A Brief Survey of Various Page Ranking Algorithms in Web Mining

[1]Riddhi A. Butala
[1]M.E. Student,
Dept. of CE, Gujarat Technological University, Gujarat, India

***Abstract*— the World Wide Web consists of million of the web pages that are interconnected to each other. Day by day the growth of the World Wide Web is increasing very rapidly. With rapid growth of web it becomes very difficult to provide the relevant information in response to user query. The search engines help the user to surf the web. Due to the vast number of web page it is highly impossible to provide the proper, relevant and quality information. Thus web search engines need efficient ranking algorithm, so that the user could retrieve the web page which is most relevant to user query. In this paper, a survey of page ranking algorithms and competition of some important ranking algorithms.**

***Index Term* — WWW, Web Mining, search engines, Page Ranking.**

I. INTRODUCTION

WWW is a huge resource of hyperlink and heterogeneous information including text, image, audio, video, and metadata. It is estimated that WWW has expanded by about 2000% since its evolution and is doubling in size every six to ten month [10] [1]. Day by day increasing the growth of information resources WWW it is difficult to manage the information on the web. That's why it has a increasing of necessary for the user to use efficient information retrieval techniques to find and order the desired information. Therefore we need some efficient search engines. The search engines will play an important role in searching a web page. Search engines collect, analyze, organize, and handled the data on internet [2].

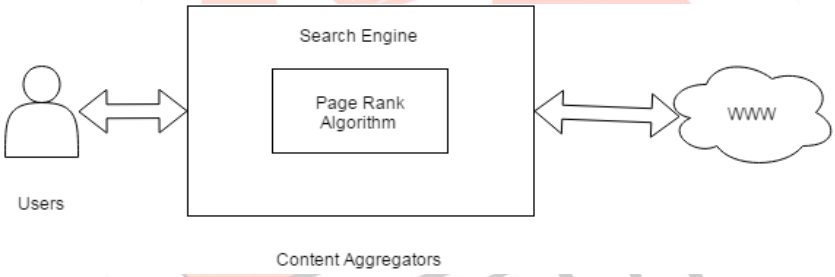

Fig 1: concept of search engine [1][2][3]

But the search engines returns thousands of results which includes a mixture of relevant and irrelevant information [2]. Fig 1 Shows the simple concept of the simply concept of the search engine. Search engine are used to find the information from the World Wide Web. Some popular search engines are Google, msn, bing, yahoo search etc [1]. They are many other things are participated in this searching techniques, downloaded, index, query process, and store the store hundreds of millions web pages [1] [2] [3]. They answer the millions of query every day, hours, minutes and every second. Query processer will act like a content aggregators and the keep a recorded of every information are available of WWW [2][1]. Fig 2 shows the architecture of search engine. The most important part of search engine is crawler. The web crawler is downloading the web data [10]. Index is generally maintained alphabetically considering the keywords [1]. When the query processor component was retrieved the user keyword, it matching the query keyword with the index return the URLs of the pages to the user. But before representing pages to user some ranking mechanism are applied in back end or front end is used by most of search engine, to make the user search engine make easier. Most relevant page is put on the top of the result list and less relevant page is put on the bottom of the result list.

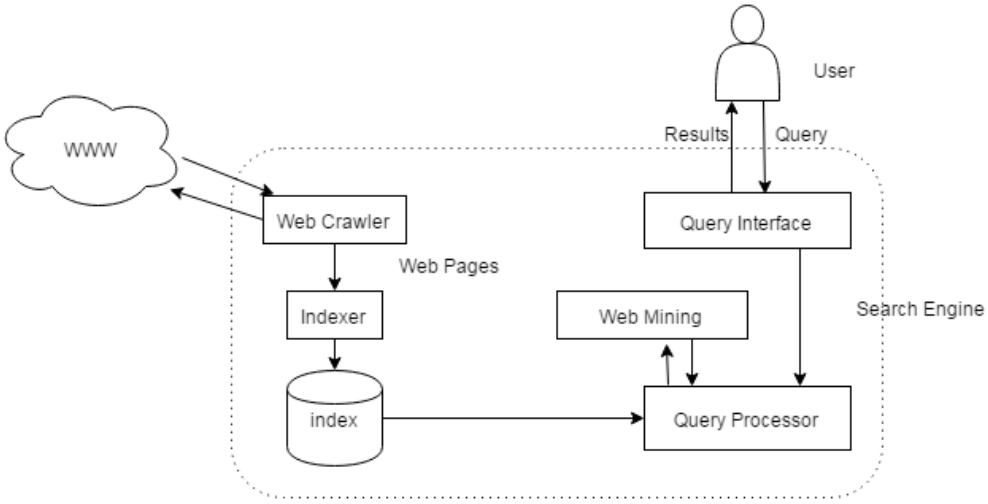

Fig 2: Architecture of Search Engine [1]

In this paper survey of different page ranking algorithm and comparison of this algorithm are carried out. The structure of this paper is follows: section II provides the overview of web mining concept. Section III provides detail overview of some important web page ranking algorithm. Section IV provides competition of this algorithm and finally, Section V provides the conclusion of this paper.

II. WEB MINING

With increase of volume of the World Wide Web (WWW), the area of mining of information from the web is very huge [4]. For this reason various search engines are trying to improve their working to give best result to users, so that's why we need some efficient ranking techniques. Web Mining is defined as the application of data mining techniques going on the World Wide Web to find hidden information [1]. An application of web mining can be seen in the case of search engines like Google, yahoo, bing etc…[1]. The web mining techniques has classified in varies three categories: Web Content Mining (WCM), Web Structure Mining (WSM), and Web Usage Mining (WUM)[4][5][1]. The WCM is finding the useful information from web content, WSM is finding the relationships between web pages by analyzing web structures, and WUM is finding the user profiles and the users' behavior recorded inside the web logfile [5].

Web Content Mining(WCM)

The content based or text based web mining is the concept of data mining for finding the more specific data. For finding the information on web is complex then finding the static database, because of the dynamic nature of web it has huge amount of documents [6]. WCM is the content of the web page and its web page itself or resulted web page will obtain on search engine. The WCM will classify two different views: i. Information retrieval view (IR) and ii. Database view (DB) [1]. In information retrieval view, the bunch of the unstructured and semi- structure text data will have HTML or XML structure inside the documents can be used. IR is an important technique for widely used in web content mining especially used for web search engine [1] [7]. IR model will defend has similarity between query and document, there are three type of IR model 1. Boolean model 2. Vector space model and 3. Language model [5]. In database view (DB), a web site can be transformed to represent a multi-level database and web mining tries to infer the structure of the web site from this database [1].

Web Usage Mining(WCM)

WSM discovers relationships between web pages by analyzing web structures [5]. Web Structure Mining makes use of graph theory to analyze the hyperlink structure, based on the hyperlink topology and It also categorized the website and interlink the website [5] [8]. Web Structure Mining is the process of inferring knowledge from the World Wide Web organization and links between references and referents in the Web. The structure of a typical web graph consists of web pages as nodes and hyperlinks as edges connecting related pages. Web Structure mining is the process of using graph theory to analyze the node and connection structure of a web site [9]. It is used to discover structure information from the web and it can be divided into two kinds based on the kind of structure information used. They are Hyperlinks and Document Structure [9] [1]. It is used to generate structural summary about the web pages in the form of web graph where web pages act as nodes and hyperlinks as edges connecting two related pages [1].

Web Structure Mining(WSM)

WSM discovers relationships between web pages by analyzing web structures [5]. Web Structure Mining makes use of graph theory to analyze the hyperlink structure, based on the hyperlink topology and It also categorized the website and interlink the website [5] [8]. Web Structure Mining is the process of inferring knowledge from the World Wide Web organization and links between references and referents in the Web. The structure of a typical web graph consists of web pages as nodes and hyperlinks as edges connecting related pages. Web Structure mining is the process of using graph theory to analyze the node and connection structure of a web site [9]. It is used to discover structure information from the web and it can be divided into two kinds based on the kind of structure information used. They are Hyperlinks and Document Structure [9] [1]. It is used to generate structural

summary about the web pages in the form of web graph where web pages act as nodes and hyperlinks as edges connecting two related pages [1].

Why Ranking is Required

Web create a new challenges of information retrieval, the amount of information retrieving in web is increased day by day. Rapidly growth of net surfing, and as well as number of new users will increased rapidly. So high quality and efficient information retrieving through the search engine is very expensive and the process of crawling, indexing and searching will be very complex and slow. When increasing the complexity of net surfing user then automatically decreased searching result. For solving that problem of efficient information retrieving, search engine will have efficient ranking algorithm. The most popular ranking algorithms are the page Rank, Weighted Page Rank, HITS, PR Based VOL, WPR Based VOL, SimRank, etc… ranking algorithm will calculate the rank value based on inlink and outlink of find the popularity of web page.

III. PAGERANKING ALGORITHM

Page Ranking algorithms are the soul of search engine and they give the best result of the user expectation. User need of the best quality results are main reason in innovation and improvement of different page ranking algorithms like Page Rank, HITS, Weighted Page Rank, SimRank, Page Rank based VOL, Weighted Page Rank based VOL, Weighted Page Rank based Zero Link Similarity. Now a day's Google search engine is very important because many web users is used.

Page Rank

S. Brin and L. Page [10] was proposed page rank algorithm at Stanford University. Now a day's page rank algorithm was used by very popular search engine GOOGLE. The main concept of page rank, marching the text value of query and find the overall score of web page and it utilize the link to improved the search result. The main goal of page rank is improve the quality of search engine [10]. PageRank is a very good way to prioritize the results of web keyword searches. Page rank is also help for full text searches in main Google system.

The basic formula of page rank is,

$$PR(A) = (1-d) + d\left(\frac{PR(T1)}{C(T1)} + \cdots + \frac{PR(Tn)}{C(Tn)}\right) \quad (1)$$

Where,

$PR(A) = page\ rank\ of\ page\ A, it\ calculated\ using\ simple\ itarative\ algo.$

$T1 \ldots Tn = page\ A\ has\ poin\ to\ it\ page\ T1 \ldots Tn$

$C(A) = number\ of\ out\ going\ link\ of\ page\ A$

$d = dampening\ fector(its\ value\ is\ 0.85(approx))$

HITS

Jon Kleinberg [15] introduced Hyperlink-Induced Topic Search (HITS) algorithm, it also known as hubs and authorities. HITS are a link analyses algorithm that rates Web pages. The hubs are serving as large directories that are not actually authoritative in the information that it held, but we used vast catalog of information that lead directly it's called authoritative page [14]. Fig.3 shows the hubs and authorities. This method assigns two scores for each page: its authority, which estimates the value of the content of the page, and its hub value, which estimates the value of its links to other pages [16]. So the authority is sum if all score of hub pages and the hub score is sum of all linking pages of authority pages.

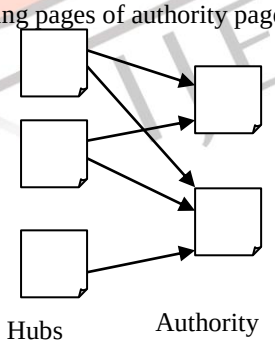

Fig. 3: Hubs and authority [16] [2]

∀p, we update the auth (p) and hub (p) to be:

$$\sum_{i=1}^{n} hub(i) \quad (2)$$

$$\sum_{i=1}^{n} auth(i) \quad (3)$$

Where,

$n = total\ number\ of\ page\ connected\ to\ p$

$i = current\ page\ connected\ to\ p$

Weighted Page Rank

Wenpu Xing and Ali Ghorbani [5] are proposed a Weighted Page Ranking algorithm. Is the improvement over PageRank Algorithm by introducing weighting scheme. In this approach inlink and outlink weights are used to calculate webpage rank value. The proposed a weighted PageRank algorithm which gives more Rank portion to the neighboring pages with more in-links. It is yet does not sufficiently reflect the actual behaviors of surfers, because only the information of topological structure is used. Strength of this algorithm is that it works offline independent to query. And limitation is, its ranking may be distinguished easily.
$W^{in}_{(v.u)}$ And $W^{out}_{(v.u)}$ are the weight of *link(v,u)* is calculated based on number of inlinks and outlinks of page *u* and the number of inlinks and outlinks of all the reference page of page *v*.

$$W^{in}_{(v,u)} = \frac{I_u}{\sum_{p \in R(v)} I_p} \quad (4)$$

$$W^{out}_{(v.u)} = \frac{O_u}{\sum_{p \in R(v)} O_p} \quad (5)$$

Where,

$I_u = number\ of\ inlink\ of\ page\ u$
$I_p = number\ of\ inlink\ of\ page\ p$
$R(v) = raference\ pagw\ list\ of\ page\ v$
$O_u = number\ of\ outlink\ of\ page\ u$
$O_p = number\ of\ outlink\ og\ page\ p$

SimRank

S. Qiao, Tianrui Li, Li and Yan Zhu, Jing Peng, Jiangtao Qiu [7] was proposed by SimRank algorithm. This algorithm is variant of weighted PageRank algorithm, called SimRank that distributes Rank value in proportion to the inter-page similarities. To apply the method, all pair-wise page similarities need to be computed earlier on. For finding the similarity between P_a and P_b, The formula behind this algorithm is given below.

$$sim(p_a, p_b) = \frac{d_a \bullet d_b}{\|d_a\|^2 + \|d_b\|^2 - d_a \bullet d_b} = \frac{\sum_{i=1}^{m} w_{ip_a} \times w_{ip_b}}{\sum_{i-1}^{m} w_{ip_a}^2 + \sum_{i-1}^{m} w_{ip_b}^2 - \sum_{i-1}^{m} w_{ip_a} \times w_{ip_b}} \quad (6)$$

Where,

$sim(p_a, p_b) = similarity\ between\ p_a and\ p_b$
$d_a = dampening\ factor\ of\ inter\ page\ A$
$d_b = damening\ factor\ of\ inter\ page\ B$
$m = number\ of\ term\ in\ query\ Q$

Page Rank Based VOL

Gyanendra Kumar, Neelam Duhan, A. K. Sharma [11] was proposed Page Rank based VOL algorithm. Unlike traditional PageRank algorithm, it does not divide page rank value equally between outgoing links. Instead of this it assign more rank value to the outgoing links which is most visited by users. So in this manner page rank is calculated based on visits of inbound links. The formula behind this algorithm is given below.

$$PR(u) = (1 - d) + d \frac{\sum_{v \in B(u)} L_u(PR(v))}{TL(v)} \quad (7)$$

Where,

$d = dampening\ factor$
$u = web\ page$
$B(u) = set\ of\ pages\ that\ point\ to\ web\ page\ u$
$PR(u) = rank\ score\ of\ page\ u$
$PR(v) = rank\ score\ of\ page\ v$
$L_u = No.of\ visits\ of\ link\ score\ of\ page\ u\ and\ page\ v$
$TL(v) = the\ total\ no.of\ visits\ of\ all\ links\ on\ page\ v$

Weighted Page Rank Based VOL

Neelam Tyagi, Simple Sharma [12] was proposed by Weighted Page Rank Based VOL algorithm. In the traditional weighted page rank algorithm, it assigned the larger rank value is more popular page. All the outgoing links is proportional to popularity. The number of outlinks and inlinkes popularity will store two function W_{out} and W_{in} respectively. But in this proposed algorithm, it is not conceder popularity of outgoing link. In proposed improved weighted page rank algorithm it assign the more rank value to outgoing link which is most visited by user. In this WPR (VOL) algorithm it calculated the user browsing behaviors. It calculates the how many time user will visited by link.
The formula behind this algorithm is

$$WPR_{vol}(u) = (1 - d) + d \sum_{v \in B(u)} \frac{L_u WPR_{vol}(v) W^{in}_{(v,u)}}{TL(v)} \quad (8)$$

Where,

$d = dampening\ factor$
$u = web\ page$
$B(u) = set\ of\ page\ that\ poin\ u$
$WPR_{vol}(u) = rank\ score\ of\ \ page\ u$
$WPR_{(vol)}(v) = rank\ score\ of\ page\ v$
$L_u = no.of\ visits\ of\ links\ which\ are\ pointing\ to\ page\ u\ from\ page\ v$
$TL(v) = total\ no.of\ visit\ of\ all\ links\ present\ on\ page\ v$

Weighted Page Rank Based Zero Link Similarity

The **Sang-yeon Lee, Young-gi Kim, Seok-Jong Lee, Keon Myung Lee [13]** was proposed WPR based Zero Link Similarity algorithm. This algorithm was improved weighted PageRank algorithm that can deal with such zero inter-page similarities, which handles them by allocating a minimum similarity to the links to the pages with the zero-similarity. The proposed algorithm has been implemented using the MapReduce paradigm for big data handling and overcome the problem of simrank algorithm. The formula behind this algorithm is,

$$\rho \frac{\min(s_{ij})}{\sum_{L_{in}(i)} s_{ik}} = \alpha\ (1-\rho) ZR \quad (9)$$

Where,

$\rho = user\ supplied\ parameter\ for\ zero\ similarity$
$ZR = number\ of\ non\ zero\ similar\ link$
$\alpha = controlled\ by\ minimum\ similarity$

IV. COMPARISION OF PAGERANKING ALGORITHMS

Table 1: Comparison of Different Page Rank Algorithm

Variants of PR	Working Approach	Advantages	Disadvantages	Efficiency
Page Rank (Larry p 1996)	Calculate page rank based upon number of backlinks	High quality results, backlink predictor, advertising business, frequently indexing	False page rank or spoof page rank, equal distribution of page rank	Moderate
HITS (C hris H. 2001)	Compute the authority score of n highly relevant page on the top of list	Hub and Authority values are calculated so that the relevant and important pages are obtained	Topic drift and efficiency problems occur. Non-relevant documents can be retrieved	More
Weighted Page Rank (Wenpu Xing 2004)	Assign more page rank value to popular page	More relevant page than traditional page rank algorithm	Does not consider user access pattern	High
SimRank (S. Qiao, 2010)	The relevance of a page to the given query which can improve the accuracy of scoring.	Improved the traditional PageRank algorithm by taking into account the weight of page to a given query.	Applying this method for large volume of pages it's computationally expensive	High
Page Rank based VOL (G. Kumar, 2011)	Assign more rank value to the outgoing links which is most visited by users	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 reduces the search space to a large scale	None	Moderate

Weighted Page Rank based VOL (N. Tyagi 2012)	Assign more rank value to the outgoing links which is most visited by users and received higher popularity from number of inlinks	find more relevant information according to user's query	Very ideal but, it is not easy to apply it to the Web scale	High
WPR based Zero link similarity (Sang-yeon Lee 2014)	Identify the keyword using TFIDF	The zero value for the inter-page similarity of neighboring pages due to the language characteristics. It's also help to handled big data	For finding the content based semantic based keyword is not possible.	V-High

V. CONCLUSION

The quality of keyword base searching is the current challenges of the web mining. The main drawback of web search engine is cannot provide high quality and intelligent service. Search engine was support the keyword, link address and content based search. For finding the relevant information retrieving (IR) in the World Wide Web. Firstly, I study the various page ranking algorithm and then compared this above algorithms. Each and every algorithm has got its own advantage and disadvantage. As per the requirements of a search engine we can utilize the above said algorithms. In future guidance, we will improve the searching result using this survey report.

VI. REFERENCES

[1] N. Duhan, A. K. Sharma, K. Bhatia, "Page Ranking Algorithms: A Survey," 2009 IEEE International Advance Computing Conference (IACC 2009).

[2] Mercy Paul Selvan, A.Chandra Sekar, A.Priya Dharshin, "Survey on Web Page Ranking Algorithms" International Journal of Computer Applications (0975 – 8887) Volume 41– No.19, March 2012.

[3]Kaushal Kumar1, Abhaya2, Fungayi Donewell Mukoko3, "PageRank algorithm and its variations: A Survey report," IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 14, Issue 1 (Sep. - Oct. 2013), PP 38-45.

[4] D. Ganeshiya, D. Sharma, "Keyword Ratio Oriented WebPage Rank Algorithm," IEEE Industrial and Information Systems (ICIIS), 2014 9th International Conference on 15-17 Dec. 2014.

[5] Wenpu Xing and Ali Ghorbani, "Weighted PageRank Algorithm," IEEE Proceedings of the Second Annual Conference on Communication Networks and Services Research on 2004.

[6] C. E. Dinuca, D. Ciobanu, "Web content mining," Annals of the University of Petroşani, Economics, 12(1), 2012, 85-92

[7] S. Qiao, Tianrui Li, Li and Yan Zhu, Jing Peng, Jiangtao Qiu, "SimRank: A Page Rank Approach based on Similarity Measure*," intelligent System and Knowledge Engineering(ISKE),IEEE 2010 international coference on 15-16 Nov.2010

[8] S. Kumar, Kumar Abhishek and M. P. Singh, "Accessing Relevant and Accurate Information using Entropy," Eleventh International Multi-Conference on Information Processing-2015 (IMCIP-2015).

[9] Seifedine Kadry and Ali Kalakech, "On the Improvement of Weighted Page Content Rank," Journal of Advances in Computer Networks, Vol. 1, No. 2, June 2013.

[10] S. Brin, L. Page, "The anatomy of a large-scale hypertextual Web search engine," Computer networks and ISDN systems, 1998, pp.107-117.

[11] G. Kumar, N. Duhan and A. K. Sharma, "Page ranking based on number of visits of links of Web page," Computer and Communication Technology (ICCCT), 2011 2nd International Conference on. IEEE, 2011.

[12]N. Tyagi, S. Sharma. "Weighted Page rank algorithm based on number of visits of Links of web page," International Journal of Soft Computing and EngineeringVol.2, Issue-3, July2012.

[13] Sang-yeon Lee, Young-gi Kim, Seok-Jong Lee, Keon Myung Lee, "An Improvement of Weighted PageRank to Handle the Zero Link Similarity," IEEE SCIS&ISIS 2014, Kitakyushu, Japan, December 3-6, 2014.

[14] C. Ding, X. He, P. Husbands, H. Zha, and H. Simon, "Link Analysis: Hubs and Authorities on the World". Technical report:47847, 2001.

[15] Chakrabarti, Soumen, Byron Dom, David Gibson, Jon Kleinberg, Prabhakar Raghavan, and Sridhar Rajagopalan. 1998. "Automatic resource list compilation by analyzing hyperlink structure and associated text". In Proc. WWW. URL: citeseer.ist.psu.edu/chakrabarti98automatic.html.

[16] Christopher D. Manning Prabhakar Raghavan Hinrich Schütze **"**Introduction to Information Retrieval". Cambridge University Press. 2008. Retrieved 2008-11-09.