National Conference on Recent Research in Engineering and Technology (NCRRET-2015)
International Journal of Advance Engineering and Research Development (IJAERD)
e-ISSN: 2348 - 4470 , print-ISSN:2348-6406

Review of Various Web Page Ranking Algorithms in Web Structure Mining

Asst. prof. Dhwani Dave
Computer Science and Engineering
DJMIT ,Mogar

Abstract: The World Wide Web contains large amount of data. These data is stored in the form of web pages .All these pages can be accessed using search engines. These search engines need to be very efficient as there are large number of Web pages as well as queries are submitted to the search engines. Page ranking algorithms are used by the search engines to present the search results by considering the relevance, importance and content score. Several web mining techniques are used to order them according to the user interest. In this paper such page ranking techniques are discussed.

Keywords: Web Content Mining, Web Usage Mining, Web Structure Mining, PageRank, HITS, Weighted PageRank

I. INTRODUCTION

The web is a rich source of information and it continues to increase in size and difficulty. Efficient and effective retrieval of the necessary web page on the web is becoming a challenge aspect now days [1]. The Web is unstructured data warehouse, which delivers the mass amount of information and also enlarges the complexity of dealing information from different perspective of knowledge searchers, business analysts and web service providers [2]. Beside, the Google report on in 2008 that there are 1 trillion unique URLs on the web [3]. Web has grown enormously and the usage of web is unbelievable so it is essential to understand the data structure of web. Because of the massive amount of information it becomes very hard for the users to find, extract, filter or evaluate the relevant information. This issue lifts up the attention to the obligation of some technique that can solve these challenges.

The paper is organized as follows- The categories of Web Mining are discussed in Section 2. Section 3 explains the important of Web Page Ranking and two important algorithms such as Hypertext Induced Topic Selection (HITS) algorithm and PageRank algorithm. In section 4, we explore the comparison between Web Page Ranking algorithms used. The Conclusion remarks are given in Section 5.

II. WEB MINING CATEGORIES

Web Mining consists of three main categories based on the web data used as input in Web Data Mining. (1) Web Content Mining; (2) Web Usage and (3); Web Structure Mining.

A. Web Content Mining

Web content mining is the procedure of retrieving the information from the web into more structured forms and indexing the information to retrieve it quickly. It focuses mainly on the structure within a web documents as an inner document level [9].

B. Web Usage Mining

Web usage mining can be defined as one of the application of data mining techniques to discover interesting usage patterns from web usage data, in order to understand and better serve the needs of web-based applications [2].Web-usage mining mines the secondary data derived from the behavior of users while interacting with the web. This includes data from Web server-access logs, proxy-server logs, browser logs, user profiles, registration data, user sessions or transactions, cookies, bookmark data etc [9].

C. Web Structure Mining

Web structure mining is defined as the process by which we discover the model of link structure of the web pages. We classify the links; generate the ease of use information such as the similarity and relations among them by taking the advantage of hyperlink topology [4]. PageRank and hyperlink analysis fall in this class.

Overview of these three web mining categories is explained and compared in the following Table 1:

Criteria	Web Mining			
	Web	Content	Web Usage	Web

	Mining	Mining	Structure Mining
View of Data	-Unstructured -Structured	User Interactivity	-Link Structure
Main Data	- Text documents -Hypertext documents	-Server Logs (log-files) -Browser Logs	-Link Structure
Representation	-Bag of words, n-gram Terms, -Phrases, Concepts or Ontology -Relational Learning	-Relational Table - Graph -User Behavior	-Graph -Web pages Hits
Method	-Machine Learning -Statistical (including: NLP)	-Machine Learning -Statistical Association rules	Proprietary algorithm -Web PageRank
Application Categories	-Categorization -Clustering -Finding extract rules patterns in text	-Site Construction -Adaptation and management -Marketing, -User Modeling	Categorization - Clustering

Table 1: Comparison of different Web Mining categories

III. WEB PAGE RANKING ALGORITHM

Page ranking algorithms are used by the search engines to present the search results by considering the relevance, importance and content score The web mining techniques are used to order them according to the user interest. Some ranking algorithms depend only on the analysis of the link structure of the documents i.e. their popularity scores (web structure mining), whereas others look for the actual content in the documents (web content mining), while some use a combination of both i.e. they use content of the document as well as the link structure to assign a rank value for a given document. There are number of algorithms proposed based on link analysis. Three important algorithms, such as PageRank, Weighted PageRank and HITS (Hyper-link Induced Topic Search) are discussed below.

A. PageRank Algorithm (Google)

The "PageRank" algorithm, proposed by founders of Google Sergey Brin and Lawrance Page, is one of the most common page ranking algorithms. This algorithm is also currently used by the leading search engine Google. The algorithm uses the linking (citation) info occurring among the pages as the core metric in ranking procedure. Existence of a link from page p1 to p2 may indicate that the author is interested in page. The PageRank metric PR(p), defines the importance of page p to be the sum of the importance of the pages that point to p. More formally, consider pages p1,…,pn, which link to a page pi and let cj be the total number of links going out of page pj. Then, PageRank of page pi is given by:

$$PR(pi)=d+(1-d)[PR(p1)/c1+\ldots+PR(pn)/cn]$$

where d is the damping factor.

This damping factor d shows that users will only continue clicking on links for a finite amount of time before they get distracted and start exploring something completely unrelated. With the remaining probability (1-d), the user will click on one of the cj links on page pj at random. Damping factor is usually set to 0.85. So it is easy to infer that every page distributes 85% of its original PageRank evenly among all pages to which it points.

Problems of PageRank Algorithm are:

- It is a static algorithm that, because of its cumulative scheme, popular pages tend to stay popular generally.
- Popularity of a site does not guarantee the desired information to the searcher so relevance factor also needs to be included.
- It should support personalized search that personal specifications should be met by the search result.

B. HITS (Hyper-link Induced Topic Search) Algorithm (IBM)

It is executed at query time, not at indexing time, with the associated hit on performance that accompanies querytime processing. Thus, the hub(going) and authority(coming) scores assigned to a page are queryspecific. It is not commonly used by search engines. It computes two scores per document, hub and authority, as opposed to a single score of PageRank. It is processed on a small subset of „relevant" documents, not all documents as was the case with PageRank. This algorithm was given by

Kleinberg in 1997. According to this algorithm first step is to collect the root set. That root set hits from the search engine. Then the next step is to construct the base set that includes the entire page that points to that root set. The size should be in between 1000-5000. Third step is to construct the focused graph that includes graph structure of the base set. It deletes the intrinsic link, (the link between the same domains). Then it iteratively computes the hub and authority scores.

Problems of HITS Algorithm are as follow: [5][6]

1 Mutually reinforced relationships between hosts. Sometimes a set of documents on one host point to a single document on a second host, or sometimes a single document on one host points to a set of document on a second host.
2. Automatically generated links. Web document generated by tools often have links that were inserted by the tool.
3. Non-relevant nodes. Sometimes pages point to other pages with no relevance to the query topic.

C. Weighted PageRank Algorithm

Wenpu Xing and Ali Ghorbani proposed a Weighted PageRank algorithm which is an extension of the PageRank algorithm. This algorithm assigns a larger rank values to the more important pages rather than dividing the rank value of page evenly among its outgoing linked pages, each outgoing link gets a value proportional to its importance. In this algorithm weight is assigned to both backlink and forward link. Incoming link is defined as number of link points to that particular page and outgoing link is defined as number of links goes out. This algorithm is more efficient than PageRank algorithm because it uses two parameters i.e. backlink and forward link. The popularity from the number of in links and outlinks is recorded as W_{in} and W_{out} respectively. $W_{in}(v, u)$ is the weight of link (v, u) calculated based on the number of in links of page u and the number of in links of all reference pages of page v. [7][8]

IV COMPARISION

Based on the literature analysis, a comparison of some of various web page ranking algorithms is shown in table 2. Comparison is done on the basis of some parameters such as main technique use, methodology, input parameter, relevancy, quality of results, importance and limitations.

Criteria	Algorithm		
	PageRank	**HITS**	**Weighted Page Rank**
Mining technique used	Web Strucutre Mining	Web Structure Mining, Web Content Mining	Web Structure Mining
Methodology	This algorithm computes the score for pages at the time of indexing of the pages.	It computes hubs and authority of the relevant pages.	Weight of web page is calculated on the basis of input and outgoing links and on the basis of weight the importance of page is decided.
Functionality	- Computes scores at index time. - Results are sorted on the importance of pages.	Computes scores of n highly relevant pages on the fly.	Computes weight of the web page at the time of indexing.
Accuracy	High	Middle	High
Input Parameter	Black links	Content, Back, forward links	Back links and Forward links

Importance	High. Back links are considered.	Moderate. Hubs and authorities scores are utilized.	High. The pages are sorted according to the importance.
Complexity	O(log N)	<O(log N)	O(log N)
Limitation	Results come at the time of indexing and not at the query time.	Topic drift and efficiency problem.	Relevancy is ignored.

V. CONCLUSION

In this paper it has been mentioned the introduction of web mining and its related categories such as web content mining, web structure mining and web usage mining and are also tabulated to provide comprision. The main goal of search engines is to provide relevant information to the users to cater to their needs based on the query they provide. Therefore, finding the relevant content of the Web and retrieving information according the user's interests and needs have become increasingly important. The different algorithms used for link analysis like PageRank (PR), Weighted PageRank (WPR), Hyperlink-Induced Topic Search (HITS) algorithms are discussed and compared depending on which the aim is to discover an efficient and better system for mining the web topology to identify relevant web pages.

REFERENCES

[1] Brin, and L. Page, The Anatomy of a Large Scale Hypertextual Web Search Engine,, Computer Network and ISDN Systems, Vol. 30, Issue 1-7, pp. 107-117, 1998.

[2] C. Ding, X. He, P. Husbands, H. Zha, and H Simon, Link analysis: Hubs and authorities on the world. Technical report: 47847, 2001.

[3] J. Hou and Y. Zhang, Effectively Finding Relevant Web Pages from Linkage Information, IEEE Transactions on Knowledge and Data Engineering, Vol. 15, No. 4, 2003.

[4] Zakaria Suliman Zubi, Marim Aboajela Emsaed. 2010. Sequence mining in DNA chips data for diagnosing cancer patients.In Proceedings of the 10th WSEAS international conference on Applied computer science (ACS'10), Hamido Fujita and Jun Sasaki (Eds.). World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, Wisconsin, USA, 139-151.

[5] Ashish Jain, Rajeev Sharma, Gireesh Dixit, Varsha Tomar ,"Page Ranking Algorithms in Web Mining, Limitations of Existing methods and a New Method for Indexing Web Pages", 2013 IEEE International Conference on Communication Systems and Network Technologies.

[6] Miguel Gomes da Costa, Júnior Zhiguo Gong," Web Structure Mining: An Introduction", proceedings of the 2005 IEEE International Conference on Information Acquisition June 27 - July 3,2005, Hong Kong and Macau, China

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

[8] Rashmi Rani, Vinod Jain ," Weighted PageRank using the Rank Improvement" International Journal of Scientific and Research Publications, Volume 3, Issue 7, July 2013.

[9] Zakaria Suliman Zubi, "Ranking WebPages Using Web Structure Mining Concepts" Proceeding of the 12th WSEAS international conference of Recent Advances in Telecommunications, Signals and Systems, March 2013.