Volume 5, Issue 7, July 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Intelligent Web Information Retrieval using Enhanced Weighted Page Rank Algorithm

Geetika Malhotra*
Department of Computer Science
Bhagwan Parshuram Institute of Technology, New Delhi, India

Abstract— The World Wide Web is a useful information resource which is used for information retrieval and knowledge discovery. To carry out information retrieval efficiently search engines need to perform multiple tasks based on their architecture. Web structure mining plays an important role in extracting the most relevant information based on user's queries. In this paper we have proposed a new algorithm, the Enhanced Weighted Page Rank (EWPR) for page rank, taking into account that most relevant web page is retrieved as per the user's queries. This algorithm works efficiently and performs better than the Page Rank as a weight factor (WF) is used to retrieve the most relevant page and finally search results are improved as the list is re-ranked by updating existing page rank values of a page.

Keywords— Information Retrieval, World Wide Web, Web Mining, Weighted Page Rank, Web Content Mining

I. INTRODUCTION

The World Wide Web is a very popular and interactive information resource acting as a storehouse for text, image, audio, video, and metadata. The amount of information available on Web is very huge and is rapidly increasing and consists of dynamic unstructured data. The web pages do not have a unified structure and there are huge document libraries which are not arranged [1]. The user community on the web is rapidly expanding day by day. A particular user is interested only in some small portion of the web, other portion of the web contains information that is not relevant to the user. Therefore, the web has become difficult for users to extract and filter the information that is more relevant. Retrieving relevant information and providing it to users has become increasingly important. So ranking algorithms are used to sort the web pages so that more relevant results are displayed at the top [3].

Various ranking algorithms developed are Page Rank, Weighted Page rank, Page Content Rank, HITS, SALSA, SUBSPACE HITS, SIMRANK etc. Most of these algorithms are either based on web structure mining or web content mining. Web content mining extracts useful information from the content of web documents whereas web structure mining is used to set links between references and referents in the web [2]. In this paper, Enhanced Weighted Page Rank (EWPR) is being proposed for search engines that works on the basis of Weighted Page Rank algorithm and takes into account Weight factor (WF). The important purpose of this proposed algorithm is to find more relevant information as per the queries of the user [8].

This paper is organized as follows: In Section II, a description of web mining is given. Then in Section III, our proposed EWPR algorithm is described stepwise followed by experimental results in Section IV. Lastly, the conclusions are stated in Section V.

II. WEB MINING

Web mining is a data mining technique in which large number of web pages are crawled and knowledge is extracted from them. Web mining can be broadly classified into Web Content Mining (WCM), Web Structure Mining (WSM) and Web Usage Mining (WUM) [4]. WCM deals with the process of extracting useful information from web content which may consist of text, images, audio, video, lists or tables. Web Content Mining usually gives a filtered information to the users based on their inference and focuses on the structure within the web document. Web Structure Mining (WSM) is the process of discovering knowledge from web pages and also discovers the link structure of the hyperlinks between the documents. WUM ascertains user profiles and finds the user's behaviour stored in the web log profile [5].
The complete process of web mining is illustrated in Fig. 1.

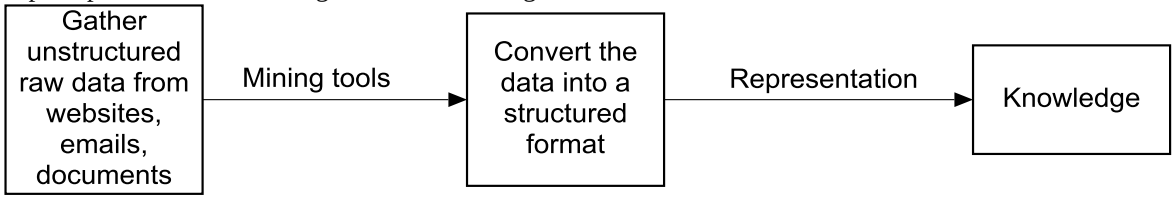

Fig. 1 Web mining process flow

The process of web mining can be described in the following steps:

1) *Finding a Resource:* The web documents should be discovered and retrieved from a relevant source.
2) *Pre- processing:* The data should be gathered from the retrieved web documents and then should be pre-processed to collect information from the documents [2]. Pre-processing involves steps like data cleaning, data integration.
3) *Generalization*: It is used to discover patterns from the web data at particular websites as well as multiple sites [9].
4) *Analysis*: In this step, the web documents are analysed and patterns are interpreted from them. This is an important step in the knowledge discovery process. As a result, knowledge is extracted.

III. ENHANCED WEIGHTED PAGE RANK ALGORITHM (EWPR)

Search engines use ranking algorithms to sort the results of search so that the first few results of search shown to the user are the most appropriate ones. The Weighted Page Rank algorithm distributes the rank of a web page among its various linked pages in accordance with their popularity. Generally it happens that the user does not get the required relevant documents easily on the top when the queries are searched. To resolve this problem, Enhanced Weighted Page Rank algorithm (EWPR) is proposed. The proposed algorithm will be an enhancement to the Weighted Page Rank algorithm as a new dimension namely Weight Factor (WF) is added for retrieving the most relevant page.

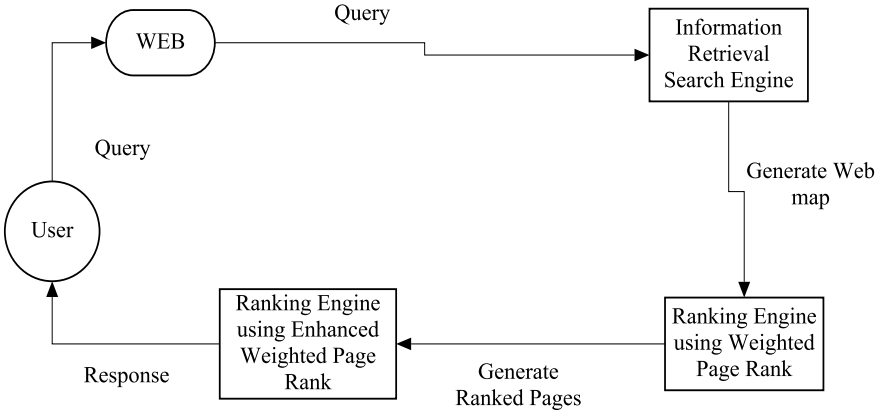

Fig. 2 Proposed Framework of EWPR

A. System Framework

The design of the system for implementing the algorithm is shown in Fig. 2. The phases of the system framework are described in the following steps:

1) *Web:* It is a system which has rich interlinked hypertext documents which can be accessed through internet. Web consists of text, images, videos and navigation between them is possible through hyperlinks. This is important as the Weighted Page Rank algorithm relies on the structure of web.
2) *Information Retrieval Search Engine:* Search engine is a website that collects and maintains content from all over the internet. The users can enter their queries to access the information they need to and the search engines provides the links that matches the contents of the user's query [5]. Hence information needed by the user can be retrieved and a web map is generated.
3) *Ranking Engine Using Weighted Page Rank:* This algorithm fully depends on the link structure of the web graph as it is used to calculate the importance of web page and how relevant it is. It determines how many pages are linked and pointed by a particular web page [7].
4) *Ranking engine using EWPR:* This is the most important step as it is used to determine how relevant a web page is by calculating the Weight Factor (WF) and adding it to the Weighted Page Rank output.

B. Implementation of Enhanced Weighted Page Rank Algorithm (EWPR)

A uniform method is proposed for efficient ranking of web pages according to user's queries. This method takes input from a query processor and the documents that match according to user's queries are considered. Then the rank score of the returned pages can be improved using EWPR algorithm. The algorithm is given in Fig. 3.

Algorithm 1: Steps of EWPR Algorithm

Input: Query text Q, Set of pages (Pi)

Output: New Pi

1. Find f(Pi)= { number of frequency of matched documents D in sequential pattern}
2. Find Weight Factor for every page $WF(Pi) = \sum f(Pi)$
3. New Rank (Pi) = Rank (Pi) + Weight Factor, where Rank (Pi) is the existing rank value (Page Rank).
4. Reorder and return new Pi.

Fig. 3 Shows the steps of EWPR algorithm

C. Relevancy Calculation

Weighted Page Rank algorithms provide information about a given query by using website structure. Relevancy calculation determines what relevant ranking a web page holds in response to a user's query. The pages shown in the results are categorized into four classes based on their relevancy [10].

1) *Very Relevant Pages (VR):* These pages contain very important information about the specified query.
2) *Relevant Pages (R):* These pages contain relevant but not important information about the specified query.
3) *Weak Relevant Pages (WR):* They contain the keywords of the query and no relevant information.
4) *Irrelevant pages (IR):* These pages neither contain keywords nor relevant information about a query [11].

How relevant the page is depends on its category and its position in the page list. The larger the relevancy, the better is the result. The Relevancy Z of a page is given by equation 1 as below:

$Z = \sum(n - i) * Wi$ (1), where i denotes the ith page in the result page list R(P), n represents the first n pages chosen from the list R(p), and Wi is the weight of the page i. The values of Wi are shown in Fig. 4.

Wi = V1, if the ith page is in PR
V2, if the ith page is in R
V3, if the ith page is in WR
V4, if the ith page is in IR
Where v1>v2>v3>v4

Fig. 4 Shows the values of weights of pages according to their category

IV. EXPERIMENTAL RESULTS

To evaluate the proposed algorithm, a user based approach is adopted [6]. User based approach is a very effective approach as it generates the most effective results. This approach emphasizes user's subjective perceptions of making most relevant decision. In this research, a survey is conducted with 100 students as sample is selected randomly from the management schools and distributed to them, the first 10 pages returned by Google using the query "Financial Planning" are evaluated and then the students are asked to score or grade each page (over 100) according to the relevance of the given query and then the average score of each page is calculated. From the survey, following results are obtained as shown in Table 1 and Table 2. The comparative analysis of EWPR with traditional method is shown in Table 3.

Table I Scores for the query "financial planning"

Pages as returned by Google	**Average Score returned by students**
Page 1	51
Page 2	91
Page 3	94
Page 4	85
Page 5	74
Page 6	62
Page 7	50
Page 8	36
Page 9	39
Page 10	26

Table II Result of the proposed algorithm

Pages as returned by Google	**Average Score returned by students**
Page 1	46
Page 2	91
Page 3	97
Page 4	87
Page 5	82
Page 6	71
Page 7	57
Page 8	43
Page 9	18
Page 10	13

Table III Google result v/s student score v/s ewpr

Pages as returned by Google	**Reorder pages returned by students score**	**EWPR**
Page 1	Page 3	Page 3
Page 2	Page 2	Page 2
Page 3	Page 4	Page 4
Page 4	Page 5	Page 5
Page 5	Page 6	Page 6
Page 6	Page 1	Page 7
Page 7	Page 7	Page 1
Page 8	Page 9	Page 8
Page 9	Page 8	Page 9
Page 10	Page 10	Page 10

V. CONCLUSION

Web mining is used to extract information from user's queries. As huge amount of information is present on the web, the users spend a lot of time to get the information that is most relevant to them. To find the relevant pages Weighted Page Rank Algorithm is used but mostly the results are not that proper. This paper presents an efficient algorithm where an additional factor called the Weight Factor (WF) is used which assigns a weight to the pages so that most relevant pages are retrieved. In this way, it helps users to get the relevant information about their queries quickly. Thus this algorithm is more accurate and efficient than traditional Weighted Page Rank.

REFERENCES

[1] R. Rani, V. Jain, " Weighted Page Rank using the Rank Improvement", *International Journal of Scientific and Research Publications*, vol. 3,issue 7,pp. 1-5, July 2013.

[2] S. Kadry, A. Kalakesh, "On the Improvement of Weighted Page Content Rank", *Journal of Advances in Computer Networks*, vol.1,no.2, pp. 110-114, June 2013.

[3] N. Tyagi, S. Sharma, "Weighted Page Rank Algorithm Based on Number of Visits of Links of Web Page", *International Journal of Soft Computing and Engineering*, vol. 2, issue 3 pp. 441-446, July 2012.

[4] S. Tuteja, "Enhancement in Weighted Page Rank Algorithm using VOL", *IOSR Journal of Computer Engineering*, vol. 14, issue 5, pp. 135-141, Sep- Oct 2013.

[5] N. Barsagade, "Web Usage Mining and pattern discovery : A survey" , CSE 8331, Dec 2003.

[6] L. Su, " A comprehensive and systematic model of user evaluation of web search engines: II an evaluation by undergraduates", *Journal of the American Society for Information Science and Technology*, vol. 54, no.13, pp. 1193-1223, 2003.

[7] B. Liu, "Web Data Mining Exploring Hyperlinks, Contents and Usage Data" *Springer Verlag* , Chapter 3, 2007.

[8] R. Kosala , H. Blockeel, "Web Mining Research: A Survey", *SIGKDD Explorations, Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining, vol.2,* no.1, pp. 1-15,Feb 2000.

[9] T. Haveliwala, "Topic sensitive page rank: a context-sensitive ranking algorithms fir web search", *IEEE Transactions on Knowledge and Data Engineering,* vol. 15, no.4, pp. 784-796, July-August 2003.

[10] A. Singh, R. Kumar, "Comparative Study of Page Ranking Algorithms for Information Retrieval", *International Journal of Electrical and Computer Engineering*, vol. 4, no. 7, pp. 469-481, 2009.

[11] P. Sharma, P.B.D. Tyagi , "Weighted page content rank for ordering web search result*", International Journal of Engineering Science and Technology,* vol. 2, no.12, pp. 7301-7310, 2010.

