

An Improving Web Page Ranking based on Visit of Links with Time Factor and Cursor Movement Algorithm

Bharat Bhushan Agarwal¹, Dr. Mahmoodul. Hasan Khan²

Research Scholar, Dept of Computer Science Engg., Teerthanker Mahaveer University (TMU), Moradabad, India¹

Associate Professor, Department of Computer Science Engineering, I.E.T., Lucknow, India²

Abstract: The web users face the problem in retrieving important and useful information from the large number of disordered and scattered information due to the rapid development of WWW. However, a current search engine does not fully satisfy the user’s need of high-quality relevant information search services. In this paper discussed various Page Rank algorithms to obtain actual page rank of a web page on the web.

Keywords: Page-Rank, Weighted-Page-Rank, Visits of Links, Reading Time, in links, out links.

1. INTRODUCTION

Web mining is the most popular field in today’s scenario where the study is going on fastly. It assimilates the material, collected by the Data-Mining techniques and methodologies in the WWW. Here mining can be understood as the extraction of somewhat important, useful and relevant from a basic material, such as mining diamond from the earth and Data Mining is defined as finding hidden information in a data base. A Data Mining technique used to extract information from WWW is known as Web-Mining.

Web-Mining is defined into various groups depending upon the collected information gathered from WWW as “Web Structure Mining”, “Web Usage Mining” and “Web Content Mining”. First dimension of web mining improves navigation. The second dimension being derived from the server logs. The last dimension provides better resources for visitors by exploring the data contained in the related sites. With the tremendous growth and increasing demand of information on web it has become quite necessary to satisfy the user’s demand, up to the level of his/her expectation. User always expects to get the most relevant results, which with such complex structure and varying queries becomes hard to provide for a search engine. Hence different ranking algorithms like, Page-Rank (PR), Weighted-Page-Rank (WPR), Page-Rank based on Visits of Links (VOL) and Page-Rank based on Visits of Links with Time Factor algorithms are used in different search engines to deal with such problems. The purpose of this paper is to use the concept of Time Factor in combination with cursor movement in Visits of Links (VOL) based on Page Ranking Algorithm (PRA) for finding more useful and relevant information according to the query of the web user’s and to present the most important and useful web pages in the outcome list.

2. RELATED WORK

2.1 PAGE RANK ALGORITHM

The Page Rank Algorithm works the concept of in links.

According to this algorithm a web page has a higher page score, if the total of the rank of its entire in links is more. Page Rank can be calculated as under.

$$PR ('A') = (1-d_f) + d_f (PR (X_1)/C (X_1) + \dots + PR (X_n)/C (X_n))$$

Where,

PR ('A') is the page rank of web page A, (d_f) is the damping factor whose value can be set between 0 and 1. It is used for normalization and usually taken as (.85), (1-d_f) denotes the distribution of the page rank from non directed linked web pages, PR (X₁) denotes back link to web page A, C (X₁) is the number of out bounds from web page X₁.

2.2 WEIGHTED PAGE RANK (WPR) ALGORITHM

After this, its extended version Weighted-Page-Rank (WPR) algorithm was introduced which works on the concept of in links and out links. In this large rank value is assigned to most important or popular web pages instead of being divided evenly among its out linked web pages (as done in Page Rank Algorithm).

The popularity which is estimated from the count of inbound and forward links is recorded as Wⁱⁿ (v,u) and W^{out} (v,u) respectively. Wⁱⁿ (v,u) denotes the weight of link(v, u).

Calculation depends on the count of inbounds of web page u and the count of inbounds links of entire reference web pages of webpage v is carried out by:

$$W^{in} (v,u) = I_u / \sum_{p \in R ('v')} I_p$$

Where,

I_u= counts of inbounds of web page u, I_p= counts of inbounds of web page p, R (v) = reference web page list of web page v, Wⁱⁿ (v,u) shows the link’s weight (v, u), Calculation depending on the count of out bounds of web page u and the count of forward links of entire reference web pages of web page v is carried out by:

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

Where,

O_p = count of forward bounds of page p, O_u = count of out bounds of page u, $R(v)$ = represents the reference web page text of web page, $W^{out}(v,u)$ = weight of out link(v,u).

Seeing the significance of web pages, the original formula of Page Rank formula is revised and is given as under-

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

The purpose of this technique is to make proper and maximum use of the understanding of these internal social assemblies of the web for link mining in application development.

2.3 PAGE RANK ALGORITHM BASED ON VISIT OF LINKS (VOL)

According to this algorithm which is based on number of Visits of Links (VOL) a web-page which is most visited by web users will get higher page rank and will be displayed at the top of the search result.

The improved form of the algorithm is given in the equation below:

$$PR(u) = (1 - d_f) + d_f \sum_{v \in B(u)} L_u(PR(v)) / TL(v)$$

Where,

d_f is the damping factor with value 0.85, u and v are the web pages, B(u) is the set of web pages that link to web page u, PR(u) is the page rank of web page u.

PR(v) is the page rank of web page v, L_u is the number of visits of link that is pointing to web page u from web page v, TL(v) denotes the total count of visits of those links which are present on web page v.

2.4 PAGE RANK ALGORITHM BASED ON VISIT OF LINKS (VOL) WITH TIME FACTOR

According to Page Rank Algorithm (PRA) based on Visit of Links (VOL) with Time Factor, focus is on ranking those web pages at higher position that are most liked by the web user which can be done by the inclusion of Time Factor into the computation of the ranking algorithm. Time Factor (TF) of a web page shows how much a web user likes a web page. Time Factor includes reading time, whereas reading time means the time spent by the user on reading a document. So inclination is towards that web page which is found to be most important for a web user or on which most of the time of the web user is spent.

The algorithm consists of following steps:-

1. Initially, page rank of all the web pages can be assumed to be 1.
2. Find all in links of u, denoted as B(u).
3. Calculate page rank of all web pages by using the following formula:

$$PR(u) = ((1 - d_f)/N) + \{ (d_f \sum_{v \in B(u)} L_u(PR(v)) / TL(v)) \} RT(u)$$

Where,

d_f is the damping factor, u and v represent the web pages,

B(u) is the set of web pages that link to web page u, PR(u) is the page rank of web page u, PR(v) is the page rank of web page v, L_u is the total of visits of link which link to web page u from web page v, TL(v) represents the total number of visits of all links that are present on web page v, RT(u) is the maximum time that the web users, takes to read a web page u, N is the total number of web pages.

Repeat step 3 until web page rank values of two consecutive iterations are same.

2.5 PAGE RANK ALGORITHM BASED ON VISIT OF LINKS (VOL) WITH TIME FACTOR AND CURSOR MOVEMENT

A new algorithm is proposed which is based on Visits of Links with Time Factor and Cursor Movement. The proposed algorithm focus is on the web user query inclination, where that web page is considered which contains most useful or relevant information. Time spent by the web user in reading a document with Cursor Movement by a web user is a necessary factor which helps to decide the significance of a web page.

3. PROBLEM

By simply counting the visits with Time Factor on the url's do not guarantee the significance of a web page for a web user, because reading time or viewing time includes the idle time where the web page is opened up, but the web user is not viewing or reading the web page instead doing other works and because of which most target oriented web pages will not be displayed at the top of the search result. Page Rank Algorithm (PRA) based on Visits of Links (VOL) with Time Factor is unable to focus on the web user query inclination, as a result of which finding or searching the material on the web and retrieving the user's interests and needs from their behavior is a crucial factor. Hence, to improve the precision of ranking score of web pages, a new algorithm is proposed which is an improvement on Visits of Links with Time Factor based on Page Rank Algorithm.

4. METHODOLOGY

For calculating page rank of all web pages by the proposed algorithm which is based on Time Factor and Cursor Movement following formula is used:

$$PR[u] = \left(\frac{1 - d_f}{N} \right) + \left\{ \left(\frac{d_f \sum_{v \in B[u]} L_u(PR[v])}{TL[v]} \right) \right\} \frac{CT[u]}{RT[u]}$$

Where,

d_f denotes the damping factor, u and v denote the web pages, B[u] is the set of web pages that links to web page u, PR[u] is the page rank of web page u, PR[v] is the page rank of web page v, L_u is the total number of visits of link which linking web page u from web page v, TL[v] denotes the total number of visits of all those links which are present on web page v, RT[u] is the maximum time which a web user takes to read a web page u, CT[u] is the cursor movement time of the web page u, where $CT[u] \leq RT[u]$, N is the total number of web pages.

5. RESULT AND DISCUSSION

To test it, consider the same example of hyperlink structure that has been used earlier where maximum amount of time spent by web user on a web page is denoted in seconds along with the count of visits on each link and Cursor Movement time of the four web pages namely, web page A, web page B, web page C and web page D is shown in the below given Fig.1.

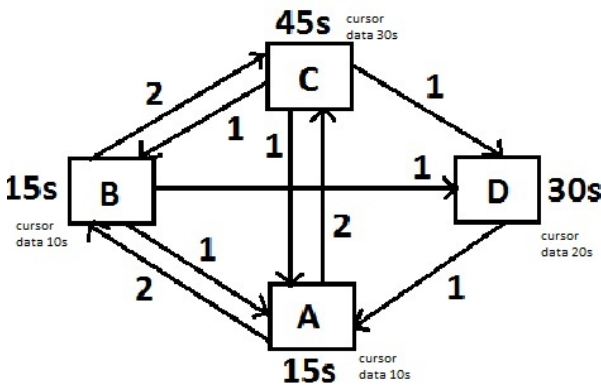


Figure 1: Hyperlinked web graph

The initial scores of Page Rank for all the web pages are assumed as 1 and damping factor value d_f as 0.85. Page rank of web pages A, B, C and D can be computed by using the given equations:

$$PR(A) = ((1 - d_f)/N) + d_f * (CT(A)/RT(A)) * ((PR(B) * LA/TL(B)) + (PR(C) * LA/TL(C)) + (PR(D) * LA/TL(D))) = 0.9365$$

$$PR(B) = ((1 - d_f)/N) + d_f * (CT(B)/RT(B)) * (PR(A) * LB/TL(A) + PR(C) * LB/TL(C)) = 0.2268$$

$$PR(C) = ((1 - d_f)/N) + d_f * (CT(C)/RT(C)) * (PR(A) * LC/TL(A) + PR(B) * LC/TL(B)) = 0.0696$$

$$PR(D) = ((1 - d_f)/N) + d_f * (CT(D)/RT(D)) * (PR(B) * LD/TL(B) + PR(C) * LD/TL(C)) = 0.0823$$

Based on the above rank scores the web pages will be displayed in the order-

Webpage A > Webpage B > Webpage D > Webpage C

Here, page A gets larger rank which shows that page A is preferred which is found most useful for a web user or on which the web user spends maximum time. Figure 2 shows the Snap shot of coding for graph (Figure) 1.

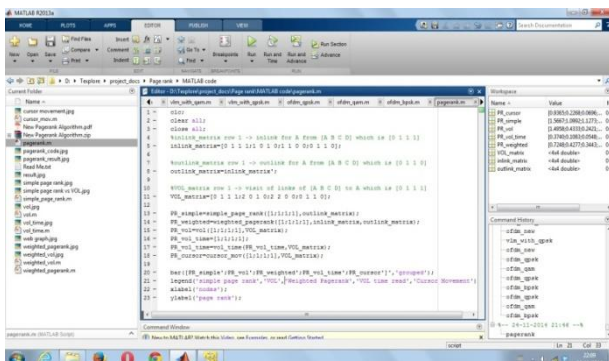


Figure 2: Snap shot for graph 1.

Figure 3 shows the graph which depicts the comparison among the “Page Rank”, “Weighted-Page-Rank” (WPR), and “Page Rank based on Visit of Links” (VOL), “Page Rank based on Visit of Links with Time Factor” and “Page Rank based on Visit of Links with Time Factor and Cursor Movement algorithms”.

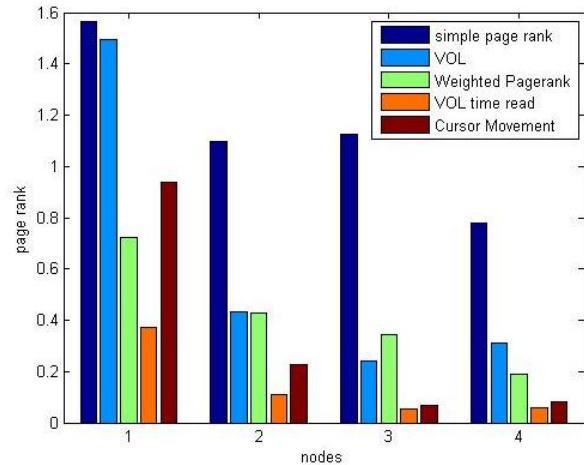


Fig3: Comparison of different algorithms

6. CONCLUSION

This algorithm provides the best results among various ranking algorithms because it is totally based on Cursor Movement including Visits of Links and Time Factor. Cursor Movement includes that actual time in which the cursor is moving on the web page or the web user is scrolling the web page and it removes the idle time from the reading time when the web page is open and the web user is not viewing the web page instead doing some other work. Thus, that web page is preferred which contains relevant information for a web user or on which the web user spends maximum Reading Time and the most useful web page is found on the top of the search results.

7. FUTURE WORK

Web Mining is that technique of data mining which discovers the information automatically from web documents. “Page Rank”, “Weighted-Page-Rank” and “Visits of Links based Page Rank algorithms” are used in Web Structure Mining to rank the relevant web pages. In this paper, focus is on the use of new algorithms which is a combination of concepts of Reading Time factor with Cursor Movement on the web pages by the use of which the web user gets more important, useful and relevant web pages easily on the top of the result list as it uses “Web Usage Mining” and “Web Structure Mining” techniques. As a part of future work, for the Cursor Movement based Page Rank Algorithm, to compute the page rank efficiently, a standard method to be applied on all the scripts could be developed. For example, if the web pages are submitted to Google, then Google should run a script on each and every web page to compute the Reading Time and Cursor Movement time. This will record both the times for every web page on same scale with same standard and hence the page rank can be implemented

efficiently. By this, a standard can be set for ranking the web page on the basis of Cursor Movement. Also, applying the algorithm uniformly will improve the authenticity of the web page rankings improving the web page ranking significantly.

REFERENCES

- [1]. Agrawal bharat bhushan and Dhall shivangi (2010). Web Mining: Information and Pattern Discovery on the World Wide Web, College of Engineering Technology, (I.F.T.M), Moradabad, India, pp.1-8.
- [2]. Agarwal Sweta and Agrawal bharat bhushan (2013). An Improvement on Page Ranking Based on Visits of Links, IJSR, 2(6):265-268.
- [3]. Bharat Bhushan Agarwal, Dr M H Khan (2013) "Analysis of Rank Sink Problem in Page Rank Algorithm" International Journal of Scientific & Engineering Research, Volume 4, Issue 11, pp.251-256.
- [4]. Bharat Bhushan Agarwal, Sonia Gupta (2013) "A Survey: World Wide Web and the Search Engines" International Conference on Advances in Computer Application (ICACA), Proceedings published in International Journal of Computer Applications (IJCA), pp.40-43.
- [5]. Neelam Duhan, A. K. Sharma, Komal Kumar Bhatia, "Page Ranking Algorithms: A Survey" Advance Computing Conference, 2009. IACC 2009 IEEE International.
- [6]. Wenpu Xing and Ghorbani Ali, "Weighted PageRank Algorithm", Proceedings of the Second Annual Conference on Communication Networks and Services Research (CNSR '04), IEEE, 2004.
- [7]. Eirinaki Magdalini and Vazirgiannis Michalis (2005). 'Usage-based page rank for web personalization', Proceeding of the fifth IEEE International Conference on Data Mining, Washington, pp.130-137.
- [8]. Gupta Ankur and Jindal Rajni (2008). 'An overview of ranking algorithms for search engines', Proceedings of the 2nd National Conference for Nation Development, India, pp .335-340.
- [9]. Kadry Seifedine and Kalakech Ali (2013). On the improvement of Weighted Page Content rank, Journal of Advances in Computer Networks, 1(2):110-113.
- [10]. Mukhopadhyay Debajyoti, Giri Debasis and Singh Sanasam Ranbir (2003). An approach to confidence based page ranking for user oriented web search, ACM SIGMOD Record, 32(2):28-33.
- [11]. Bharat Bhushan Aggarwal & Sumit Prakash Tayal, "Software Engineering", University Science Press, 2007.
- [12]. Bharat Bhushan Agarwal and Sumit Prakash Tayal, "Data Mining and Data warehousing", Laximi Publications Ltd., 2009.