ISSN: 2277-3754
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 1, Issue 2, February 2012

Analysis of Link Algorithms for Web Mining

T.Munibalaji, C.Balamurugan
t.munibalaji@gmail.com, bmurugan.c@gmail.com

Abstract- In present scenario web mining is the most active area where the research is going on rapidly. Web mining is the integration of information gathered by traditional data mining methodologies and techniques with information gathered over the World Wide Web. (Mining means extracting something useful or valuable from a baser substance). Based on the information gathered over the WWW web mining is categorized into three: Web content mining, Web structure mining and Web usage mining. In search engines an web mining application can be seen. Most of the search engines are ranking their search results in response to user's queries to make their search navigations easier. In this paper we give a survey of page ranking algorithms and 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.

Keywords- **Data mining; Web mining, web content, Page rank, Weighted Page rank, hits, weighted page content rank, web structure.**

I. INTRODUCTION

The **World Wide Web** ("**WWW**" or the "**Web**") is a rich source of voluminous and heterogeneous information continues to expand in size and complexity. Retrieving of the required web page on the web, efficiently and effectively, is becoming a challenge. User always wants to have relevant pages when he/she performs searching on the web. Because of the bulk amount of information user is feeling very difficult to find, extract, filter or evaluate the relevant information. Web mining deals with extracting these interesting patterns & developing useful abstracts from diversified sources. The following are the challenges [2] [8] in web mining:

1) The amount of information is huge on the web
2) The coverage of web information is very wide and diverse
3) Information/Data of most all types exist on the web
4) Much of web information is semi structured
5) Much of the web information is linked
6) Much of the web information is redundant
7) The web is noisy
8) The web is also about services
9) The web is dynamic
10) The web is a virtual society

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 an application of data mining. It involves the analysis of Web server logs of a Web site. The Web server logs contain the entire collection of requests made by user through their browser and responses by the Web server. The information in the logs varies depending on the log file format and option selected on the Web server.

Web Mining Process: The complete process of extracting knowledge from Web data [5] [8] is as follows in Figure-1:

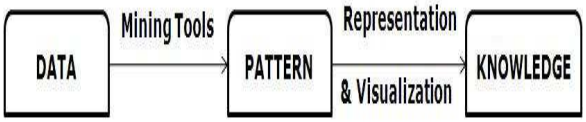

Fig.1. Web mining process

Web mining can be decomposed into the subtasks, namely:

1. **Resource finding**: the task of retrieving intended Web documents. By resource finding we mean the process of retrieving the data that is either online or offline from the text sources available on the web such as electronic newsletters, electronic newswire, the text contents of HTML documents obtained by removing HTML tags, and also the manual selection of Web resources.
2. **Information selection and pre-processing**: automatically selecting and preprocessing specific information from retrieved Web resources. It is a kind of transformation processes of the original data retrieved in the IR process. These transformations could be either a kind of pre-processing that are mentioned above such as stop words, stemming, etc. or a pre-processing aimed at obtaining the desired representation such as finding phrases in the training corpus, transforming the representation to relational or first order logic form, etc.
3. **Generalization**: automatically discovers general patterns at individual Websites as well as across multiple sites. Machine learning or data mining techniques are typically used in the process of generalization. Humans play an important role in the information or knowledge discovery process on the Web since the Web is an interactive medium.
4. **Analysis**: validating and/or interpretation of the mined patterns.

III. WEB MINING CATEGORIES

Web mining aims at finding and extracting relevant information that is hidden in Web-related data, in particular hypertext documents published on the Web. Web Mining is the extraction of interesting and potentially useful patterns

ISSN: 2277-3754
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 1, Issue 2, February 2012

and implicit information from artifacts or activity related to the World Wide Web [1-2-4-8]. Web mining research overlaps substantially with other areas, including data mining, text mining, information retrieval, and Web retrieval. There are roughly three knowledge discovery domains that pertain to web mining: Web Content Mining, Web Structure Mining, and Web Usage Mining as shown in Figure-2. However, there are two other different approaches to categorize Web mining. In both, the categories are reduced from three to two: Web content mining and Web usage mining.

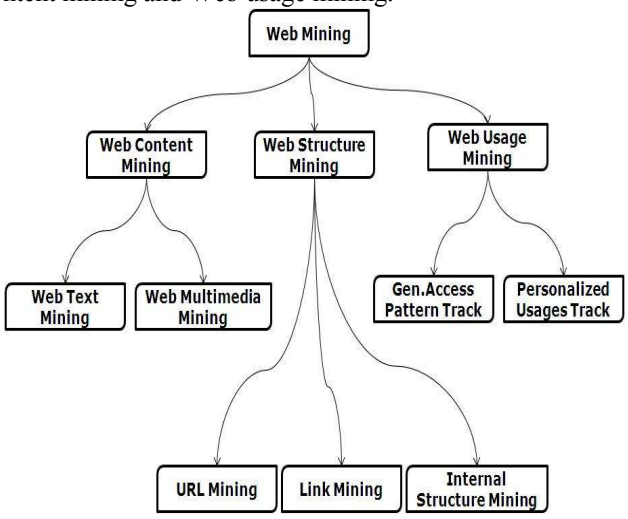

Fig.2. Web Mining Categories

A. Web Content Mining

Web Content Mining [1-2-4] deals with discovering useful information or knowledge from web page contents. Web content mining analyzes the content of Web resources. Content data is the collection of facts that are contained in a web page. It consists of unstructured data such as free texts, images, audio, video, semi-structured data such as HTML documents, and a more structured data such as data in tables or database generated HTML pages. The primary Web resources that are mined in Web content mining are individual pages. They can be used to group, categorize, analyze, and retrieve documents. Web content mining could be differentiated from two points of view:

1. Agent-Based Approach: This approach aims to assist or to improve the information finding and filtering the information to the users. This could be placed into the following three categories:

a. Intelligent Search Agents: These agents search for relevant information using domain characteristics and user profiles to organize and interpret the discovered information.

b. Information Filtering/ Categorization: These agents use information retrieval techniques and characteristics of open hypertext Web documents to automatically retrieve, filter, and categorize them.

c. Personalized Web Agents: These agents learn user preferences and discover Web information based on these preferences, and preferences of other users with similar interest.

2. Database Approach: Database approach aims on modeling the data on the Web into more structured form in order to apply standard database querying mechanism and data mining applications to analyze it. The two main categories are

a. Multilevel databases: The main idea behind this approach is that the lowest level of the database contains semi-structured information stored in various Web sources, such as hypertext documents. At the higher level(s) meta data or generalizations are extracted from lower levels and organized in structured collections, i.e. relational or object-oriented databases.

b. Web query systems: Many Web-based query systems and languages utilize standard database query languages such as SQL, structural information about Web documents, and even natural language processing for the queries that are used in World Wide Web searches.

B. Web Structure Mining

Web structure mining [2-4] is the process of discovering structure information from the web. The structure of a typical web graph consists of web pages as nodes, and hyperlinks as edges connecting related pages as shown in Figure-3.

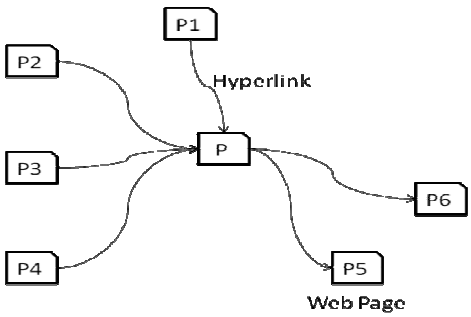

Fig.3 Web graph structure

a) Hyperlinks

A hyperlink is a structural unit that connects a location in a web page to a different location, either within the same web page or on a different web page. A hyperlink that connects to a different part of the same page is called an Intra-document hyperlink, and a hyperlink that connects two different pages is called an inter-document hyperlink.

b) Document Structure

In addition, the content within a Web page can also be organized in a tree structured format, based on the various HTML and XML tags within the page. Mining efforts here have focused on automatically extracting document object model (DOM) structures out of documents.

Web structure mining focuses on the hyperlink structure within the Web itself. The different objects are linked in some way. Simply applying the traditional processes and assuming that the events are independent can lead to wrong conclusions. However, the appropriate handling of the links could lead to potential correlations, and then improve the predictive accuracy of the learned models.

ISSN: 2277-3754
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 1, Issue 2, February 2012

C. WEB USAGE MINING

Web usage mining [2] is the process of finding out what users are looking for on the Internet. Web usage mining focuses on techniques that could predict the behavior of users while they are interacting with the WWW. It collects the data from Web log records to discover user access patterns of Web pages. Usage data captures the identity or origin of web users along with their browsing behavior at a web site. There are two main tendencies in Web Usage Mining driven by the applications of the discoveries: General Access Pattern Tracking and Customized Usage Tracking. The general access pattern tracking analyzes the web logs to understand access patterns and trends. Customized usage tracking analyzes individual trends. Its purpose is to customize web sites to users. Figure 4 is shown in Appendix.

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 labeled 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 [7-8]. Three important algorithms Page Rank, Weighted Page Rank and Weighted Page Content Rank are discussed below:

A. PAGE RANK

Page Rank [8] is one of the methods Google uses to determine a page's relevance or importance. The page rank value for a page is calculated based on the number of pages that point to it. This is actually a measure based on the number of back links to a page. Page Rank is displayed on the toolbar of your browser if you've installed the Google toolbar (http://toolbar.google.com/). But the Toolbar Page Rank only goes from 0 – 10 and seems to be something like a logarithmicscale:

Toolbar PageRank (log base 10)	Real PageRank
0	0 – 100
1	100 - 1,000
2	1,000 - 10,000
3	10,000 - 100,000
4	and so on...

Following are some of the terms used:

(1) PR: Shorthand for Page Rank: the actual, real, page rank for each page as calculated by Google.

(2) TOOLBAR PR: The Page Rank displayed in the Google toolbar in your browser. This ranges from 0 to 10.

(3) BACKLINK: If page A links out to page B, then page B is said to have a "back link" from page A.

We can't know the exact details of the scale because the maximum PR of all pages on the web changes every month when Google does its re-indexing! If we presume the scale is logarithmic then Google could simply give the highest actual PR page a toolbar PR of 10 and scale the rest appropriately.

Page Rank is a "vote", by all the other pages on the Web, about how important a page is. A link to a page counts as a vote of support. If there's no link there's no support (but it's an abstention from voting rather than a vote against the page). The another definition given by Google is as follows: We assume page A has pages T1...Tn which point to it (i.e., are citations). The parameter d is a damping factor, which can be set between 0 and 1. We usually set d to 0.85. Also C(A) is defined as the number of links going out of page A. The Page Rank of a page A is given as follows:

PR (A) = (1-d) + d (PR (T1)/C (T1) + ... + PR (Tn)/C (Tn))

Note that the Page Ranks form a probability distribution over web pages, so the sum of all web pages' Page Ranks will be one. Page Rank or PR (A) has the following sections:

1. **PR (Tn)** - Each page has a notion of its own self-importance. That's "PR (T1)" for the first page in the web all the way up to "PR (Tn)" for the last page.
2. **C (Tn)** - Each page spreads its vote out evenly amongst all of it's outgoing links. The count, or number, of outgoing links for page 1 is "C (T1)", "C (Tn)" for page n, and so on for all pages.
3. **PR(Tn)/C(Tn)** - so if our page (page A) has a backlink from page "n" the share of the vote page A will get is "PR(Tn)/C(Tn)".
4. **d(...** - All these fractions of votes are added together but, to stop the other pages having too much influence, this total vote is "damped down" by multiplying it by 0.85 (the factor "d").

B. Weighted Page Rank

The more popular web pages are the more linkages that other web pages tend to have to them or are linked to by them. The proposed extended Page Rank algorithm–a Weighted Page Rank Algorithm[7]–assigns larger rank values to more important (popular) pages instead of dividing the rank value of a page evenly among its out link pages. Each out link page gets a value proportional to its popularity (its number of in links and out links). The popularity from the number of in links and out links is recorded as $W^{in}(v,u)$ and $W^{out}(v,u)$, 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.

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

Where I_u and I_p represent the number of in links of page u and page p, respectively. $R(v)$ denotes the reference page list of page v. $W^{out}(v,u)$ is the weight of *link*(v, u) calculated based on the number of out links of page u and the number of out links of all reference pages of page v.

ISSN: 2277-3754
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 1, Issue 2, February 2012

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

Where O_u and O_p represent the number of out links of page u and page p, respectively. $R(v)$ denotes the reference page list of page v. Figure-5 shows an example of some links of a hypothetical website.

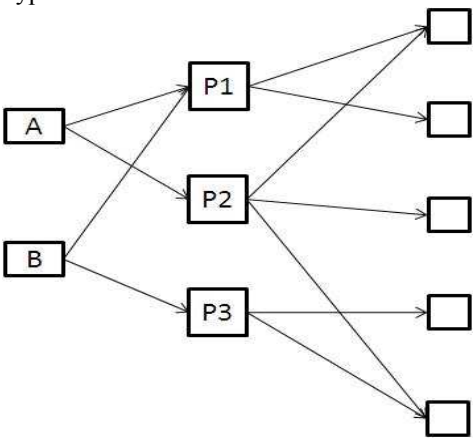

Fig.5. Links of a Website

C. Weighted Page Content Rank

Weighted Page Content Rank Algorithm (WPCR)[7-8] 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 in links and out links 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, in link and Out link Weights of
All back links 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 back links of P (say set B).

b) PR (P) = (1-d) +d

c) Output PR (P) i.e. the Rank score

V. COMPARISON OF PAGE RANKING ALGORITHMS

The following table shows the difference between Page Rank, Weighted Page Rank, and Page Content Rank. Table is shown in Appendix.

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 focused on comparitative study of page rank Algorithms .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 Back links, Weighted Page Rank uses Back links and Forward Links as Input Parameter and Weighted Page Content Rank uses Back links, 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] S. Chakrabarti et al., "Mining the Web's Link Structure". Computer, 32(8):60-67, 1999.

[2] Raymond Kosala, Hendrik Blockee, "Web Mining Research: A Survey", ACM Sigkdd Explorations Newsletter, June 2000, Volume 2.

[3] Cooley, R., Mobasher, B., and 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), Newport Beach, CA, 1997.

[4] Pooja Sharma, Deepak Tyagi, Pawan Bhadana “Weighted Page Content Rank for Ordering Web Search Result” International Journal of Engineering Science and Technology Vol. 2 (12), 2010, 7301-7310.

[5] Companion slides for the text by Dr. M. H. Dunham, "Data Mining: Introductory and Advanced Topics", Prentice Hall, 2002.

[6] Wenpu Xing and Ali Ghorbani, "Weighted Page Rank Algorithm", Proceedings of the Second Annual Conference on Communication Networks and Services Research (CNSR'04), 2004 IEEE.

[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.

[8] Tamanna Bhatia, ” Link Analysis Algorithms For Web Mining “, IJCST Vol. 2, Iss ue 2, June 2011.

ISSN: 2277-3754
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 1, Issue 2, February 2012

AUTHOR BIOGRAPHY

T.Munibalaji received MCA from Andhra University, Visakahapatnam, India in 2001 and M.Tech (CS) from Acharya Nagarjuna University in 2010 and pursuing M.Phil in Computer Science from SV University, Tirupati, India. I have got 10 years of teaching and industrial experience. Served as the Head, Dept of MCA, S V College of Engineering, Karakambadi, Tirupati, India during 2008-2009. My areas of interests include Data Mining and Data warehousing, Intelligent Systems and Cloud Computing.

.

C.Balamurugan received MCA from Dr.M.G.R University, Chennai, India in 2009 and pursuing M.B.A from S.V. University, Tirupati, India. I have 3 years of teaching and industrial experience. My areas of interests include Data Mining and Data warehousing, Computer Networks, Information Security, Database Management Systems and Cloud Computing.

APPENDIX

ISSN: 2277-3754
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 1, Issue 2, February 2012

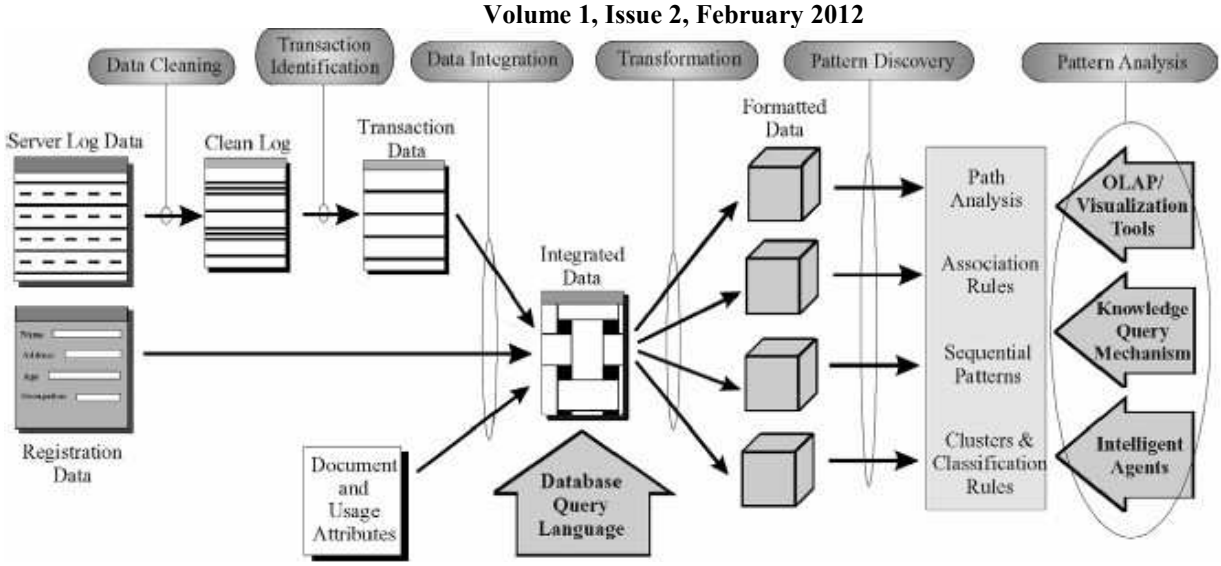

Fig.4 Web Usage Mining Process [3]

Algorithm	Page Rank	Weighted Page Rank	Page Content Rank
Mining Technique used	Web Structure Mining	Web Structure Mining	Web Content Mining
Working Procedure	Computes scores at indexing time not at query time. Results are sorted according to importance of pages	Computes scores at indexing time, unequal distribution of score. Pages are sorted according to importance. In other words Assigns large value to more important pages instead of diving the rank value of a page evenly among its Out link pages.	Computes new scores of the top n* pages on the fly. Pages returned are related to the query i.e. relevant documents are returned. I n other words 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, forward links	Content
Complexity	O(log N)	<O(log N)	O(m^*)
Limitations	Computes scores at indexing time. Results are sorted according to importance of pages.PR is equally distributed to outgoing links	Determination of relevancy is ignored. Some pages may be irrelevant to the given query.	Importance of WebPages is ignored. .Best when compared with Page Rank and Weighted Page Rank

*n: number of pages chosen by the algorithm, N: number of web pages, m: Total number of occurrences of query terms in n pages