International Journal of Engineering Inventions
e-ISSN: 2278-7461, p-ISSN: 2319-6491
Volume 4, Issue 1 (July 2014) PP: 55-59

Page Rank Link Farm Detection

Akshay Saxena[1], Rohit Nigam[2]

[1, 2] Department Of Information and Communication Technology (ICT) Manipal University, Manipal - 576014, Karnataka, India

Abstract: *The PageRank algorithm is an important algorithm which is implemented to determine the quality of a page on the web. With search engines attaining a high position in guiding the traffic on the internet, PageRank is an important factor to determine its flow. Since link analysis is used in search engine's ranking systems, link based spam structure known as link farms are created by spammers to generate a high PageRank for their and in turn a target page. In this paper, we suggest a method through which these structures can be detected and thus the overall ranking results can be improved.*

I. Introduction

1.1 Page Rank

We will be working primarily on PageRank[1] algorithm developed by Larry Page of Google. This algorithm determines the importance of a website by counting the number and quality of links to a page, assuming that an important website will have more links pointing to it by other websites.

This technique is used in Google's search process to rank websites of a search result.

It is not the only algorithm used by Google to order search engine results, but it is the first algorithm that was used by the company, and it is the best-known. The counting and indexing of Web pages is done through various programs like webcrawlers[1] and other bot programs. By assigning a numerical weighting to each link, it repeatedly applies the PageRank algorithm to get a more accurate result and determine the relative rank of a Web page within a set pages.

The set of Web pages being considered is thought of as a directed graph with each node representing a page and each edge as a link between pages. We determine the number of links pointing to a page by traversing the graph and use those values in the PageRank formula.

Pagerank is quite prone to manipulation. Our goal was to find an effective way to ignore links from pages that are trying to falsify a Pagerank. These spamming pages usually occur as a group called link farms described below.

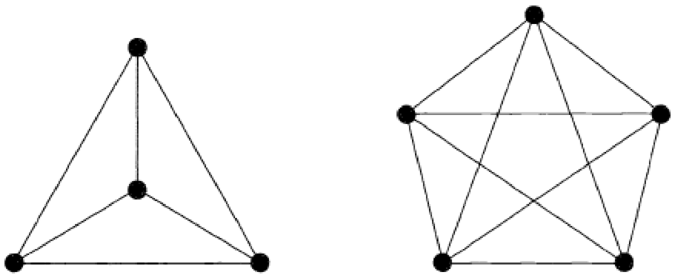
1.2 Complete Graph

In the mathematical field of graph theory, a complete graph[2] is a simple undirected graph in which every pair of distinct vertices is connected by a unique edge. A complete digraph is a directed graph in which every pair of distinct vertices is connected by a pair of unique edges (one in each direction).

1.3 Clique

In the mathematical area of graph theory, a clique[2] an undirected graph is a subset of its vertices such that every two vertices in the subset are connected by an edge. Cliques are one of the basic concepts of graph theory and are used in many other mathematical problems and constructions on graphs. Cliques have also been studied in computer science: the task of finding whether there is a clique of a given size in a graph (the clique problem) is NP-complete, but despite this hardness result many algorithms for finding cliques have been studied.

FIG 1.1



1.4 Link Farm

On the World Wide Web, a link farm[3] is any group of web sites that all hyperlink to every other site in the group. In graph theoretic terms, a link farm is a clique. Although some link farms can be created by hand, most are created through automated programs and services. A link farm is a form of spamming the index of a search engine (sometimes called spamdexing or spamexing). Other link exchange systems are designed to allow individual websites to selectively exchange links with other relevant websites and are not considered a form of spamdexing.

FIG 1.2 Normal Graph of Web Pages

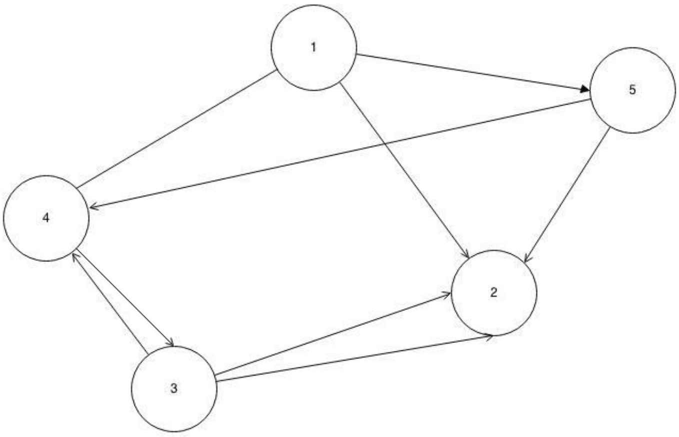
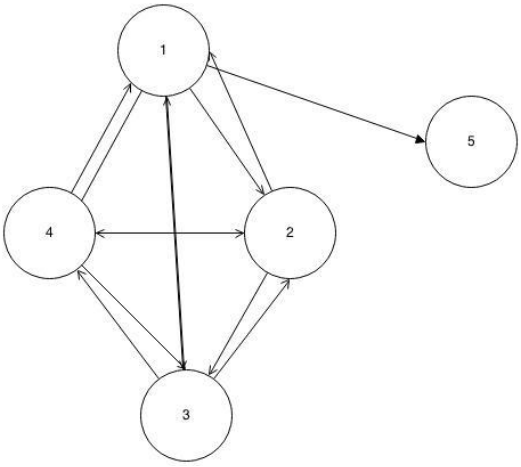


FIG 1.3 Link Farm of 1, 2, 3 & 4



II. Existing System

Academic citation literature has been applied to the web, largely by counting citations or backlinks to a given page. This gives some approximation of a page's importance or quality. PageRank extends this idea by not counting links from all pages equally, and by normalizing by the number of links on a page. PageRank is defined as follows:

We assume page A has pages T1...Tn which point to it (i.e., are inbound links). The parameter d is a damping factor which can be set between 0 and 1. We usually set d to 0.85. There are more details about d in the next section. Also C(A) is defined as the number of links going out of page A. The PageRank of a page A is given as follows:

PR(A) = (1-d)/N + d (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))

Note that the PageRank values form a probability distribution over web pages, so the sum of all web pages' PageRank values will be one.

PageRank or *PR(A)* can be calculated using a simple iterative algorithm, and corresponds to the principal eigenvector of the normalized link matrix of the web. Also, a PageRank for 26 million web pages can be computed in a few hours on a medium size workstation

III. Previous Work

Link spamming is one of the web spamming techniques that try to mislead link-based ranking algorithms such as PageRank and HITS. Since these algorithms consider a link to pages as an endorsement for that page, spammers create numerous false links and construct an artificially interlinked link structure, so called a spam farm, to centralize link-based importance to their own spam pages.

To understand the web spamming, **Gyöngyi** et al[4]described various web spamming techniques in. Optimal link structures to boost PageRank scores are also studied to grasp the behavior of web spammers. Fetterly[5] found out that outliers in statistical distributions are very likely to be spam by analyzing statistical properties of linkage, URL, host resolutions and contents of pages. To demote link spam, **Gyöngyi** et alintroduced TrustRank[4] that is a biased PageRank where rank scores start to be propagated from a seed set of good pages through outgoing links. By this, we can expect spam pages to get low rank. Optimizing the link structure is another approach to demote link spam.

Carvalho[6] proposed the idea of noisy links, a link structure that has a negative impact on link-based ranking algorithms. Qi et al. also estimated the quality of links by similarity of two pages. To detect link spam, Bencz´ur introduced SpamRank[7]. SpamRank checks PageRank score distributions of all in-neighbors of a target page. If this distribution is abnormal, SpamRank regards a target page as a spam and penalizes it.

Becchetti[8] employed link-based features for the link spam detection. They built a link spam classifier with several features of the link structure like degrees, link-based ranking scores, and characteristics of outneighbors. Saito[9] employed a graph algorithm to detect link spam. They decomposed the Web graph into strongly connected components and discovered that large components are spam with high probability. Link farms in the core were extracted by maximal clique enumeration.

IV. Proposed System

The PageRank algorithm, as described in the previous section, works upon a given set of pages found via search using Web spiders or other programs. Now to rank these pages, we consider the links of each page and apply the PageRank algorithm to it. However, the page rank algorithm can be easily fooled by link farms. Link Farms work as follows - We know that the PageRank is higher for pages with more links pointing to it i.e. more the in-line, higher the rank. A link farm forms a complete graph. This means, we have a large set of pages pointing to each other, thus increasing the number of in-links and out-links even they are not relevant, hence they falsify the Page Rank values obtained. On applying the algorithm to this graph, an interesting fact came to notice that the Page Rank values of all the pages in a link farm was the same and remained so even after multiple iterations of the algorithm. One could easily use this fact to identify link farms and remove such sets of pages from our whole set. But the problem occurs when a page which is not a part of the link farm and is actually relevant, it happens to have the same or nearly same PageRank as that of the link farm pages. So if we consider pages with same page ranks for elimination, some relevant pages might also get discarded unintentionally.

To solve this problem, we introduce a new factor we call Gap Rank. GapRank is based on PageRank and in a way inverse of it. While PageRank is based on inbound links of a page, GapRank is calculated using outbound links of pages.

Gaprank Formula

$GR(A) = (1-d)/N + d\ (GR(T1)/L(T1) + ... + GR(Tn)/L(Tn))$

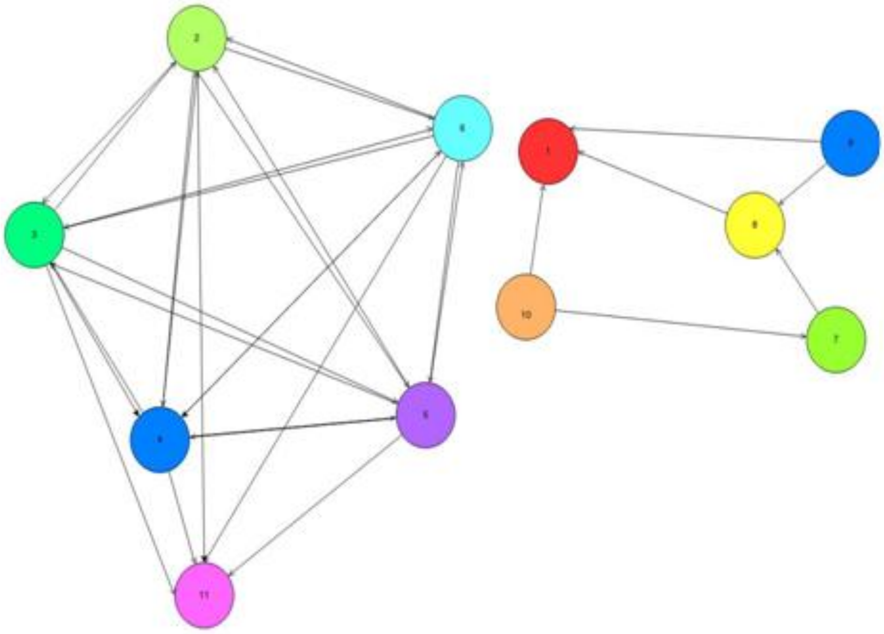
where A is the page under consideration, T1,T2….Tn the set of pages that link from A i.e. outbound pages,L(Tn) is the number of inbound links on page Tn, and N is the total number of pages.

The purpose of this formula is to rank the set of pages based on the number of outbound links. We then use the GapRank to identify the Link Farm. Since all the pages in a link farm have same number of inbound links and outbound links, the GapRank of these pages will be equal, just like their PageRanks. This can now be used to differentiate a link farm from a page having almost same PageRank as those of link farm pages, but it's GapRank will be different.

This analysis can further be used to reduce the ranks of these spam pages or eliminate them altogether so that we get proper ranks for pages that suffered because of the link farm**.**

V. Methodology

Consider a given dataset consisting of 11 pages P1, P2,..,P11. The links between the pages can be seen in the FIG 5.1, depicting a graph of nodes for each page and edges for links between them.



[FIG 5.1 Graph of initial dataset]

5.1 Calculating Pagerank

Initial PageRank to be given to all pages will be 1/n, where n is the size of dataset, i.e., 1/11. Set the initial damping factor. In our example we take damping factor as 0.85.

Calculate the PageRank values of all the pages according to the formula :

PR(A) = (1-d)/N + d (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))

where PR represents the PageRank of the represented page, C represents the total number of out bound links of the page and d represents the damping factor.

For the given dataset, we obtain the following PageRank values after 30 iterations (Table 5.1.1):

Page Name	PageRank
P1	0.0557838
P2	0.0426136
P3	0.0426136
P4	0.0426136
P5	0.0426136
P6	0.0426136
P7	0.0194318
P8	0.0359489
P9	0.0136364
P10	0.0136364
P11	0.049858

[Table 5.1.1 PageRank values of all pages in the dataset.]

We consider the PageRank values obtained after 30 iterations of PageRank calculation for all pages as final because variations in the values after this will be negligible. Negligible changes reflect final values allotted to each page.

5.2 Searching For Duplicate Pagerank Values

Scan the result for all the duplicate PageRank values using the best scanning technique available. The pages with duplicate values are shown in Table 5.2.1. These pages may or may not belong to the spam group (link farm) as some pages may have the same value as a spam page just by mere coincidence. To eliminate these pages from the suspected pages, we perform the next step.

Page Name	PageRank
P2	0.0426136
P3	0.0426136
P4	0.0426136
P5	0.0426136
P6	0.0426136

[Table 5.2.1 Pages having duplicate values]

5.3 Calculating Gaprank Values

Gap Rank values for the selected pages are calculated in the same way as PageRank values, by using the formula shown in fig 4.1. The Gap Rank values obtained are shown in the Table 5.3.1

Page Name	GapRank
P2	0.106365
P3	0.106365
P4	0.106365
P5	0.106365
P6	0.106365

[Table 5.2.1 GapRank values of pages having duplicate PageRank values]

The pages having the same Gap Rank and PageRank values are identified as spam pages which constitute the link farm.

5.4 Remove Link Farms From The Dataset

The pages constituting the link farm are removed from the data set and a new dataset is obtained without the spam pages. The PageRank values of the pages in the new dataset are calculated from the beginning with new initial PageRank values for all pages determined by 1/n', where n' is the new dataset size. The new PageRank values are shown in the table 5.4.1 given below.

Page Name	**PageRank**
P1	0.122724
P7	0.04275
P8	0.0790875
P9	0.03
P10	0.03

[Table5.4.1 New PageRank Values for non-spam pages]

REFERENCES

[1]. Sergey Brin and Lawrence Page: The Anatomy of a Large-Scale Hypertextual Web Search Engine

[2]. Richard D. Alba: A graph-theoretic definition of a Sociometric Clique

[3]. Baoning Wu and Brian D. Davison:Identifying Link Farm Spam Pages

[4]. Z. Gy¨ongyi and H. Garcia-Molina. Web spamtaxonomy. In Proceedings of the 1st internationalworkshop on Adversarial information retrieval on theWeb, 2005.

[5]. D. Fetterly, M. Manasse and M. Najork. Spam, damnspam, and statistics: using statistical analysis tolocate spam Web pages. In Proceedings of the 7thInternational Workshop on the Web and Databases,2004.

[6]. A. Carvalho, P. Chirita, E. Moura and P. Calado. Sitelevel noise removal for search engines. In Proceedingsof the 15th international conference on World WideWeb. 2006

[7]. A. A. Bencz´ur, K Csalog´any, T Sarl´os and M. Uher.SpamRank-fully automatic link spam detection. InProceedings of the 1st international workshop onAdversarial information retrieval on the Web, 2005

[8]. L. Becchetti, C. Castillo, D. Donato, S. Leonardi andR. Baeza-Yates. Link-based characterization anddetection of Web spam. In Proceedings of the 2ndinternational workshop on Adversarial informationretrieval on the Web, 2006.

[9]. H. Saito, M. Toyoda, M. Kitsuregawa and K. Aihara: A large-scale study of link spam detection by graphalgorithms In Proceedings of the 3rd internationalworkshop on Adversarial information retrieval on theWeb, 2007.