Global Journal of Pure and Applied Mathematics.
ISSN 0973-1768 Volume 13, Number 10 (2017), pp. 7229-7235

http://www.ripublication.com

New web page rank method using HITS Centrality

N. Punithavelan
Division of Physics, School of Advanced Sciences
VIT University, Chennai-600127, India.

B. Jaganathan
Division of Mathematics, School of Advanced Sciences
VIT University, Chennai-600127, India.

Abstract

Internet is the world's richest and most dense source of information. Users face the problem of getting the most relevant and useful information from large collection of disordered information user wants not only the most relevant data but they also want the data as quickly as possible. World Wide Web (WWW) is expanding rapidly and thus there is a necessity of ranking these web pages. Hence role of ranking web pages is crucial. In this paper we give a brief overview of the HITS algorithm proposed by Jon Kleinberg. We proposed new page rank method for web page ranking using hubs and authority based matrix functions. Brief analyses of our page rank method are also discussed.

Keywords: web page rank; HITS algorithm; matrix function; matrices; centralites.

I. INTRODUCTION

Webpage ranking helps us a lot when we need to search on a particular matter as it gives us the most popular and important pages and saves our time in opening all the webpages related to our matter. It ranks the webpages using Page Rank algorithm in which the webpages with top ranks seems to have more traffic than the ones with ranks below, that is more people view these websites which means they are updated

regularly and contain most of the matters related to the subject.

Search Engine Ranking is the position at which a particular site appears in the results of a search engine query. Each page of the search results typically lists 10 websites, although they are sometimes augmented with local listings, videos and images. A site is said to have a high ranking when it appears at or near the top of the list of results. Thus a higher ranking corresponds to a lower number. Search engine ranking is influenced by a multitude of factors including age of site, the quality of a site's link portfolio, relevancy of the page, social signals and level of competition, among others. For example, Google admits to using 200 factors when determining a site's search engine ranking. Search engines rank individual pages of a website, not the entire site. This means that the homepage might rank number 1 for certain keywords, while a deep internal page might be listed on the third page.

The structure of this paper is as follows: section II deals with the important of ranking algorithms. In section III the relationship between web pages and web graph is presented. In section IV, the HITS based page rank algorithm is presented.

In section V, we present our proposed new web page rank method using hubs and authority based matrix functions and its illustration. In section VI the conclusion and possible future works are presented.

II. WEB PAGE RANKING

Ranking is a key element to any SEO strategy and cannot be understated. Especially on dominant search engines with billions of searches every day. Therefore knowing where your site ranks on the biggest search engines in cyberspace is essential when researching your current standing and developing a strong strategy for the future. Research show that 40% likely to click on your site if you are ranked 1 by google, 30% more likely to click if you are ranked 2 and 24% more likely if you are ranked 3. It is very easy for more experienced surfers to type in keywords or a business name and see the results that come back, however if your aren't ranking at the top or even on the first page of results then you are going to miss out on traffic, site hits and most importantly potential business. So you need to devise a plan of action to increase your ranking. Knowing your websites ranking position is just as important as the content on your page because if customers cannot find you then the information on your site or the great service you offer, is unattainable and effectively worthless.

III. RELATION BETWEEN WEB GRAPH AND WEB PAGES

A directed graph or diagraph G=(V,E) is formed by a set of vertices V and edges E formed by ordered pairs of vertices. That is $(i, j) \in E$ does not implies $(j, i) \in E$. In the

case of diagraphs, which model directed networks. There are two types of degree, the in-degree of node is given by the number of edges which point to i. The out-degree is given by the number of edges pointing out from i.

A walk is a sequence of vertices $v_1, v_2, ...v_k$ such that for $1 \le i < k$ there is an edge between v_i and v_{i+1} (a directed edge from v_i to v_{i+1} in the case of diagraph) vertices and edges may be repeated. A walk is closed if $v_1 = v_k$. A path is walk consisting only of distinct vertices.

A graph G is connected if each pair of vertices is linked by a pair in G. A diagraph is strongly connected if for any pair of vertices v_i and v_k. There is a walk starting at v_i and ending at v_k. A diagraph is weakly connected it the graph obtained by disregarding the orientation of its edges is connected unless otherwise specified evenly diagraph in this paper is simple (un-weighted with no multiple edges or loops and connected). Note, however that most of the techniques and results in the paper can be extended without difficulty to more general diagraphs.

The adjacency matrix of a graph is a matrix $A \in R^{|v|\times|v|}$ defined In the following way:

$$A = (a_{ij}), a_{ij} = \begin{cases} 1, f(i,j) \text{ is an edge} \\ 0, otherwise \end{cases} \qquad (1)$$

under the conditions imposed on *G*. *A* has zeros on the diagonal if G is an undirected graph. *A* will be a symmetric matrix and the eigenvalues will be real.

IV. HITS ALGORITHM

HITS was proposed by Jon Kleinberg who was a young scientist at IBM in Silicon Valley and now a professor at Cornell University

Each node i in the network is assigned two non-negative weights an authority weight x_i and sub weight y_i. To begin with each x_i and y_i is given an arbitrary non-zero value. Then the weights are updated in the following ways:

$$x_i^{(k)} = \sum_{j:(j,i)\in E} y_j^{(k-1)} \text{ and } y_i^{(k)} = \sum_{j:(j,i)\in E} x_j^{(k)} \qquad (2)$$

for $k = 1, 2, 3, ...$

The weights are then normalized so that $\sum_j (x_j^{(k)})^2 = 1$ and $\sum_j (y_j^{(k)})^2 = 1$.

The above iterations occur sequentially and it can be shown that under mild conditions both sequences of vector $[x^{(k)}]$ and $[y^{(k)}]$ converge as $k \to \infty$. In practice, the iterative process is continued until there is no significant change between

consecutive iterates.

This iteration sequence shows the natural dependence relationship between hubs and authorities if a node 'i'points to many nodes with large x-values, it receives a large y-values and if it is pointed to by many nodes with large y –values it receives a large x-value .

In terms of matrices the above equation (2) becomes

$$x^{(k)} = A^T y^{(k-1)} \quad and \quad y^{(k)} = Ax^{(k)} \tag{3}$$

Followed by normalization in the 2-norm. this iterative process can be expressed as

$$x^{(k)} = c_1 A^T A x^{(k-1)} \quad and \quad y^{(k)} = c_2 A A^T y^{(k-1)} \tag{4}$$

Where $c_1$and c_2 are normalization factors A typical choice for the initialization vectors $x^{(0)}$, $y^{(0)}$ would be constant vector

$$x^{(0)} = y^{(0)} = [\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}} ...., \frac{1}{\sqrt{n}}]$$

Hence HITS is just an iterative power method to compute the dominant eigen vector for A^TA and for AA^T. The authority scores are determined by the entries of the dominant eigen vector of the matrix A^TA. Which is called the authority matrix and the hub scores are determined by the entries of the dominant eigenvector of AA^T called the hub matrix.

V. PROPOSED NEW WEB PAGE RANK METHOD USING HUBS AND AUTHORITY BASED MATRIX FUNCTIONS

We propose a new web page rank method using hubs and authority based matrix functions. In this method we first calculate the new matrices using hub and authority matrices.

$$\zeta = \begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix} \tag{5}$$

$$\zeta' = transpose\ of\ \zeta \tag{6}$$

We calculate the matrix function using the equation (5) and (6)

$$Hub\ score\ matrix\ function = e^{\zeta} \tag{7}$$

$$Authority\ score\ matrix\ function = e^{\zeta'} \tag{8}$$

We make use of the above matrix function given in equation (7) and (8) to solve the equation (9) and (10) we get an page rank values.

$$[I - c_1\zeta]X = 1 \quad (9)$$

$$[I - c_2\zeta']Y = 1 \quad (10)$$

Where $c_1 = \frac{1}{\rho + 0.1}$ and $c_2 = \frac{1}{\eta + 0.1}$

where, ρ *and* η are spectrums of ζ *and* ζ' respectively.

Both X and Y are giving the hub and authority scores of the web pages. We calculate the hub and authority scores using the equation (11) and (12).

$$Hub\ score\ value = [X]_{ii}\ or [Y]_{ii} \quad (11)$$

$$Authority\ score\ value = [X]_{n+i,n+i}\ or [Y]_{n+i,n+i} \quad (12)$$

A. Experiments

We now explain the working of the proposed new web page rank method using hubs and authority based matrix functions by considering the hyperlinked web graph in Fig.1

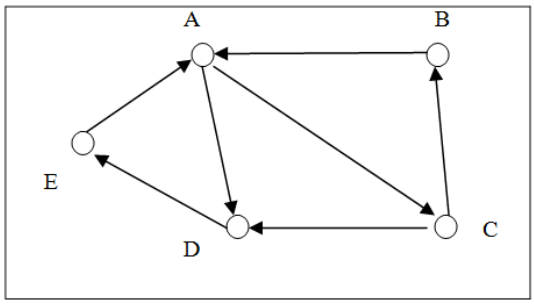

Fig.1. Web graph

The tables given below show the page ranks computed for the web pages using the hub and authority scores.

Table 1: Pagerank computations using our proposed Page Rank method using hubs and authority values for Web Graph in Fig.1

Web pages	**Web page ranking**			
	Hub score value	***Hub score based Ranking values***	***Authority score value***	***Authority score based Ranking values***
A	3.1583	1	4.0139	2
B	2.5070	4	0.1857	4
C	3.1583	2	0.1857	5
D	1.6011	5	4.9451	1
E	2.5070	3	1.6011	3

VI. CONCLUSION

From the previous section on new web page rank method using hubs and authority based matrix functions. We notice that our method is more efficient when compared to the HITS algorithm with respect to hub and authority. This due to fact in our proposed method both hub and authority matrices are involved in the matrix function. It can also be seen that the ranking of the webpages using our method no iterative techniques involved hence our proposed method more efficient than HITS method with respect to time.

In our future work, based on this method, we envisage to work with bigger web graphs. We also propose to introduce other HITS centrality techniques for calculating the ranks of web pages.

a.

REFERENCES

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

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

[3] 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.

[4] Weapu Xing,Ghorbani Ali,"Weighted PageRank Alogithm", Proceedings of the Second Annual Conference on Communication Networks and Services Research (CNSR'04),IEEE,2004.

[5] B.Jaganathan,Kalyani Desikan, "Category-Based Pagerank Algorithm," International Journal of Pure and Applied Mathematics., vol.101,No.5, pp. 811-820, August 2015.

[6] B.Jaganathan,Kalyani Desikan, "Penalty -Based Pagerank Algorithm," ARPN Journal of Engineering and Applied Sciences, vol.10,No.5, pp. 2000-2003, March 2015.

[7] B.Jaganathan,Kalyani Desikan, "Weighted Pagerank Algorithm based on In-Out weight of webpages," Indian Journal of Science and Technology, vol.8,No.34, pp. 1-6, December 2015.

[8] B.Jaganathan,Kalyani Desikan, "Hermition matrix based Pagerank Algorithm," Global Journal of Pure and Applied Mathematics, July 2016.

[9] Marco Bressan., Enoch Peserico.,"Choose the damping, choose the ranking? "Journal of Discrete Algorithms vol.8 pp.199–213, 2010.

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

[11] D. Sepandar, H. Kamvar Taher, Haveliwalla Christopher, D. Manning Gene and H. Golub, "Exploiting the Block structure of the web for computing Page Rank", Stanford University Technical Report.

[12] S. Brin and C. Page, "The Anatomy of a large scale Hypertentud Web Search Engine", Computer Network and ISDN Systems, Vol. 30, Issue 1-7, pp 107-117, 1998.