Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 50 (2015) 157 – 162

Procedia Computer Science

2nd International Symposium on Big Data and Cloud Computing (ISBCC'15)

Reachability Based Web Page Ranking Using Wavelets

S Hariharan[a,]*, S Dhanasekar[a], Kalyani Desikan[b]

[a]*Assistant Professor, Vellore Institute of Technology Chennai Campus, Chennai 600127, India.*
[a]*Professor, Vellore Institute of Technology Chennai Campus, Chennai 600127, India.*

Abstract

A naïve approach has been made by applying the concept of reachability for web page ranking and implementing multi resolution analysis using Haar wavelet to order the web pages. In this article, page ranking has been done by developing a structured signal using in links, out links and reachability values of the web pages of network graphs. Using Haar wavelet, the page ranking is proposed and developed. The average and detailed coefficients of the input signal and the down sampling process provides the necessary page ranking of web pages. This approach does not involve any iterative technique, damping factor or initialization of the page ranks. In this paper, comparison between the original page rank, category-based page rank and the proposed approach have been made. The result reflects the role of paths between the pages in page rankings.

Peer-review under responsibility of scientific committee of 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15)

Keywords: Page rank; World Wide Web; Wavelet Rank; Information Retrieval; Search Engines; Measure Rank.

1. Introduction

In this internet world, Web page ranking is one of the important aspects and is very much needed to obtain the relevant pages based on the interest of the user from the large collection of disordered information. Search engines are an important tool for users to retrieve information for a particular query. However, current search engines cannot fully satisfy the user's need for high-quality information search services; and it raises many new challenges for information retrieval. Page Rank algorithms are well known for ordering web pages.

* Corresponding author. Tel.: +91-44-39931235; fax: +91-44-39932555.
E-mail address: s.hariharan@vit.ac.in

Peer-review under responsibility of scientific committee of 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15)
doi:10.1016/j.procs.2015.04.078

Currently, the most classic Web structure algorithm is PageRank algorithm [1] that Sergey Brin and Larry Page proposed at Stanford University. In order to verify the performance of the algorithm, they successfully applied it to the Google search engine prototype, and now Google has become the world's most well-known search engine. However, the page rank algorithms [2 - 11] based on structural analysis, citation based ranking, weighted page rank, mining approach and block structure for measuring the relative importance of web pages make use of the hyperlinks-based approach and completely ignoring factors like content, topic and reachability. It is difficult to achieve most relevant page /more informative page from the large collection of web pages. Recently the category-based page rank algorithm was studied by Jaganathan and Kalyani Desikan [12].

Many of the existing page ranking algorithms are based on connectivity. Graph theory based on networks plays an important role in page ranking and many algorithms use the in links and out links of a page for ranking them. One more important aspect in graph theory involves the concept of Eccentricity. The more the eccentricity of a page, the more will be its reachability. That is, the rank will be high for those pages which have small eccentricity value. Behzad and Simpson [13] gave some information on eccentric sequences and eccentric sets in graphs. The computation of eccentricity for large graphs is given by Takes and Walter [14]. Unfortunately, not all the web graphs are strongly connected to have finite values as eccentricities. In order to exploit the concept of eccentricity, we have considered a similar approach which involves reachability for page ranking. In this article, the normalized length of the longest path from each vertex is considered as one of the component of the signal. To the knowledge of the authors, no attempt has been made so far to rank the web pages using the concept of reachability. In this paper we have considered a vertex (webpage) as a compact signal and made use of the wavelet [15-16] concept in ranking the pages.

2. Multi resolution analysis

A multiresolution analysis (MRA) of the set of square integrable functions denoted by $L^2(R)$, equipped with the standard inner product $(\cdot,\cdot)$, is a chain of closed subspaces indexed by all integers

$$\ldots V_{-1} \subset V_0 \subset V_1 \subset \ldots \tag{1}$$

such that

(i) $\overline{\lim_{n\to\infty} V_n} = L^2(R)$

(ii) $\overline{\lim_{n\to-\infty} V_n} = \{0\}$

(iii) $f(\cdot) \in V_n \Leftrightarrow f(2\cdot) \in V_{n+1}$

(iv) Let $\varphi(\cdot)$ be a scaling function such that $\{\varphi(\cdot - k) : k \in Z\}$ constitutes a complete orthonormal basis of V_0.

To obtain a multi resolution analysis, it suffices to construct the scaling function $\varphi(x)$. The entire space chain can then be reconstructed from $\varphi(x)$ according to (iii) and (iv). Since $V_0 \subset V_1$ and from (iii) and (iv), it is easy to see that $\varphi(\cdot)$ must be a linear combination of $\{\varphi(\cdot - k) : k \in Z\}$, leading to the two scale relation

$$\varphi(\cdot) = 2\sum_{k\in Z} h_k \varphi(2\cdot - k), \tag{2}$$

for a suitable set of coefficients $(\cdots, h_{-1}, h_0, h_1, \cdots)$.

Let W_0 denote the orthogonal complement of V_0 in V_1. A function $\psi(x)$ whose integer translates $\{\psi(\cdot - k) : k \in Z\}$ constitutes an orthonormal basis of W_0 is called a wavelet. This wavelet function $\psi(x)$ satisfies the two scale relation

$$\psi(\cdot) = 2\sum_{k\in Z} g_k \varphi(2\cdot - k), \tag{3}$$

for a suitable set of coefficients $(\cdots, g_{-1}, g_0, g_1, \cdots)$. From (i) – (iv) it is clear that

$$\{\psi_{jk}(\cdot) = 2^{j/2}\psi(2^j \cdot - k; j, k \in Z) \tag{4}$$

is an orthonormal basis of $L^2(R)$.

For the well known Haar wavelet, the scaling and wavelet functions are defined by

$$\varphi(x) = \begin{cases} 1, 0 \le x < 1 \\ 0, otherwise \end{cases} \qquad \psi(x) = \begin{cases} 1, 0 \le x < \frac{1}{2} \\ -1, \frac{1}{2} \le x < 1 \end{cases} \tag{5}$$

For any function $f \in L^2(R)$, define $P_J : L^2(R) \to V_J$ to be the projection of f onto the resolution space V_J.

$$f(x) = \sum_{k\in Z} c_{0k}(f)\varphi_{0,k}(x) + \sum_{0\le j\le J-1, k\in Z} d_{j,k}(f)\psi_{jk}(x) \tag{6}$$

is the synthesis of the signal f on the space V_J. The analysis of the signal can be done using the coefficients c_{0k}, $d_{j,k}$ called the scaling and wavelet coefficients of f. These coefficients are also called the average and detailed coefficients of the corresponding signal.

For a detailed introduction to wavelet theory, we refer to Strang [15], and Hernandez and Weiss [16].

3. Reachability based Page Ranking

Considering the Web network as a graph $G(V, E)$, where the vertex set V represents the web pages and the edge set E represents the corresponding link between the pages. When the graph G is strongly connected, the eccentricity $\varepsilon(v); v \in V$ can be calculated as the maximum graph distance between v and any other vertex $u \in V$. The minimum graph eccentricity in a graph is called the graph radius and the maximum eccentricity is called the graph diameter.

By the nature of the page ranking concept in strongly connected Web graphs, one can identify the relationship between the rank of a vertex and its eccentricity. The rank will be high when the eccentricity of a vertex (webpage) matches with the graph radius and the rank will be the lowest when its eccentricity matches with the graph diameter. So, the weightage of the eccentricity values has to be reversed for each and every vertex of a Web graph. When the Web graph is not strongly connected, we can consider the reversed values (normalized values) of the possible longest path from each and every vertex as its reachability component. In this, we need to avoid the cycles where the

longest path ends at the same vertex. Of course, the in links and the out links are also deciding factors in web page ranking.

Considering a web page (vertex) as a compact signal with its components as the in links, out links and the normalized reachability values, we can calculate the corresponding average and detailed coefficient values using (6). Since the in links are considered as recommendations and the out links are considered as authorities, a weight has been introduced as the ratio between the number of in links and the number of out links of a given vertex. The resulting values based on the product of the average coefficients and the corresponding weights in descending order will give the corresponding page ranking for each web page. If there is a tie between the web pages we use the detailed coefficients to rank them.

Consider the compact signal constructed in Fig 1. where l represents the number of in links, m represents the number of out links and n represents the reversed reachability value of a web page. When we synthesize and analyze a compact signal (web page) using multi resolution analysis, the maximum information of a signal (web page) can be captured from the unique average coefficient through the down sampling process and the high frequency components can be analyzed using the detailed coefficients at each and every level of down sampling.

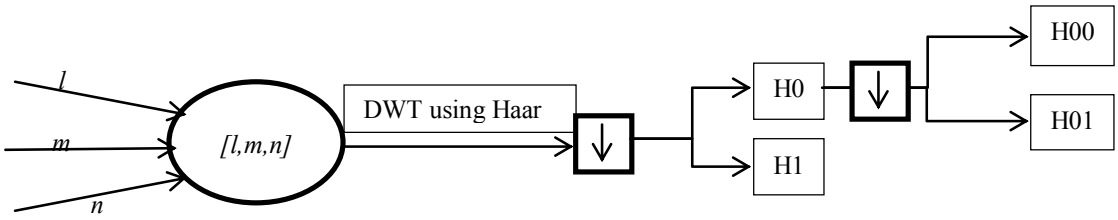

Fig. 1. Down Sampling using Multi Resolution Analysis

We used MATLAB to compute the average and detailed coefficients using Discrete Wavelet Transform (DWT) through Haar wavelet.

4. Comparison between page ranks

Given below are two sample graphs showing web pages as vertices, grouped based on three different categories.

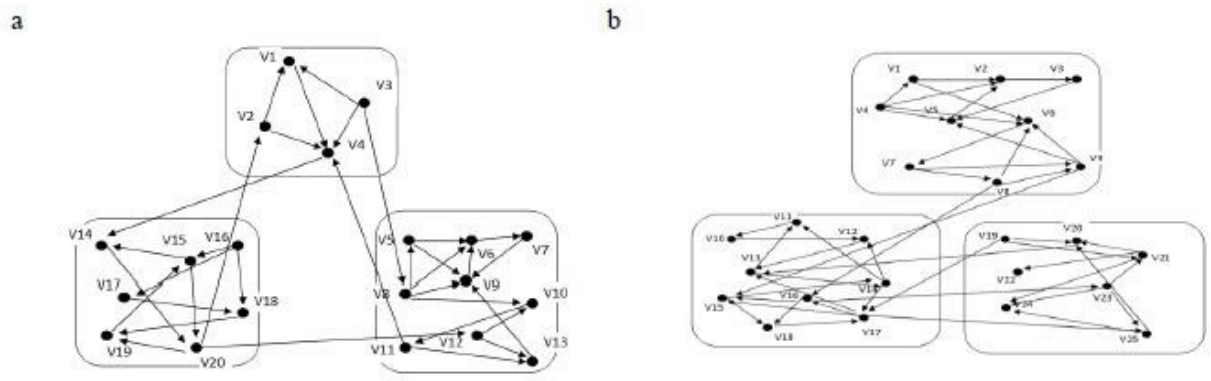

Fig. 2. (a) Web graph 1; (b) Web graph 2.

We shall now explain how to consider a webpage as a signal and to incorporate the notion of reachability to compute page ranks. We have illustrated this for the vertex $v9$ of Web graph 1 of Fig 2(a).

Consider the web page (vertex) $v9$ of Web graph 1 with 4 in links, 1 out link and 2 as its longest path length. The corresponding signal of $v9$ will have the values [4,1,8] where 8 is the reversed reachability value based on the length of the longest path of Web graph 1. Note that the length of the longest path of Web graph 1 is 10 which starts at $v17$ and for Web graph 2 it is 9 which starts at $v11$. Using MATLAB, we get the average coefficient for this vertex (signal) as 10.5, the first level detailed coefficient as -2.1213 and the second level detailed coefficient as -5.500. When we multiply the average coefficient with its weight 4 (the ratio between in links and out links of $v9$), the corresponding weight of the web page becomes 42. Similarly, we can calculate the weights for all the Web pages.

In Table 1, we see that the web pages (vertices) $v19$ and $v7$ have the same rank. When there is a tie between two pages the detailed coefficients are considered for breaking the tie and ranking them. The corresponding detailed coefficients at the first and second levels are 1.4142 and -3.0 for $v19$, 0 and -7.0 for $v7$. Hence, in the page ranking order $v19$ is ranked higher than $v7$. It can also be noted that vertex $v7$ of Web graph 1 is not ranked high in our proposed algorithm because it is part of the cycle $v6 - v7 - v9 - v6$ but its rank is artificially boosted in the original page rank algorithm.

Similarly in Table 2, though $v11$ and $v14$ have the same rank, $v11$ comes ahead of $v14$ based on their detailed coefficients 1.4142, 1.0 for $v11$ and -0.4714, -1.6667 for $v14$. In Web graph 2 we see that vertex $v3$ is part of the cycle *v5-v2-v3-v5* and hence its rank has been artificially boosted in the original page rank algorithm. We also observe that vertex *v6* of Web graph 2 is ranked higher in our proposed algorithm because it has more connectivity.

In the tables below, the web pages are arranged in decreasing order of page rank values. The existing page ranks based on the original page rank algorithm and category-based page rank algorithm are tabulated in the first four columns of each table with their corresponding page rank values. The reachability based page rank of web pages are given in the last two columns of both the tables with their corresponding page rank values.

Table 1. Comparison between the page ranks for web graph – 1.

Web Page	Page Rank	Web Page	Category-based Page Rank	Web Page	Proposed Page Rank
v9	2.4446	v9	2.0365	v9	42
v6	2.3504	v6	1.9934	v6	30
v7	2.1478	v7	1.8444	v4	22
v20	1.8103	v19	0.7432	v13	17
v14	1.5974	v20	0.6836	v10	11
v4	1.3469	v11	0.6199	v14	11
v19	1.2221	v10	0.5528	v19	9
v11	0.9458	v4	0.5454	v7	9
v12	0.9373	v15	0.5086	v1	7
v10	0.9362	v13	0.4877	v18	7
v2	0.9194	v12	0.3814	v15	6
v13	0.8163	v14	0.3736	v20	5
v15	0.7119	v18	0.3561	v5	4.75
v1	0.5832	v1	0.278	v11	3.25
v18	0.3561	v2	0.2013	v12	2.25

Table 2. Comparison between the page ranks for web graph – 2.

Web Page	Page Rank	Web Page	Category-based Page Rank	Web Page	Proposed Page Rank
v5	2.5431	v5	2.3059	v5	30
v2	2.4379	v2	2.237	v2	30
v3	2.2222	v3	2.0514	v6	22
v13	1.9569	v13	1.1953	v20	15
v12	1.6422	v12	1.1624	v18	13
v11	1.3619	v9	1.015	v13	9.75
v14	1.3405	v6	0.9736	v15	9.75
v9	1.3395	v7	0.9256	v12	9
v10	1.3086	v10	0.9118	v3	9
v7	1.2628	v11	0.8956	v17	7.125
v6	1.2471	v14	0.8373	v16	5.67
v17	1.0952	v17	0.7336	v11	5
v15	0.8437	v8	0.5549	v14	5
v8	0.7681	v15	0.4217	v9	4.333
v18	0.5899	v18	0.3942	v21	4.333

v8	0.1925	v17	0.1925	v2	1.75
v17	0.1925	v5	0.1835	v8	1.375
v5	0.1909	v8	0.1575	v17	1
v3	0.15	v3	0.15	v3	0
v16	0.15	v16	0.15	v16	0

v24	0.4852	v20	0.3822	v25	4
v20	0.4834	v24	0.3642	v24	3.8295
v25	0.4483	v16	0.3059	v22	3.4965
v16	0.3828	v21	0.2595	v10	3
v21	0.2793	v22	0.2235	v7	2.75
v23	0.2313	v25	0.2233	v8	2.33
v22	0.2291	v1	0.2138	v1	1.75
v1	0.2138	v23	0.1615	v23	1.6667
v4	0.15	v4	0.15	v4	0
v19	0.15	v19	0.15	v19	0

It can be noted that the web pages with the highest and lowest page ranks computed using the original, category-based and reachability based methods match for both the sample graphs.

4. Conclusion

In this paper, a new approach based on the concept of reachability in graphs has been applied to rank web pages. The in links, out links and reversed reachability values of vertices in the web graph are considered as the components of a compact signal and by applying multi resolution analysis to the constructed signal we obtained the average and detailed coefficients. The introduced page ranking method based on reachability is easy to calculate and effective because of the simplicity of the Haar wavelet. The procedure does not require any initial assignment of the rank for pages. It avoids the iteration process and hence the computational complexity is reduced. The concept of eccentricity can be implemented when the Web graph is strongly connected. The ranking can be improved by assigning proper weights to web pages containing relevant information and that can be included as a factor in the given signal. It will be an interesting study if we apply other existing wavelets for page ranks. The idea can be extended using Fuzzy logic by appropriately making each and every signal (web page) as a fuzzy number.

References

1. S.Brin and L.Page . The anatomy of a large scale Hyper textual web search engine. *Computer networks and ISDN systems* .1998;**30**:107-117.
2. Page.L,Brin.S,Motwani.R and Winograd.T. The page rank citation ranking : bringing order to the web. *Technical report,.Stanford Digital Library technologies.* 1998.
3. Wenpu Xing and Ali Ghorbani.Weighted page rank algorithms. *Proceedings of the second annual conference on communication networks and services research (CNSR'04).* IEEE 2004.
4. Klienberg.J.M. Authoritative sources in a hyperlinked environment. *Journal of the ACM.*1999:46(5):604-632.
5.Auth Dell Zhang and Yisheng Dong.A novel web usage mining approach for search engines. *Computer Networks.*2002:39:303-310.
6.Zhicheng Dou,Ruihua Song,Ji-Rong Wen and Xiaojie Yuan. Evaluating the effectiveness of personalized web research. 2009:21(8)
7.Sepander.D,Kanver,Taher.H,Havliwala Christopher.D and Manning Gene,Golub.H. Exploiting the block structure of the web for computing page rank. *Standford University Technical Report*.
8.N.Duhan,A.K.Sharma and K.K.Bhatia. Page ranking algorithms –A survey. *Proceedings of the IEEE international conference on advanced computing*.2009
9.Dilipkumar Sharma and Sharma.A.K. A comparative analysis of web page ranking algorithms. *Journal on computer science engineering*.2010:2(8):2670-2676.
10.Mander Kale and Shanthi Thilagam.P. DYNA-RANK efficient calculation and updation of page rank. *International conference on computer science and information technology*.2008 .
11.Neelam Tyagi and Simple Sharma. Weighted page rank algorithms based on number of visits of links of web page. *International journal of soft computing and Engineering*.2012:2(3).
12. B. Jaganathan and Kalyani Desikan, Category-based page rank algorithm, *Proceedings of ICMCE* 2014:933-937.
13. M. Behzad and J E Simpson, Eccentric sequences and eccentric sets in graphs, *Discrete Math.* 1976:6:187-193.
14. Frank. W. Takes and Walter A. Kosters, Computing the eccentricity distribution of large graphs, *Algorithms,* 2013:6:100-118.
15. G. Strang and T. Nguyen, Wavelets and filter banks, *Wesley-Cambridge Press,* 1996.
16. E. Hernandez and G. Weiss, A first course on wavelets, *CRC Press,* 1996.