Journal of Informatics and Mathematical Sciences
Vol. 9, No. 3, pp. 649–656, 2017
ISSN 0975-5748 (online); 0974-875X (print)
Published by RGN Publications

http://www.rgnpublications.com

Proceedings of the Conference
Current Scenario in Pure and Applied Mathematics
December 22-23, 2016
Kongunadu Arts and Science College (Autonomous)
Coimbatore, Tamil Nadu, India

Research Article

Penalty Matrix-based PageRank Algorithm

B. Jaganathan* and Kalyani Desikan

Division of Mathematics, School of Advanced Sciences, VIT University, Chennai 600127, India
***Corresponding author:** jaganathan.b@vit.ac.in

Abstract. In this paper we give a brief overview of the adjacency matrix based page rank algorithm and eigen vector based page rank that are used in the Google search engine. In this paper a new approach has been introduced by considering the web as a mixed graph rather than a simple graph. We propose an improved method for the computation of page rank on the basis of penalty assigned to web pages which are accessed through Advertisement links/pages. Consequently, we have applied the concept of column-stochastic Penalty Matrix to web page ranking. This approach does not involve any iterative technique. This method is based only on the concept of Eigen values and Eigen vectors of the Penalty matrix.

Keywords. Page rank; Eigen values; Eigen vector; Penalty matrix

MSC. 05C99; 68M10; 94C15

Received: January 5, 2017 **Accepted:** March 10, 2017

1. Introduction

In the world wide web, web page ranking is an important aspect, and is very much needed to retrieve the relevant/informative pages from the huge collection of unordered information, based on the interest of the user. Search engines play an important role in retrieving information for a given query. But search engines cannot fully satisfy the user's need for relevant information search services.

Many of today's search engines use query terms to retrieve pages related to a user's query. With the huge size of the Web, this step normally results in thousands of retrieved pages related to the query. To make this list manageable, many search engines sort this list by some ranking criterion. One popular way to create this rank is to make use of the additional information inherent in the Web due to its hyper linking structure. Thus, link analysis has become the means to ranking. One successful and popular link based ranking system is PageRank, the ranking system used by the Google search engine. Actually, for pages related to a query, an information retrieval score is combined with a PageRank score to determine an overall score, which is then used to rank the retrieved pages [1,2].

The structure of this paper is as follows: Section 2 deals with the different types of ranking algorithms. In Section 3 the relationship between web pages and web graph is presented. In Section 4, the Iterative adjacency matrix based page rank algorithm is presented. Section 5 deals with the eigen value and eigen vector of the adjacency matrix based page rank algorithm. In Section 6, we present our proposed.

Penalty matrix based page ranking method and its illustration. In Section 7 the comparisons between the different PagerRank algorithms are given. In Section 8 the conclusion and possible future work are presented.

2. Different Types of Ranking Algorithms

Sergey Brin and Larry Page proposed the PageRank algorithm at Stanford Universtiy [3]. A new approach known as weighted PageRank algorithm was put forth by Wenpu Xing and Ghorbani Ali [4]. The algorithm is an extension of the original PageRank algorithm. Recently, we developed three PageRank algorithms: Category based PageRank algorithm, Penalty based PageRank algorithm and Weighted PageRank algorithm based on in-out weight of webpages [5–7]. Generally, users search for web pages within a particular category or topic only. For example, individuals normally search the web for information related to say a particular topic or category of information like music, movies, and different kinds of sports etc. It depends on an individual's interest. Hence, it would be ideal to calculate the page rank based on categories of web pages. Hence we proposed Category base pagerank algorithm [5]. 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 proposed Penalty-based PageRank Algorithm [6]. In the weighted page rank algorithm, more important (popular) web pages are assigned larger page rank values. The popularity of a web page depends on the number of its in links and out links and each web page gets a proportional page rank value. The popularity of each page can be obtained using the in and out weights [7]. In this paper we propose a efficient PageRanking algorithm based on Eigenvalue and Eigenvector of the Penalty matrix associated with the web graph. This approach does not involve any iterative technique.

3. Webgraph and Graph Theory Relationship

Before proceeding further, we must understand the relation between web pages and a web graph. The World Wide Web (WWW) is generally represented as a directed web graph. The vertices are the web pages and directed edges represent the hyperlinks between web pages (out-link/in-link) [5–7]. A degenerate edge of a graph which joins a vertex to itself is called a loop. A web graph with no loops is called a simple directed graph.

An example of a simple directed graph representing 4 web pages connected by hyperlinks is given below:

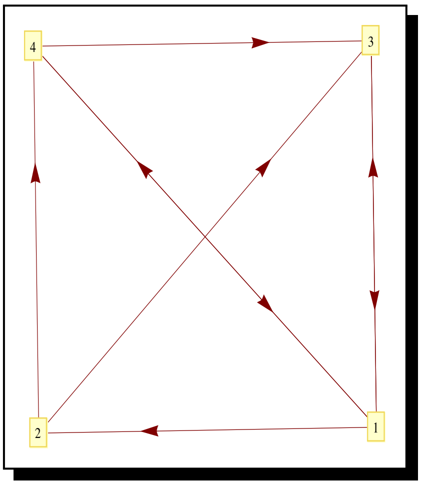

Figure 1. Web graph-1

In graph theory, an adjacency matrix [3–5], for a directed graph is a matrix for representing adjacent vertices (or nodes) of the graph. For any directed graph G with n vertices, the adjacency matrix is a $n \times n$ matrix with matrix elements being 1 for vertices (or nodes) which point to other vertices and 0 otherwise. This can be mathematically represented as:

$$a_{i,j} = \begin{cases} 1 & \text{if } i \neq j \text{ and } v_i \text{ is pointing/links to } v_j \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

and it is denoted by $A(G)$.

In simple graphs with no loops the adjacency matrix consists of only zeros and ones with diagonal entries being zero.

Adjacency Matrix for Figure 1 is as follows:

$$A(G) = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{bmatrix}. \tag{3.2}$$

We now explain the linear algebra behind the original Google page rank and our proposed Penalty page rank algorithms by considering the hyperlinked web graph shown in Figure 1, consisting of four pages 1, 2, 3 and 4.

4. Original Page Rank Algorithm: Iterative Method

The PageRank algorithm [3] was originally proposed by Larry Page and Sergey Brin. A brief explanation of Adjacency matrix based page rank algorithm used in the Google search engine is given below.

Consider the web as a directed graph on n vertices (webpages).

The initial page rank for each of these n web pages is denoted by the vector $\mathrm{PR}_0 = (\mathrm{PR}_0(1), \mathrm{PR}_0(2), \dots, \mathrm{PR}_0(n))$ and the value of each element of this vector (web page) is set as 1.

The formula for computing the mth iteration of the pagerank is given by

$$\mathrm{PR}_m = (1-d) + d * A(G) * \mathrm{PR}_{m-1}, \tag{4.1}$$

where d is the damping factor [6]. The probability, at any step, that the person will continue to click is known as the damping factor or teleport probability, denoted as d. In the above formula, the parameter d is the damping factor or teleport probability whose value is between 0 and 1, and it is usually set to 0.85. Page rank value of each page, at each iteration, is calculated using equation (4.1). To determine the final Page rank of a web page, iterations are carried out until they converge.

5. Linear Algebra behind Google PageRank Algorithm

The linear algebra behind Google Pagerank algorithm was originally proposed by Kurt Bryan and Tanya Leise [9]. A brief explanation of linear algebra behind Google PageRank algorithm is given below.

The objective behind this approach is to rank webpages acoording to their importance scores. A page's importance score is derived from the links made to that page from other webpages. Importance scores and there by the ranking of the webpages are based on an eigenvector of a weighted link matrix. Importance score for any web page will be a non-negative real number.

We now explain the linear algebraic approach behind Google PageRank algorithm by considering the hyperlinked webgraph shown in Figure 1.

In Figure 1 an arrow, say, from page 1 to page 2 indicates a link (hyperlink) from page 1 to page 2. A page that is pointed to is considered to be important. If X_i indicates the number of links that point to page i, for Figure 1 we get $X_1 = 2$, $X_2 = 1$, $X_3 = 3$ and $X_4 = 2$. This implies that page 3 is the most important page since it has the maximum number pages pointing to it. Pages 1 and 4 tie for the second position and page 2 is least important. This approach is very simplistic and it is purely based on the number of inlinks that a webpage has. But it is logical to note that a link to page k from an important page should boost page k's importance score more than a link from an unimportant page. In our web graph pages 1 and 4 have 2 in-links, also each of these pages links to the other. We can see that page 1's second in link is from a seemingly important page 3 while page 4's second in link is from a relatively unimportant page 1. So possibly page 1 must be ranked higher than page 4. Hence, taking this into consideration and thereby overcoming the drawbacks of the earlier approach based merely on the number of links Kurt Bryan and Tanya Leise [9] introduced the following improvement:

Compute score of page k as sum of scores of all pages linking (pointing) to page k.

So, page 1's score must be given by scores of page 3 and page 4 that link (point) to page 1

$$X_1 = X_3 + X_4\,.$$

Similarly, we can write the remaining equations as follows

$$X_2 = X_1,$$

$$X_3 = X_1 + X_2 + X_4,$$

$$X_4 = X_1 + X_2\,.$$

But this approach has its own drawback. The importance score of a webpage can be artificially boosted by just being linked to a lot of important pages. Hence, it has to be ensured that a web page does not gain extra influence by merely linking to a lot of important webpages. This is achieved by boosting page k's score by X_j/n_j, if page j links to page k, where n_j is the number of outlinks from page j.

Hence, for page 1 in Figure 1, we have

$$X_1 = X_3/1 + X_4/2$$

as page 3 contains only one outlink, while page 4 contains two outlinks (splitting its vote into half).

Similarly, for the remaining pages we have

$$X_2 = X_1/3,$$

$$X_3 = X_1/3 + X_2/2 + X_4/2,$$

$$X_4 = X_1/3 + X_2/2\,.$$

This system of linear equations can be written as

$$AX = X \tag{5.1}$$

where $X = [X_1\ X_2\ X_3\ X_4]'$.

We see that the Webpage ranking problem has been transformed into the "standard" linear algebra problem of finding an eigenvector corresponding to eigenvalue 1 for a square (link) matrix A. Here, the components of the eigenvector give the importance scores (ranks) of the corresponding pages.

Matrix A is referred to as the "link matrix" for the given web graph. Link matrix is column-stochastic matrix. A square matrix is said to be a column-stochastic matrix if all its elements are nonnegative and the elements in each column sum to one. Also, we know that every column-stochastic matrix has 1 as an eigenvalue. For the web graph in Figure 1, the link matrix is

$$A = \begin{bmatrix} 0 & 0 & 1 & 1/2 \\ 1/3 & 0 & 0 & 0 \\ 1/3 & 1/2 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix}. \tag{5.2}$$

6. Proposed Penalty Matrix based Page Ranking Algorithm

We now present our penalty-based PageRank method and its illustration. 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 matrix that we propose is as follows:

$$P_{ij} = \text{Transpose of } A_{ij}, \tag{6.1}$$

$$A_{ij} = \begin{cases} 0.85 & \text{if } (i,j) \in E \text{ and } j \notin Ad(V), \\ 0.15 & \text{if } (i,j) \in E \text{ and } j \in Ad(V), \\ 0 & \text{otherwise} \end{cases}$$

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

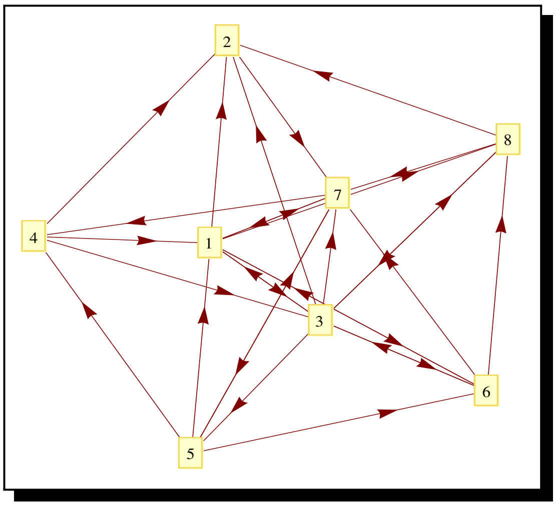

Figure 2. Web graph-2

Also, we can find the column stochastic Penalty matrix by dividing each entry of a column by the sum of the corresponding column and it is denoted by P. For the web graph in Figure 2, we consider 1, 3 and 8 web pages are advertisement web pages then the column stochastic Penalty matrix is

$$P = \begin{bmatrix} 0 & 0 & 0.0405 & 0.1304 & 0.0556 & 0.1154 & 0.0811 & 0 \\ 0.2982 & 0 & 0.2297 & 0.7391 & 0 & 0 & 0 & 0.4595 \\ 0.0526 & 0 & 0 & 0.1304 & 0 & 0.1154 & 0 & 0.0811 \\ 0 & 0 & 0 & 0 & 0.3148 & 0 & 0.4595 & 0 \\ 0 & 0 & 0.2297 & 0 & 0 & 0 & 0.4595 & 0 \\ 0.2982 & 0 & 0.2297 & 0 & 0.3148 & 0 & 0 & 0 \\ 0.2982 & 1 & 0.2297 & 0 & 0.3148 & 0.6538 & 0 & 0.4595 \\ 0.0526 & 0 & 0.0405 & 0 & 0 & 0.1154 & 0 & 0 \end{bmatrix}.$$

Penalty matrix based web ranking problem has been transformed into the "standard" linear algebra problem of finding an eigenvector corresponding to eigenvalue 1 for a square matrix P:

$$PX = X. \tag{6.2}$$

Components of the eigenvector give the importance score (rank) of the corresponding page.

After computing the column stochastic Penalty matrix we make use of equation (6.2) to compute the page rank of the web pages.

7. Comparison between Page Rank Algorithms

We have calculated the page rank for web pages in the web graph (Figure 2) using the original adjacency matrix-based iterative page rank algorithm [3], link matrix based original page rank algorithm [10] and our proposed Penalty matrix based page rank algorithm. The absolute values of the elements of the eigen vector of the Penalty matrix, corresponding to eigen value one, give the page rank values. Table 1 shows the page ranks computed for the web pages. The web pages are arranged in the table in decreasing order of page rank value. It can be noted that page rank scores for the advertisement pages 1, 3 and 8 are the lowest, as it must be, in our proposed algorithm.

Table 1. Page rank computations using iterative, link matrix based and our proposed penalty matrix based page rank algorithms

Adjacency matrix based iterative page rank algorithm	Proposed Penalty matrix based page rank algorithm
3	7
1	4
7	2
5	5
6	6
4	1
8	3
2	8

8. Conclusions and Future Work

This paper focuses on penalty based page rank score method for calculating the page ranks 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 matrix 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. Also we applied our Penalty based page rank algorithm for 1000 web pages web graph result shows that the page rank scores for the advertisement pages are the lowest, as it must be, in our proposed algorithm. As a future scope, we propose to analyze the performance of penalty based page rank methods in other web structure mining algorithms.

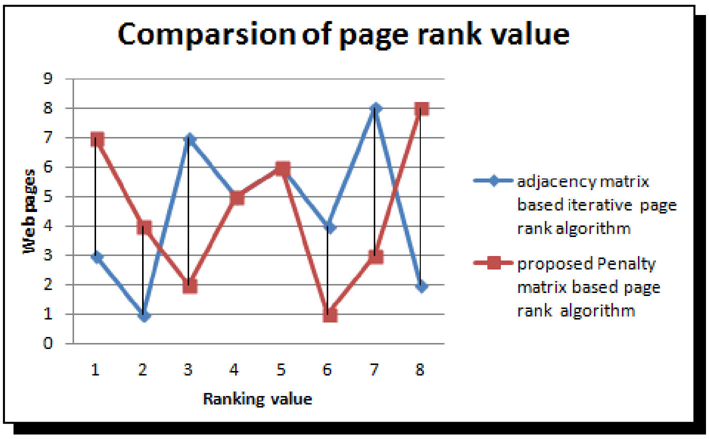

Figure 3. Page rank values obtained using the two algorithms

Competing Interests

The authors declare that they have no competing interests.

Authors' Contributions

All the authors contributed significantly in writing this article. The authors read and approved the final manuscript.

References

[1] A.N. Langville and C.D. Meyer, *Google's PageRank and Beyond, The Science of Search Engine Rankings*, Princeton University Press, Princeton (2006).

[2] F. Schneider, N. Blachman and E. Fredricksen, *How to Do Everything with Google?*, McGraw-Hill, New York (2003)

[3] L. Page, S. Brin, R. Motwani and T. Winograd, The page rank citation ranking: bringing order to the web, *Technical Report*, Stanford Digital Library Technologies Project (1998).

[4] W. Xing and G. Ali, Weighted pagerank alogithm, in *Proceedings of the Second Annual Conference on Communication Networks and Services Research* (CNSR'04), IEEE (2004).

[5] B. Jaganathan and K. Desikan, Category-based pagerank algorithm, *International Journal of Pure and Applied Mathematics* **101** (5) (2015), 811 – 820.

[6] B. Jaganathan and K. Desikan, Penalty-Based Pagerank Algorithm, *ARPN Journal of Engineering and Applied Sciences* **10** (5) (2015), 2000 – 2003.

[7] B. Jaganathan and K. Desikan, Weighted Pagerank Algorithm based on In-Out weight of webpages, *Indian Journal of Science and Technology* **8** (34) (2015), 1–6.

[8] B. Jaganathan and K. Desikan, Hermition matrix based Pagerank Algorithm, *Global Journal of Pure and Applied Mathematics* **12** (3) (2016), 277–280.

[9] M. Bressan AND E. Peserico, Choose the damping, choose the ranking?, *Journal of Discrete Algorithms* **8** (2010), 199 – 213.

[10] K. Bryan and T. Leise, *The $25,000,000,000 eigenvector: the linear algebra behind Google*, Society for Industrial and Applied Mathematics Philadelphia, PA, USA, Vol. **48** (3) (2006), 569 – 581.