



Issues in Sequential Web Page Ranking Algorithms

¹Atul Kumar Srivastava*, ²Rakhi Garg, ³P. K. Mishra

^{1,3}Department of Computer Science, Banaras Hindu University, Varanasi, Uttar Pradesh, India

²Computer Science Section, MMV, Banaras Hindu University, Varanasi, Uttar Pradesh, India

Abstract- *With the massive and diverse information available on the Web it is very tedious job for the user to get relevant information from Web. There are several search engines which help users to surf the Web pages according to their query. It is extremely difficult for the Web user to get the required information in efficient time as the Web pages available are very large in number. Web page ranking algorithms plays an important role to rank the Web pages so that Web user get the Web pages relevant to their query search. Two important ranking algorithms: HITS and PageRank have been proposed to get relevant Web pages. Both algorithms computes the rank in iterative nature, which takes more computation time and resources. Because of their iterative nature the issues in HITS and PageRank algorithms are efficiency, relevancy and scalability.*

In this paper, we mainly focus on these issues in detail. Such issues can be resolved by designing algorithms for Parallel, Distributed and Grid environment. Now day's researchers and scientist are working in this direction to design algorithm that not only handles these issues but also enhances the performance as well.

Keywords— *Web Mining, Web Structure Mining, PageRank algorithm, HITS algorithm*

I. INTRODUCTION

Web is one of the most popular medium to access information. Due to exponential growth of data on Web, accessing relevant information is becoming a key issue for Web users as well as Web site developer. There are several issues like scalability, efficiency, relevancy, multimedia data, etc. due to huge, dynamic, heterogeneous (Structured data, Semi-structured Web pages, unstructured text and multimedia files) nature of Web data [1,2]. Due to all these characteristics of Web data it is not an easy task for the user to find relevant information from Web. To solve these problems several techniques: Database (DB), Natural language processing (NLP), Information Retrieval (IR) and Web mining have been evolved [4, 5]. Among them Web mining is a converging research area to solve these issues in last few decades.

Web mining is used to discover useful information or knowledge from hyperlink structure of Web pages, contents of Web pages and stored user's activity log [5]. It uses basic data mining techniques like Association rule, Classification, Clustering, Regression etc. to extract the useful pattern from Web data. Apart from that it also uses some techniques of Information Retrieval, NLP, and Social Network Analysis etc. Based on the heterogeneity of data, Web mining can be categorized into three types: Web Content Mining, Web Usage Mining and Web Structure Mining [2, 5].

Web Content Mining extract useful information or knowledge from the contents of Web pages. Web Usage Mining is used to discover the navigation behaviour of user from the Web server log data. Web Structure Mining is used to extract the relevant information or knowledge from the hyperlink structure of Web [4, 18]. From the hyperlink structure of Web we can extract the relevant Web pages, and discover the communities of Web pages which share common interest [2, 4, 5].

Due to heterogeneous and huge data available on the Web, it is very tedious job to find relevant information from the Web. Web structure mining uses social network analysis and hyperlink structure analysis to provide the relevant information to user [2, 5]. Earlier search engines extract relevant Web pages based on the similarity of content of Web pages and user's query. But this mechanism alone was no longer sufficient for search engines due to following reasons [7, 11]:

- Web data increases exponentially, so for any query the search engines provide 10 million number of pages in search result page.
- It is easy to spam content similarity methods.

In the year 1998, two famous algorithms: PageRank and HITS based on hyperlink structure of Web have been proposed by S. Brin & L. Page, J. Kleinberg to overcome above two issues. PageRank is the basic algorithm on which Google search engine has been developed. In this algorithm the importance of a Web page is proportional to the importance score of all Web pages links to it. HITS algorithm uses a query to select a sub graph from the Web graph. From this sub graph two types of importance score for a Web page is identified: authoritative score and hub score [3, 8].

In this paper, we give overview of Web structure mining and discuss various link analysis ranking algorithm. Section II cover the overview of web structure mining and link analysis algorithm. In section III and IV we discuss about HITS and PageRank algorithm in detail. Section V focus on the issues involved in both the algorithms and discuss various improvements on sequential PageRank algorithm proposed by different researchers.

II. WEB STRUCTURE MINING

Web structure mining is used to extract the relevant information or knowledge from the hyperlink structure of Web pages. Basically it aspires to obtain insights on how Web pages are structured, how Web pages are linked among each other (known as link analysis) [2,5, 19]. It is further categorized into intra-page structure mining and inter-page structure mining as shown in figure 1.

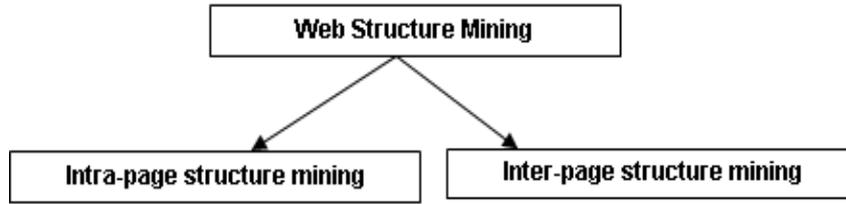


Figure 1: Categorization of Web Structure Mining

III. HITS ALGORITHM

It is a query dependent algorithm and computes the rank of Web pages online i.e. whenever a Web user enter a query in search engine, then first it extract the Web pages which are relevant to user’s query and then calculate two rank score: authoritative score and hub score for each Web pages [8, 9]. Authoritive score measures the number of incoming links of that page i.e. a Web page with many incoming links. Hub score measures the number of outgoing link of that page i.e. a Web page with many outgoing links. The main idea of these two score is that a good hub page should be links to many good authoritative Web pages and a good authoritative Web pages should be linked by many good hub pages. Thus, authoritative and hub score have a mutual reinforcement relationship between them. By using this relation the two score iteratively updates and is computed as shown in figure 2 [8]:

$$a_p = \sum_{q:(q,p) \in E} h_q, \quad h_p = \sum_{q:(q,p) \in E} a_q$$

Figure 2: Iterative relation to Compute Hub and Authoritive Score

Where **h** and **a** represents the n-dimensional hub and authority score of each Web pages.

Disadvantage

- It does not have capability to fight with spam Web pages. Spammers easily add out-links from one’s own page to point several good authoritative Web pages. It increases their hub score.
- Due to query dependent nature and online computation this algorithm takes more time.
- This algorithm suffers from topic drift problem.

IV. PAGERANK

This algorithm is static algorithm as it computes the rank score of each Web pages offline and does not dependent on the search query. In PageRank, the importance of Web page *u* (pagerank score) is calculated by adding up the pagerank score of all Web pages that point to Web page *u*. Since a Web page point to many other Web pages so its pagerank score is shared among all the Web pages which it points. Mathematical equation to compute pagerank value iteratively is shown in figure 3 [3, 7, 9]:

$$R(u) = \sum_{v \in B(u)} R(v) / O(v)$$

Figure 3: Iterative equation to compute PageRank Score

Where *R(u)* and *R(v)* pagerank score of Web pages of *u* and *v*, *B(u)* is the set of Web pages which point to page *u* and *O(v)* is the number of outgoing links of Web page *v*. There are some issues with sequential Web PageRank algorithm as follows [5, 9]:

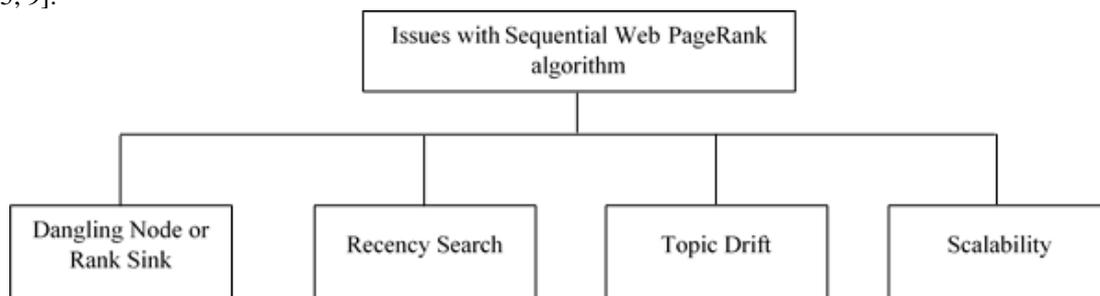


Figure 4: Categorization of PageRank Issues

V. ISSUES

We discuss above issues in detail and some improved PageRank algorithm proposed by researchers based on above issues as follows:

A. Dangling Node or Rank Sink Issue

Dangling nodes are those nodes in Web graph which have no outgoing link. So during the iteration they only accumulate rank score and does not distribute their ranks to other Web pages [5]. To overcome with this issue several researchers proposed their algorithms. Some algorithms are discussed below:

Sergey Brin and Lawrence Page [3] uses the stochastic matrix to find the dangling links. They simply remove all dangling link until all PageRank are calculated and after calculation simply add those dangling links. They uses the notion of damping factor to get out of the loop to solve cyclic issues shown in Equation 1:

$$PR(u) = \frac{1-d}{n} + d \sum_{v \in B(u)} \frac{PR(v)}{Od(v)} \quad (1)$$

Where $PR(u)$ and $PR(v)$ PageRank score of Web pages u and v , $B(u)$ is the set of Web pages which point to page u and $Od(v)$ is the number of outgoing links of Web page v and d is the damping factor ($0 < d < 1$).

Sung Jin Kim and Sang Ho Lee [11] proposed improved PageRank algorithm which face the norm-leak phenomena that occur due to the dangling nodes. They define new hyperlink matrix H in which all dangling column have value $[1/n]$ and then compute the PageRank (A^*) value by using this equation shown in Equation 2:

$$A^* = (1 - d)A + d[1/n]_{n \times n} \quad (2)$$

Where denotes PageRank matrix, d is the damping factor ($0 < d < 1$) and n is the number of Web pages

B. Recency Search Issues

The important factor in PageRank algorithm is the number of incoming links to any Web page. The old Web pages in Web have more incoming link than newly crawled Web pages so old Web pages get higher rank value than new ones. But sometimes it may be that relevant information in newer Web pages [5, 6]. To solve this issue many research proposed improved PageRank algorithm:

Jing Wan and Si-Xue Bai [16] proposed improved PageRank algorithm based on the time activity model. They combines the time activity model (based on Web site, user interest, Web pages contents and Web site developer) with the traditional PageRank algorithm. Through this time activity model given in equation (3), high time activity Web pages can be extracted in time:

$$PR_{final} = \text{Time - activity score of web page} \times PR \quad (3)$$

Where PR_{final} is the final pagerank score of Web pages.

Shiguang Ju, Zheng Wang and Xia Lv [14] improved the PageRank method by proposing a new method. This method combines the last modification time of Web pages with the in-link and out-link weight of concerned Web pages and calculated by using equation (4):

$$PR_t(p) = \alpha \frac{1}{n} + (1 - \alpha) \sum_{q \in B(p)} PR(q) W^i W^o W^{decay} \quad (4)$$

W^i and W^o is the In-link weight and out link weight of link (q, p) and W^{decay} is links decay weight.

Wenpu Xing and Ali Ghorbani [10] proposed a weighted pagerank algorithm in which they considered the importance of both number of incoming and number of outgoing links of Web pages and distributes the rank based on the importance of Web pages, and they observed that this algorithm discovers larger number of Web pages regarding user's query shown in equation (5):

$$PR(u) = \left(\frac{1-d}{n} \right) + d \sum_{v \in B(u)} PR(v) W_{(v,u)}^{in} W_{(v,u)}^{out} \quad (5)$$

Where $W_{(v,u)}^{in}$ is the weight of incoming link and $W_{(v,u)}^{out}$ weight of outgoing link.

C. Topic Drift Issues

PageRank is query independent algorithm. So it is unable to predict whether a hyperlink is related to user’s query or subject of Web pages [5, 7]. Due to this it may suffers with topic drift problem. Many researches have been proposed improved PageRank algorithm which avoid the topic drift problem.

Xiaoyun Chen, BaojunGao and Ping Wen [12] solve the issue of topic drift problem by using latent semantic model. This method distributed the rank score of Web page to its outgoing Web pages based on content similarity between Web pages. Latent semantic model is used to determine the similarity between Web pages. They analysed the result obtained by applying this algorithm and observed that it provide the more relevant Web pages than basic PageRank, but this algorithm takes more computation time than PageRank algorithm.

Zhou Cailan and Chen Kai [13] uses the feedback of user clicks and improved the PageRank method. This method stores the user search log and click log file based on search query, which contains click time and information about URL’s. This method add an attribute with PageRank algorithm i.e. weight of user click and computes pagerank shown in equation 6:

$$PR(i) = PR(i) \frac{S(i)}{T(i)} \tag{6}$$

Where $S(i)$ denotes click weight and $T(i)$ denotes weight of click time.

D. Scalability

Several researchers proposed the method which solves the scalability issue of PageRank computation. Yuan Wang and David J, DeWitt [15] proposed a framework which computes PageRank in distributed manner i.e. every Web server computes PageRank locally over its own data and then merged the result from every web server to determine global ranking. They compute local Page Rank vector in individual Web server by equation (7).

$$\gamma_{l(m)} = PR(G_{(m)}, V_{n_m}, V_{n_m}) \tag{7}$$

Where, $G_{(m)}$ represents the server m that contains n_m pages and V_{n_m} is the uniform column vector of dimension n_m compute the server rank vector.

$$\gamma_s = PR(G_s, V_{n_s}, V_{n_s}) \tag{8}$$

Where G_s represent the server link graph.

Shumingshi, Jin Yu, and Guang Wen Yang [17] proposed distributed page rank. For the computation of rank they uses open system PageRank method. To reduce the communication overhead between servers they uses indirect transmission.

Table 1 summarizes the various issues discussed above handled by the respective algorithms designed by the authors.

Table 1: Issues handled by the respective algorithm designed by authors

Algorithm	Algorithm designed by Authors	Handling issue
Basic PageRank algorithm	Sergey Brin and Lawrence Page [3]	Rank Sink Problem, Cyclic Problem
PageRank algorithm using damping factor	Sung Jin Kim and Sang Ho Lee [11]	Dangling link
Timestamp and link based PageRank algorithm	ShiguangJu, Zheng Wang and Xia Lv [14]	Recency Search
Time-activity-curve based PageRank algorithm	Jing Wan and Si-XueBai [16]	Recency Search
Feedback of user-click based Pagerank algorithm	Zhou Cailan and Chen Kai [13]	Topic Drift Problem
LSM-PageRank algorithm	Xiaoyun Chen, BaojunGao and Ping Wen [12]	Topic Drift Problem
DISS-PageRank algorithm	Yuan Wang and David J, DeWitt [15]	Scalability
Distributed PageRank algorithm	Shumingshi, Jin Yu, and Guang Wen Yang [17]	Scalability

VI. CONCLUSION

Web Page ranking algorithms play an important role to find the quality of Web pages provided by Web search engine. Various link analysis ranking algorithms like HITS, PageRank were proposed to rank the Web pages. This paper mainly focus on PageRank algorithm and their improvements based on the issues of the basic PageRank algorithm. Sequential improved PageRank algorithms almost resolved issues like: *Rank Sink, Dangling node, topic drift etc.* observed in the basic PageRank algorithms. However, some issues like scalability, relevancy etc. are still present and work should be done in this direction to resolve these aforesaid issues. To overcome these issues research is going on to apply sequential PageRank algorithm on Parallel, Distributed and Grid environment.

REFERENCES

- [1] Kosala, Raymond, and Hendrik Blockeel. "Web mining research: A survey." ACM Sigkdd Explorations Newsletter 2.1 (2000): 1-15.
- [2] Graubner-Müller, Alexander. Web Mining in Social Media: Use Cases, Business Value, and Algorithmic Approaches for Corporate Intelligence. No. 3. BoD—Books on Demand, 2011.
- [3] Brin, Sergey, and Lawrence Page. "The anatomy of a large-scale hypertextual Web search engine." Computer networks and ISDN systems 30.1 (1998): 107-117.
- [4] Jicheng, Wang, et al. "Web mining: knowledge discovery on the Web." Systems, Man, and Cybernetics, 1999. IEEE SMC'99 Conference Proceedings. 1999 IEEE International Conference on. Vol. 2. IEEE, 1999.
- [5] Liu, Bing. Web data mining: exploring hyperlinks, contents, and usage data. Springer Science & Business Media, 2007.
- [6] Desikan, Prasanna, et al. "Hyperlink Analysis—Techniques & Applications." Army High Performance Computing Center Technical Report (2002).
- [7] Borodin, Allan, et al. "Link analysis ranking: algorithms, theory, and experiments." ACM Transactions on Internet Technology (TOIT) 5.1 (2005): 231-297.
- [8] Kleinberg, Jon M. "Authoritative sources in a hyperlinked environment." Journal of the ACM (JACM) 46.5 (1999): 604-632.
- [9] Berkhin, Pavel. "A survey on pagerank computing." Internet Mathematics 2.1 (2005): 73-120.
- [10] Xing, Wenpu, and Ali Ghorbani. "Weighted pagerank algorithm." Communication Networks and Services Research, 2004. Proceedings. Second Annual Conference on. IEEE, 2004.
- [11] Kim, Sung Jin, and Sang Ho Lee. "An improved computation of the PageRank algorithm." Advances in Information Retrieval. Springer Berlin Heidelberg, 2002. 73-85.
- [12] Chen, Xiaoyun, Baojun Gao, and Ping Wen. "An Improved PageRank Algorithm Based on Latent Semantic Model." Information Engineering and Computer Science, 2009. ICIECS 2009. International Conference on. IEEE, 2009.
- [13] Cailan, Zhou, Chen Kai, and Li Shasha. "Improved PageRank algorithm based on feedback of user clicks." Computer Science and Service System (CSSS), 2011 International Conference on. IEEE, 2011.
- [14] Ju, Shiguang, Zheng Wang, and Xia Lv. "Improvement of page ranking algorithm based on timestamp and link." Information Processing (ISIP), 2008 International Symposium on. IEEE, 2008.
- [15] Wang, Yuan, and David J. DeWitt. "Computing pagerank in a distributed internet search system." Proceedings of the Thirtieth international conference on Very large data bases—Volume 30. VLDB Endowment, 2004.
- [16] Wan, Jing, and Si-Xue Bai. "An improvement of PageRank algorithm based on the time-activity-curve." Granular Computing, 2009, GRC'09. IEEE International Conference on. IEEE, 2009.
- [17] Shi, ShuMing, et al. "Distributed page ranking in structured p2p networks." Parallel Processing, 2003. Proceedings. 2003 International Conference on. IEEE, 2003.
- [18] Srivastava, Mitali, Rakhi Garg, and P. K. Mishra. "Preprocessing Techniques in Web Usage Mining: A Survey." International Journal of Computer Applications 97.18 (2014): 1-9.
- [19] Srivastava, A. K., Srivastava, M., Garg, R., & Mishra, P. K. International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS) www.iasir.net/algorithms, 3(7), 14.