Discernment of Search Engine Spamming and Counter Measure for It

Sukrati Agrawal
M.Tech. Scholar CS Dept.
MIST, Indore (MP), India

Antriksha Somani
Assistant Professor CS Dept.
MIST, Indore (MP), India

Vishal Chhabra
Asst. Prof. CS/IT Dept.
MIST, Indore (MP), India

ABSTRACT
In today's world everyone is glancing for online information through search engine. As there are lots of providers for information searched by the user, and it is not possible to display all the information on the first page of search engine. In current scenario website owners uses SEO(search engine optimization) techniques to be on first page of search engine result. But there are some owners who approach illegal techniques to gain high ranking for their website which is known as Spamdexing or search engine spamming. Spamdexing is a black hat SEO technique in which spammers spoofs the user to a web page which is not high ranked by the search engine in order to earn more profit or degrade the efficiency of search engine. This paper presents taxonomy of current Black Hat SEO techniques through which a web spammer gains high rank for his web pages & also the counter measures to overcome these spammed result.

Keywords
Search Engine, SEO, Black hat SEO, Search engine Spamming, cloaking, page ranking algorithm, Spamdexing, Countermeasure against spamming.

1. INTRODUCTION
A search engine is a type of system software which search WebPages' content in World Wide Web (WWW) based on the user query (combination of keywords). Search engine build the index of website content used for information retrieval. When a user submits a query to search engine then it returns a list of abounded pages that are most relevant to the specified query. Search Engine ranks the result based on its algorithm. Most of the internet the internet users are concerned for first page result. Some SEO experts also use unethical techniques known as Black Hat SEO techniques in order to drive high traffic on their website which ensure that a site is accessible to a search engine and improves the probability that the site will be at high rank in search engine result page (SERP) [1].

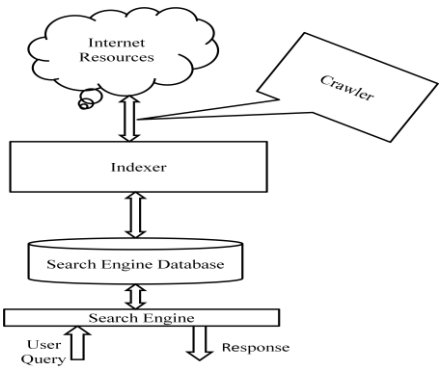

Fig 1: SEO working Process

1.1 Search Engine Optimization
SEO is a process of increasing traffic to a given website by increasing the site's visibility in search engine results. SEO experts improve webpage relevancy by improving content, making sure that the pages are able to be indexed correctly. Figure 2 depicts the classification of Search Engine Optimization.

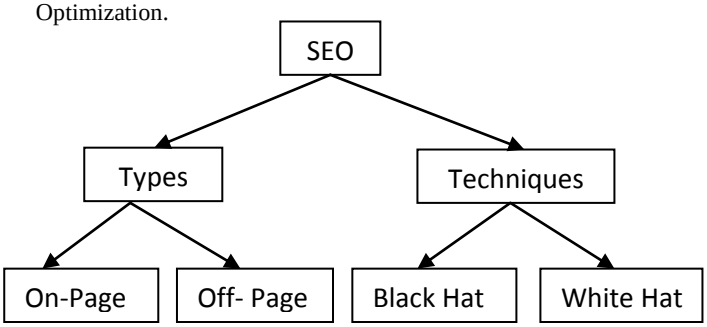

Fig 2: classification of SEO

1.1.1 On page Optimization:- On-page SEO (also known as "on-site" SEO) is the act of optimizing content of your web site that affect your search engine rankings. It includes providing good content, good keywords selection, putting keywords on correct tags and calculating the keyword density of a page.

1.1.2 Off page Optimization: Off-page SEO focuses on increasing the inbound links to go websites. It includes link building, increasing link popularity by submitting open directories, search engines, link exchange, etc.

1.1.3 White Hat SEO - It's good optimization techniques that fulfill search engines guidelines and lasts long time. It includes keyword research, keyword analysis, re- writing of Meta tags or Meta contents. Its main aim is to promote accessibility of both users and Search engine.

1.1.4 Black Hat SEO – Search engine techniques used to get higher search rankings in an unethical manner by going against current search engine guidelines are known as Black hat SEO. These techniques are also known as spamdexing. Black hat SEO techniques include stuffing of keywords, doorway and cloaked pages, link farming, hidden texts and links and blog comment spam.

1.2 State-Of-The-Art Spamdexing Techniques
Spamdexing is a combination of two word "spam" and "indexing," that refers to the art of search engine spamming which causes deliberate manipulation of search engine rankings result in order to get more users' hits on an undeserving website to acquire higher ranking in major search engines. Fig 3 depicts the classification of spamming techniques.

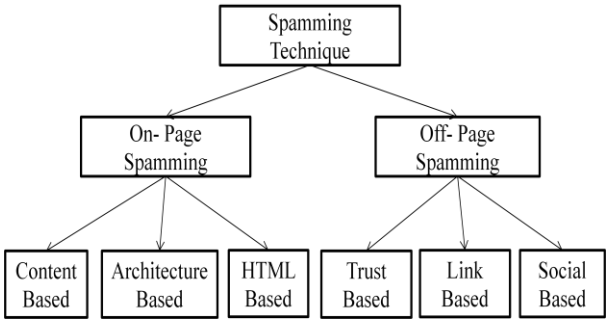

Fig 3: Types of Spamming Technique

Spamdexing techniques are classified into two major categories, On page spamming & off- page Spamming. They are further classified as follows.

1.2.1 On-Page Spamming

*1.2.1.1 Content based Spamming***:** Content based spam refers to changes in the content of the pages. Its classification includes the number of words in page text, the number of hyperlinks and the redundancy of the content [2].

A. Keyword stuffing: This technique is also known as keyword spamming. In this technique words are repeated and the frequency of words on a page is high which increases the probability of getting high search engine ranking results.

B. Duplicate content: Content is called duplicate if there are two or more pages that shares common content. This is further classified into three broad categories: True Duplicates, Near Duplicates, and Cross-domain Duplicates.

C. Doorway pages: These are low quality content pages which consist of keywords & phrases rather than relevant information in order to increase the page rank. It is also known as thin pages.

1.2.1.2 Architecture Based: It focuses on the design of website. It checks what a search engine crawl. The architecture is further classified into following.

A. Cloaking: Cloaking is a SEO technique which misguides the user by sending them to a page which is a different version of a web page crawl by search engine for indexing. Claoking includes different types such as IP cloaking, User agent cloaking & repeat cloaking [1,3].

B. Typosquatting: Typosuqatting is derived from two words, Typo & Squatting. The word Typo means "writing mistake" & Squatting means "to sit on ones heels". In this technique typosquatter registers its domain name which is just similar to a popular brand name website, like google.com, if visitors mistyped the domain name googel.com (rather than google.com) then it redirect the user to googel.com which is of not user choice. This technique is used to snare unaware web users.

C. Cybersquatting: Cyber Squatting is also a combination of two words cyber (computer network) & squatting (one that settles on property without right). It is the practice of registering a domain name which is may be a trademark, or a replica of like one, for one's profit. That means using other identity to earn more profit.

1.2.1.3 HTML based: This technique is based on text hiding and Meta tag stuffing.

A. Hidden or invisible text: In this technique color of back ground & text, font size is too small in order to hide it within the HTML code. This is useful to make a page appear to be relevant in a way that makes it more likely to be found.

B. Meta tag stuffing: Redundancy of keywords in the Meta tags and use of keywords that are not related to the site's content in order to redirect user on an undeserving website.

1.2.2 Off-Page Spamming Technique : In off- page spamming technique web spammer tires to increase the number of linking sites to his site rather than applying spamming technique to his own website. These techniques deal with the promoting strategies of websites which leads to higher ranking in search engine result.

1.2.2.1 Trust Based: In this off-page spamming technique the link is created with a pages with already has higher ranking in SERP. This includes host parasiting attack and piracy ads.

1.2.2.2 Link Based: In Link Based spamming an outbound of more the hundred link is created to acquire high ranking. This type of link outbound is known as link farm. Link building is another measure on which the search engine relies to update page rank of a website.

1.2.2.3 Social based: In this technique a renowned social site is used for the promotion of spammed website.

2. LITERATURE SURVEY

In paper "Counter Measure Against Evolving Search Engine Spamming Techniques"[1] Link Based Spam Detection along with Page Rank Algorithm is used which helps in identification of link from social through which the target website & its spam graph pattern can be discovered.

Alexandros Ntoulas et al. [2] investigated the web spam using a content analysis technique. For the detection of spam they took different attribute in consideration which is no of words in a page, no of words in the title of a page, average word length, and many more. In order to achieve the spam detection they did an experimental work on the MSN data set. when they used these attribute in isolation then they are not able to identify the all of spam web page so in order to achieve high accuracy they combined these methods with C4.5 classifier which works on decision tree. The amount of spam web page content can be identify using a content analysis technique.

Patel Trupti et al [3] have discussed different cloaking detection tool on the basis of some parameter through which search engine can identify the presence of fraudulent pages in website. Cloaking is spamming technique in which page deliver to user browser is different from the page at bot [2].

Nipendra Narayan Das et al [4] provides comparative study of different page ranking algorithm with their pros & cons. Whenever user search for online information then search engine provides a colossal list of relevant page link as a result and it's not possible for a user to go through each and every page In order to provide best related information to user in quick time the search engine uses different page rank algorithms to make user's navigation easier and faster. The sequence of the resulting link of pages depends upon the type of the page ranking algorithm used.

Web spam manipulates the existing web pages with the intention to raise their ranking in search engine. As better

rankings affect the number of visitors to a site, attackers use different techniques to boost their pages to higher ranks. Web spam pages provide undeserved advertisement revenues to the page owners. Sometimes it also poses a threat to Internet users by hosting malicious content and launching drive-by attacks against unsuspecting victims. In the paper, "Removing Web Spam Links from Search Engine Results" Web spam detection is discussed on the basis of search engine results by Manuel Egele et al [5]. By removing spam sites from the results, more slots are available to links that point to pages with useful content.

3. PROPOSED SPAM DETECTION TECHNIQUE

In this electronic era everyone is sought for information through search engine and the information providers want to be at the apex ranking of search engine hence they uses different illegal techniques to acquire high ranking for their web pages. In order to redirect user to deserving web pages rather than the undeserving one our proposed system is providing counter measure for such search engine spamming.

3.1 Detection of Spammed websites

3.1.1 Creating a root database: Root consists of links of previously caught spammed website. The Database is prepared using content based as well as link based detection. Content Based Spam detection deals with on-page data analysis while Linked based concern with the off-page spam analysis. Content based spam detection marks a website as spammed when it identifies the greater magnitude of keyword density while link based spammed detection identifies the number of inbound and outbound links of webpage [5]. When this link for a websites is found to be more than hundred then it is marked as spammed website.

3.1.2 Refining the Result: The overall spammed detection cannot be finalized based on root as root database may contain the false positive results also. The root may contain links of website which are really good [1]. So, a weighted page rank algorithm is applied on the root result.

3.2 Combining Spam detection with weighted page rank

Weighted Page rank is an important constituent in search engine result. It is a link based algorithm which assigns rank to a page based on number of incoming and outgoing link to that page. In comparison to page rank algorithm it divides the page rank unevenly [4]. Now, we will compare the result generated by weighted page rank algorithm with the root result to eliminate the false positive result from the root .

We propose an architecture which compares the webpage detected as spammed with its corresponding weighted page rank as shown in Fig. 4. A new spam detection technique is introduced in this architecture, which takes root database, High rank sites database, web and weighted page rank into consideration in marking a website as spammed. Finally when a website is detected as spammed after applying this overall technique then the spammed web site is penalize with the negative page rank which leads to the lowering the website position the SERP.

Suppose a user fires a query on "example.com" then for given user's query the result is searched in the web which has lots of pages for searched result. Instead of giving this result directly to the user, example .com first compare these result pages with the links in the root and if any match is found then the resulting page is penalized with the decrement in its page value. Now server fires the query: link in order to obtain all the linked pages with this spammed website and the entire graph responsible for this will be penalized. Finally the High rank sites data along with weighted page rank is preferred to display the result as per their ranking.

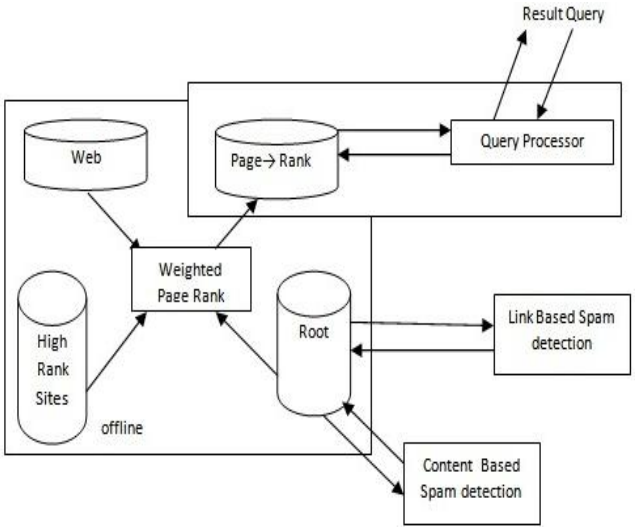

Fig 4: Proposed Solution

4. CONCLUSION

In this work we surveyed current black hat SEO techniques and proposed an architecture that can be used to identify the target spammed websites with the entire graph responsible for spreading spam. The entire graph responsible for spreading spam will e panelized with negative page rank hence lowers the position of entire graph position in SERP. We improve our search engine result by the use of content based analysis and weighted page ranking. Thus it will provide a more efficient spam free search engine result.

5. ACKNOWLEDGEMENT

With due respect we would like to inform, that the article which we had proposed is prepared for academic scenario, that is not feasible for industrial research purpose the references which we had taken from listed here, one or not only references there are many other references which we had taken from real world and from daily life aspects. So it is vary through to remember all those references.

6. REFERENCES

[1] Ugrasen Suman "Counter Measures against Evolving Search Engine Spamming Techniques" published in Electronics Computer Technology (ICECT) in 3rd International Conference,April 2011, Volume-6

[2] Alexandros Ntoulas, Marc Najork, Mark Manasse, Dennis Fetterly "Detecting Spam Web Pages through Content Analysis" published in 15th International World Wide Web Conference (WWW) May 2006.

[3] Patel Trupti1, Kachhadiya Kajal2, Panchani Asha3, Mistry Pooja "Search engine optimization: Black hat Cloaking Detection technique" published in International Journal of Engineering Research & Technology (IJERT) Vol. 2 Issue 10, October – 2013

[4] Nripendra Narayan Das, Ela Kumar,Sheetal "Approaches of page ranking algorithm" on International Journal of

Computer Applications, Volume 82 – No. 2, November 2013

[5] Manuel Egele, Clemens Kolbitsch, Christian Platzer "Removing web spam links from search engine results" Journal in Computer Virology on Volume 7 Issue 1, February 2011

[6] Pooja Devi,Alshlesha Gupta,Ashutosh Dixit "A comparative study of HITS vs Page Rank Algorithms for twitter users analysis" in Computational Science and Technology (ICCST), 2014 International Conference on 27-28 Aug. 2014

[7] Gurpreet Singh Bedi, Ms. Ashima Singh "Analysis of Search Engine Optimization (SEO) Techniques" in International Journal of Advanced Research in Computer Science and Software Engineering , Volume 4 on March 2014

[8] Ong Kok Chien, Poo Kuan Hoong, Chiung Ching Ho "Comparative study of HITS vs Page Rank Link based Ranking Algorithms" in International Journal of Advanced Research in Computer Science and Communication Engineering , Volume 3 on February 2014.