ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

Web Spam Detecting using Ranking with using email id

Amit Dahiya, Reema, Shabnam Kumari

M.Tech Scholar, Department of CSE, Sat Kabir Institute of Technololy and Management Bahadurgarh, Haryana, India

Assistant Professor, Department of CSE, Sat Kabir Institute of Technology and Management Ladrawan, Bahadurgarh, Haryana, India

Assistant Professor, Department of CSE Sat Kabir Institute of Technology and Management Ladrawan, Bahadurgarh, Haryana, India

ABSTRACT: Web Spam is a big problem now a day. Every one want to get the necessary information as soon as possible after enter the keyword in the search engine. But the web spam is try to permutes different web pages for the profit or for the advertisement and due to this user not able to get the correct information in the required time.

In this paper, I use a new approach or technique thorough we can solve this problem. Here I use rating system and provide rate to the different web pages according to the user stay time and also link it to the email id of the particular user so the particular user can get the required information which he/she want to search.

KEYWORDS: Web page, Web Spam, Optimize Search Engine, Ranking

I. INTRODUCTION

There are different web pages which are available on the internet and these web pages are provided with the help of search engine and can search any information which we wants. When the user enter any information on the search engine then search engine provide the different web pages according to the ranking of the web page. There are wide area of the information which are search by the user such as net banking, online shopping, social networking etc. But some of the organization miss us it and want to show their own page for the intention of profit or for getting large number of viewer on their own page. These are the spammers who wants to want to divert the user for the promotion of their own web site or web page. Spammer uses high ranking keyword or high ranking keywords as links and provides these links to their own web page and this result in to the spammer's web page would shown on the top 10 to 15 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."

Internet is the fast growing field and today millions of users use internet and search different information from the wide variety of topic. There are lot of search engines but popular are Google, Alta vista, yahoo etc. These search engine try to find the best result for the user input keyword and show these best result at the top of the search engine instead of non-informative information. Search engine act as a bridge between the user and the web database. When a user enter any keyword in the search engine then thousand of result are shown with 10-12 links per page. These links are providing on the basics of their popularities. There are only few 2 to 3 pages are having essential information but the rest or the web pages are on web only for attracting high traffic to their own page and wants to get high hits to their own pages.

The spammer main purpose is to use various technique some are high ranking keywords, high ranking links, clocking spam so that these pages are shown in the first 10-12 links on the search engine and "Keyword Stuffing" is the most popular method or practice for creating web pages. These type of web pages uses memory and waste time of the user

DOI: 10.15680/IJIRCCE.2016. 0405288

and also they do not provide the required information. These type of activities may divert the user to different search engine because the user not able to get the required information and spent lot of time for the required information.

The main motive of the spammers is to misguide the search engine to show non informative information on the top level so the user have to click on these web pages. We have to find to remove these type of non informative pages but it is totally not possible by the human to specifies that the page is spam page or not. There are many techniques which are used to find the web page is spam page or not such as cloaking, link analysis, content analysis technique. If we use all these technique to detect the web page is spam page then we found that these techniques are still not efficient for the detection of the web page is spam page or not. All these techniques are using additionally space for storing page, ranking time, indexing of search engine.

Overview of our newly introduced approach:- In this paper, we introduced a new method or approach to find or detect the spam page and this technique is providing ranking to the web page. In this technique, we categories the pages into different categories such as web page is very bad, bad, medium, good and very good. When user enter into any web page then he/she gives rank to the given page and this page is link to the email id of the particular user and a link is send to the search engine for the verification of the web page. Here user provide their view about the particular web page from the available categories and then through this the user can get the right information page when the user search the same page.

When the user click on the verification then only the rank point are added to that page and it is impossible to any spammer to manipulate our result because this technique just focus on the user on the web page and when user leave then the point given to the web page is also stop and rank is allotted to that particular page.

Different techniques having different advantages but my technique main advantage is for the both search engine and the users. The benefit for the search engine is that the search engine provide the informative pages rather than non informative. The benefit for the user is that user get the necessary page result in the top once and get the information very quickly.

II. LITERATURE REVIEW

Web spam is not a new problem that a search engine faces. But actually is old as search engine. Today the web spam is a big problem not only for the users who not able to get the informative page but also for the search engine. There are different techniques which are used to detect the web spam and a lot of pagers had been surveyed. Through this lot of information was collected and we found that most of techniques uses the ranking method for the 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) Cloaking spam technique.

(2) Link Spam based technique.

(3 Content Spam based technique.

2.1.1 Clocking based spam

In this technique, clocking is user in this case and assign a rank to the page on the clock bases.It is used to getting hits to the pages even they are not correct or deserving it. In this case, after user enter their query in the search box then search engine delivering different content to the user which is used by the spammers. Clocking can be used with conjunction with different available techniques. By using client side scripting, some part of the web page is invisible to rewrite the web page after it had been delivered, then serve the user immediately the different page other then which the user is request to deliver.

2.1.2 Link Spam based technique

Another method or technique used to optimize search engine results or to detect web spam is Link spam. It can putting high ranking as a hyper link and divert user from the correct link. Some of the examples of the hyper links which are used by the web spammer are www.computer.com, www.twitter.com, www.onlinebanking.com,www.facebook.com,

DOI: 10.15680/IJIRCCE.2016. 0405288

www.Software.com etc. In this technique ,link based spam used search engine page rank technique which counts the number of links and also count the rank of the page.

2.1.3 Content based spam technique

In Content based spam, it uses the ranking technique, it assign rank to the keywords as content in the web page and putting high ranking keywords as content in web page. When a user search for a keyword then this page come into query result given by the search engine. There are so many keywords some of are ONLINE BANKING,COMPUTER,TWITTER,FACEBOOK,SOFTWARE etc. These some popular keywords are used by spammers as their content of the web page so the spam page can attract high traffic and even get high ranking and hits and this all makes a user fool and may divert user from the relevant result.

2.2 Different types of Spam Detection Technique

There are different algorithms or techniques which are used to optimize the search engine and detection of the web spam. But all these techniques are developed to help the user to get the appropriate result for their queries .Heuristic methods can be used for the identify or detect the web spam so to optimize the search engine. Some of the techniques are discussed below:-

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

(2) Ant Colony Optimization technique for web spam detection technique.

(3) Content based technique for Web spam detecting technique.

Before applying these techniques ,we firstly have to collect the information or date and for testing the web spam algorithms performance we have to collect the all processes. For the collection of the information or data we should have to consider the things which are given below:-

(1) The collection should be available for researchers collection should include many examples of spam and non-spam content.

(2) The collection should contains errors i.e. classification error.

(3) The collection contains different examples of spam free content or spam affected content.

(4) The collection should follows different web spam techniques as much as possible.

(5) The collection should be available for different researchers.

2.2.1 Ant colony optimization technique

As the name suggest that the working of this techniques is based on the behavior of ants .It is an optimizing technique for the detection of the web spam. Ant colony technique also used features of both content and link based techniques. Ant colony technique is basically widely used to identifies that the page is web spam page or not.

2.2.2 Content spam detection technique

In this technique, it is used to detect whether a web page is spam page or a normal web page There are so many words in the web page and also number of words in the web page title which are used for detecting the web page is a spam page or not. There are some very common words which are used such as “A”,“AN”,“THE”. These words are used nearly by all the web pages. If e found any page who does not contains these words then we consider the web page as a spam page. There is also a another method which works on the anchor text and detecting the spam pages.

III. PROPOSED WORK

In this paper, a new technique is used in which Timer is set which is used to calculate the time which a user take to stay in the particular page. The main purpose of this new technique is to save user time and user get the relevant result as fast as possible. Here we uses the ranking system, means the user we reaches the site then the timer starts working and timer will provide rank to different webpage depend on the stay time of the user on the webpage. Therefore, the webpage where the user stay for long time can get high rank and the page where user stay for less time can get less rank. Through this technique, the high rank web pages shows on the top of the search engine and the less rank webpage shows on the bottom. There are basically 10-15 links on a search engine and through this technique, user get the

DOI: 10.15680/IJIRCCE.2016. 0405288

appropriate web page on the top links and get the relevant information from that web page and save their time in finding the correct information. Therefore, an any time when user search that keyword, the user get the correct result.

To implement this technique, we have to used minimum i3 processor with minimum 2GB RAM and 320GB hard disk. The front end is the Visual Studio 2008 and backend is SQL Server 2008

3.1 Algorithm

The algorithm has been implemented using ASP.Net as a frontend and SQL Server 2008 as a backend. Here, I use a table to implement this technique: cloud provider table.

Cloud computing is using a network of remote servers to host on the Internet to store the data, manage the data, and process the data, rather than a local server or a personal computer. 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.

The algorithm works as follows:

In this initially we starts with zero rank or point to each page. Firstly the search engine searches according to the user entered keyword and then provide the user 10-15 links according to the rank of the web page. The rank also be updated according to the user need or requirements. There are many different condition or situation which are applied for the updating of the rank of any web page. These are the following steps which is need to follow are given below:-

(1) First of all, user have to enter a the keyword which hi'/she want to search in the search engine.

(2) After this, Firstly the search engine search the particular keyword on the title and then in the database. Here the database is made in SQL Server 2008

(3) Then, the search engine gives 10-15 link result. Search engine give the result in the combination of 10 link per page

(4) When the user click on any particular link then the new page is open in new tab.

(5) At the upper most part of the page there are 5 categories to choose. The categories are to assign the rank to the page by the help of 5 options such as Very Bad, Bad, Medium, Good, Very Good same as technique used in fuzzy logic.

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

(7) When user click on any raking option, then one text box will open which is for the user to enter their email id.

(8) After filling the user email id , a verification mail is send to the user email id and hence this is done because of the verification of the email id.

(9) After user verify the link then ranking point will be given and according to the user ranking option the point is giver to the page and the data for that particular user is stored in the database.

(10) Here, only one email id is valid for that day for one website link and hence with this we van rate or rank any web page and just for that particular day and this is due to avoid the misuse of the ranking system.

(11) Here we also that if any user click on any other tab then timer stop working and value is stored in the database and for every 10 seconds we assign 1 point or rank to the page.

3.2 Implementation Work

To implement this work, front end is the Visual Studio 2008 and backend is SQL Server 2008ASP.NET use for front end and SQL Server 2005. we have to used minimum i3 processor with minimum 2GB RAM and 320GB hard disk. The front end is the Visual Studio 2008 and backend is SQL Server 2008.

Table 1:- Cloud Provider Table

ID	TITLE	DETAIL	WEBSITE	POINT
1	RESTAURANT	RESTAURANT IS THE PLACE WHERE WE CAN EAT FOOD	www.resta.com	0
2	ANITA	ANITA IS A GIRL WHO LEARN ENGLISH	www.antgoogle.com	0
3	SEARCH ENGINE	SEARCH ENGINE IS USED TO SEARCH ANYTHING	www.searche.com	0
4	HUNUMAN	HANUMAN TEMPLE LOCATED ALLOVER INDIA	www.hunuman.com	0
5	SEARCH RESTAURANT	TECHNIQUE USED TO SEARCH RESTAURANT	www.restas.com	0

At the initial stage the rank field is zero for each row. First column tells the name of the field, second column tells the details about the title, the the 3rd shows the website of the particular title and the last column shows the point or rank. When a user search a keyword such as “RESTAURANT” then the keyword enter by the user is first search by the search engine in the title and then in the database ,here we use the SQL Server 2008 and then the result shows on the combination of 10 links. When a user click on any link ,the particular link will open in a separate new tab. After this timer starts and calculate the time and assign one point for every 10 second and then the point of the “RESTAURANT” increases from zero to one after 10 seconds and the rank is increases with increase in the stay time on that particular page. When the user searches same keyword again in the future, then the respective page shows at the top of the search result in the search engine. Here I also uses one or more concepts as if the user leave the page or any other place in the system or any non-client side then timer stop working and the value is then stored in the database of SQL server 2008. In previous techniques like content based which are focusing on the particular field i.e. the content which the user want and user get the correct web page or not on the basics of content

Below is the second table which show the different user email id and IP addresses. Table1 is the master table, used for website links and the table2 is the child table, used for.Here a second table is used for URL as reference key.

DOI: 10.15680/IJIRCCE.2016. 0405288

Table2:- Email id and IP address table

UID	URL	User_Emailid	User_IP	Date
1	www.xyz.com	xyz@xyz.com	167.124.16.1	13/12/2012
2	www.ahype.com	sdbc@xyz.com	167.124.16.2	13/12/2012
3	www.xsyz.com	sdfs@jii.com	167.124.16.2	13/12/2012
4	www.cabc.com	ccdf@fds.com	167.124.16.4	13/12/2012
5	www.shype.com	sadf@gsdf.com	167.124.16.1	13/12/2012
6	www.cabc.com	aiew@aglsg.com	167.124.16.6	13/12/2012

When the user click on any particular link then the new page is open in new tab. At the upper most part of the page there are 5 categories to choose. The categories are to assign the rank to the page by the help of 5 options such as Very Bad, Bad, Medium, Good, Very Good same as technique used in fuzzy logic and user can rank the point to any particular web page and explained in above algorithm and solve this problem through this technique.

IV. CONCLUSION

There are many techniques which are used to detect the web spam. There are different approaches for the different techniques to detect the web spam or to optimizes the search engine result. With the help of this technique we gives rank to different web pages and through this the informative page shows at the top in the search engine and help the user to get the correct or effective result of their query.

REFERENCES

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

[2] 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].

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

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

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

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

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

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

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

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