# Rating Approach For Web Spam Detection

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

Dr. R. S. Chhillar
Head of Department of Computer Science and Applications Maharshi Dayanand University, Rohtak, Haryana-124001

## ABSTRACT

Web Spam is not a new problem and it is not likely to be solved in near future. Web Spam can decrease the quality of any search engine and also can waste the user valuable time. There is a large number of techniques that must be use by search engine to detect spam and increase the performance of the search engine. In this paper we purpose a approach in which rating system is used with previously introduced approaches. Rating is given to any page by user. In this approach we combine the Content based spam detection technique and link based spam detection technique with time factor. Time factor is basically the amount of time spend by user on a particular page

In this paper, we purpose a new technique in which first content based detection is used then title based searching is done and according to rating to a particular page which is given by user points are collected in database. And on the basic of points collected by a page is observed whether a page is spam or not.

## Keywords

Web, Spam, Cloud, Content spam, Link Spam

## 1. INTRODUCTION

Everything have its advantage or disadvantage. On the one side, web play a good role in human life for providing the information about various topic via search engine. User enter the related keywords or data into search engine, then search engine shows the result from the web to the user. Peoples or organization use web for advertising their product or to share information with the user. And user use the web for various purpose such as for gathering information about any relevant topic , for Net banking, for purchasing various things, for booking e-ticket and vice versa. But on the other side some people or organization misuse it only for getting profit or for getting high collection of hits to their pages even they are not deserving. These types of peoples or organization are called as spammers. Spammers means people who flood the internet in an attempt to force the message on people who would not otherwise choose to receive it. Spammer basically use high ranking keyword or high ranking keywords as links in their web pages and many other things so that when a user search for a particular page, the spammer's page would shown on the top 10 position.

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."

Today, human use Internet for searching information about various topic and according to their need. People uses popular search engine such as Google, Alta vista, yahoo etc. for searching about various things. Now it is the duty of search engine to use best approach so that they can show informative information at the top position instead of non-informative information. Actually search engine work as a bridge between the web database and the user. When user enter the keyword into search engine , then search engine get thousand of result and shows the result 8-10 links per page. But out of thousands results, only few are good for user and a large collection of result are bad. These pages are on the web only for attractive high traffic so that they can get high hits from different user. The main purpose of attractive high traffic or getting high hits only for profit. The spammer use various technique such as high ranking keywords, high ranking links, clocking spam at their web page so that they can include their page in searching process of search engine. One popular practice when creating a spam page is "Keyword Stuffing"

The main drawback of these type of pages is that these pages use memory and time of search engine and user , even they are not valuable. On the one side, these type of pages decrease the performance of search engine by showing non informative information on the top of the position and misguide the user and to force the user to click on their link.

Generally web spammers mislead the search engine to show non informative information to the users. Our work 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 we use all method in combination to detect or to take judgment about a web page then we 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 spam. There are many approaches which was developed till now to find those type of pages which are non informative for the user but still they are on top position such as content based detection technique, title base detection technique, link base technique etc. 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.

**Overview of our newly introduced approach:-** In this paper, we introduced a new approach to detect web spam i.e. Ranking approach for web spam detection. In this technique web spam is detected using Ranking which work on user focus for that particular page and the rank given by user to page . In this technique the ranking are divided into 5 category like fuzzy logic such as Very Bad, Bad, Medium, Good, Very Good. if any user give ranking to page then in this approach email is required and after filling user email id, one link is send by search engine for verification. When user click on that verification mail then only rank points are added to a page. If any web spammer want to manipulate our result then it is impossible for him/her because our 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 and rating is giving to web page in one time in one day.

The advantage 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. LITERATURE REVIEW

Web Spam is old as commercial search engines [4]. Web Spam can be viewed as a binary classification problem mean a classifier is used to predict whether a web page is spam or not[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.

### 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.**One popular method use by spammer is keyword stuffing. In this the content of web page is stuffed with number of popular keyword which are not irrelevant to the rest of the page. The main purpose of doing so is so that by adding extra keyword with their legitimate content, their page also become result of search engine queries and they can attract high users and can get high profit.
Content spam mean a technique in which silent keywords are inserted into their web pages so that their web page can be result of the search engine query. 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 a user fool.

**2.1.2 Link Spam based technique.** Link spam refer to web spam technique that tries to get link base high score. This work is done by spammer. Link spam is used by putting high ranking keywords as a hyper link in web page which targeted on our 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 much technique such as some of the part of the web page make invisible by the web spammer by using font color and background color same, 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 we should consider this thing into consideration that:-

(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 we can consider this page 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 modern world, we know that time is an important constraint that need to be focus upon when we develop any spam detection technique but along with time, user behavior for a page also be consider. This is done with the help of rating system. New approaches that have been developed to take less time in searching process as compared to the previous existing approaches. so considering rating as an important factor a new approach has been proposed in this paper using rating system. Rating system is actually divided into five categories such as Very Bad, Bad, Medium, Good, Very Good. User of a page give rating to page according to his/her beneficial of the page. In this approach, when a user search for a keyword, search engine search the keyword on various cloud and return the query result in the form of links. When a user click on link then relevant page will open on next tab. Then there is a option of rating from the user such as Very Bad, Bad, Medium, Good, Very Good. When a user click on any option for give rating then in new approach it require email id of the user and when user fill his/her email id then a verification link will send and when user click on their verification link then only the rating points will added and according to rating point the position of page on search engine result is decided.

To implement this task we have used 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.

## 3.1 Algorithm

This work has been implemented in ASP.Net with the help of SQL Server 2008 as backend. Regarding this

work two tables have been used: cloud provider table, email id and IP address maintence table.

The algorithm works as follows:

Initially we give 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 Rating 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) At the above side of page, there is a 5 option button in series which indicate the different rating to page such as Very Bad, Bad, Medium, Good, Very Good same as fuzzy logic.

(6) For Very Bad 1 point, for bad 2 point, for medium 3 point, for good 4 point and for Very Good 5 points are added into the database.

(7) When user click on any rating option, then one text box will open which want email id of the user .

(8) After filling the user email id , one verification mail will send at the user mail id.[ This work is done because of verification of Email id]

(9) When user click on the verification link, then his/her rating points will considered and according to the user choice, points are allotted to the page and database will updated.

(10) In the above approach, one email id is valid for one day for one website link. It mean with the help of one email id, we can rate any web page only one time in one day [it is because to avoid the concept of misusing the rating system]

(11) In the above approach 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.

(12) But if user click on another tab then timer will become stop and that value become store in database.

## 3.2 Implementation Work

To implement this work, ASP.NET use for front end and SQL Server 2005 used for data cloud. In this paper, two table is used. One is cloud provider table, in which Title, detail, website link, ranking are used and in second table, Website link, Email id , IP address and date fields are used. The demo of table are as follows:-

Table 1:- Cloud Provider Table

| ID | TITLE | DETAIL | WEBSITE | Rating |
|---|---|---|---|---|
| 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, Go Grid, etc offer services from clouds. The main drawback of the previously introduced approaches are that in previous approach user rating point system is not consider with another approaches.

Here a second table is used in which email_id and User_IP address and date are managed. Table1 is used as a master table and in which website links are used as primary key and Table2 is used as a child table in which URL are as reference key.

Table2:- Email id and IP address table

| UID | URL | User_Emailid | User_IP | Date |
|---|---|---|---|---|
| 1 | www.abc.com | abc@xyz.com | 167.124.16.1 | 13/12/2012 |
| 2 | www.hype.com | abc@xyz.com | 167.124.16.2 | 13/12/2012 |
| 3 | www.xyz.com | sfs@jii.com | 167.124.16.2 | 13/12/2012 |
| 4 | www.abc.com | cdf@fds.com | 167.124.16.4 | 13/12/2012 |
| 5 | www.hype.com | sdf@gsdf.com | 167.124.16.1 | 13/12/2012 |
| 6 | www.abc.com | iew@aglsg.com | 167.124.16.6 | 13/12/2012 |

In this, user first enter a keyword in the search box. Search engine search the keyword from the cloud provider table and show the result in the combination of 8 to 10 links per page. When a user click on any link, the link page will open in second tab. At the top of the page, we have used rating system i.e. Very bad, bad, medium, good, very good. Five option buttons are used to get the user feedback about the page. If user get informative information on the page then he/she can give rating very good. when a user start giving rating then a text box will become visible in which email id of user is require. When a user fill his/her email id , then a verification email will send to relevant email id. And when user do click on the verification link, then only the rating point given by user will added on the data base with respective website link. Second thing, when user start giving rating to page, then IP address of the page is also taken into account so that with the help of one IP address , one time rating can be give to one web link. The benefit of this is that, spammer of the page cant give high value rating to page in continuous manner. Spammer must wait for next day to give rating to the web page.

## 4. CONCLUSION

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 we want to say that if 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. In this paper an effort was made to find spam pages with the help of user thought or user focus. Still there is a need to develop new techniques for detecting web spam so that we can stay ahead of spammers.

## 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].