A Review Paper on Page Ranking Algorithms

Sanjay* and Dharmender Kumar

Department of Computer Science and Engineering,Guru Jambheshwar University of Science and Technology.

Abstract – Page Rank is extensively used for ranking web pages in order of relevance by mostly all search engines world-wide. There are many algorithms for page ranking such as Google Page Rank algorithm, Hyperlink-Induced Topic Search (HITS) algorithm etc. Some search engine uses link structure based page ranking algorithm while some uses content based. The page ranking algorithm reflects the popularity of a web page in its page rank score. But with the growing requirements of ordering more and more relevant web pages, the traditional page rank algorithm undergo several enhancements and improvements. The main aim of this paper is to discuss the various existing page ranking algorithms and the modification done to the standard page rank algorithm.

Keywords – Page Rank, Weighted Page Rank, Modified PageRank, HITS

I. INTRODUCTION

Vast pool of information related to the user queries is stored in different web pages and is growing day by day. But all information is not relevant to user. Sometimes while surfing on web for some information, user may end to a web page displaying irrelevant data or insufficient information. So, this problem can be solved by displaying the more relevant pages by the search engines. In the early days of internet, the results for any user query by the search engines are displayed on the basis of keyword search mechanism. But this approach is not sufficient in returning useful information in all cases. Sometimes the problem of topic drifting is faced by the users. Therefore a rank score or rank value is associated with the listed web pages by any search engine in response to a user query. The web pages having higher page ranks are listed in the top and thus helps the user in collecting required and important information in the least possible time. This method of ranking web pages is started by the Google search engine and also adopted by several other search engines like Bing, Yahoo etc. Page Rank is being used in various fields like social networking sites, research papers digital libraries, etc apart from the search engines. Over the time this Google Page Rank algorithm is also modified and becomes the base for different Page Ranking algorithms.

The page rank plays important role in the process of Web Mining. Web Mining may be defined as the process of mining useful or important information related to a user query. Web Mining is a Data Mining technique used in discovering various patterns from the web.

Web Mining is divided into 3 main categories:

1. **Web Content Mining**: WCM is the process of extracting the useful information and knowledge from the web page content.
2. **Web Structure Mining**: WSM mainly analyse the node and link structure of a web site.
3. **Web Usage Mining**: WUM is the process of discovering the data stored in server access logs, user profile and pattern in user browsing the web pages.

Page Rank can be computed on the basis of backlinks, forward links, and topic sensitivity. The topic sensitive page ranking algorithm makes use of web structure mining and web content mining. The traditional Page ranking algorithm uses only the linking structure of web pages whereas Hyperlink- Induced Topic Search (HITS) algorithm uses both the linking structure and the content of web pages for calculating the page ranks. Web Structure Mining and Web Content Mining both come under the Web Mining.

II. PREVIOUS WORKS ON PAGE RANKING ALGORITHMS

Page ranking algorithms play important role in displaying results of user query according to the page rank score of the web pages. A large number of researchers have worked in developing an efficient page ranking algorithm and comparing its performance with the various existing algorithms. Wenpu Xing and Ali Ghorbani, introduced

ISSN: 2278 – 1323

Weighted PageRank (WPR) algorithm [4]. The proposed algorithm is based on concept of standard page rank algorithm. WPR gives larger rank score to popular (most important) pages rather than dividing the rank score of a page equally among all outlinks. WPR considers both inlinks (links to a page) and outlinks (links from a page) of the web pages in computing the page rank instead of inlinks used by standard page rank algorithm. The performance of Weighted Page Rank algorithm is better than the standard page rank algorithm in pointing large number of relevant and important pages to a user query. H. Dubey and B.N. Roy, proposed a modified page rank algorithm using normalization technique [1]. The main goal of this proposed algorithm is to minimize the number of iterations executed in Page Rank algorithm. Thereby minimizing the time complexity of page rank algorithm and converging point is obtained much earlier as compared to the traditional Page Rank algorithm. The proposed algorithm includes normalization technique using mean value of the page rank of the web pages. TIAN Chong, presented a new page ranking algorithm using classified tree [2]. Here a classified tree is made according to the similar searching results of different users. This approach eliminates the problem of Topic–Drift and older outdated pages caused by conventional page rank algorithm. It adds the advantage of increased the search efficiency without compromising the searching speed factor of algorithm. G. Kumar and N. Duhan et al., proposed a page ranking algorithm based on visits of links [5]. Mostly the various page ranking algorithms are link or content based. The proposed approach deviates from the common trend by considering the additional factor of the number of visits to the inbound links of page. Larger the rank value of web page means that page is mostly visited by users. Hence the presented method displays the most important pages based on the user's interest and minimizes the search space. M. Sehgal and Priya, extended the standard page rank algorithm by using the Time factor [3]. Basically the rank of any web page increases as the more number of users click on that web page. But in this approach the page rank does not change only by user clicks, because sometimes the visited page may not be useful for user. So, this drawback can be overcome by making the use of total time spend by any user on any web page along with the number of user clicks on that page. Therefore the proposed algorithm is found better than the page rank algorithm using only number of clicks. N. Tyagi and S. Sharma, proposed an algorithm called Weighted Page Rank Algorithm based on number of visits of links [6]. The proposed algorithm is an extension to Weighted Page Rank Algorithm and makes comparison between original Weighted Page Rank method and Visits of Links method. The original WPR algorithm assigns page rank value to each outlink page directly proportional to its popularity. But more is the number of visits to a links by user, more is the importance of web page. In the proposed algorithm the most visited outlinks have higher rank value and thus higher popularity. P. Patel, discussed the basic idea of page rank based on number of visits by user to any web page [18]. The dependence of page rank on the damping factor is also analysed and concluded that most favourable value of damping factor is 0.85, and is used by majority of ranking algorithms.

Y. Qin and D. Xu, considered the human factor and introduced a balanced rank algorithm on the basis of page belief recommendation and page rank [10]. This algorithm includes the credit evaluation mechanism and helpful in eradicating the problem of topic drift. S. N. Mishra et al., devised a Topic sensitive weighted page rank algorithm [23]. This algorithm is based on web structure mining and produces better results against a user query. The basic page rank algorithm is independent of user search query. This algorithm based on Topic Sensitive Link Analysis gives better scores to the important pages. S. Setayesh and A. Harounabadi et al., presented a modified Page Ranking Algorithm based on the concept of Ant Colony [8]. Here each user is assumed as an ant and the user's interest or choice for any web page is considered as the pheromone left by the ants. In the end the tour from all the ants are considered and the amount of pheromone associated with each web page is used for the calculation of page ranks. Thus the proposed method leads to better rank values and less prone to errors than that of traditional Page Rank. Yen, Chia-Chen, et al., proposed an associated page rank algorithm by using page relevance measurement of web documents [7]. This algorithm consider the efficiency in addition to the accuracy. The link spam problem can be examined in the web pages from the analysis of web pages relevance values. Selection of common words and using techniques of word stemming can make the algorithm more precise. This algorithm can further minimizes the complexity of topic-sensitive page rank. R. Jain et al., proposed a Dynamic Page Ranking algorithm which is an extension to the basic page rank algorithm [13]. Different Word Sense Disambiguation (WSD) approaches are used in this algorithm. WSD deals with the identification

ISSN: 2278 – 1323

of the senses of word having multiple meanings in the textual context.

S. Mulla and M. Takalikar, formulated a new method to minimize the delay in server requests [11]. The proposed algorithm includes the iterative diagram based standing page rank calculations. The Modified Page Rank reduces the processing and request time of server by making use of similarity profiling of sentences. N. Preethi and T. Devi, presented a new Integration Case and Relation Based Page Rank algorithm [12]. The Case Based Reasoning (CBR) is learning and solving problems of some specific situations on the basis of past experiences of same situations. This algorithm makes use of textual case based reasoning. The number of incorrect result pages can be reduced by relation based and textual case based ranking algorithms. S. Gupta et al., devised a new combined page ranking scheme for efficient information retrieval [16]. The proposed architecture used the old tf/idf weight of terms and context based indexing to compute the page rank of web documents. Thesaurus was used to create context based index. Young-Chon Kim et al., used syntactic classification of pages and devised a new page ranking algorithm [20]. The proposed method includes the following 3 steps: first, properties of pages are selected on the basis of user's demand. Second, measure those properties and third, each property is given a specific weightage while ranking pages. P. Sharma and D. Tyagi et al., proposed a Weighted Page Content Rank Algorithm that uses the concept of web content and structure mining in addition to the weighted page rank algorithm [25]. WPCR is denoted by a numerical value representing the rank of a particular web page. The proposed method solves the problems with the traditional algorithms and gives higher ranks to the most relevant pages. M. Shamiul Amin et al., proposed a new score based page ranking algorithm [19]. The proposed algorithm involves usage information and web content mining. Both Semantic and syntactic matches are considered in this approach. Syntactic score is computed on the basis of total number of words exactly matched in the page and semantic score is calculated by making the use of synonym matched. Finally both scores are added to compute the total relevancy score of each web page. This algorithm produces better page rank scores to the web pages in comparison to the various existing page ranking algorithms.

III. PAGE RANK

In 1990s, the text based searching of user query is the only mechanism of displaying results by the search engine. But with the extensive growth in the information stored on the web, this method is not fully sufficient to satisfy the user need of the best useful web pages. Then the concept of ranking a web page is introduced. After that different ranking algorithms evolved based on different factors like some are link structure based and some are content based or combination of both. Page Rank is a type of "vote", by all web pages depicting the importance of a particular web page. A single link to a web page is considered as a single vote of support. The value of Page Rank varies from 0 to 10. The page rank gives the probability of landing a user on a web page by clicking on links at random. Larger the number of inbound links to a page, more is the page rank value of that page and higher is the probability that user will reach to that web page. Also if the inbound links to a web page are coming from important websites then the page rank score is higher and vice versa. The different page ranking algorithms are as follows:

1. PAGE RANK ALGORITHM

In 1998, the founder of Google search engine, Larry Page and Sergey Brin invented the Page Rank Algorithm to quantize the importance of millions of web pages comprising the World Wide Web (WWW). Page Rank of a web page is a numerical number representing the importance of that web page based on the number of inbound links. The basic concept of PageRank is that the importance of a page is directly proportional to the number of web pages linking to that page. So Page Rank algorithm considers a page more important if large number of other web pages are linking to that page or if links are coming from some of most important and popular web pages. Page Rank of whole website is not valid because page rank is associated with every web page on the web. Page Rank of a web page X is calculated by the page rank of those pages that links to page X using formula given below:

$$PR(X) = (1-d) + d\left[\frac{PR(Y_1)}{C(Y_1)} + \ldots\ldots + \frac{PR(Y_n)}{C(Y_n)}\right]$$

where,

PR (X) = Page Rank of web page X,

ISSN: 2278 – 1323

PR (Y_i) = Page Rank of pages Y_i that links to a web page X

C (Y_i) = Number of outbound links on web page Y_i

d = Damping Factor (value between 0 and 1, but usually value is 0.85)

Repeat the above step involving the calculation of page rank until two consecutive same values are obtained.

Page rank algorithm used by search engine displays the web page according to their page rank. Page rank algorithm is non- keyword specific and link structure based algorithm. The major disadvantage of page rank algorithm is that it is calculated and stored at the time of indexing and not at the time of query. The Search Engine Ranking Positions of any website depends upon the actual current status of that website whereas page rank calculated by using page ranking algorithm depends upon the stored database which is updated only once in a period of two-three months. The page rank Algorithm divides the page rank value equally to all the outlinking web pages but all outlinks are not equally important or relevant to user query. This problem can be solved by considering an additional weight factor in the calculation of page rank and gives birth to a new algorithm known as Weighted Page Rank algorithm.

2. WEIGHTED PAGE RANK ALGORITHM

Wenpu Xing and Ali Ghorbani proposed an algorithm called weighted page rank (WPR). This weighted page rank algorithm is different from the traditional page rank algorithm in the fact that each outlink page has a page rank value proportional to its importance (number of inlinks and outlinks) instead of dividing it equally [4].

W^{in} (v, u) = weight of link (v, u) or importance of web page due to inlinks

W^{out}(v, u) = weight of link (v, u) or importance of web page due to outlinks

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

where, $I_u$and I_p denote the no. of inlinks of page u and page p respectively

R(v) represents the reference page list of page v

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

where, O_u and O_p denote the no. of outlinks of page u and page p respectively

R(v) represents the reference page list of page v

After calculating the importance of web pages, the modified page rank formula is given as:

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

This Weighted Page Rank algorithm solves the problem of ranking web pages based on their relevancy or importance by considering the weight factor. But the problem of query independency and calculation of page ranks at indexing time still remain with WPR and with the traditional Page Ranking algorithm. To solve the problems of query independency and indexing time page rank calculation, a new algorithm is introduced which is known as HITS algorithm.

3. HITS ALGORITHM

Hyperlink- Induced Topic Search Algorithm is proposed by Jon Kleinberg. The most popular social networking website Twitter makes use of HITS algorithm in suggesting user accounts to follow. This algorithm calculates the page rank at the query time and solves the problem of indexing time page rank calculation faced in previous page ranking algorithms. HITS algorithm is a link structure analysis algorithm. It ranks the web page based on two scores Hubs and Authority instead of a single score. HITS algorithm is executed at the query time rather than at the time of indexing. Hubs denote the web pages acting as the resource lists and Authority denotes the pages having useful information. Most of the web pages act as hubs as well as authorities simultaneously. HITS algorithm is having following two steps:

(i) **Sampling Step:** collect the set of relevant pages for the given query.

(ii) **Iterative Step:** Found the Hubs and Authorities using output of step 1.

The weight of Hub (H_p) and the weight of Authority (A_p) can be calculated using following formulae:

ISSN: 2278 – 1323

$$H_p = \sum_{q \in I(p)} A_q$$

$$A_p = \sum_{q \in B(p)} H_q$$

Where, H_q = Hub score of a web page

A_q = Authority score of a web page

I (p) = set of reference pages of page p

B (p) = set of referrer pages of page p

The hub weight of a web page = sum of authority weights of pages that it links to.

The authority weight of a web page = sum of hub weights of all the pages that links to it.

HITS algorithm makes use of both backlinks and forward links whereas traditional page rank algorithm uses only backlinks. HITS solves the problem of indexing time page rank calculation. But there is a problem of topic drift and efficiency problem with the HITS algorithm.

IV. CONCLUSIONS

In this paper, we discussed the various algorithms and techniques mainly used by search engines in ranking web pages on the internet. With the course of time the traditional page rank algorithm has been modified by adding many different factors. Google Page Rank Algorithm computes the page ranks of web pages only at the time of indexing and HITS algorithm computes the page ranks at the time of user query. But these modification are not sufficient to cope with the increasing data or information on every web page day-by-day. There is a need of some kind of modified algorithm that can give results at the time of indexing as well as at the time of user query. The existing algorithms may consider the bookmarked web pages in calculating the Page Rank of web pages. The Page Ranking algorithms are now finding applications not only in ranking web pages but extensively used in ranking research papers, suggesting user accounts to follow and in many other fields.

REFERENCES

[1] H. Dubey and B. N. Roy. "An improved page rank algorithm based on optimized normalization technique," *International Journal of Computer Science and Information Technologies*, ISSN: 0975-9646 Vol. 2 (5), pp. 2183-2188, 2011.

[2] T. I. A. N Chong, "A kind of algorithm for page ranking based on classified tree in search engine." *Computer Application and System Modelling (ICCASM), 2010 International Conference on*. IEEE Vol. 13, pp. V13-538, 2010.

[3] S. Monica, Priya. "Enhanced Page Rank Algorithm Using Time Factor.*"International Journal of Engineering and Computer Science*, ISSN: 2319-7242 Volume 03 Issue, pp. 6990-6995, 2014.

[4] Wenpu Xing and A. Ghorbani. "Weighted pagerank algorithm." *inProc.Second Annual Conference, Communication Networks and Services Research*,IEEE,pp. 305-314, 2004.

[5] G. Kumar et al. "Page ranking based on number of visits of links of Web page." *Computer and Communication Technology (ICCCT), 2011 2nd International Conference on*. IEEE, 2011.

[6] N. Tyagi and S. 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, 2012

[7] Yen, Chia-Chen, et al. "Pagerank algorithm improvement by page relevance measurement." *Fuzzy Systems, FUZZ-IEEE. IEEE International Conference on*. IEEE, 2009.

[8] S. Setayesh and A. Harounabadi et al. "Presentation of an Extended Version of the PageRankAlgorithm to Rank Web Pages Inspired by Ant ColonyAlgorithm."*International Journal of Computer Applications*, Volume 85 – No 17, ISSN: 0975 – 8887 January 2014.

[9] A. Humad and V. Solanki. "A New Context Driven Page Ranking Algorithm."*International Journal of Latest Trends in Engineering and Technology (IJLTET)*, Volume 4, Issue 1, ISSN: 2278-621X, May 2014.

[10] Y. Qin and D. Xu. "A Balanced Rank Algorithm based on page rank and page belief Recommendation." *IEEE*, 2010.

[11] S. Mulla and M. Takalikar. "Latest information summarization using modified page rank algorithm."*International Journal of Computer Technology and Applications (IJCTA)*, Vol 5 (6), pp. 1845-1848, Dec 2014.

[12] N. Preethi and T. Devi "New Integrated case and relation based page rank algorithm." *International Conference on Computer Communication and Informatics (ICCCI)*, IEEE, 2013.

[13] R. Jain et al. "Enhanced retrieval of web pages using improved page rank algorithm." *International Journal on Natural Language Computing (IJNLC)*, Vol. 2, No. 2, April 2013.

[14] J. Beel and B. Gipp. "Google scholar's ranking algorithm: the impact of citation counts." *Proceedings of the 3rd IEEE International Conference on Research Challenges in Information Science (RCIS'09)*, IEEE, pages 439–446, April 2009.

[15] A. Jain et al. "Page ranking algorithm in web mining, limitations of existing methods and a new method for indexing web pages." *International Conference on Communication Systems and Network Technologies*, IEEE, 2013.

ISSN: 2278 – 1323

[16] S. Gupta et al. "New combined page ranking scheme in information retrieval system." *International Journal of Scientific and Research Publications*, Volume 4, Issue 4, ISSN: 2250-3153, April 2014.

[17] KadrySeifedine and Kalakech Ali. "On the improvement of weighted page content rank" Journal of Advance in Computer Networks, Vol.1, No.2, June 2013.

[18] P. Patel. "Research of Page ranking algorithm on search engine using Damping factor" International Journal of Advance Engineering and Research Development (IJAERD), Volume 1, Issue 1, ISSN: 2348-4470, Feb 2014.

[19] M. Shamiul Amin et al. "A score based web page ranking algorithm" *International Journal of Computer Applications*, Volume 110, No.12, January 2015.

[20] Young-Chon Kim et al. "A syntactic classification based web page ranking algorithm." *6th International Workshop on MSPT Proceedings*, 2006.

[21] EirinakiMagdalini and VazirgiannisMichalis "UPR: Usage based page ranking for web personalization", *5th IEEE International Conference on Data Mining (ICDM) Proceedings*, 2005.

[22] F. Lamberti et al. "A relation-based page rank algorithm for semantic web search engines" *IEEE Transactions on Knowledge and Data Engineering*, Vol. 21, No. 1, pp. 123-136, January 2009.

[23] S. N. Mishra et al. "An effective algorithm for web mining based on topic sensitive link analysis" *International Journal of Advanced Research in Computer Science and Software Engineering*, Volume 2, Issue 4, April 2012.

[24] T. S. Govada and N. L. Prasanna "Comparitive study of various page ranking algorithms in web content mining" *International Journal of Advanced Research (IJAR)*, Volume 2, Issue 7, pp. 457-464, 2014.

[25] P. Sharma et al. "Weighted page content rank for ordering web search result" International Journal of Engineering Science and Technology (IJEST), Vol. 2 (12), pp.7301-7310, 2010.

ISSN: 2278 – 1323