AN IMPROVEMENT OF WEIGHTED PAGE CONTENT RANK TO HANDLE THE LINK SIMILARITY

Riddhi A. Butala[1], Mrs. Pooja Mehta[2]

[1]*M.E. Student, Dept. of CE, Gujarat Technological University, Gujarat, India*
[2]*Assistence Professor, Dept. of CE, Gujarat Technological University, Gujarat, India*

ABSTRACT

The World Wide Web consists of millions of the web pages that are interconnected to each other. Day by day the growth of the World Wide Web is increasing very rapidly. With rapid growth of web it becomes very difficult to provide the relevant information in response to user query. The search engines help the user to surf the web. Due to the vast number of web page it is highly impossible to provide the proper, relevant and quality information. Thus web search engines need efficient ranking algorithm, so that the user could retrieve the web page which is most relevant to user query. In this paper, propose an improvement of weighted page content rank to handle the link similarity as well as also handled the zero link similarity.

Keywords -- *Page Rank, Weighted Page Content Rank, similarity, TF-IDF, etc…*

I. INTRODUCTION

The web mining is technique of data mining. The web mining techniques is classified in various three categories: WSM (Web Structure mining), WCM (Web Content Mining) and WUM (Web Usage Mining) [4] [3] [1]. The WCM is finding the useful information from web content, WSM is finding the relationships between web pages by analysing web structures, and WUM is finding the user profiles and the users' behaviour recorded inside the web logfile [4]. WWW is a huge resource of hyperlink and heterogeneous information including text, image, audio, video, and metadata. It is estimated that WWW has expanded by about 2000% since its evolution and is doubling in size every six to ten month [2] [1]. So day by day increasing the size of World Wide Web people wont to something new, and they may find the information on internet. So the people get the results of number of relevant and irrelevant web documents. Information retrieval (IR), ranking is the main issue of the web mining.

In this paper is organized as follows: Section II presents the related work and explains the various PageRank and its variants in more detail. Section III introduces an proposed methodology of improvement to the weighted Page Content Rank, and Section IV shows some experiment results for the my proposed methodology. And in Section V shows conclusion.

II. RELATED WORKS

Page Ranking algorithms are the soul of search engine and they give the best result of the user expectation. User need of the best quality results are main reason in innovation and improvement of different page ranking algorithms like Page Rank, HITS, Weighted Page Rank, SimRank, Page Rank based VOL, Weighted Page Rank based VOL, Weighted Page Rank based Zero Link Similarity. Now a day's Google search engine is very important because many web users is used.

S. Brin and L. Page [2] was proposed page rank algorithm at Stanford University. Now a day's page rank algorithm was used by very popular search engine GOOGLE. The main concept of page rank, marching the text value of query and find the overall score of web page and it utilize the link to improve the search result. The main goal of page rank is improve the quality of search engine [2]. PageRank is a very good way to prioritize the results of web keyword searches. Page rank is also help for full text searches in main Google system. Advantage of this algorithm is High quality results, backlink predictor, advertising business, frequently indexing. Disadvantage of this algorithm is False page rank or spoof page rank, equal distribution of page rank

Jon Kleinberg [5] introduced Hyperlink-Induced Topic Search (HITS) algorithm, it also known as hubs and authorities. HITS are a link analyses algorithm that rates Web pages. The hubs are serving as large directories that are not actually authoritative in the information that it held, but we used vast catalog of information that lead directly it's called authoritative page [6]. This method

assigns two scores for each page: its authority, which estimates the value of the content of the page, and its hub value, which estimates the value of its links to other pages [7]. So the authority is sum if all score of hub pages and the hub score is sum of all linking pages of authority pages. The main advantage this algorithm is Hub and Authority values are calculated so that the relevant and important pages are obtained. The main drawback of this algorithm is Topic drift and efficiency problems occur, Non-relevant documents can be retrieved

Wenpu Xing and Ali Ghorbani [4] are proposed a Weighted Page Ranking algorithm. Is the improvement over PageRank Algorithm by introducing weighting scheme. In this approach inlink and outlink weights are used to calculate webpage rank value. The proposed a weighted PageRank algorithm which gives more Rank portion to the neighbouring pages with more in-links. It is yet does not sufficiently reflect the actual behaviours of surfers, because only the information of topological structure is used. Strength of this algorithm is that it works offline independent to query. And limitation is, its ranking may be distinguished easily.

S. Qiao, Tianrui Li, Li and Yan Zhu, Jing Peng, Jiangtao Qiu [8] was proposed by SimRank algorithm. This algorithm is variant of weighted PageRank algorithm, called SimRank that distributes Rank value in proportion to the inter-page similarities. To apply the method, all pair-wise page similarities need to be computed earlier on. The main disadvantage of this algorithm is when applying this method for large volume of pages it's computationally expensive.

Gyanendra Kumar, Neelam Duhan, A. K. Sharma [9] was proposed Page Rank based VOL algorithm. Unlike traditional PageRank algorithm, it does not divide page rank value equally between outgoing links. Instead of this it assign more rank value to the outgoing links which is most visited by users. So in this manner page rank is calculated based on visits of inbound links.

Neelam Tyagi, Simple Sharma [10] was proposed by Weighted Page Rank Based VOL algorithm. In the traditional weighted page rank algorithm, it assigned the larger rank value is more popular page. All the outgoing links is proportional to popularity. The number of outlinks and inlinkes popularity will store two function Wout and Win respectively. But in this proposed algorithm, it is not conceder popularity of outgoing link. In proposed improved weighted page rank algorithm it assign the more rank value to outgoing link which is most visited by user. In this WPR (VOL) algorithm it calculated the user browsing behaviours. It calculates the how many time user will visited by link. The limitation of this algorithm is Very ideal but, it is not easy to apply it to the Web scale.

Seifedine Kadry and Ali Kalakech [12] proposed SWPCR algorithm. The simplified Weighted Page Content Rank algorithm is based on two main classes of web mining techniques web structured mining and web content mining. The proposed Simplified Weighted Page Content Rank (SWPCR) algorithm is an improved by very well-known Page Rank algorithm by adding the Content Weight Factor (CWF) scheme for find the important web pages on top of the searching result. The main disadvantage of this algorithm is personalizing relevance ranking. They can't utilize the user information to "personalize" web search engine results.

The Sang-yeon Lee, Young-gi Kim, Seok-Jong Lee, Keon Myung Lee [11] was proposed WPR based Zero Link Similarity algorithm. This algorithm was improved weighted PageRank algorithm that can deal with such zero inter-page similarities, which handles them by allocating a minimum similarity to the links to the pages with the zero-similarity. The proposed algorithm has been implemented using the MapReduce paradigm for big data handling and overcome the problem of simrank algorithm.

III. THE PROPOSED METHODOLOGY

The proposed methodology will base on Weighted Page Content Rank algorithm (WPCR). The proposed methodology handled the link similarity as well as zero link similarity of inter web page using content weight factor. Fig 1 illustrates different components involved in implementation and evaluation of the proposed methodology.

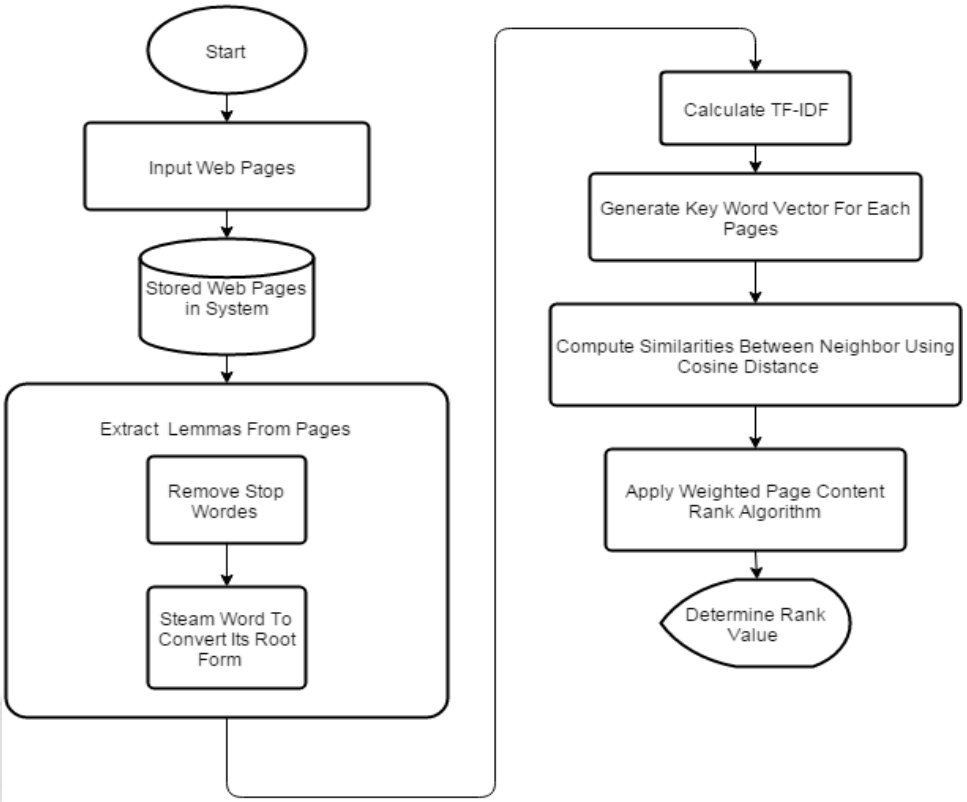

Fig: 1 Architecture Diagram of Proposed Methodology

Based on link similarity measured of inter page firstly I conduct the morphological analysis to extract the lemmas from the web pages. For extracting the lemmas from the web pages, firstly removed the stop word and secondly convert the steam word to its root form. Due to the language characteristics, noun words were extracted to identify keywords. Keywords were identified using the term-frequency and inverse-document frequency information. They determined the number of time the key word appears in the document.

$$TF - IDF = \frac{TF}{\log\left(\frac{DF}{n}\right)} \ldots\ldots\ldots\ldots \text{ Eq. (1)}$$

After the extracting the keywords, applied the weighted page rank algorithm $W_{(i.j)}^{in}$ And $W_{(i.j)}^{out}$ are the weight of *link(i,j) for* calculate the total number of total numbers of links in web page and calculate the inlink and outlink of web pages.

$$W_{(i,j)}^{in} = \frac{I_u}{\sum_{p \in R(i)} I_p} \ldots\ldots\ldots\ldots \text{Eq. (2)}$$

$$W_{(i.j)}^{out} = \frac{O_u}{\sum_{p \in R(i)} O_p} \ldots\ldots\ldots\ldots \text{Eq. (3)}$$

Where, R(i) reference page list of page I, and Iu and Ip are inlink of page u and page p. Ou and Op are outlink of page u and page p.

Generating the keyword vector for each web page. It is an algebraic model for representing the text documents as indexing terms. For k_i and k_j key word vector for page i and j, where $k_i = \{w_{1\,i}, w_{2\,i}, \ldots, w_{n\,i}\}$ and $k_j = \{w_{1\,j}, w_{2\,j}, \ldots, w_{n\,j}\}$

Computing the simulations "between" neighbouring web page using cosine distance. Cosine similarity its gives a useful measure of how similar two web pages are likely to be in terms of their subject matter. Cosine distance of:

$$S_{i\,j} = \frac{k_i \,.k_j}{|k_i||k_j|} \ldots\ldots\ldots\ldots\ldots \text{ Eq. (4)}$$

At last, applied the content Weight Factor CWF(Pi)= GPA(f(Pi)).in the end rank value are determined as in merging the output of both results of similarity of web page and CWF.

$$WPCR = output\ of\ WPR + CWF \ldots\ldots \text{ Eq. (5)}$$

IV. EXPERIMENTS

From the experiments, we analysed the relevancy of page and execution time. Figure 2 shows the experiment result for the existing methodology and proposed methodology of the numbers of

relevant pages are retrieved of a given keyword. We observed that the proposed method WPCR are produced slightly improved the results of compared on existing method of WPR, for handling the links similarity as well as zero links similarity. The relevancy of a page is depended on the keyword. Figure 3 shows the experiment results of the execution time of finding the relevant pages. Execution time was slightly increased when numbers of links are increased. We have tested various set of web pages data sets.

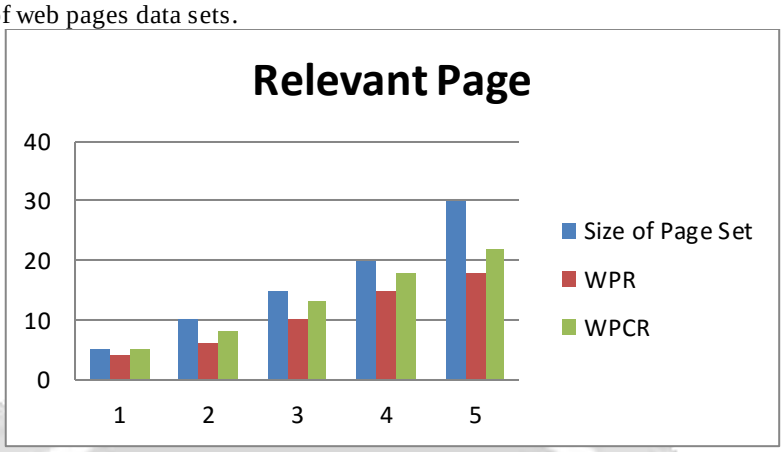

Chart 1: Relevant page Analysis

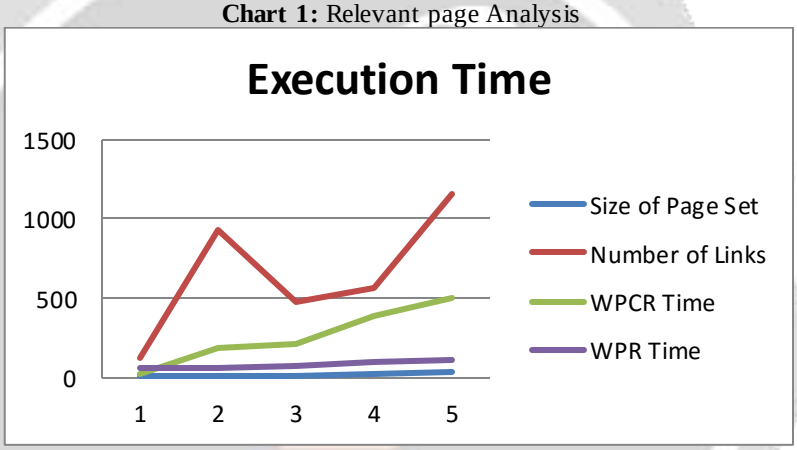

Chart 2: Execution time analysis

V. CONCLUSIONS & FUTURE WORK

Thus I have concluded that improving the weighted Page content Rank algorithm for the link similarity of inter-page as well as also zero link inter-page similarities of neighbouring web pages. To take care of link similarity as well as zero link similarity of inter-page and to adjust the weights of inlinks and outlinks values and also adding the content weight values. My current work is on improving the ranking result of finding the larger number of relevant web pages on top of the search engine result. So my proposed methodology is depending on mainly the WSM and WCM techniques. We observed that proposed method has given an improved based on relevancy page but increased the execution time of finding the number of relevant page.

Thus the future work on this proposed work is decreased the time complexity of finding the relevant web pages and experiments the more number of web pages.

VI. REFERENCES

[1] N. Duhan, A. K. Sharma, K. Bhatia, "Page Ranking Algorithms: A Survey," 2009 IEEE International Advance Computing Conference (IACC 2009).

[2] S. Brin, L. Page, "The anatomy of a large-scale hypertextual Web search engine," Computer networks and ISDN systems, 1998, pp.107-117.

[3] D. Ganeshiya, D. Sharma, "Keyword Ratio Oriented WebPage Rank Algorithm," IEEE Industrial and Information Systems (ICIIS), 2014 9th International Conference on 15-17 Dec. 2014.

[4] Wenpu Xing and Ali Ghorbani, "Weighted PageRank Algorithm," IEEE Proceedings of the Second Annual Conference on Communication Networks and Services Research on 2004.

[5] Chakrabarti, Soumen, Byron Dom, David Gibson, Jon Kleinberg, Prabhakar Raghavan, and Sridhar Rajagopalan. 1998. "Automatic resource list compilation by analyzing hyperlink structure and associated text". In Proc. WWW. URL: citeseer.ist.psu.edu/chakrabarti98automatic.html.

[6]C. Ding, X. He, P. Husbands, H. Zha, and H. Simon, "Link Analysis: Hubs and Authorities on the World". Technical report:47847, 2001.
[7] Christopher D. Manning Prabhakar Raghavan Hinrich Schütze "Introduction to Information Retrieval". Cambridge University Press. 2008. Retrieved 2008-11-09.
[8] S. Qiao, Tianrui Li, Li and Yan Zhu, Jing Peng, Jiangtao Qiu, "SimRank: A Page Rank Approach based on Similarity Measure*," intelligent System and Knowledge Engineering(ISKE),IEEE 2010 international coference on 15-16 Nov.2010
[9] G. Kumar, N. Duhan and A. K. Sharma, "Page ranking based on number of visits of links of Web page," Computer and Communication Technology (ICCCT), 2011 2nd International Conference on. IEEE, 2011.
[10]N. Tyagi, S. Sharma. "Weighted Page rank algorithm based on number of visits of Links of web page," International Journal of Soft Computing and EngineeringVol.2, Issue-3, July2012.
[11] Sang-yeon Lee, Young-gi Kim, Seok-Jong Lee, Keon Myung Lee, "An Improvement of Weighted PageRank to Handle the Zero Link Similarity," IEEE SCIS&ISIS 2014, Kitakyushu, Japan, December 3-6, 2014.
[12] Seifedine Kadry and Ali Kalakech, "On the Improvement of Weighted Page Content Rank," Journal of Advances in Computer Networks, Vol. 1, No. 2, June 2013.