International Journal of Computer Applications Technology and Research
Volume 3– Issue 3, 180 - 185, 2014, ISSN: 2319–8656

A Survey of Web Spam Detection Techniques

Mahdieh Danandeh Oskuie
Department of Computer, Shabestar Branch,
Islamic Azad University,
Shabestar, Iran

Seyed Naser Razavi
Computer Engineering Department, Faulty of
Electrical and Computer Engineering,
University of Tabriz, Iran

Abstract: Internet is a global information system. Most of the users use search engines due to high volume of information in virtual world in order to access to required information. They often observe the results of first pages in search engines. If they cannot obtain desired result, then they exchange query statement. Search engines try to place the best results in the first links of results on the basis of user's query.

Web spam is an illegal and unethical method to increase the rank of internet pages by deceiving the algorithms of search engines. It involves commercial, political and economic applications. In this paper, we firstly present some definitions in terms of web spam. Then we explain different kinds of web spam, and we describe some method, used to combat with this difficulty.

Keywords: HITS; Machine learning; PageRank; Search Engine; Web pam.

1. INTRODUCTION

Nowadays, with regard to increasing information in web, search engines are considered as a tool to enter the web. They present a list of results related to user query. A legal way to increase sites rank in the list results of search engines is increasing the quality of sites pages, but this method is time-consuming and costly. Another method is use illegal and unethical methods to increase the rank in search engines. The effort of deceiving search engines is called web spam.

Web spam has been considered as one of the common problems in search engines, and it has been proposed when search engines appeared for the first time. The aim of web spam is to change the page rank in query results. In this way, it is placed in a rank higher than normal conditions, and it is preferably placed among 10 top sites of query results in various queries.

Web spam decreases the quality search results, and in this way it wastes users' time. When the number of these pages increases, the number of pages investigated by crawlers and sorted by indexers increases. In this case, the resources of search engines are lost, and the time of searching in response to user query increases.

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[1]. Wu and et al. have introduced web spam as a behavior deceiving search engines [2].

The successes that have been achieved in terms of web spam decrease the quality of search engines, and spam pages are substituted for those pages whose ranks have increased by using legal method. The negative effect of increasing the number of pages spam in internet has been considered as crucial challenge for search engines [3]. It reduces the trust of users and search engine providers. Also, it wastes computing resources of search engines [4]. Therefore, if an effective solution is presented to detect it, then search results will be improved, and users will be satisfied in this way.

Combatting with web spam involves web spamming detection and reducing its rank while ranking or its detection depending on the type of policy [5].

2. VARIOUS KINDS OF WEB SPAM

The word "spam" has been used in recent years to point to unwanted and mass (probably commercials) massages. The most common form of spam is email spam. Practically, communication media provide new opportunities to send undesired messages [6].

Web spam has been simultaneously emerged with commercial search engines. Lycos is the first commercial search engine, and has emerged in 1995. At first, web spam was recognized as spamdexing (a combination of spam and indexing). Then, search engines tried to combat with this difficulty [5]. With regard to article presented by Davison in terms of using machine learning methods to detect web spam, this subject has been taken into account as a university discussion [7]. Since 2005, AIRWeb[1] workshops have considered a place for idea exchanging of researchers interested in web spam [5].

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].

Researcher have detected and identified various type of web spam, and they have been divided into three categories:

- Content based spam
- Link based spam
- Page-hiding based spam

[1] Adversarial Information Retrieval on the Web

2.1 Content-based web spam

Content-based web spam has changed the content of page to obtain higher rank. Most of content spamming techniques target ranking algorithms based on TF-IDF. Same of the methods used in this spam is as follows [1]:

- *Body spam:*
  One of the most popular and the simplest methods of spamming is body spam. In this method, terms of spam are placed in documents body.
- *Title spam:*
  Some search engines consider higher weights for the title of documents. Spammers may fill in this tag with unrelated words. Therefore, if higher weight is dedicated to the words of tag from search engine, then the page will receive higher rank.
- *Meta tag spam:*
  The HTML meta tag explanations allow the page designer to provide a short explanation about the page. If unrelated words are placed here, and search engine algorithms consider these pages on the basis of these explanations, then page will receive higher rank for unrelated words. Nowadays, search engines consider lower performance to this tag or ignore it.
- *URL spam:*
  Some search engines break URL of a web page into the terms, sometimes; spams create long URLs containing spam terms. For example, one of URLs created by this method is follows:

  Buy-canon-rebel-20d-lens-case.camerasx.com
- *Anchor text spam:*
  Like document title, search engines dedicate higher weight to anchor text terms, and it presents a summary about the document to which is pointed. Hence, spam terms are sometimes placed in anchor text of a link.
- *Placing spam terms into copied contents:*
  Sometimes, spammers copy the texts on web, and place spam terms in random places.
- *Using many unrelated terms:*
  Spammers can misuse these methods. The page that has been created by this spamming method is displayed in many query words.
- *Repetition of one or more special words:*
  Spammers can obtain high rank for considered page by repeating some the key words. If ranking algorithms of search engines it will be effective.

2.2 Link-based web spam

Link-based web spam is manipulation of link structure to obtain high rank. Some of them have been mentioned as follows[10]:

- *Link farm:*
  Link farm is a collection of pages or sites connected to each other. Therefore, each page will have higher link by creating link farms.
- *Link exchange:*
  Web site owners help each other to add a link to your site. Usually, web site owners obviously show this intention on web pages, or they may be sent to other site owners to request link exchange.
- *Buying the link:*
  Some owners of web sites buy their own web sites from other sites providing this service.
- *Expired domains:*
  Spammers buy expired domains, and unused content is placed over it. Some expired domains may not be already admired, and the links of other sites may remain in these domains, and the validity of those domains is misused.
- *Doorway pages:*
  Web pages involve links. Usually links in this doorway page point to the page of web site. Some spammers may create many doorway pages to obtain higher rank.

2.3 Page-hiding based web spam

Page hiding-based web spam presents a different content to search engines to obtain high rank. Two samples have been mentioned here [11]:

- *Cloaking:*
  Some web sites present different content to search engine rather than to users. Usually, web server can detect and identify company's robots of search engines by IP address, and sends a content different form a page presented to normal users.
- *Redirection:*
  Main page uses different web spamming techniques to be seen by the search engine. When a user refers to a page through search result link, redirection is performed during loading a page.

3. The METHODES OF COMBATTING WITH WEB SPAM

The experts of search engine combat with web spam methods, and they have presented various methods to combat with it, Such as machine learning method and link-based algorithms.
In machine learning method, the classifier predicts that whether the web page or web site has spam or not. This is predicted on the basis of web pages features.
In link-based method, link-based ranking algorithms are used such as HITS and PageRank.

3.1 Machine learning method

One of the methods used to identify web spam is machine learning method. Since web spam methods are continuously changing, the classification of these methods should be necessarily temporary. However, there are some fixed principles [5]:

- Each successful spam, target one or more characteristics used by ranking algorithms of search engine.
- Web spam detection is a classification problem. Through using machine learning algorithms, search engines decide whether a page has spam or not.
- Generally, innovations in web spam detection are followed by statistical anomalies, and are related to some observable features in search engines.

Spam and nonspam pages have different statistical features [12], and these differences are used in terms of automatic classification. In this method at first, some features have been considered for spam page. Through using classification method and on the basis of these features, a method is learnt. On the basis of this method, search engine can classify pages into spam and nonspam page.

Ntoulas et al. took into account detection of web spam through content analysis [13]. Amitay et al. have considered categorization algorithms to detect the capabilities of a

website. They identified 31 clusters that each were a group of web spam [14].

Prieto et al. presented a system called SAAD in which web content is used to detect web spam. In this method, C4.5, Boosting and Bagging have been used for classification [15]. Karimpour et al. firstly reduced the number of features by using PCA, and then they considered semi-supervised classification method of EM-Naive Bayesian to detect web spam [16]. Rungsawang et al. applied ant colony algorithm to classify web spam. The results showed that this method, in comparison with SVM and decision tree, involves higher precision and lower Fall-out [17]. Silva et al. considered various methods of classification involving decision tree, SVM, KNN, LogitBoost, Bagging, adaBoost in their analysis[18]. Tian et al. have presented a method based on machine learning method, and used human ideas and comments and semi-supervised algorithm to detect web spam [19].

Becchetti et al. considered link based features such as TrustRank and PageRank to classify web spam [20]. Castillo et al. took into account link-based features and content analysis by using C4.5 classifier to classify web spam [21]. Dai et al. classified temporal features through using two levels of classification. The first level involves several SVM^{light}, and the second level involves a logistic regression [22].

3.2 Link-based method

With regard to emerging HITS and PageRank and the success of search engines in presenting optimized results by using link-based ranking algorithms, spammers tried to manipulate link structure to increase their own ranking.

PageRank method was introduced by Page et al. in 1998. This method was considered as one of the best solutions to combat with web spam. In this method, all links do not have the same weight in rank determination; instead, links from high rank sites present higher value in comparison with link of sites having fewer visitors. As a result, sites created by spammers rarely have a rule in determining the rank. Due to this issue, Google search engine has been preferred over years [23].

HITS method has been presented by Kleinberg. In this algorithm, sites are divided into two group; namely, Hubs and Authorities sites. In this algorithm, Hub sites refer to those sites involving many links in Authorities sites. These two group effect ranking [24]. Figure 1 show Hub and Authority sites.

Bharat and Henzinger presented imp algorithm proposed as HITS development to solve the problem of mutual reinforcement. Their idea is that if there is K edge on one site in the first host to one document in the second host, and then Authority weight is computed as 1/K. In contrast, if there is L edge from one document over the first host to a set of pages over the second host, then Hub weight is computed as 1/L [25].

Zhang et al. used the quality of both content and link to combat with web spam. They presented a repetitive procedure to distribute the quality of content and link in other pages of the web. The idea proposed in terms of combining content and link to detect link spam seems logical [26].

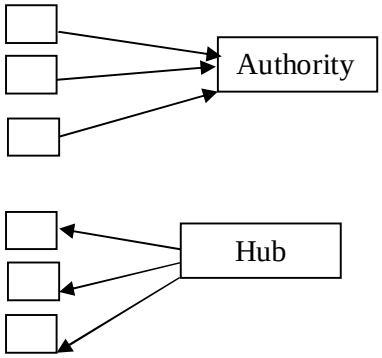

Figure 1. Hub and Authority

Acharya et al. proposed using historical data to detect spam pages for the first time. Heterogeneous growth rate in back links may be a signal of spam [27]. Also, shen et al. features extracted from various reports of web graph are growth rate in input link and death rate in input link.
Eiron et al. proposed HostRank that is more resistant against link spam in comparison with PageRank [28]. Lempel et al. proposed "TKC effect" for the first time. In this method, connected pages obtain high rank for iterative processes. Link farms misuse TKC effort to increase their own rank in search engines. They proposed SALSA algorithm that is more resistant against TKC effect in comparison with HITS [29].

Ng et al. proposed two algorithms; namely, random HITS and subspace THIS for the instability of HITS [30]. Zhang et al. proposed damping factors to compute PageRank to detect the collusion between web pages [31]. Li et al. presented some method to improve HITS results. According to HITS, these pages having less input links and more output link, undesirable results will be obtained. They proposed weighted setting for such pages in adjacently matrix to solve this problem [32]. Chakrabarti et al. created the model of DOM for each web page, and they found out that sub trees that correspond with searching more than other parts, show special behavior against the process of mutual reinforcement [33].

Gyngyi et al. used the concept of trust to combat with link spam, and proposed TrustRank algorithm. TrustRank is one of the most popular and successful anti-spamming techniques. TrustRank is based on trust concept in social networks. In this way, good pages usually point to good pages, and good pages rarely have links to spam pages. Therefore, at first, a group of valid pages are selected, and trust score is dedicated to them. Then, it is followed like distribution scheme of PageRank. Algorithm 1 shows TrustRank algorithm. This is not very different from computing main PageRank. In this algorithm, selecting the seed set is very important. Selection is performed in a way that those pages that have high PageRank score and connection are selected. Here, inverse PageRank is selected in order to select connected and seed pages.

Also, Gyngyi et al. presented different value of PageRank and TrustRank to precisely detect spam pages. In this way, the pages involving good PageRank score and weak TrustRank score are considered as link-based spam pages [34].

Input: T transition matrix
N number of pages
L limit of oracle invocations
α_B decay factor for biased PageRank
M_B number of biased PageRank iterations
Output:
t^* TrustRank scores
Begin
1 s ← SelectSeeds(. . .) ;
2 σ ← Rank({1, . . . , N}, s) ;
3 d ← 0_N ;
4 for i← 1 to L do
if O(σ (i)) = 1 then
d(σ (i)) ← 1 ;
5 d ← d/|d|;
6 t^* ← d ;
for i = 1 to M_B do
$t^* = \alpha_B \cdot T \cdot t^* + (1- \alpha_B)d$
return t^*
End

Algorithm 1. TrustRank

One of anti-spamming algorithms is BadRank. In this algorithm, bad initial page collection is selected, and a value is dedicated to each page in bad pages collection. In this algorithm, like PageRank, a bad value can be distributed via web graph repeatedly. In each repetition, bad value is dedicated to each page pointing to bad pages. Finally, spam pages will have bad and high scores[35].

Guha et al. proposed an algorithm of distributing trust and distrust values at one time [36]. Wu et al. as well as Krishnan and Raj proposed distrust distribution to combat with web spam [2,37]. Both results showed that using distrust distribution in reducing spam rank is more useful than using the trust alone.

Benczur et al. proposed SpamRank. According to their proposition, PageRank values of input link in normal pages should follow power rule distribution. They investigated PageRank distribution of all input links. If, a normal pattern is not followed by distribution, then a penalty will be considered for this page [38].

Becchetti et al. proposed Truncated PageRank algorithm to combat link-based spam. They suppose that link farm spam pages may involve many supporters in web graphs in short intervals, but they don't have any supporters in long intervals, or they have few supporters. Based on this assumption, they presented Truncated PageRank. The first level of links is ignored, and nodes of next stages are computed [39].

Another anti-spamming algorithm is "anti- TrustRank", and it is supposed that if a page points to bad pages, then it may be bad. This algorithm is inverted TrustRank. Anti-TrustRank distributes "bad" scores. In comparison with TrustRank, anti-TrustRank selects "bad" pages instead of good pages [37].

Spam Mass Estimation was introduced following TrustRank. Spam Mass is a measurement of how a page rank is created via linking by spam page. It computed and combines both scores involving regular and malicious scores [34].

Wu and Davison proposed Parent Penalty to combat with link farms [40]. Their algorithm involves three stages.

- Producing a seed set from all data collection
- Development stage
- Value ranking

Algorithm 2 shows that how initial collection is selected. Here, IN(P) shows a collection input links in page P. INdomain(P) and OUTdomain(P) show the domain of input links and output page of P respectively. d(i) is the name of link domain of i.

1 **for** p **do**

2 **for** i in IN(p) **do**

3 **if** d(i) ≠ d(p) and d(i) not in INdomain(p) **then** add d(i) to INdomain(i) ;

4 **for** k in IN(p) **do**

5 **if** d(k) ≠ d(p) and d(k) not in OUTdomain(p) **then** add d(k) to OUTdomain(i) ;

6 X ← the intersection of INdomain(p) and OUTdomain(p);

7 **if** size(X) ≥T_{IO} **then** A[p] ← 1 ;

Algorithm 2. ParentPenAlty: Seed Set

Pages in link farms usually have several nodes common between input and output links. If there is just one or two common nodes, then this page will not be marked as a problematic page. If there is more common nodes, then page may be a part of link farm. In this stage, T_{Io} threshold is used. When the number of common links of input and output links is equal to T_{Io} or greater than T_{Io}, page will be marked as spam, and it is placed in seed set.

Development stage has been shown in algorithm 3. In this stage, bad initial value is distributed for page. It is supposed that if a page only points to a spam page, then no penalty will be considered for it, while if a page involves many output links in spam pages, then the page may be a part of link farm. Hence, another threshold (T_{PP}) is used to detect a page. In this way, if the number of output links in spam pages is equal to threshold or more than threshold, then that page will be marked as spam.

Data: A[N], T_{PP}

1 **while** A do change **do**

2 **for** p : A[p] = 0 **do**

3 badnum ← 0 ;

4 **for** k ϵ OUT(p) **do if** A[k] = 1 **then** badnum ← badnum + 1 ;

5 **if** badnum ≥ T_{PP} **then** A[p] ← 1 ;

Algorithm 3. ParentPenalty: Seed Set Expansion

Finally, bad value is combined with normal link based ranking algorithms. In this way, adjacent matrix of web graph is changed in data set. There are two possibilities to consider a penalty for spam links. They are as follows: reducing the weight of adjacent matrix elements or removing link.

4. CONCLUSION

In the paper, web spam has been considered as a crucial challenge in the world of searching. We explained various methods of web spamming and algorithms to combat with web spam. Up to now, many methods have been created to combat with web spam. However, due to its economical profit and attractiveness, on one side, researchers have presented new methods to combat with it, and in another side, spammers present some methods to overcome these limitations. As a result, a certain method has not been proposed up to now. We hope that we can observe spam pages reduction by presenting character algorithms to detect web spams.

5. REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

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

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

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

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

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

[10] 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.

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

[12] 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.

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

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

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

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

[21] 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.

[22] 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.

[23] 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.

[25] 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.

[26] 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.

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

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

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

[36] 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.

[37] 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.

[38] 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.

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

[40] 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.