Sentimental Analysis and LSI Similarity Measure for Efficient Page Ranking

S. Anto, S.P. Siddique Ibrahim, S. Siamala Devi

Abstract: -Supplementary factor to the general pagerank calculation which is utilized by Google chrome to rank sites in their web index results is tended to in this paper. These extra factors incorporate couple of ideas which expressly results to build the precision of evaluating the PageRank value. By making a decision about the likeness between the web page content with the text extracted from different site pagesresulted in topmost search using few keywords of the considered page for which the rank is to be determined by utilizing a comparability measure. It results with a worth or rate which speaks to the significance or similarity factor. Further, in a similar strategy if sentimental analysis is applied the search results of the keywords could be analysed with keywords of the page considered, it results with a Sentimental Analysed factor.In this way, one can improve and execute the Page ranking procedure which results with a superior accuracy.Hadoop Distributed File System is used to compute the page rank of input nodes. Python is chosen for parallel page rank algorithm that is executed on Hadoop.

Key-Words: -PageRank, Sentimental Analysis,similarity factor, Hadoop, Python.

I. INTRODUCTION

PageRank[1] is an algorithm used to rank webpages in the search engine results. PageRank is a technique used to calculate the importance of a webpage. In general, the algorithm uses the number of inlinks and number of outlinks along with their importance to calculate the PageRank. Importance can be represented with respect to their weightage where weight can be calculated again through number of inlinks and outlinks initially.

Page rank is an important web graph algorithm which ranks the internet users based on their importance.

Page rank ranks the set of hyperlinks and it suggests the random surfer about the probability of accessing the particular hyperlink. Markov chain is useful in understanding the concept of page rank. The present iteration values are dependent on the previous iteration values. The process should be repeated until the number of iterations reaches the maximum iteration given.

The process can be stopped if the Euclidean distance between successive iterations is less than the predefined threshold value.

A typical example is shown as follows.

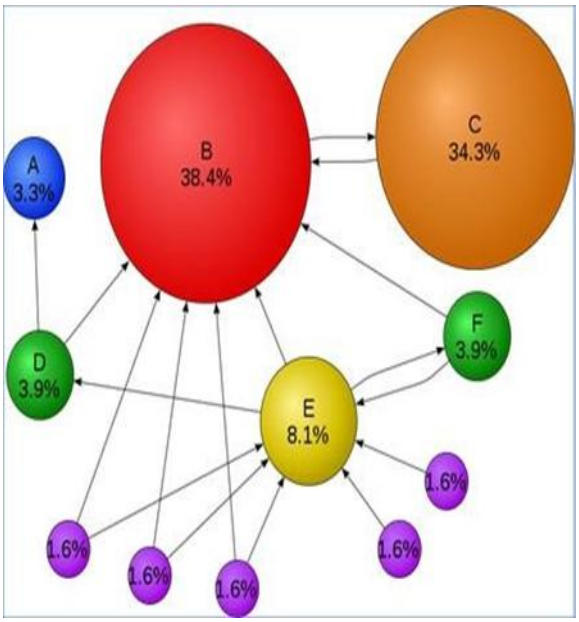


Fig 1. Page Rank Example

The above diagram shows the page rank of set of 7 nodes. Among all these B is given high priority then C, then E and so on.

It means that for a random surfer the probability of selecting the node B, is higher than that of all other nodes present.

The page rank for the above diagram can be computed on desktop or laptops. But it proceeds in a sequential way. If we want to find out the page rank for millions of nodes single computer or laptop is not suitable.

For computing such huge number of nodes, we need parallel as well as distributed systems[10] to increase the efficiency of the program.

Typically, formula for PageRank is represented as follows,

$$PR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)}$$

Here d is damping factor. It is a probability that an imaginary surfer will continue further without stopping.

Generally, the favourable value for ‘d’ is 0.85.

N is the number of nodes present in the set. Pi is page rank of node ‘i’. Pj is page rank of node ‘j’ and L(Pj) is number out going edges of node j.

where, PR is PageRank, a is the damping factor (typically considered 0.85), u is the selected page and v is the page having its inlink towards u.

Revised Manuscript Received on 14 August, 2019.

Dr. S. Anto, Associate Professor, School of Computer Science and Engineering, VIT University, Vellore, Tamilnadu, India.(Email: anto.s@vit.ac.in)

S.P. Siddique Ibrahim, Assistant Professor, Department of Computer Science and Engineering, Kumaraguru College of Technology, Coimbatore, Tamilnadu, India.(Email: siddiqueibrahim.sp.cse@kct.ac.in)

Dr. S. Siamala Devi, Associate Professor, Department of Computer Science and Engineering, Sri Krishna College of Technology, Coimbatore, Tamilnadu, India.(Email: siamalamagesh@gmail.com).

The World Wide Web has developed exponentially in the most recent decade. A lot of website pages are added to the World Wide Web each day. The data present on the web is tremendous and for the most part chaotic and unstructured[3]. Consequently, it turns into a testing work for the search motors to exhibit important data to the clients. The working design of a search engine has appeared in Figure 2. There is a neighborhood store which figures out how to download the pages amid the creeping procedure.

A crawler resembles a web specialist, in charge of visiting URL against the search-key data. The file of these website pages is kept up by the indexer. The record is a gathering of watchwords alongside the pointers to the location of these catchphrases on the web. The inquiry processor forms the client queries and matches the watchwords with the record made by indexer.

The ranking[4] module at that point allocates ranks to these website pages in the request of their relevance and importance with the assistance of a ranking calculation These site pages are then shown in the search motor interface in diminishing request of their relevance and importance.

PageRank was developed by Google's founders [2] in 1998 and it seems to be very partisan that people and search engines still adapt this method to rank pages. Although the algorithm was designed so precise such that it classifies and ranks every page with respect to its importance, there are still few drawbacks in this method (which are to be mentioned in the problem statement section) and to overcome this, there need to be a new enhancement made to the existing PageRank algorithm[5,6] such that it overcomes the drawback which have been noticed.

It can be resolved by adding few factors which lets one to calculate PageRank such that they can even find out if the webpage is worth reading and if the webpage is worth of its uniqueness and if the webpage is worth with respect to its PageRank which is calculated initially using normal PageRank algorithm[5,6].

II. PROBLEM STATEMENT

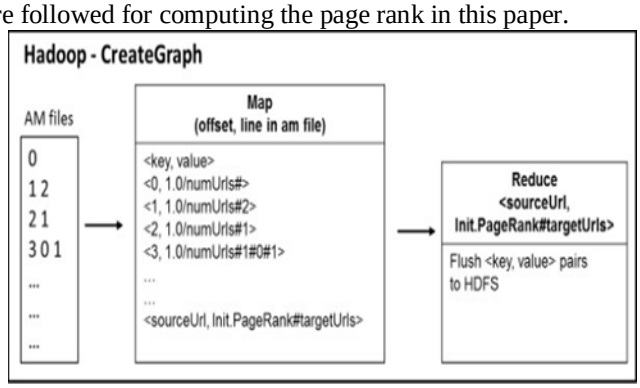
In real time, Page Rank is calculated just with respect to the number of inlinks and outlinks count. Whereas, practically, a page which isn't worthy enough to hold a good page rank also holds a good page rank (for some extent). It happens when a page which strictly does nothing has an inlink from a very standard page which holds a very high page rank. Being a page which holds high value as page rank, its outlink makes a major change in calculating the page rank of the resultant pages which receives the outlink[8].

Thus, in a case if there are only one or two outlinks from the webpage with high page rank, it holds a major contribution while calculating the page rank of the resultant web pages which have an inlink[9] from that page. This leads to a problem where a page which isn't worthy enough is also holding a better page rank that what it actually deserves.

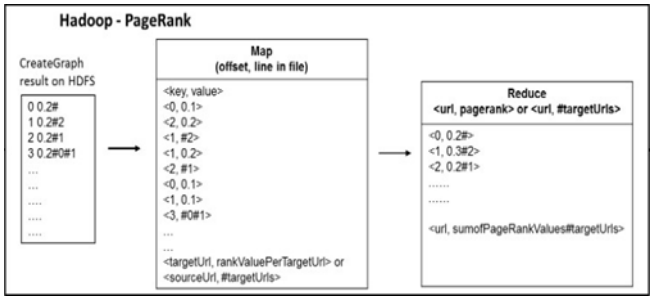
Thus, to overcome this problem, a couple of factors can be added in the calculation of the page rank which increases the standard of page rank and which helps to resolve this drawback.

III. PROPOSED SYSTEM& RESULTS

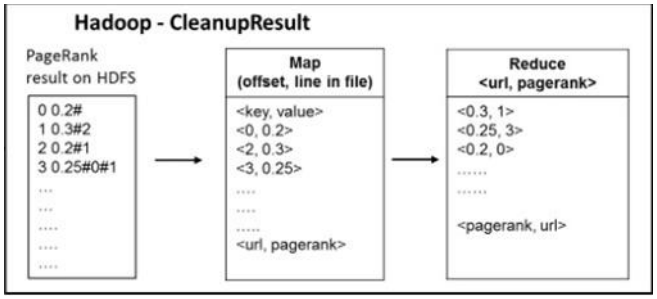
Here, the PageRank calculation includes few factors which leads to a better accuracy by additionally implementing few analysis and comparison. In real time, page rank will be calculated using number of inlinks and outlinks. It leads to a resultant where a webpage which don't even work would get a good page rank if it has an inlink from a page which has a better page rank. The three steps are followed for computing the page rank in this paper.



Step 1: Hadoop Create Graph



Step 2: Hadoop Page rank



Step 3: HadoopCleanup Result

To overcome the flaw of an ineligible webpage to get a good pagerank, we propose a system which applies sentimental analysis[7] i.e. for a considered webpage, we extract few keywords, and from the top 5 web results from the search of that set of keywords, we require a text file.

The text file represents all the relevant data for the page which we want to calculate the page rank. For the required text file, we apply sentimental analysis which gives a resultant scalar between -1 and 1. This can be added as a factor in calculation of pagerank which represents if the considered page is worth getting pagerank which it would get by using general PageRank algorithm.

Also, we consider an another factor which helps to improvise the accuracy pagerank for any considered pagerank. Here, Considering the similar methodology as if it was considered in the above procedure, we extract keywords, analyse top 5 webpages from the internet and then, we apply similarity measure to check if the considered page is worth enough to have a rank which is required using default pagerank algorithm.

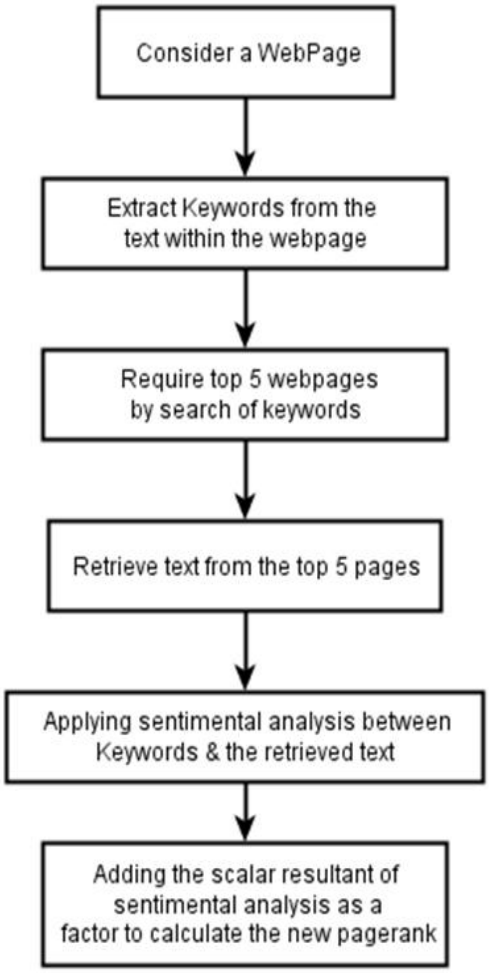


Fig2. Applying Sentimental Analysis

How can one require keywords?

Keywords of a webpage take major part in representing the entire webpage. Retrieving keywords from a given text file follows a sequential procedure. In which, firstly extraction of words from the text takes place which can be done by tokenizing the input text.

From the required tokens, removal of stop words takes place (say removal of prepositions or adverbs etc.), which further follows by a process named stemming in which various verb forms are together considered as a single word mostly, v1 form i.e. first verb form.

Thus, by representing all the verb forms using only a single word, it leads in minimizing the count of words to a large extent. And the required set of words after stemming procedure is said to be set of keywords.

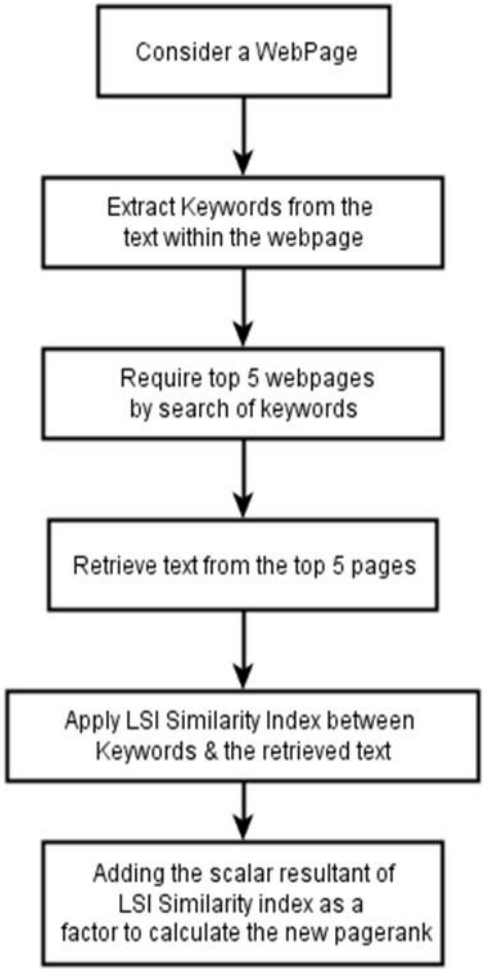


Fig 3. Applying LSI Similarity Index

IV. CONCLUSION

Finally, one can conclude that the considered problem statement will be resolved for a major extent and can establish a new, enhanced pagerank by implementing the considered algorithm. The enhanced algorithm considers both Sentimental analysis and LSI similarity index both as two different factors and the resultant Enhanced pagerank will be calculated as follows,

EPR = PR*0.6 + (SeF)*0.2 + (SiF)*0.2

where, EPR is Enhanced PageRank, PR is PageRank, SeF is Sentimental Factor, SiF is Similarity Factor.To be precise the factors which have been considered are in ratio 3:1:1 for the default pagerank algorithm, Sentimental factor and the Similarity factor. Yet, it can still be modified and set those factors i.e. the ratios in such a way that they can be equipped in the pagerank more precisely.

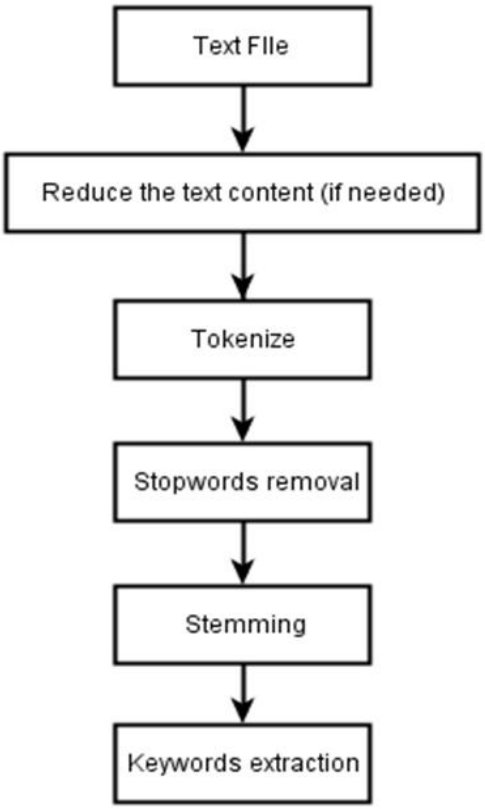


Fig 4. Extraction of Keywords

Thus, it can be concluded that the proposed enhanced PageRank algorithm can be considered as a model which helps out the search engines to overcome the existing drawback and also to precisely calculate the PageRank in a new, modified weighted model such that the common problem noticed will be supressed for a large extent and also, it will not affect the current PageRank algorithm used by google search engines in default.

REFERENCES

1. Prajapati, R., & Kumar, S. (2016, March). Enhanced weighted PageRank algorithm based on contents and link visits. In Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on (pp. 1850-1855). IEEE.
2. Berkhout, J. (2016, May). Google's PageRank algorithm for ranking nodes in general networks. In Discrete Event Systems (WODES), 2016 13th International Workshop on (pp. 153-158). IEEE.
3. Huang, C. C., & Ku, L. W. (2013, December). Interest analysis using semantic PageRank and social interaction content. In Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on (pp. 929-936). IEEE.
4. Ji-Lin, Z., Yong-jian, R., Wei, Z., Xiang-Hua, X., Jian, W., & Yu, W. (2010, December). Webs ranking model based on pagerank algorithm. In Information Science and Engineering (ICISE), 2010 2nd International Conference on (pp. 4811-4814). IEEE.
5. Suzuki, S., Aman, H., Amasaki, S., Yokogawa, T., & Kawahara, M. (2017, August). An Application of the PageRank Algorithm to Commit Evaluation on Git Repository. In Software Engineering and Advanced Applications (SEAA), 2017 43rd Euromicro Conference on (pp. 380-383). IEEE.
6. Agalya, A., B. Nagaraj, and K. Rajasekaran. "Concentration control of continuous stirred tank reactor using particle swarm optimization algorithm." Trans Eng Sci 1, no. 4 (2013): 57-63.
7. Chung, F. (2014). A Brief Survey of PageRank Algorithms. IEEE Trans. Network Science and Engineering, 1(1), 38-42.
8. Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: tasks, approaches and applications. Knowledge-Based Systems, 89, 14-46.
9. Esuli, A., &Sebastiani, F. (2007). Pageranking wordnetsynsets: An application to opinion mining. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (pp. 424-431).
10. Mehta, R., Mehta, D., Chheda, D., Shah, C., &Chawan, P. M. (2012). Sentiment analysis and influence tracking using twitter. International Journal of Advanced Research in Computer Science and Electronics Engineering (IJARCSEE), 1(2), pp-72.
11. Yan, Z., Xing, M., Zhang, D., & Ma, B. (2015). EXPRS: An extended pagerank method for product feature extraction from online consumer reviews. Information & Management, 52(7), 850-8.