

PAGERANK ALGORITHM BASED ON A NEW STOCHASTIC WEIGHT MATRIX

B.Jaganathan¹

¹Division of Mathematics, School of Advanced Sciences
VIT University
Vandalur-Kelambakkam Road, Chennai 600127, India
e-mail: jaganathan.b@vit.ac.in

August 9, 2017

Abstract

In this paper a new approach has been introduced by considering the web as a simple directed graph. We propose an improved method for the computation of page rank on the basis of weight assigned to web pages. Consequently, we have applied the concept of column-stochastic Weighted 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 column-stochastic Weight matrix.

AMS Subject Classification: 05C99, 68M10, 94C15

Key Words: PageRank, Eigen values, Eigen vector, Weight Matrix.

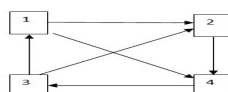
1 Introduction

In the world wide web, web page ranking plays an important role, and is very much needed to retrieve the relevant/informative pages from the vast collection of unordered information, based on the interest of the user. Though Search engines play an important

role in retrieving information for a given query, they cannot fully satisfy the users need for relevant information. 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.

2 Webgraph And Graph Theory Relationship

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) [1-4]. 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:



In graph theory, an adjacency matrix [1-3], 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 matrix with matrix elements being 1 for vertices (or nodes) which point to other vertices and 0 otherwise. This can be mathematically represented as:

$$a_{ij} = \begin{cases} 1 & \text{if } i \text{ pointing to } j \text{ and } i \neq j \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

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

3 Linear Algebra Behind Google Pagerank Algorithm

The linear algebra behind Google Pagerank algorithm was originally proposed by Kurt Bryan and Tanya Leise [4]. A brief explanation of linear algebra behind Google PageRank algorithm is given below. The objective behind this approach is to rank web pages according to their importance scores. A pages importance score is derived from the links made to that page from other web pages. Importance scores and there by the ranking of the web pages 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 web graph shown in Fig.1 In Fig.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 Fig.1 we get $X_1 = 1, X_2 = 2, X_3 = 1$ and $X_4 = 2$. This implies that pages 4 and 2 tie for the first position since it has the maximum number pages pointing to it. Pages 1 and 3 tie for the second position. This approach is very simplistic and it is purely based on the number of in-links that a web page 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. 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 [4] 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 that link (point) to page 1 $X_1 = X_3$ Similarly, we can write the remaining equations as follows $X_2 = X_1 + X_3$ $X_3 = X_4$ $X_4 = X_1 + X_2$

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 web pages. This is achieved by boosting page k 's score by $\frac{X_j}{n_j}$ if page j links to page k , where n_j is the number of out-links from page j Hence, for page 1 in Fig.1, we have $X_1 = \frac{X_3}{2}$ as page 3

and 1 contains two out-links (splitting its vote into half), while page 4 and 2 contains only one out-links. Similarly, for the remaining pages we have

$$X_2 = \frac{X_1}{2} + \frac{X_3}{2} \quad (2)$$

$$X_3 = \frac{X_4}{1} \quad (3)$$

$$X_4 = \frac{X_1}{2} + \frac{X_2}{1} \quad (4)$$

This system of linear equations can be written as $AX = X$ where

$$X = (X_1, X_2, X_3, X_4)$$

We see that the Web page ranking problem has been transformed into the "standard" linear algebra problem of finding an eigen vector corresponding to eigenvalue 1 for a square (link) matrix A. Here, the components of the eigen vector 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 non negative and the elements in each column sum to one. Also, we know that every column-stochastic matrix has 1 as an eigenvalue.

4 Proposed Weight Matrix Based Page Ranking Algorithm

Let $G = (V, E, W)$ be a weighted directed network (or weighted directed graph) with set of edges E where each edge $e = (v_i, v_j)$ is attached with a weight w_{ij} . If there are no directed edges between vertex i and vertex j , then $w_{ij} = 0$. Since we consider directed network/web graph without loops, we have

- (i) $w_{ii} = 0$.
- (ii) $w_{ij} = w_{ij}^{in} + w_{ij}^{out}$

where

$$w_{ji}^{in} = \frac{I_i}{\sum_{p \in R(j)} I_p}$$

where I_i and I_p represent the number of inlinks of page i and page p , respectively. $R(j)$ denotes the reference page list of page j .

$$w_{ji}^{out} = \frac{O_i}{\sum_{p \in R(j)} O_p}$$

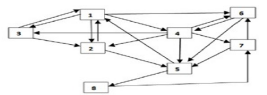
where O_i and O_p represent the number of out links of page i and page p , respectively. $R(j)$ denotes the reference page list of page j . Reference page of j means a page which is referring or pointing to page j . We define

$$W(G) = \begin{pmatrix} 0 & w_{12} & \cdots & w_{1n} \\ w_{21} & 0 & \cdots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \cdots & 0 \end{pmatrix}$$

Also, we can find the column stochastic Weight matrix by dividing each entry of a column by the sum of the corresponding column and it is denoted by $W_s(G)$. Stochastic Weight matrix based web ranking problem has been transformed into the "standard" linear algebra problem of finding an eigenvector corresponding to Eigen value 1 for a square matrix $W_s(G)$

$$W_s(G)X = X \quad (5)$$

Components of the eigenvector give the importance score (rank) of the corresponding page. After computing the column stochastic Weight matrix we make use of equation (5) to compute the page rank of the web pages.



5 Conclusions And Future Work

This paper focuses on Stochastic Weight 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 Weight matrix based page rank methods are better than the existing page rank method. This is due to the fact that in the weight matrix method we consider both in weight and out weight (Popularity) of the web pages but in adjacent matrix method only in weight is considered. Our finding suggests that the Weight matrix based page rank method boosts up the page ranks of the most relevant pages and pulls down the page ranks of irrelevant pages. Also we applied our Weight matrix based page rank algorithm to a 1000 web pages web graph and the result shows that the page rank scores given based on in and out weight of web pages are similar to the original page rank algorithm rank score. As a future scope, we propose to analyze the performance of Weight matrix based page rank methods to other web structure mining algorithms.

References

- [1] Page, L., Brin, S., Motwani, R., Winograd. T., The Page Rank Citation Ranking: Bringing Order to the Web, *Technical report, stanford Digital Library Technologies Project*, (1998).
- [2] Weapu Xing, Ghorbani Ali, Weighted PageRank Alogithm, Proceedings of the Second Annual Conference on Communication Networks and Services Research (CNSR04),IEEE, (2004).
- [3] Jaganathan.B.,Kalyani Desikan, Category-Based Pagerank Algorithm, *International Journal of Pure and Applied Mathematics*, **101** No. 5 (2015), 811-820.
- [4] Kurt Bryan,Tanya Leise, The \$25,000,000,000 Eigenvector: The linear algebra behind google, *Society for Industrial and Applied Mathematics Philadelphia, PA, USA*, **48** No. 3 (2006), 569-581.

