AN IMPROVED PAGERANK ALGORITHM BASED ON A HYBRID APPROACH

Fares Anwar*

Faculty of Engineering, Science & Technology
Infrastructure University Kuala Lumpur
Malaysia

Koo Kwong Ze

Faculty of Engineering, Science & Technology
Infrastructure University Kuala Lumpur
Malaysia

Rozilawati Razali

Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia
Malaysia

Abudhahir Buhari

Faculty of Engineering, Science & Technology
Infrastructure University Kuala Lumpur
Malaysia

Elisha Tadiwa

Faculty of Engineering, Science & Technology
Infrastructure University Kuala Lumpur
Malaysia

***Corrospounding author's Email:** fares.anwar@iukl.edu.my

Peer-review under responsibility of 4th Asia International Multidisciplinary Conference 2020 Scientific Committee

<http://connectingasia.org/scientific-committee/>

© 2020 Published by Readers Insight Publisher,

lat 306 Savoy Residencia, Block 3 F11/1, 44000 Islamabad, Pakistan,

editor@readersinsight.net

This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Abstract

PageRank is an algorithm that brings an order to the Internet by returning the best result to the users corresponding to a search query. The algorithm returns the result by calculating the outgoing links that a webpage has thus reflecting whether the webpage is relevant or not. However, there are still problems existing which relate to the time needed to calculate the page rank of all the webpages. The turnaround time is long as the webpages in the Internet are a lot and keep increasing. Secondly the results returned by the algorithm are biased towards mainly old webpages resulting in newly created webpages having lower page rankings compared to old webpages even though new pages might have comparatively more relevant information. To overcome these setbacks, this research proposes an alternative hybrid algorithm based on optimized normalization technique and content based approach. The proposed algorithm reduces the amount of iterations required to calculate the page rank hence improving efficiency by calculating the mean of all page rank values and normalising the page rank value through use of the mean. This is complemented by calculating the valid links of web pages based on the validity of the links rather than the conventional popularity.

Keywords: *PageRank, Optimized Normalized Technique, Content-based Approach, Biasness, Iteration*

Research Highlights

PageRank algorithm is employed in search engines for retrieving the best results for a query by using the structure of the internet graph to rank the importance of a webpage (1). The problems of the PageRank algorithm are the algorithm is favoring the old webpage making new informative webpage become less prioritize and the algorithm required a lot of time to calculate the true page rank of a webpage (2). The aim of this paper is to address the above issues by proposing an enhanced PageRank algorithm that combines optimized normalized technique (3) and content-based approach (4). The reason to use optimized normalized technique is by using the technique it can speed up the convergence process making the page rank to be converging faster than conventional PageRank algorithm to get the final true page rank. Moreover, the content-based approach is letting the rank distribution to become validity manner as to de-bias to old webpages compared to conventional popularity based.

Graphical Abstract

Error! Reference source not found. shows the iterations needed for the proposed PageRank algorithm to calculate the final true page ranks of all webpages. It is apparent that the proposed algorithm is 74% faster than other PageRank algorithms in terms of iteration. Moreover, the

proposed algorithm requires only 27 iterations to calculate the final true page ranks of all webpages.

Figure1. Line graph of Proposed PageRank Algorithm

Research Objectives

This research is aim to propose an enhanced algorithm to get the result without favoring to old webpage. This objective is test by ranking position of PageRank algorithm and hence the ranking position obtained by using the proposed algorithm will be compare with the ranking position with the PageRank algorithm. In additon the an enhanced algorithm will decrease the iteration needed to calculate the page rank. This objective is test by the number of iterations needed to calculate the final page rank of all the webpages.

Methodology

The proposed algorithm starts with calculating the mean of the all page rank value and normalizing the page rank value through the mean. Then calculate the valid links of the webpages. This technique used because of the valid links and the rank distribution will be based

on the validity of the links instead of the conventional popularity. The proposed algorithm is illustrated by the flow chart in Figure below.

Figure 2. Flowchart of the proposed algorithm

Results

The proposed algorithm reduces the number of iterations required to calculate the page rank, hence improves the efficiency, by calculating the mean of all page rank values and normalizes them through the use of the mean. Through this approach, the algorithm is also able to determine the relevancy of webpages based on validity of links rather than popularity. These claims are demonstrated by an experiment conducted on the proposed algorithm using a

dummy web structure consisting of 12 webpages. The results showed that the traditional PageRank algorithm has 74% more iterations than the proposed algorithm. The proposed algorithm returned a mean value of 1.00 compared to 1.32 for the traditional algorithm. These results confirm that the proposed algorithm saves a substantial amount of computing power while being more precise and not biased.

Findings

This paper has proposed an alternative hybrid PageRank algorithm based on optimized normalization technique and content-based approach. The proposed algorithm has outperformed other PageRank algorithms in terms of iterations needed and the quality of webpages returned. In term of iterations needed, the proposed algorithm requires less iterations to calculate the final true page ranks of all webpages, thus it becomes faster than other PageRank algorithms. In terms of quality, the proposed algorithm takes into account the valid links of the webpages and hence, it has the same results as the Web Content based PageRank algorithm but different compared to the traditional PageRank algorithm

REFERENCES

- [1]. Page L, Brin S, Motwani R, Winograd T. The pagerank citation ranking: Bringing order to the web. Available from: <http://ilpubs.stanford.edu:8090/422>
- [2]. Kumari T, Gupta A, Dixit A. Comparative Study of Page Rank and Weighted Page Rank Algorithm. *Int J Innov Res Comput Commun Eng.* 2014;(2):2929–37. Available from: www.ijircce.com
- [3]. Dubey H, Roy B. An improved page rank algorithm based on optimized normalization technique. *Int J Comput Sci Inf Technol.* 2011;2(5):2183–8. Available from: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.228.938>
- [4]. Hao Z, Qiumei P, Hong Z, Zhihao S. An improved PageRank algorithm based on web content. In: *International Symposium on Distributed Computing and Applications for Business Engineering and Science. 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science; 2015 p. 284–7.* Available from: <https://ieeexplore.ieee.org/abstract/document/7429612/>