On the Improvement of Weighted Page Content Rank

Seifedine Kadry and Ali Kalakech

***Abstract*—The World Wide Web has become one of the most useful information resource used for information retrievals and knowledge discoveries. However, Information on Web continues to expand in size and complexity. Making the retrieval of the required web page on the web, efficiently and effectively, is a challenge. Web structure mining plays an effective role in finding or extracting the relevant information. In this paper we proposed a new algorithm, the Simplified Weighted Page Content Rank (SWPCR) for page rank, based on combination of two classes of Web mining "Web structure mining" and "Web content mining". This algorithm will be an enhancement to the well-known Page Rank algorithm by adding to this algorithm a content weight factor (CWF) to retrieve more relevant page.**

***Index Terms*—Web mining, PageRank, weighted PageRank, simplified weighted page content rank.**

I. Introduction to Web Mining

With the dramatically explosive growth of the amount of information available over the internet, the World Wide Web has become a more useful environment to store, spread and retrieve information. The amount of information on the web is huge, diverse in meaning, dynamic, mostly unstructured data, different degree of quality of the information extracted and how much interest knowledge from information extracted. Therefore, the web has become more difficult for users to find extract, filter or evaluate the relevant information precisely and for content providers to catalog documents.

Web mining is the Data Mining technique [1] that automatically extracts the information from web documents and sorts them into identifiable patterns and relationships. Web mining can be easily executed with the help of other areas like Database (DB), Information retrieval (IR), Natural Language Processing (NLP) and Machine Learning.

II. Web Mining Process

The Web mining process is similar to the data mining process. The difference [2] usually lies in the data collection. In traditional data mining, the data is often already collected and stored in a database. For Web mining, data collection can be a fundamental task, especially for Web structure mining and Web content mining, which implies crawling a large number of Web pages. The complete process of extracting knowledge from Web data is illustrated Fig. 1:

Manuscript received December 1, 2012; revised February 8, 2013. This work was supported by the American University of the Middle East, School of Engineering and Technology.

S. Kadry is with the Engineering and Technology school, American University of the Middle East, Kuwait (e-mail: skadry@gmail.com).

A. Kalakech was with the Arts, Sciences and Technology University, Lebanon (e-mail: alikalakech@hotmail.com).

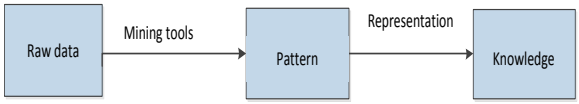

Fig. 1. Web mining process

It consists of following tasks:

1) Resource finding: the function of retrieving relevant web documents.
2) Information selection and pre-processing: the automatic selection and preprocessing of specific information from retrieved web resources. This process transforms the original retrieved data into information.
3) Generalization: It automatically discovers general patterns at individual web sites as well as across multiple sites. Data Mining techniques and machine Learning are used in generalization.
4) Analysis: the validation and interpretation of the mined patterns. It plays an important role in pattern mining. A human plays an important role in information on knowledge discovery process on web.

III. Web Mining Taxonomy

Web mining can be broadly divided into three distinct categories [3]-[4], according to the types of data to be mined, Web Content Mining, Web Structure Mining and Web Usage Mining as shown in Fig. 2. These will be explained in the following subsections.

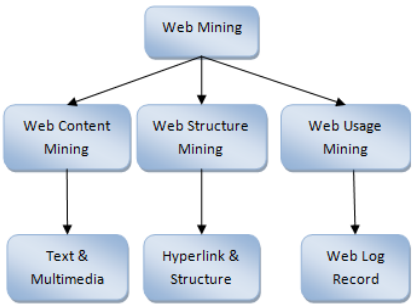

Fig. 2. Web mining categories

A. Web Content Mining (WCM)

Web Content Mining is the process of extracting useful information from the contents of Web documents. Content of Web documents may consist of text, images, audio, video, or structured records such as lists and tables. Web content mining is related to data mining because many data mining techniques can be applied in web content mining. It is also related to text mining because much of web contents are text based. However, it is also different from these because web data is semi structured in nature and text mining focuses on unstructured text.

Web content mining can be viewed from two different points of view: IR (Information Retrieval) and DB (Database)

DOI: 10.7763/JACN.2013.V1.23

views. The goal of Web content mining from the IR view is mainly to assist or to improve the information finding or filtering the information to the users usually based on either inference or seek user profiles, while the goal of Web content mining from the DB view mainly tries to model the data on the Web and to integrate them so that more advanced queries other than the keywords based search could be performed.

B. Web Usage Mining (WUM)

Since Web is a reaction media between Web users and Web pages, user navigational behavior needs to be fully concerned during Web mining. Web usage mining, is able to capture , analysis and model the interaction between users and pages during browsing, in turn, providing complementary assistance for advanced Web applications, such as adaptive Web design and Web recommendation. Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data, in order to understand and better serve the needs of Web-based application.

A Web server log is an important source for performing Web Usage Mining because it explicitly records the browsing behavior of site visitors. The data recorded in server logs reflects the access of a Web site by multiple users.

C. Web Structure Mining (WSM)

Web Structure Mining is the process of inferring knowledge from the World Wide Web organization and links between references and referents in the Web. The structure of a typical web graph consists of web pages as nodes and hyperlinks as edges connecting related pages. Web Structure mining is the process of using graph theory to analyze the node and connection structure of a web site. It is used to discover structure information from the web and it can be divided into two kinds based on the kind of structure information used. They are Hyperlinks and Document Structure.

IV. Link Analysis Algorithms

The analysis of hyperlinks and the graph structure of the Web have been helpful in the development of web search Link analysis is one of many factors considered by web search engines in computing a composite score for a web page on any given query. Many researchers suggested solutions to the problem of searching or querying the Web, taking into account its structure as well as the meta-information included in the hyperlinks and the text surrounding them. There are three important algorithms proposed based on link analysis: PageRank [5]-[7], Weighted PageRank (WPR) [8]-[10] and Hypertext Induced Topic Search (HITS) [10]. We will discuss the WPR algorithm because our proposed improvement is based on it.

A. Weighted PageRank

Weighted PageRank algorithm is an extension of the PageRank algorithm proposed by Xing and [8]. This algorithm assigns a larger rank values to the more important pages rather than dividing the rank value of a page equally among its outgoing linked pages. Each outgoing link gets a value proportional to its importance. The importance is assigned in terms of weight values to the incoming and outgoing links and is denoted:

$W^{in}_{(m,n)}$: is the weight of link(m, n) calculated based on the number of incoming links of page n and the number of incoming links of all reference pages of page.

$$W^{in}_{(m,n)} = \frac{I_n}{\sum_{p \in R(m)} I_P} \quad (1)$$

I_n is number of incoming links of page n, I_p is number of incoming links of page *p*, *R(m)* is the reference page list of page m.

$W^{out}_{(m,n)}$: is the weight of link (m, n). It is calculated on the basis of the number of outgoing links of page n and the number of outgoing links of all the reference pages of page m.

$$W^{out}_{(m,n)} = \frac{O_n}{\sum_{p \in R(m)} O_p} \quad (2)$$

O_n is number of outgoing links of page n, O_p is number of outgoing links of page p, and then the weighted PageRank is given by:

$$WPR(n) = (1-d) + d \sum_{m \in B(n)} WPR(m) W^{in}(m,n) W^{out}(m,n) \quad (3)$$

B. PageRank VS Weighted PageRank

In order to compare the WPR with the PageRank, the resultant pages of a query are categorized into four categories based on the relevancy to the given query.

- Very Relevant Pages (VR): These are the pages that contain very important information related to a given query.
- Relevant Pages (R): These Pages are relevant but not having important information about a given query.
- Weakly Relevant Pages (WR): These Pages may have the query keywords but they do not have the relevant information.
- Irrelevant Pages (IR): These Pages are not having any relevant information and query keywords.

The PageRank and WPR algorithms both provide ranked pages in the sorting order to users based on the given query. So, in the resultant list, the number of relevant pages and their order are very important for users. Relevance Rule is used to calculate the relevancy value of each page in the list of pages. That makes WPR different from PageRank.

Relevancy Rule: The Relevancy of a page to a given query depends on its category and its position in the page-list. The larger the relevancy value, the better is the result.

$$K = \sum_{k \in R(p)} (n-i) \times W_i \quad (4)$$

where *i* denotes the i^{th} page in the result page-list *R*(*p*), n represents the first n pages chosen from the list R(p), and W_i is the weight of i^{th} page as given in (5).

$$W_i = (v1, v2, v3, v4) \quad (5)$$

where *v*1, *v*2, *v*3 and *v*4 are the values assigned to a page if the page is VR, R, WR and IR respectively. The values are always *v*1>=*v*2>=*v*3>=*v*4. Experimental studies show that WPR produces larger relevancy values than the PageRank.

V. Simplified Weighted Page Content Rank Algorithm (SWPCR)

The World Wide Web has become a new communication medium with informational, cultural, social and evidential values after a few decades since its inception. Search engines are widely used for Web information access and they are making more information easily accessible than ever before. For example Google Web search receive 34,000 queries per second (2 million per minute; 121 million per hour; 3 billion per day; 88 billion per month) for most queries, there exist thousands of documents containing some or all of the terms in the query. A search engine Google needs to rank them using PageRank algorithm so that the first few results shown to the user must be the ones that are most pertinent to the user's need but the users may not get the required relevant documents easily on the top few pages. To resolve the problems found in PageRank algorithm, Simplified Weighted Page Content Rank is a new algorithm for page rank based on combination of two classes of Web mining "Web structure mining" and " Web content mining", The proposed algorithm will be an enhancement to the well-known web structure mining algorithm Page Rank which is used by the most famous search engine Google .By adding to this algorithm a content weight factor (CWF) to retrieve more relevant page.

A. System Design

Search engines are the key to finding specific information on the vast expanse of the World Wide Web .with our proposed SWPCR algorithm the search engine system is modified in order to add more components for calculating the importance and relevancy of pages. The modified system is displayed in Fig. 3.

The various components and search process are explained below to have an understanding of the existing as well as modified system.

- Web: is a system of interlinked hypertext documents accessed via the Internet. Web that may contain text, images, video, and other multimedia and navigates between them using hyperlinks.
- IR search engine: is the practical application of information retrieval techniques to large scale text collections. Search engine is a web site that collects and organizes content from all over the internet. Those wishing to locate something would enter a query about what they would like to find and the engine provides links to content that matches what they want.
- Ranking engine using WPR: is used to calculate the importance of the page, how many pages are pointing to or are referred by this particular page.
- Ranking engine using SWPCR: is used to calculate the importance and relevance of page by calculated content weight factor then combine the output of WPR with the output of CWF.

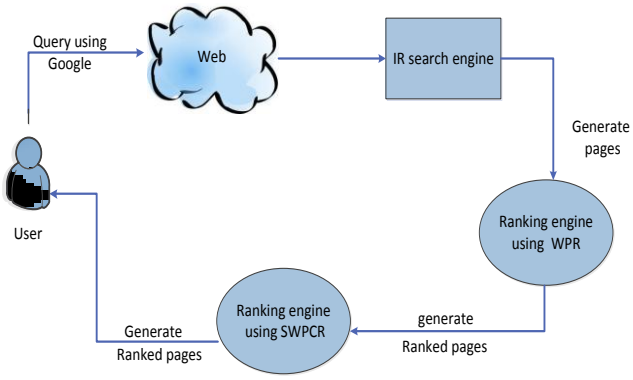

Fig. 3. Modified system architecture

B. Implementation

Algorithm SWPCR

Input:

- Query text Q
- Set of pages {Pi} ← Google (Q)

Output:

- New (Pi)

Relevance calculation

- Find f(Pi) = {number of frequency of logical combination of Q}
- Find content weight factor CWF(Pi)= GPA(f(Pi))
- Reorder and return the new {Pi}

C. Relevance Calculation

Relevance calculation concerns the problem how to determine a relevance ranking of web pages with respect to a given query. For this problem there are many different proposals to measure the relevance of a page, the most important of these features are matching functions which determine the term similarity to the query. Some of these matching functions depend only on the frequency of occurrence of query terms; others depend on the page structure, term positions, graphical layout, etc. but no consensus has been reached yet on the best way to calculate the relevance ranking of web pages with respect to a given query. The proposed algorithm SWPCR design new methods to calculate the relevance of a page based on two factors:

1) Find *f(Pi):* the frequency of logical combination of query text, the number of times that term appears in page *Pi*.
2) Find content weight factor *CWF(Pi)) = GPA (f(Pi))* that is consider the core of SWPCR proposed algorithm based on :

Given a matrix with $m \times n$

- n = number of words in a given query, each column contains the frequency of n words f(n) in each of the given pages
- m = number of pages
- sort the rows of the following array:

P1	f(n)	f(n-1)	f(n-2)	…….	f(1)
P2	f(n)	f(n-1)	f(n-2)	…….	f(1)
P3	f(n)	f(n-1)	f(n-2)	…….	f(1)
…	f(n)	f(n-1)	f(n-2)	…….	f(1)
Pm	f(n)	f(n-1)	f(n-2)	…….	f(1)

D. Simulation

In order to evaluate the proposed algorithm we adopted the user-based approach [9]-[11]. Many researchers adopted the user-based approach to develop effectiveness evaluation of relevancy search engine because the system-based approach ignores user's Perception, needs, and searching behavior in real-life situations. While user-based approach emphasizes user's subjective perceptions of relevance judgment. In this research, we conducted survey with 100 students as sample selected randomly from the engineering schools and distribute to them the 10 first pages returned by Google using the query "Advanced Software Engineering", then we ask the students to score or grade each page (over 100) according to the relevancy of our given query and then we calculate the average score of each page. From the survey we obtain the following results:

TABLE I: SCORES FOR THE QUERY "ADVANCED SOFTWARE ENGINEERING"

Pages as returned by Google	Average Score returned by students
page1	50
Page2	90
Page3	93
Page4	84
Page5	73
Page6	61
Page7	49
Page8	35
Page9	38
Page10	25

The result obtained by our proposed algorithm is given by the following:

TABLE II: RESULT OF THE PROPOSED ALGORITHM

Pages as returned by Google	Score returned by SWPCR GPA (f(Pi))
Page1	45
Page2	90
Page3	96
Page4	86
Page5	81
Page6	70
Page7	56
Page8	42
Page9	17
Page10	12

We compare the new order resulting from the evaluation with the result of our SWPCR algorithm

TABLE III: GOOGLE RESULT V/S STUDENT SCORE V/S SWPCR

Pages as returned by Google	Reorder Pages Returned by students score	SWPCR
Page1	**Page3**	**Page3**
Page2	**Page2**	**Page2**
Page3	**Page4**	**Page4**
Page4	**Page5**	**Page5**
Page5	**Page6**	**Page6**
Page6	**Page1**	**Page7**
Page7	**Page7**	**Page1**
Page8	**Page9**	**Page8**
Page9	**Page8**	**Page9**
Page10	**Page10**	**Page10**

The simulation indicates that relevant pages determined by SWPCR are more relevant to the students than returned by Google.

VI. CONCLUSION AND FUTURE WORKS

Web mining is used to discover the content of the Web, the users' behavior in the past, and the Webpages that the users want to view in the future. Web mining consists of Web Content Mining (WCM), Web Structure Mining (WSM), and Web Usage Mining (WUM). Web structure mining plays an effective role in finding the relevant information. Three commonly used algorithms in web structure mining are HITS, PageRank and Weighted PageRank, which are used to rank the relevant pages.

Several algorithms have been developed to improve the performance of these PageRank algorithms. This thesis introduces the SWPCR algorithm, that is enhancement to the well-known algorithm "Weighted Page Rank" which is used by the most famous search engine Google by adding to this algorithm a content weight factor (CWF) to retrieve more relevant page . The survey studies using the query "Advanced software engineering" show that SWPCR is able to identify a larger number of relevant pages to a given query compared to Weighted PageRank.

Research continues to improve Page Rank algorithm and the relevance features f search engine. This research has led to a continuous improvement of search engine relevancy. Considerable research is centered today on discovering new types of features which can noticeably improve search quality. Only three of the most promising areas are mentioned here:

- Synonyms Dictionary: The classical retrieval models are based on term matching, matching terms in the user query with those in the documents. However, many concepts can be described in multiple ways (using Different words) due to the context and people's language customs. If a user query uses different words from the words used in a document, the document will not be retrieved although it may be relevant because the document uses some synonyms of the words in the user query. This causes low recall. For example, "picture", "image" and "photo" are synonyms in the context of digital cameras. If the user query only has the word "picture", relevant documents that contain "image" or "photo" but not "picture" will not be retrieved, thus this research proposed new method to find the synonyms problem by using Synonyms Dictionary that enhanced the relevancy of search engine result by retrieving the documents matching query terms and its synonym.
- Understanding the user's intent of query: There are many different types of query may require very different types of relevance ranking algorithms. For example, a technology query may require very different types of analysis from a health query. Work on algorithms that understand the intent of a query and select different relevance ranking methods accordingly could lead to significantly increases in the quality of the ranking.
- Personalized Web Search: it is possible to utilize user information to "personalize" web search engine results.

For example. There are many different ways to personalize results: with respect to the user search history, user community, questionnaires or external sources of knowledge about the user, etc. Many scientific papers have been written on Personalized Web Search, but the problem remains unsolved. Web search engines have mainly declined to away from personalized algorithms. Google has proposed several forms of personalized search to its users, but this feature has not had much success. Nevertheless, the search continues for the right way to personalize relevance ranking.

REFERENCES

[1] B. Liu, *Web Data Mining Exploring Hyperlinks, Contents, and Usage Data*, New York: Springer-Verlag, 2007, ch. 3.

[2] A. Scime, *Web mining: Applications and Techniques*, London: Idea Group Publishing, 2005, ch. 5.

[3] G. Xu, Y. Zhang, and L. Li, *Web Mining and Social Networking. Techniques and Applications*, Australia: Springer-Verlag, 2011, ch. 2.

[4] R. Kosala and 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.

[5] R. Jain and G. N. Purohit, "Page ranking algorithms for web mining," *International Journal of Computer Application*, vol. 13, pp. 22-25, Jan 2011.

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

[7] A. Singh and 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.

[8] P. Sharma and 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.

[9] 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.

[10] W. Xing and A. Ghorbani, "Weighted pagerank algorithm," in *Proc. of the Second Annual Conf. on Communication Networks and Services Research (CNSR '04),* IEEE, pp. 305 - 314 , 2004.

[11] Y. Chuang and L. Wu, "User-based evaluations of search engines: hygiene factors and motivation factors," in *Proc. of the 40th Hawaii International Conf. on System Sciences*, pp. 82, 2007.

Ali Kalakech is an associate professor at the Information Systems Department in the Lebanese University, Faculty of Economics and Business Administration. He got his Master Degree in Computer Systems from the National Institute of Applied Sciences, Toulouse, France in 2001. He received the Doctor degree from the National Polytechnic Institute, Toulouse, France in 2005. His Research interests include testing and evaluating the dependability and the performance of computer systems.

Seifedine Kadry is an associate professor at the American university of the Middle East, Faculty of Engineering. He got his Master Degree in Computer Science and Applied Math from AUF-EPFL-Inria, Lebanon in 2002. He received the Doctor degree from the Clermont Ferrand II University, France in 2007. His Research interests include software testing and security.