Link Analysis Algorithms For Web Mining

Tamanna Bhatia
Dept. of Computer Science, Desh Bhagat Engineering College, Mandi Gobindgarh, Punjab, India

Abstract
with the explosive growth of Web, the primary goal of website owner is to provide relevant information to the users to fulfill their needs. Hyperlink analysis was important methodology used by famous search engine Google to rank the pages. There are ranking algorithms Page Rank, Weighted Page Rank and Weighted Page Content Rank. Weighted Page Rank also takes the importance of the inlinks and outlinks of the pages but the rank score to all links unequally distributed as compare to Page Rank. In this paper we give description about Weighted Page Content Rank (WPCR) based on web content mining and structure mining that shows the relevancy of the pages to a given query is better determined, as compared to the Page Rank and Weighted Page Rank algorithms and also providing the difference between Page Rank, Weighted Page Rank and Weighted Page Content Rank.

Keywords
Web mining, web content, Page rank, Weighted Page rank and weighted page content rank, and web structure.

I. Introduction

The World Wide Web is a rich source of information and continues to expand in size and complexity. Retrieving of the required web page on the web, efficiently and effectively, is becoming a challenge. Whenever a user wants to search the relevant pages, he/she prefers those relevant pages to be at hand. The bulk amount of information becomes very difficult for the users to find, extract, filter or evaluate the relevant information. This issue raises the necessity of some technique that can solve these challenges.
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 etc. The following challenges [1] in Web Mining are:
1) Web is huge.
2) Web pages are semi structured.
3) Web information stands to be diversity in meaning.
4) Degree of quality of the information extracted.
5) Conclusion of knowledge from information extracted.

This paper is organized as follows- Web Mining is introduced in Section II. The areas of Web Mining i.e. Web Content Mining, Web Structure Mining and Web Usage Mining are discussed in Section III. Section IV describes the various Link analysis algorithms. Section IV (A) defines Page Rank, IV (B) defines Weighted Page Rank and IV(C) defines Weighted Page Content Rank Algorithm. Section V provides the comparison of various Link Analysis Algorithms.

II. Web Mining

Web mining is the Data Mining technique that automatically discovers or extracts the information from web documents. It is the extraction of interesting and potentially useful patterns and implicit information from artifacts or activity related to the World Wide Web.

A. Web Mining Process

The complete process of extracting knowledge from Web data [2] is follows in Fig.1:

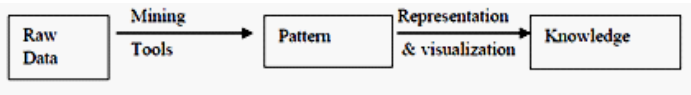

Fig. 1: Web Mining Process
The various steps are explained as follows.

1. Resource finding: It is the task of retrieving intended web documents.
2. Information selection and pre-processing: Automatically selecting and pre- processing specific from information retrieved Web resources.
3. Generalization: Automatically discovers general patterns at individual Web site as well as multiple sites.
4. Analysis: Validation and interpretation of the mined patterns.

III. Web Mining Categories

Web mining research overlaps substantially with other areas, including data mining, text mining, information retrieval, and Web retrieval. The classification is based on two aspects: the purpose and the data sources. Retrieval research focuses on retrieving relevant, existing data or documents from a large database or document repository, while mining research focuses on discovering new information or knowledge in the data. On the basis of this, Web mining can be classified into web structure mining, web content mining, and web usage mining as shown in Fig. 2.

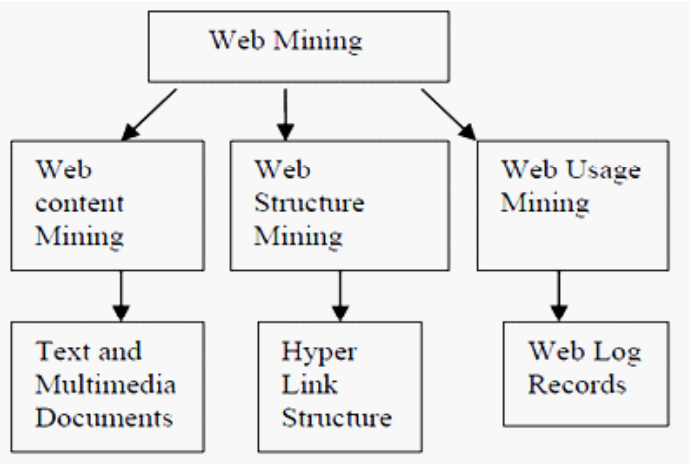

Fig. 2: Web Mining Categories

A. Web Content Mining

Web Content Mining [4-6] is the process of extracting useful information from the contents of Web documents. Content data corresponds to the collection of facts a Web page was designed to convey to the users. Web content mining is related but is different from data mining and text mining. It is related to data mining. It is related to text mining because much of web contents are text based. It is different from data mining because web data are mainly semi-structured and or unstructured. Web content mining is also different from text mining because of the semi-structure nature of the web, while text mining focuses on unstructured texts. The technologies that are normally used in web content mining are NLP (Natural language processing) and IR (Information retrieval).

B. Web Structure Mining

It is the process by which we discover the model of link structure of the web pages. The goal of Web Structure Mining [1,3] is to generate structured summary about the website and web page. It tries to discover the link structure of hyper links at inter document level. The other kind of the web structure mining is mining the document structure. It is using the tree-like structure to analyze and describe the HTML (Hyper Text Markup Language) or XML (Extensible Markup Language).

C. Web Usage Mining

Web Usage Mining [1,3] is the process by which we identify the browsing patterns by analysing the navigational behaviour of user. It focuses on techniques that can be used to predict the user behaviour while the user interacts with the web. It uses the secondary data on the web. This activity involves the automatic discovery of user access patterns from one or more web servers. Through this mining technique we can ascertain what users are looking for on Internet. It consists of three phases, namely pre-processing, pattern discovery, and pattern analysis. Web servers, proxies, and client applications can quite easily capture data about Web usage.

IV. Link Analysis Algorithms

Web mining technique provides the additional information through hyperlinks where different documents are connected. We can view the web as a directed labelled graph whose nodes are the documents or pages and edges are the hyperlinks between them. This directed graph structure is known as web graph. There are number of algorithms proposed based on link analysis. Three important algorithms Page Rank, Weighted Page Rank and Weighted Page Content Rank are discussed below:

A. Page Rank

Page Rank is a numeric value that represents how important a page is on the web. Page Rank is the Google's method of measuring a page's "importance." When all other factors such as Title tag and keywords are taken into account, Google uses Page Rank to adjust results so that more "important" pages move up in the results page of a user's search result display. Google Fig.s that when a page links to another page, it is effectively casting a vote for the other page. Google calculates a page's importance from the votes cast for it. How important each vote is taken into account when a page's Page Rank is calculated. It matters because it is one of the factors that determine a page's ranking in the search results. It isn't the only factor that Google uses to rank pages, but it is an important one. The order of ranking in Google works like this:

Find all pages matching the keywords of the search.

Adjust the results by Page Rank scores.

The algorithm of Page Rank [7] as follows:

Page Rank takes the back links into account and propagates the ranking through links. A page has a higher rank, if the sum of the ranks of its backlinks is high. Fig. 3 shows an example of back links wherein page A is a backlink of page B and page C while page B and page C are backlinks of page D.

The original Page Rank algorithm is given in following equation

$$PR(P)=(1-d)+d(PR(T1)/C(T1)+.....PR(Tn)/C(Tn)) \quad .... (1)$$

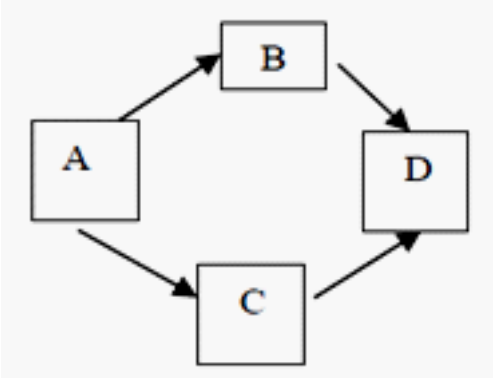

Fig. 3 : Example of Backlinks

Where, PR (P)= PageRank of page P

PR (Ti) = PageRank of page Ti which link to page

C (Ti) =Number of outbound links on page T

D = Damping factor which can be set between 0 and 1.

B. Weighted Page Rank

Extended Page Rank algorithm- Weighted Page Rank assigns large rank value to more important pages instead of dividing the rank value of a page evenly among its outlink pages. The importance is assigned in terms of weight values to incoming and outgoing links denoted as and respectively. is calculated on the basis of number of incoming links to page n and the number of incoming links to all reference pages of page m. (2)

In is number of incoming links of page n, Ip is number of incoming links of page p, R(m) is the reference page list of page m. 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. (3)

On is number of outgoing links of page n, Op is number of outgoing links of page p, Then the weighted Page Rank is given by following formula

$$WPR(n)=(1-d)+d \quad (4)$$

C. Weighted Page Content Rank

Weighted Page Content Rank Algorithm (WPCR) is a proposed page ranking algorithm which is used to give a sorted order to the web pages returned by a search engine in response to a user query. WPCR is a numerical value based on which the web pages are given an order. This algorithm employs web structure mining as well as web content mining techniques. Web structure mining is used to calculate the importance of the page and web content mining is used to find how much relevant a page is? Importance here means the popularity of the page i.e. how many pages are pointing to or are referred by this particular page. It can be calculated based on the number of inlinks and outlinks of the page. Relevancy means matching of the page with the fired query. If a page is maximally matched to the query, that becomes more relevant.

Algorithm: WPCR calculator

Input: Page P, Inlink and Outlink Weights of all backlinks of P, Query Q, d (damping factor).

Output: Rank score

Step 1: Relevance calculation:

a) Find all meaningful word strings of Q (say N)

b) Find whether the N strings are occurring in P or not? Z= Sum of frequencies of all N strings.

c) S= Set of the maximum possible strings occurring in P.

d) X= Sum of frequencies of strings in S.

e) Content Weight (CW)= X/Z
f) C= No. of query terms in P
g) D= No. of all query terms of Q while ignoring stop words.
h) Probability Weight (PW)= C/D
Step 2: Rank calculation:
a) Find all backlinks of P (say set B).
b)PR(P)=(1-d)+d[
c) Output PR(P) i.e. the Rank score

V. Comparison of Algorithms

Table1 shows the difference between above three algorithms:

Table 1: Comparison of Page Rank, Weighted Page Rank and Weighted Page Content Rank

Contents	Comparison		
	Page Rank	Weighted Page Rank	Weighted Page Content Rank
Mining Technique Used	WSM	WSM	WSM and WCM
Complexity	O(logn)	<O(logn)	<O(logn)
Working Procedure	Computes Scores at index time. Results are sorted on the importance of pages	Assigns large value to more important pages instead of diving the rank value of a page evenly among its outlink pages.	Gives sorted order to the web pages returned by a search engine as a numerical value in response to a user query.
Input/ output parameters	Backlinks	Backlinks and forward links	Backlinks, forward links and content
Advantages	It provide important information about given query by diving rank value equally among its outlink pages	It provide important information about given query and assigning importance in terms of weight values to incoming and outgoing links	It provide important information and relevancy about a given query by using web structure and web content mining
Search Engine	Google	Google	Research Model
Limitations	(1) Page Rank is equally distributed to outgoing links (2) It is purely based on the number of inlinks and outlinks.	(1)While some pages may be irrelevant to a given query, it still receives the highest rank (2) There is a less determination of the relevancy of the pages to a given query	No limitation best as comparison to Page Rank and Weighted Page Rank

VI. Conclusion

Web mining is the Data Mining technique that automatically discovers or extracts the information from web documents. Page Rank and Weighted Page Rank algorithms are used in Web Structure Mining to rank the relevant pages. In this paper we focussed that by using Page Rank and Weighted Page Rank algorithms users may not get the required relevant documents easily, but in new algorithm Weighted Page Content Rank user can get relevant and important pages easily as it employs web structure mining and web content mining. The input parameters used in Page Rank are Backlinks, Weighted Page Rank uses Backlinks and Forward Links as Input Parameter and Weighted Page Content Rank uses Backlinks, Forward Link and Content as Input Parameters.
As part of our future work, we are planning to carry out performance analysis of Weighted Page Content and working on finding required relevant and important pages more easily and fastly.

References

[1] Rekha Jain, Dr G.N.Purohit, "Page Ranking Algorithms for Web Mining," International Journal of Computer application,Vol 13, Jan 2011.
[2] Cooley, R, Mobasher, B., Srivastava, J."Web Mining: Information and pattern discovery on the World Wide Web". In proceedings of the 9th IEEE International Conference on tools with Artificial Intelligence (ICTAI' 97).Newposrt Beach,CA 1997.
[3] Pooja Sharma, Pawan Bhadana, "Weighted Page Content Rank For Ordering Web Search Result", International Journal of Engineering Science and Technology, Vol 2, 2010.
[4] R. Kosala, H. Blockeel "Web mining research" A survey. ACM Sigkdd Explorations,2(1):1-15, 2000.
[5] Wang jicheng, Huang Yuan,Wu Gangshan, Zhang Fuyan, "Web mining: Knowledge discovery on the Web Systems", Man and Cybernetics 1999 IEEE SMC 99 conference Proceedings. 1999 IEEE International conference
[6] Raymond Kosala, Hendrik Blockee, "Web Mining Research : A Survey", ACM Sigkdd Explorations Newsletter, June 2000, Volume 2.
[7] Taher H. Haveliwala, "Topic-Sensitive Page Rank: A Context-Sensitive Ranking Algorithms for Web Search", IEEE transactions on Knowledge and Data Engineering Vol.15, No 4 July/August 2003.

ISSN : 2229-4333(Print) | ISSN : 0976-8491(Online)

Tamanna Bhatia received her B.tech degree in Computer Science and Engineering from Bhai Gurdas Institute of Engineering and Technology, Sangrur ,affilated to Punjab Technical University, Jalandhar in 2010 and M.tech(Pursuing) in Computer Science and Engineering from Guru Nanank Dev Engineering College, Ludhiana, affilated to Punjab Technical, University Jalandhar. At present,She is working as an assistant professor, with Department of Computer Science and Engineering, in Desh Bhagat Engineering College,Mandi Gobindgarh, Punjab Technical University, Jalandhar in 2010 respectively. Her research interests include Web Mining, algorithms related to Web Structure Mining, Data Mining.