Analysis of Rank Sink Problem in PageRank Algorithm

Bharat Bhushan Agarwal, Dr M H Khan

ABSTRACT

Web is expanding day by day and people generally rely on search engine to explore the web .It is a challenge for service provider to provide proper relevant and quality information to the internet user by using the web page contents and hyperlink between the web pages and accordingly, ranking the desired pages[11]. One successful and well-publicized link-based ranking system is PageRank, which is used by the Google search engine [2]. The PageRank algorithm has been developed to overcome so called "abundance problem". The World Wide Web continues to grow rapidly at a rate of a million pages per day [4] and searching a right document from such an enormous number of Web pages has brought about the need for a new ranking scheme beyond the relevance matching, the traditional information retrieval scheme. In this paper we discuss the Page Rank Algorithm and deal with the Rank Sink problem associated with the algorithm.

Keywords

Page Rank, Web pages, Rank Sink, Hyperlink Structure.

———————— ★ ————————

1 INTRODUCTION & BACKGROUND

Page Rank was developed at Stanford University by Larry Page and Sergey Brin in 1996 as part of a research project about a new kind of search engine. Sergey Brin come up with the idea that information on the web could be ordered in a hierarchy by "link popularity": a page is ranked higher as there are more links to it. It is patented by the Stanford, and its name Page Rank comes from Larry Page.

In comparison with other Web ranking methods such as HITS [7] or social network analysis methods such as betweenness centrality [10], some attractive features of PageRank include its query-independence, its potential scalability and its virtual immunity to spamming [1].

In [6], Taher calculated PageRank according to the topic-sensitivity of Web pages. That is, they replaced the uniform vector in the PageRank equation by a biased vector considering topic relevance. In [3], Deng Cai *et al* regarded a web page as a set of blocks which can be partitioned by the layout of a web page. They extracted the page-to-block and block-to-page relationships from link structure and page layout analysis and constructed a new semantic graph. Then, they calculated a block based PageRank algorithm.

In [8], Glen and Jennifer encorded personalized views as partial vectors. Then, they calculated PageRank with partial vectors according to user's preference. In [9], Yizhou Lu *et al* proposed a novel framework to calculate PageRank exploiting the power-law distribution of in-degrees of web pages and the hierarchical structure of the web graph. In [5], although the PageRank algorithm is based on a simple idea, They present the block-based strategy for efficiently computing PageRank, a ranking metric for documents, and discuss the number of iterations required to yield a useful PageRank assignment. There are many research attempts to improve the original version of PageRank [6][3].

- *Bharat Bhushan Agarwal is Research Scholar in computer Science engineering in Teerthanker Mahaveer University (TMU) Moradabad, India, TEN 10 Ph.D. /14. E-mail: bharat_innocent@yahoo.co.in*
- *D r.* Mahmoodul Hasan Khan, CSE department I. E. T, Luck now, NDIA.mhkhan.ietfaculty@yahoo.com

In this paper, we address the task of improving the rank of web pages which are more important to user query of backlinks: page A is a back link of page B and page C while page B and page C are backlinks of page D

based on the link structure by improving PageRank algorithm. In Section 2, we described the standard PageRank method used by Google for computing PageRank scores of a given Web network. In Section 3, we discussed the a problem which lead to misleading ranks of web pages. In Section 4, we discuss how to improve the PageRank algorithm by removing the rank sink problem. In Section 5 we discuss some related work and future directions

2 PAGERANK

Page Rank is one of the methods that Google uses to determine the importance or relevance of a web page. The Page Rank of a web page is based on the link structure of the web graph and does not depend on the content of web pages. To rank web pages with their popularity, this algorithm uses number of pages that points to it, also known as in degree algorithm (since it ranks web pages according to their in degree). This concept was used by S. Brin (et al. 1998 &1999) during their PhD at Stanford University. This algorithm is used in most famous search engine 'Google' named as Page Rank Algorithm. It uses the concept of citation analysis and treats incoming links as citations. But as only citation analysis was not giving efficient and relevant result because this gives some approximation of importance of page. So, page rank provides a better approach that can compute the importance of web page by simply counting the number of pages that are linking to it. These type of links are called as back links. If this type of link comes from an important page then this link has higher weightage than those which are coming from non-important pages. The link from one page to another is considered as a vote. Not only the total number of votes that a page receives is important but the relevancy and popularity of pages that casts the vote is also important.

This is actually a measure based on the number of back links to a page. Page Rank is displayed on the toolbar of the browser if the Google toolbar (http://toolbar.google.com/) is installed. But the toolbar will display a Page Rank of 0 to 10 for the page, 0 being an unnoticeable page and 10 a highly visible page.
Therefore, a page has a high rank if the sum of the ranks of its backlinks is high. Figure 1 shows an example

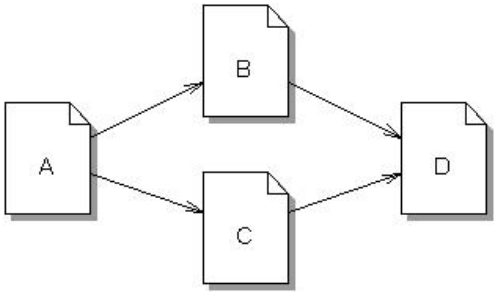

Figure 1: An example of back links

Working of the Page Rank algorithm depends upon link structure of the web pages. The Page Rank algorithm is based on the concepts that if a page contains important links towards it then the links of this page towards the other page are also to be considered as important pages. The Page Rank gave importance to the back link in deciding the rank score. If the addition of all the ranks of the back links is large then the page it is provided has large rank. A simplified version of Page Rank is given in Equation (1.0):

$$PR(u) = \sum_{v \varepsilon B_u} \frac{PR(v)}{N(v)} \quad \ldots\ldots\ldots\ldots (1.0)$$

Where,

u represents a web page,
B (u) is the set of pages that point to u,
PR (u) and PR (v) are rank scores of page u and v respectively,
N(v) indicates the number of outgoing links of page v,
c is a factor applied for normalization.

In Page Rank, The values assigned to the outgoing links of page p are in turn used to calculate the ranks of the pages to which page p is pointing.

Later Page Rank was modified observing that not all users follow the direct links on WWW.

Therefore, Page Rank provides a more advanced way to compute the importance or relevance of a web page than simply counting the number of pages that are linking to it. In a simple way, link from one page

increase the importance of a web page unnecessarily. This problem can be solved by adding an additional term viz. d*E (V). Here E (v) is a vector that adds an artificial link. a This simulates a random surfer, which periodically decides to stop following links and jumps to a new page. E (v) adds

to another page may be considered as a vote. However, not only the number of votes a page receives is considered important, but the importance or the relevance of the ones that cast these votes as well.
Thus, the modified version is given in equation (2.0)

PR (A) = (1-d) + d (PR (T_1)/C (T_1) + …… PR (T_n)/C (T_n)) ……..(2.0)

Where,

PR (A) = the page rank of page A.

PR (T_i) = Page Rank of page T_i which link to page A [i.e., Page A has pages T_1 ……T_n, which points to it (like citations)] which is also called inbound links.

C (T_i) = number of links going out of page T_i (Outbound Links) [i.e., number of pages T_i referred].

d = damping factor which is 0.85 and is used for normalization.

The Page Rank forms a probability distribution over the web pages so the sum of Page Ranks of all web pages will be one. The Page Rank of a page can be calculated without knowing the final value of Page Rank of other pages. It is an iterative algorithm which follows the principle of normalized link matrix of web. Page Rank of a page depends on the number of pages pointing to a page.

3 PROBLEM OF RANK SINK IN PAGE RANK

There are some problems which we have studied during our research and one of them is discussed in the following sub-section.

3.1 Rank Sink

During the research, in the PageRank calculations, problem of cyclic reference occurred due to which the PageRank value for these pages increases which led to links of small probabilities between every pair of nodes.

Within a website, two or more pages might connect to each other to form a loop. If these pages did not refer to but are referred to by other WebPages outside the loop Figure (2), they would accumulate rank but never distribute any rank. This scenario can be termed a *rank sink.*

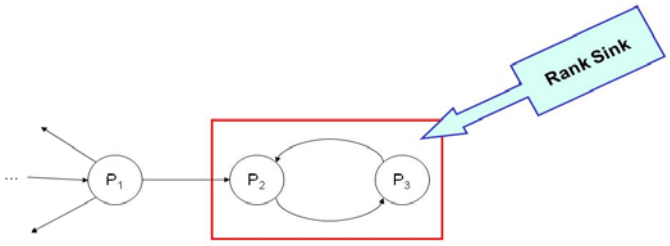

Figure 2: An example of Rank Sink

Example

Consider two web pages that point to each other but to no other page. And if there is number of web pages, which points to one of them. Then, during iteration, this loop according to the figure, will accumulate rank but never distribute any rank Figure (3).

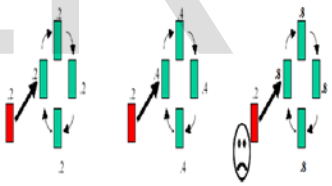

Figure 3: An example of Rank Sink

4 REMOVING RANK SINK BY IMPROVED PAGERANK ALGORITHM

A problem, called rank sink that exists with these PageRank calculations is that when a cyclic reference occurs (Page A points to Page B and Page B points to Page A), and the PR value for these pages increases. This

PageRank Algorithm

When we applied the PageRank algorithm to Figure (4) the PageRank of all the pages can be calculated by the following:-

1) PR(A) = (1-0.85) + 0.85(PR(C)/1)

problem is solved by adding an additional term to the formula.

PR (A) = (1-d) + d (PR (T_1)/C (T_1) + …… PR (T_n)/C (T_n)) + d*E (v) where,

$$E(v) = \frac{(1-d)}{N\,d}$$

Here E (v) is a vector that adds an artificial link. This simulates a random surfer, which periodically decides to stop following links and jumps to a new page.

N = Number of nodes in the graph (link structure)

d = Damping factor used for normalization

To solve the *rank sink* problem, we observed the users' activities. This states that not all users follow the existing links. For example, after viewing page (a), few users may not decide to follow the existing links but directly to page b, which is not directly linked to page a. For this purpose, the users just type the URL of page b into the URL text field and jump to page b directly. In this case, the rank of page b should be affected by page a even though these two pages are not directly connected. Therefore, there is no absolute *rank sink*.

5 Simulation Results

EXAMPLE 1

Let us consider the link structure of three web pages A, B, C shown in the Figure(4):

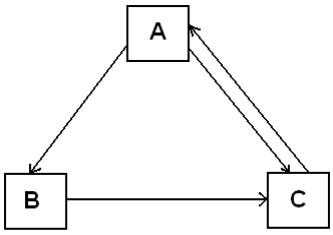

Figure (4): Link structure of three web pages

2) PR(B) = (1-0.85) + 0.85(PR(A)/2)

3) PR(C) = (1-0.85) + 0.85 (PR(A)/2 + PR(B)/1)

After 20 iterations the rank of pages is as follows:-

PR (A) = 1.163375

PR (B) = 0.644418

PR(C) =1.192206

Improved PageRank Algorithm

When we applied the Improved PageRank algorithm to Figure (4) the Page Rank of all the pages can be calculated by the following:-

1) PR(A) = (1-0.85) + 0.85(PR(C)/1)+(0.15/3)

2) PR(B) = (1-0.85) + 0.85(PR(A)/2)+(0.15/3)

3) PR(C) = (1-0.85) + 0.85 (PR(A)/2 + PR(B)/1)+(0.15/3)

After 20 iterations the rank of pages is as follows

PR (A) = 1.537987

PR (B) = 0.852656

PR(C) = 1.576412

The following table & graph Figure (5) shows the variation amongst the various algorithms for computing the rank of

web pages shown in link structure of Figure (4)

The problem of rank sink is present in general PageRank algorithm which led the user to less relevant pages while searching query on WWW because the rank will sink and increase by forming the loop. The improved PageRank algorithm is achieved by removing the rank sink problem in which the different web pages forms a loop and distribute rank unnecessarily to the web pages by the help of which more accurate importance of the web pages is retrieved.

Pages / Stages	Before PageRank	PageRank	Improved PageRank
PR(A)	1	1.163375	1.537987
PR(B)	1	0.644418	0.852656
PR(C)	2	1.192206	1.576412

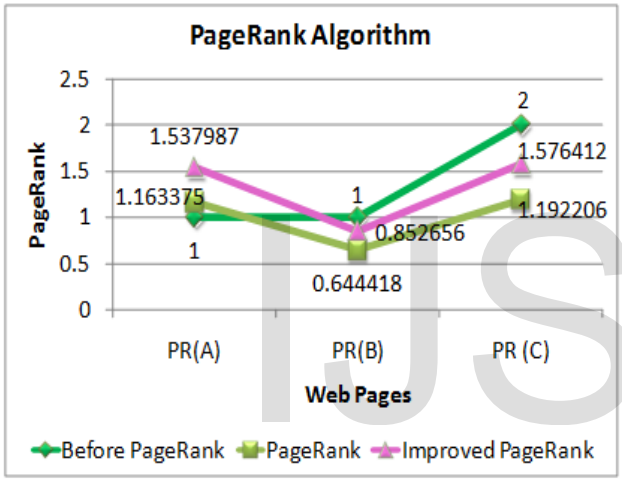

Figure 5: Comparative study of different methods to calculate rank of web pages

6. Conclusion

PageRank, the popular link-analysis algorithm for ranking web pages. The primary goal is to provide high quality search results over a rapidly growing World Wide Web. Earlier the Page Rank gave importance to the back link in deciding the rank score. If the addition of all the ranks of the back links is large then the page it is provided has large rank. Later on, link from one page to another page was considered as a vote, not only the number of votes a page receives is considered important, but the importance or the relevance of the ones that cast these votes as well.

7. Future work

In this paper, we have discussed three scenario of computing the rank of web pages: *first*, how the rank of web pages is computed before general PageRank was developed. *Second*, general PageRank, where all the outlinks are having equal probability of being accessed. *Third*, improved algorithm which removes the problem of cyclic reference by simulating a random surfer, which periodically decides to stop following links and jumps to a new page. This work can also be extended by applying this algorithm on a link structure including large number of web pages on the WWW. The proposed work can be further extended for the study of Weighted Page Rank.

REFERENCES

[1]. Langville A. N. and Meyer C. D (1995) ,Deeper inside PageRank, *Internet Mathematics*, **1(3):** 335–380.

[2]. S. Brin, and L.Page (1998) 'The anatomy of a large-scale hypertextual Web search engine'*Proc. 7th Int. World Wide*

[3]. *Conference, Computer Networks and ISDN Systems, Brisbane,* Australia, 30: pp. 107–117.

[4]. *Cai D., He X., Wen J.R.,*and *Ma W.Y.(2004)* 'Block-level Link Analysis', *Proceedings of the 27th Annual ACM SIGIR 04*, New York ,pp. 440–447.

[5]. S. Charkrabarti(2003), Mining the web Discovering Knowledge from Hypertext Data, *Morgan Kaufmann, San Francisco.*

[6]. T Haveliwala(1999),Efficient Computation of PageRank,*Technical Report*; Stanford University.

[7]. Haveliwala, T (2003), Topic-Sensitive Page Rank: A Context-Sensitive Ranking Algorithm for Web Search, *IEEE Transactions on Knowledge and Data Engineering*, 15(4): 784–796.

[8]. Jeh G.,and Widom G. (2002), Scaling Personalized Web Search, *Technical Report*; 9817799

[10] Lu Y., Zhang B., Xi W., Chen Z., Liu Y., Lyu M.R.,and Ma W.Y. (2004) 'The Power Rank Web Link Analysis Algorithm.' *13th WWW conference*, New York, pp. 254–255.

[11] Newman M.E.J. (2003), The structure and function of complex network, *SIAM Review*, **45(2):**167–256.

[12] Sharma Dilip Kumar and Sharma A K(2010), A comparative analysis of web Page ranking algorithm , *Inter*

National Journal on computer Science and Engineering,2(8):2670-2676.

Bharat Bhushan Agarwal is a Research Scholar in Teerthanker Mahaveer University (TMU) Moradabad, Uttar Pradesh, (INDIA) **and** completed his M.Tech degree in Computer Science and Engineering from U . P. Technical University, Lucknow in 2008. and completed his B.Tech degree in Computer Science and Engineering with Honours from Moradabad Institute of Technology, Moradabad in 2003 and published so many books of computer engineering in Laxmi Publications, Delhi. Now he is working with IFTM University as an Assistant Professor.