ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 3, Issue 2, pp: (998-1004), Month: April - June 2015, Available at: www.researchpublish.com

ACTION-SENSITIVE ENHANCED WEIGHTED PAGE RANK ALGORITHM (ASE-WPRA)

[1]Suman Kumari, [2]Ritu Maheshwari Bansal

[1]Research Scholar, M.Tech (CSE), MRIU, Faridabad, Haryana, India
[2]Assistant Professor (CSE), Faculty of Engineering & Technology Engineering, MRIU, Faridabad, Haryana, India

Abstract: **With the development of internet enormous web pages have generated. It is very much critical for user to access web pages that satisfy user demands. Hence ranking of pages come into role to provide relevant information to cater their needs. Important pages receive a higher PageRank and are more likely to appear at the top of the search results. The PageRank of a page is defined recursively and depends on the number and PageRank metric of all pages that link to it. This article provides an acquaintance of traditional weighted page rank algorithm. Then we propose a new method to yield more accurate positive results. Firstly, we adding the concept of duplicate links. In addition to that, we accumulate visitors' action performed per web page. Secondly, priority is assigned to each action .After analyzing priority, rank is updated. Following that by incorporating the concept of gini index of SPRINT algorithm, improved method turns out to be better than traditional page rank algorithm. Gini index provides weightage to each action. The relevancy of the webpages returned is more because the user behavior is also considered to rank the webpages.**

Keywords: **Duplicate links, Gini index, link ratio analysis, Enhanced Weighted PageRank algorithm, SPRINT algorithm, user action, weighted PageRank algorithm.**

I. INTRODUCTION

Every day the web grows by roughly a million electronic pages, adding to hundreds of millions pages already on- line. Page Ranking is an important component for information

Retrieval system. It is used to measure the importance and behaviour of web pages**.** For such engines, it is essential that the search results not only consist of web pages related to the query terms, but also rank the pages properly so that the users quickly have access to the desired information. Due to the size of web and requirements of users creates the challenge for search engine page ranking [7]. PageRank has been developed by Google and is named after Larry Page, Google's co-founder and president [3].The PageRank algorithm at Google is one of the successful algorithms that quantify and rank the importance of each web page. Google, one of the world's most popular search engines, state that PageRank is an important part of their ranking

Function [3].This algorithm was initially proposed in [7], and an overview can be found in, e.g., [13], [6].

In this paper, we follow the approach in which user actions are considered. When the visitor search query in search engine then the search engine result list get displayed on the screen. Results are according to the ranking algorithm, that search engine is following. It is very much crucial to get most appropriate webpage at the top of the list. When a visitor open a webpage then it performs certain action for e.g. save, download, print, add to bookmarks/add to favorites, copy certain text. In proposed work there are three main feature as follows: First, consider the action performed by visitor on each and every webpage. Second, check the priority given to each action. Third, based on priority rank is updated by incorporating previous rank. Fourth, websites with duplicate links get higher priority than others.

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 3, Issue 2, pp: (998-1004), Month: April - June 2015, Available at: www.researchpublish.com

The organization of this paper is as follows: In Section II, we provide an overview of related work. In Section III, we provide an overview of the Weighted PageRank algorithm, its link ratio. In section IV, we propose the concept of duplicate links. In section V, we discuss the overview of the gini index of SPRINT algorithm. Then,we discuss the proposed algorithm by finding relation of duplicate links and gini index in section VI. In section VII, we are extracting the conclusion.

II. RELATED WORK

The relevancy of a web page is calculated by search engines using page ranking algorithms. Most of the page ranking algorithm use web structure mining and web content mining to calculate the relevancy of a web page. In [12], the standard Weighted PageRank algorithm is being modified by incorporating Visits of Links(VOL).The proposed method takes into account the importance of both the number of visits of inlinks and outlinks of the pages and distributes rank scores based on the popularity of the pages. So, the resultant pages are displayed on the basis of user browsing behavior. In [16], For ordinary keyword search queries, topic sensitive PageRank scores for pages satisfying the query using the topic of the query keywords is computed. For searches done in context (e.g. when the search query is performed by highlighting words in a Web page), topic-sensitive PageRank scores using the topic of the context in which the query appeared is computed. In [17], a new page rank algorithm is introduced which uses a normalization technique based on mean value of page ranks. The proposed scheme reduces the time complexity of the traditional Page Rank algorithm by reducing the number of iterations to reach a convergence point.

III. WEIGHTED PAGERANK ALGORITHM

The Weighted PageRank algorithm (WPR), an extension to the standard PageRank algorithm, is introduced, this material can be found in, e.g.,[1], [4], [13], [6].

WPR takes into account the importance of both the inlinks and the outlinks of the pages and distributes rank scores based on the popularity of the pages in [15].Unlike standard PageRank, it does not evenly distribute the page rank of a page among its outgoing linked pages. The page rank of a web page is divided among its outgoing linked pages in proportional to the importance or popularity (its number of inlinks and outlinks). Win (v, u), the popularity from the number of inlinks, is calculated based on the number of inlinks of page u and the number of inlinks of all reference pages of page v as given in equation 1.

$$W^{in}(v,u) = \frac{I_u}{\sum_{p \in R(v)} I_p} \qquad (1)$$

Where Iu and Ip are the number of inlinks of page u and p respectively. R(v) represents the set of web pages pointed by v. Wout (v, u), the popularity from the number of outlinks, is calculated based on the number of outlinks of page u and the number of outlinks of all reference pages of page v as given in equation 2.

$$W^{out}(v,u) = \frac{O_u}{\sum_{p \in R(v)} O_p} \qquad (2)$$

Where Ou and Op are the number of outlinks of page u and p respectively and R(v) represents the set of web pages pointed by v. The page rank using Weighted PageRank algorithm is calculated by the formula as given in equation 3.

$$PR(u) = (1-d) + d \sum_{v \in B(u)} PR(v) W^{in}(v,u) W^{out}(v,u) \qquad (3)$$

Traditional Weighted PageRank algorithm works on two parameters-inlinks and outlinks.

IV. DUPLICATE LINKS

In page ranking algorithm , RatioRank is discussed which contains the inlink weights and outlink weights,this material can be found in [5],[9]. This algorithms treat all links equally when distributing rank scores. Every website owner wants its site to be at topmost in search result list. All commercial website owner like travel, online booking, online shopping etc.

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 3, Issue 2, pp: (998-1004), Month: April - June 2015, Available at: www.researchpublish.com

want to gain more profit. This is possible if their home page, customer care no, contact details are easily available to the visitors. So to facilitate this we provide the concept of duplicate links. Duplicate links means providing same links at the top and the bottom of the page. In this section, we propose the concept of duplicate links. We come across certain websites which are very lengthy when we see it till the last of the same webpage. So when visitor wants to move to website inlinks then they scrollup which is inconvenient. So to make inlinks more user friendly to visitors, inlinks can be provided both at the top and bottom of the webpage. Till now if a website provides an outlink for other website then rank of outlink url increases. Now in this proposed work we are proposing that if a website is providing an outlink to another website then former website should also get benefited. so if it provides duplicate outlinks for another website then rank of former website should increase by factor of 2.Also for each duplicate inlinks rank of former webpage should increase by factor of 2.Till now inlinks and outlinks are considered to be equal so page ranking increase by 1.Now in proposed work if a webmaker uses this concept of duplicate links then webpage with duplicate links get more ranking than single link.

Suppose the ratio of duplicate inlinks and duplicate outlinks as $W^{din}(u)$

So the formula for duplicate in link is as follow:

$$W^{din}(u) = \frac{\text{Duplicate inlinks}}{\text{Duplicate outlinks}} \qquad (4)$$

V. SPRINT

A decision-tree-based classification algorithm, called SPRINT [9], [17], [14], [15]. It removes all of the memory restrictions, and is fast and scalable.The algorithm has also been designed to be easily parallelized, allowing many processors to work together to build a single consistent model.Full form of SPRINT is A Scalable Parallel Classifier for Data Mining. This material can be found in [14].In [14], gini index is used which is as follows:

$$\text{Gini(S)} = 1 - \Sigma p_j^2 \qquad (5)$$

−Where p_j is the relative of class j in S

VI. PROPOSED METHODOLGY

In this section we discuss the proposed algorithm. Action performed by user is used to improve the ranking of search results. When a visitor open a webpage then it performs certain action for e.g. save, download, print, add to bookmarks/add to favorites, copy certain text. So for this we evaluate the weightage of action per webpage. That calculated weightage is then incorporated with the previous weightage of that corresponding webpage. Priority is provided to each action as follows:

Action performed	Priority
save/download/Print	1
add to favorites/add to bookmarks	2
copy text	3

Figure.1 Priority based on user action

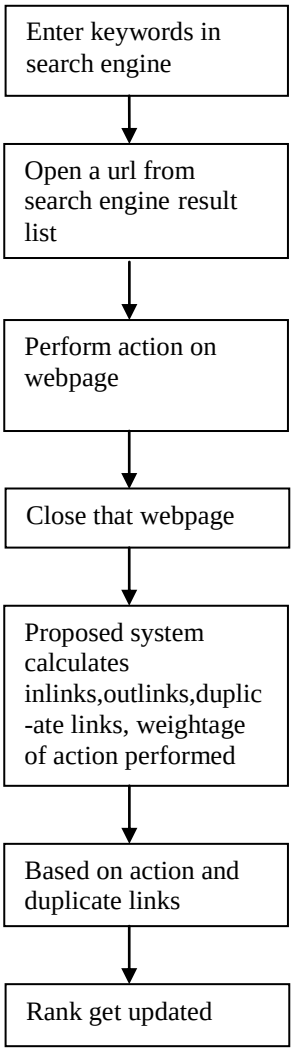
As we can see from the above table. Priority of save/download/print is highest among all action while priority of copy text is lowest. This weightage will be calculated by formula of gini index as follow:

$\text{Gini(S)} = 1 - \Sigma p_j^2$ by using (5)

In the set of gini index we provide the actions. Based on the priority of action,weightage is assigned to P.

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 3, Issue 2, pp: (998-1004), Month: April - June 2015, Available at: www.researchpublish.com

Flow chart for the proposed methodology is as follows:



For checking the duplicate inlinks formula is as follows:

$$W^{din}(u) = \frac{\text{Duplicate inlinks}}{\text{Duplicate outlinks}}$$ by using (4)

By combining (4) and (5), Proposed algorithm will become

$$PR'(u) = \left[(1-d) + d \sum_{v \in B(u)} PR(v) W^{in}(v,u) W^{out}(v,u)\right] \times W^{din}(u) \times Gini(S)$$

The proposed algorithm works on four parameters-inlinks, outlinks, duplicate inlinks, gini weightage of visitor action.

VII. CONCLUSION

Page Ranking is an important component for information retrieval system. It is used to measure the importance and behavior of web pages.

1. Consider the action of visitor on webpage.
2. Better accuracy of rank.
3. User friendly in case of large website.
4. Reliable system.

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 3, Issue 2, pp: (998-1004), Month: April - June 2015, Available at: www.researchpublish.com

Comparison of previous Ranking Algorithm with proposed Ranking algorithm (ASE-WPRA)

Algorithm	Page Rank	Weighted Page Rank(WPR)	Enhanced Weighted Page Rank Algorithm
Main Technique	Web Structure Mining	Web Structure Mining	Web structure and Web content mining
Methodology	This algorithm computes the score for pages at the time of indexing of the pages.	Weight of web page is calculated on the basis of input and outgoing links and on the basis of weight the importance of page is decided.	Weight of web page is calculated on basis of inlinks, outlinks, duplicate links, weightage of action.
Input Parameter	Back Links	Back Links and Forward links.	Inlinks, outlinks, duplicate links, weightage of action performed.
Relevancy	Less(this algo,rank the pages on the indexing time)	Less as ranking is based on the calculation of weight of the web pages at the time of indexing.	Duplicacy and weightage of action concept so relevancy is more.
Quality of results	Medium	Higher than PR	Better than WPR
Importance	High. Back Links are considered.	High. The pages are sorted according to the importance.	High
Limitation	Result come at the time of indexing and not at query time	Relevancy is ignored.	It will take more time as it take 4 parameters instead 2 (in case of WPR)

Following is the graph in which accuracy in performance of existing and proposed page ranking method is compared.It depicts that proposed comes out to be better than existing page ranking method.

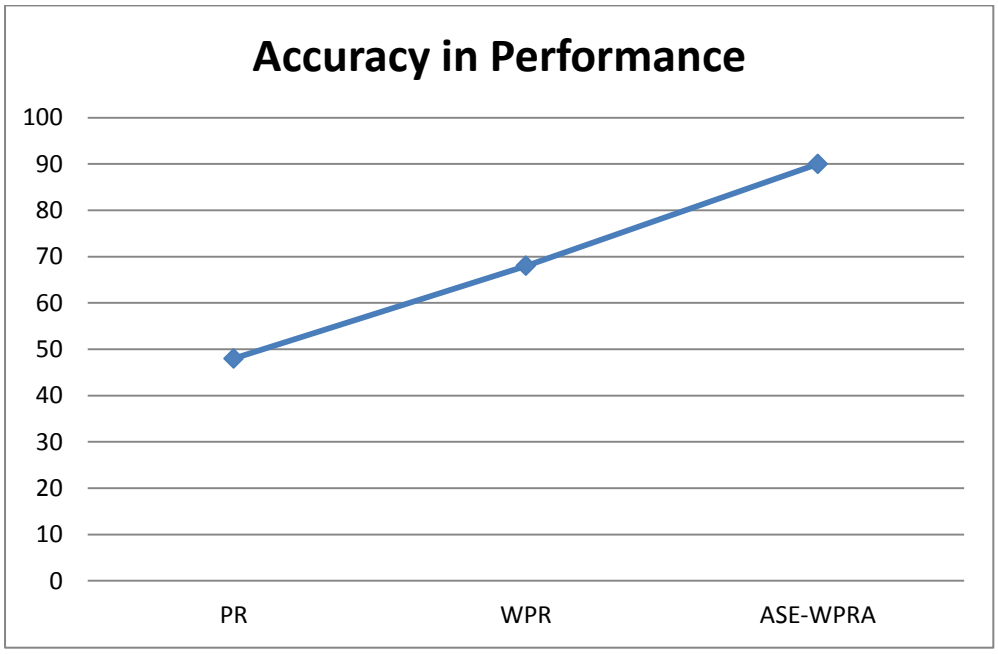


Figure.2 Accuracy in performance

Where PR=PageRank Algorithm

WPR=Weighted PageRank Algorithm

Following is the graph in which efficiency of existing and proposed page ranking method is compared.It depicts that proposed comes out to be better than existing page ranking method.

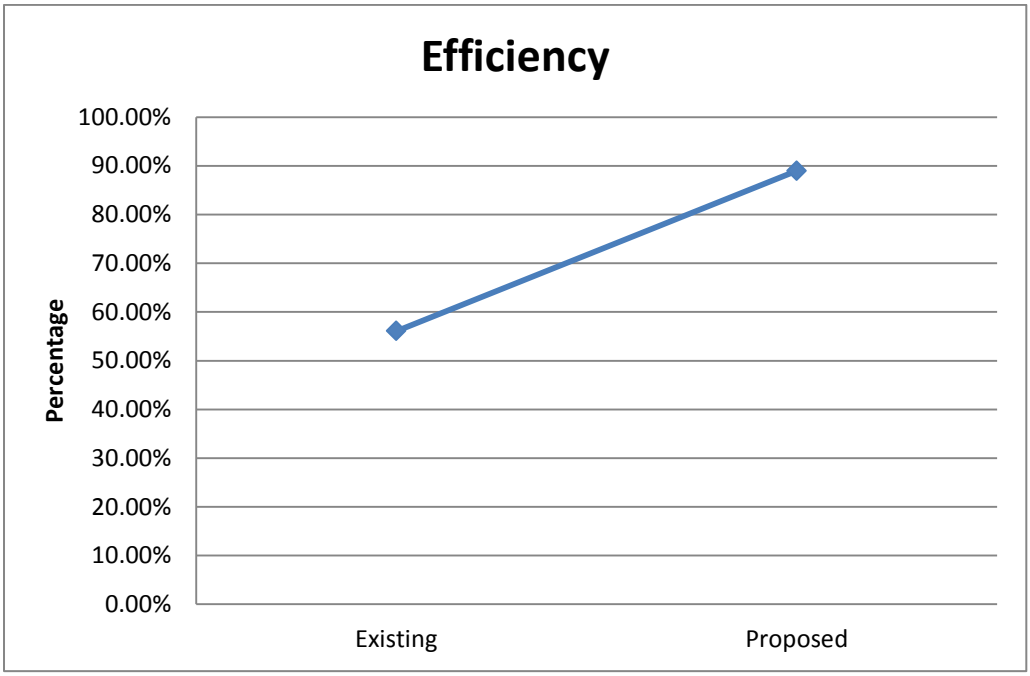


Figure.3 Efficiency graph

REFERENCES

[1] J. Wang, Z. Chen, L. Tao, W. Ma, and W. Liu. Ranking user's relevance to a topic through link analysis on web logs. WIDM, pages 49-54, 2002.

[2] S. Brin and L. Page, "**The anatomy of a large-scale hypertextual Web search engine**," Computer Networks ISDN Syst., vol. 30, pp. 107–117, 1998.

[3] L. Page, S. Brin, R. Motwani, and T. Winograd. **The Pagerank citation ranking: Bring order to the web**. Technical report, Stanford University,1998.

[4] Wenpu ,A.Xing;Ghorban,"**Weighted PageRank algorithm**", Communication Networks and Services Research,2nd Annual IEEE Conference,DOI:10.1109/DNSR.2004.1344743,pages 305-314,2004.

[5] Singh,R.;Sharma,D.K.,"**Information And Communication Technologies**," IEEE Conference,DOI:10.1109/CICT. 2013.6558107,2013,Pages 287-291.

[6] A. N. Langville and C. D. Meyer, **Google's PageRank and Beyond**: The Science of Search Engine Rankings. Princeton, NJ: Princeton Univ. Press, 2006.

[7] B. Liu. **Web Data Mining: Exploring Hyperlinks**, Contents, and Usage Data. Springer, 2006.

[8] Singh,R.;Sharma,D.K.,"Advance Computing Conference(IACC),"IEEE 3rd International,2013,pages 794-799.

[9] Yan-Wen Wu;Li,Li;Zhao,Sheng-Yi;Ai,Xue-Yi, "**Application of Improved SPRINT Algorithm in the Graduation Design Process Management** ," IEEE conference, Pages 252 – 255,2007

[10] Neelam Duhan, A. K. Sharma, Komal Kumar Bhatia, "**Page Ranking Algorithms: A Survey**," Advance Computing Conference,IACC IEEE International,2009.

[11] D. J. Fifield. **Distributed tree construction from large data-sets**. Bachelor's Honours Thesis, Australian National University, 1992.

[12] Sonal Tuteja, "**Enhancement in Weighted PageRank Algorithm Using VOL**,"IOSR Journal of Computer Engineering (IOSR-JCE), vol. 2, issue 6, Sept-Oct 2013 , pp. 135-141.

[13] K. Bryan and T. Leise, "The $25,000,000,000 eigenvector: **The linear algebra behind Google**," SIAM Rev., vol. 48, pp. 569–581,2006.

[14] John Shafer_ Rakesh Agrawal Manish Mehta,"**SPRINT: A Scalable Parallel Classi_er for Data Mining**,"IBM Almaden Research Center,CA 95120.

[15] D. J. DeWitt, S. Ghandeharizadeh, D. A. Schneider, A. Bricker, H.-I. Hsiao, and R. Rasmussen. " **The Gamma database machine project**," IEEE Transactions on Knowledge and Data Engineering,March 1990, pages 44-62.

[16] Taher H. Haveliwala," **Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search**," IEEE Transactions On Knowledge And Data Engineering, VOL. 15, NO. 4,pages 784-796 JULY/AUGUST 2003

[17] Hema Dubey ,Prof. B. N. Roy," **An Improved Page Rank Algorithm based on Optimized Normalization Technique**," International Journal of Computer Science and Information Technologies,Vol. 2 (5), pages 2183-2188 ,2011.