Web Spam Detection Study

Sajan Aggarwal
Department of Computer Science and Engineering, UIET, MDU,Rohtak, Haryana-INDIA
sajanmca.1985@gmail.com

Dr. Yogesh
Department of Computer Science and Engineering, UIET, MDU,Rohtak, Haryana-INDIA
yogs_crsce@yahoo.com

Gaurav
University Institute of Engineering and Technology, MDU,Rohtak, Haryana-INDIA
durejagaurav@gmail.com

ABSTRACT

Today all we are totally dependent on the Information provided by the search engine. A user searches any information on the search engine and follow top results to answer his/her queries. if he not able to find the desire information, instead of click on the next link, he changes the search query to get the desired result.

In World wide web, web spamplays a major role to change the result and provide unsolicited message at the top position. Basically, the Web Spam is the combination of two words i.e. Web and Spam. The first word WEB contains a major collections of web pages those are used by various search engine to answer the queries of the user who demand some information on the net and the second word SPAM means unrequested or unwelcomed messages which are basically not used for the user. The person who do the WEB SPAM activity is called SPAMMER. The main purpose of the spammer is either for earning high traffic or for profit.

The WEB SPAM is totally anillegal activity There are different types of techniques available for detecting web pages such as content based technique, Link based technique, Title based technique etc.

Keywords: Web, Spam, Cloud

I. INTRODUCTION

Today Internet plays a vital role in the human life. We are totally dependent on Internet for search any information and search engine plays role to search the required information from the net and give the result to the user who demand the information and the result shown by the search engine is about 8-10 links per page. And the result shown by the search engine is totally depending on the sites traffic. There are many ways to increase the site traffic. A simple way to increase site traffic is by providing relevant information on the website. This is a simple way but time consuming. Itconsumes lots of time to get the high site traffic so that the website can be come on the top links. On the other hand, shortest way to get high traffic is web spamming. The result of this is that the web site come on the first 4-5 links without informative information on the page. This is the work of web spammer. WEBSPAM is basically combination of two words i.e. WEB+ SPAM. The first word Web means a lot of pages which are used by various search engine to answer the query for the user and the second term SPAM means unwanted web pages or messages. Web spam means web pages which are created by the user with the intention to get high traffic so that his page can come on the 1stpage of the search engine query result. If any user searches any query on the search engine and instead of getting informative information, he/she get non-informative information on the top result.

According to Henzingeret 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.”

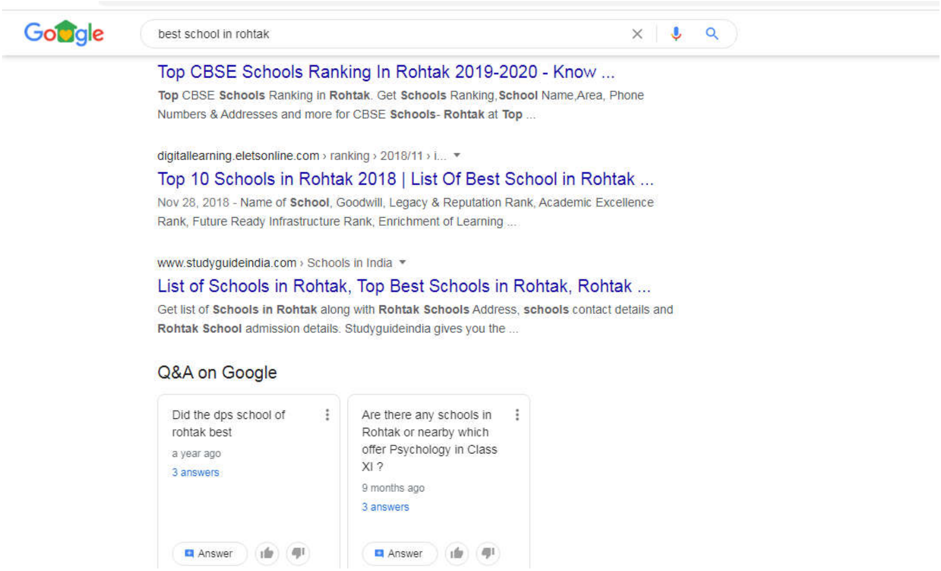
According to a definition presented by Gyongyi and Garcia, it refers to an activity performed by individuals to increase the rank of web page illegally[11]. Wu and et al. have introduced web spam as a behavior deceiving search engines [12].

The negative effect of increasing the number of pages spam in internet has been considered as crucial challenge for search engines [13]. It reduces the trust of users and search engine providers. Also, it wastes computing resources of search engines [14].

There are so many techniques already available which are used by search engine programmer to identify such pages and put them at the last of the search engine query but still web spammer are so much intelligent and uses different ideas to get his non informative pages at the top of the result. The main purpose of the spammer is for getting high traffic so that profit can be earn by his non informative websites.

Search engine should block these pages and put them at the last or remove from search engine query so that a user who required some information via search engine can get informative information instead of non-informative information. The benefit of these is for user is that user can trust on the search engine query result, time saving- instead of wasting of any informative information in non-informative result, user can get demanded information at the first page of the search engine result, resources saving- instead of wasting resources such as electricity, net connectivity or bandwidth, memory, if any user get information on the top then he can save CPU memory, bandwidth, RAM or much more. In short, we can say that search engine should use some major policy so that search engine can detect web spam pages or reduces its rank or delete its pages from the search engine query result. But this thing is only can possible if any search engine uses different policy and couple two or more policy to get the better result.

In this picture, if any user search best school in Rohtak in google search engine then instead of getting best result, other companies put their pages via web spamming on the first page or top result of the search engine query.



best school in rohtak

Top CBSE Schools Ranking In Rohtak 2019-2020 - Know ...

Top CBSE **Schools** Ranking in **Rohtak**. Get **Schools** Ranking,**School** Name,Area, Phone Numbers & Addresses and more for CBSE **Schools**- **Rohtak** at **Top** ...

digitallearning.eletsonline.com › ranking › 2018/11 › i...

Top 10 Schools in Rohtak 2018 | List Of Best School in Rohtak ...

Nov 28, 2018 - Name of **School**, Goodwill, Legacy & Reputation Rank, Academic Excellence Rank, Future Ready Infrastructure Rank, Enrichment of Learning ...

www.studyguideindia.com › Schools in India

List of Schools in Rohtak, Top Best Schools in Rohtak, Rohtak ...

Get list of **Schools in Rohtak** along with **Rohtak Schools** Address, **schools** contact details and **Rohtak School** admission details. Studyguideindia gives you the ...

Q&A on Google

Did the dps school of rohtak best

a year ago

3 answers

Are there any schools in Rohtak or nearby which offer Psychology in Class XI ?

9 months ago

3 answers

So main task of search engine should be identify that pages which are not required or not informative as per the user query and put them at the last of the result, and result of the that user can trust on the search engine query result and saves his time and get informative information.

II. LITERATURE REVIEW

Today we are totally dependent on the Internet for getting any information and Web spam is a grave issue for the search engine and the person who required information from the NET. The first word Web means a lot of pages which are used by various search engine to answer the queries for the users and the second term SPAM means unwanted web pages or messages. Web spam means web pages which are created by the user with the intention to get high traffic so that his page can come on the first page of the search engine query result. The main aim of the web spammer to get high volume to his website which not providing any informative information to the user.

Web Spam is old as commercial search engines[4].

Web spam is the result of using unethical methods to manipulate search results [1, 8, 9].

Perkins has defined web spam as follows: “The attempt to deceive algorithms related to search engines” [9].
There are so many techniques which are used by various search engine so that they can stop web spammer activity and can stand in front of them. But sorry to say, still web spammer plays a vital role and make fool to the search engine and to the user as well. Some of the techniques used by spammer is as under:

Classification of spam

Some of the followingtechniques is used by spammer for web spam to either for high ranking to their web sites or for profit to their pages are:-

(1) Content Spam based technique.

(2) Link Spam based technique.

(3) Cloaking spam technique.

CONTENT BASED SPAM TECHNIQUE

In this type of technique, spammer used high ranking keywords in their web page so that he can make fool to the search engine and get high volume traffic to their web page. In this technique spammers used currently high ranked keyword and place them in their webpage in different different areas. Some of the area are as under
BODY SPAM:- In this type the spammer used keyword in the body part of the HTML page of the web page. It is the very easiest way to put the popular keyword in the web page and attract the traffic to the web page.

META KEYWORD:- In this type the spammer used keywords in the Meta tag of the HTML page. The main advantage of this is the keywords which are used by the spammer in meta tag is not shown during open web page. These type of keywords used in meta tag is only for getting traffic but today search engine not using that tags as for answer the user query.

URL SPAM: In this type spammer used long URL for getting high traffic. Search engine breaks that URL and uses each word as a separate keyword

TITLE SPAM: Some spammers use this technique to get traffic on their web page. In this type the spammers used popular keyword in the title of the HTML code. Search engine reflect and uses each word match with the user query.

REPETITON OF WORDS: Some spammers uses this technique and in this technique spammer put popular keyword in different places so that he can fool search engine working. By using proper algorithm or programming we can detect this words.

LINK SPAM BASED TECHNIQUE

This technique is widely used by the spammer and in this type of spamming the spammer uses links as per user choice and when any users click on that link it reflect and send the user to their spamming pages such as if we a user want to book ticket from railway and link shown www.railwayticketbooking.com and user click on that link and it target page some other page instead of railway official page. Link-based web spam is manipulation of link formation to obtain the high rank. Some of them are mentioned as follows[10]:

Link farm: It is a group of web pages or websites which are connected to each and each web sites or web page have excessive links by creating link farms.

Link exchange: In This type the web sites owner help each other via adding their websites link to their web site so that each other can get advantage of both parties. This technique is called link exchange which are widely used by the spammer who want high volume to their websites

Buying the link: In This type the web sites owner buy some other websites for providing their link to their websites and for this the owner of the websites payes some rupees to the 2nd person

Many link based spam technique usesGoogle's page rank technique which count the number of links into a page and also count page rank of the referring page[4].

CLOCKING BASED SPAM

In this type of technique spammer uses a collection of techniques and provide different result to the search engine query and different result to the user which are non-informative pages and the main uses is only to make fool to the search engine and shown their pages at the top position so that user can click on it and they can get high traffic without providing information to the userClocking spam is basically used with content spam [4].

TECHNIQUE OF HANDLING WITH WEB SPAM

There are so many techniques by which we can stand in front of spammer and can easily detect their web spam activity. The result of the using technique is user can easily trust on search engine query result and search engine optimization increase. We can use more than one technique at the same time but should be keep in mind that we must uses these techniques in this way so that we can stand and detect web spam. Some of these techniques are as under

(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.

(4) Web spam detection using timer approach

The first step before applying the web spam detection techniques, is to collect the data and follow the collection process which is required for testing the web spam detection algorithms performance. For selectin 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.

CONTENT SPAM DETECTION TECHNIQUE

Content based techniques [1], number of the words in the web page, number of the words in the 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 theseindividual words are used a number of times. If any web page does not contain these common words then this page can be considered as spam. There is 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.

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 mainlyworking on the behavior of the ant for exploring web spam.

There islots of method for detecting 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 and powerful technique which is used for fighting with web spam.

WEB SPAM DETECTION USING TIMER

In this type of technique a timer is used by the search engine and if any user click on any page and found its non informative then as user close the page, the timer save the time spend by the user on a page and give indexing to the page as per time spend by the user. The main benefit of this approach is that if any user sees the page is non informative then he closes within 5 to 10 seconds and the timer attached with search engine saves the time and give indexing to the page

III.CONCLUSION

In this paper, we noticed that spammer uses each and every activity to fool the search engine and result of this is high traffic to their non informative web pages. web spam is a major challenge and we can see that there are various methods or techniques available in the market but still there is a need of more advancement of technique so that we can stand in the front of web spammer and put their pages either at the last of search engine query result or after detecting that these pages are web spam with no informative information available on the pages, completely deleted

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]. ArnonRungsawang, ApichatTawesiriwate, BunditManaskasemsak, “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]. SumitSahu, 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, KarolyCsalogany, 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].

[12]. Gyongyi, Z. and H. Garcia-Molina, Web Spam Taxonomy, in First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb 2005). 2005: Chiba, Japan.

[13]. Wu, B., V. Goel, and B.D. Davison. Topical trustrank: Using topicality to combat web spam. in Proceedings of the 15th international conference on World Wide Web. 2006. ACM.

[14]. Gyngyi, Z. and H. Garcia-Molina, Link spam alliances, in Proceedings of the 31st international conference on Very large data bases. 2005, VLDB Endowment: Trondheim, Norway. p. 517-528.

[15]. Abernethy, J., O. Chapelle, and C. Castillo, WITCH: A New Approach to Web Spam Detection, in In Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web (AIRWeb}. 2008.

[16]. Najork, M., Web Spam Detection. Encyclopedia of Database Systems, 2009. 1: p. 3520-3523.

[17]. Castillo, C., et al., A reference collection for web spam. SIGIR Forum, 2006. 40(2): p. 11-24.

[18]. Davison, B.D., Recognizing nepotistic links on the web. Artificial Intelligence for Web Search, 2000: p. 23-28.

[19]. Collins, G. Latest search engine spam techniques. Aug 2004; Available from: http://www.sitepoint.com/article/search-enginespam-techniques.

[20]. Perkins, A. The classification of search engine spam. 2001; Available from: http://www.silverdisc.co.uk/articles/spamclassification.

[21]. Sasikala, S. and S.K. Jayanthi. Hyperlink Structure Attribute Analysis for Detecting Link Spamdexing. in International Conference on Advances in Computer Science–(AET-ACS 2010), Kerela. 2010.

[22]. Wu, B. and B.D. Davison. Cloaking and Redirection: A Preliminary Study. in AIRWeb. 2005.

[23]. Fetterly, D., M. Manasse, and M. Najork. Spam, damn spam, and statistics: Using statistical analysis to locate spam web pages. in Proceedings of the 7th International Workshop on the Web and Databases: colocated with ACM SIGMOD/PODS 2004. 2004. ACM.

[24]. Ntoulas, A., et al. Detecting spam web pages through content analysis. in the 15th International World Wide Web Conference. May 2006. Edinburgh, Scotland.

[25]. Amitay, E., et al. The connectivity sonar: Detecting site functionality by structural patterns. in the 14th ACM Conference on Hypertext and Hypermedia. Aug 2003. Nottingham, UK.

[26]. Prieto, V., et al., Analysis and Detection of Web Spam by Means of Web Content, in Multidisciplinary Information Retrieval, M. Salampasis and B. Larsen, Editors. 2012, Springer Berlin Heidelberg. p. 43-57.

[27]. Karimpour, J., A. Noroozi, and S. Alizadeh, Web Spam Detection by Learning from Small Labeled Samples. International Journal of Computer Applications, 2012. 50(21): p. 1-5.

[28]. Rungsawang, A., A. Taweesiriwate, and B. Manaskasemsak, Spam Host Detection Using Ant Colony Optimization, in IT Convergence and Services, J.J. Park, et al., Editors. 2011, Springer Netherlands. p. 13-21.

[29]. Silva, R.M., A. Yamakami, and T.A. Alimeida. An Analysis of Machine Learning Methods for Spam Host Detection. in 11th International Conference on Machine Learning and Applications (ICMLA). 2012.

[30]. Tian, Y., G.M. Weiss, and Q. Ma. A semisupervised approach for web spam detection using combinatorial feature-fusion. in GRAPH LABELLING WORKSHOP AND WEB SPAM CHALLENGE. 2007.

[31]. Becchetti, L., et al. Link-Based Characterization and Detection of Web Spam. in AIRWeb 2006. 2006. Seattle, Washington, USA.

[32]. Castillo, C., et al., Know your neighbors: web spam detection using the web topology, in Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. 2007, ACM: Amsterdam, The Netherlands. p. 423-430.

[33]. Dai, N., B.D. Davison, and X. Qi, Looking into the past to better classify web spam, in Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web. 2009, ACM: Madrid, Spain. p. 1-8.

[34]. Page, L., et al., The PageRank citation ranking: bringing order to the web. 1999. [24] Kleinberg, J.M., Authoritative sources in a hyperlinked environment. Journal of the ACM (JACM), 1999. 46(5): p. 604-632.

[35]. Bharat, K. and M.R. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. in Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. 1998. ACM.

[36]. Zhang, L., et al. Exploring both content and link quality for anti-spamming. in Computer and Information Technology, 2006. CIT'06. The Sixth IEEE International Conference on. 2006. IEEE.

[37]. Acharya, A., et al., Information retrieval based on historical data. 2008, Google Patents.

[38]. Eiron, N., K.S. McCurley, and J.A. Tomlin, Ranking the web frontier, in Proceedings of the 13th international conference on World Wide Web. 2004, ACM: New York, NY, USA. p. 309-318.

[39]. Lempel, R. and S. Moran, The stochastic approach for link-structure analysis (SALSA) and the TKC effect. Computer Networks, 2000. 33(1): p. 387- 401.

[40]. Ng, A.Y., A.X. Zheng, and M.I. Jordan. Stable algorithms for link analysis. in Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. 2001. ACM.

[41]. Zhang, H., et al., Making eigenvector-based reputation systems robust to collusion, in Algorithms and Models for the Web-Graph. 2004, Springer. p. 92-104.

[42]. Li, L., Y. Shang, and W. Zhang. Improvement of HITS-based algorithms on web documents. in Proceedings of the 11th international conference on World Wide Web. 2002. ACM.

[43]. Chakrabarti, S., M. Joshi, and V. Tawde, Enhanced topic distillation using text, markup tags, and hyperlinks, in Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. 2001, ACM: New Orleans, Louisiana, USA. p. 208-216.

[44]. Gyongyi, Z., et al., Link spam detection based on mass estimation, in Proceedings of the 32nd international conference on Very large data bases. 2006, VLDB Endowment: Seoul, Korea. p. 439- 450.

[45]. Sobek, M., Pr0-google's pagerank 0 penalty. badrank. 2002.

[46]. Guha, R., et al., Propagation of trust and distrust, in Proceedings of the 13th international conference on World Wide Web. 2004, ACM: New York, NY, USA. p. 403-412.

[47]. Krishnan, V. and R. Raj. Web spam detection with anti-trust rank. in the 2nd International Workshop on Adversarial Information Retrieval on the Web (AIRWeb 2006). 2006. Seattle, USA.

[48]. Benczur, A.A., et al. SpamRank–Fully Automatic Link Spam Detection Work in progress. in Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web. 2005.

[49]. Becchetti, L., et al. Using rank propagation and probabilistic counting for link-based spam detection. in Proc. of WebKDD. 2006.

[50]. Wu, B. and B.D. Davison. Identifying link farm spam pages. in Special interest tracks and posters of the 14th international conference on World Wide Web. 2005. ACM.