Spam Link Detection using Graph Mining

Akankasha Mishra
P.G Student
Parul Institute of Engineering & Technology

Sheetal Mehta
Asst. Prof,
Parul Institute of Engineering & Technology

ABSTRACT
Web deals with huge, diverse, unstructured and dynamic data. The Search Engines are thus an effective way to fetch users query result. Spam poses a significant role in misguiding the web users utilizing spamming techniques on content and link. Thus we need a development of effective and efficient tool that can serve this purpose and thereby minimizes the effect of spam. Link spam can be filtered efficiently using graph based detection. In Graphs based classification nodes are web pages and links are hyperlinks to redirect.

General Terms
Information Retrieval, Search Engine, Machine Learning

Keywords
Spam Detection, Link Spam, Page-Rank, Classification.

1. INTRODUCTION
World Wide Web is an efficient platform which stores, disseminates and fetch information also mine useful knowledge. Web data is highly distributed and heterogeneous, dynamic, huge and unstructured in nature. Also huge amount of search and browse log data resides in various search engines. This massive data gives great deal of opportunities in mining of data to get knowledge and improve web search results. Search engines receive plenty of queries based on user's interest. Web search system consists of four major components – query understanding, document understanding, query-document matching, and user understanding. We researches and application includes- finding relevant information (query-based search), learning from useful knowledge (knowledge discovery), finding needed information (web data semantics), personalization of information (web pattern knowledge), web communities and social networking. Since information is massive manually finding relevant information is difficult. Web search spam's are way to mislead search engines and prevent user from relevant information. Web spammer continues developing tactics that influences results of search ranking algorithm. Thus designing of efficient and effective algorithms and tools that process, model, clean large scale log data is a challenge.

Search Engines are the main source of information used to gather knowledge. In search engines, spamming plays a main role in affecting the quality of a search engine. Spam means to send the same message or unwanted message indiscriminately to large numbers of recipients on the Internet. Spam spread through any information system such as email, instant message news group, web and blog. Spam in web search engine is known as web spam. Search engines are intended to help users find relevant information on the Internet. Typically users summit a query to a search engine, which returns a list of links to pages that are most relevant to this query. Search Engine Optimization (SEO) is popular methods, which are employed to improve rankings. These methods include optimizing page contents, and site structure. There are some cases in which SEO methods are misused to mistake search engine and to acquire higher rankings than appropriate. Such activity is called as search engine spam or web spam or spamdexing, which is the deliberately manipulation of search engine indexes. Most of the web sites are interested to display the web pages within ten search results. Because of this reason they are intended to improve the ranking of web pages through artificially manipulate ranking factor for their pages. There are two types of web spam in search engine. One is the content spam, which is indicate that to manipulate the content of their web pages such as repeated keywords, background color is same as text color, have tiny text at bottom of page etc. The another one is link spam, which is specify the densely connected links with one another, reciprocal links, artificially manipulate the incoming links, outgoing links, getting connection with expiry links for increasing their pagerank. A densely linked set of Hubs and Authorities [6] is represented in figure 1.

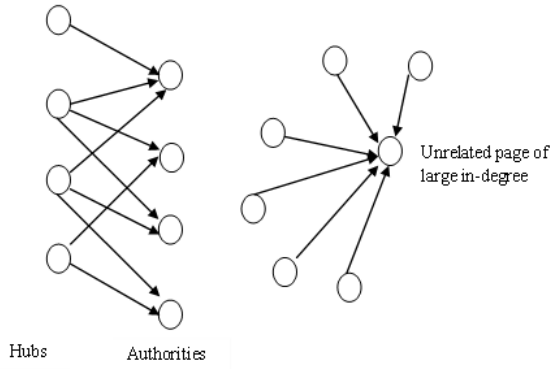

Figure 1. A densely linked set of Hubs and Authorities

2. RELATED WORK
In this section, we describe web spam and spamming techniques. Then, various web spam detection methods are discussed.

2.1 Web Spam
Web search has become very important in the information age. People frequently use search engines to find a company or product, so getting a high ranking position in a search engine's result becomes crucial for businesses. By studying ranking algorithms of various search engines, a lot of techniques have been proposed for a web page to be ranked high in a search engine's results. But these techniques result in web spamming which refers to mislead search engines into ranking some pages higher than they deserve

Content-based spamming methods basically tailor the contents of the text fields in HTML pages to make spam pages more relevant to some queries. This kind of spamming is also called term spamming. There are two main content spamming techniques.

Link spamming misuses link structure of the web to spam pages. There are two main kinds of link spamming. Out-link

spamming tries to boost the hub score of a page by adding out-links in it pointing to some authoritative pages. One of the common techniques of this kind of spamming is directory cloning, i.e., replicating a large portion of a directory like Yahoo! in the spam page. In-link spamming refers to persuading other pages, especially authoritative ones, to point to the spam page. In order to do this, a spammer might adopt these strategies: creating a honey pot, infiltrating a web directory, posting links on user-generated content, participating in link exchange, buying expired domains, and creating own spam farm [4].

Hiding techniques is also used by spammers who want to conceal or to hide the spamming sentences, terms, and links so that web users do not see those [3]. Content hiding is used to make spam items invisible. One simple method is to make the spam terms the same color as the page background color.

In cloaking, spam web servers return an HTML document to the user and a different document to a web crawler. In redirecting, a spammer can hide the spammed page by automatically redirecting the browser to another URL as soon as the page is loaded. In two latter techniques, the spammer can present the user with the intended content and the search engine with spam content [6].

The Web is both an excellent medium for sharing information as well as attractive platform for delivering products and services [3]. This platform is, to some extent, mediated by search engines in order to meet the needs of users seeking information. Search engine has enormous amount of information and return a customary answer for given query with only small set of results. Most of the web sites are interested to display the web pages within ten search results. Because of this reason they are intended to improve the ranking of web pages through artificially manipulate ranking factor for their pages [4]. There are two types of spam in search engine [2]. One is the content spam, which is used by spammers to manipulate the content of their web pages such as repeated keywords, background color is same as text color, have tiny text at bottom of page etc. The another one is link spam, which is specify the densely connected links with one another, reciprocal links, artificially manipulate the incoming links, outgoing links, getting connection with expiry links for increasing their pagerank.

In [1], removing web spam links from search engine results based on the ranking of a page determines the importance of different page features to the ranking in search engine results. In [8], the author used Graph Theory algorithms for finding the web spam in search engine. In [9], there is a new approach for Link spam detection using contents page feature. In [6], the author introduced Anti-Trust Rank algorithm to detect spam pages with relatively high pageranks. In this research, it has been focussed primarily to develop a generic tool for link spam detection in search engine results using Graph mining and to improve the quality without removing any valid results.

3. SYSTEM ARCHITECTURE

The web is an outstanding medium for sharing information using the search engines in order to satisfy the users seeking information. The ideas and the concepts identified this research would benefit the users who search information in the search engine. Since the users will get quality web links and no need to spend much amount of time to search the information in the web. The following are the important factors to create the link spam in search engine.

- Traditional Search engines are combating with the Spamdexing (Web spam).
- Spammer may create the artificial links to improve the pagerank.
- Spammers can create densely connected links with one another.
- Links farms used by spammers to raise popularity of spam pages.
- Spammers can have reciprocal links between the links.
- Spammers have large number of in degree and less number of outdegree.
- Spammer can repeat the same keyword many times in links.

The main intention of this research is to reduce the links spam in search engine results and user gets the quality results for the given query. This system is used to detect the link spam in Web search engine results. The input for the system is top ten search engine results. Web pages available in various Web server of the Web are downloaded by the crawler using either depth first crawler or breadth first crawler and the document corpus collected are stored in the web page repository i.e., a Web warehouse. Then generates the web graph from the search engine results. The Web Graph can represent in the figure 2.

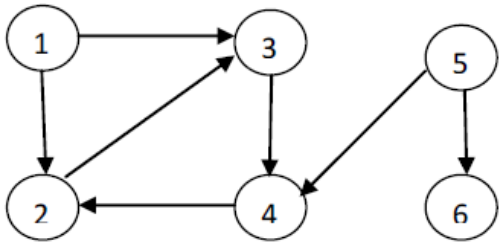

Figure 2. Sample Web Graph

A Graph G is a triple consisting of vertex set V (G), an edge set E (G), and a relation that associates with each edge two vertices [11]. Let us take a small graph and then converted into adjacency matrix representation of graph G. For the adjacency matrix representation of Graph G = (V, E), assume that the vertices are numbered 1, 2, 3,….|V| in arbitrary manner. Then the adjacency matrix representation of a Graph G consists of a |V| × |v| matrix A = (aij) such that (3)

$$a_{ij} = \begin{cases} 1 & \text{if } (i, j) \in E, \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

Table 1.Adjacency Matrix Representation

Node	1	2	3	4	5	6
1	0	1	1	0	0	0
2	0	0	1	0	0	0
3	0	0	0	1	0	0
4	0	1	0	0	0	0
5	0	0	0	1	0	1
6	0	0	0	0	0	0

Let the start node 1, if there is a link to node 2, then adjacency matrix table place "1", otherwise place "0". Likewise complete all the nodes and represented in adjacency matrix is given in Table 1. After generated web graph, classify the links

based on above specified factors. If the above specified features are equivalent to spam pages, then declared as 'spam' pages and finally remove it from web page. The non spam links would have Web pages URL addresses are indexed using indexing methods. The structural connectivity of the Web page URL addresses is stored in the link server and the indices with respect to their Web page ID are stored in the index server. Finally the quality links are returned to client interface.

4. IMPLEMENTATION

The system takes search engine results as input and generates the web graph for the links. The following parameters are used to identify and filter the link spam in the web through checking of each level of parameters. Its mainly focus on the following:

- Degree related features
  - In-degree
  - Out-degree
- PageRank related features
  - PageRank
  - Indegree
  - Outdegree

4.1 Algorithm

- Take a data set.
- Generate the Web Graph for links.
- Check links whether the Degree related features are equivalent to link Spam, remove from the graph.
- Ensure links whether the PageRank related features are equivalent to link spam, remove from the graph.
- Test links which are related to supporter's features are equivalent to link spam, remove from the Graph.
- Check Degree, PageRank and Supporters features for link s, which provides the results are equivalent to link spam, remove from the graph.
- Otherwise declare as "Non Spam links"
- Repeat until completion of all edges in the Graph

The evaluation of the overall process is as in Table 2

Table 2.Classifier Formulation

Label/ Prediction	Non-Spam	Spam
Non-Spam	A	B
Spam	C	D

Where 'a' represents the number of non-spam examples that were correctly classified, 'b' represents the number of nonspam examples that were falsely classified as spam, 'c' represents the spam examples that were falsely classified as non-spam, and 'd' represents the number of spam examples that were correctly classified.

5. RESULTS

Following fig 3 shows results generated for 10367 nodes in NodeXL simulator for graph generation.

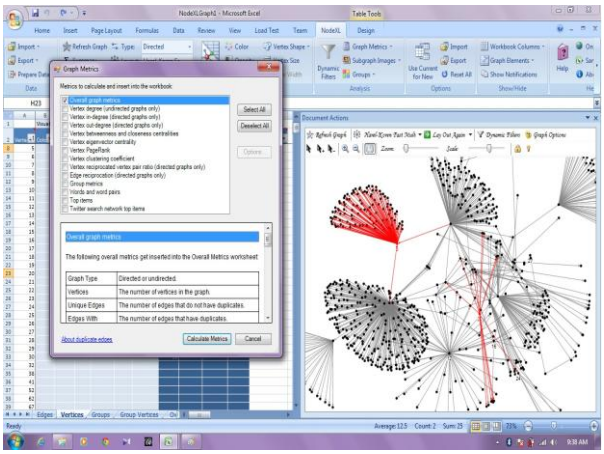

Figure 3 : NodeXL based graph generation

Figure 4 give number of different parameter generated in Nodexl for the given dataset and also gives results for indegree and outdegree

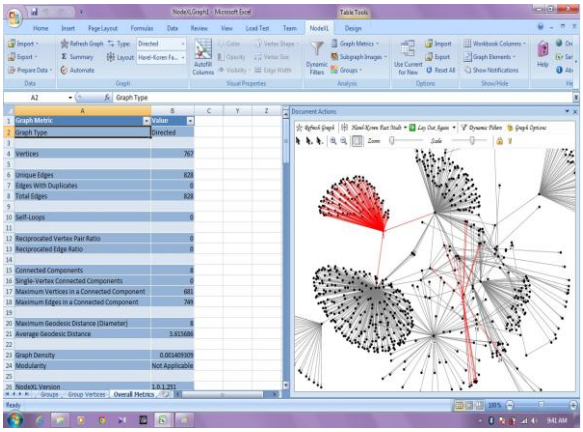

Figure 4 : NodeXL Result

5.1 Results Obtained

Table 3 presents number of non-spam and spam results generated by weka tool

Table 3.Classifier Formulation Result

Label/ Prediction	Non-Spam	Spam
Non-Spam	533	15
Spam	249	240

6. CONCLUSION AND FUTURE WORK

Web Search is entirely based on ranking. Link Spam has significant effect in affecting the page rank. Thus the spamming links need to be filtered and removed. The filtration is based on graph which stores links of only those pages that are having high page ranking. The traditional PageRank suffers from dangling node issue and looping

cycles between hyper links. Moreover normalized PageRank computes the iteration required for convergence which determines time complexity.

This work will be completed in future by calculating detection rate of spam links by calculating true positive, false positive and false negative ratio.

7. REFERENCES

[1] Manuel Egele, Clemens Kolbitseh,Chritian Platzer., "Remove Web Spam Links from Search Results," Springer- Verlag France, 2009.

[2] Isabel Drost and Tobias Scheffer., "Thwatting the Nigritude Ultramine: Learning to identify Link spam," springer-verlag Berlin Heidelberg 2005.

[3] Jyoti Pruthi, Ela Kumar, "Anti-Trust Rank: Fighting web spam, International Journal of Computer Science Issues IJCSI)," Vol.8, Issue 1, January 2011, ISSN (Online): 1694-0814, Faridabad, India, 2009.

[4] Gyongyi, Z., Garcia-Molina, H. 2005. Web Spam Taxonomy. First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb'05).

[5] J. M. Kleinberg., "Authoritative sources in a hyperlinked environment," Journal of the ACM, 46 Sept. 1999.

[6] Wang, W., Zeng, G. Tang, D. 2010. Using evidence based content trust model for spam detection. Expert Systems with Applications. 37: 5599-606.

[7] Y.T. Yu, M.F. Lau, "A comparison of MC/DC, MUMCUT and several other coverage criteria for logical decisions", Journal of Systems and Software, 2005, in press.M

[8] Michael Hilberer, "Development of algorithms for Web Spam Detection Based on Structure And link Analysis," IADIS International journal on WWW/Internet, vol,3, no.2 pp 11-24, ISSN 1645-7641, 2010.

[9] Bin Zhou and Jian Pei, "Link Spam Target Detection using Page Farms, ACM Transactions on Knowledge Discovery from Data," vol.3, NO.3 Article 13, Publication date : July 2009.

[10] Vijay Krishnan and Rashmi Raj, "Web spam detection with Anti-Trust rank," In AIRWeb'06, August 2006.