EFFECTIVE WEB SEARCH USING PROGRESSIVE WEIGHTED PAGE RANK ALGORITHM WITH HIGH RELEVANCY AND ACCURACY

Rakhi Kumawat[#1], Dr.Mayank Patel[#2]

M.Tech Scholar[1] , Associate Professor [2]

Computer Science & Engineering Department , Rajasthan Technical University(RTU)
Geetanjali Institute of Technical Studies,Udaipur,Rajasthan ,India

[1]rakhi.kumawat17@gmail.com

[2]mayank999_udaipur@.yahoo.com

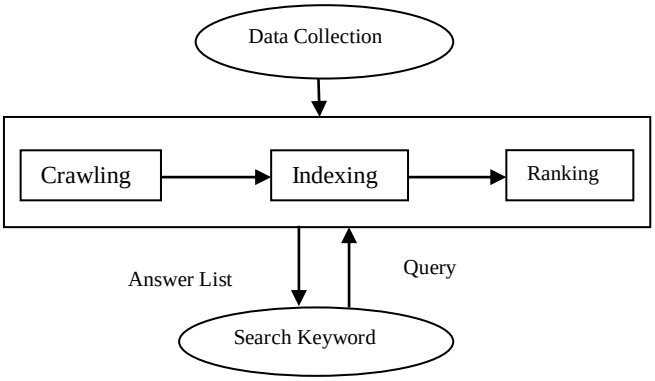
***Abstract*— In this world of Internet, information is spread and has enormous sources which generate information. One of the major web mining goals is high level of relevancy of the information as results. There is a tremendous need of methods to find the appropriate and accurate content with behavioral aspect covered with keyword search. The field of web mining deals with categorizing the information according to user's interest and search with the help of some relevancy calculation techniques and algorithms, like HITS, page ranking etc. still there is some gaps and inequality with these approaches thus we require to review. Weighted Page Rank (WPR) can be an innovative solution standard in information retrieval industry. WPR and take care of both in-links and the out-links to get the rank with accuracy in scores and dynamically updated weights of the links and nodes. Although the visions of WPR were good but there is a need of review of other methods and comparing the problems associated with some of its new solution like agent based approach. Thus this work suggest Progressive Weighted Page Rank (PWPR) algorithm to improve high relevancy and accuracy in retrieved results. We have designed a prototype website for the testing the concept, at this early stage, idea is producing satisfactory results with better performance.**

***Keywords*—** Web Mining, Page Rank, Weighted Page Rank (WPR), Agent, Progressive Weighted Page Rank (PWPR), Accuracy, Relevancy.

I. INTRODUCTION

Internet is wildly used now days, search engines are the way to interact or explore the world of Internet. The results that search engines provide need to be relevant, accurate and also in appropriate form of representation. Normally user provide query in the form of keywords and submit to the search engine. It is expected from the service provider to provide appropriate, relevant and quality information to the user. Web is the most well known means of information exchanges and retrieval for various types of content like text, video, images and so forth.

The search engines set aside huge effort to measure the relativity of user query and the displayed information. The searched results are measured using ranks of various pages which were dynamically updated using lots of parameters. Calculating the relevancy is a run of the mill task because it covers total analysis of pages and their behaviour and ranks them accordingly. This retrieval process is totally powerful in nature and continuously gets updated with changes deriving the search results [16].



Fg:1 Information Retrieval(IR) System

For analyzing the page rank the ranking engine uses incoming and outgoing links alongside the content quality and users feedbacks. Google, Rediff, Bing are some of the search engines which uses in excess of 1000 parameters updated each multi month to get quality and most related results to the users.. Web Mining techniques such as clustering, classification, association rule discovery and categorization to filter, classify as well as gathering 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 depend just on the link structure of the documents for example their popularity scores (web structure mining), some search for the content of the documents with respect to the user query (web content mining), while others use a combination of both for example they use links as well as the content of the report to assign a rank value to the concerned archive [2].The algorithm used to perform these tasks is page ranking algorithms. They are additionally divided into two major types: Page Rank and Weighted Page Rank.

II. BACKGROUND

Web structure mining is used to calculate the significance of the page and web content mining is used to check the page is how much related. It can be evaluated based

on the number of in-links and out links of the page. Relevancy of the web page means how the content of the webpage is related to the query or the field. If a page is mostly matched to the given query, that becomes more relevant.

Page ranking algorithms are envolve to further improve the results and reduces the time and resource requirements towards getting the effective outcomes. Thus they gives Weighted Page Rank (WPR) which assumes that more popular the web pages then there will be more linkages of other web pages. This algorithm gives larger rank values to more important pages and not dividing the rank value of a page equally 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 [16]. The weights to the incoming and outgoing links are basically measure of popularity. The process of calculating the WPR starts with selecting the web with rich hyper links to design the correct web structure. Once the structure was finalized then the web map is prepared using certain web tools like JSPider. 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 [16].

Page Rank algorithm is the most commonly used algorithm for ranking the various pages. Working of the Page Rank algorithm depends upon link structure of the web pages. The basis of Page Rank algorithm is that if a page is having main links towards it then the links of this page towards the other page are also to be considered as important pages. If the addition of the all the ranks of the back links is large then the page then it is provided a large rank [7][8]. A simplified version of PageRank is given by:

$$PR(u) = \sum_{v \in B_u} \frac{PR(v)}{L(v)},$$

Where the PageRank value for a web page u is dependent on the PageRank values for each web page v out of the set Bu (this set contains all pages linking to web page u), divided by the number L(v) of links from page v.

- **Weighted Page Rank Algorithm (WPR)**

Weighted Page Rank [1] Algorithm is proposed by Wenpu Xing and Ali Ghorbani. Weighted page rank algorithm (WPR) is the modification of the original page rank algorithm. WPR chooses the rank score dependent on the popularity of the pages by taking into thought the significance of both the in-links and out-links of the pages. This algorithm gives high estimation of rank to the more well known pages and does not similarly isolate the rank of a page among its out-link pages. Each out-link page is given a rank esteem dependent on its popularity. Popularity of a page is chosen by observing its number of in links and out links.

- **Weighted Links Rank Algorithm**

A modification of the standard page rank algorithm is given by Ricardo Baeza-Yates and Emilio Davis [13] named as weighted links rank (WLRank). This algorithm provides weight value to the link based on three parameters i.e. length of the anchor text, tag in which the link is contained and relative position in the page.

- **EigenRumor Algorithm**

Page rank and HITS are very promising in providing the rank value to the blogs but some limitations arise, if these two algorithms are applied directly to the blogs The rank scores of blog entries as decided by the page rank algorithm is often very low so it cannot allow blog entries to be provided by rank score according to their importance. To resolve these limitations, a EigenRumor algorithm [14] is proposed for ranking the blogs.

- **Distance Rank Algorithm**

A keen ranking algorithm named as distance rank is proposed by Ali Mohammad Zareh Bidoki and Nasser Yazdani [15]. It depends on reinforcement learning algorithm. In this algorithm, the distance between pages is considered as a discipline factor. In this algorithm the ranking is done based on the shortest logarithmic distance between two pages and ranked by them. The Advantage of this algorithm is that it can discover pages with high caliber and all the more rapidly with the utilization of distance based arrangement.

- **Time Rank Algorithm**

An algorithm named as TimeRank, for improving the rank score by utilizing the visit time of the site page is proposed by H Jiang et al.[16] Authors have estimated the visit time of the page in the wake of applying unique and improved techniques for website page rank algorithm to think about the level of significance to the clients. This algorithm uses the time factor to expand the precision of the page ranking.

- **Relation Based Algorithm**

Fabrizio Lamberti, Andrea Sanna and Claudio Demartini [18] suggested a relation based algorithm for the ranking the web page for semantic web search engine. Various search engines are presented for better information extraction by using relations of the semantic web. This algorithm proposes a relation based page rank algorithm for semantic web search engine that depends on information extracted from the queries of the users and annotated resources. Results are very encouraging on the parameter of time complexity and accuracy.

III. RELATED STUDY

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. [7] Even though the algorithms are working well but some performance parameters was not showing the effective results. 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.

In the paper [8] 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 Progressive gain and average precision. The author had claimed to applied 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.

Some of the researchers had focused their intentions towards developing the new approach likewise given with [9]. This paper gives 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.

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. In the paper [10] the 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, HITS uses Back links, Forward Link and Content as Input Parameters. Complexity of PageRank algorithm is O(log N) where as complexity of Weighted PageRank and HITS algorithms are <O(log N).

Some of the basic understanding of web mining and it categories was given with paper [11]. Mainly the paper focused its directions 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. 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. But the problem is to develop the simulation or actual program or comparing the output of different approaches. Thus the paper [12] had 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 ha made here to get the in-depth analysis of both the approaches.

The paper [13] 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.

IV. PROBLEM IDENTIFIED

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, PageRank, 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 the its always suffer from the problem of page sink. 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 [16].

- 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.
- 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 [15].

V. PROPOSED SOLUTION

Our work giving a novel Progressive Weighted Page Rank (PWPR) 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.

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 factors like mostly relevant (MR), Poorly relevant (WR), fairly relevant (NR), lightly relevant (LR), 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 factors. The work named as Progressive because it is having combined functionality of HITS, Standard Page Rank and Weighted Page Rank algorithms.

Calculations

(i) Progressive weighted page rank PWin (v, u): Then calculate the PWin (v, u) for each node present in web graph by applying the equation given below. PWin(m, n)= In Ip $p\epsilon R$ (m) Where

- *Win*(*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.

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

Where

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

(i) 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:

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

Where i denotes the ith 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 =
- v1, if the ith page is Mostly Relevant(MR)
- v2, if the ith page is Poorly Relevant(PR)
- v3, if the ith page is Fairly Relevant(FR)
- v4, if the ith page is Lightly relevant (LR)
- v5, if the ith page is Securely relevant (SR)
- v6, if the ith page is Irrelevant pages(IR)

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

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

VI. EXPECTED OUTCOMES

Following are the expected outcomes of proposed approach,

- Quality of the pages returned by this algorithm is high as compared to page rank algorithm.
- It is more efficient than page rank because rank value of a page is divided among it's outlink pages according to importance of that page.
- They can be joined with some content based ranking algorithm to improve relevancy of the web pages.
- As page rank is a query autonomous algorithm for example it pre-computes the rank score so it requires less investment.
- This algorithm is increasingly doable as it computes rank score at indexing time not at query time.
- It returns vital pages as Rank is determined based on the notoriety of a page.
- Less vulnerability to localized links: - For figuring rank estimation of a page, it consider the whole web chart, as opposed to a little subset, it is less defenseless to localized connection spam.

VII. 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 webpages 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 prototype for proposed approach. 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,VarunTalwar and PabitraMitra, "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] NadavEiron, Kevin S. McCurley 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] YajuanDuan, 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] LaxmiChoudhary 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, SulochanaNathawat 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

[16] Er.Manika Dutta and Dr.Kishori Lal Bansal ," Improvement in Weighted Page Rank Algorithm using Efficiency and Precision",in International Journal of Computer Science & Engineering Technology (IJCSET) Vol. 8 ,No. 01, Jan 2017.

[17] Rakhi Kumawat and Dr.Mayank Patel "Result Evaluation Of Web Search Using Progressive Weighted Page Rank Algorithm with High Relevancy and Accuracy",in International Journal of Research and Analytical Reviews (IJRAR), Vol. 6, Issue 2, June 2019.