Efficient calculation of PageRank using TrustRank and Anti-TrustRank

Jignesh Patel[1], Swati Patel[2], Hiteishi Diwanji[3]

L. D. College of Engineering, Ahmedabad, India
[1]Jignesh.patel1662@gmail.com, [2]swati.ldce@gmail.com, [3]hiteishi.diwanji@gmail.com

Abstract **- Web is collection of millions of pages and Web is growing and changing very rapidly, millions of pages are added and deleted every day in web. The information present on the World Wide Web is of great need, the world is full of questions and the web is the major source of gaining information about user's specific query. As per the web search engine for the query, a millions of pages are retrieved among which the quality of the pages that are retrieved is questioned. Numbers of mathematical algorithms are used for the efficient ranking purpose. In this paper we have explained algorithm which uses content for calculating PageRank called web content mining. In this paper we are trying to reduce the problem of theme drift. For that we have added new parameters TrustRank and Anti-TrustRank. These parameters are used to calculate relevancy and irrelevancy of the page to user query. In our proposed work we use web content mining for calculating the PageRank values of the pages. Anti-TrustRank is measurement of antitrust of page. TrustRank is measurement of trust of page.**

Index Terms **- Web Structure Mining, Web Content Mining, TrustRank, Anti-TrustRank, Spam Page, PageRank**

I. INTRODUCTION

PageRank is a prototype of a large-scale search engine which makes heavy use of the structure present in hypertext [1]. Google is designed to crawl and index the Web efficiently and produce much more satisfying search results than existing systems [1]. PageRank is an algorithm for rating web pages objectively and mechanically. It measures the human interest and attention devoted to them effectively [2].

Web structure mining is used to identify the relationship between Web pages linked by information or direct hyperlink connection. This structured data is discoverable by the provision of web structure schema through database techniques for Web pages. This connection allows a search engine to search data relating to a user's search query directly to the linking Web page from the Web site the content rests upon. This task is accomplished through use of spiders scanning the Web sites, retrieving the home page and linking the information through reference links to bring forth the specific page containing the desired information. The higher the PR value, the more forward of the sorting result: (1) If the more number of the hyperlink pointing to a page, more important of the page, the higher of the PR value; (2) If there is an important hyperlink of the page pointing to another page, the another page is also important and PR value is high [3]. Weights are evenly distributed to all the inlinks of a webpage and to any webpage rank is given high if the sum of weight of inlinks of that webpage is high [4].

The main purpose for structure mining is to extract previously unknown relationships between Web pages. This web structure data mining provides use for a business to link the information of its own Web site to enable navigation and cluster information into site maps. This gives ability to access the desired information through keyword association and content mining to users. Hyperlink structure is also determined to path the related information within the sites to the relationship of competitor links and connection through search engines and third party co-links.

Structure mining uses minimize two main problems of the World Wide Web due to its vast amount of information. The first problem is irrelevant search results. Relevance of search information become misinterpreted due to the problem that search engines often only allow for low precision criteria. The second problem is the inability to index the vast amount if information provided on the Web. This causes a low amount of recollect with content mining. This minimization comes with the function of discovering the model underlying the Web structure provided by web structure mining.

Today internet has become the main source of retrieving information. To find the relevant information regarding a particular topic from the huge amount of data and information present on internet, an efficient searching technique is required. The visitor of a web page tends to visit web pages with similar content rather than content irrelevant pages [5]. The traditional web searching techniques like Keyword-based search adopted by many web search engines are having some advantages like simple, quick and easy to implement but on the other hand they have shortcomings like less precision in results, large set of search results, which is time consuming for user to find relevant information out of it and cannot judge the meaning of the user's query.

There is one more reason for irrelevant search results which comes mainly when the search is performed using keywords stored in meta-tags of the web pages. Many web designers add the keywords in meta-tags irrespective of the content of the web pages to increase the rank of web page so that the link of the web page is shown at the top of the search result. In our work, we are proposing algorithm in which web page ranks are calculated considering the frequency factor of the keywords searched by user's

query and associative factor of those main keywords with the other main keywords. Frequency factor means how many times the keyword is repeated in the content of the web page. By doing this, the above problem is very well sorted as search is performed on the web page content.

II. RELATED WORK

In existing algorithm [6], author has proposed a ranking algorithm for indexing the web pages for effective Semantic information retrieval by ranking the pages. The input of algorithm is the keyword specified in the meta-tags of the web page, set of interrogative words and web pages. The output of algorithm is the rank of the web pages with respect to the keyword and interrogative word. The ranking results can be further used for indexing the web pages according to the ranks calculated. In the indexed database, the links of the web pages, keyword, interrogative word and its rank corresponding to the keyword and the interrogative word is stored. So the indexing of the web pages become very simple which further makes the retrieval more fast and efficient. Hence the interrogative words and main keywords in user's query are searched in the indexed database and where the words matched in the database, the corresponding links are shown in descending order of the ranks.

In our algorithm, we have added two new parameters TrustRank [7] and Anti-TrustRank [8]. TrustRank is measurement of trust of page. Anti-TrustRank is measurement of antitrust of page. These parameters are stored in database. In this algorithm, input parameters are user's query string, web page, Anti-TrustRank and TrustRank of web page. Output is PageRank for web page. TrustRank and Anti-TrustRank values are stored in database for each page. If page contains one of keyword of user's query string in metadata then that page is related to user's query. If keywords are found in web page then web page is not spam page [9]. So TrustRank is increased inversely proportional to number of words in web page. If none keywords are found in web page then web page is spam page. So Anti-TrustRank is increased inversely proportional to number of words in web page. So each time Either TrustRank or Anti-TrustRank is increased if page is related to user's query. Modified TrustRank and Anti-TrustRank values are stored in database again. So it keeps effect of past results. Owner of web page can insert only related keywords in metadata.

Input: Keywords, WP, Q, AntiTrustRank, TrustRank
Output: PageRank [WPj]
Algorithm:

1. for each WPj of set WP do,
2. for each Keywords[k] of set Keywords do,
3. $Freq_k = 0$
4. for each $WP_j[i]$ do,
5. increase $Words_j$ by 1
6. if (Keyword[k] = $WP_j[i]$)
7. increase $Freq_k$ by 1
8. for each Q_p of set Q do,
9. $Dis_p = \infty$
10. for (l=i-5 to i+5) do,
11. if(Q_p= $WP_j[l]$)
12. $temp_p = WP_j[i] – WP_j[l]$
13. Dis_p=minimum(Dis_p, $temp_p$)
14. End for
15. for each Dis_p do,
16. find closeness
17. End for
18. End for
19. End for
20. for each $close_p$ do,
21. if($close_p$!=0)
22. Rank [WP_j, Keywords[K], Q_p] = $close_p$*($Freq_k$ / $Words_j$)
23. End for
24. End for
25. if(spamPage) Increase AntiTrustRank[WP_j]
26. else Increase TrustRank[WP_j]
27. PageRank [WP_j] = calculateFinalPageRank()
28. End for

Let say a keyword K is selected from the set of keywords. For K we select a web page WP on which searching is done. So we scan the whole web page WP comparing each word with K. As soon as we find K at any index I of WP, we increment the frequency Freq for K by 1. If keyword K is found at index I, we try to find other keywords from Q from index i-5 to i+5. If keyword from Q is found at index l then distance (l-i) is measured. For whole web page calculate these distances and find closeness for each keyword in Q. We also calculate the total number of words for web page WP as Words. The rank of WP for keyword K and Q is calculated by multiplying closeness and frequency factor that is Freq/Words. If none keyword K is found in webpage WP then its PageRank value is 0. This webpage is considered as spam page and its Anti-TrustRank value is increased by

1/Words. If even single keyword is found in web page then web page is not spam page and its TrustRank is increased by 1/Words. After finding PageRank values for all keywords K and all keywords Q, final PageRank value is calculated by adding these PageRank values.

III. ANALYSIS

For analysis, we have tested existing and our algorithm 5 times. Existing algorithm gives same value for same query and modified algorithm gives different value according to previous queries due to modification in TrustRank and Anti-TrustRank parameters. Result is displayed in figure 1. From figure 1 we can see modified algorithm considers previous queries from user and existing algorithm gives constant value.

We have tested modified algorithm for 5 different pages 5 times. Results are shown in Table 1. Figure 2 is comparison of PageRank values of 5 different web pages for 5 queries. TrustRank and Anti-TrustRank parameters were 0 for all web pages before analysis.

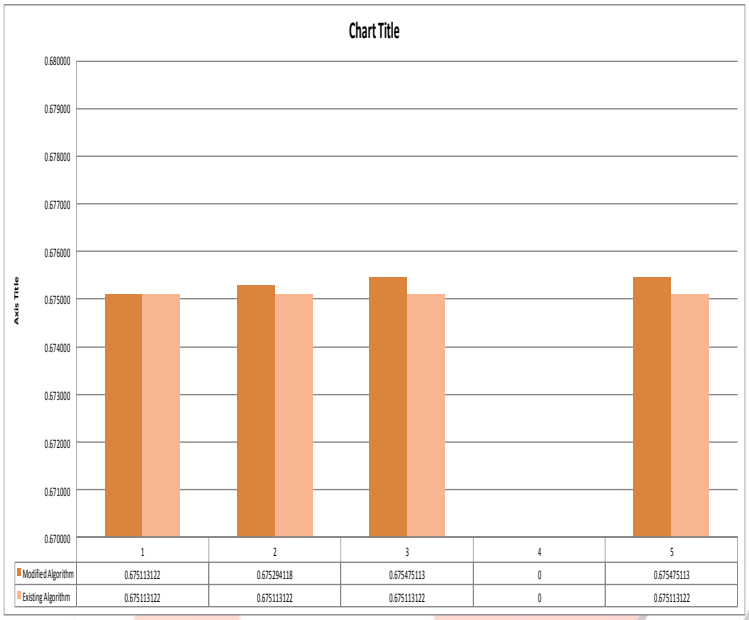

Figure 1 Comparison of Existing algorithm and Modified algorithm

Table 1 PageRank for 5 different web pages for 5 Queries

Query	Sachin.htm	SachinBiography.htm	SachinESPN.htm	SachinYahoo.html	SachinInfo.html
sachin tendulkar	0.675113122	0.656500803	0.08553906	0.140889831	0.9463415
sachin tendulkar	0.675294118	0.658105939	0.08586185	0.141949153	0.9468835
sachin tendulkar	0.675475113	0.659711075	0.08618464	0.143008475	0.9474255
Football	0	0	0	0	0
sachin tendulkar	0.675475113	0.659711075	0.08618464	0.143008475	0.9474255

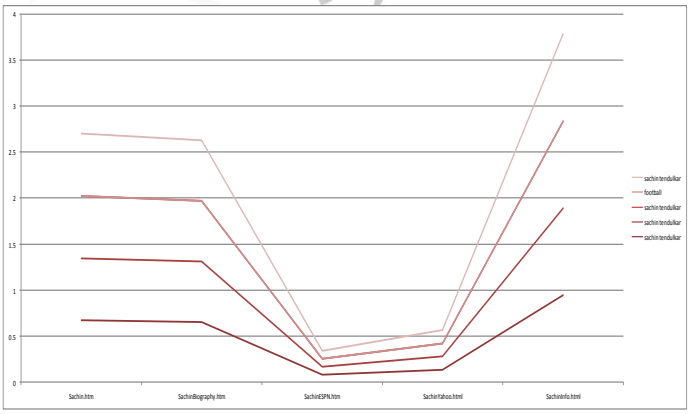

Figure 2 Comparison of PageRank values

During analysis, TrustRank and Anti-TrustRank values were changed, which are shown in Table 2 below. For example for query "sachin tendulkar" for web page "Sachin.htm" TrustRank and Anti-TrustRank values are 0.00018 and 0 respectively. Here "football" is stored in metadata for all web pages. But it is not found in contents of the page. All pages are considered as Spam pages and Anti-TrustRank values were increased for all web pages.

Table 2 TrustRank and Anti-TrustRank values for 5 different web pages for 5 Queries

Query	Sachin.htm	SachinBiography.htm	SachinESPN.htm	SachinYahoo.html	SachinInfo.html
sachin tendulkar	0.00018 / 0	0.00161 / 0	0.00032 / 0	0.00106 / 0	0.00054 / 0
sachin tendulkar	0.00036 / 0	0.00321 / 0	0.00064 / 0	0.00212 / 0	0.00108 / 0
sachin tendulkar	0.00054 / 0	0.0048 / 0	0.00097 / 0	0.00318 / 0	0.00163 / 0
Football	0.00054 / 0.000181	0.0048 / 0.0016	0.00097 / 0.00032	0.00318 / 0.00106	0.00163 / 0.00054
sachin tendulkar	0.00072 / 0.000181	0.00642 / 0.0016	0.00129 / 0.00032	0.00424 / 0.00106	0.00217 / 0.00054

IV. CONCLUSION AND FUTURE SCOPE

In this research, two new parameters TrustRank and Anti-TrustRank are added. TrustRank is value of trustiness of web page. Anti-TrustRank is value of mistrustfulness of web page. This algorithm scans all contents of web page and decrease problem of theme drift. We can find spam pages using Anti-TrustRank values. This algorithm provides web security.

As future work, we can improve accuracy of algorithm by providing synonyms. Algorithm can be modified to use indexing to improve performance. Meta-tags can be used instead of metadata. This algorithm uses text as content. Algorithm can be enhanced to media files, too.

REFERENCES

[1] S. Brin and L. Page, "The Antonomy of a Large Scale HyperTextual Web Search Engine,"7th International WWW Conference Proceedings, Auatralia, April 1998.

[2] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web" Technical report, Stanford Digital Library Technologies Project, 1998.

[3] Zhou Cailan and Chen Kai,Li Shasha," Improved PageRank Algorithm Based on Feedback of User Clicks" IEEE 2011.

[4] Ranveer Singh and Dilip Kumar Sharma," RatioRank: Enhancing the Impact of Inlinks and Outlinks" 3rd IEEE International Advance Computing Conference (IACC) 2013.

[5] Apostolos Kritikopoulos, Martha Sideri and Iraklis Varlamis, "Wordrank: A Method for Ranking Web Pages Based on Content Similarity", Databases, 2007. BNCOD '07. 24th British National Conference 2007.

[6] Robin Sharma, Ankita Kandpal, Priyanka Bhakuni, Rashmi Chauhan, R.H. Goudar and Asit Tyagi, " Web Page Indexing through Page Ranking for Effective Semantic Search", Intelligent Systems and Control (ISCO), 2013 7th International Conference.

[7] Z. Gy ¨ongyi, H. Garcia-Molina, and J. Pedersen, "Combating web spam with trustrank," in Proc. 30th international conference on Very large data bases, vol. 30. VLDB Endowment, 2004, pp. 576–587.

[8] V. Krishnan and R. Raj, "Web spam detection with anti-trust rank," in Proc. ACM SIGIR workshop on Adversarial Information Retrieval on the Web, 2006.

[9] Olivier Fercoq," PageRank optimization applied to spam detection ", Network Games, Control and Optimization (NetGCooP), 2012 6th International Conference.

[10] Bing-Yuan Pu, Ting-Zhu Huang and Chun Wen, "An Improved PageRank Algorithm: Immune to Spam ", Network and System Security (NSS), 2010 4th International Conference 2010.