International Journal of Advance Engineering and Research Development (IJAERD)
Volume 1,Issue 5,May 2014, e-ISSN: 2348 - 4470 , print-ISSN:2348-6406

Improving Spam Link Detection based on Graph for Search Engine Result

Mishra Akanksha AshwaniKumar[1], Prof. Sheetal Mehta[2]

[1]*Computer Science & Engineering, Parul Institute of Engineering & Technology, akanksha10391@gmail.com*
[2]*Computer Science & Engineering, Parul Institute of Engineering & Technology, prof.sheetal.mehta@gmail.com*

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 .It employs calculation of PageRank and Normalized PageRank based on mean value of the traditional PageRank algorithm that filters the spam pages. The resultant numeric value employed is used to obtain rank the page and generate the graph.

Keywords- web graph; link spam; degree feature; pagerank feature; boosting

I. 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. The researches and application includes- finding relevant information (query-based search), learning from useful knowledge (knowledge discovery), finding needed information (web data semantics), and personalization of information (web pattern knowledge), web communities and social networking. Since information is massive manually finding relevant information is difficult.

Search engines are significant for discovering and retrieving information. The web users get thousands of result for keyword search made, but not all are of equivalent significance .But the surfer is also interested in top ranked results to get knowledge he wants. The creation of such spam sites via pages with densely connected links between them leads to problem in spam detection technique. It also makes the surfing too troublesome and As the spam's are made surrounded to the normal pages with implication of multiple techniques .This makes the spam detection difficult. Spam's are widespread in almost every domain from links, mails, social, to blogs.

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.

II. PROBLEM STATEMENT

Information on the Web is massive and search engines are meant for information acquisition. In search engines, spamming has great significance in degrading search engine results (quality).

Spam also deprives legitimate websites, weakens users trust, is means of malware, and wastes significant amount of computational and storage resources. During recent past few years, great advances in detecting fraudulent pages has been noticed but, in response new spams have been noticed.

Thus it is necessary to improve anti-spam techniques to get over these attacks. It's nowadays a great deal for web retrieval systems which have drastic influence on the performance of such systems. Thus we need to refine the anti-spamming technique to serve the user effectively browsing web and solving issue of spamming by increasing efficiency of the spam filtering technique based on graph mining. This improvement focus on incrementing the detection rate of the filtration technique using graph based analysis.

III. OBJECTIVE

The main objective of this work is to improve the ranking of search results being selected. The spamming techniques are tightly coupled to ranking algorithms employed by the major search engines. Since the users can not browse through all pages in result and is interested in few top ranked pages in search results. So, web page ranking in search engine is widely addressed for referencing results.

Its main goal is to solve serious problem of search ranking as abundance of information is available on the web to satisfy users. Also content based methods are easily spammed e.g. by adding some important words to a web page. To solve these problems, researches have resorted to hyperlinks. Hyperlinks embed useful information and useful in organizing information within the websites.

IV. LITERATURE SURVEY

Search engines are the source of acquiring information. Spam can be part of any information system either email, web, social, blog or reviews platform. SEO method employs methods (optimizing page contents, and site structure) to improve ranking of the Web page. In case when a spam page acquire higher rank than appropriate its Spamdexing, e.g. deliberately manipulation of search engine indexes. It is key challenge for search engine industry.

The manipulation can be in any form either with content on the page, excessive and undeserved link creation, click fraud and tag spam. The reason for spam is due to phenomenon which mainly takes into consideration top ranked results. Its main purpose is deliberately triggering unjustifiable and favorable relevance to some target web pages.

4.1 A Generic Tool for Link Spam Detection in Search Engine Results using Graph Mining [1]

The main motive besides web spam is increasing ranking of target page in search engine results. In many cases noted websites intend to increase the popularity of the links. Spam consumes much of the bandwidth and wastes user's time. SEO is a popular method employed to improve ranking based factor taking into consideration site structure and optimization of web page content. But there are cases where SEO method is misused and there is deliberate manipulation of search engine indexes.

Link spam shows densely connected links with each other, artificial, reciprocal links, incoming, outgoing, connection with expiry links to gain high pagerank score. The research cites the concept of detecting link spam based on the concept of web graph model (graph mining).

The important factors creating 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.

Thus we aim to reduce the link spam and guarantee quality search to users.

Parameters considered –

Degree related feature : indegree, outdegree

Pagerank related feature : pagerank, indegree/pagerank, and outdegree/pagerank

4.2 Existing System Architecture

The Search engine consists of:

- Crawler : used for retrieving web contents and page
- Indexer : stores and indexes information on the retrieved pages
- Ranker : measure importance of web page retrieved
- Retrieval Engine : performs lookup on indexes table against query

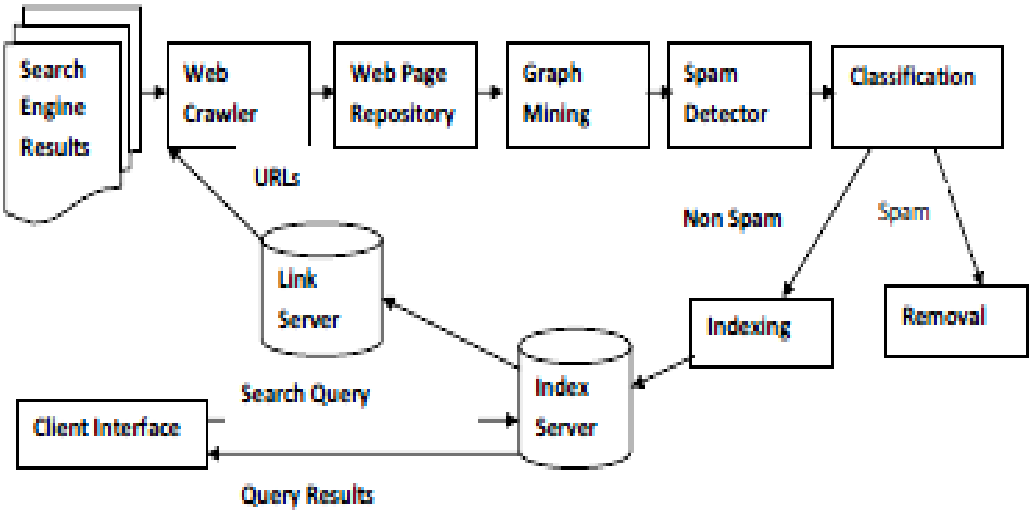


Figure 1: System Architecture [1]

4.3 Evaluation of the Process

Confusion Matrix:

Table 1: Confusion Matrix

		Prediction	
		non-spam	**spam**
True Label	**non- spam**	a	b
	spam	c	d

- ***a*** represents the number of non-spam examples that were correctly classified
- ***b*** represents the number of examples of non-spam that were falsely classified as spam
- ***c*** represents the spam examples that were falsely classified as non-spam
- ***d*** represents the number of spam examples that were correctly classified success

4.4 Measures Considered

True Positive-rate (or Recall) :- $R = d / (c + d)$

False Positive-rate :- $FP = b / (b + a)$

Precision :- $P = d / (b + d)$

F-Measure :- $F = 2PR / (P+ R)$

In this way the graph can be pruned and efficient detection of spam can be computed as per the classification applied and matrices computed. Thus the evaluation of entire process can be analyzed focusing on spam detection task. It is unlike for any analysis method to be accurate in detecting spam links as differentiating spam and non-spam is highly difficult.

4.5 Limitation of Existing System

There are various parameters affecting the detection rate of link spam detection while classifying them into spam and non-spam. The study shows that formulation of detection rate is related to PageRank algorithm which employs measure of score and count of incoming to compute link quality. By working on the issue of page rank, i.e. dangling node monitoring, the efficiency of PageRank could be increased.

As such dangling nodes do not have any outgoing links to any Web page they need to be taken into consideration as it imposes a significant remark in affecting the PageRank computation, allocation. Moreover PageRank feature can thus enhance detection rate effectively. Also computational issues are not addressed.

PageRank is based on link structure solely. Moreover link based algorithm depends on ranking (indegree/inlinks and outdegree/outlinks) which can be easily boosted making web page acquiring irrelevant higher page ranking. This issue needs to be addressed effectively.

V. PROPOSED WORK

Web search engines are requested to provide an efficient mechanism to collect and manage the Web data. Web Search Engines have the capabilities to match user queries with the background indexing database quickly and rank the returned web contents in an efficient way. This forces search engine to waste significant amount of computation al and storage resources. The user is always interested in viewing top ranked pages of search results to get information. Thus spammers pose a significant impact in misleading users by exercising the PageRank properties the web pages. Thus the users need to be aware of such spamming links to have guaranteed and effective search. The link manipulation takes various forms.

The detection rate needs to be made efficient to classify the legitimate sites from malicious sites directed through hyperlinks. The various parameters for testing Web Spam includes such as degree feature, PageRank feature. The efficiency of detection rate as experimented in [1] shows that rate of detection is effective when the features are combined to classify the links.

The main issue of PageRank in evaluation of hyperlinks suffers from the looping and dangling nodes feature. Thus we can optimize the detection rate by handling this issue. The analysis of link structure effectively can be done based on boosting node detection. The boosting nodes are deliberately created by spammers in order to boost the pagerank and provide higher indegree to target spam pages. These nodes can have access and links generated from normal pages too.

Thus boosting nodes have lower indegree and outlinks to spam nodes. In link farm some boosting pages may point to other boosting pages. The boosting nodes are detected based on boosting threshold (c) comparison value which is c usually > 0.5 and between (0, 1).

VI. RESULTS

The following results are generated for two different dataset for which existing link spam detection technique has been applied in which degree related, page-rank related and all features (degree and page-rank) has been tested.

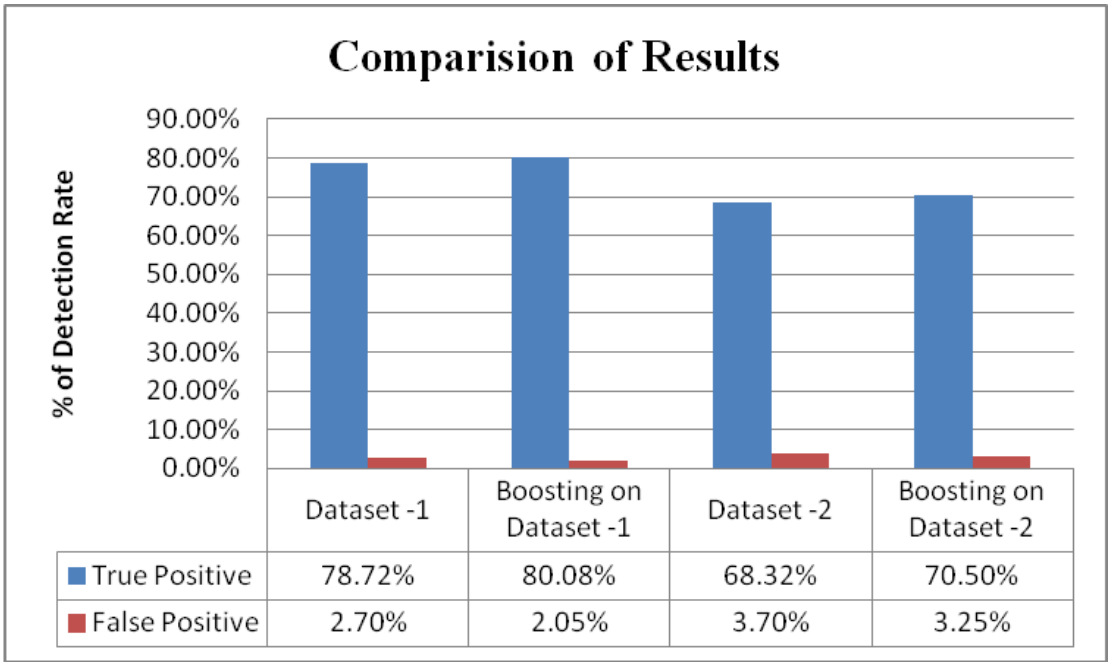
6.1 Comparison of Existing and Proposed Method

Table 2: Comparison of Existing and Proposed Method for Different Dataset

Parameters	Dataset - 1	Dataset - 2
Both Features TP Rate	78.72%	68.72%
Both Features FP Rate	2.7%	3.7%
Both Features With Boosting TP Rate	80.08%	70.50%
Both Features With Boosting FP Rate	2.05%	3.25%

It can be noticed from the above comparison that spam detection based on link analysis is highly dependent on the dataset size and analysis based on page ranking cannot be so effective in determining spam.

Thus application of boosting methodology can to some extent minimize the effectiveness and efficiency of spam detection. The accuracy of detection is increased with application of boosting techniques. This enhances the spam link filtering based on graph mining.



	Dataset -1	Boosting on Dataset -1	Dataset -2	Boosting on Dataset -2
True Positive	78.72%	80.08%	68.32%	70.50%
False Positive	2.70%	2.05%	3.70%	3.25%

Figure 2: Representation of Result Comparison

REFERENCES

[1] D. Saraswathi and A.Vijaya , "A Generic Tool for Link Spam Detection in Search Engine Results using Graph Mining", (PRIME) Pattern Recognition, Informatics and mobile Engineering ,February 21-22, 2013 IEEE

[2] Hema Dubey, Prof. B. N. Roy, "An Improved Page Rank Algorithm based on Optimized Normalization Technique" (IJCSIT) International Journal of Computer Science and Information Technologies, 2011, Vol. 2 (5) 2183-2188

[3] Mr. Dushyant Rathod, "A Review On Web Mining", International Journal of Engineering Research and Technology (IJERT)., April – 2012, Vol. 1 Issue 2.

[4] Renu Sharma, "A Framework to Compare Web Mining Types", , July 2013 ISSN: 2277 *5. Guandong Xu , Yanchun Zhang and Lin Li ,"Web Mining and Social Networking", ISBN 978-1-4419-7734-2 e-ISBN 978-1-4419-7735-9 DOI 10.1007/978-1-4419-7735-9*

[5] Shipra Srivastav,Rinkle rani Aggrawal,."A Modified Algorithm to Handle Dangling Pages using Hypothetical Node", International Journal of Computer Application, Vol 48, June 2012

[6] Hendrik Blockeel , Raymond Kosala," Web Mining Research: A Survey", SIGKDD Explorations. Copyright 2000 ACM SIGKDD, July 2000.

[7] Chakrit Likitkhajorn, Athasit Surarerks, Arnon Rungsawang ," An Approach of Two-Way Spam Detection Based on Boosting Page Analysis ",2012 IEEE

[8] Antriksha Somani, Ugrasen Suman, "Counter Measures against Evolving Search Engine Spamming Techniques", 2011 IEEE

[9] The Anatomy of a Search Engine - The Stanford University InfoLab infolab.stanford.edu/~backrub/google.html?

