Journal of Informatics and Mathematical Sciences
Vol. 9, No. 3, pp. 529–538, 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

Web Page Ranking Algorithms Through Centrality Measures

Saroj Kumar Dash* and B. Jaganathan

Department of Mathematics, School of Advanced Sciences, Vellore Institute of Technology University, Chennai, Tamil Nadu, India
***Corresponding author:** sarojkumar.dash@vit.ac.in

Abstract. In this paper we give a brief overview of the adjacency matrix based page rank that we used in the Google search engine. In this paper various centrality measures discussed like in-degree centrality, out-degree centrality, degree centrality and Eigen centrality. We have applied the above mentioned centrality measures to web page ranking. This approach does not involve any iterative technique. This centrality measures are better than original iterative based page rank algorithm for ranking the web pages.

Keywords. In-degree; Out-degree; Eigen centrality; Page rank

MSC. 65YF15; 65F30; 68Q30

Received: January 27, 2017 **Accepted:** March 4, 2017

1. Introduction

With the huge number of web pages that exist today, search engines assume an important role in the current Internet. But even if they allow finding relevant pages for any search topic, now a days the number of results returned is often too big to be carefully explored. Moreover, the needs of the users vary, so that what may be interesting for another. The role of ranking algorithm is the crucial: select the pages that are most likely be able to satisfy the user's needs and bring them in the top positions.

The analysis of cross-referencing patterns-link analysis has come to play an important role in modern information retrieval [2, 7]. Link analysis algorithms have been successfully applied to web hyperlink data to identify authoritative information sources.

The *World Wide Web* (WWW) is growing tremendously on all aspects and is a massive, explosive, diverse, dynamic and mostly untrusted data repository. As on today, www is the largest repository for knowledge reference [2, 7, 19]. There are a lot of challenges in the web: web is huge, web pages are semi structured, and web information tends to be diversity in meaning degree of quality of the information extracted and the conclusion of the knowledge from the extracted information.

A Google report on 25th July, 2008 says that there are one trillion (1,000,000,000,000) unique URLs (*Universal Resource Locators*) on the web. The actual number could be more than that and Google could not index all the pages. When Google first created the index in 2000, Google index reached 1 billion pages. In the last nine years, web has grown tremendously and the usage of the web is unimaginable so it is important to understand and analyze the underlying data structure of the web for effective Information Retrieval [2, 7, 19].

With the rapid growth of the *World Wide Web* (WWW) and the user's demand on knowledge, it is becoming more difficult to manage the information on www and satisfy the user's needs. Therefore, the users are looking for better information retrieved techniques and tools to locate extract, filter and find the necessary information. Most of the users use information retrieval tools like search engines to find information from the WWW. There are tens and hundreds of search engines available but some are popular like Google, Yahoo, Bing, etc., because of their crawling and ranking methodologies.

The search engines download, index and store hundreds of millions of webpages. They answer tens of millions of queries every day. So web mining and ranking mechanism becomes very important for effective information retrieval [16].

2. Link Analysis

The goal of information retrieval is to find all documents relevant for a query in a collection of documents. Decades of researches in information retrieval were successful in developing and refining techniques that are solely word-based. With the advent of web new sources of

information become available, one of them being the hyperlinks [8, 19] between documents and records of user behavior. To be precise, hypertexts (i.e. collections of documents connected by hyperlinks) have existed and have been studied for a long time. What was new was the large number of information for web information retrieval as we will show in this article. This area of information retrieval is commonly called link analysis [3, 5, 13, 19].

Why would one expect hyperlinks to be useful? A hyperlink is a reference of a web page that is contained in a web page. When the hyperlink is clicked-on in a web browser, the browser displays page. This functionality alone is not helpful for web information retrieval. However, the way hyperlinks are typically used by authors of web pages can give them valuable information contents. Typically, authors create links because they think that will be useful for the readers of the pages. Thus, links are usually either navigational aids that, for example, bring the readers back to the homepage of the site, or links that point to pages whose content augments the content of the current page. The second kinds of links tend to print to pages whose content augments the content of the current page. The second kinds of links tend to point to high quality pages that might be on the same topic as the page containing the link. Based on this motivation, link analysis makes the following simplifying assumptions.

(i) A link from page A to page B is a recommendation of page B by the author of page A.

(ii) If page A and page B are connected by a link the probability that they are on the same topic is higher than if they are not connected.

Link analysis has been used successfully for deciding which web pages to add to the collection of documents (i.e. which pages to crawl), and how to order the documents matching a user query (i.e. how to rank pages). It has also been used to categorize web pages, to find pages that are related to given pages, to find duplicated web sites and various other problems related to web information retrieval.

2.1 A Graph Representation for the Web

In order to simplify the description of the algorithms below, we first model the web as a graph. This can be done in various ways connectivity based ranking techniques usually assume the most straight forward representation. The graph contains a node for each page u and there exists a directed edge (u,v) if and only if page u contains a hyperlink to page v. We call this directed graph the link graph (web graph) G. Some algorithms make use of the undirected co-citation.

2.2 Web Page Rank Algorithms Need and Importance

With the increasing number of web pages and users on the web. The number of queries submitted to the search engines are also increasing rapidly. Therefore, the search engines needs to be more efficient in its process. Web mining techniques are employed by the search engines to extract relevant documents from the web database and provide the necessary information to

the users. The search engines [1] become very successful and powerful if they use efficient ranking mechanism. Google search engine is very successful because of its page rank algorithm. Page ranking algorithms are used by the search engines to present. The search results by considering the relevance, importance and content score and web mining techniques to order them according to the user interest. Some ranking algorithms depend only on the link structure of the documents i.e. their popularity scores (web structure mining) where as others look for the actual content in the documents while some combination of both. That is they use content of the document as well as the link structure to assign a rank value for a given document. If the search results are not displayed according to the user's interest then the search engine will lose its popularity. So the ranking algorithms became very important [4, 6, 17, 18].

2.3 Page Rank Algorithm

S. Brin and L. Page proposed an algorithm called page rank algorithm [9, 14, 15]. Page rank algorithm is used by the famous search engine, Google. They applied the citation analysis in web search by treating the incoming links as citations to the web pages. However by simply applying the citation analysis techniques to the diverse set of web documents did not result in efficient outcomes. Therefore, page rank provides a more advanced way to compute the importance or relevance of a web page than simply counting the number of page that are linking to it (called backlinks). If a backlink comes from an "important" page then that backlink is given a higher weighting than those backlinks come from non-important pages. Simple way, link from one page to another page may be considered a vote. However, not only the number of votes a page receives is considered important, but the "importance" or the "relevance" of the ones that last these votes as well.

2.4 Mathematical form of Page Rank Algorithm

Assume any arbitrary page A has pages T_1 to T_n pointing to it (incoming link) page rank can be calculated by the following:

$$PR(A) = (1-d) + d\left[\frac{PR(T_1)}{C(T_1)} + \ldots + \frac{PR(T_n)}{C(T_n)}\right].$$

The parameter d is a damping factor, usually sets it to 0.85 the other pages having too much influence this total vote is "damped down" by multiplying it by 0.85. $C(A)$ is defined as the number of links going out of page A. The page ranks from a probability distribution over the web pages, so the sum of all web pages, Page rank will be one. Page rank can be calculated using a simple iterative algorithm and correspond to the principal eigenvector of the normalized link matrix of the web.

2.5 Zero-One Gap Problem in Page Rank Algorithm

The basic page rank assumes each row of the matrix M has at least one non-zero entry, i.e. corresponding node in G has at least one out-link. But in reality it does not true. Many web

pages does not have any out-links and many web applications only consider a sub graph of the whole web even if a page has out-linked it might web is projected to a sub graph. Removing all the pages without out-links is not a solution, because it generates new zero out-link pages. This dangling page problem has been described by Brin, Page and Bianchini et al. [9, 14, 15]. The probability of jumping to random pages is 1 in zero out-link page, but it drops to d (in most cases $d = 0.15$) for a page with a single out-link. This problem is referred as "zero-one gap".

3. Various Methods of Calculation Page Rank

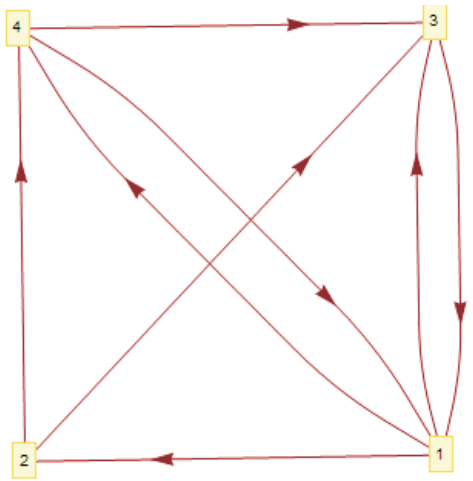

Figure 1. Web Graph-1

Suppose that initially the importance is uniformly distributed among the 4 nodes, each getting $\frac{1}{4}$. Denote by v the initial rank vector, having all entries equal to $\frac{1}{4}$. Each incoming link increases the importance of a web page, so at *Step* 1, we update the rank of each page by adding to the current value the importance of the incoming links. This is the same as multiplying the matrix A with v. At *Step* 1, the new importance vector is $v_1 = Av$. We can iterate the process, thus at *Step* 2, the updated importance vector is:

$$v_1 = A(Av) = A^2v.$$

Numeric computations give:

$$v = \begin{pmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{pmatrix}, Av = \begin{pmatrix} 0.37 \\ 0.08 \\ 0.33 \\ 0.20 \end{pmatrix}, A^2v = \begin{pmatrix} 0.43 \\ 0.12 \\ 0.27 \\ 0.16 \end{pmatrix}, \ldots, A^6v = \begin{pmatrix} 0.38 \\ 0.13 \\ 0.29 \\ 0.19 \end{pmatrix}, A^7v = \begin{pmatrix} 0.38 \\ 0.12 \\ 0.29 \\ 0.19 \end{pmatrix}, A^8v = \begin{pmatrix} 0.38 \\ 0.12 \\ 0.29 \\ 0.19 \end{pmatrix}.$$

We notice that the sequences of iterates $v, Av, \ldots, A^kv$ tends to the equilibrium value $v^* = \begin{pmatrix} 0.38 \\ 0.12 \\ 0.29 \\ 0.19 \end{pmatrix}$. We call this the PageRank vector of our web graph.

3.1 Linear Algebra Point of View

Let us denote by x_1, x_2, x_3 and $x_4$the importance of the four pages. Analyzing the situation at each node we get the system:

$$\begin{cases} x_1 = 1 \cdot x_3 + \frac{1}{2} \cdot x_4 \\ x_2 = \frac{1}{3} \cdot x_1 \\ x_3 = \frac{1}{3} \cdot x_1 + \frac{1}{2} \cdot x_2 + \frac{1}{2} \cdot x_4 \\ x_4 = \frac{1}{3} \cdot x_1 + \frac{1}{2} \cdot x_2 \end{cases}$$

This is equivalent to asking for the solutions of the equations $A \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}$, which is nothing but the eigenvector $c \begin{bmatrix} 12 \\ 4 \\ 9 \\ 6 \end{bmatrix}$ corresponding to the eigenvalue 1. Since Page Rank should reflect only the relative importance of the nodes and since the eigenvectors are just scalar multiples of each other, we can choose any of them to be our PageRank vector. Choose v^* to be the unique eigenvector with the sum of all entries equal to 1. (We will sometimes refer to it as the probabilistic eigenvector corresponding to the eigenvalue 1.) The eigenvector $\frac{1}{31} \cdot \begin{bmatrix} 12 \\ 4 \\ 9 \\ 6 \end{bmatrix} \sim \begin{bmatrix} 0.38 \\ 0.12 \\ 0.29 \\ 0.19 \end{bmatrix}$ is our PageRank vector.

3.2 Probabilistic Point of View

Since the importance of a web page is measured by its popularity (how many incoming links it has), we can view the importance of page i as the probability that a random surfer on the Internet that opens a browser to any page and start following hyperlinks, visits the page i. We can interpret the weights we assigned to the edges of the graph in a probabilistic way. A random surfer that is currently viewing web page 2, has $\frac{1}{2}$ probability to go to page 3 and $\frac{1}{2}$ probability to go to page 4. We can model the process as a random walk on graphs. Each page has equal probability$\frac{1}{4}$ to be chosen as a starting point. So, the initial probability distribution is given by the column vector $\begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix}$. The probability that page i will be visited after one step is equal to Ax, and so on. The probability that page i will be visited after k steps is equal to $A^k x$. The sequence $Ax, A^2x, \ldots, A^k x, \ldots$ converges in this case to a unique probabilistic vector v^*. In this context v^* is called the stationary distribution and it will be our Page Rank vector. Moreover, the i-th

entry in the vector $v^*$is simply the probability that at each moment a random surfer visits page i. The computations are identical to the ones we did in the dynamical systems interpretation, only the meaning we attribute to each step being slightly different.

3.3 Centrality Measures

(a) ***In-Degree Centrality*:** In-degree of a node v_i is defined as:

$$\sum \deg^{in}(v_i),$$

where $\deg^{in}(v_i)$ is the number of in-links for the node v_i.

(b) ***Out-Degree Centrality*:** Out-degree of a node v_i is defined as:

$$\sum \deg^{out}(v_i),$$

where $\deg^{out}(v_i)$ is the number of in-links for the node v_i.

(c) ***Degree Centrality***: Sum of in-degree and out-degree is called degree centrality.

(d) ***Eigen Centrality*:** The i-th component of the eigenvector for the highest eigenvalue of the adjacency matrix is called the Eigen centrality of the node v_i, that is:

$$Ax = \lambda_i x,$$

where λ_i is the eigenvalue of A with highest real part.

4. Graph Comparison between Various Centrality Measures and Page Rank

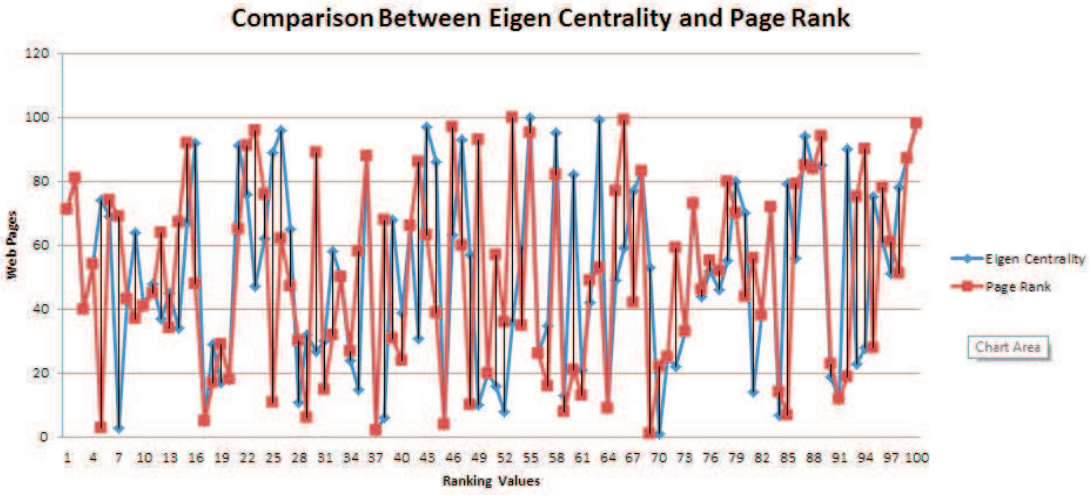

Figure 2

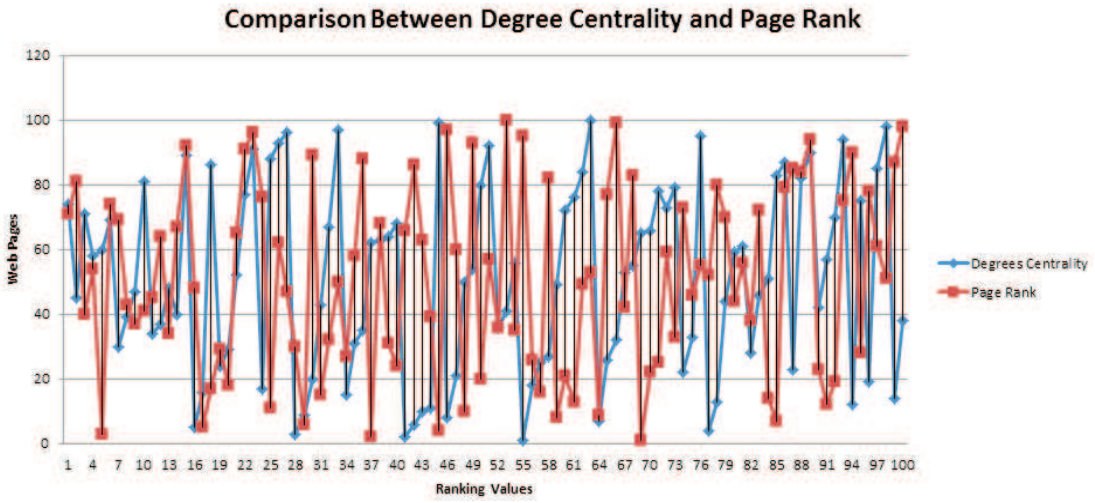

Figure 3

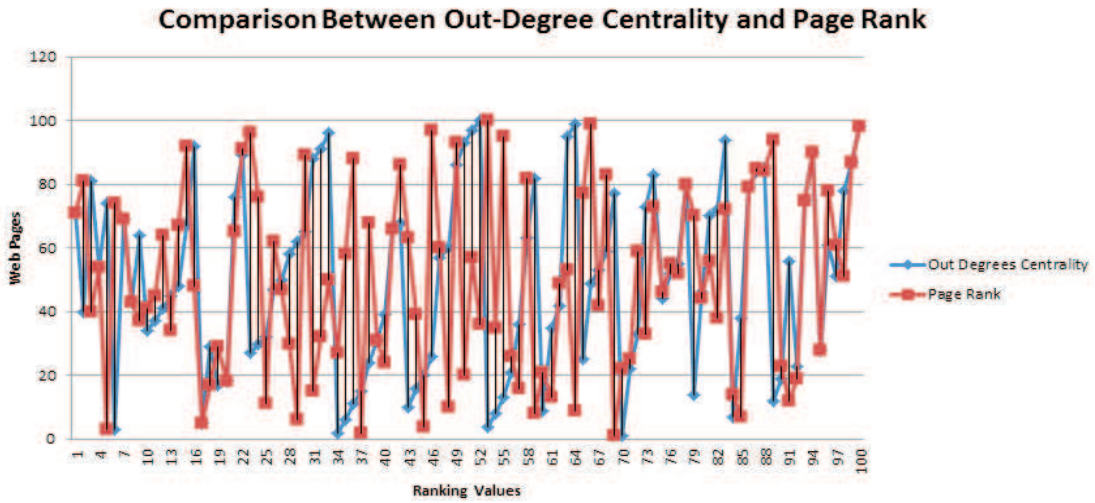

Figure 4

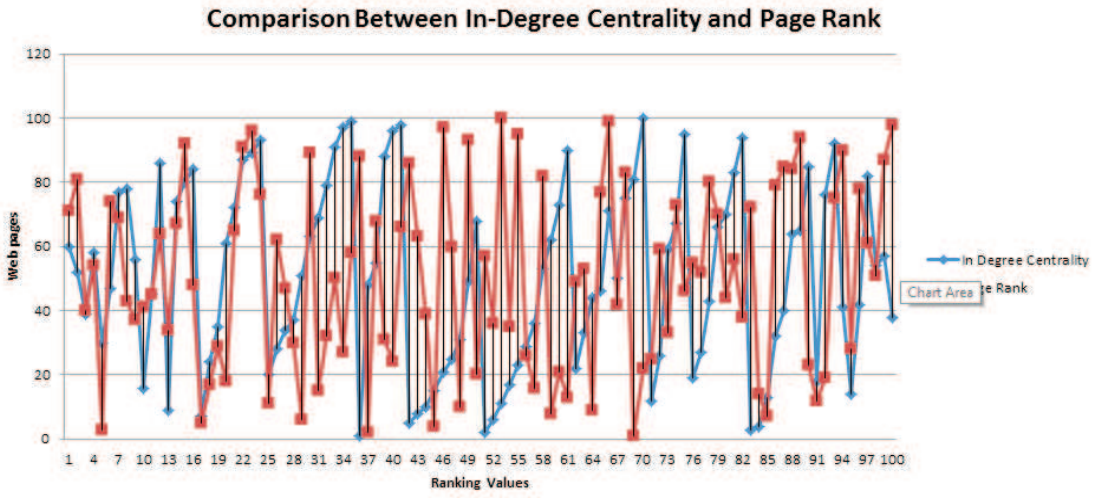

Figure 5

5. Conclusion and Future Works

We notice that Eigen centrality measure is more efficient when compared to the adjacency matrix based iterative page rank algorithm with respect to time. Calculating page rank using Eigen vector centrality measure is better than other Three centrality measures like degree centrality, in and out-degree centralities. Since Eigen vector centrality gives almost same rank with respect to page rank algorithm.

In the Adjacency based iterative page rank algorithm we get the page rank value after "n" iterations only, where "n" may be very large. Hence Eigen centrality measure we can apply for web page ranking instead of Adjacency based iterative page rank algorithm.

In our future work, we will introduce other centrality measures for calculating web page algorithm.

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] K. Bharat and A. Broder, A technique for measuring the relative size and overlap of public web search engines, *Computer Networks and ISDN Systems* **30** (1) (1998), 379 – 388.

[2] T. Bhatia, Link Analysis Algorithms for web mining, *International Journal of Computer Application* **2** (1) (2011) 243 – 246.

[3] M. Bianchini, M. Gori and F. Scarselli, Inside page rank, *ACM Transactions on Internet Technology* **5** (1) (2005), 92 – 128.

[4] S. Brin and C. Page, The anatomy of a large scale hypertentud web search engine, *Computer Network and ISDN Systems* **30** (1-7) (1998), 107 – 117.

[5] S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan and Tomkins, Mining the link structure of world wide web, *IEEE Computer* **32** (1999), 60 – 67.

[6] L. Choudhary and B. Shankar, Burdark role of ranking algorithms for information retrieval, *International Journal of Artificial Intelligence and Application* **3** (4) (2012), 21 – 34.

[7] C.H.Q. Ding, X. He, P. Husbands, H. Zha and H.D. Simon, Page rank: HITS and a unified framework for link Analysis, in *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development on Information Retrieval*, Tampere, Finland, pp. 353 – 354.

[8] `http://googleblog.blogspot.com/2008`.

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

[10] B. Jaganathan and K. Desikan, Hermition matrix based pagerank algorithm, *Global Journal of Pure and Applied Mathematics*, **12** (1-3) (2016), 271 – 280.

[11] B. Jaganathan and K. Desikan, Penalty-based pagerank algorithm, *ARPN Journal of Engineering and Applied Sciences* **10** (5) (2015), 2000 – 2003.

[12] 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.

[13] J.M. Klienberg, Authoritative sources in a hyperlinked environment, *Journal of the ACM* **46** (5) (1999), 604 – 632.

[14] A.N. Langville and C.D. Meyer, *Google Page Rank and Beyond: The Science of Search Engine Rankings*, Princeton University Press, Princeton, New Jersey (2006).

[15] L. Page, S. Brin, R. Motwani and T. Winograd, *The Page Rank Citation Ranking: Bringing Order to the Web*, Stanford Digital Library Technologies Project (1998).

[16] P. Ravikumar and A.K. Singh, Web structure mining exploring hyperlinks and algorithms for information retrieval, *American Journal of Applied Sciences* **7** (6) (2010), 840 – 845.

[17] D. Sepandar, H.K. Taher, H. Christopher, D.M. Gene and H. Golub, *Exploiting the Block Structure of the Web for Computing Page Rank*, *Stanford University Technical Report* (2003).

[18] D.K. Sharma and A.K. Sharma, A comparative analysis of web page ranking algorithms, *Journal on Computer Science and Engineering* **2** (8) (2010), 2670 – 2676.

[19] J. Wang, Z. Chen, L. Tao, W. Ma and W. Liu, Ranking user's relevance to a topic through link analysis on web logs, *WISM* (2002), pp. 49 – 54, `dl.ac.org`.