International Journal of Pure and Applied Mathematics
Volume 119 No. 15 2018, 2763-2770
ISSN: 1314-3395 (on-line version)
url: http://www.acadpubl.eu/hub/
Special Issue

http://www.acadpubl.eu/hub/

Study of Preference Based Page Ranking Algorithm

[1]Sumit Kumar, [2]Amit Kumar Jakhar , [3]Yugal Kumar

[1] Department of Information Technology, KIET Group of Institutions, Muradnagar, Ghaziabad

[2-3] Department of Computer Science and Engineering, JayPee University of Information Technology, Waknaghat, Solan, India

sumitkumarbsr19@gmail.com, amitjakhar69@gmail.com, yugalkumar.14@gmail.com

Abstract

With massive explosion of digital data and many people's relying on more and more on search engines to get all kinds of information. Due to huge amount of data it has become steadily more difficult for the search engines to produce most relevant data links to the users. Therefore, producing the data on the web and abiding with the user interest or desires from their activities is very essential. Since, web mining is a utility of data mining techniques that is used to classify users and web pages by analyzing the user's behavior and content of the web pages. Web Structure mining plays a significant role in this field of data collection. In the past years, several algorithms have been developed to improve the performance is Weighted Page-Rank Algorithm that is an extension to the standard Page-Rank Algorithms. Therefore, the processing of such large amount of data on the internet as refined version of weighted page rank is considered over distributed systems using Hadoop Map-Reduce framework.

Keywords: Web Mining, Web Structure Mining, Weighted Page Rank, Page Rank, Hadoop MapReduce

1. Introduction

As digital information get popularity, then the retrieving information has become a bottleneck in information integration and sharing. Nowadays, the internet is imposing immense impact in every aspect of human life. Therefore, summarizing and analyzing user's patterns of behavior become highly important. If the web address, or the URL, of the site is known, then it is easier to access the information. But, if it is not available, then the page can be found by following the links from other pages or another way is to search for things using a search engine. The majority of the internet users utilize the information retrieval tools like: search engines to find the desired information of off the internet [7]. Several search engines are available for the retrieval purpose but the most commonly used search engines are Google, yahoo, Bing, etc. For searching the result of a given query of any user, a search engine usually scan its index of web pages for the content related to the search engine and the entire process is shown in Figure 1 and 2.

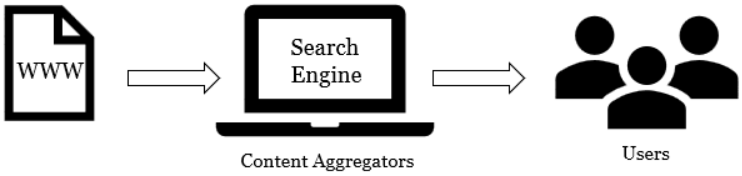

Figure1. Concept of a Search Engine

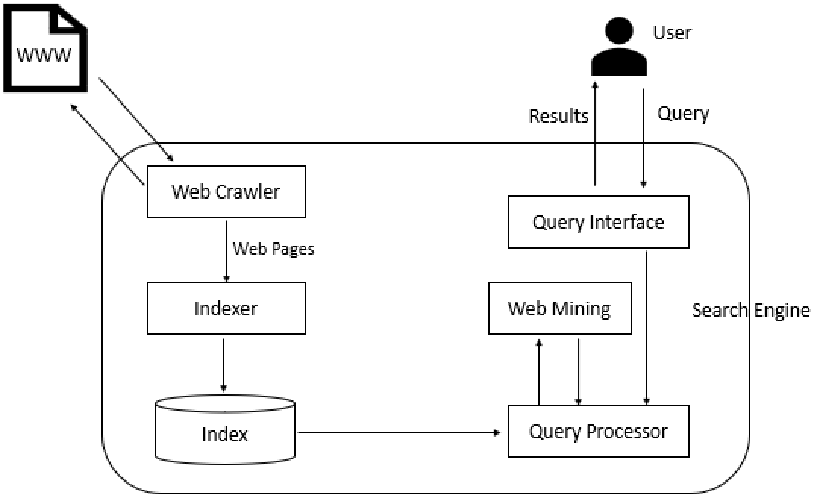

Figure2. Architecture of a Search Engine

2. Web Mining

The extraction of helpful information or patterns from massive databases is referred as Data Mining and the web mining discovers the content of the web on the basis of the past behavior of user and retrieves the useful information from World Wide Web [4, 11, 12]. The whole process can be classified into followings: Web Content Mining (WCM), Web Usage Mining (WUM), and Web Structure Mining (WSM).

- Web content mining: is used to extract knowledge from web [6]. It is somehow related but different from data and text mining. It is different in the context as it uses structured data and unstructured data respectively. Due to semi structured nature of web data, web content mining technique uses creative application of both data and text mining. The techniques used in this process are derived from information retrieval and natural language processing [4].
- Web structure mining: is a process of analyzing the structure of a website, i.e. nodes and connection structure through graph theory [4]. We can get two things from this website, first how the website is connected from other websites and the document structure, i.e. how the pages are connected.
- Web usage mining: it looks on the patterns and information from the server logs. It extracts the patterns and checks the activity of user from server. For example: the location of the user, selected items by users and the number of times [4]. Basically, it checks every activity of user. This is done in three steps namely; pre-processing, pattern discovery and pattern analysis.

These three categories have their own application areas such as Web personalization, site modification, usage classification and characterization, ranking of pages etc. The ranking of each page is used by the search engines to find the most useful pages.

Constraints on the Web Mining:

1. Complex structure of WWW: it contains a large amount of unstructured data that makes searching an extremely complex task [13].
2. Dynamic Nature of Internet
3. User Diversity
4. Relevant Information is time consuming

3. Existing Ranking Algorithms

In last decade, the huge growth of information over the internet identified thereafter page ranking is become an undoubtedly challenging task for better retrieval. So, the important goal of the search engines is to generate an interactive interface for all the internet users to retrieve relevant information in an ordered manner and within time bound. The retrieval information is based on some page ranking strategies and these are classified into two groups as the view of search engine. The first is the query dependent that rely upon the frequency of accessing page and the positioning in the terms of generated query. The second is query independent factors, which take into account the link status of the demand page. Among the entire existing page ranking techniques the Weighted Page Rank is the most appropriate to work with.

Yadav et al. [9] have indentified the Page level keywords that can be found in each standalone pages of any website on the web. The search engine generates the results by accessing the page level keyword that is an important factor for identifying the suitable page, which contains the desired keywords. The search engine responded with resulted set that may contain a large number of unwanted web pages against the query. The authors have shown that the retrieved result from the search engine which is relevant for a particular query increases the precision and accuracy while producing the desired pages. They also evaluate their model which is based on page level keywords on educational queries and get better result. Later, Jain et al. [10] described numerous algorithms that are used for link analysis like Page Rank (PR), Weighted Page Rank (WPR), Hyperlink-Induced Topic Search (HITS) and CLEVER and thereafter a comparative study on these algorithms is done by other researchers [1, 2].

4. Weighted Page Rank Algorithm

The algorithm was introduced by Wenpuet et al., in which some improvement is done in the original Page Rank Algorithm [7]. The idea of this algorithm is to use page links as votes count. It also simplifying page with more links is more important. The prominence of the webpage is decided by evaluating both in-links and out-links of the web pages. In contrast to the Page Rank algorithm, WPR does not evenly divide the rank of the page among in-links and out-links; rather it provides high rank value to the more popular pages on internet. The web page popularity is defined by assigning different values to incoming links and outgoing links, and the values is calculated by using the following equations 1 [6].

$$\mathbf{W}^{\mathbf{in}}_{(\mathbf{u},\mathbf{v})} = \frac{\mathbf{I_u}}{\sum_{\mathbf{p}\in\mathbf{R(v)}} \mathbf{I_p}} \tag{1}$$

where, I_u represents the number of in-links of a page u,

I_p represent the number of in-links of a page p and

$R_{(v)}$ represent the reference page list of page v

$$\mathbf{W}^{\mathbf{out}}_{(\mathbf{u},\mathbf{v})} = \frac{\mathbf{O_u}}{\sum_{\mathbf{p}\in\mathbf{R}(\mathbf{v})} \mathbf{O_p}} \tag{2}$$

where O_u represents the number of out-links of page u,

O_p represent the number of out-links of page p, and

$R_{(v)}$ represent the reference page list of page v.

The following equation 3 gives the Weighted Page Ranking value [6].

$$\mathbf{WPR(u) = (1-d) + d\sum_{v\in R(u)} WPR(v) W^{in}_{(v,u)} W^{out}_{(v,u)}} \tag{3}$$

5. Proposed Algorithm

A different algorithm namely Preference Based Page Rank Algorithm is proposed which employs Web Structure mining, Web Content mining and Web Usage mining techniques with the help of specialized crawlers and agents to rank web pages. User always has a vague idea of their preferences, but they cannot express them exactly (by any desired set of rules or equation). So, the preferences are obtained by the method of data mining. Software agents learn from the environment and accordingly react to it. It is efficient in ways of relevancy because of the fact that it considers relevancy and user behavior while web page ranking. The resulting in easier navigation and more satisfied information produced [8].

User based weighted page rank calculates the relevancy in two ways, probability weightage (probability of the queries) and content weightage (weight of the content with respect to the query), incorporating the user behavior by keeping in the account of how many times user has visited the in-links and out-links of any webpage while ranking.

The following equation of the proposed algorithm is given below:

$$\mathrm{PBR(u)} = (1-\mathrm{d}) + \mathrm{d}\sum_{\mathrm{v}\in\mathrm{R(u)}}(\mathrm{Vu} * \mathrm{W}^{\mathrm{in}}_{(\mathrm{v,u})} + \mathrm{W}^{\mathrm{out}}_{(\mathrm{v,u})})\mathrm{PBR(v)} * (\mathrm{C_w} + \mathrm{P_w})/\mathrm{T_v} \qquad 4)$$

PBR(u)= preference-based rank score of webpage u.

PBR(v)= preference-based rank score of webpage v.

$W^{in}_{(v,u)}$= weight of in-link(v,u)

$W^{out}_{(v,u)}$= weight of out-link(v,u)

Vu= number of visits of link pointing from v to u.

D= dampening factor (usually set between 0 and 1)

C_w= content weight of page u.

P_w = probability weight of page u

T_v = total number of visits of all links present on v,

R(u)= set of pages pointing to webpage u.

It calculates the relevancy in terms of Content Weight, Probability Weight and user's visit count. Content weight C_w is the ratio of frequency of query string(x) and sum of frequency of all possible significant query strings(z)

$$C_w = x/z \tag{5}$$

Probability weight, P_w is the ratio of query present in document(c) and total number of query terms ignoring stop words (d).

$$P_w = c/d \tag{6}$$

The working of the proposed algorithm has been depicted in the following steps and the Figure 3.

Obtain hyperlink structure of the web pages which are retrieved by the crawler.

1. Take the web graph of retrieved web pages into consideration.

2. Initialize equal rank to every web page i.e. one.
3. In-links and Out-links are to be computed by using the equation (1) and (2).
4. Content weight and Probability weight are to be computed using equation (5) and (6).
5. Using the values from step 4 and 5, to the equation of proposed algorithm i.e. equation (4).
6. The process must be repeated till stable page ranks are achieved i.e. consecutive pages shouldn't have same rank.

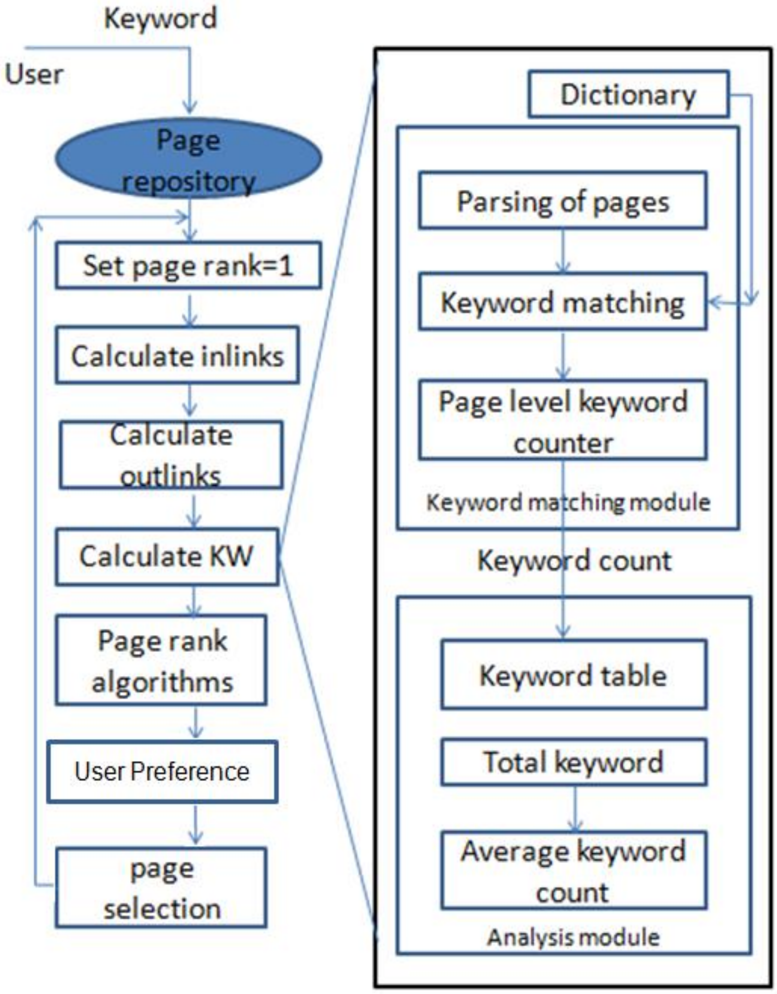

Figure3. Methodology of Preference Based Algorithm

6. Comparative Analysis

After analyzing the detailed technical study of two important ranking algorithms, it is concluded that each of them has some relative importance and limitations. It is also found that the weighted Page Rank algorithm mainly focuses on the hyperlink rather than the content in the web pages. Moreover, it computes rank at the time of indexing. Whereas, the Preference Based algorithm provides the ranking to the web pages is more target-oriented because it considers the fact that the trend of user's history or behavior is very significant. The preference based algorithm uses web structure, web usage and web content mining techniques to find out the relevancy of the web page as per the query mentioned.

Table I. Comparative study of web Page Ranking Algorithms

Algorithms	**Parameter(Input)**	**Mining Techniques**	**Complexity**	**Importance**	**Limitations**
Weighted Page Rank	Backward links and Forward links	Web Structure Mining	< O(log N)	Importance is evaluated by considering weights of in-links and out-links.	Relevancy is ignored, and the method computed score at a single level.
Preference Based Ranking	In-links, Out-links, agents and count of user visits	Web Structure, Content and Usage mining	< O(log N)	Disregard the similarity of ranking and become more target-oriented	Requires specialized crawler

7. Conclusion

In this work, a new and modified ranking algorithm is proposed by considering user based preference with handy considerations of the content of web page to efficiently rank the web pages. The comparison of the proposed algorithm is done with one of the page ranking algorithms, i.e. Weighted Page Rank Algorithm. The preference based algorithm makes search navigation easier and ranking algorithms which consider web structure mining are comparatively less relevant to user query as they have no idea about user trends on the topic. The ranking algorithm based on web content mining usually disregards the importance of the other webpage. On the other hand, our proposed algorithm tries to overcome the above-mentioned limitations by taking both users trends and content of webpage hand in hand. Moreover, it avoids similarity in ranking and it is also more dynamic in nature.

References

1. Neelam Duhan et al, “Page ranking algorithm: A survey” in IEEE International Advance Computing Conference, Patiala, 2009.
2. C. Deisy et al, “A Novel Relation-Based Probability Algorithm for PageRanking in Semantic Web Search Engine” presented at Department of Computer Science and Engineering, Thiagarajar College of Engineering, Madurai, TamilNadu, 2011.
3. AdityaPratap Singh et al, “An Efficient Algorithm for Ranking Research Papers Based on Citation Network” in 3rd Conference on Data Mining and Optimization (DMO), Selangor, Malaysia, 28-29 June,2011.
4. Lissa Rodrigues et al, “Hybrid Model for Improvised Page Ranking Algorithm” in International Conference on Control, Instrumentation, Communication and Computational Technologies (lCCICCT), Mumbai, Maharashtra, 2015.
5. Lissa Rodrigues et al, “An Efficient Page Ranking Approach Based on Hybrid Model” in Second International Conference on Advances in Computing and Communication Engineering, Mumbai, Maharashtra, 2015.
6. Dr. Daya Gupta et al, “User Preference Based Page Ranking Algorithm” in International Conference on Computing, Communication and Automation, Delhi, India, 2016.
7. W.Xing and A.Gorbani, “Weighted PageRank Algorithm”, Proceedings Of the Second Annual Conference on Communication Networks and Services Research, May 2004, pp.305-314.

8. 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), July 2012.
9. S. Goel, S. Yadav, “Search engine evaluation based on page levelkeyword”, Proceedings of the 2013 3rd IEEE International AdvanceComputing Conference, IEEE Computer Society (2013), pp. 870–876.
10. 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 International Conference on Communication Systems and Network Technologies.
11. R.Cooley, B.Mobasher and J.Srivastava, "Web Mining: Information and Pattern Discovery on the World Wide Web", Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97), 1997.
12. Companion slides for the text by Dr. M. H. Dunham, "Data Mining: Introductory and Advanced Topics", Prentice Hall, 2002.
13. J. Han, Kevin and C. Chang, "Data Mining for Web Intelligence", IEEE, 2002.