VOL. 10, NO. 5, MARCH 2015 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

PENALTY-BASED PAGERANK ALGORITHM

B. Jaganathan and Kalyani Desikan
Division of Mathematics, School of Advanced Sciences, V.I.T University, Chennai, India

ABSTRACT

In this paper we give a brief overview of the original PageRank algorithm that is used in the Google search engine. This algorithm exploits the link structure of the web and greatly improves the results of Web search. We propose a new method for the computation of page rank on the basis of penalty scores assigned to web pages which are accessed through Advertisement links. We compare the page ranks obtained using the original page rank algorithm and our proposed penalty-based page rank method.

Keywords: PageRank, world wide web, search engines, information retrieval.

1. INTRODUCTION

The Internet, as we know, is the world's richest and most intense source of information. From the large collection of disordered information that is available on the internet, the foremost problem that users face is in fetching the most relevant and useful information that they are looking for. Current search engines do not fully satisfy the user's need for high-quality information search. It is hard to retrieve the most relevant and appropriate page from the huge collection of web pages. This leads to numerous challenges in information retrieval.

PageRank algorithm [1] that Sergey Brin and Larry Page proposed is the most popular Web structure based page ranking algorithm. But it has its drawbacks. The PageRank algorithm completely ignores factors like content, topic and relevancy of a web page. It only makes use of the hyperlinks-based structural analysis for measuring the relative importance of web pages. In this paper we revisit the PageRank algorithm and its computational steps. We propose a new technique for finding page ranks by assigning penalty scores to web pages which are accessed through Advertisement links.

This paper is organized as follows: in section 1 on, "Web Page Ranking Algorithm" different types of ranking algorithms are presented. In section 2, "PageRank Algorithm" computation and its illustration are presented. In section 3 on, "Penalty Based PageRank Algorithm", our proposed penalty based page ranking method and its illustrations are discussed. In section 4, "the comparison between page rank algorithms'' is given. In section 4, the conclusion and the possible future work are presented.

1.1 Web page ranking algorithm

We notice that the size of the World Wide Web is growing by leaps and bounds. At the same time the number of users of the internet is also growing incredibly. The mounting number of users on the web submitting innumerable queries puts a huge load on the search engines. The search engines must be capable of scaling up and processing these queries efficiently. Web ranking techniques are employed to extract only the most relevant documents from the database and provide the users with the desired information.

Different ranking algorithms adopt different techniques. In the sections that follow, we discuss in brief the following algorithms:

- Link analysis algorithm
- Personalized web search ranking algorithm
- Page segmentation algorithms

1.2 Link analysis algorithm

Link analysis algorithms exploit the link structure of the web graph. The quality of results from search engines that are based on these algorithms does not meet the user's expectations. In order to offset this and to substantially improve the quality of the search results, web pages must be ranked not only based on the links between the pages but on other criteria as well.

1.3 Personalized web search ranking algorithm

Personalized web search greatly differs from generic web search. In a generic web search identical queries fetch identical search results for all users. It completely ignores varied user interests and their information needs. A personalized Web search, on the other hand, either provides different search results to different users or organizes search results differently for each user. This is done based upon user interests, preferences, and information needs.

1.4 Page segmentation algorithms

Webpage segmentation has several applications. Informative vis-a-vis non-informative content on a web page can be segregated using Segments. Different types of information can also be differentiated by segmentation. This proves to be very useful in web page ranking. Consider a multiword query whose terms match across different segments of a web page. This information can obviously be used to adjust the relevance of the page to the query.

In the next section we discuss the PageRank algorithm that was originally proposed by Larry Page and Sergey Brin.

2. PAGERANK ALGORITHM

2.1 PageRank computation

Consider the web as a directed graph G={V,E}, where V is the set of vertices or nodes, i.e., the set of all web pages, and E is the set of directed edges in the graph, i.e., hyperlinks. Let N be the total number of pages in the web graph. Define an adjacency matrix A for the web graph G by numbering the web pages, say from 1 to N and then assigning $A_{ij} = 1/n$, if there is a hyperlink from page i to j, n is the number of out links from node i, and $A_{ij} = 0$ otherwise, it can be denoted as

$$A_{ij} = \begin{cases} 1/n & if\ (i,j) \in E \\ 0 & otherwise \end{cases}$$

The initial page rank for each of the N nodes is PR_0= (PR_0 (1), PR_0 (2),… PR_0 (N)) and their value is set as 1.
Page rank formula for the m^{th} iteration is

$$PR_m = (1-d) + d * A * PR_{m-1}. \quad (1)$$

where d is the damping factor whose value lies between 0 and 1, and it is usually set as 0.85. The basic premise of the PageRank Algorithm is that an imaginary surfer who randomly clicks on hyperlinks will eventually stop clicking. The probability that the surfer continues to click, at any step, is the damping factor, denoted as d in the above formula. For our analysis in this paper we would be using the following web graph.

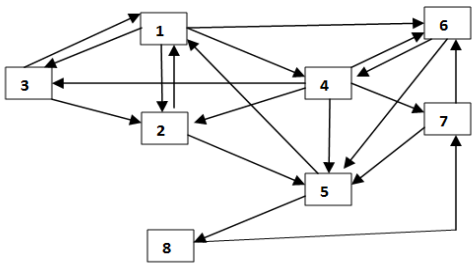


Figure-1. Hyperlink structure of web graph.

2.2 PageRank illustration

Let us assume the hyperlink structure for the web graph of eight pages given in Figure-1. In this hyperlink structure, the pages are denoted by rectangles. The links are denoted by directed lines from one page to another. Every page in this hyperlink structure has at least one outgoing edge. Its corresponding Adjacency matrix representation is presented in Table-1.

Table-1. Adjacency matrix for the hyperlink structure of web graph 1.

i/j	1	2	3	4	5	6	7	8
1	0	1/2	1/2	0	1/2	0	0	0
2	1/4	0	1/2	1/5	0	0	0	0
3	1/4	0	0	1/5	0	0	0	0
4	1/4	0	0	0	0	1/2	0	0
5	0	1/2	0	1/5	0	1/2	1/2	0
6	1/4	0	0	1/5	0	0	1/2	0
7	0	0	0	1/5	0	0	0	1
8	0	0	0	0	1/2	0	0	0

In the above matrix, each entry a_{ij} represents the reciprocal value of total number of links from j to i. The initial page rank value of each page is taken to be 1. Page rank value of each page at each iteration is calculated by using equation (1). To determine the final Page rank of a web page iterations are carried out until they converge.

3. PENALTY BASED PAGERANK

We now present our penalty-based Page Rank method and its illustration

3.1 Penalty-based PageRank method

Within a web graph some pages may be identified as advertisement pages. These pages may not be relevant to search queries. But many pages may point to the same advertisement page and this might lead to Advertisement pages having an artificially higher page rank. To offset this, we can assign penalty scores for advertisement pages. One Penalty based Adjacency matrix that we propose is as follows:

$$A_{ij} = \begin{cases} 1/n & if\ (i,j) \in E\ and\ j \notin Ad(V) \\ -1/n & if\ (i,j) \in E\ and\ j \in Ad(V) \\ 0 & otherwise \end{cases}$$

where $Ad(V)$ is the set of all advertisement pages.

3.2 Penalty based PageRank illustration

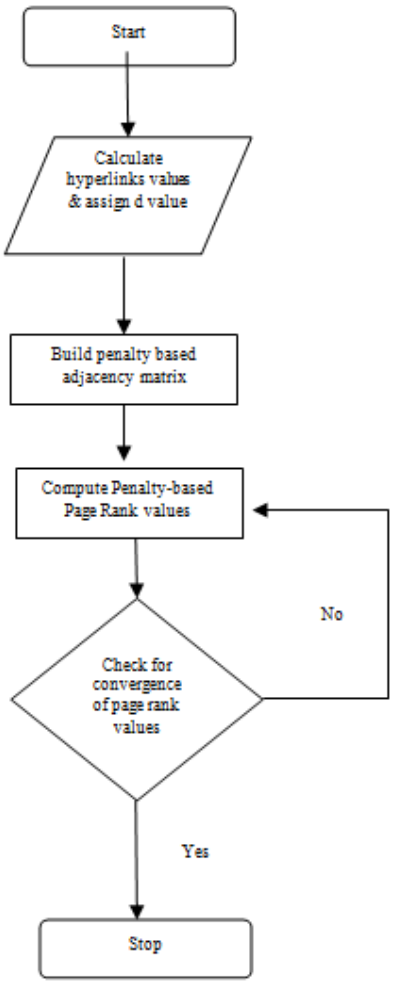
For the web graph shown in Figure-1, suppose that the pages 3 and 7 are advertisement pages, we assign the penalty scores for these advertisement pages. Now the Penalty based adjacency matrix for the graph is as given below:

Table-2. Penalty based Adjacency matrix for the hyperlink structure of Figure-1.

i/j	1	2	3	4	5	6	7	8
1	0	1/2	1/2	0	1/2	0	0	0
2	1/4	0	1/2	1/5	0	0	0	0
3	-1/4	0	0	-1/5	0	0	0	0
4	1/4	0	0	0	0	1/2	0	0
5	0	1/2	0	1/5	0	1/2	1/2	0
6	1/4	0	0	1/5	0	0	1/2	0
7	0	0	0	-1/5	0	0	0	1
8	0	0	0	0	1/2	0	0	0

3.3 Flow chart for Penalty-based PageRank algorithms

Given below is the flow-chart for our Penalty-based PageRank algorithm



3.4 Algorithm to compute PageRank values using penalty based PageRank method

We now present the algorithm for computing the PageRank making use of our Penalty-based algorithm.

Input: Set of web pages, in-links and out-links of pages, damping factor d=0.85
Output: Page Rank Score.
Process:

- Assign an initial page rank value for each node/web pages
- Assign penalty scores for advertisement pages
- Generate adjacency matrix for the web graph based on the penalty scores
- Calculate ranks scores using penalty based adjacency matrix

4. COMPARISON BETWEEN PAGERANK ALGORITHMS

We calculated the page ranks for the hyper link structure given in Figure-1 using both the original Page rank algorithm as well as our proposed penalty based algorithm. In the Table given below we have shown the web page ranks computed using the two algorithms. We have arranged the web pages (vertices) in the increasing order of page rank values.

Table-3. Comparison between the page rank algorithms for web graph in Figure-1.

Web page	Penalty-based PageRank	Web page	Original PageRank
5	1.595	5	1.5021
1	1.425	1	1.4042
6	0.9575	6	1.0095
2	0.9575	7	0.9693
4	0.7875	4	0.8774
8	0.575	2	0.8515
3	-0.2325	8	0.7884
7	-0.87	3	0.5976

The graphs below show the page rank values obtained using the two algorithms for the web graph.

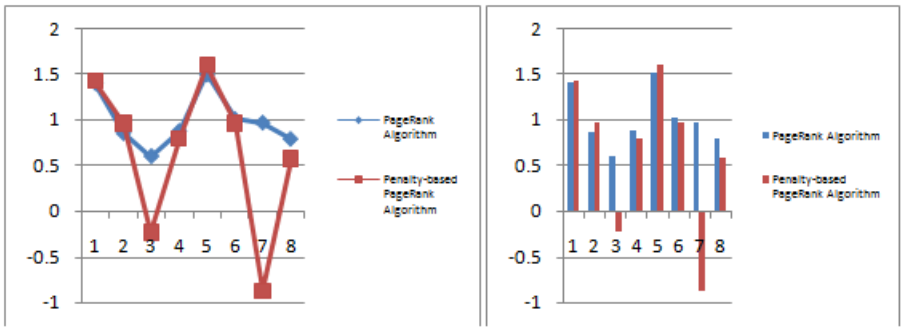


Figure-2. Comparison of Page rank values for the web graph.

It can be noted that the page ranks computed using the two algorithms match with respect to the top three web pages. Also, we see that the page rank scores for the advertisement pages are the lowest (negative values), as it must be, in our proposed algorithm.

5. CONCLUSION/RESULTS AND DISCUSSIONS

This paper focuses on penalty based PageRank score method for calculating the PageRank of web pages. Our algorithm enables us to distinguish between the most significant web pages and the least significant ones. This paper shows that penalty based page rank methods are better than the existing page rank method when we are able to identify the advertisement pages in a web graph. Our finding suggests that the penalty based page rank method boosts up the page ranks of the most relevant pages and pulls down the page ranks of irrelevant/ advertisement pages. As a future scope, we propose to analyse the performance of penalty based page rank methods in other web structure mining algorithms.

REFERENCES

[1] Page, L., Brin, S., Motwani, R., Winograd. T. 1998. The Page Rank Citation Ranking: Bringing Order to the Web. Technical report, Stanford Digital Library Technologies Project.

[2] Kleinberg, J.M. 1999. Authoritative sources in a hyperlinked environment. Journal of the ACM. 46 (5), 604-632.

[3] Dilip Kumar Sharma, Sharma A. K. 2010. A Comparative Analysis of Web Page Ranking Algorithms. Journal on Computer Science and Engineering. Vol. 02, No. 08, pp. 2670-2676.

[4] Laxmi Choudhary., Bhawani Shankar Burdak. 2012. Role of Ranking Algorithms for Information Retrieval. International Journal of Artificial Intelligence and Applications (IJAIA); 3, 4, 21-34.

[5] Mandar Kale., Santhi Thilagam, 2008. P. DYNA-RANK: Efficient calculation and updation of PageRank. International Conference on Computer Science and Information Technology.

[6] Sepandar, D., Kamvar Taher, H., Haveliwala Christopher, D., Manning Gene, Golub H. 2003. Exploiting the Block Structure of the Web for Computing Page Rank, Stanford University Technical Report.