Web Spam Detection using Timer with Ranking Technique

Sajan Aggarwal
M-Tech Scholar
Department Computer Science and Applications
Maharshi Dayanand University,
Rohtak, Haryana-124001

Rajender Singh Chhillar,Ph.D
Head
Department Computer Science and Applications
Maharshi Dayanand University,
Rohtak, Haryana-124001

ABSTRACT
Today, Web Spam is a very serious problem for search engine and for the user. The word Web Spam is the combination of two words i.e. Web and Spam. The word web contains thousands of web pages which are used by various search engines to answer the query of the user and the word Spam means the unsolicited messages. The unsolicited message means the web pages are created with the intention of attracting high traffic for getting high score and for fun or for profit. There are various techniques available for detecting web pages such as content based technique, Link based technique, Title based technique etc.

This paper purpose a new technique in which how much time a user stay on a particular page is noted and on the behalf of that time it is observed whether a page is spam or not.

Keywords
Web, Spam, Cloud

1. INTRODUCTION
In Today world internet is a central and very important part of human life. In modern life, a lot of people use web pages for various purposes such as for gathering information, for shopping, for booking ticket, for online banking, for advertise their product and so on. Search engine acts as a mediator between user and the web database. A search engine is used to deliver the information to the user. Users enter the required keyword in the search engine and then the search engine transfers all the keyword's related links in the collection of 10 to 15 links per page and total of thousand links. But out of these links shown by the search engine, only top 10 to 15 links are informative for a user and rest of link are not informative. These non informative links or web page are not of any use for the user and are referred as Web Spam. The word Web spam is a made of two words i.e. Web and Spam. The word web contains thousands of web pages which are used by various search engines to answer the query of the user and the word Spam means the bulk messages. The bulk message means the web pages are created with the intention of getting high ranking and for fun or for profit. Identifying spam pages and take prevention from it, is a serious problem , one side web spammer waste the user valuable time and on the other side they decrease the search engine performance by putting a lot number of high ranking keywords in a page or by putting a large number of link and getting high hits for those information which are not deserve it. In addition to these, one major drawback is that web spam using resources such as space for storing page, indexing and ranking time of search engine. All these resources are occupy or used by those pages even they are not deserving it. Search engines would like to avoid spam pages that might be used for ranking, storing and indexing content. It is observe that major intention for creating spam pages is for money. Some time spam pages are contain malware. Malware mean when user open this type of page, malware silently get installed in the user system and do harmful in user system or stole important information from the user system or damage the computer system functionality. Web spammers work these types of work only for fun or for profit but generally the main reason behind all these tasks is for income by attracting high traffic or by getting high hits.

Generally web spammers mislead the search engine to show non informative information to the users. The main work of newly introduced approach is to remove those pages which are not giving informative information but still they are shown on the top position by getting high hits. But it is totally impractical to judge by human that a page is spam or not. So, there are various other method which are used to judge a web page is spam or not such as content analysis, link analysis, cloaking technique. If all method are used in combination to detect or to take judgment about a web page then search engine optimizer can say that a web page is spam page or not. But all previously develop technique still are not enough to stay ahead of web spammers and there is a need to develop new techniques or approaches to detect web spam.

According to Henzinger et al.[10] "Spamming has become so prevalent that every commercial search engine has had to take measures to identify and remove spam. Without such measures, the quality of the rankings suffers severely."

This paper introduced a new approach to find web spam i.e. web spam detection using timer. In this technique web spam is detected using timer which work on user focus for that particular page. In this technique if any user stay on a particular page then only points are added but as soon as if user click on another tab or in non-client area then no additional points are added. If any web spammer want to manipulate web result then it is impossible for him/her because this technique want the full focus of user on web page and as soon as focus is lost from web page then points which are allotted to page are also stopped. In this approach it is consider into observation that if a page is informative for a user then the user will read or must be doing his/her task and he/she will not loss its focus for page.

The benefit of this approach is for both the user and for the search engine. For user, user gets informative information on the top position and save the time. And for the search engine, with the help of this technique, the performance of search engine increase and resources are not misused by the spammers.

2. RELATED WORK

Web Spam is old as commercial search engines [4]. Today Web spam is a serious problem for both the user and for the search engine. A number of techniques are there for Web Spam. Hence, to carry out the work, large number of papers had to be surveyed, lots of information was collected. All these technique are used for giving high ranking to their web pages so that the spammer can get their page at top position.

2.1 Classification of spam

The name of the some techniques for web spam that is used by spammer to give high ranking to their pages are :-

(1) Content Spam based technique.

(2) Link Spam based technique.

(3) Cloaking spam technique.

2.1.1 Content based spam technique

Content based spam is used by putting high ranking keywords as content in web page. This work is done so that if a user search for a keyword then this page also come into query result of search engine. Some of the popular keywords are ONLINE, RESERVATION, COMPUTER, HARDWARE, RAILWAY etc. These types of popular keywords are used by web spammers as content of web page so that web spammer can attract high traffic and can get high hits for these type of web page which make user fool.

2.1.2 Link Spam based technique

Link spam is used by putting high ranking keywords as a hyper link in web page which targeted on spam page. Some of the example of these type of technique which is used by spammer is www.onlinerailway.com , www.reservation.com, www.computer.com, www.abc.com, www.hardware.com etc. Many link based spam technique used google's page rank technique which count the number of links into a page and also count page rank of the referring page[4].

2.1.3 Clocking based spam

Clocking spam is a technique which is used for getting high hits to their pages even they are not deserving it. In this technique, delivering different content via search engine to the user is used by the spammers. Clocking can be use with conjunction of many technique. Such as some of the part of the web page make invisible by the web spammer, by using client side scripting to rewrite the page after it has been delivered, by serving a page that immediately redirects the user's browser to a different page. Clocking spam is basically used with content spam [4].

2.2 Classification of Spam Detection Technique

There are also some algorithms or techniques which were developed for detecting web spam. All these techniques are developed with the intention for detecting web spam. Heuristic methods can also applied for detecting web spam. Some of these techniques are as follow

(1) Content based Web spam detecting technique.

(2) With the help of Ant Colony Optimization web spam detection technique.

(3) Anti-Trust ranking method for detecting web spam.

Before applying web spam detection techniques, the first step is to collect the data and follow the collection process which is required for testing the web spam detection algorithms performance. For collecting the data these things should be consider:-

(1) the collection should include many examples of spam and non-spam content. [2]

(2) The collection should contain little classification error. [2]

(3) The collection should be freely available for researchers. [2]

(4) The collection should include many different web spam techniques as possible.[2]

(5) The collection should represent a uniform random sample over a dataset.

2.2.1 Content spam detection technique

Content based techniques[1], number of words in the web page, number of words in page title are used to detecting whether a web page is spam or not. There are some words such as "THE", "A", "AN" which are used mostly each and every page and the these individuals words are used a number of times. If any web page not contains these common words then this page can consider as spam. There are also further method of content based i.e. amount of anchor text is used for taking decision about the web page is spam or not.

2.2.2 Ant colony optimization technique

Ant colony optimization technique[3] was used for detecting web spam. They also used content and link based feature with the ant colony technique. This technique is basically works on the behavior of the ant for detecting web spam.

There are various method for detect link spam. Link Spam detection problem can be used with ranking method or with the machine learning of classification of directed graph.[9] Anti Trust rank algorithm is latest or powerful technique which is used for fighting with web spam.

3. PROPOSED WORK

In today world, time is an important constraint that need to be focus upon when we develop any spam technique. New techniques that have been developed take less time in searching process as compared to the previous existing techniques. so considering time as an important factor a new technique has been proposed using timer.

Timer will evaluate the time user stays on a particular web page. The main objective of this proposed technique is to rank the web pages so that user can get informative information on the top instead of non-informative information. whenever user search for required topic this technique will display the web pages on the basis of their rank. The best pages with relevant information on the top to least interested irrelevant pages on the bottom In this approach, suppose a user enter any keyword then search show the related link in the combination of 10-15

links per page. Suppose user click on the first link and user stay on that page for only little time then this newly introduced technique consider this page as a spam page and give ranking less.
Next time when user search for that keyword then this page link will be show at last. The main motive of this technique to learn how much time a user stays on a particular page. If a user stay on a page for some time let say 1 minute then this page consider as informative page and then increase the rank of that page. To implement this task laptop with configuration i3 processor, 2 GB RAM, 320 GB HD with software Visual Studio 2008 as front end and SQL Server 2008 as backend are used.

3.1 Algorithm

This work has been implemented in ASP.Net with the help of JAVAScript tools and SQL Server 2008 as backend. Regarding this work one table have been used: cloud provider table.

The algorithm works as follows:

In this initially set zero rank or point to each page. Search engine first search according to title base and content based method and then order by according to the rank or point which is given to each and every page. The rank or point of pages will updated according to user behavior. There are many circumstances which play role for updating the rank of each page.

(1) In this technique, a user first enter a keyword in search engine
(2) Search engine search that particular keyword first in title and then in detail section of database which is made in SQL Server 2008
(3) Search engine give the result in the combination of 10 link per page
(4) When a user click on any link, the link will open in new tab
(5) As soon as user stay on that page, timer will count seconds and for each 10 seconds 1 point or rank is allotted to that page.
(6) But if user click on non-client area then timer will become stop and that value become store in database.
(7) In this purpose work also try to get the IP address of the each server, which help us to find pages which are actually non-informative but vendor of the page try to give high rank by open it again and again. This is possible by getting an IP address and valid it is for 4 to 5 days. If the same page is open by same IP address with in 4 to 5 days then no updating on rank will be done.\
(8) Continue from point 1

3.2 Implementation Work

Table 1:- Cloud Provider Table

ID	TITLE	DETAIL	WEBSITE	POINT
1	HOTEL	HOTEL IS THE PLACE WHERE WE CAN STAY AND EAT FOOD	www.abc.com	0
2	RAM	RAM IS A BOY WHO LEARN EVERYTHING	www.googlee.com	0
3	SEARCH	IF YOU WANT TO SEARCH ANYTHING THEN COME IN THIS SITE	www.xyz.com	0
4	RAMSHYAM	RAMSHYAM TEMPLE LOCATED IN INDIA	www.meits.com	0
5	SEARCH HOTEL	THERE IS A TECHNIQUE	www.hype.com	0

Cloud computing is a computing model in which resources are provided to end users as a service over internet. Many companies such as Google, Amazon, GoGrid, etc offer services from clouds.

According to my new developed technique, SQL Server 2008 as backend and one table with 5 columns or field has made. Initially value of point field is set to "0" for every newly row. The name of the field is in the above table 1. if any user search for a keyword i.e. "RAM" then it search first on title and then on detail and show the result in one page in the combination of 10 link per page. When a user click on a particular link, then this link will open in another tab and as soon as user is stay on that page, the timer will calculate the second and for each 10 second 1 point or rank is increased in database for that particular link. And when the same keyword is search again in future, then this page is show on top position as compare to its previous position. In this newly approach, This paper have used one more concept that is if any user either click on any non-client area after opening a link or add another tab to do other way side by side, then timer of that particular link will become stop. The benefit of this is that if a user open a page or link and if it is informative then he/she must be stay on that page but when a user jumps on second page or tab then timer will become stop and no additional point will be added. The main criteria used in this technique are that user focus for a page is considers and points are allotted to that page. In previously detection technique i.e. content based detection technique, if user enter any keyword then in content based technique only according to content decision are made whether a page is informative or not.

But the main disadvantage of previous developed technique is that user thought or behavior is not taken into observation. In

this newly approached technique user behavior or thought is also observe and on the behalf of that decision are made and points are allotted.

In the previous approach i.e. content base approach, if user enter a keyword i.e. RAM, then the search engine search the result only in detail section of the cloud and return the result of this query is www.googlee.com and www.hype.com [from table 1]. Let say the result of this query is one is informative and one is non-informative. If user click on the both the links, then each and every time, the search of above result is come in same order and may informative page on second position instead of non-informative page. But in newly approach, if user click on both the links but user find second page is informative then user will stay on second page. As well as user stay on second page, the points are allotted to the page and when next time user search again the same keyword then second page will come on top position instead of second position.

Table 2:- Comparison of two approaches i.e. Content approach and Timer with ranking approach

Values	**Content technique**	**Timer with ranking**
Search Engine Performance	Less	High
User Time Consume	High	Less
Resource wastage	High	Less
Effectiveness	Less	High

Thus, it is observe that the implementation of the purpose work yields search engine performance increase and save user time for searching information.

4. CONCLUSION AND FUTURE WORK

There are various different techniques which are used to make fool the web spammers. The results of some techniques are best and effective. At last if we want to stay in the front of web spammer then combination of technique must be used. By using this newly introduced approach, we can give high ranking to those page those are deserve it so that we can stay ahead of spammers. This paper an effort was made to find spam pages with the help of user thought or user focus. The future work is to modify into web spam techniques and to develop new technique so that search engine performance increase, the misuse of resources should be minimize and saves user's valuable time

5. REFERENCES

[1] Alexandros Ntoulas, Marc Najork, Mark Manasse, Dennis Fetterly, "Detecting Spam Web Pages through Content Analysis", International World Wide Web Conference Committee[2006].

[2] Carlos Castillo, Debora Donato, Luca Becchetti, Paolo Boldi, Stefano Leonardi, Massimo Santini, Sebastiano Vigna, "A Reference Collection for Web Spam".

[3] Arnon Rungsawang, Apichat Tawesiriwate, Bundit Manaskasemsak, "Spam Host Detection Using Ant Colony Optimization", Springer [2012].

[4] Marc Najork, "Web Spam Detection",

[5] Yiqun Liu, Min Zhang, Shaoping Ma, Liyun Ru, "User Behavior Oriented Web Spam Detection", National Science Foundation and National 863 High Technology Project, China [2008].

[6] Sumit Sahu, Bharti Dongre, Rajesh Vadhwani, "Web Spam Detection Using Different Features", International Journal of Soft Computing and Engineering [IJSCE], [2011].

[7] Luca Becchetti, Carlos Castillo, Debora Donato, Stefano Leonardi, Ricardo Baeza-Yates," Link Based Characterization and Detection of Web Spam", AIRWEB, Washington [2006].

[8] Andras Benczur, Istvan Biro, Karoly Csalogany, Tamas Sarlos, "Web Spam Detection via Commercial Intent Analysis", AIRWEB, Canada [2007].

[9] Dengyong Zhou, Christopher J.C. Burges, Tao Tao, "Transductive link Spam Detection", AIRWEB, Canada [2007].

[10] Carlos Castillo, Debora Donato, Aristides Gionis, Vanessa Murdock, Fabrizio Silvestri, "Know your Neighbors: Web Spam Detection using the Web Topology", SIGIR [2007].

[11] Jyoti Pruthi, Dr. Ela Kumar, "Anti-Trust Rank:- Fighting Web Spam", International Journal of Computer Science Issues,(IJCSI) [2011].