International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 14 (2015) pp 34856-34860

Performance Analysis Of Page Access Coefficient Algorithm For Information Filtering

Mrs. L.Rajeswari

Programmer, Computer Centre Alagappa University, Karaikudi 630 003 Tamilnadu. India
04565-223241 lr_eswari@yahoo.co.in

Dr.S.S.Dhenakaran

Professor, Department of Comp. Sci. & Engg. Alagappa University, Karaikudi 630 003 Tamilnadu. India
ssarvind@yahoo.com

ABSTRACT
Data mining, one of the thrust areas of research explores the unknown but potentially valid information from high volume of data already available. Web mining is prominent area of data mining that includes exploration of information from the web pages available on the world wide networks. Hence, information retrieval from the available web pages requires specific techniques with reduced complexity and increased speed. Online social networks involve large quantity of information transmission that creates increased impact on it. Social networks can be defined as the network of interactions or instance of relationship between multiple numbers of nodes connected to it. In this social network connection the nodes are commonly referred as actors, edges indicate the relationship among the actors and interaction between them. Page Ranking, a common method used for information filtering involves high computational complexities and number of iterations to find out particular pages are also more. Page Access Coefficient (PAC) is used for the information retrieval from the social network websites. This paper compares the performance of PAC with Page Rank and Weighted Page Rank. The experimented analysis clearly shows that the performance of PAC is better than Page Rank and Weighted Page Rank.

Keywords: Social Networks, Page Ranking, Weighted Page Ranking, Information Filtering, Page Access Coefficient.

1. Introduction

The first recognizable online social network (OSN) SixDegrees.com launched in 1997, abundant applications such as Facebook, Twitter and LinkedIn became popular Internet platforms, which connect people around the globe [1]. World Wide Web is the collection of repository of interlinked documents that are in various forms such as text, images, audios as well as videos. This multimedia documents uses, Hypertext Transfer Protocol to exchange information between the nodes. The primary goal of the web site owner is to provide the relevant information to the users to fulfill their needs [2]. Web mining is the technique used to categorize the web pages by ranking the web pages, which plays important role in page access. Information retrieval from the available websites becomes a challenging issue, which involves complexities. Ranking WebPages is an important mission as it assists the user look for highly ranked pages that are relevant to the query [3]. From the social network web pages it becomes more complex to search for the relevant web pages since it is interconnected. Hence, page ranking algorithms are involved for the purpose. The most famous search engine Google used Hyperlink structure for ranking the web pages. Various algorithms are exists in the context, to get the desired results. Different page ranking algorithms Page Rank (PR), WPR (Weighted Page Rank), HITS (Hyperlink Induced Topic Search) are used for the relevant information retrieval. Some of these algorithms are user-neutral and measure the relevance of documents primarily based on the contents and relationships of documents [4]. Most of these algorithms largely focus on local activities of the user, and fail to embrace the large social contexts of the user.

The paper is structured as follows. Section 1 contains basic information about social networks, page ranking and searching of relevant web pages. Section 2 provides the background study. Section 3 consists of Experimental study followed by comparative analysis. Section 4 shows results and discussion. Section 5 concludes the outcomes of the performance analysis.

2. Background Study

2.1 Information Filtering

Filtering [5] is the tool used to find most valuable and potentially valid information from the huge volume of information found on the web pages in such a way that limited time span is spent on the listening to the web pages searching. Filters are also used to organize and structure information found on the multiple documents in various formats. Rather than for the use of individual addressing, filtering often used for groups. Filtering is also needed on the search results from Internet search engines.

Information retrieval [6] can be characterized in varied ways, whereas the goal is to eliminate the redundant information and information overloading. It helps the user to choose the needed document from high voluminous data within short span of time only with the relevant documents. Relevant information on the other hand, defined solely for a specific

user and placed under the context of a particular domain or a topic.

Social networks are used for group communication rather than the individual usage. Hence, the social information can be used to improve the task of retrieving relevant information and refining each agent's knowledge. The information filtering techniques were employed at various applications to make suitable use of the pages.

2.2 Ranking Web Pages

As the use of web is increasing more day by day, the web users get easily lost in the web's rich hyper structure. The main aim of the owner of the website is to give the relevant information according their needs to the users. Different Page Rank based algorithms like Page Rank (PR), WPR (Weighted Page Rank), HITS (Hyperlink Induced Topic Selection), Distance Rank and Eigen Rumor algorithms are also available in the context.

With swift growth of the World Wide Web, providing user with highest quality of information becomes a challenging issue to be overcome. The main reason for the drawback is that some of the web pages are mostly self – descriptive and purely used for the navigational purpose. Therefore, finding appropriate pages relies on the content present in the web pages.

Different web page ranking algorithms [7] are also compared based on their methodology; relevancy, results yielded and their drawbacks are explained clearly. The goal of search engines is to provide relevant information to the users to cater to their needs. Therefore, finding the content of the Web and retrieving the user's interests and needs have become increasingly important. Depending upon the need and aim of the work, topology and technique can be used.

Weighted Page rank [8] takes into account the importance of both the in links and the out links of the pages and distributes rank scores based on the popularity of the pages. The results of the author's simulation studies show that WPR performs better than the conventional Page Rank algorithm in terms of returning larger number of relevant pages to a given query. Their experimental results ended with the current version of WPR, only the in links and out links of the pages in the reference page list are used in the calculation of the rank scores.

The author [5] makes an overview of methods and problems including social filtering, where people help each other with filtering objects on the net. Filters are also used to organize and structure information on the web pages for the proper access on it. Filters are also used to organize and structure information within short span of time. Various approaches and current research area on information filtering on the social networks are been carried out in.

In [9] shows the concept of role of ranking algorithms for information filtering. Their experimental study includes different page ranking algorithms and compares those algorithms and makes a study on it. From the results, it is noted that Simulation Interface has been designed for Page Rank algorithm and Weighted Page Rank algorithm but Page Rank is the only ranking algorithm on which Google search engine works. From the analysis, it is concluded that existing algorithms have limitations in terms of time response, accuracy of results, importance of the results and relevancy of results. An efficient web page ranking algorithm should meet out these challenges efficiently with compatibility with global standards of web technology.

The author [10] shows the Modeling Social Network Sites with Page Rank and Social Competences and provided with the results. The main idea consists in defining a personalization vector that is a linear combination of some prescribed social skills or social competences. In this model, a social competence is any feature of the user related to social skills that can enhance its Page Rank as a node of a Social Networking Sites.

Page ranking and Weighted Page Ranking are the two commonly used algorithms on the mining of web content to improve the maximized utilization of the web pages.

2.3 Page Rank

Page Rank is the method used to rank vast number of web pages available on the network. The page ranking algorithm [11] depends on the linking structure of the web pages. The page rank algorithm is based on the concept that if a page contains an important link towards it, then the page where the link exceed in also considered as an important navigation for the current page. It provides an advanced way to compute the importance or the relevance of web pages than finding the back links. Page Rank equation is given as follows:

$$PR(A) = (1-d)+d(PR(T_1)/ O(T_1) + \ldots\ldots\ldots\ldots +PR(T_n)/O(T_n)) \quad (1)$$

The parameter d is the damping factor, usually set as 0.85. O(A) is defined as the number of links that are considered as the out links for the page A.

2.4 Weighted Page Rank Algorithm

Weighted Page Rank algorithm [9] is an extension of the page ranking algorithm. This algorithm assigns a larger rank values to the more important pages rather than dividing the rank value of a page evenly among its outgoing linked pages. Each outgoing links gets a value proportional to its importance. The importance is assigned in terms of weight values to the incoming and outgoing links and are denoted as $W^{in}(m, n)$ and $W^{out}(m, n)$ as shown in the equation 2 and 3 below. The weight link (m, n) is calculated on the basis of number of incoming links of a page n and the number of incoming links of all reference pages of page *n*.

$$W^{in}_{(m,n)} = \frac{I_{\text{In}}}{\sum_{p \in R(m)} I_P} \quad (2)$$

$$W^{out}_{(m,n)} = \frac{O_{\text{n}}}{\sum_{p \in R(m)} O_P} \quad (3)$$

Where I_n and I_{out} are the number of incoming number of page n and page p respectively. $R(m)$ denotes the reference page list

on a page m. weight of the link is calculated on the basis of outgoing links for $W^{out}(m, n)$, is the weight of the link calculated based on the number of outgoing links of page p and the number of outgoing links for all reference pages of m. The O_n and O_p are the number of outgoing links of page *n* and *p* respectively. The formula given below is the modification of the page rank algorithm, called as weighted page rank algorithm.

$$\mathbf{WPR(n) = (1-d) + d} \sum_{m \in B(n)}$$

$$\mathbf{WPR\ (m)}\ W^{in}_{(m,n)}\ W^{out}_{(m,n)} \qquad \mathbf{(4)}$$

2.5 Page Access Coefficient Calculation

The rank of the web page is calculated based on the number of incoming links of pages, number of outgoing links of page and the total number of pages. The formula for calculating PAC [12] is as given below:

$$\mathbf{PAC = I_A + (O_A / n)} \qquad \mathbf{(5)}$$

where

PAC Page Access Coefficient
A Web Page
I_A Number of pages referring the page A
O_A Number of pages referred by page A
n Total number of Pages

2.6 Example

Let us consider an example of hyperlinks structure of four pages A, B, C, D and E as shown in the Figure 1.

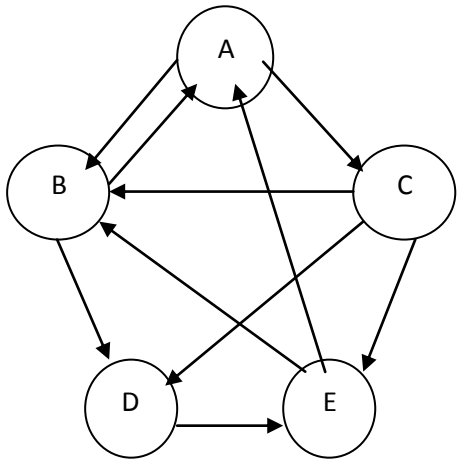

Figure 1 Hyperlink Structure for 5 pages

Pages	Inner link	Outer link
A	**2**	**2**
B	**3**	**2**
C	**1**	**3**
D	**2**	**1**
E	**2**	**2**

Page Rank, Weighted Page Rank and Page and Access Coefficient Values for Figure 1 are shown in Table 1.

Table 1 – Ranks for the Pages using PR, WPR and PAC

Pages	Page Rank	Weighted Page Rank	Page Access Coefficient
A	**1.67**	**0.66**	**2.4**
B	**1.93**	**1.59**	**3.4**
C	**0.86**	**0.22**	**1.6**
D	**1.33**	**0.39**	**2.2**
E	**1.65**	**0.44**	**2.4**

PR calculated values for the Figure 1, WPR calculated values for the Figure 1 and PAC calculated for the Figure 1 are shown in the above Table. From this, it is shown that the results for PAC are obtained in single iteration. The order of PR values of page A, B, C, D, E; WPR values for the pages A, B, C, D, E and PAC values for the pages A, B, C, D, E are remains same.

Page B is having the top most rank value, then A and E followed by D and the last one is Page C.

But in the PR calculation many iterations are involved which increases the calculation time.

In the WPR calculation the weight of each vertex has to be calculated, it increases the calculation complexity. But as for the PAC method, the results are obtained in single step; therefore it reduces the time and complexity of the calculation. The time complexity for the proposed algorithm is *O(n)*, since for *n* vertices the iteration is only *n*.

3. Experimental Study

The PAC method is implemented in C# language, development platform is Visual Studio 2010, Framework version 4.0 under the configuration of Core-i3 with 3 GB RAM. The datasets for the implementation is taken from the Stanford educational website, social web graph dataset [13]. The total number of record set varies so as to demonstrate the strength and performance of the PAC algorithm. The following table shows the results from the experimental analysis of three algorithms. The outcomes of the experimental study are shown in the Table 2 given below. The execution time is taken as the parameter for comparison here.

Table 2 – Comparative Analysis between Existing and PAC Algorithm

Number of Pages	Page Rank (sec)	Weighted Page Rank (sec)	Page Access Coefficient (sec)
200	0.832	0.511	0.043
500	1.049	0.849	0.071
1000	2.508	1.718	0.100
2000	10.472	9.002	0.192
5000	55.449	54.800	0.548
10000	192.108	183.968	1.676
15000	418.248	387.150	3.609
20000	967.239	698.901	5.878

The results obtained above illustrates that the PAC method stands good for larger number of records. Even though the datasets becomes large the execution time becomes low with using the PAC Algorithm. The PAC algorithm produces better results without iterations and complex computations. The algorithm's efficiency lies in the above said parameters. The unit measurement for the algorithm is given in seconds for better understanding of the outcomes.

4. Results and Discussions

From the Table 2, time delay distribution

	Time saved while using PAC than PR	Time saved while using PAC than WPR
Pages 200 →	**0.789**	**0.468**
Pages 500 →	**0.978**	**0.778**
Pages 1000 →	**2.408**	**1.618**
Pages 2000 →	**10.280**	**8.810**
Pages 5000 →	**54.901**	**54.252**
Pages 10000 →	**190.432**	**182.292**
Pages 15000 →	**414.639**	**383.541**
Pages 20000 →	**961.361**	**693.023**

As, we all know Internet as well as Social Network websites are having enormous number of pages. So the execution time of PAC is very low when compared to PR and WPR. It is also demonstrated in the following chart. Being used at the social networks in spite of continuous changing of behavior, the algorithm works with low execution time since there is no iterations and complex calculations. The chart displays the performance of the PAC compared with the existing algorithms.

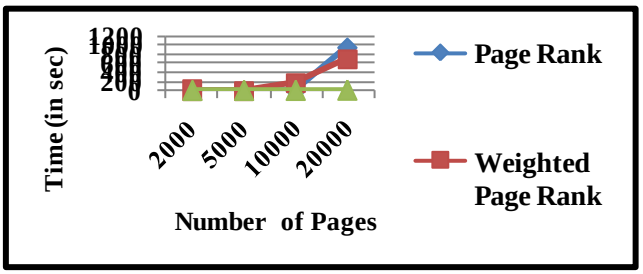

Chart 1 – Comparative Analysis between the Existing and PAC Method

5. Conclusion

The rapid growth of internet and the social networks made an important issue, whether the retrieved information were relevant or not. It also checks whether the given query satisfies the user need or not. This paper comes out with the solution for the page access using the page access coefficient algorithm, that makes simpler computations to retrieve information that are more closer enough to the search query with no iterations. Another important advantage of the proposed algorithm is that the time taken to filter the relevant web pages from high voluminous data was minimized which is on the other hand increases the execution speed. The page rank and weighted page ranking algorithms results were compared with the PAC algorithm. The comparative analysis shows that the PAC efficiently retrieves relevant web pages quickly than PR and WPR.

References

[1] Julia Heidemann and Mathias Klier, "Identifying Key Users in Online Social Networks: A PageRank Based Approach", as Completed Research Paper, pp:1-21.

[2] Rekha Jain and Dr. G. N. Purohit, Page Ranking Algorithms for Web Mining", in International Journal of Computer Applications (0975 – 8887),Volume 13– No.5, January 2011, pp:22-25.

[3] NeelamTyagi, Simple Sharma, "Weighted Page Rank Algorithm Based on Number of Visits of Links of Web Page", in International Journal of Soft Computing and Engineering (IJSCE),ISSN: 2231-2307, Volume-2, Issue-3, July 2012, pp:441 – 446.

[4] Zakaria SulimanZubi, "Ranking WebPages Using Web Structure Mining Concepts", in Recent Advances in Telecommunications, Signals and Systems, Vol 2, pp:21- 28.

[5] Jacob Palme, "Information Filtering", in http://www.dsv.su.se/~jpalme/select/information-filtering.pdf, pp:1-10.

[6] DELGADO et al. "content-based Collaborative Information Filtering", Nagoya 466 Japan (jdelgado.ishii.tomkey)

[7] Pardakhe N.V. and Prof. Keole R. R., "Analysis of Various Web Page Ranking Algorithms in Web Structure Mining", in International Journal of Advanced Research in Computer and Communication

Engineering,Vol.2, Issue 12, December 2013,pp:4798-4803.

[8] Wenpu Xing and Ali Ghorbani, "Weighted PageRank Algorithm", Proc of the Second Annual Conference on Communication Networks and Services Research (CNSR'04), IEEE,2004.

[9] LaxmiChoudhary and Bhawani Shankar Burdak, "Role of Ranking Algorithms for Information Retrieval", pp: 1-17.

[10] Francisco Pedroche, "Modelling Social Network Sites with PageRank and Social Competences", in International. Journal of Complex Systems in Science, vol. 1 (2011), pp. 65–68.

[11] Ashutosh Kumar Singh, Ravi Kumar P. "A Comparative Study of Page Ranking Algorithms for Information Retrieval". IJECE,4.7.2009.

[12] Rajeswari L., Prof. Dhenakaran S.S., "Page Access Coefficient Algorithm for Information Filtering in Social Network", IJCET@IAEME, Volume 4, Issue 3, May-June (2013), pp.60-69.

[13] Leskovec J., Lang K., Dasgupta A., Mahoney M.. Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters. Internet Mathematics 6(1) 29--123, 2009. http://snap.stanford.edu/data/web-Stanford.html.