Analysis of Web Pages through Link Structure

Sameena Naaz
Department of Computer Science
Jamia Hamdard, New Delhi, India

M Hayat Khan
Department of Computer Science
Jamia Hamdard, New Delhi, India

ABSTRACT
As we know that web is a collection of huge amount of data, it is not very easy to find relevant information. To find the desired data, user visits different web pages. Most Web users typically use a Web browser to navigate a Web site. They start with the home page or a Web page found through a search engine or linked from another Web site, and then follow the hyperlinks they think relevant in the starting page and the subsequent pages, until they have found the desired information in one or morepages.

The aim of this work is to study the different characteristics of various ranking algorithms. Here the factors affecting the ranking of pages of a website are considered and it has been studied that how the popularity of a site can be raised and how spam pages can be tracked. Firstly the importance of different characteristics responsible for Page Ranking are determined. Then by taking this information into consideration a technique is developed that successfully distinguishes spam pages from licit pages.

General Terms
Search Engine Optimization

Keywords
PageRank, Inbound Links, Outbound Links, Spam Page.

1. INTRODUCTION
With the rapid increase in the use of Internet, and the amount of data that now available on the Web, it has become very important to determine which data is relevant and which is irrelevant. Web search has become amazingly powerful in its ability to discover and exploit nearly any kind of information within the billions of pages that comprise the Web. While conventional search engines for algorithmic search have been very flourishing in dealing with moderately simple keywords related web search, nowadays there has been tremendous increase in the exploration of new areas of web due to appearance of new web users. So web search needs advancement and development of many new search applications for web [1].

Traditionally for navigating a website, web users uses web browsers. They start with the home page or a Web page found through a search engine or linked from another Web site, and then follows the hyperlinks they think relevant in the starting page and the subsequent pages, until they have found the desired information in one or more pages. They may also use search facilities provided on the Web site to speed up

information searching. For a Web site consisting of a very large number of Web pages and hyperlinks between them, these methods are not sufficient for users to find the desired information effectively and efficiently. For constructing a link structure of a website we can take web pages as nodes and the hyperlinks between web pages as directed edges. Search engines are the enabling technology for finding information on the Internet [2]. They provide regularly updated snapshots of the Web and keep track of every web page accessed. Its size is increasing day by day, there are about more than 80 millions of web pages. While searching this huge amount of web pages or a datasets, a few amounts of web pages are received as a response of that search query. Among these received pages some pages are relevant and some are irrelevant. So this is responsible for various researches in Page ranking. Google's PageRank is one of the most popular and important page ranking algorithm. This algorithm recursively calculate rank of a page and determines the significance of web page by considering the importance of all pages that are linked to it.

Within the recent years many adjustment, modifications, advancements and research related to PageRank algorithm have taken place, but page rank is still very crucial ranking algorithm because of the basic idea behind the algorithm as it is used in the development of many ranking algorithms.

This paper describes the components which are responsible for affecting the PageRank of the web pages and helps in increasing the popularity of web site and also finding the spam pages. Firstly we determine the importance of different characteristics responsible for Page ranking. So by taking into consideration the information obtained by these characteristics, we develop a technique that uses important characteristics to successfully distinguish spam pages from licit pages.

2. PAGERANK ALGORITHM
PageRank was developed by Google founders Larry Page and Sergey Brin at Stanford. PageRank represents the importance of a page on a web by a numeric value. PageRank is the method used by Google to measure the significance of a web page. The factors of a search such as keywords and Title tag are considered and taken as input by the Google PageRank to determine the importance of a web page, according to which Google adjust and display the result of searched pages, showing the important pages at the top of the list. Whenever a page is pointed by another page it means that the other page is casting a vote for it.It is one of the important factor that is used by Google for ranking of a web page.

The order of ranking in Google works like this: First of all find the pages that are matched by the keywords meant for searching. The upshots are adjusted according to the PageRank scores.PageRank takes the inlinks as input and determines the rank through links. PageRank of a web page will be higher if the sum of the PageRank of its inlinks is high.

The formula propounded by Page and Brin to calculate the PageRank is -

$$PR(A)= (1-d)+d(PR(T1)/C(T1)+\ldots..+PR(Tn/C(Tn))$$

Where PR(Ti) is the PageRank of the Pages Ti which links to

page A, C(Ti) is number of outlinks on page Ti and d is

damping factor. Damping factor is used to neutralize the effect of other pages on a page whose pageRank is being calculated. As suggested by Page and Brin the damping factor is taken as 0.85.

PageRank provides an approach that can be used to compute the significance of web page by just counting the number of inlinks. The importance of these inlinks depends on a fact that if an inlink comes from an important page than this inlink will be given more weightage than the inlink coming from less important or non-important page. The link from one page to another is considered as a vote. So a vote cast by an important page is of much significance[3] [4].

PageRank does not rank a web site as a whole, but it determines the PageRank of each page individually. The PageRank of any page is recursively described by the PageRank of all of its outlinks. The PageRank of page is not influenced uniformly by the PageRanks of its outlinks. Also the PageRanks of these outlinks on their outlinks respectively. Suppose a page A has a link on it from a page T, the PageRank of page A will be influenced by the PageRank of its outlink i.e PageRank of page T. Now the PageRank of page T is weighted by the number of outlinks C(T) on it. Therefore if page T has more outlinks than it will be less beneficial for page A. However an additional inbound link on page A will be beneficial for its PageRank as this inlink will increase Page A's rank [5].

2.1 Illustration of PageRank

The PageRank of a web page is related to the number links to and from it. PageRank forms a probability distribution over the web pages so the sum of PageRanks of all pages in a website will be one. The PageRank of a page is computed iteratively. The Pagerank for various pages shown in figure 1 is shown in table 1.

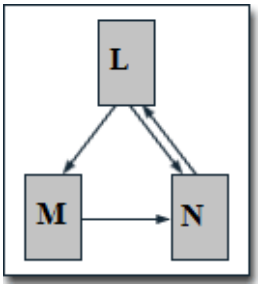

Figure 1:Website Clip 1

Suppose we have a website consisting of three web pages L, M and N. Now page L links to the pages M and N, page M links to page N and page N links to page L. According to Larry Page and Sergey Brin the damping factor should be 0.85, but for simplicity we have taken damping factor as 0.5 as it is not going to change the basic principles of PageRank algorithm. Evaluating Page Rank for the above website, we have:

d=0.5

PR(L) = (1-d) + d* PR(N)
PR(M) = (1-d) + d* (PR(L) / 2)
PR(N) = (1-d) + d* (PR(L) / 2 + PR(M))

On solving above equations iteratively, we get the following PageRank values for each page:

PR(L) = 1.0769231
PR(M) = 0.7692308
PR(N) = 1.1538462

Table 1: PageRank of Web Pages

Iteration	PR(L)	PR(M)	PR(N)
0	1	1	1
1	1	0.75	1.125
2	1.0625	0.765625	1.1484375
3	1.07421875	0.76855469	1.15283203
4	1.07641602	0.76910400	1.15365601
5	1.07682800	0.76920700	1.15381050
6	1.07690525	0.76922631	1.15383947
7	1.07691973	0.76922993	1.15384490
8	1.07692245	0.76923061	1.15384592
9	1.07692296	0.76923074	1.15384611
10	1.07692305	0.76923076	1.15384615
11	1.07692307	0.76923077	1.15384615
12	1.07692308	0.76923077	1.15384615

According to Page and Brin the sum of PageRank of all Pages should be equal to the number of web pages so the sum of Page Rank of all the pages is 3 and therefore equals to the total number of web pages.

As the size of the web is actually very big, the Google search engine will have an iterative computation of PageRank values. Each page is assigned an initial value and the PageRank of all pages are then calculated in several iterations which are based on the equations determined by the PageRank algorithm. From above example we conclude that we get a good approximation of the PageRank values after only a few iterations. But according to Lawrence Page and Sergey Brin, to get a better approximation of the values of PageRank in a system of web,about 100 iterations are necessary.

2.2 The Effect of Inbound Links

If we observe the PageRrank algorithm we will find that each additional inbound link for a web page always increases PageRank of that page. PageRank algorithm, which is given by:
PR(A) = (1-d) + d (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))

We may assume that an additional inbound link from page T increases the Rank of page A by

$d \times PR(T) / C(T)$

Where PR(T) is the PageRank of page T and C(T) represents the total number of its outbound links. As we can see that page A usually links to other pages itself. Therefore, the page rank will also be distributed among these pages. If these pages link back to page A, page A will have higher PageRank benefit from its additional inbound link.

To understand the effects of an additional inbound link, let us illustrate by an example.

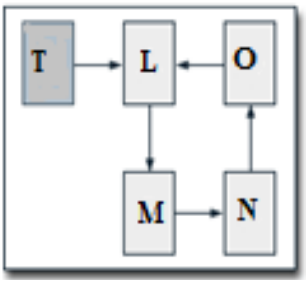

Figure 2 :Website Clip 2

Let us consider a website consisting of four pages L, M, N and O which are linked to each other in circular manner. Initially all these pages will have PageRank of 1 without any external inbound link to any of these pages. Now we will add a page T to our example, assuming that the PageRank of this additional page T is 10 (PR(T) =10). Also, page T links to page L by its outbound link. Taking the damping factor d as 0.5, we get the following equations for the PageRank evaluation:

d=0.5

PR(L) = (1-d) + d*(PR(T) + PR(O))

PR(M) = (1-d) + d*PR(L)

PR(N) = (1-d) + d*PR(M)

PR(O) = (1-d) + d*PR(N)

On solving above equations we get:

PR(L) = 6.33333333

PR(M) = 3.66666666

PR(N) = 2.33333333

PR(O) = 1.66666666

As we know that the initial PageRank of each of the four pages of website is 1, so the PageRank of the additional inbound link of page L is distributed among all the pages accordingly. So we find that the primary effect of the additional inbound link of page L, is distributed by the link on our site, which is given by d × PR(T) / C(T) = 0,5 × 10 / 1 = 5.

2.3 Effect of Outbound Links

As we know that the link structure of the web is the basic principle behind the PageRank algorithm, it is certain that if the PageRank is influenced by inbound links of a page, than its outbound links do also have some other effect.

Let us take an example to illustrate the effects of outbound links.

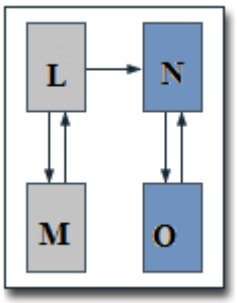

Figure 3: Website Clip 3

Let us assume two websites, each having two web pages. First site have two pages L and M, the Second site is also of two pages N and O. Initially, both pages of each website exclusively link to each other. It is certain that the Page Rank of each page is 1. Now we add a link from page L to page N. Taking the damping factor d as 0.75, now solving the following equations for the PageRank evaluation we get:d=0.75

PR(L)=(1-d)+d*PR(M)
PR(M)=(1-d)+d*PR(L)
PR(N)=(1-d)+d*(PR(O)+PR(L))
PR(O) = (1-d) + d*PR(N)

On solving the above equations the PageRank values for the first website are:

PR(L)=0.60869565
PR(M) = 0.47826086

So the overall PageRank of first website is 1.08695652.

When we evaluate the PageRank values for the second website we get:

PR(N)=1.52173913
PR(O) = 1.39130434

Therefore, the accumulated PageRank of the second website is 2.91304347. The total PageRank for both websites is 4. Therefore, the overall PageRank of a web shall remain same even after the addition of a link. Besides, the increment in PageRank of one site is equals to the decrement in the PageRank of the other site.

As it has already been shown that on adding an additional inlink(inbound link) in a closed system of web pages the overall PageRank benefit is equals to: (d / (1-d)) * (PR(T) / C(T)), where T is the inbound link(linking page), PR(T) is Page Rank of page T, and C(T) representsthe number of outbound links of page T. Therefore, the value thus obtained shows the Page Rank loss in an earlier closed system of web pages, when a page T inside this closed system now points through a link to an external page.

The reasonability of the above formula needs that the web page which receives the link from the earlier closed system of web pages should not link back to that system, because the page then gains back some of the lost Page Rank [6].

According to Random Surfer model a random surfer after getting bored by various clicks is likely to switch to a random page. So thePageRank value of a page is the probability of the surfer reachingon that page by clicking on a link. A web page having no link from it to any other page will be treated as a sink and thus stops the random surfing process. Now if somehow therandom surfer reaches at a sink page, it chooses another URL at random and continues surfing again [7][8].

3. WEB SPAM

There are many deliberate human actions that are meant to trigger an unjustifiably favorable relevance or importance for some web pages, considering the page's true value [9] .Web contains very large number of ways that are attracted by panorama of connecting with large number of web users at a very cheap cost. The most probable way of visiting a website is depended on the result obtained after searching, and most of the web users would like to click on the first few results in a search engine. Therefore, there is an economical advantage for manipulating list of search by developing web pages that can score high independent of their actual merit [10].

A spam page or host is a page or host that is used for spamming or receives a substantial amount of its score from other spam pages. There are many techniques for Web spam, and they can be broadly classified into content (or keyword) spam and

link(topological) spam.To detect spam pages only by content analysis is not possible every time, as some spam pages only differ from normal pages because of their links, not because of their contents. Many of these pages are used to create link farms. A link farm is a densely connected set of pages, created explicitly with the purpose of deceiving a link-based ranking algorithm.

Link-based analysis does not capture all possible cases of spamming, since some spam pages appear to have spectral and topological properties that are statistically close to those exhibited by non spam pages [11] [12].

3.1 Spam Rank

SpamRank is an approach which can be used to find out the negative aspect of the webpage. The foundation of the SpamRank is based on linking to spam pages.A link from a spam page to a normal page increases its SpamRank and thus decreases its PageRank. The SpamRank is similar to PageRank but the main difference is that,SpamRank is defined by the number of outbound links of a web page whereas PageRank is defined by number of inbound links. .i.e. PageRank is proportional to the number of inbound links.

By using the following formula we can evaluate SpamRank of web pages:

SR(A) = SPF(A) (1-d) + d (SR(T1)/I(T1) + ... + SR(Tn)/I(Tn))

Where: SR(A) represents the SpamRank of page A,SR(Ti) is the Spam Rank of pages Ti which are outbound links of page A, I(Ti) represents the number of inbound links of page Ti, d is used to denote the necessary damping factor.

SPF(A) represent the spamming factor for the special evaluation of certain web pages.

By means of the SpamRank algorithm, Spam pages can be evaluated. A Spamming filter assigns a numeric value SPF(A) to them, which can be based on the degree of spamming. The sum of all SPF(A) shouldbe equal to the total number of web pages. By using SpamRank we can find out the spam pages just like by using PageRank we can locate important regions of the web.

3.2 Experimental Setup and Results

Considering some of the characteristics of web pages we can find out the most probable spam pages in a website by using SpamRank formula.

We have assumed a link structure of a website of seven web pages, now this link structure (graph) is taken as input for the above Spam Rank algorithm. Initially we have taken the spam rank of each page as 0 and considering the damping factor equals to 0.85, we get the following equations for the evaluation of Spam Rank of each page.

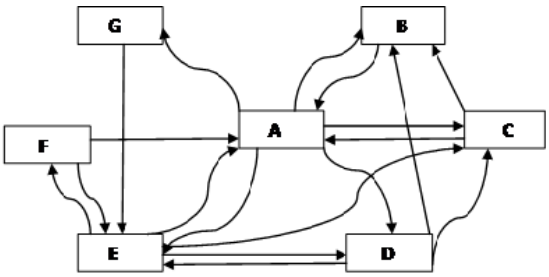

Figure 4: Link Structure of a Website

d=0.85;
SRA = SPF*(1-d)+d*((SRB/Ib)+(SRC/Ic)+(SRD/Id)+
(SRE/Ie)+(SRG/Ig))
SRB = SPF*(1-d)+d*(SRA/Ia)
SRC = SPF*(1-d)+d*((SRA/Ia)+(SRB/Ib))
SRD = SPF*(1-d)+d*((SRB/Ib)+(SRC/Ic)+(SRE/Ie))
SRE = SPF*(1-d)+d*((SRA/Ia)+(SRD/Id)+(SRF/If)+(SRC/Ic))
SRF = SPF*(1-d)+d*((SRA/Ia)+(SRE/Ie))
SRG = SPF*(1-d)+d*(SRE/Ie)
On solving above equations we get the following Spam Rank values for each page:

Table 2: SpamRank of Web Pages

ITR	SR(A)	SR(B)	SR(C)	SR(D)	SR(E)	SR(F)	SR(G)
1	0.15	0.1755	0.3247	0.4372	0.4373	0.2684	0.2429
2	0.8605	0.2963	0.5481	0.7277	0.8495	0.4768	0.3305
3	1.3024	0.3714	0.6871	0.9382	1.1319	0.6119	0.3905
4	1.5960	0.4213	0.7795	1.0799	1.1387	0.7015	0.4302
5	1.7913	0.4545	0.8409	1.1739	1.4426	0.7611	0.4566
6	1.9210	0.4766	0.8816	1.2363	1.5250	0.8006	0.4741
7	2.0071	0.4912	0.9087	1.2778	1.5797	0.8269	0.4857
8	2.0643	0.5009	0.9267	1.3053	1.6161	0.8443	0.4934
9	2.1023	0.5074	0.9387	1.3236	1.6402	0.8559	0.4985
10	2.1276	0.5117	0.9466	1.3358	1.6562	0.8636	0.502
11	2.1443	0.5145	0.9519	1.3439	1.6669	0.8688	0.5042
12	2.1555	0.5164	0.9554	1.3492	1.6740	0.8721	0.5057
13	2.1629	0.5177	0.9577	1.3528	1.6787	0.8744	0.5067
14	2.1678	0.5185	0.9593	1.3552	1.6818	0.8759	0.5074
15	2.1710	0.5191	0.9603	1.3567	1.6839	0.8769	0.5078
16	2.1732	0.5197	0.9610	1.3578	1.6852	0.8776	0.5081
17	2.1747	0.5199	0.9614	1.3585	1.6862	0.8780	0.5083
18	2.1756	0.5200	0.9617	1.3589	1.6868	0.8783	0.5084
19	2.1763	0.5200	0.9619	1.3592	1.6872	0.8785	0.5085
20	2.1767	0.5201	0.9612	1.3594	1.6874	0.8786	0.5086
21	2.1770	0.5201	0.9622	1.3596	1.6876	0.8787	0.5086
22	2.1771	0.5201	0.9622	1.3597	1.6877	0.8788	0.5086
23	2.1773	0.5201	0.9622	1.3598	1.6878	0.8788	0.5087
24	2.1774	0.5202	0.9623	1.3598	1.6879	0.8788	0.5087
25	2.1774	0.5202	0.9623	1.3598	1.6879	0.8788	0.5087
26	2.1775	0.5202	0.9623	1.3598	1.6879	0.8788	0.5087
27	2.1775	0.5202	0.9623	1.3598	1.6879	0.9789	0.5087
28	2.1775	0.5202	0.9623	1.3598	1.6880	0.9789	0.5087
29	2.1775	0.5202	0.9623	1.3598	1.6880	0.9789	0.5087
30	2.1775	0.5202	0.9623	1.3598	1.6880	0.9789	0.5087

ITR: Represents Iterations
SR: Represents SpamRank

Page	Spam Rank
A	2.1775
B	0.5202
C	0.9623
D	1.3598
E	1.6880
F	0.8789
G	0.5087

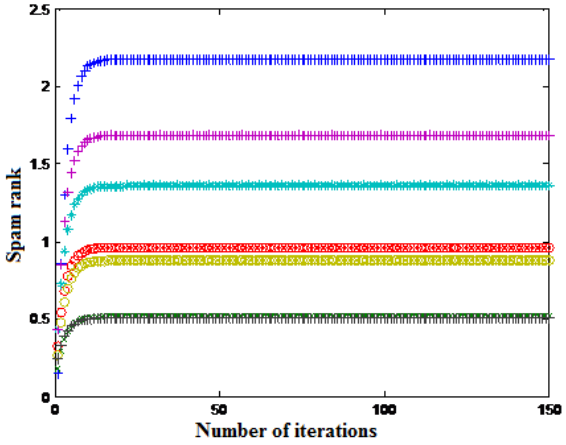

Figure 5: SpamRank of Pages vs Number of iterations.

3.3 Discussion

From the above experiment we can see that the page with higher number of outlinks as compared to number of inlinks has higher SpamRank. So we can say that with the addition of an outbound link, the SpamRank of the page increases. SpamRank does not determine the importance of a web page in a website but it measures its negative characteristics.

SpamRank is based on the principle of linking to spam pages. So, if a page is linked by a page having high SpamRank its SpamRank will also be increased.

4. CONCLUSIONS AND FUTURE WORK

If a webpage has a high PageRank, the influence of its SpamRank can be very low almost negligible. But if it is linked by another page, this could have very serious aftermath.

There is also a problem in the direct reversion of the PageRank algorithm and that is, Just as an additional inbound link on a page can do nothing but can increase its PageRank, an additional outbound link can also increase the SpamRank of that page. It happens because of the addition of SpamRank values in the SpamRank formula. Therefore it is not important that how many good outlinks a page have, even a single link to a spam page can be enough for decreasing the PageRank and resulting in the increment of SpamRank.Undeniably, this problem may appear in few special cases only.

There is also another problem and that occurs when all links are weighted uniformly within the SpamRank computation**.** Consider a case in which two pages differ mainly in PageRank and there is a link on each page from a same page having high SpamRank, then the page with higher PageRank will suffer less from the transferred SpamRank than the page having low PageRank now regarding the procedure presented in this paper,the conclusion is that outbound links are responsible in decreasing the efficiency of the website and the analysis of link structures similar to the PageRank technique help us to understand to some extent that how to deal with topological spam.

In future it will be interesting to see the combination of PageRank and SpamRank. Also we will try to calculate SpamRank first and, then find the PageRank of the pages and also divide the SpamRank each time in the calculation of PageRank iteratively.

5. REFERENCES

[1] Nidhi Grover , Ritika Wason , "Comparative Analysis Of Pagerank AndHITS Algorithms", (IJERT), ISSN: 2278-0181 ,Vol. 1 Issue 8, October -2012

[2] Olston, C. and Chi, E. H., "Integrating Browsing and Searching on the Web", ACM Transactions on Computer-Human Interaction (TOCHI), Vol. 10, No. 3, pp. 177-197.

[3] Larry Page and Sergey Brin , "The anatomy of a large scale hyper-textual Web search engine", Computer Networks and ISDN Systems, 30(1-7):107–117.

[4] Rekha Jain, Dr. G. N. Purohit, "Page Ranking Algorithms for Web Mining", International Journal of Computer Applications (0975 – 8887) Volume 13– No.5, January 2011.

[5] Nan Ma, Jiancheng Guan, Yi Zhao, "Bringing PageRank to the citation analysis" Information Processing and Management 44 (2008) 800–810.

[6] Ji-Rong Wen, "Enhancing Web Search through Web Structure Mining" 2009, IGI Global.

[7] C. Cooper and A. Frieze. "A general model of web graphs. Random Struct. Algorithms", 22(3):311-335, 2003.

[8] Felix Ukpai Ogban, Prince Oghenekaro Asagba, Olumide, "The Illusion in the Presentation of the Rank of a Web Page with Dangling Links", J. Appl. Sci. Environ. Manage. December 2013 Vol. 17 (4) 551-558

[9] Z. Gy□ongyi and H. Garcia-Molina, "Web spam taxonomy", First International Workshop on Adversarial Information Retrieval on the Web, 2005

[10] R. Baeza-Yates, P. Boldi, and C. Castillo, "Generalizing PageRank: Damping functions for link-based ranking algorithms", Proceedings of SIGIR, Seattle, Washington, USA, August 2006. ACM Press

[11] Junghoo Cho, Hector Garcia-Molina and Lawrence Page, "Efficient Crawling Through URL", rdering (PDF, 1998).

[12] Stefano Leonardi,Carlos Castillo,Debora Donato and Ricardo BaezaYates, "LinkBased Characterization and Detection of Web Spam", AIRWEB'06, August 10, 2006, Seattle, **Washington, USA.**

[13]