Link Analysis Algorithms for Web Structure Mining

T.Nithya[1], V.Kumutha[2]

ABSTRACT

As the web is growing rapidly, the users get easily lost in the web's rich hyper structure. The primary goal of the web site owner is to provide the relevant information to the users to fulfill their needs. Web mining technique is used to categorize users and pages by analyzing users behavior, the content of pages and order of URLs accessed. Web Structure Mining plays an important role in this approach. In this paper we discuss and compare the commonly used algorithms i.e.PageRank, Weighted Page Rank, Weighted Page Content Rank and HITS.

Keywords-Web mining, web content, Page rank, Weighted Page rank and weighted page content rank, and Hyper-link Induced Topic Search

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 IV(D) defines Hyper-link Induced Topic Search 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 is follows in Fig.1:

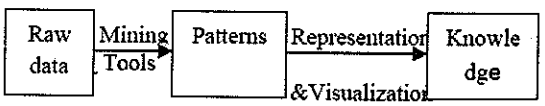

Fig. 1: Web Mining Process

1. Research Scholar
Department of Computer Science, D.J.Academy for Managerial Excellence, tnithyamca@gmail.com
2. Assistant Professor
Department of Computer Science, D.J.Academy for Managerial Excellence, kumutha.v@djacademy.co.in

To clarify the confusion of what forms Web mining. Kosala and Blockeel [2] had suggested a decomposition of Web mining in the following tasks:

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

In general, Web mining tasks can be classified into three categories [2; 3]: Web content mining, Web structure mining and Web usage mining. 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

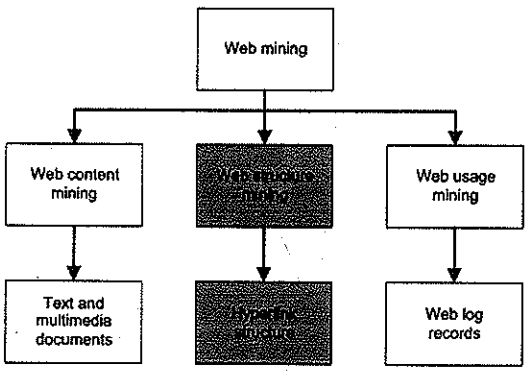

Fig. 2: Web Mining Categories

A. Web Content Mining

Web content mining [2][5] aims to extract/mine useful information or knowledge from web page contents. Web content mining is related but different from data mining and text mining. It is related to data mining because many data mining techniques can be applied in Web content mining. It is related to text mining because much of the web contents are texts. However, it is also quite different from data mining because Web data are mainly semi-structured and/or unstructured, while data mining deals primarily with structured data. 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. Web content mining thus requires creative applications of data mining and/or text mining techniques and also its own unique approaches. The technologies that are normally used in web content mining are NLP (Natural language processing) and IR (Information retrieval).

B. Web Structure Mining

The challenge for Web structure mining is to deal with the structure of the hyperlinks within the Web itself. Link analysis is an old area of research. However, with the growing interest in Web mining, the research of structure analysis had increased and these efforts had resulted in a newly emerging research area called Link Mining [4], which is located at the intersection of the work in link analysis, hypertext and web mining, relational learning and inductive logic programming, and graph mining. There is a potentially wide range of application areas for this new area of research, including Internet. The Web contains a variety of objects with almost no unifying structure, with differences in the authoring style and content much greater than in traditional collections of text documents. The objects in the WWW are web pages, and links are in-, out-and co-citation (two pages that are both linked to by the same page). Attributes include

HTML tags, word appearances and anchor texts [4]. This diversity of objects creates new problems and challenges, since is not possible to directly made use of existing techniques such as from database management or information retrieval. Link mining had produced some agitation on some of the traditional data mining tasks. As follows, we summarize some of these possible tasks of links mining which are applicable in Web structure mining.

1. Link-based Classification. Link-based classifications the most recent upgrade of a classic data mining task to linked domains [6]. The task is to focus on the prediction of the category of a web page, based on words that occur on the page, links between pages, anchor text, html tags and other possible attributes found on the web page.

2. *Link-based Cluster Analysis.* The goal in cluster analysis is to find naturally occurring sub-classes. The data is segmented into groups, where similar objects are grouped together, and dissimilar objects are grouped into different groups. Different than the previous task, link-based cluster analysis is unsupervised and can be used to discover hidden patterns from data.

3. *Link Type.* There are a wide range of tasks concerning the prediction of the existence of links, such as predicting the type of link between two entities, or predicting the purpose of a link.

4. *Link Strength.* Links could be associated with weights.

5. *Link Cardinality.* The main task here is to predict the number of links between objects.

Web Usage Mining

Web Usage Mining [7][8] is the process by which we identify the browsing patterns by analyzing the navigational behavior of user. It focuses on techniques that can be used to predict the user behavior 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 preprocessing, 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 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. Four important algoPage Rank, Weighted PageRank and Weighted Page Content Rank, HITS are discussed below:

IV (A) PageRank

This algorithm was developed by Brin and Page at Stanford University which extends the idea of citation analysis [5]. In citation analysis the incoming links are treated as citations but this technique could not provide fruitful results because this gives some approximation of importance of page. So PageRank provides a better approach that can compute the importance of web page by simply counting the number of pages that are linking to it. These links are called as backlinks.If a backlink comes from an important page than this link is given higher weightage than those which are coming from non-important pages. The link from one page to another is considered as a vote. Not only the number of votes that a page receives is important but the importance of pages that casts the vote is also important. Page and Brin proposed a formula to calculate the PageRankof a page A as stated below-

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

Here *PR(Ti)* is the PageRank of the Pages Ti which links to page A, *C(Ti)* is number of outlinks on page Ti and *d* is damping factor. It is used to stop other pages having too much influence.

The PageRank forms a probability distribution over the web pages so the sum of PageRanks of all web pages will be one. The PageRank of a page can be calculated without knowing the final value of PageRank of other pages. It is an iterative algorithm which follows the principle of normalized link matrix of web. PageRank of a page depends on the number of pages pointing to a page.

IV(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. It 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..... (4)

IV(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=Setof 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

IV (D) Hyper-link Induced Topic Search (HITS)

Klienberg gives two forms of web pages called as hubs and authorities. Hubs are the pages that act as resource lists. Authorities are pages having important contents. A good hub page is a page which is pointing to many authoritative pages on that content and a good authority page is a page which is pointed by many good hub pages on the same content. A page may be a good hub and a good authority at the same time [10][11].

Following expressions are used to calculate the weight of Hub *(Hp)* and the weight of Authority *(Ap)*.

$$H_p = \sum_{q \in I(p)} A_q \qquad A_p = \sum_{q \in B(p)} H_q$$

Here Hq is Hub Score of a page, *Aq* is authority score of a page, *I(p)* is set of reference pages of page p and *B(p)* is set of referrer pages of page p, the authority weight of a page is proportional to the sum of hub weights of pages that link to it. Similarly a hub of a page is proportional to the sum of authority weights of pages that it links to.

4.3.1 Constraints with HITS algorithm [12]

Hubs and authorities: It is not easy to distinguish between hubs and authorities because many sites are hubs as well as authorities.

Topic drift: Sometime HITS may not produce the most relevant documents to the user queries because of equivalent weights.

Automatically generated links: HITS gives equal importance for automatically generated links which May not have relevant topics for the user query

Efficiency: HITS algorithm is not efficient in real-time.

V. Comparison of Algorithms

Table1 shows the difference between above four algorithms:

Table 1: Comparison of algorithms

Algorithm	Page Rank	Weighted Page Rank	Weighted Page Content Rank	HITS
Mining Technique Used	WSM	WSM	WSM and WCM	WSM and WCM
Complexity	O(log N)	<O(log N)	<O(log N)	<O(log N)
Input/ Output Parameters	Backlinks	Backlinks and forward links	Backlinks, forward links and content	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	It provide important information and relevancy about a given query by using web structure and web content mining
Search Engine	Google	Google	Research Model	Clever

VI. Conclusion

Web Mining is powerful technique used to extract the information from past behavior of users. Web Structure Mining plays an important role in this approach. Various algorithms are used in Web Structure Mining to rank the relevant pages. PageRank, Weighted PageRank, Weighted Page Content Rank and HITS treat all links equally when distributing the rank score. PageRank and Weighted PageRank are used in Web Structure Mining. Weighted Page Content Rank and HITS are used in both structure Mining and Web Content Mining. The input parameters used in Page Rank are BackLinks, Weighted PageRank uses Backlinks and Forward Links as Input Parameter, HITS and Weighted Page Content Rank uses Backlinks, Forward Link and Content as Input Parameters. Complexity of PageRank algorithm is O(log N) where as complexity of Weighted PageRank , Weighted Page Content Rank and HITS algorithms are <O(log N).

As part of future work, the planning to carry out performance analysis of Weighted Page Content and HITS and working on finding required relevant and important pages more easily and fastly.

REFERENCES

[1] M.G. da Gomes Jr. and Z. Gong, Web Structure Mining:An Introduction, Proceedings of the IEEE International Conference on Information Acquisition, 2005.

[2] Raymond Kosala, Hendrik Blockeel, Web Mining Research: A Survey,*ACM SIGKDD Explorations Newsletter*, June 2000, Volume 2 Issue 1.

[3] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pag-Ning Tan,Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, *ACM SIGKDD Explorations Newsletter,* January 2000, Volume 1 Issue 2.

[4] L. Getoor, Link Mining: A New Data Mining Challenge. *SIGKDD Explorations*, vol. 4, issue 2, 2003.

[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] Q. Lu, and L. Getoor. Link-based classification. In *Proceedings of ICML-03*, 2003.

[7] Rekha Jain, Dr G.N.Purohit, "Page Ranking Algorithms for Web Mining," International Journal of Computer application,Vol 13, Jan 2011.

[8] Pooja Sharma, Pawan Bhadana, "Weighted Page Content Rank For Ordering Web Search Result", International Journal of Engineering Science and Technology, Vol 2, 2010.

[9] 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.

[10] Ding, X. He, P. Husbands, H. Zha, and H. Simon, Link analysis: Hubs and authorities on the world. Technical report: 47847, 2001.

[11]J. M. Klienberg, Authoritative sources in a hyperlinked environment. Journal of the ACM, 46(5):604-632, September 1999.

[12]Chakrabarti, B.Dom, D.Gibson, J. Kleinberg, R. Kumar, P. Raghavan,S. Rajagopalan, and A. Tomkins, Mining the Link Structure of the World Wide Web, IEEE Computer, Vol. 32, pp. 60-67, 1999.

Author's Biography

Ms. T.Nithya received her Professional degree MCA., and M.Phil., in Computer Application from D.J Academy for Managerial Excellence, Bharathiar University, Coimbatore. Her area of Research includes Data Mining and Web Mining. She has presented 2 papers in National Conferences. She has 2 years of teaching experience.

Ms. V. Kumutha received her Professional degree MCA. and M.Phil degree in Computer Science. Her area of Research includes Data Mining, Genetic algorithm and Neural Network. She has presented papers in National and International Conferences She has 9 years of experience in teaching and currently working as Assistant Professor in MCA department at D.J. Academy for Managerial Excellence, Coimbatore.