International Journal of Engineering & Technology, 7 (4.10) (2018) 566-569

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Enhanced Web Page Ranking Method Using Laplacian Centrality

B Jaganathan[1]*, Kalyani Desikan[2]

[1,2] Division of Mathematics, School of Advanced Sciences
Vellore Institute of Technology
Vandalur-Kelambakkam Road, Chennai 600127,India
**Corresponding author E-mail: jaganathan.b@vit.ac.in*

Abstract

In today's era of computer technology where users want not only the most relevant data but they also want the data as quickly as possible. Hence, ranking web pages becomes a crucial task. The purpose of this research is to find a centrality measure that can be used in place of original page rank. In this article concept of Laplacian centrality measure for directed web graph has been introduced to identify the web page ranks. Comparison between the original page rank and Laplacian centrality based Page rank has been made. Kendall's τ correlation co-efficient has been used as a measure to find the correlation between the original page rank and Laplacian centrality measure based page rank.

Keywords: *Centrality Measures; Laplacian centrality; PageRank; Web graph.*

1. Introduction

Whenever a user enters a query, he/she expects the appropriate web pages to be returned by the search engine. Web Page ranking ensures this and enables the display of better ranked pages that have better content, quality and relevance, .on the top Also, with the constantly increasing size of the web, it becomes harder to find quality content. Page Ranking ensures that irrelevant pages which users rarely visit do not clutter the search results. In this paper, we propose a new centrality strategy for weighted directed networks (web graph) which permits one to consider more "intermediate" environmental information around a vertex. This new centrality strategy gives a different perspective to the estimation of the web page ranks. This strategy is called "Laplacian centrality" method for directed web graph/network because it employs a matrix valued function that describes the so called "Laplacian energy" of the directed network. The basic idea is that the importance (centrality) of a vertex/web page is related to the ability of the network to respond to the deactivation of a vertex/web page from the directed network. This is used as the indicator to show the importance of a web page in the network. Laplacian centrality for undirected graph was introduced by Xingqin Qi et.al [7].

This paper is organized as follows. In section 1 we present the definition of various centrality measures of a vertex. In section 2 we present the definition of Laplacian energy and Laplacian centrality of a vertex. Also, we present an illustration to show a structures description of Laplacian centrality. Analytical, numerical and graphical results based on Laplacian centrality measures applied to a directed web graph are given in section 3.

1.1 Graph Centrality/Centrality Measures

Centrality measures are computed to assign a score to each node/web page. Let *G = (V,E)* be a graph with a set of vertices/nodes *V* and a set of directed edges *E*. Centrality Measures are used to find the most important vertices within a graph, to assign important score to each node/web page and it is used in social networking analysis. Various centrality measures have been proposed for weighted directed networks. Some of the centrality measures are degree centrality, closeness centrality, betweenness centrality and eigen vector centrality. Starting with degrees centrality, we will discuss these centrality measures in this section.

1.2 Degree Centrality

Degree centrality is a simple centrality measure that counts the number of neighbors of a node. In the case of directed networks, we have: in-degree and out-degree. In-degree refers to the number of in-coming links, or the number of predecessor nodes while out-degree refers to the number of out-going links, or the number of successor nodes. In general, in-degree is of interest as in-links are determined by other nodes in the network, while out-degree is dependent on the node alone.

In the case of undirected network, if a node has many neighbors, it is considered to be important. In the case of directed network the importance of a node is based on the number of nodes that link to it or the number to nodes to which the node is linked.

Degree centrality is determined by the number of edges that are incident upon a node/vertex. It is based on in-links and out-links. Applied to a web graph, the degree centrality is classified into two:

- In degree centrality
- Out degree centrality

The In degree centrality of the vertex v_i is given by

$$C_D(v_i) = \frac{\deg^{in}(v_i)}{n} \tag{1}$$

where $\deg^{in}(v_i)$ is the in-degree of vertex v_i and n is the total number of directed edges in *G*.
The out degree centrality of the vertex v_i is defined as

$$C_d(v_i) = \frac{\deg^{out}(v_i)}{n} \tag{2}$$

where $\deg^{out}(v_i)$ is the out-degree of vertex v_i and n is the total number of directed edges in *G*.

1.3 Betweenness Centrality

It is a measure of importance of a vertex within a graph. Betweenness centrality measures the number of times the node acts as a bridge between two other nodes along the shortest path between them. This measure was introduced by Linton Freeman[5]. It was used to measure the control of a human on the communication between other humans in a social network.

Betweenness centrality determines the number of times a vertex acts as a bridge along the shortest path between two other nodes/vertices. Let $\sigma(v_j,v_{k)}$ be the number of shortest paths from node v_j to node v_k and $\sigma(v_j,v_k|v_i)$ the number of those paths that pass through node v_i. The betweenness centrality of the vertex v_i is given by

$$C_B(v_i) = \frac{\sum \frac{\sigma(v_j, v_k \mid v_i)}{\sigma(v_j, v_k)}}{(n-1)(n-2)/2} \tag{3}$$

where *n* is the total number of vertices in directed graph *G*.
Betweenness centrality indicates the extent to which a vertex lies on the shortest paths between other vertices. Within a network, vertices with high betweenness may have considerable influence because of the control that they have over information passing between other vertices. In a communication network, the removal of vertices with high betweenness centrality from a network will cause the maximum disruption in communications between other vertices as they lie on the largest number of paths that are taken by messages in the network.

1.4 Closeness Centrality

The closeness centrality of a directed graph is defined as the inverse of farness, that is, sum of shortest distances between a node/vertex and all the other nodes/vertices. The closeness centrality of a node/vertex *vi* is given by

$$C_C(v_i) = \frac{n-1}{\sum_j d_{ij}} \tag{4}$$

where d_{ij} is the shortest distance between vertices v_i and v_j and n is the total number of vertices in G. Closeness centrality gives the mean distance from a vertex to other vertices. Closeness centrality differs from degree and eigenvector centrality. For instance, consider a node A in a network that is connected to a single node B. If node B is very close to the other nodes in the network, its closeness score will be high. It automatically follows that node A would also have a relatively large closeness score, as we can reach all the nodes that B reaches in one additional step from A. However, the degree of A is only 1 and hence, its degree centrality score will be low and its eigenvector score may not be impressive.
The above mentioned measures are defined for weighted directed networks.
We now look at the Kendall Tau rank correlation coefficient that we have used to compare the ranking of the web pages.

1.5 Kendall Rank Correlation Coefficient

The Kendall tau rank correlation coefficient measures the extent to which two sets of ranks given to the same set of objects are similar. This coefficient depends on the number of inversions that are required to convert one rank order into the other. For performing the inversions, the rank orders are considered in pairs and a value of 1 or 0 is assigned to this pair depending on whether they correspond or not. This provides a set of binary values that are then used to compute the Pearson correlation coefficient.

Kendall’s tau correlation coefficient is a correlation measure that measures the strength of the relationship between two variables X and Y that are paired observations. Kendall’s tau, like Spearman’s rank correlation Crichton [2] is computed based on the ranks assigned to the observations. The values for each variable are ordered separately and ranked. Conover [1] details how to compute Kendall’s tau correlation coefficient. Like the other measures of correlation, Kendall’s tau values lie between −1 and +1. The correlation is positive if the ranks of both variables increase together while the correlation is negative if the rank of one variable increases as the other decreases.

Similar to the Spearman’s rank correlation []we can calculate confidence intervals and perform hypothesis tests on Kendall’s tau correlation coefficient. Since Spearman’s rank correlation coefficient is much easier to compute than Kendall’s tau, it is a more widely used measure of rank correlation. The key advantage of using Kendall’s tau correlation coefficient is that the statistical properties of the distribution of this statistic is slightly better and a direct interpretation of Kendall’s tau can be given in terms of probabilities of observing concordant and discordant pairs [2]. Most often the values of Spearman’s rank correlation and Kendall’s tau are very close and would eventually lead to the same conclusions. The following formula is used to calculate the value of Kendall’s rank correlation. Kendall’s rank correlation coefficient is given by

$$\tau = \frac{N_c - N_d}{\frac{1}{2}n(n-1)}$$

where N_c is the number of concordant entries, N_d is the number of discordant entries and n is the number of observations. Concordant indicates that the objects are ordered in the same way and discordant implies that they are ordered differently.

2. Laplacian Centrality

In this section we discuss the Laplacian centrality measure for directed graphs.

2.1 Laplacian Energy for a Network/Web Graph

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 assigned a weight w_{ij}. If there are no directed edges between vertex i and vertex j, then $w_{ij} = 0$. For directed networks/web graphs without loops, we have

(i) $w_{ii} = 0$

(ii) $w_{ij} = w_{ij}^{in} + w_{ij}^{out}$

Where $w_{ij}^{in} = \dfrac{I_j}{\sum_{p \in R(i)} I_p}$

where I_i and I_p represent the number of in-links of pages *i* and *p*, respectively. *R(i)* denotes the reference page list of page *j*.

$$w_{ij}^{out} = \frac{O_j}{\sum_{p \in R(i)} O_p}$$

where O_i and O_p represent the number of out-links of pages i and p, respectively. R(i) denotes the reference page list of page j. Reference page of j means a page which is referring or pointing to page j[3]. We define

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

$$X(G) = \begin{pmatrix} X_1 & 0 & \cdots & 0 \\ 0 & X_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & X_n \end{pmatrix}$$

with

$$X_i = \sum_{j=1}^{n} w_{ij}$$

X_i is referred to as the sum-weight of vertex v_i.

2.2 Laplacian Matrix

The matrix L (G) = X (G) − W (G) is called the Laplacian matrix of the weighted directed graph G.

2.3 Laplacian Energy

Let *G = (V, E, W)* be a weighted directed network on n vertices and $\lambda_1, \lambda_2, ...\lambda_n$ be the eigen values of its Laplacian matrix.

The Laplacian energy of *G* is an invariant that is given by:

$$E_L(G) = \sum_{i=1}^{n} \lambda_i^2 \tag{5}$$

We now illustrate the computation of the Laplacian centrality based web page ranks in the next section.

3. Illustration

Let us consider a 5 node web graph,

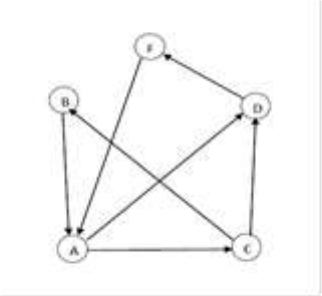

Fig. 1: Directed network/web graph

For Fig. 1 the adjacency, weighted, weighted sum diagonal and Laplacian matrices are given below:

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

$$w(G) = \begin{pmatrix} 0 & 0 & 1 & 3/2 & 0 \\ 3 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 7/12 \\ 3 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$X(G) = \begin{pmatrix} 5/2 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 7/12 & 0 \\ 0 & 0 & 0 & 0 & 3 \end{pmatrix}$$

$$L(G) = \begin{pmatrix} 5/2 & 0 & -1 & -3/2 & 0 \\ -3 & 3 & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & 0 & 7/12 & -7/12 \\ -3 & 0 & 0 & 0 & 3 \end{pmatrix}$$

We now show how to compute the Laplacian centrality for each node of the given web graph.

Laplacian centrality for web pages *A, B, C, D* and *E* of Fig 1 is calculated using below equation.

For example, for page A we have

$$C_L(A,G) = 1 - \frac{E_L(G_A)}{E_L(G)}$$

Where

$$E_L(G) = \sum_{i=1}^{n} \lambda_i^2$$

G_A is obtained from *G* by deleting the web page *A*, that is the row and column corresponding to web page *A* are deleted from *L(G)*, and we get

$$L(G_A) = \begin{pmatrix} 3 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & 0 & 7/12 & -7/12 \\ 0 & 0 & 0 & 3 \end{pmatrix}$$

We compute $E_L(G_A)$ by summing the eigenvalues of $L(G_A)$. For the graph given in Fig. 1 we have $E_L(G_A) = 22.3403$ and $E_L(G) = 28.5903$ and consequently, we get $C_{L(}A,G) = 0.2186$.
Similarly, we calculate the Laplacian centrality for all other web pages (nodes). We rank the web pages based on their Laplacian centrality scores.
For the web graph given in Fig 1, we calculated the page rank for the web pages of using the original Page rank algorithm and our Laplacian centrality based page rank algorithm. The page ranks computed for the web pages using the page rank algorithm and Laplacian centrality based page rank algorithm are given in Table 1. Laplacian centrality based page ranks are assigned based on the Laplacian centrality values. The ranks are assigned such that the web page with the least Laplacian centrality value is assigned rank 1 and so on.

Table 1: Page ranks using original PageRank algorithm and Laplacian Centrality based ranks

Web pages	Original Page rank scores	Page rank	Laplacian Centrality scores	Laplacian Centrality based page rank
A	0.300129	1	0.2186	2
B	0.096961	5	0.3148	5
C	0.15755	4	0.1399	4
D	0.224516	2	0.0119	1
E	0.220839	3	0.3148	3

For Table 1, we see that the Kendall’s τ correlation co-efficient between the original page rank and Laplacian centrality based page rank is +0.8. Here the number of concordant entries N_c is 9, the number of discordant entries N_d is 1 and the number of observations is 5

4. Conclusion

In this article a new approach based on the Laplacian centrality measure for directed graphs has been applied to rank web pages. The Laplacian centrality based page rank is easy to calculate and effective because it avoids the iteration process. Hence, the computational complexity is reduced. From the Kendall Tau correlation values for the web graph that we have presented to illustrate the working of our method, we see that our ranking method gives a different ranking compared to the original page rank method for some networks. This has to be further explored by considering larger networks.

References

[1] Conover W.J (1980), Practical Non-Parametric Statistics, 2nd edn., *John Wiley and Sons*, New York.

[2] Crichton N.J (1999) Information point: Spearman's rank correlation, *Journal of Clinical Nursing* 8,763.

[3] Jaganathan.B.,Kalyani Desikan (2015), Weighted Page Rank Algorithm Based on In-Out Weight of Webpages, *International Journal of Science and Technology*, 8 No. 34, 1-6.

[4] Kurt Bryan,Tanya Leise (2006), The $25,000,000,000 Eigenvector: The linear algebra behind google, *Society for Industrial and Applied Mathematics Philadelphia*, PA, USA, 48 No. 3, 569-581.

[5] Linton C. Freeman (2004), The Development of Social Network Analysis: A Study in the Sociology of Science,*Social Networks*, Empirical Press, Vancouver, BC, 27, 377–384.

[6] 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*, 1–17.

[7] Xingqin Qi , Eddie Fuller,Qin Wu,Yezhou wu, Cun-Quan Zhang (2012), Laplacian centrality: A new centrality measure for weighted networks, *Journal of Information Sciences*, 194, 240-253.