International Journal of Engineering Works
Kambohwell Publisher Enterprises
ISSN: 2409-2770 Vol. 4, Issue 2, PP. 41-48, February 2017
www.kwpublisher.com

# Extended Weighted Page Rank Based on VOL by Finding User Activities Time and Page Reading Time, Storing them Directly on Search Engine Database Server

Isha Mahajan, Sachin Gupta, Ms. Harjinder Kaur, Dr. Darshan Kumar

***Abstract*—** Searching on the web can be considered as a process of user enters the query and search system returns a set of most relevant pages in response to user's query. But results returned are not mostly relevant to user's query and ranking of the pages are not efficient according to user requirement. In order to improve the precision of ranking of the web pages, after analyzing the different algorithms like Page Rank, Weighted Page Rank, Page Rank based on VOL, Weighted Page Rank algorithm based on VOL. In this paper, we are proposing enhancement by including "User Activities Time" and "Page Reading Time" in Weighted Page Rank based on VOL algorithm ($WPR_{VOL}$). Page Reading Time (PRT) is the total time page remains focused in browser tab. User Activities Time (UAT) is the total time user does activities like Key Press, Mouse Click, Touch the Screen and Scrolling the page etc. $WPR_{VOL}$ Algorithm signifies the importance of a web page for a user and thus helps in increasing the accuracy of web page ranking. Our proposed Extended Weighted Page Rank based on Visit of links ($EWPR_{volT}$) algorithm is a page ranking mechanism, which considers user browsing behavior / user using trends into account. Other algorithms discussed in literature are either link or content oriented. $WPR_{VOL}$ has already being devised for search engines, which works very much similar to weighted page rank algorithm and takes number of visits of inbound links of web pages into account. Also we are making one more improvement in our algorithm ($EWPR_{volT}$) by storing the no of visits on links, PRT and UAT information directly on Search Engine database server instead of storing it on client's web server in the form of logs which was suggested in earlier literature. The proposed improvement in algorithm finds more relevant information according to user's query. So, this concept is very useful to display most important and useful pages on the top of the result list on the basis of user usage trends, which reduce the search space to a large scale for user.

***Keywords*—** Weighted Page Rank based on Visits of links, Weighted Page Rank, Page Rank, Page Rank based on visit of links, User Activities Time, Page Reading Time, User Activity Time, reading time, Search Engine, Web Crawler, Crawling, Information Retrieval, World Wide Web, Backlinks, Inlinks, Outlinks, Inbound Links, Outbound Links, Visit of Links.

## I. INTRODUCTION

World Wide Web is growing rapidly day by day, the number of web pages is increasing into millions and billions around the world [11]. With the rapid growth of the Web, users get easily lost in the rich hyperlink structure [9]. Providing relevant information to the users to cater to their needs is the primary goal of website owners. Therefore, finding the content of the web and retrieving the user's interests and needs from their behavior have become increasingly important. For the purpose mentioned it is important to understand and analyze the underlying data structure of web for effective and efficient information extraction with the increasing demand of users [6]. Search Engine has become the most considerable means for people to get the required information or services [10]. So it has become necessary for the search engines to give most specific and user need satisfying results.

There are lots of search engines but few like Google, Yahoo, Bing, AOL, ASK, Baidu, Excite etc. are famous because of their crawling and ranking methodology. All search engines employ ranking methodologies and follow unique strategies to rank a website. Every day they solve and satisfy millions of queries. So Ranking methodology becomes a very important aspect of web mining in all the three components of search engine [11] (i.e. Crawler, Indexer, Ranking mechanism).

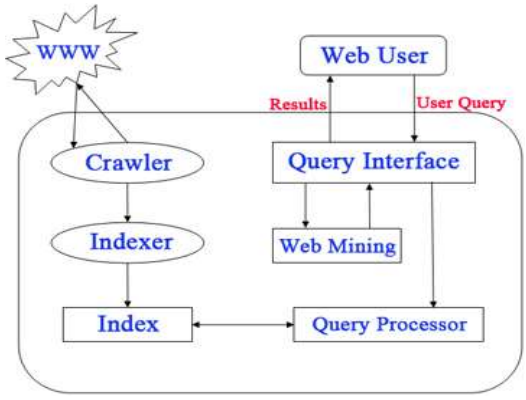


Fig. 1. Architecture of Search Engine [11]

Figure 1 shows the Search engines Query resolution process, which are used to find information from the WWW. They download, index and store hundreds of millions of web pages. They answer thousands of queries every day. They act

like content collector as they keep record of all the information available on www [1].

In web search, ranking algorithms play an important role in ranking web pages so that the user could get the good result which is more relevant to the user's query. When a user makes a query from search engine, it generally returns a large number of pages in response to user queries. The search engines will move through millions of pages and bring the most related results [10]. This result-list contains many relevant and irrelevant pages according to user's query. As user impose more number of relevant pages in the search result-list. To assist the users to navigate in the result list, various ranking methods are applied on the search results. The search engine uses these ranking methods to sort the results to be displayed to the user. In this way user can find the most important and useful result first. There are a variety of algorithms developed, few of them are PAGE RANK, HITS, SALSA, RANDOMZE HITS, SUBSPACE HITS, SIMRANK, WEIGHTED PAGE RANK etc [12]. Most of these ranking algorithms are either link or content oriented in which consideration of user usage trends are not available. There are some other algorithms like Page Rank based on VOL ($PR_{VOL}$), Weighted Page Rank based on VOL ($WPR_{VOL}$) which also considers the user usage trends (i.e. user browsing behavior of websites). So $WPR_{VOL}$ mechanism 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 into account. The original Weighted Page Rank algorithm is an extension to the standard Page Rank algorithm. WPR takes into account the importance of both the inlinks and outlinks of the pages and distributes rank scores based on the popularity of the pages. The main purpose of the proposed algorithm is finding more relevant information according to user's query. So, this concept is very useful to display most valuable pages on the top of the result list on the basis of user browsing behavior, which reduce the search space to a large scale [1].

## II. LITERATURE SURVEY

Data mining can be defined as the process of extracting useful information from large amount of data. The application of data mining techniques to extract relevant information from the web is called as web mining [8]. So the Web mining is a data mining technique used to extract information from World Wide Web. It plays a vital role in web search engines for ranking of web pages and can be divided into three categories: [6]:

i. ***Web Content Mining (WCM):*** It is the process of extracting useful information from the contents of web documents. This mining technique is used on the web documents and results page that are obtained from a search engine. WCM is to mine the content of web pages.

ii. ***Web Structure Mining (WSM):*** It is the processes of discovering link structure of the hyperlinks in inter documents level from the web. It is used in many application areas.

iii. ***Web Usage Mining (WUM):*** It is the process to discover interesting usage patterns from web data in order to understand and better serve the needs of web-based applications. Also WUM to extract information from the web server logs.

## III. RELATED WORK

Following are some ranking algorithms discussed with their varying nature of web mining category, working, and input parameters etc.

### *A. Page Rank Algorithm*

Page Rank was developed at Stanford University by Larry Page and Sergey Brin (PHD Research Scholars, also founders of Google) in 1996. Page Rank uses the hyper link structure of Web [2]. It is based on the concepts that if a page contains important links towards it then the links of this page towards the other page are also to be considered as important pages. It considers the back link in deciding the rank score. If the addition of all the ranks of the back links is large then the page it is provided has large rank [6]. A simplified version of Page Rank is given below:

$$PR(u) = c \sum_{v \epsilon B(u)} \frac{PR(v)}{N_v}$$

Notations used are:

- u and v represents the web pages.
- B(u) is the set of pages that point to u.
- PR(u) and PR(v) are rank scores of page u and v respectively.
- $N_v$ indicates the number of outgoing links of page v.
- c is factor applied for Normalization.

In Page Rank, the rank of page P, is evenly divided among its outgoing links. Later Page Rank was modified observing that not all users follow the direct links on WWW. Therefore, it provides a more advanced way to compute the importance or relevance of a web page than simply counting the number of pages that are linking it [6]. If a backlink comes from an important page, then that backlink is given a higher weighting than those backlinks comes from non-important pages. Thus, the modified version of Page Rank is given as :

$$PR(u) = (1-d) + d(\frac{PR(T1)}{C(T1)} + \frac{PR(T2)}{C(T2)} + \cdots + \frac{PR(Tn)}{C(Tn)})$$

Where, d is a damping factor which set its value to 0.85. d can be thought of as the probability of users following the links and could regard (1 − d) as the page rank distribution from non-directly linked pages. We assume several web pages T1 … Tn which point to u web page. T1 is the incoming link page to page u and C(T1) are the outgoing links from page T1.

$$PR(u) = (1-d) + d \sum_{v \epsilon B(u)} \frac{PR(v)}{N_v}$$

Page Rank algorithm is used by the famous search engine "Google". Page Rank algorithm is the most frequently used algorithm for ranking billions of web pages. During the processing of a query, Google's search algorithm combines pre


ISSN: 2409-2770

computed Page Rank scores with text matching scores to obtain an overall ranking score for each web page [1].

### *B. Weighted Page Rank Algorithm*

Wenpu Xing et. al. [3] discussed a new approach known as weighted page rank algorithm (WPR). This algorithm is an extension of Page Rank 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.

WPR performs better than the conventional Page Rank algorithm in terms of returning larger number of relevant pages to a given query. According to author the more popular web pages are the more linkages that other web pages tend to have to them or are linked to by them. A Weighted Page Rank Algorithm – assigns larger rank values to more important (popular) pages instead of dividing the rank value of a page evenly among its outlink 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 $W^{in}_{(v,u)}$ and $W^{out}_{(v,u)}$.

$W^{in}_{(v,u)}$ given in following equation 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.

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

Where Iu and Ip represent the number of inlinks of page u and page p, respectively. R(v) denotes the reference page list of page v. $W^{out}_{(v,u)}$ given in following equation is the weight of link(v, u) calculated based on the number of outlinks of page u and the number of outlinks of all reference pages of page v.

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

Where Ou and Op represent the number of outlinks of page u and page p, respectively. R (v) denotes the reference page list of page v. Considering the importance of pages, the original Page Rank formula is modified in following equation:

$$WPR(u) = (1-d) + d \sum_{v \epsilon B(u)} WPR(v) W^{in}_{(v,u)} W^{out}_{(v,u)}$$

Notations used are:

- u and v represents the web pages.
- d is the damping factor. Its value is 0.85.
- B(u) is the set of pages that point to u.
- WPR(u) and WPR(v) are rank scores of page u and v respectively.
- $W^{in}_{(v,u)}$ 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 (i.e. outlinks) of page v.
- $W^{out}_{(v,u)}$ is the weight of link(v, u) calculated based on the number of outlinks of page u and the number of outlinks of all reference pages (i.e. outlinks) of page v.

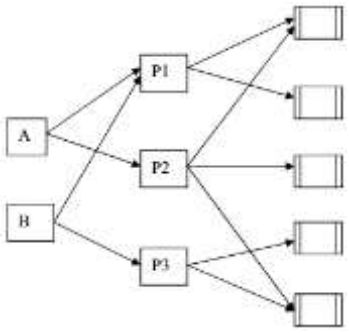


Fig. 2. Web interlinked structure

### *C. Page Rank based on Visits of Links Algorithm*

Gyanendra Kumar et. al. [4] proposed a new algorithm in which they considered user's browsing behavior. As most of the ranking algorithms proposed earlier are either link or content oriented in which consideration of user usage trends are not available. They propose in their paper, a page ranking mechanism called Page Ranking based on Visits of Links ($PR_{VOL}$) is being devised for search engines, which works on the basic ranking algorithm of Google, i.e. Page Rank and takes number of visits of inbound links of web pages into account. This concept is very useful to display most valuable pages on the top of the result list on the basis of user browsing behavior, which reduce the search space to a large scale. In this paper as the author describe that in the original Page Rank algorithm, the rank score of page p, is evenly divided among its outgoing links or we can say for a page, an inbound links brings rank value from base page p. So, he proposed an improved Page Rank algorithm. In this algorithm we assign more rank value to the outgoing links which is most visited by users. In this manner a page rank value is calculated based on visits of inbound links.

The modified version of Page Rank based on VOL is given in following equation:

$$PR_{vol}(u) = (1-d) + d \sum_{v \epsilon B(u)} \frac{PR_{vol}(v) L_u}{TL(v)}$$

Notations used are:

- u and v represents the web pages.
- d is the damping factor. Its value is 0.85.
- B(u) is the set of pages that point to u.
- $PR_{vol}(u)$ and $PR_{vol}$ (v) are rank scores of page u and v respectively.
- $L_u$ is the number of visits of link which is pointing page u from v.
- TL(v) denotes total number of visits of all links present on v.

Isha Mahajan: Student, ishamahajan90@gmail.com, Pathankot, 7508736757
Sachin Gupta: sgs.engineer@gmail.com, Pathankot, 9988639937
Ms. Harjinder Kaur: Department Head at SSIET, harrysaini988@gmail.com
Dr. Darshan Kumar: Director at SSIET, Dinanagar



ISSN: 2409-2770

*D. Weighted Page Rank based on Visits of links Algorithm*

Neelam Tyagi et. Al. [1] discussed that the original Weighted PageRank algorithm assigns larger rank values to more important (popular) 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 recorded as $W^{in}_{(v,u)}$ and $W^{out}_{(v,u)}$., respectively. Here we proposed an improved Weighted Page Rank algorithm. In this algorithm we assign more rank value to the outgoing links which is most visited by users and received higher popularity from number of inlinks. We do not consider here the popularity of outlinks which is considered in the original algorithm. The advanced approach in the new algorithm is to determine the user's usage trends. The user's browsing behavior can be calculated by number of hits (visits) of links.

The modified version based on $WPR_{VOL}$ is given in following equation:

$$WPR_{vol}(u) = (1-d) + d \sum_{v \epsilon B(u)} \frac{WPR_{vol}(v) W^{in}_{(v,u)} L_u}{TL(v)}$$

Notations used are:

- u and v represents the web pages.
- d is the damping factor. Its value is 0.85.
- B(u) is the set of pages that point to u.
- $WPR_{vol}(u)$ and $WPR_{vol}$ (v) are rank scores of page u and v respectively.
- $W^{in}_{(v,u)}$ 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 (i.e. outlinks) of page v.
- $L_u$ is the number of visits of link which is pointing page u from v.
- TL(v) denotes total number of visits of all links present on v.

## IV. Problem Definition

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) and WPR (Weighted Page Rank) are used in different Search Engines to deal with such problems but fail to focus on the user query preference, therefore finding the content of the Web and retrieving the user's interests and needs from their behavior is a crucial factor. There are algorithms for Search Engines like $PR_{VOL}$ (Page Rank based on Visits of Links) and $WPR_{VOL}$ (Weighted Page Rank based on Visits of Links) which considers user's interests and needs from their behavior. These algorithms are missing a crucial factor for how long a webpage has been actually used by users. This can be a great factor to decide the Significance of page. This User Activities Time (UAT) actually reflects the usefulness of information in the page as conceived by the user. This User Activities Time is the total time user has done activities on page like mouse move, keys press, page touched, page scrolled etc. This UAT can be calculated by subtracting Idle Time from Page Reading Time. Page Reading Time (PRT) is the total time page has been actively opened in browser tab. In technical terms, we can say PRT is the total time duration for which focus was on that page.

Storing Visit of Links (VOL) information on client's web server in the form of log was proposed in earlier approach by Neelam Tyagi [1], creates extra work to be done by crawler. As crawler will crawl that information along with other data first from Web Server Logs and then only It will be stored in search engine database, so search engine will not always have fresh information. New fresh information will be available only to search engine when crawler will crawl that page again. So Search Engine will have to depend on crawler for new information. Also the worst case could be the information can be manipulated by client when present on his server to improve its rankings. More over there can be hardware failure or malicious attacks on Client web server, which can lead to loss of information. These drawbacks can be removed by storing Visit of Links (VOL), Page Reading Time (PRT) and User Activities Time (UAT) information directly on search engine database server.

## V. Our Approach

To enhance the quality of search and to display the most target oriented pages at the top of the search list, we propose a new approach which focuses on the user query preference, where consideration is done on the most useful or important pages. To determine the usefulness of pages, we take time spent on webpage i.e. User Activities Time (UAT) and Page Reading Time (PRT) as an essential factor along with Visits of Links (VOL) on that webpage. These all will decide the importance of a page. User Activities Time is the actual time spent by the user to read the webpage, which we suppose reflects the usefulness of information in the page as conceived by the user. This proposed approach will compute the rank according to visits of links of inbound links as well as user attention given to the web page. This algorithm behaves completely different from other page ranking algorithms because it takes users usage trends or user browsing behavior in its working.

Our modified version based on $WPR_{VOL}$ is given in following equation:

$$EWPR_{volT}(u) = (1-d) + d \left[ \frac{UAT(u)}{PRT(u)} \left( \sum_{v \epsilon B(u)} \frac{EWPR_{volT}(v) W^{in}_{(v,u)} L_u}{TL(v)} \right) \right]$$

Notations used are:

- u and v represents the web pages.
- d is the damping factor. Its value is 0.85.
- B(u) is the set of pages that point to u.
- UAT(u) is the User Activities Time i.e. total time user actually spends on that webpage by doing activities like cursor movement with mouse, Key Press, Touch etc.



ISSN: 2409-2770

- PRT(u) is the Page Reading Time i.e. the time page has been actively opened in browser tab. In more technical words, we can say when Focus is on that page.
- $EWPR_{volT}(u)$ and $EWPR_{volT}$ (v) are rank scores of page u and v respectively.
- $W^{in}_{(v,u)}$ 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 (i.e. outlinks) of page v.
- $L_u$ is the number of visits of link which is pointing page u from v.
- TL(v) denotes total number of visits of all links present on v.

### *A. How to calculate the Visits of Links (VOL), Page Reading Time (PRT) and User Activities Time (UAT).*

To count the hits or visits of an outgoing links on a web page a client side script is used. A client side script will be given to webmaster of website to add in his website (Just like Google Analytics), that script will track the User Activities Time (UAT), Page Reading Time (PRT) and Links activity data(to calculate VOL) and sends this data to our search engine hosting server where these data will be stored. Whenever a web page (containing our Analysis script) will load on system, the Our Analysis script will be loaded. This Script will start tracking various events like screen touch, key press, mouse move, click, page scroll etc. This script will calculate the idle time i.e. the time when no such event (mentioned above) happened and this script will also subtract the idle time from active page focus time. When a mouse click etc event happen over hyperlink / hotspot or Webpage got closed then our Analysis Script will calculate the UAT and PRT and it will send a message to web server of Search Engine (hosting Analysis Script). This Message will contain information of current web page, hyperlink and at the same time it sends the UAT, PRT. JavaScript, JQuery and AJAX are the most famous client side scripting languages.

In Earlier Approach, suggested by Neelam Tyagi et. al. [1]. She suggests, on client's server a data base or log file will be used to record the web page id, hyperlinks of that page and hit count of hyperlinks. Hit count will incremented every time a hit occur on hyperlink. The database or log files will be accessed by crawler (i.e. Bot of Search Engine [7]) 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.

In our enhanced approach instead of storing the link visit count activity and UAT, PRT information on client server, we will save that information on the server of Search Engine.

Benefits of storing the information on Search Engine database or Server rather on Client Server are follows:

- It will also save crawler's work, as crawler no longer need to crawl such data for search engine. Therefore reduces extra work or overhead for crawler.
- Search engine will always have updated data, in the earlier approach, as crawler will crawl the data first and then data will be stored in search engine. Data will be updated only when crawler will provide data next time, until then outdated data.
- Having data on our server gives us more control to access the data than on client web server.
- Also having data on our server means it is safe from any manipulation from client. Client cannot tamper with data.
- No risks of data loss imagine if the client web server goes down due to any reason like Hardware failure or malicious attacks. Data can be lost with the client website. Data loss risks can be minimized by storing data on our server.
- It will certainly save space on client web server. Client might need extra space to save data in logs which is no longer required. Also it will save the cost for client.
- We can easily analyze the data and take back up of data for future needs.

To avoid the large value set of user activities time, active page reading time values as well as to less complicate or simplify our calculations, the values of UAT, PRT which will be sent to the server will be compared with the last updated UAT, PRT values (if exists), if the new UAT, PRT values will be larger than already existing values then the existing values will get replaced with the new values in database. If this User n Activities Time and Page Reading Time for that webpage does not exists in database, new UAT, PRT values are stored to database.

### *B. Code to calculate the Visits of Links (VOL), Page Reading Time (PRT) and User Activities Time (UAT).*

```
<script type="text/javascript">

//Variable Declaration and Initialization
var timeoutID = 0;
var start, end, pagefocustime = 0;
var opentime, midtime = 0;
var exacttime = 0;
var inactivetime = 10000, idletime = 0;
var window_focus = true;

//WHEN DOCUMENT READY
document.addEventListener("DOMContentLoaded",
function(event) {
  start = performance.now(); //Recording Page Load Time
  midtime = start;

  //When off Focus
  window.onblur=function(){
       window_focus = false;
       end = performance.now();
       pagefocustime += end - midtime;
       exacttime += end - midtime;
  };

  //When on Focus
```



ISSN: 2409-2770

```
window.onfocus=function(){
    window_focus = true;
    midtime = performance.now();
};


if(window_focus){
    function setup() {
        this.addEventListener("mousemove",
resetTimer, false);
        this.addEventListener("mousedown",
resetTimer, false);
        this.addEventListener("keypress",
resetTimer, false);
        this.addEventListener("DOMMouseScroll",
resetTimer, false);
        this.addEventListener("mousewheel",
resetTimer, false);
        this.addEventListener("touchmove",
resetTimer, false);
        this.addEventListener("MSPointerMove",
resetTimer, false);
        startTimer();
    }

    setup();

    function startTimer() {
        // wait 10 seconds before calling goInactive
        timeoutID = window.setTimeout(goInactive,
10000);
    }

    function resetTimer(e) {
        window.clearTimeout(timeoutID);
        goActive();
    }

    function goInactive() {
        idletime = idletime + inactivetime;
        if(window_focus){startTimer();}

    }

    function goActive() {
        startTimer();
    }

}   //end of if focus condition


//MAKE AJAX CALL to  DATABASE SERVER -
WITH REQUIRED SENDING INFORMATION
window.onbeforeunload=function(){
    end = performance.now();
    pagefocustime += end - midtime;
    opentime = end - start;
    exacttime = pagefocustime - idletime;

    //AJAX CALL to SERVER

    // CALL SETUP
    if (window.XMLHttpRequest) {
        xmlhttp=new XMLHttpRequest();
     } else {
        xmlhttp=new
ActiveXObject("Microsoft.XMLHTTP");
     }
    xmlhttp.onreadystatechange=function()
    {
    if (xmlhttp.readyState==4 && xmlhttp.status==200)
        {
        console.log(xmlhttp.responseText);
        }
    }

    //SENDING INFORMATON WITH AJAX
    if (document.referrer != "") {
        xmlhttp.open("GET","http://<Domain
Name>/<filename>?url="+location.href+"&caller_url="
+document.referrer+"&page_focus_time="+pagefocustim
e+"&exact_time="+exacttime, true);
        xmlhttp.send();
    }
    else{
        xmlhttp.open("GET","http://<Domain
Name>/<filename>?page_focus_time="+pagefocustime+
"&exact_time="+exacttime, true);
        xmlhttp.send();
    }

};

});
</script>
```

## RESULT ANALYSIS

To explain the working of original and our proposed extended algorithm, let's take an example of Hyperlink Structure that consists of four web pages A, B, C and D with maximum user activities time (UAT) and maximum page reading time (PRT) on a page is mentioned in seconds along with number of visits of each link as shown in fig 3.


ISSN: 2409-2770

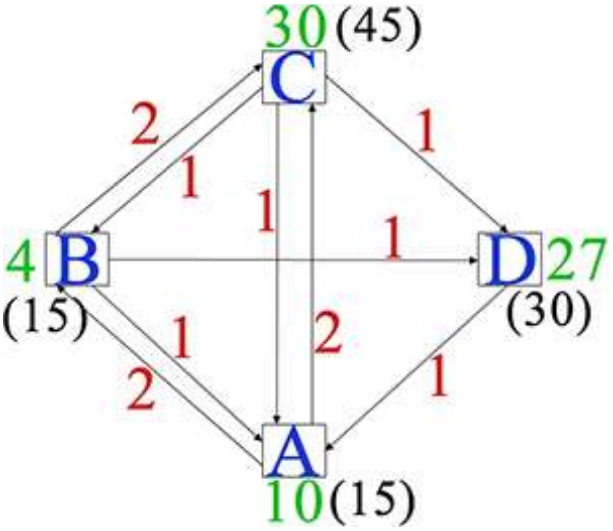


Fig. 3. World Wide Web hyperlink Structure with No. of visits of Links, User Activities Time and Page Reading Time

Name of web pages A, B, C and D are indicated with blue color in fig 3. No. of visits of links are indicated with red color in fig 3. User Activities Time is provided with green color in fig 3. Page Reading Time is given in Black color in fig3.

Let us assume the initial page rank of all pages as 1 and the value of damping factor d is 0.85. The rank for pages A, B, C and D can be calculated by the following equation:

$$EWPR_{volT}(u) = d + (1-d)\left[\frac{UAT(u)}{PRT(u)}\left(\sum_{v\in B(u)} \frac{EWPR_{volT}(v)W^{in}_{(v,u)}L_u}{TL(v)}\right)\right]$$

| Iterations | Webpage A | Webpage B | Webpage C | Webpage D |
|---|---|---|---|---|
| **1** | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| **2** | 1.1719 | 0.4434 | 0.4839 | 0.2916 |
| **3** | 0.4774 | 0.4382 | 0.4528 | 0.2161 |
| **4** | 0.4102 | 0.2881 | 0.3046 | 0.2132 |
| **5** | 0.3821 | 0.2618 | 0.2721 | 0.1921 |
| **6** | 0.3591 | 0.2532 | 0.2630 | 0.1879 |
| **7** | 0.3540 | 0.2476 | 0.2570 | 0.1866 |
| **8** | 0.3519 | 0.2460 | 0.2553 | 0.1858 |
| **9** | 0.3510 | 0.2454 | 0.2546 | 0.1856 |
| **10** | 0.3507 | 0.2452 | 0.2544 | 0.1855 |
| **11** | 0.3506 | 0.2451 | 0.2543 | 0.1855 |
| **12** | 0.3506 | 0.2451 | 0.2543 | 0.1855 |

Table 1. Iterative Calculation for $WPR_{VOL}$ Algorithm without UAT, PRT.

According to the results shown in Table 1, the web pages will be display in search results in the following order -

**Page A > Page C > Page B > Page D**

Here, web page A gets higher rank and will be display in top of search results. A have 3 inbound links over others A,B,C with 2 inbound links.

| Iterations | Webpage A | Webpage B | Webpage C | Webpage D |
|---|---|---|---|---|
| **1** | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| **2** | 0.8313 | 0.2283 | 0.3726 | 0.2774 |
| **3** | 0.3412 | 0.2052 | 0.2862 | 0.1896 |
| **4** | 0.2853 | 0.1755 | 0.2149 | 0.1821 |
| **5** | 0.2754 | 0.1708 | 0.2046 | 0.1752 |
| **6** | 0.2707 | 0.1700 | 0.2028 | 0.1742 |
| **7** | 0.2700 | 0.1697 | 0.2021 | 0.1741 |
| **8** | 0.2699 | 0.1697 | 0.2020 | 0.1740 |
| **9** | 0.2698 | 0.1697 | 0.2020 | 0.1740 |
| **10** | 0.2698 | 0.1697 | 0.2020 | 0.1740 |

Table 2. Iterative Calculation for $EWPR_{volT}$ Algorithm with Visits of Links, User Activities Time and Page Reading Time.

According to the results shown in Table 2, the web pages display in search results in the following order -

**Page A > Page C > Page D > Page B**

Here, also page A gets higher rank, so will be display in top of search results list. Here result shows Page D comes before Page B because user activities time (UAT) of Page D is more than Page B, both pages B and D have same inbound links. It means here in scenario, preference is given to that page which is found useful for a user or on which user spends most of the time and also has most back links.

By applying various algorithms on the web graph shown in Figure 3, we get variations in algorithms which are shown in figure below:

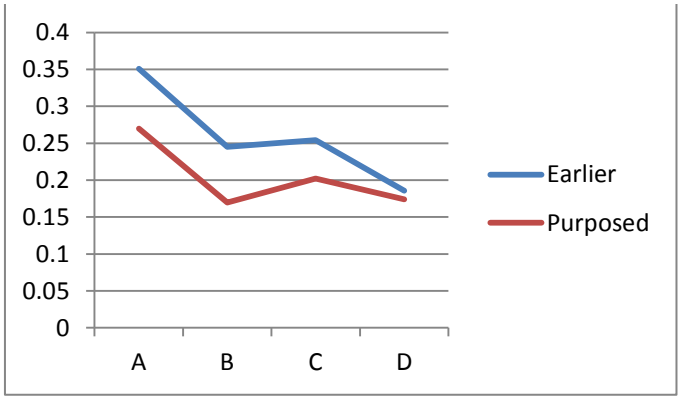


Fig. 4. Variations in Earlier and Purposed Work

## CONCLUSION & FUTURE WORK

Our proposed ranking algorithm provides the best results among other ranking algorithms because it is totally based on User browsing behavior (i.e. Total no. of Visits on Links, Page Reading Time and User Activities Time). This User Activities Time is the total time user has done activities on page like mouse move, keys press, page touched, page scrolled etc. This UAT can also be calculated by subtracting Idle Time from Page Reading Time. Page Reading Time (PRT) is the total time page has been actively opened in browser tab.

Thus, the webpage preferred by this proposed algorithm contains relevant information, highly linked by other pages and mostly used by users in terms of large UAT and PRT. Most


ISSN: 2409-2770

useful web page is found on the top of the search results, reducing the search space for web user.

Web Mining is the data mining technique that automatically discovers or extracts the information from the documents available on World Wide Web. Page Rank (PR), Weighted Page Rank (WPR) uses web structure mining (WSM) to rank the pages whereas other algorithms like Page Rank based on VOL ($PR_{VOL}$), Weighted Page Rank based on VOL ($WPR_{VOL}$) uses both web usage mining (WUM) and web structure mining (WSM) for ranking of the web pages.

Here in this paper, we have focused that other algorithms like Page Rank, Weighted Page Rank, Page Rank Based on Visit of Links and Weighted Page Rank based on Visits of Links may not get the required relevant document easily. To solve this problem, we have used the web page User Activities Time (UAT), Page Reading Time (PRT) to increase the precision or accuracy of web page ranking. Combining the concepts of Page Reading Time factor with User Activities Time factor of the web pages, user gets more important, useful and relevant web pages easily on the top of the result list. Our proposed algorithm ($EWPR_{volT}$) uses both WSM and WUM techniques to give satisfactory results and is in agreement with the applied theory for developing the algorithm. To reduce the risk of data loss, to avoid tempering of data, always to have fresh / updated data and to free the crawler from job of crawling the link of visits and user activities time related data from website server log. We are storing link of visits, page reading time, user activities time related data directly on Search Engine database server.

As a part of future work, to calculate the rank more efficiently and accurately, User Activities Time and Page Reading Time can be made better.

## REFERENCES


[1] Neelam Tyagi and Simple Sharma, "Weighted Page Rank Algorithm Based on Number of Visits of Links of Web Page", International Journal of Soft Computing and Engineering (IJSCE), ISSN: 2231-2307, vol. 2, issue 3, pp. 441–446, July 2012.

[2] S. Brin, and Page L., "The Anatomy of a Large Scale Hypertextual Web Search Engine", Computer Network and ISDN Systems, vol. 30, issue 1-7, pp. 107-117, 1998.

[3] Wenpu Xing and Ali Ghorbani, "Weighted PageRank Algorithm", Proceedings of the Second Annual Conference on Communication Networks and Services Research (CNSR '04), IEEE, 2004.

[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] Tamanna Bhatia," Link Analysis Algorithms For Web Mining ", International Journal of Computer Science and Technology ( IJCST), ISSN : 0976-8491, vol. 2, issue 2, pp. 243-246, June 2011

[6] Shweta Agarwal and Bharat Bhushan Agarwal, "An Improvement on Page Ranking Based on Visits of Links", International Journal of Science and Research (IJSR), ISSN: 2319-7064, vol. 2, issue 6, pp. 265-268, June 2013.

[7] Sachin Gupta, Sashi Tarun and Pankaj Sharma, "Controlling access of Bots and Spamming Bots", International Journal of Computer and Electronics Research (IJCER), ISSN: 2278-5795, vol. 3,issue 2, pp. 87-92, April 2014.

[8] Sonal Tuteja, "Enhancement in Weighted PageRank Algorithm Using VOL", IOSR Journal of Computer Engineering (IOSR-JCE), ISSN: 2278-0661, vol. 2, issue 6, pp. 135-141, Sept-Oct 2013.

[9] Rekha Jain and Dr. G. N. Purohit, "Page Ranking Algorithms for Web Mining", International Journal of Computer applications, ISSN: 0975 – 8887, vol. 13, no. 5, pp. 22-25, Jan 2011.

[10] Sachin Gupta and Pallvi Mahajan, "Improvement in Weighted Page Rank based on Visits of Links (VOL) algorithm", International Journal of Computer & Communication Engineering Research (IJCCER), ISSN: 2321-4198, vol. 2, issue 3, pp. 119-124, May 2014.

[11] Sachin Gupta and Shashi Tarun, "Extended Architecture of Web Crawler", International Journal of Computer and Electronics Research (IJCER), ISSN: 2278-5795, vol. 3, issue 3, pp. 147-169, June 2014.

[12] Anushree Gambhir and Arushi Goyal, "Weighted Page Rank Algorithm Based on Number of Visits of Links of Web Pages in Time Duration", International Journal of Enhanced Research in Science Technology & Engineering, ISSN: 2319-7463, vol. 3, issue 7, pp. 387-391, July 2014.



**Isha Mahajan** is pursuing her M.Tech in Swami Sarvanand institute of engineering & technology (SSIET), Dinanagar, Punjab (India). She has done B.Tech in Information Technology from Sri Sai College of Engineering and Technology. She is having 3 Years of Teaching Experience and 1 Year Industrial Experience. She belongs to Pathankot city of Punjab, India.