Survey on Internet Spam: Classification and Analysis

Article *in* International Journal of Computer Applications in Technology · May 2013

CITATIONS
5

READS
43

3 authors, including:

Geerthik S
SSCE
2 PUBLICATIONS **5** CITATIONS

SEE PROFILE

ISSN:2229-6093

Geerthik S , Int.J.Computer Technology & Applications,Vol 4 (3),384-391

Survey on Internet Spam: Classification and Analysis

Geerthik.S
Asst.professor, Sree Sastha College of Engineering, Chennai
geerthiks@gmail.com

Abstract

In recent years spam detection and dealing with spam in information retrieval systems is a difficult task. This article surveys the classification of various spam in the internet based on their properties. The impact of various spams in social networks, email, image, content and links is discussed and the technique applied to prevent the spam in various areas is listed. A detailed analyzes of cloaking and redirection in web search is also given in this paper. Finally this article listed the things to be considered to construct the spam algorithms.

Keywords: email spam, content spam, spam filters, link spam.

1. Introduction.

Spam is an endless repetition of worthless text or image. Spam can spread out in any information systems like emails, web, social network sites, and blogs or in review platforms. The concept of web spam was introduced in 1996 [1] and it soon become key challenges for search engine industry [2].Nowadays the major search engine companies have identified adversarial information retrieval [14] as top priority because of multiple negative effects caused by spam, and also the appearance of new challenges in the field of research. First spam spoils the quality of research and prevents the legitimate websites of revenue that might earn in the absence of spam. Second it weakens the trust of user in a search engine provider which is a notable issue since the user can easily continue his search form one search engine to other.

In the websites the spam spreads by adult content distribution and malware spreading and also leads for phishing. For example [15] ranked 100 million webpage's using page rank algorithm [5] and found that out of that 20 results were pornographic websites that achieves high ranking by content and web link manipulation. Web spam phenomenon occurs mainly by the following fact. The fraction of webpage referrals that come from search engines is very high and also the users examine only top ranked results. Thus [23] showed that 85% of queries only the first result page is clicked and only 3 to 5 links are requested [24].So the website owners attempt to manipulate search engine rankings. The manipulation is done in different forms like undesired link creation, cloaking, click fraud and tag spam. [32] Shows that 6% of English language websites are classified as spam. [4] Reports that 22% of spam are found on host level and [25] states that it is 16.5%. The goal of this survey is to give ideas about various spam in the internet and the algorithms used in spam detection and also to create awareness in further research in area of adversarial information retrieval.

2. Classification of Spam

In this chapter we analyze various spam and the ways they can be prevented.With the recent researches and the impact of spam in various areas in internet, the spam classification is given in the fig.1.

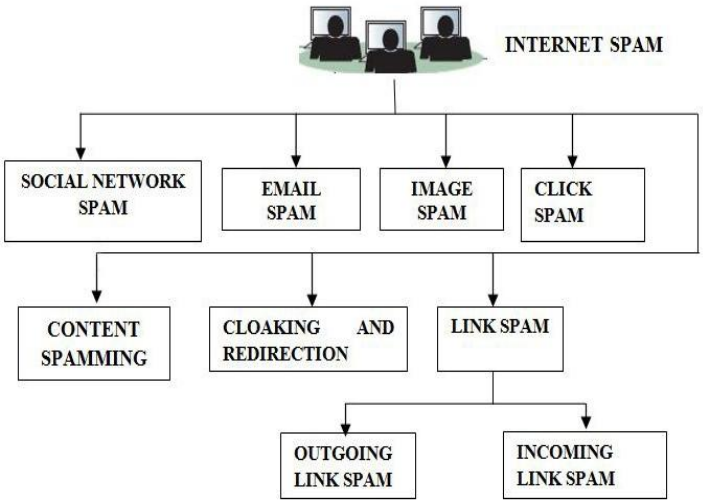

Fig.1. Classification of internet spam

2.1 Social network spam

In past few years the development of social networking sites is very high. The people communicate with their friends and chat or share multimedia contents with them. Sites like facebook, twitter are constantly among top 20 most viewed websites on the internet [36]. Statistics shows that average people spent more time on social network compared with other sites. The increase in popularity of social networks allows them to collect a huge amount of personal information about the users, their friends, habits and also their wealth information. In social network a person can reach any person which is attracted by the malicious parties. In 2008, 83% of users received minimum one unknown friend request or message [37]. Unfortunately social networking sites do not provide strong authentication mechanisms, and It is easy to act like a person and sneak into person's network of trust [16].

The explosive growth of unsolicited emails has prompted the development of numerous spam filtering techniques.[17] showed how it would be possible for spammers to craft targeted spam by leveraging the information available in online social networks. As for Twitter, [18] ran an experiment on Twitter spam. It created a popular hash tag on Twitter, and observed that spammers started using it in their messages. It also discuss some features that might allow one to distinguish a spammer from legitimate users, such as node degree and frequency of messages. Spam detection and the behavior of users have been studied for a long time. Bayesian classification algorithm is applied to distinguish the suspicious behaviors from normal ones The analyze of data set and performance of social network twitter is analyzed in [26] and the result is by Bayesian classifier is the best algorithm in finding spam email filters are widely implemented both on modern email clients and servers.

A Bayesian spam filter is superior to a static keyword based spam filter because it can continuously evolve to tackle new spam by learning keywords in new spam emails. However, Bayesian spam filters can be easily poisoned by avoiding spam keywords and adding many innocuous keywords in the emails. In addition, they need a significant amount of time to adapt to a new spam based on user feedback. Moreover, few current spam filters exploit social networks to assist spam detection. Regarding the drawbacks in Bayesian spam filter an user-friendly spam filter called Social network Aided Personalized and effective spam filter (SOAP) [27] is used. Unlike previous filters that focus on parsing keywords (e.g., Bayesian filter), SOAP exploits the social relationship among email correspondents to detect the spam adaptively and automatically. SOAP integrates three components into basic Bayesian filter. social closeness spam filtering, social interest-based spam filtering, and adaptive trust management. Experimental results show that SOAP can greatly improve the performance of Bayesian spam filters in terms of the accuracy, attack-resilience and efficiency of spam detection

2.2 Email spam

The most common communication in the internet is using email communication. With the vast growth in email and its popularity unsolicited e-mail (spam) also emerged very quickly with almost 90% of all email messages. i.e., over 120 billion of these messages are sent each day [25]. The cost of sending these e-mails is very close to zero being easy to reach a high number of potential consumers [13]. In this context, spam consumes resources; time spent reading unwanted messages, bandwidth, CPU, disk, being also used to spread malicious content.

The email system design can easily be exploited by spammers who send inaccurate information. All email on the Internet is sent via a protocol called Simple Mail Transfer Protocol ("SMTP").SMTP is designed to capture information about the route that an email message travels from its sender to its recipient. In actuality, the SMTP protocol provides no security, email is not private, it can be altered en route, and there is no way to validate the identity of the email source. In other words, when a user receives an email Message, there is no way to tell who sent the email and who has seen it. The lack of security in SMTP, and specifically the lack of reliable information identifying the email source, is regularly exploited by spammers and allows for considerable fraud on the Internet (such as identity theft or "phishing").

For handling email spam [6] introduced a novel hybrid model, Partitioned Logistic Regression, which has several advantages over both naive Bayes and logistic regression. This model separates the original feature space into several disjoint feature groups. [7] Proposed a decentralized privacy preserving approach to spam filtering. The solution exploits robust digests to identify messages that are a slight variation of one another and a structured peer-to-peer architecture between mail servers to collaboratively share knowledge about spam.[21] developed an ant-colony based spam filter to evaluate and predict spam messages. The developed spam filter is compared with other popular techniques like multi-layer perception, naive Bayes and Ripper classifiers the developed filter is alternate tool in predicting the spam and also yield better accuracy .[22] developed a rule based filter for light weight and accurate detection of email spam. This filter cascade three filters, one for fingerprints of message bodies, another for white and black list of email address in form of header and last for the words specific to the spam and legitimate email in message header. The method has high performance of about 90 emails per seconds.

2.3 Image spam

Recently, spammers have proliferated "image spam", emails which contain the text of the spam message in a human readable image instead of the message body. It consists in embedding the spam message into images which are sent as email attachments. Its goal is to circumvent the Analysis of the emails' textual content performed by spam filters, including automatic text classifiers. Since attached images are displayed by default by most email clients, the message is directly conveyed to the user as soon as the email is opened. The simplest kind of image spam can be viewed as a screen shot of a plain text written using a standard text editor. An example of image spam is given in fig.2.

Making detection to image spam by conventional content filters is very difficult. New techniques are needed to filter these messages. Often spam images are constructed by introducing random changes to a given template image, to make signature-based detection techniques ineffective, and are obfuscated to prevent optical character recognition (OCR) tools from reading the embedded text.

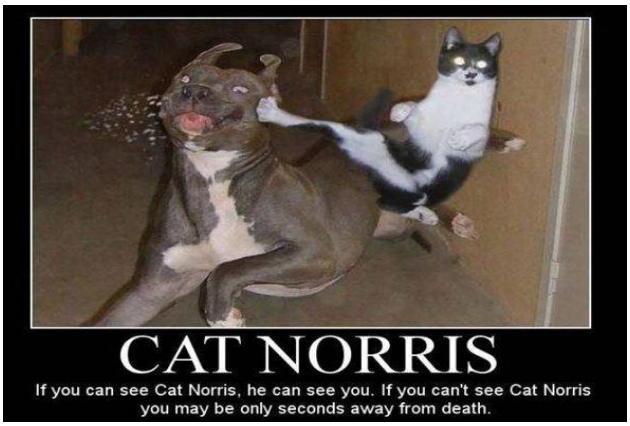

Fig 2.An real image spam received in email

Ironically, some text obfuscation techniques used against OCR tools are very similar to the ones exploited to design CAPTCHAs (Completely Automated Public Turing test to tell Computers and Humans Apart). A common type of CAPTCHA requires the user to type letters and/or digits from a distorted image that appears on the screen. Such tests are commonly used to prevent unwanted internet bots from accessing websites, since a normal human can easily read a CAPTCHA, while the bot cannot process the image letters and therefore, cannot answer properly, or at all. [19] Introduces a technique in which presented a comprehensive solution to image spam filtering, which combines cluster analysis of spam images on the server side and active learning classification on the client side for effectively filtering image spam.

Extensive experimental evaluations of both server-side algorithms and client-side algorithms on a real image spam dataset collected from an e-mail server demonstrated the efficiency of this method. [20] Proposed a machine learning method to detect the spam images from images in normal emails. The proposed method extracts efficient global image features to train an advanced binary classifier to distinguish the spam images, which achieves promising preliminary results on the limited sample database. The proposed method achieves fairly good performance in 5-fold cross-validation on the data set with 928 spam images and 810 natural images.

2.4 Click spam.

Here the spammers generate fraud clicks and make the control function towards their websites. To achieve the goal spammers submit queries to search engine and click on the links point to the target pages [28, 34]. Online advertising is other incentive for spammers to generate fraudulent clicks [30].

2.5. Content Spamming

Content spamming involves changing the logical view that the search engine has over the page contents. An example of content spamming is keyword stuffing which involves placement of keywords within the webpage to raise the keyword count. When we search video songs hd on google and click the link filetube as the result of google search it is directed to the page given in the fig.3.the page contains only add and it does not contain the search result is an example of content spamming.

Spam target pages are thus stuffed with a large number of keywords that are either of high advertisement value or highly spammed, including misspelled popular words such as “googel” or “accomodation” as seen among the top hits of a major search engine in Fig. 3.The content spamming involves title spamming, Body tag spamming or Meta tag spamming. With the advent of link based ranking algorithms [5,68] content spam phenomena’s has partially over come. However spam is constantly evolving and soon afterwards spammers started creating link farms [31, 32] .

Data mining techniques have been already applied to spam detection problems [38]. Content-based spam detection is the automation of recognition of spammers message as spam and it works based on discovering the patterns in the message and by classifying a message.

[39] Introduce a novel content-based spam detection approach, which first uncovers patterns hidden in both spam and legitimate messages, and then it associates the discovered patterns with the corresponding class .Not only it uses words and symbols for classification of message, but it uses the combination of them. This approach is efficient in terms of computational complexity, being able to classify more than one hundred messages per second.

Fig .3 A page with no content other than ads.

2.6 Cloaking and Redirection

Cloaking is a search engine optimization technique in which the content presented in the search engine spider is different from that is presented in the user browser. Redirection is used to send users automatically to another URL after loading the needed current URL. Both Cloaking and redirection are used in the search engine spamming [33]. To distinguish users from the crawlers, spammers analyze a user field of HTTP request and keep track of IP address used by search engine crawlers. The other way to redirect user to malicious pages by executing Java script activated by page OnLoad() event or timer. The thing to be noticed in Java script is the most wide spread and difficult to detect by crawlers, since most crawlers are script agnostic [34].Cloaking is very hard to detect; the method is described in [8] which aid cloaking detection method by using the most frequent words from the MSN query log and highest revenue generating words from the MSN advertisement log. In theory cloaking could be detected by comparing crawls with different user agent strings and IP addresses of the robots. Spammers however tackle robot behavior, collect and share crawler IP addresses and hence very effectively distinguish robots from human surfers. Spam may also circumvent a post-processing Web spam filter since the crawler may believe the spam site to be honest based on outdated information. The widespread use of parking domains for spamming purpose illustrates this phenomenon. Spammers purchase sites that terminate their operation and fill them with spam. For some additional time these sites appear with their previous content both in the search engine index and also in the input for the spam classificatory. The crawler will meanwhile fetch the new Content believed to be honest, follows its links and prioritizes its processes in favor for the spammer's target. The HTML code excerpt in Fig. 4 shows the use of a parking domain for spamming, in combination with hiding content from human users by using style sheets.

```
<div style="position:absolute; top:20px; width:600px;
height:90px; overflow:hidden;">
<font size=-1>atangledweb.co.uk currently offline<br>
atangledweb.co.uk back soon<br></font><br><br>
<a href="http://www.atangledweb.co.uk"><font size=-
1>atangledweb.co.uk</font></a><br><br><br>
Soundbridge HomeMusic WiFi Media Play<a class=l
href="http://www.atangledweb.co.uk/index01.html">-
</a>>...
SanDisk Sansa e250 - 2GB MP3 Player -<a class=l
href="http://www.atangledweb.co.uk/index02.html">-
</a>>...
AIGO F820+ 1GB Beach inspired MP3 Pla<a class=l
href="http://www.atangledweb.co.uk/index03.html">-
</a>>...
Targus I-Pod Mini Sound Enhancer<a class=l
href="http://www.atangledweb.co.uk/index04.html">-
</a>>...
```

Sony NWA806FP.CE7 4GB video WALKMAN ->...

Fig.4 The use of a parking domain to impute spam pages into a Web crawl.

2.7 Link spam

Link spam is also called as comment spam or blog spam. It is the form of spamming which targets weblogs, guestbook's, discussion boards and also wikis (wiki spam). Any web application that displays hyperlinks submitted by visitors or the referring URL's of the web visitors may be the target. Link spamming occur in the internet guestbook's where spammers repeatedly fill a guest book with link to their own site and no relevant to increase their search engine rankings. The two types of link spam are outgoing link spam and incoming link spam.

Outgoing link spam is the easiest and less cost link spam. Here the server has direct access to his pages and therefore add any items to them and create a large set of authoritative links.

In incoming link spam the spammer simply increases the number of incoming links to his target page. The easiest way to redirect to the target page is set page refresh time to zero and initialize the refresh URL attribute with a URL of target page. Another approach is to use page level scripts that aren't usually executed by crawlers and it is more effective from spammer's point of view.

Recently several results have appeared that apply rank propagation to extend initial trust or distrust judgments over a small set of seed pages or sites to the entire web, such as trust [9], distrust[10] propagation in the neighborhood or their combination [35] as well as graph based similarity measures [11].Recent researches are detecting the link spam using temporal information and with genetic programming features.

3. Things to be considered

Having analyzed all the related work devoted to the topic of web spam mining, we identify a set of underlying principles that are frequently used for algorithms construction.

•The filter designed for spam should only filter the spam and not any legitimate message.

• Spammers mostly target popular queries and queries with high advertising value.

• Spam pages deviate from power law distributions based on numerous web graph statistics such as Page Rank or number of in-links.

• Spammers build their link farms with the aim to boost ranking as high as possible, and therefore link farms have specific topologies that can be theoretically analyzed on optimality.

• According to experiments, the principle of approximate isolation of good pages takes place: good pages mostly link to good pages, while bad pages link either to good pages or a few selected spam target pages.

•It has also been observed that connected pages have some level of semantic similarity – topical locality of the Web, and therefore label smoothing using the Web graph is a useful strategy.

• Numerous algorithms use the idea of trust and distrust propagation using various similarity measures, propagation strategies and seed selection heuristics.

• Because one spammer can have a lot of pages under one website and use them all to boost ranking of some target pages, it makes sense to analyze host graph or even perform clustering and consider clusters as a logical unit of link support.

• In addition to traditional page content and links, there are a lot of other sources of information such as user behavior or HTTP requests. There is hope that more will be developed in the near future. Clever feature

engineering is especially important for web spam detection.

• Despite the fact that new and sophisticated features can boost the state-of-the-art further, proper selection and training of a machine learning models is also of high importance.

4. Conclusion and Future Work

The different spam which affects the internet is classified and the techniques used to fight against the spam is surveyed in this paper. The problems caused by different spam in the websites and the solutions available to filter the spam is also discussed .For filtering spam in social networks advanced Bayesian filter technique SOAP , for filtering email spam a rule based filter using data mining concepts is suggested. Having analyzed all the types of spam in the internet the things to consider for designing effective spam filter is also listed.

Further work involves producing effective spam filters with multiple spam filtering in a single filter, since spammers are aggressive and new spam are originating with the development of internet. Further work also includes creating algorithm for tracing spammer location and makes him to pay for his activities in internet.

References

[1]. Eric Convey ."Porn sneaks way back on Web", The Boston Herald, May 22, 1996

[2]. M.R.Henzinger, R.Motwani, and C. Silverstein."Challenges in web search engines", SIGIR Forum, 36, 2002
.
[3].L.Becchetti,C.Castillo,D.Donato,S.Leonardi, and R.Baeza-Yates."Link-based characterization and detection of web spam", In Proceedings of the Second International Workshop on Adversarial Information Retrieval on the Web, AIRWeb'06, Seattle, USA, 2006.

[4] C.Castillo, D.Donato, A.Gionis, V. Murdock, and F.Silvestri. "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, SIGIR'07, Amsterdam, The Netherlands, 2007.

[5]. L.Page, S. Brin, R. Motwani, and T. Winograd. "The page rank citation ranking: Bringing order to the web",1998.

[6]. M.-T.Chang, W.-T.Yih, and C. Meek."Partitioned Logistic Regression for Spam Filtering" Proc. 14th ACM SIGKDD Int'l Conf.Knowledge Discovery and Data mining (KDD), pp. 97-105, 2008.

[7]. E.Damiani, S.D.C.Vimercati, S. Paraboschi, and P.Samarati. "P2PBased Collaborative Spam Detection and Filtering" Proc. Fourth IEEE Int'l Conf. Peer-to-Peer Computing, pp. 176-183, 2004

[8]. K.Chellapilla and D.M.Chickering. "Improving cloaking detection using search query popularity and monetizability."In Proceedings of the 2nd International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), pages 17–24, Seattle, WA, August 2006.

[9]. B. Wu, V. Goel, and B. D. Davison. Topical TrustRank. " Using topicality to combat web spam." ,In Proceedings of the 15th International World Wide Web Conference (WWW), Edinburgh, Scotland, 2006.

[10]. Drost and T.Scheffer. "Thwarting the nigritude ultramarine:Learning to identify link spam." In Proceedings of the 16th European Conference on Machine Learning (ECML), volume 3720 of Lecture Notes in Artificial Intelligence, pages 233–243, Porto, Portugal, 2005

[11].A.Benczúr, K.Csalogány, and T.Sarlós. "Link-based similarity search to fight web spam." In Proceedings of the 2nd International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), held in conjunction with SIGIR2006, 2006

[12].C.Kanich, Spamalytics."An Empirical Analysis of Spam Marketing Conversion", in Computer and Communications Security Conference (CCS08). ACM, pp. 27-31, 2008

[13].V.Cheng and C.Li, "Personalized Spam Filtering with Semisupervised Classifier Ensemble", in IEEE/WIC/ACM International Conference on Web Intelligence, 2006.

[14].D.Fetterly."Adversarial Information Retrieval: The Manipulation of Web Content", 2007.

[15].N.Eiron, K. S. McCurley, and J. A. Tomlin. "Ranking the web frontier",In Proceedings of the 13th International Conference on World Wide Web, WWW'04,New York, NY, 2004.

[16].S.Moyer and N.Hamiel ."Satan is on my friend list: Attacking social networks".http://www.blackhat.com/html/bh-usa-08/bh-usa-08-archieve.html,2008

[17].G. Brown, T. Howe, M. Ihbe, A. Prakash, and K. Borders. "Social networks and context-aware spam". In ACM Conference on Supportive Cooperative Work, 2008

[18].B. Krishnamurthy, P. Gill, and M. Aritt. "A few chirps about twitter". In USENIX Workshop on Online Social Networks, 2008

[19].Gao, Yan .“ A Comprehensive Approach to Image Spam Detection: From Server to Client Solution, Information Forensics and Security.”, IEEE Transactions , Dec. 2010

[20].Gao, Yan “Image spam hunter, Acoustics, Speech and Signal Processing.”, 2008. ICASSP 2008. IEEE International Conference on April 2008

[21].El-Alfy, El-Sayed M. “Discovering classification rules for email spam filtering with an ant colony optimization algorithm.” CEC '09 ,IEEE Conference Publications ,May 2009

[22].Takesue, Masaru . “Cascaded Simple Filters for Accurate and Lightweight Email-Spam Detection .”, Emerging Security Information Systems and Technologies (SECURWARE), Fourth International Conference , July 2010

[23].C.Silverstein, H.Marais, M. Henzinger, and M. Moricz.” Analysis of a very large web search engine query log” SIGIR Forum, 33, Sept. 1999.

[24].T.Joachims,L.Granka,B.Pan,H.Hembrooke, and G.Gay.“Accurately interpreting click through data as implicit feedback”, In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR’05, Salvador, Brazil, 2005.

[25]. A. A. Bencz´ur, K. Csalog´any, T. Sarl´os, and M.Uher. Spamrank. “Fully automatic link spam detection work in progress”, In Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web, AIRWeb’05, May 2005.

[26].Wang, Alex Hai .“ DON’T FOLLOW ME: SPAM DETECTION IN TWITTER”, Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on , July 2010

[27]. Li, Ze , Shen, Haiying Helen. “SOAP: A Social network Aided Personalized and effective spam filter to clean your e-mail box.” INFOCOM, 2011 Proceedings IEEE Conference Publications , April 2011.

[28]. F. Radlinski. “Addressing malicious noise in click through data.” ,In Proceedings of the 3rd international workshop on Adversarial information retrieval on the web, AIRWeb’07, Canada, 2007.

[29]. Z. Dou, R. Song, X. Yuan, and J.-R. Wen. “Are click through data adequate for learning web search rankings?.”In Proceedings of the 17th ACM conference on Information and knowledge management, CIKM, 2008.

[30]. N. Immorlica, K. Jain, M. Mahdian, and K. Talwar. “Click Fraud Resistant Methods for Learning Click- Through Rates.” Technical report, Microsoft Research, Redmond, 2006.

[31]. Z. Gy¨ongyi, H. Garcia-Molina. “ Link spam alliances.”, In Proceedings of the 31st International Conference on Very Large Data Bases, VLDB’05, Trondheim, Norway, 2005.

[32]. S. Adali, T. Liu, and M. Magdon-Ismail. “ Optimal Link Bombs are Uncoordinated.” In Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web, AIRWeb’05, Chiba, Japan, 2005.

[33].Z.Gyongyi, H.Garcia-Molina. “ Web Spam technology.”,In First International Workshop on Adversarial Information Retrieval on the Web(AIR Web),2005

[34].K.Chellapilla,A.Maykov.“A taxonomy of javascript redirection spam.”,In Proceedings of the 3rd international workshop on Adversarial information retrieval on the web, AIRWeb’07, Banff, Canada, 2007.

[35].B.Wu, V.Goel, and B. D. Davison. “Propagating trust and distrust to demote web spam”, In Workshop on Models of Trust for the Web, Edinburgh, Scotland, 2006

[36]. http://www.alexa.com/topsites/global

[37]. Harris Interactive Public relation research .A study of social networks scams, 2008.

[38]. T.Fawcet. “A challenge problem for data mining. KDD Explorations”, in vivo’ spam filtering, December 2003.

[39].Veloso,Adriano.“Lazy Associative Classification for Content-based Spam Detection”, Web Congress, 2006,IEEE conference publication, Oct. 2006