International Journal of Enhanced Research in Science Technology & Engineering, ISSN: 2319-7463
Vol. 3 Issue 7, July-2014, pp: (387-391), Impact Factor: 1.252, Available online at: www.erpublications.com

Weighted Page Rank Algorithm Based on Number of Visits of Links of Web Pages in Time Duration

Anushree Gambhir[1], Arushi Goyal[2], Sumit Singh[3], Yatharth Srivastava[4]
[1,2 ,3,4]Information Technology Department, JSS Academy of Technical Education, C -20 / 1, Sector 62, Noida, India

Abstract: To support the users to navigate the result list of their search on WWW, various ranking methods are applied on the search results. Most of the ranking algorithms presented till now, do not consider user usage trends over a period of time. In this paper, a page ranking mechanism called Weighted Page Rank Algorithm based on Visits of Links Over a Time Duration (VOLTD) is being devised for search engines, which works on the basis of weighted page rank algorithm and takes number of visits of inbound links of web pages over a prespecified period of time into account . This proposed concept is very useful to display most valuable pages on the top of the result list on the basis of user browsing behaviour, which reduce the search space to a large scale. This algorithm suggests a new approach so that a new website carrying more relevant data well as web pages gets an equal opportunity to appear amongst the legacy websites in the search result.

General Terms: Information Retrieval, page rank.

Keywords: Page rank, visit of links in time duration, search engine, worldwide web, medical domain.

1. INTRODUCTION

The World Wide Web (WWW) is a vast sea of information. WWW is a huge, dynamic and globally distributed information centre, which caters to the endless requests of the users. As on today WWW is the largest information repository for knowledge reference. With the rapid growth of web, the users generally tend to get lost in the wide structured information. So it is essential to provide a mechanism which can serve relevant data to the users. Therefore making websites in a form so as to cater to the users' demands is the primary objective of the website developers. The mechanism of search engines was developed for this purpose. A search engine returns a list of pages, both relevant as well as irrelevant in response to user queries. Search engine returns many pages for a single query. The user is concerned only with the relevant pages; therefore it is essential to impose relevancy check on the search engine result. To assist a user to navigate the search results, ranking mechanism is provided. This ranking mechanism sorts the search results and places the most relevant pages on the top of the search. In this way, the user can find the most important and useful result on the top, followed by the other results.ie-the ranking mechanism ranks all the results in decreasing order of their relevance.

There are a variety of algorithms developed, few of them are Page Rank, HITS, SALSA, RANDOMZE HITS, and SIMRANK etc . A wide research has been performed in this field. Some of the remarkable works are those done by S. Brin and Page [11] who developed Page Rank Algorithm, Wenpu Xing et. al.[12], who developed the Weighted Page Rank Algorithm (WPR) , Gyanendra Kumar et. al[4] who developed the Page Ranking based on Visits of Links(VOL), Chongchong Zhao , Zhiqiang Zhang[2] who proposed a new keyword based approach to improve the web search. In most of these ranking algorithms, considerations of user usage trends are not available. In this paper, a page ranking mechanism called Weighted Page Rank Algorithm based on Visits of Links over a Time Duration (VOLTD) is proposed for search engines, which works on the basis of Weighted Page Rank Algorithm and takes number of visits of inbound links of web pages in a predecided period of time into account. This algorithm suggests rank modification over a predecided period of time, so that a new website launched on the WWW gets an equal chance of appearing on the top of the search list along with the legacy site if the new site has relevant web pages as per the user's query. The details of the proposed algorithm will clarified later in this paper.

2. RELATED WORKS

S. Brin and L. Page [11] developed **Page Rank Algorithm** at Stanford University based on the hyper link structure. PageRank algorithm is used by the famous search engine. The PageRank algorithm is based on the concepts that if a page surrounds important links towards it then the links of this page near the other page are also to be believed as imperative pages. The Page Rank imitate on the back link in deciding the rank score. Thus, a page gets hold of a high rank if the addition of the ranks of its back links is high.

Wenpu Xing et. al.[12] discussed a new approach known as **Weighted Page Rank** Algorithm (WPR). This algorithm is an extension of PageRank algorithm. 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. Each outlink page gets a value proportional to its popularity (its number of inlinks and outlinks). The popularity from the number of inlinks and outlinks is recorded as Win(v,u) and Wout(v,u), respectively. Win(v,u) is the weight of link(v, u) calculated based on the number of inlinks of page u and the number of inlinks of all reference pages of page v.

Gyanendra Kumar et. al. [4] proposed a new algorithm in which they considered user's browsing behaviour. As most of the ranking algorithms proposed are either link or content oriented in which consideration of user usage trends are not available. In this paper, a page ranking mechanism called Page Ranking based on Visits of Links(VOL) is being devised for search engines, which works on the basic ranking algorithm of Google, i.e. PageRank and takes number of visits of inbound links of web pages into account.

Monica Peshave along with **Kamyar Dezhgosha[8]** proposed a project to present the anatomy of a large scale Hypertext Transfer Protocol (HTTP) based Web search engine by using the system architecture of large search engines such as Google, Yahoo as a prototype. The paper describes in detail the basic tasks a search engine performs. An overview of how the whole system of a search engine works is provided.

Neelam Tyagi and Simple Sharma[9] proposed an improved Weighted PageRank algorithm. In this algorithm they assigned more rank value to the outgoing links which is most visited by users and received higher popularity from number of in links. They did not consider the popularity of outlinks which is considered in the original algorithm. The advanced approach in this algorithm was to determine the user's usage trends. The user's browsing behavior can be calculated by number of hits (visits) of links.

Jyoti Gautam and Ela Kumar[6] proposed An Improved Framework for Tag-Based Academic Information Sharing and Recommendation System. The paper proposes a framework for a tag-based Academic Information Sharing and Recommender System which shares information such as question papers, assignments, tutorials and quizzes on a specific area. The approach is based on the set of tags for recommending academic information to each user on the area of his choice. User self-defined tags could be attached to academic information and it can be used further to calculate tag score. The improved index (based on the improved TFIDF algorithm) is used to calculate tag score to give the optimised results.

Chris Ridings and Mike Shishigin[1] provided description and working technology of page rank algorithm. Their paper "PageRank Uncovered" provides a comprehensive study on google's working mechanism.

Chongchong Zhao, Zhiqiang Zhang, Xiaoqin Xie, Xiaoqin Xie, Tingting Liang[2] provided a research paper on A New Keywords Method to Improve Web Search . The contributions of their research are twofold. First, the existing ranking algorithms of search engine are classified. And it extends the expression of queries by "keyword and ", instead of keywords only. Second, a new ranking algorithm based on user feedback and semantic tags is proposed, and it is also compared with Google by several evaluation methods.

Emil Gatial, Zoltan Balogh, Michal Laclavik, Marek Ciglan and Ladislav Hluchy[3] This article proposes a system for focused web crawling used in a domain of job offers. This work is performed within the scope of NAZOU project. The result of this paper focuses on algorithms for web page content analysis in order to specify the depth of crawling and identification of page relevance. In this article, the simple text analysis technique is described, which is based on searching the keywords from the specified domain, its synonyms and prescribed word orders.

Jidong Wang, Zheng Chen , Li Tao, Wei-Ying Ma, Liu Wenyin[5] , this paper proposes to develop a novel approach that utilizes Web logs to compute the relevance of a web-user to a given query. In contrast to traditional methods that are purely based on textual analysis, this approach calculates the web-user's relevance through link analysis under a unified framework where the importance of web-pages and web-users mutually reinforce each other in an iterative way.

3. Weighted Page Rank Algorithm Based On Number of Visit Of Links Of Web Pages In Time Duration (Voltd)

We have seen that the original Weighted PageRank algorithm based on no. of visits of links (VOL) assigns larger rank values to more important (popular) pages without considering the time factor. Each outlink page gets a value proportional to its popularity (its number of inlinks and outlinks). The popularity from the number of inlinks and outlinks is recorded as Win(v,u) and Wout(v,u), respectively. Here we proposed an improved Weighted PageRank algorithm in time duration (VOLTD):

In this algorithm the improvement proposed by us is the incorporation of the time factor along with the number of times the page is visited ie- the number of times a particular page is visited in a given time period is calculated, and the rank is calculated accordingly. As we start with a offline rank updater, we initially allot a rank to the website being crawled .Gradually as the hits on the website increases, the rank is updated over a particular period of time. Rank is updated at a particular time span, so that the new websites launched on the web and having appropriate data as per search are given equal probability to come up the queue, rather than just the old ones.

3.1 Algorithm Suggested

Step 1: Finding a Website
Find a website which has rich hyperlinks because the weighted PageRank and WPR (VOL) methods rely on the web structures.

Step 2: Building a Web Map
Then generate the web map from the selected website.

Step 3: Calculate Win (v,u)
Then calculate the Win (v,u) for each node present in web graph by applying the equation as below.

$$W^{in}_{(u,v)} = \frac{I_n}{I_p}$$

,Where

- Win(v,u) is the weight of link(v, u) calculated based on the number of inlinks of page u and the number of inlinks of all reference pages of page v.
- In and Ip are the number of incoming links of page n and page p respectively.

Step 4: Apply proposed formula
Now calculate the PageRank value of the nodes present in web graph by using the proposed formula:

$$WPR_{vol}(u) = \left\{ \left[(1-d) + d \sum_{v \in B(u)} \frac{L_u WPR_{vol}(v) W^{in}_{(v,u)}}{TL(v)} \right] * t + k \right\}$$

Where

- t represents year under consideration.
- k constant
- u represents a web page,
- d, is the dampening factor.
- WPR (u) and WPR (v) are rank scores of page u and v respectively,
- L(u) denotes number of visits of link which is pointing page u form v.
- TL(v) denotes total number of visits of all links present on v.

Step 5: Repeat by going to step 4
Final step will be used recursively until the values are to be stable.

4. How To Calculate Hits In Time Duration (Voltd)

Here, to count the hits or visits of an outgoing links on a web page in a particular time period, a client side script is used. Whenever a web page is accessed the script will be loaded on the client side from web server and time duration will start counting at that point. Script will monitor the click as well as keyboard event. When an event occurs over the hyperli0nk then it will send a message to web server with information of current web page, date and hyperlink. On server side a data base of log file will be used to record the web page id, hyperlinks of that page, date and hit count of hyperlinks with respect to time duration. Hit count will be incremented every time a hit occurs on the hyperlink. The database or log files will be accessed by crawler at the time of crawling. This (hit count) crawled information will be stored in search engine’s database which is used to calculate the rank value of different web pages or documents as per the time period. In the next duration of time, new hyperlinks will be introduced and the database will be modified as per the number of hits occurred on hyperlinks in which time factor is going to be updated for each hyperlink (including old and new) as per our suggested algorithm.

5. RESULT ANALYSIS

For result analysis, the suggested algorithm was applied to medical domain. The categories which were analyzed under this domain were ear, nose, throat, chest congestion, infection, cancer. Database of over more than ten thousand web links was analyzed. For each category, the relevant, irrelevant and total retrieved documents were retrieved and precision, recall values were calculated for each category. Using the precision and recall values, the F1 value was calculated for each category. The summarized result analysis is presented in a tabular form in the following table:

Table 1: Precision, Recall, F1 values

CATEGORIES	PRECISION	RECALL	F1
EAR	0.800	0.750	0.774
NOSE	0.815	0.720	0.764
THROAT	0.883	0.699	0.760
CHEST CONGESTION	0.851	0.801	0.825
INFECTION	0.836	0.784	0.809
CANCER	0.879	0.709	0.785

As depicted in the table, the precision value lies in between 0.800-0.883 and the recall value lies in between 0.699-0.801. correspondingly, the F1 value lies in between 0.760-0.809.

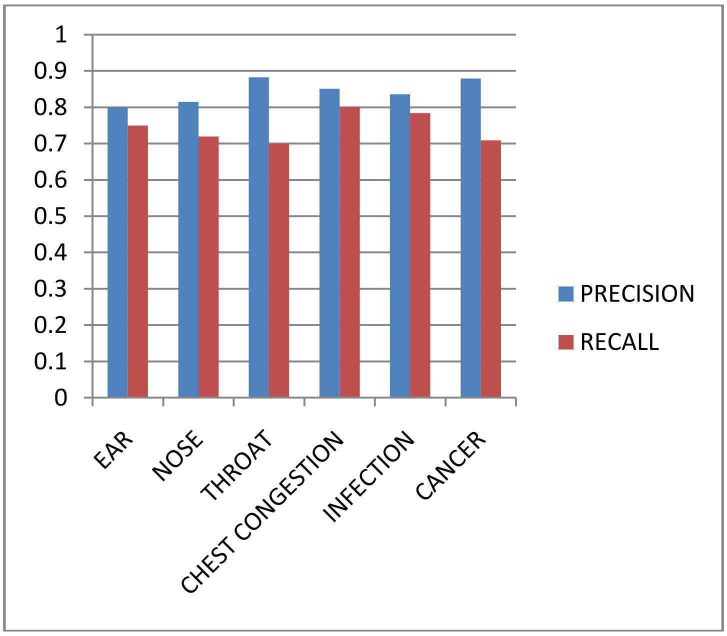

Figure 1: Graphical representation of PRECISION AND RECALL values

6. BENEFITS OF WEIGHTED PAGE RANK BASED ON WPR IN TIME DURATION (VOLTD)

WPR in Time Duration (VOLTD) supplies the most important web pages or information in front of users.
Following are the advantages:

I. As WPR (VOLTD) method uses link structure of pages, the popularity of in links and their browsing information in time duration, the top returned pages in the result list is supposed to be highly relevant to the user information needs.

II. A link with high probability of visit contributes more towards the rank of its out linked pages.

III. The rank value of any page by original Weighted PageRank method will be same either it is seen by user or not. While the ordering of pages using WPR (VOLTD) is more target-oriented.

IV. In WPR(VOLTD), all new websites launched on the web having appropriate data as per search, are given equal probability to come up the queue, rather than just the old ones.

V. In WPR(VOLTD), a user can not intentionally increase the rank of a page by visiting the page multiple times because the rank of the page depends on the probability of visits (not on the count of visits) in a specified time span on back linked pages.

Conclusion and Future Scope

This paper proposes a modified weighted page rank algorithm which considers the number of visit of links as well as the time duration. The suggested algorithm is more specific in nature as compared to the original algorithms suggested. It modifies the page rank based on the number of incoming and outgoing links calculated and modified over a pre specified period of time. The main concern here is the calculation of the incoming and the outbound links. For that purpose the method of recording the number of hits or visits on a particular site is adapted. The proposed algorithm was applied over medical domain for result analysis and the precesion, recall , F1 values were recorded. This algorithm provides better results than the original weighted page rank algorithm as in the suggested algorithm, the inclusion of time domain leads to proving new website an equal opportunity to rank itself amongst the legacy websites. The method of ordering the pages using the suggested algorithm, leads to a better quality results for the user.

Some of the future work in this algorithm includes:

i. The time duration of observation can be varied, to perform a comparative analysis.

ii. Some more improvements can be done, by adding some new experiments to this suggested algorithm.

iii. The experiments can be done taking another domain and analysing the results.

Acknowledgments

We would like to acknowledge Ms. Jyoti Gautam, for her constant support and guidance . Ms. Jyoti Gautam is Associate Professor with the Department of Computer Science and Engineering, JSS Academy of Technical Education, C-20/1, Sector- 62, NOIDA, Uttar Pradesh, India (e-mail: jyotig@jssaten.ac.in).

References

[1]. C. Ridings and M. Shishigin. Page rank uncovered. Technical report, 2002.

[2]. Chongchong Zhao, Zhiqiang Zhang. "A New Keywords Method to Improve Web Search", 12th International Conference on High Performance Computing and Communications, IEEE, (1-3rd Sept, 2010), pages 477-484.

[3]. Emil Gatial1, Zoltan Balogh1, Michal Laclavik1, Marek Ciglan1 and Ladislav Hluchy1- "Focused Web Crawling Mechanism based on Page Relevance" on 23 july 2010.

[4]. Gyanendra Kumar, Neelam Duahn, and Sharma A. K., "Page Ranking Based on Number of Visits of Web Pages", International Conference on Computer & Communication Technology (ICCCT)-2011, 978-1-4577-1385-9.

[5]. J. Wang, Z. Chen, L. Tao, W. Ma, and W. Liu. Ranking user's relevance to a topic through link analysis on web logs.

[6]. Ms Jyoti Gautam and Ms. Ela Kumar on "An Improved Framework for Tag-Based Academic Information Sharing and Recommendation System.." Proceedings of the World Congress on Engineering 2012 Vol II , WCE 2012, July 4 - 6, 2012, London, U.K.

[7]. L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Libraries SIDL-WP-1999-0120, 1999.

[8]. Monica Peshav advised by Kamyar Dezhgosha. How Search Engines Work And A Web Crawler Application."- 2005

[9]. Ms Neelam Tyagi, Ms Simple Sharma- "weighted page rank algorithm based on number of visits of link of web page "- International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-3, July 2012

[10]. Soumen Chakrabarti, Martin van dens Berg, Byron Dom -"Focused crawling: a new approach to topic specific Web Resource discovery". Published by Elsevier Science B.V. in 1999.

[11]. S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 30(1–7):107–117, 1998.

[12]. Wenpu Xing and Ali Ghorbani Faculty of Computer Science University of New Brunswick Fredericton, NB, E3B 5A3, Canada. At Communication Networks and Services Research, 2004. Proceedings. Second Annual Conference on 19-21 May 2004.