[4] and add to the threshold value, resulting in a noisy count.
5. Delete from the histogram the pairs (k,c_k) with noisy counts $c_k \leq \tau'$.

ZEALOUS algorithm enables the search logs available without disclosing their user's sensitive information. It achieves much stronger privacy guaranteed search logs.

A. Choosing Parameters

ZEALOUS requires the data publisher to specify two more parameters: τ, the first threshold used to eliminate keywords with low counts and m, the number of contributions per user. These parameters affect both the noise added to each count as well as the second threshold τ'.

1) Choosing threshold value τ: The threshold value is assigned based on the laplace distribution and the number of distinct items, m. A randomly generated value called noise gets iterated till all the keywords, i.e. the user's sensitive information gets deleted.

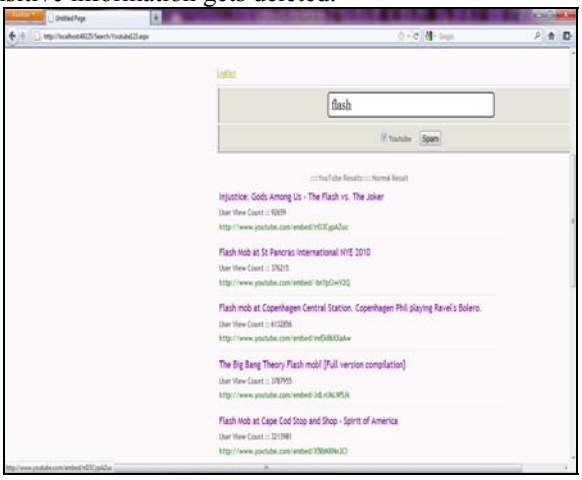

Fig.3 You Tube search result

Once the user login the system, the user can access the search engine. By giving the keyword, the query analyzer gets the keyword and brings back the respective search engine results. The Admin updates the database with the blacklisted URL's.

Fig.4 Anti Spammed results

You Tube search result for the keyword flash after implementing the Link Based Spam Detection Technique, i.e. without Spamming Links.

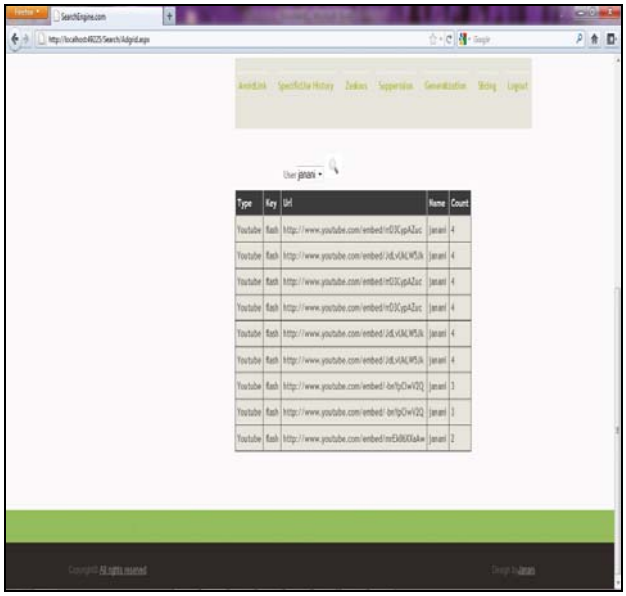

Fig.5 Specific User History

Admin maintains the individual search history of the user. After applying the Zealous Algorithm the infrequent URL's will be hid from the search log.

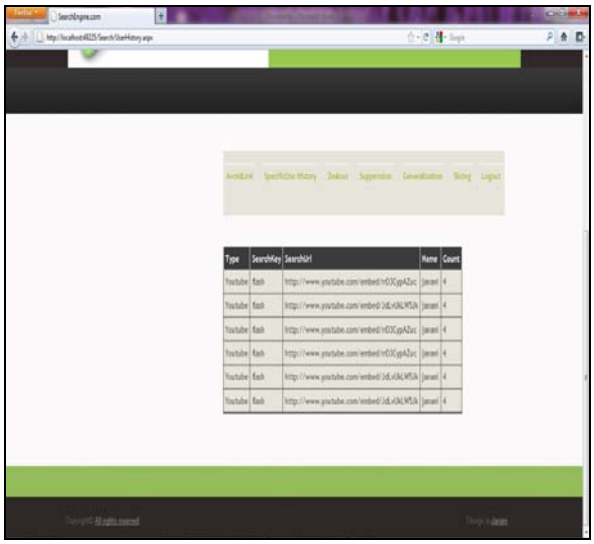

Fig.6 Zealous Algorithm

VI. SLICING ALGORITHM FOR PRIVACY GUARANTEED SEARCH LOGS

Slicing partitions the data set both vertically and horizontally. Vertical partitioning is done by grouping attributes into columns based on the correlations among the attributes. Each column contains a subset of attributes that are highly correlated. Horizontal partitioning is done by grouping tuples into buckets. Finally, within each bucket, values in each column are randomly permutated (or sorted) to break the linking between different columns. The basic idea of slicing is to break the association cross columns, but to preserve the association within each column. This reduces the dimensionality of the data and preserves better utility than generalization and bucketization.

A. Attribute Partition

An attribute partition consists of several subsets of A, such that each attribute belongs to exactly one subset. Each subset of attributes is called a column.

B. Tuple Partition

A tuple partition consists of several subsets of T, such that each tuple belongs to exactly one subset. Each subset of tuples is called a bucket.

C. Slicing

Given a micro data table T, a slicing of T is given by an attribute partition and a tuple partition.

Slicing preserves better utility than generalization and is more effective than bucketization in workloads involving the sensitive attribute.

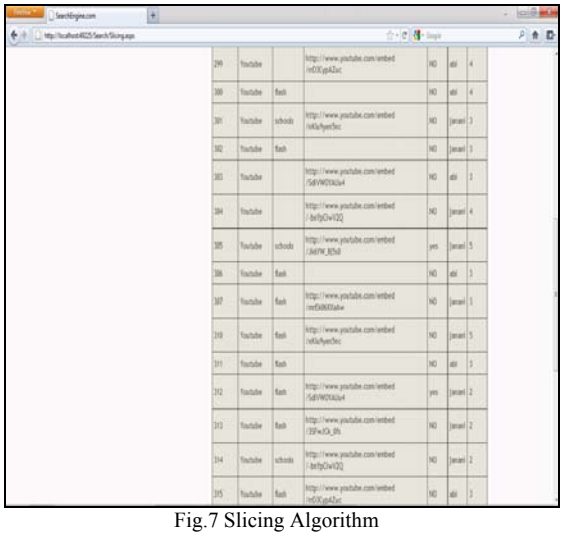

Fig.7 Slicing Algorithm

VII. CONCLUSION

With Changing WWW, various web spamming techniques are evolving. Present search engines should be an intelligent enough to know right data to mine upon. Researchers are required to make use of knowledge over knowledge to capture and counter the new minds creating the untrusted search engine results. Current search engines do not just collect and index web pages, they also collect and mine information about their users. The proposed Spam Detection technique can be used to identify and terminate the spammed websites and their entire graph patterns. A novel search log publishing algorithm called ZEALOUS has been introduced to publish privacy guaranteed search logs. These approaches filters out the malicious sites from the search results and enable the search logs available without disclosing their user's sensitive information.

REFERENCES

[1] Gotz.M, Machanavajijhala.A, Guozhang Wang, Xiaokui Xiao, "Publishing Search Logs- A comparative study of privacy guarantees", IEEE Trans, March 2012.

[2] Eun-Ae Cho, Kyungroul Lee, Kangbin Yim, "A Privacy Preserving Model for Personal Information in Search Engine", IMIS, 2012.

[3] Meng Cui, "Search engine optimization research for website promotion", ICM 2011.

[4] Antriksha Somani, Ugrasen Suman, "Counter Measures against Evolving Search Engine Spamming Techniques", 978-1-4244-8679-3/11 IEEE, 2011.

[5] Pedram Hayati, Vidyasagar Potdar."Toward Spam 2.0: An Evaluation of Web 2.0 Anti-Spam Methods", Digital Ecosystem and Business Intelligence (DEBI) Institute, Curtin University of Technology, Perth, WA, Australia, 2011.

[6] Yuan Hong, Xiaoyun He, Jaideep Vaidya, Nabil Adam, and Vijayalakshmi Atluri, "Effective anonymization of query logs". In CIKM, 2009.

[7] Michaela G"otz, Ashwin Machanavajjhala, Guozhang Wang, Xiaokui Xiao, and Johannes Gehrke. "Privacy in search logs". CoRR, abs/0904.0682v2, 2009.

[8] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith, "Calibrating noise to sensitivity in private data analysis". In TCC, 2006.

[9] M.E. Nergiz, M. Atzori, and C. Clifton, "Hiding the Presence of Individuals from Shared Databases," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), pp. 665-676, 2007.

[10] Mon-Fong Jiang, Shian-Shyong Tseng, Shan-Yi Liao, "Data Types Generalization for Data Mining Algorithm" National Chiao Tung University, 2009.