ISSN:0975-9646

Secure and Effective Web Searching Experience Using Cumulative Weighted Page Rank Algorithm

Megha Bhawsar[1], Mr. Tejalal Choudhary[2]

[1]*Department of Computer Science,*
Sushila Devi Bansal College of Technology, Indore (M.P), India
[2]*Assistant Professor, Department of Computer Science,*
Sushila Devi Bansal College of Technology, Indore (M.P), India

***Abstract*—In today's world the internet users are growing very rapidly due to high pace in information exchanges and rich content sources. Users requires high relevancy of information to cater the need of web mining goals. Thus finding the appropriate and accurate content with behavioral aspect covered with keyword search is of high interest these days. The web mining is the field which works in categorizing the information according to user's keyword and interest on the basis of some relevancy algorithms such as page ranking and HITS. But still there is some performance inequality with these algorithms thus we require to revise the efforts. Thus Weighted Page Rank (WPR) is suggested as an innovative solution standard in information retrieval industry. It takes into account both in-links and out-links to get the rank with accurate scores and dynamically updated weights of the links and nodes. Even though the directions of WPR were good but still there are some problems associated with some of its new solution like agent based approach. Thus this work suggest Cumulative Weighted Page Rank (CWPR) algorithm to improve high relevancy and accuracy in retrieved results. At the qualitative evaluation the approach with simulation analysis is showing positive results and leading towards developing the solution**.

***Keywords*—Information Retrieval, Web Mining, Page Rank, Weighted Page Rank (WPR), Agent, Cumulative Weighted Page Rank (CWPR), Accuracy, Relevancy;**

I. INTRODUCTION

Internet or World Wide Web is the most popular means of information exchanges and retrieval for different types of content like text, video, images etc. Massive data is continuously searched and traversed through several types of user queries aims towards getting thee related results. The search engines take large time to measure the relativity of user query and the displayed information. The searched results are measured using ranks of different pages which were dynamically updated using lots of parameters. Calculating the relevancy is a typical task because it covers complete analysis of pages and their behavior and ranks them accordingly. Web content mining aims towards extracting the useful content from massive sources of data like Internet. This retrieval process is completely dynamic in nature and continuously gets updated with changes deriving the search results. This result makes the users navigation easy and effectives with faster responses. Web mining deals with all the issues of information retrieval using three of its defined types i.e. web content mining, web structure mining and web usage mining. Mainly the search of web based data is handled by ranking engines and their defined algorithms. These algorithms calculate some relevance between the user queries and generate output in the form of ordered list known as page rank with some factors used to defined or filter these ranks.

For analyzing the page rank the ranking engine uses incoming and outgoing links along with the content quality and users feedbacks. Google, Yahoo, Bing are some of the search engines which uses more than 1000 parameters updated every two month to get quality and most related results to the users. The most difficult thing to analyze is hyper-links because user can navigate from one web to another using these hyper-links. Web Mining techniques such as clustering, classification, association rule discovery and categorization to filter, classify as well as group their search results. Many page ranking algorithms have been proposed in the literature such as HITS, Clever, PageRank, Weighted PageRank, and Page Content Rank [1]. Some algorithms rely only on the link structure of the documents i.e. their popularity scores (web structure mining), some look for the content of the documents with respect to the user query (web content mining), while others use a combination of both i.e. they use links as well as the content of the document to assign a rank value to the concerned document [2].The algorithm used to perform these tasks is page ranking algorithms. They are further divided into two major types: Page Rank and Weighted Page Rank.

Among most of the current search engines Google is very popular and gives satisfactory results with minimum time [3]. It retrieves a list of relevant web pages by analyzing the keywords and tags given by the user in search queries along with several other parameters to get the high accuracy and relevancy. Later on the PageRank algorithms are developed to improve the search rank and get the appropriate results earlier in the retrieved records. It works on phenomenon that if page contains the important back links and incoming links then its outgoing links to other pages also become important, thus it takes back links into account and propagates the ranking through links. When some query is given, Google combines pre-computed PageRank scores with text matching scores to obtain an overall ranking score for each resulted web page in response to the query. Although many factors determine the ranking of Google search results but PageRank continues to provide the basis for all of Google's web search tools.

II. WEIGHTED PAGE RANK (WPR) ALGORITHM

Later on this page ranking algorithms are evolve to further improve the results and reduces the time and resource requirements towards getting the effective outcomes. Thus they give Weighted Page Rank (WPR) which assumes that more popular the web pages then there will be more linkages of other web pages. This algorithm assigns larger rank values to more important pages instead of dividing the rank value of a page evenly among its outgoing linked pages. Each out link page gets a value proportional to its popularity or importance and this popularity is measured by its number of incoming and outgoing links [4]. The popularity is assigned in terms of weight values to the incoming and outgoing links. The process of calculating the WPR starts with selecting the web with rich hyperlinks to design the correct web structure. Once the structure was finalized then the web map is prepared using certain web tools like JSpider [5]. This result was compare with the retrieved record and will match their relevancy called as root set detection. Now once the root set is identified then the in-links and out-links of root set is separated to measures the weights of each links. Finally the values are passed to the formula to get the ranking status of different pages.

Further improvements have made in WPR to improve its performance and get more accuracy with results. Here the agents are defined for measuring the rank values to the pages which are more important and divide the rank to the separate content zones of individual pages also [6]. A software agent is specific program which is goal driven and react accordingly to the defined functions using actuator or effectors. The operation initiates the effectors functionalities are called as action which can be grouped into multiple sequences. Thus the pages having specific content can be of high rank then its other content and the retrieval measures the important and relevance with users query. It was named as agent based weighted page rank (AWPR) algorithm [15]. This algorithm is used for web structure mining as well as web content mining techniques. Web structure mining is used to compute the importance of the page and web content mining is used to check the page is how much related. Importance here means the popularity of the page i.e. how many pages are pointing to or referred by this particular page. It can be evaluated based on the number of in-links and out links of the page. Relevant means similar of the page with the excited query. If a page is mostly matched to the given query, that becomes more relevant.

Later section of this paper will covers the complete details require to suggest the novel approach along with its parametric and qualitative evaluations. In Section-II literature survey of last few years are given to analyze the working culture followed by problem definition, its solution and benefits. In the last conclusion as a summary is given.

III. LITERATURE SURVEY

During the last few decades the web mining techniques are evolved very rapid and innovative. It covers all the aspect of performance improvements and the cost reduction along with time efficient work. With this survey we had taken a step to look over the direction of current work.

V.K Nagappan and Dr. P. Elango (Feb, 2015): Web content mining goals this problem with the help of agent by retrieving explicit information from different web sites for its access and knowledge discovery. Most of the search engines are ranking their search results in response to users queries to make their search navigation easier .This paper also explores agent based weighted page ranking algorithms for web content mining to retrieve more relevant information. The proposed extended Page Rank algorithm is Agent based Weighted Page Rank Algorithm. The Agent assigns larger rank values to more important pages instead of dividing the rank value of a page evenly among its content. AWPR algorithm retrieves the most important content information or web pages in front of end users.

Rekha Jain, Sulochana Nathawat and Dr. G.N. Purohit (April, 2013): Proposes a novel Dynamic PageRank Algorithm to resolve the ambiguity of polysemous words entered during search. It reduces the irrelevancy among the displayed result and searched query. The step wise process includes tokenization to remove stop words with query enhancer and finally the dynamic rank calculation. Once the process is applied then the results are filtered dynamically according to their relevancy. The proposed algorithm resolves the ambiguity of polysemous words and presents the results according to user preferences. Results shows that proposed Dynamic Page Rank algorithm is more efficient than existing Page Rank algorithm.

Claudia Elena Dinuca, (2011): Gave some of the basic understanding of web mining and it categories. Mainly it focuses direction towards exploring the structure using relationship measurement through some existing tools. It captures all the direct connections and integrates the information about the pages linking and gives search outcomes. Detailed view of paper also put a light on block level link mining issues and reviewed with some popular algorithm. Since this is a huge area, and there a lot of work to do, and hope this paper could be a useful starting point for identifying opportunities for further research.

Laxmi Choudhary and Bhawani Shankar Burdak (Jul, 2012): The problem is to develop the simulation or actual program or comparing the output of different approaches. They worked on this phenomenon and designs a tool which gives step wise execution and analysis of approaches. It calculates the distance rank, page rank and Eigen values though simulation interface and let them compare with different approaches. The simulation program is developed in JAVA for two approaches: PageRank and Weighted PageRank. Comparison has made here to get the in-depth analysis of both the approaches. Normally the web rankings are measured by forming the directed labeled graphs with all the links and nodes. These structures is known as web graphs and used for the link analysis purposes. Measuring the rank of pages must have these graphs along with other details used to discover the structure of web page.

Rekha Jain and Dr. G. N. Purohit (Jan, 2011): Rank distribution and relevancy measurement is performed for PageRank, Weighted PageRank and HITS algorithm which treats all links equally on the basis of rank score. The input parameters used in Page Rank are Back Links, Weighted PageRank uses Back links and Forward Links as Input Parameter and HITS uses Back links, Forward Link and Content as Input Parameters. Complexity of PageRank algorithm is O(log N) whereas complexity of Weighted PageRank and HITS algorithms are <O(log N).

Yajuan Duan, Long Jiang et.al (Aug, 2010): Focused their intentions towards developing the new approach likewise given with a new approach for ranking measurement of well-known tweet database. It identifies the content relevancy of tweets and their URL inclusion. The paper also demonstrated the tweets with URL, length and account authority. Here the ranking model is RankSVM and toolkit was svmstruct7. The comparison of newly developed approach has given by considering three scenarios i.e. chronological order, account authority and content relevance. The paper also explores query expansion approaches to improve the recall of the search results.

Marc Najork, Hugo Zaragoza and Michael Taylor (2004): Large scale evaluation of well-known HITS algorithm is measured and compared with other algorithm. It applies in combination with the other retrieval algorithm and overcomes the issues of anchor text. The selected parameters for performance evaluation are mean reciprocal ran, normalize cumulative gain and average precision. The author had claimed to apply the examination on two large datasets. The experiments found that the HITS algorithm outperform the PageRank. The effectiveness is identified in web page degree and the selected features links. Some more extensive study will prove the performance on the basis of different query sets.

Web mining aims to partition the categorization logic of user from the traversed pages by analyzing the users search queries and behaviors along with the content of pages to rank or order the URL. Mainly it is handled by web structure mining phenomenon. The most famous algorithms are HITS and PageRank. They work on distribution of the rank scores.

Wenpu Xing and Ali Ghorbani (2004): Even though the algorithms are working well but some performance parameters was not showing the effective results. Later on Weighted PageRank algorithm (WPR) comes and works as extension of existing algorithms. It uncovers the use of both incoming and the outgoing links and give them rank according to their popularity of the traversed pages. The paper also presented with simulation results which shows the effectiveness of the developed approach.

IV. PROBLEM DEFINITION

After having a deep look inside the working and outlined features we have found some of the problems associated with existing web mining algorithms like HITS, Page Rank, WPR and AWPR. Some of them is purely based on links only and depends on content quality to generate the scores. Once the pages are configured and integrated then HITS ignores the page structure which may mislead the ranking. While PageRank is considered then it is always suffer from problem of page sink. A phenomenon is found that not all users follow the existing links. Even though we have found numerous directions we have restricted to work on following points to cover the work in given time and cost boundaries.

1) Existing algorithms depends mainly on incoming and outgoing links which might not give the correct result because here the relevance calculation is affected by these links and their popularity [14]. Thus the search results are not real and some crawler may get benefited from this weakness.
2) They assign equivalent weights to all outgoing links which was not necessary because these links may have some unrelated information posted by the similar content links as in [15].
3) The Relevance factors of AWPR [15] only give relatedness with the query and it does not consider the deriving factors like response time, security, trusted server etc. type of feedbacks parameter after which actually the calculated page rank gets reduced. Thus the relevance must be calculated with reduced noise in it.

V. PROPOSED APPROACH

This work proposes a novel Cumulative Weighted Page Rank (CWPR) algorithm using some additionally incorporated factors affecting the search results. Apart from existing factors of AWPR and PR approach it covers the popularity of incoming and out links instead of just distributing the weights equally among all the contents and links of pages. It also integrates the factors related with feedback and users experience towards getting the search results like response time, security and trustworthiness of servers. The proposed algorithm allots higher values to the more popular and socially trusted pages with lightweight nature. It also focuses on the pages which is rich in hyperlinked contents with their web structure and URL analysis. The tool builds the content map of each page using open source spider software like JSpider or ASSpider so as to get the deep analysis of content relevance with the searched query. Somewhere the underdeveloped concept uses complete link analysis, security grievance calculation, response time measurement and popularity assessments with user search history relevance to get better results of each query. Also the work will reduces the impact of noise by removing the irrelevant search items on the basis of six classes like highly relevant (HR), weekly relevant (WR), normally relevant (NR), lightly relevant (LR: reduced response time and normalized no. of links), securely relevant (SR) and irrelevant pages (IR). The concern behind this categorization is to filter the searched results and integrates the feedback experienced by users before final outcome rather than just counting the hits of pages. It also controls the weight distribution according to the above defined classes. The work named as cumulative because we would integrate functionality of these algorithms HITS,

Standard Page Rank and Weighted Page Rank algorithms to improve relevancy & quality of search.

Calculations

(i) Cumulative weighted page rank CWin (v, u): Calculate the CWin (v, u) for each node present in web graph by applying the equation given below.

$$CWin(m, n) = In\ Ip\ p \epsilon R(m)$$

Where

- $W^{in}(v, u)$ is the weight of link (v, u) calculated based on the number of incoming links of page *u* and the number of incoming links of all reference pages of page *v*.
- In 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.

$$CWPR_{vol}(u) = (1 - d) + d \sum_{v\ to\ B(u)} [L_u * WPR_{vol}(v) * W^{in}(v, u) * W^{rt}(v, u) * W^{s}(v, u) / TL(v)]$$

Where

- u represents a web page,
- B(u) is the set of pages that point to u,
- d, is the dampening factor.
- $CWPR_{vol}(u)$ and $CWPR_{vol}(v)$ are rank scores of page u and v cumulatively,
- 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.
- W^{rt} denotes the response time between the visited links
- W^{s} denote the security grievances of visited links by user's feedbacks.

(ii) Relevance: the relevancy of a page to a given query depends on its category and its position in the page-list. The larger the relevancy value is, the better is the result. The relevancy, K, of a page-list is a function of its category and position:

$$K = \sum(n - 1)X\ Wi \quad \text{(for all i belongs to R (p))}$$

Where, i denotes the i^{th} page in the result page-list R (p), n represents the first n pages chosen from the list R (p), and Wi is the weight of page i.

$$Wi = \begin{cases} v1, \text{ if the } i^{th} \text{ page is HR} \\ v2, \text{ if the } i^{th} \text{ page is WR} \\ v3, \text{ if the } i^{th} \text{ page is NR} \\ v4, \text{ if the } i^{th} \text{ page is LR} \\ v5, \text{ if the } i^{th} \text{ page is SR} \\ v6, \text{ if the } i^{th} \text{ page is IR} \end{cases}$$

Where: v1 > v2 > v3 > v4 > v5 > v6

The value of Wi for an experiment could be decided through experimental studies.

VI. CONCLUSION

Web mining deals with getting the appropriate content in near optimal time and efforts by considering the users behavior and searching patterns. But organization and extraction of content from the resources also requires web structure to be effectively mined. PageRank and HITs are the most common algorithms used for measuring the popularity of web pages and will work in getting relevancy from searched keyword. This paper deals with detailed study of some of the existing page ranking algorithms and puts a light on the remaining issues and directions for researcher's Along with the problems the paper also take a step to develop the solution for given solution. Qualitative proof of concept along with predicted calculations is presented with the paper.

REFERENCES

[1] The PageRank Citation Ranking: Bringing Order to the Web, 1998

[2] Jon M. Kleinber, "Authoritative Sources in a Hyperlinked Environment", in ACM-SIAM Symposium on Discrete Algorithms, 1998

[3] Sankar K. Pal, Varun Talwar and Pabitra Mitra, "Web Mining in Soft Computing Framework: Relevance, State of the Art and Future Directions", in IEEE Transactions on Neural Networks, Vol. 13, No. 5, Sep 2002

[4] C. Pahl, "Data mining for the analysis of content interaction in web-based learning and training systems", Book Chapter Published at Dublin City University, Ireland.

[5] Ziyang Wang, "Improved Link - Based Algorithms for Ranking Web Pages", in ACM, NSF grant #IIS-0097537, 2003.

[6] Nadav Eiron, Kevin S. Mc Curley and John A. Tomlin, "Ranking the Web Frontier", in ACM, Doi: 158113844X/04/0005., 2004

[7] Wenpu Xing and Ali Ghorbani, "Weighted PageRank Algorithm", in Second Annual Conference on Communication Networks and Services Research (CNSR'04), IEEE, 2004

[8] Marc Najork, Hugo Zaragoza and Michael Taylor, "HITS on the Web: How does it Compare?", in SIGIR ACM Conference, Doi: 78-1-59593-597-7/07/0007, 2007

[9] Yajuan Duan, Long Jiang, Tao Qin, Ming Zhou, and Heung-Yeung Shum, "An Empirical Study on Learning to Rank of Tweets", in Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 295–303,Beijing, August 2010

[10] Rekha Jain and Dr. G. N. Purohit, "Page Ranking Algorithms for Web Mining", in International Journal of Computer Applications (0975 – 8887, Volume 13– No.5, January 2011

[11] Claudia Elena Dinuca, "Web Structure Mining", in Annals of the University of Petroşani, Economics, 11(4), 2011

[12] Laxmi Choudhary and Bhawani Shankar Burdak, "Role of Ranking Algorithms for Information Retrieval", in International Journal of Artificial Intelligence & Applications (IJAIA), Vol.3, No.4, July 2012

[13] Rekha Jain, Sulochana Nathawat and Dr. G.N. Purohit, "Enhanced Retrieval of Web Pages using Improved Page Rank Algorithm", in International Journal on Natural Language Computing (IJNLC) Vol. 2, No.2, April 2013

[14] T. Nithya, "Link Analysis Algorithm for Web Structure Mining", in International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 8, August 2013

[15] V.K Nagappan and Dr. P. Elango, "Agent Based Weighted Page Ranking Algorithm for Web Content Information Retrieval", in IEEE International Conference on Computing and Communications Technologies (ICCCT'15), doi: 78-1-4799-7623-2/15, 2015