International Journal of Scientific Research & Engineering Trends
Volume 4, Issue 2, Mar- Apr 2018, ISSN (Online): 2395-566X

Intelligent Search System Based on WPRVOL and Feedback Mechanism

Rupali Tadolge
Department of Computer Engineering, University Of Mumbai, Palghar,India
tadolge.rupali1996@gmail.com

Dhanashree Parulekar
Department of Computer Engineering, University Of Mumbai, Palghar,India
dhanashreesp14@gmail.com

Umesh Patil
Department of Computer Engineering, University Of Mumbai, Palghar,India
pumesh891@gmail.com

Tina D'abreo
Department of Computer Engineering, University Of Mumbai, Palghar,India
tinad@sjcet.co.in

Abstract – **Now-a-days, huge amount of digital information is stored in trillions of webpages which are interlinked in World Wide Web. Due to rich hyper structure of web, user get easily lost. Search engine performs lots of computations to obtain specific information against user query and returns many web pages.These webpages are then ordered by ranking algorithm which can be based on number of links or count of visits of links. In this paper, a modified ranking mechanism considering both visits of links as well as time spent on web pages by user is proposed. So, this algorithm can be used to find useful result list based upon user browsing behavior.**

Keywords – **Page Rank, Search Engine, Time Feedback ,World Wide Web, Web Pages,**

I. INTRODUCTION

Researchers have focused on extracting large amount of structured information from large corpora like the Web. The process of collecting such a bulk of information is always tedious. Data mining is the process of sorting through large data sets to identify patterns and establish relationships to solve problems through data analysis. To evaluate the probability of future events and to segment the data sophisticated mathematical algorithms are used by Data mining. Data mining tools allow enterprises to predict future trends. Data mining is also known as Knowledge Discovery in Data (KDD). The application of data mining techniques to extract knowledge from Web data is called Web mining. The huge dataset of Web data includes many different kinds of information, including, web documents data, web structure data, and user profiles data. Web mining, a type of data mining used in customer relationship management, integrates information gathered by traditional data mining methods and techniques over the web. Web mining aims to understand customer behavior and to evaluate how effective a particular website is. Web mining is the collection of information collected by data mining methodologies as well as techniques with information gathered over World Wide Web. Web mining has three categorizations such as Web usage mining, Web content mining and Web structure mining. Web usage mining [1] is the application of data mining techniques to discover usage patterns from web data in order to understand and better serve needs of Web based applications. It composed of three phases, specially pre-processing, sequence innovation and variety analysis. Web Content Mining [1] is the process of extracting useful information from the contents of web documents. Content data corresponds to the collection of facts a Web page was designed to convey to the users. Web content mining is related but is different from data mining and text mining. It is related to data mining because many data mining techniques can be applied in web content mining. The goal of web structure mining [1] is to generate structural summary about the website and web page. The first kind of web structure mining is extracting patterns from hyperlinks in the web. A hyperlink is a structural component that connects the web page to a different location. The new type of web structure mining is mining the document structure. It is using the tree-like structure to analyze and describe the HTML (Hyper Text Markup Language) or XML (Extensible Markup Language). For example, Web pages can be classified and clustered according to their topics. Discovering information or knowledge from hyperlinks (or links for short) is called web structure mining. This represents the Web structure. For example, search engines use technology where useful web pages can be found from such hyperlinks.

II LITERATURE REVIEW

"Weighted Page Rank Algorithm Based on Number of Visits of Links of Web Page", [2] describes that search engine performs various retrieval process to extract information required by user from large amount of data present in web pages. To ful fil User's criteria multiple web pages are responded to user's query. Ranking algorithm orders the web pages either link oriented or content oriented and helps the user for navigating the result. They proposed WPR based on VOL (Visits Of Links). It takes into account number of inbound links of web pages. WPR considers both in links and out links of pages and assigns the rank based on popularity of pages. Due to consideration of user browsing behavior this algorithm reduces search space and provides more valuable pages on the top of the result list. Their proposed

algorithm determines Page Rank value or the importance of web pages based on the visits of the incoming links on the page and popularity of in links web page. The web pages are ordered according to their relevancy and provide the user with quality search results.

"Page Ranking Based on Number of Visits of Web pages" [3] proposed a new algorithm in which user's browsing behavior is considered. User usage trends are not available in most ranking algorithms as they are either link or content oriented. In this paper, a page ranking mechanism called Page Ranking based on Visits of Links (VOL) which considers number of inbounds links of web pages is being devised for search engines. As VOL method uses link structure of pages and browsing information, the web pages which contains relevant information according to user's requirement is returned at the top of the result list. The ordering of pages using VOL is target-oriented.

"Weighted Page Rank Algorithm" [4] stated that relevant information should provide by website owner to the users. According to user's requirement and interest it is important to find appropriate information. They explained that HITS and Page Rank, algorithm are used in web structuring mining for accomplishing concept of web mining. The algorithms assign rank considering all of the links equal. They proposed an algorithm called Weighted Page Rank algorithm as Weighted Page Rank algorithm (WPR). It examines both in links and out links of pages and gives the rank based on popularity. They suggested to use more than one reference page list for the calculation of rank scores.

In paper [5], authors developed Page Rank algorithm based on hyperlink structure. Google search engine uses Page Rank algorithm. Most frequently algorithm used for ranking billions of web pages is Page Rank algorithm. This algorithm combines pre computed Page Rank scores with text matching scores to obtain an overall ranking of web page. Page Rank algorithm is based on the link structure of the web pages and the concepts that if page surrounds important links towards it then the links of this page near the other pages are also considered as important pages. The Page Rank emulate 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.

"Web information Retrieval using Query Independent Page Rank algorithm" [6] proposed the analysis of trust level of web pages which depends on the facts, deceit, opposition and assumptions of the web pages on the web. To represent the quality and level of trust for user purpose, they assigned a rank to every document present on the web. They used page ranking concepts and implemented the algorithm analyzing the behavior of algorithm for different values of moister factor. They demonstrated the work which explains that ranking of multiple documents using link structure of web providing the ranking offline. Thus, user is able to set priority of the documents that are on the web without query dependency. For the analysis of the algorithm they used dampening factor which is selecting as 0.5 or greater. Considering only convergence speed factor the performance evaluated has the best moister factor as 0.95. Their proposed algorithm is query independent without considering query during the ranking.

"Improvement of Page Ranking Algorithm Based on Timestamp and Link", [7] explained that normal ranking algorithms or static algorithms when ranking the pages according to dynamic web works on the concept of favoring old pages and making them appear on the top of ranking results. They proposed temporal link analysis algorithm for solving the issue by using last modification time returned by HTTP response as timestamp of nodes and links. Overall weight of the pages is computed through weight of in-link and out links. Their proposed scheme as WTPR is able to decline old pages and increase the ranking of the new pages and also the old pages having higher quality are given high rank. WTPR is a dynamic self-adapted page sorting algorithm. The WTPR works upon two methods decrease in the ranking value of old pages and increase in ranking value of new pages increases reflecting the user's demand and interests effectively. Also the higher quality pages get higher ranking thereby consequently improving the ranking results.

III.PROPOSED IDEA

First user will enter the keyword. When keyword is provided by user the system will fetch web pages according to the query. In WPRVOL module, at first the initial WPRVOL is 1, if it is not the initial one then it calculates the page rank on the basis of latest WPRVOL calculations and calculated WPRVOL is then stored in the database. After that it will extract the feedback from the database and normalizes it. After normalization, feedback will store in database.

1 Components of Proposed System Architecture

1.1 User Interface-This layer enables user to perform search operation by providing keyword. User can also view the result.

1.2 World Wide Web (WWW)-The communication between the user and the system is done through WWW (World Wide Web). WWW acts as an information space and consists of web resources like documents. These web resources are found by the URLs.

1.3 WPRVOL with Feedback-It has two major modules namely WPRVOL module and Feedback module.

WPRVOL module:

Initial WPRVOL check-In this step, according to query system fetches web pages for the keywords entered by the user. At this level, the system will retrieve the data related

to that web pages i.e. Initial WPRVOL (for the first time it is 1), in links, out links, etc. from the database.

Latest WPRVOL Calculation-Once all the data related to web pages is retrieved from the database, WPRVOL algorithm is applied to each web page to obtain revised page rank.

Store WPRVOL- Revised WPRVOL is then stored in the database by updating the previous rank value of each page.

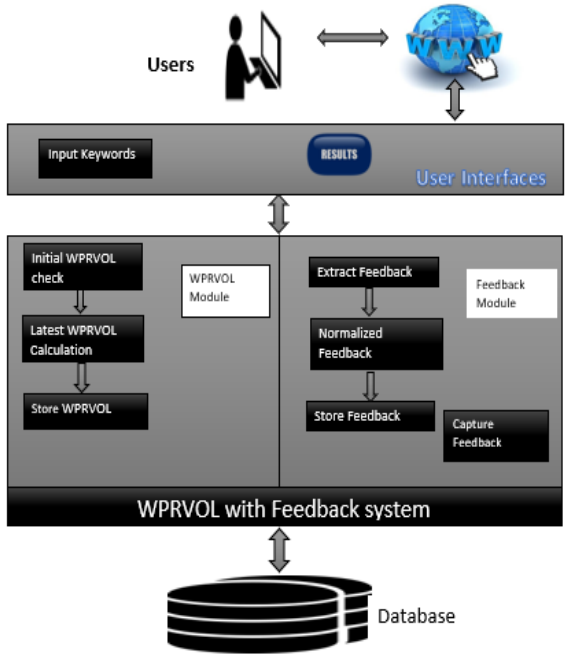

Fig 1. System Architecture

Feedback module:

Extract Feedback-Feedback module extract the previous feedback which is time spent by the user on the web pages.

Normalize the feedback: The database contains total feedback for each page with a count of total numbers of users. The feedback is normalized by taking the average. The formula for normalizing the feedback-

Average feedback=total feedback/total number of user

Store the feedback: The normalized feedback is stored in the database. Capture feedback-It will capture the time spent by the user on web pages.

1.4 Database: Database stores the data required by the application as well as the newly generated data by the WPRVOL with feedback module.

2. Time Feedback Calculation

The feedback module records user feedback for a page by capturing the time spent by that user on that web page. This module captures the time twice; first, when the user visits the page. It is called as Arrival Time (AT). And Second, when the user leaves the page. It is called as Leaving time (LT). These timings will be stored into the database against that page along with the IP address of the user.

S. No.	Time Range(Seconds)	Feedback scale
1	0 to 10	1
2	11 to 30	2
3	31 to 60	3
4	61 to 120	4
5	Greater than 120	5

Table 1: Feedback Scale.

After that the feedback scale module starts operating on this data generated by feedback module. The feedback scale module calculates the difference between leaving time and arrival time which will be the time spent by the user on a web page. Then the relevant feedback scale is obtained by mapping this time duration to feedback scales as shown in the Table 1. Finally, the feedback scale and feedback count are inserted into the database.

Suppose, the user clicks on the link to web page 'Introduction to Java'. When that page is loaded into system's memory, feedback module records the current system time which will be referred as Arrival time (AT). In this hypothetical case, say AT = 06:31:15. Arrival time and leaving time are represented using hh:mm:ss format. When the user has done with the page and closes the browser/tab, feedback module detects the closure event and records the current system time i.e. Leaving time (LT). Say LT = 06:32:01.

After that, feedback module saves these two timings to the database along with the page information and user's IP address. Feedback scale module calculates the difference between AT and LT which is, in this case is 46 seconds. Then, feedback module maps this difference with available scales as shown in the table above to find out relevant scale. In this case, the feedback scale would be 3. This scale is then stored to the database against that page. The proposed system will have a Time Feedback module. This module will record visit duration of the users. Visit duration will then be stored in the database. Then that duration is normalized as follows. Based on the research, we will have five time ranges directly mapped to the Feedback scale. If user spends say 9 Seconds on a web page and leaves the page then system will record the feedback as 1 for that page. The more time user stays on the page the greater feedback that page will receive and vice versa.

3. Algorithm

The various steps of the proposed algorithm are given below**:**

Step1: Find a Website**:** Find a website which has large number of hyperlinks because this relies on the web structures.

Step2: Build a Web Map: Then build the web map/graph of the selected website as shown in the Fig 2.

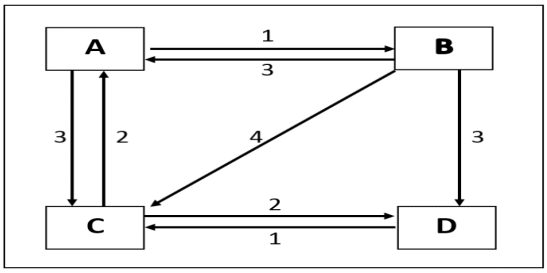

Fig. 2 Hyperlinked Web graph.

Where

- A, B, C, D are the web pages
- Directed edges represents links
- Numbers on links represents their respective weight or visit count.

Step3: Calculate Win (v, u): Then calculate the Win (v,u) for each node in web map by applying the equation (1) as follows:

$$W^{in}_{(v,u)} = \frac{Iu}{\sum_{p \in R(v)} Ip} \quad (1)$$

Where

- Win (v, u) is the weight of link (v, u) calculated based on the number of in links of page u and the number of in links of all reference pages of page v.
- Iu and Ip are the number of incoming links of page n and page p respectively.
- R(m) denotes the reference page list of page m

Step4: Obtain Feedback: Obtain Feedback from User using the method described above.

Step5: Apply Modified WPRVOL formula: Calculate the Page Rank value of the web pages by using the modified equation (2) as follows:

$$WPRvol(u) = (1-d) + \sum v \in B(u) \frac{LuWPRvol(u)Win(v,u)}{TL(v)} + Fd(u) \quad (2)$$

Where

- u represents a web page,
- B (u) that point to u is the set of pages.
- d, is the dampening factor.
- WPR vol (u) and WPR vol (v) are rank scores of pages u and v respectively,
- Lu 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.
- Fd(u) denotes the average feedback calculated by time spent by user for Page u.

Step6: Save Page Ranks: Store the generated Page Ranks to the Database.

Step7: Generate Results: Generate results & display it to user.

IV.RELATED WORK

Page ranking algorithms are the heart of search engine and give results that suites best in user expectation. Page Rank is a way of measuring the importance of website pages. The Need of best quality results are the main reason in the innovation of different page ranking algorithms, HITS, Page Rank, Weighted Page Rank, Distance Rank, Dirichlet Rank Algorithm, Page content ranking are different examples of page ranking used in the different scenario. Page Rank is a way of calculating the importance of website pages. It works by calculating the number and quality of links to a page to determine that how much important the website is.

1 Concept of in links and out links

To Web page W, an in link is a link of another Web page which contains a link pointing to W.

To Web page W, an out link is a link (URL) appearing in W which points to another Web page.

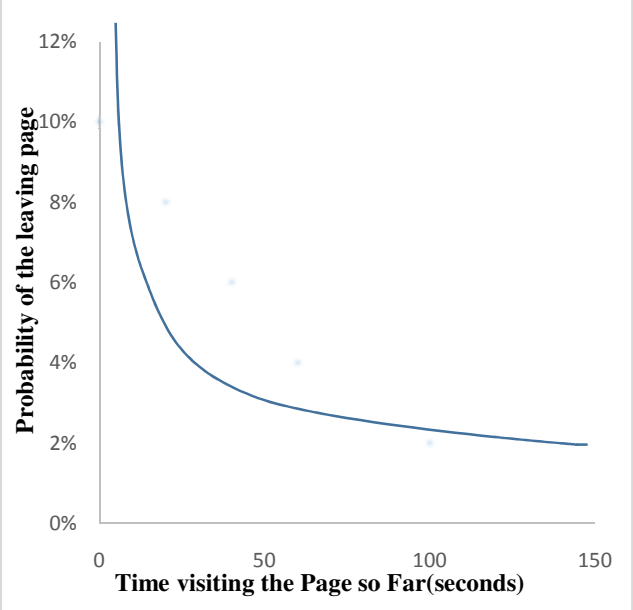

Fig. 3 Graph for Feedback Scale.

2 Time Feedback Concept

Users often leave web pages in 10–20 seconds, people's attention for much longer can hold web pages with a clear value proposition. Hence, we can conclude that the amount of time spent on the web page is directly

proportional to the relevancy or importance of the web page. The following chart shows the likelihood of leaving web pages:

It is obvious from the Fig 3. That the first 10 seconds of the web page visit are critical for user's decision to stay or leave. The possibility of leaving is incredibly sharp all through the above mention early few seconds for the reason that users are exceedingly doubtful, having suffered lots badly designed web content inside the past. People know that most web pages are pointless and they behave accordingly to avoid wasting more time than absolutely necessary on bad pages.

If the web page survives this first extremely harsh 10-second judgment, users will look around a bit. However, they are still likely to leave during the subsequent 20 seconds of their visit. Only after people have stayed on a page for about 30 seconds does the curve become relatively flat. During the first 30 seconds people continue to leave every second, but at a much slower rate. So generally, there are two cases here:

- Bad pages, which get the chop in a few seconds.
- Good pages, which might be allocated a few minutes.

Note: "good" vs. "bad" is a decision that every individual user makes within those first few seconds of arriving.

V.CONCLUSION

The proposed methodology calculates Page Rank value or importance of web pages based on the visits of incoming links on a page as well as the time spent by the users on that web page. It not only considered as link structure but also it includes users focus on a particular web page. This modified algorithm will provide more relevant results than original WPRVOL as it gives the information about which web pages does the user choose from the complete list of results by capturing time spent by user on that particular web pages. The ordering of web pages in this way increases the relevancy of pages and thereof provides the user with quality search results.

V.REFERENCES

[1]. Wikipedia," Web mining-Wikipedia, the free encyclopedia," December 2014. [Online] Available:http://en.wikipedia.org/wiki/Web_mining
[2]. Neelam Tyagi and Simple Sharma "Weighted PageRank Algorithm Based on Number of Visits of Links of Web Page", IJSCE, 2012
[3]. 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
[4]. Wenpu Xing and Ali Ghorbani "Weighted PageRank Algorithm", CNSR'04, IEEE, 2004
[5]. 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
[6]. Harmunish Taneja and Richa Gupta "Web Information Retrieval using Query Independent Page Rank Algorithm", International Conference on Advances in Computer Engineering, IEEE, 2010
[7]. Shiguang Ju, Zheng Wang and Xia Lv "Improvement of Page Ranking Algorithm Based on Timestamp and Link", IEEE, 2008.
[8]. https://www.nngroup.com/articles/how-long-do-users-stay-on-web-pages/