An Analysis of Web Mining and its types besides Comparison of Link Mining Algorithms in addition to its specifications

B. Rajdeepa, Dr. P. Sumathi

Abstract---Due to fast development of online data, World Wide Web data has becoming one of the most important sources to discover the knowledge and to extract the information. LinkMining technologies are the right solutions for knowledge discovery on the Web. Web Mining is mainly concerned with Web Usage, Web Structure and Web Content Mining.This paper gives the introduction about Web Mining and their types and mainly concentrate onWeb Structure Mining.This paper gives the comparative analysis of different algorithms in Web Structure Mining and also give some review about HITS and PageRank and their applications.

Keywords---Web Mining, Web Structure Mining, HITS, PageRank.

I. INTRODUCTION

Nowadays, the World Wide Web has becoming one of the most comprehensive information resources. It probably, if not always, gives the information need for any user. However, the Web demonstrates many radical differences to traditional information containers such as databases, in schema, volume and topic-coherence. These differences make it thought-provoking to fully use Web information in an effective and efficient manner. Web mining is right for this need [1].

In fact, Web mining can be considered as the application of the general data mining techniques to the Web. However, the essential properties of the Web make us have to tailor and extend the traditional methodologies considerably. Firstly, even though Web contains huge bulk of data, which is distributed on the internet. Before mining, we need to gather the Web document organized. Secondly, Web pages are semi-structured, in order for easy processing, documents should be extracted and represented into some format. Thirdly, Web information tends to be of diversity in meaning, training or testing data set should be large enough. Even though the difficulties above, the Web also provides other ways to support mining, for instance, the links between Web pages are important resource to be used.

Kosala and Blockeel [2] had suggested a decomposition of Web Mining in the following tasks:

- Resource finding: the duty of retrieving planned Web documents.
- Information selection and pre-processing: mechanically selecting and pre-processing specific information from retrieved Web resources.
- Generalization: automatically determines general patterns at individual Web sites as well as across multiple sites.
- Analysis: authentication and/or interpretation of the mined patterns.

Generally Web Mining Data is classified into three types and it is shown in figure 1.

- Web Content mining,
- Web Structure mining and
- Web Usage mining

Web Usage mining is the process of extracting useful information from server logs i.e. users past. Web usage mining is the process of finding out what users are looking for on the Internet. Some users might be seeing at only textual data, whereas some others might be fascinated in multimedia data. This technology is basically focused upon the use of the web technologies which could help for betterment [3].

Web content mining means mining, extraction and integration of useful data, information and knowledge from Web page substances. Content mining is the scanning and mining of text, pictures plus graphs of a Web page to determine the relevance of the content to the search query. This scanning is done after the clustering of web pages through structure mining and provides the results based upon the level of relevance to the suggested query. With the gigantic amount of information that is available on the World Wide Web. Content mining provides the results lists to search engines in order of highest relevance to the keywords in the query. [2]

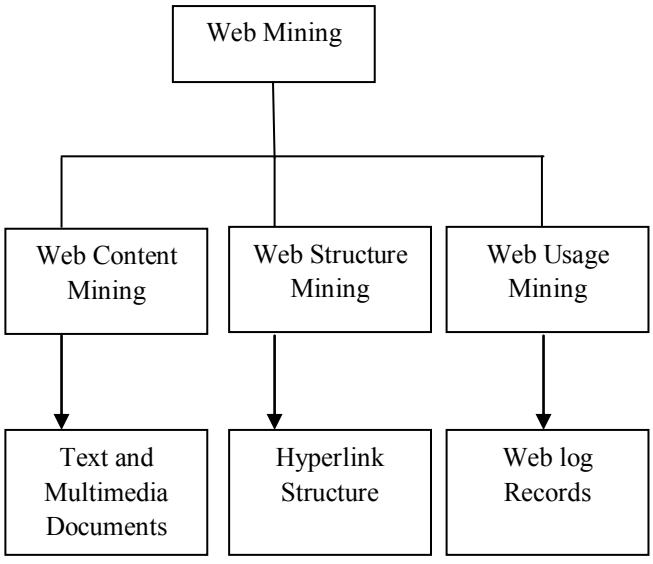

Figure 1. Overview of Web Mining

ISSN: 2278 – 1323

Web structure mining focuses on the hyperlink structureof the Web Page. The different objects are linked in variousways.Simply applying the old processes and assuming thatthe events are independent can lead to incorrect conclusions.However, appropriate handling of the links could lead topotential associations, and then improves the predictive accuracyof the learned models [4]. Two algorithms that have beenproposed to lead with those potential correlations: HITS [5]and PageRank [6].

II. WEB STRUCTURE MINING

The challenge for Web structure mining is to deal withthe structure of the hyperlinks within the Web itself. Linkanalysis is an old area of research. However, with the growinginterest in Web mining, the research of structure analysis hadincreased and these efforts had resulted in a newly emergingresearch area called Link Mining [4], which is located at theintersection of the work in link analysis, hypertext, webmining, relational learning and inductive logic programmingand graph mining. There is a actually a wide range of applicationareas for this new area of research, including Internet.

The Web contains a variety of objects with almost no similar structure i.e, with differences in the authoring style andcontent much greater than in traditional collections of textdocuments. The objects in the WWW are web pages, and links. The linksare in-, out- and co-citation (two pages that are both linked to by the same page). Attributes include HTML tags, word appearance and anchor texts [4]. This diversity of object 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 anxiety on some of the traditional data mining tasks. In this paper, we summarize some of these possible tasksof link mining which are applicable in Web structure mining.

Web structure mining is the process of using graph theory to analyze the node and connection structure of a web site.

According to the type of web organisational data, web structure mining can be dividedinto two kinds:

- Extracting patterns from hyperlinks in the web: a hyperlink is a structural component that connects the web page to a different location.
- Mining the document structure: analysis of tree-like structure of the page structures to describe HTML or XML tag usage [7].

Web structure mining is also known as “Link Analysis”. It is an old area of research but with the increasing interest in Web mining the research of structure analysis had also increased and these efforts had resulted in a newly emerging research area named Link Mining. The Web contains a variety of objects with almost no unifying structure but with differences in the style and Content muchgreater than in traditional collections of text documents. Link mining is divided into four parts and is shown in following figure:

The objects in the WWW are web pages and links. The links are in-, out- and co-citation i.e. two pages that are both linked to same page. There are some possible tasks [8] of link mining which are applicable in Web structure mining and are described as follows:

Link-based Classification: - is the most recent upgrade of a classic data mining task to web hyperlink Domains. This task is to focus on the prediction of the category of a web page, based on words that occurs on the page, links between pages, anchor texts, html tags and other possible attributes found on the web page.

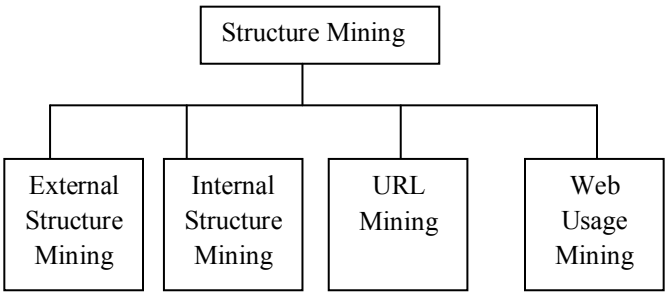

Figure 2. Types of Web Structure Mining

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 an unsupervised and can be used to discover hidden patterns from data.

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.

Link Strength:- Links could be associated with weights.

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

There are some areas where web structure mining is applied and some them are listed below:

- Used to rank the user’s query
- To deciding what page will be added to the collection
- page categorization
- to find related pages
- to find duplicated web sites and also to find out similarity between them

The structure of a typical web graph consists of web pages as nodes and hyperlinksas edges connecting related pages. Web structure mining is the processof discovering structure information from the web. This can be further dividedinto some kinds based on the kind of structure information used.

Hyperlinks : A hyperlink is a structural unit that connects a location in a web page to adifferent location i.e. 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 anintradocument hyperlink, and a hyperlink that connects two different pages iscalled an interdocument hyperlink.

Document Structure :-The content within a Web page can also be organized as a treestructuredformat, based on the various HTML and XML tags within the page.Mining efforts here have focused on automatically extracting document objectmodel (DOM) structures out of documents.

TABLE I

COMPARISON OF DIFFERENT ALGORITHM

Algorihtm	Pagerank	HITS	Weighted Content Pagerank	Weigted Pagerank	Topic Sensitive Pagerank
Main Technique	Web Structure Mining	Web Structure Mining	Web Structure Mining and Web Content Mining	Web Structure Mining	Web Structure Mining
I/O Parameters	Backlink	Content, Backlink, forward link	Content, Backlink, forward link	Content, Backlink, forward link	Content, Backlink, forward link
Complex	O(login)	<O(Login)	<O(login)	<O(login)	<O(login)
Working	This algorithm computes the score for pages at the time of indexing of the pages	It computes hubs and authority of the relevant pages. It relevant as well as important pages as the result	It gives different weight to web links based on three attributes: Releative Position in page, tag where link is contained, length of anchor text.	Weight of web page is calculated on the basis of input and outgoing links and on the basis of weight the importance of the page is decided	It computes the scores of pages according to the importance of content available on particular web page
Efficiency	Very less	Moderate	Average	Average	Good
Importance	High, black links are considered	Moderate, Hub authorities scores are utilized	High	High. The pages are sorted according to the importance	High. Score according to importance is calculated
Limitations	Results come at the time of indexing and not at the query time.	Topic drift and efficiency problem	Comparative position was not so effective, indicating that the logical position not always matches the physical position	Relevancy is ignored	Only available to text, images are not taken into account
Search	Google	Google	Google	Google	Google
Quality of result	Medium	Less than pagerank	Higher	Higher than pagerank	Higher than all
Relevancy	Less	More	Less	Less	More

TABLE II

WEB STRUCTURE MINING SPECIFICATION

Specification	Web Structure Mining
View of Data	Linking of Structure
Type of data used in mining	Primary
Main Data	Link Strucuture
Representation	Graph
Method	Proprietary Algorithm
Tasks	It tries to discover the model underlying the link structure of the web
Scope	Global
Application categories	Categorization, clustering

Table I gives the Advantages and limitations of some important web page ranking algorithms. Table II gives the specification for the web struture mining. That's how the data is avaliable on the web, what type of data is used for mining, Main data of mining, then how the data is shown on the web, what type of method is used for mining the web data, atlast their task, scope and their application categories.

III. WEB STRUCTURE TERMINOLOGY

The Web as a whole can be modeled as a directed graph containing a set of nodes and directed edges between them. Broder et al [10] studied the web graph and described some of the basic terminology necessary for a web graph model. The nodes represent the Web pages and the directed edges are the hyperlinks. We now define a set of terms that are frequently used to describe the Web graph structure and other more abstract concepts about the Web.

- Web-graph: A directed graph that represents the Web.
- Node: Each Web page is a node of the Web-graph.
- Link: Each hyperlink on the Web is a directed edge of the Web-graph.
- Indegree: The indegree of a node, p, is the number of distinct links that point to p.
- Outdegree: The outdegree of a node, p, is the number of distinct links originating at p that point to other nodes.
- Directed Path: A sequence of links, starting from p that can be followed to reach q^2.
- Shortest Path: Of all the paths between nodes p and q, which has the shortest length, i.e. number of links on it.
- Diameter: The maximum of all the shortest paths between a pair of nodes p and q, for all pairs of nodes p and q in the Web-graph.
- Average Connected Distance: Average of the lengths of the shortest paths from node p to node q, for all pairs of nodes p and q [9]. Broder et al. [10] observed that this definition could result in an infinite average connected distance, if there is at least one pair of nodes p and q that have no existing path between them. And they proposed a revised definition: "the average connected distance is the expected length of the shortest path, where expectation is uniform choices from a set of all ordered pairs, (p,q) such that there exists a path from p to q"

The importance of information contained in the hyperlink pointing to a page has been familiar. Anchor texts (texts on hyperlinks in a HTML document) of predecessor pages were already indexed by the World-Wide Web. In [12] suggested a taxonomy of different types of (hyper-)links that can be found on the Web and discusses how the links can be exploited for various information retrieval tasks on the Web.

HITS have been usedfor identifying relevant documents for topics in web catalogues [13, 14] forimplementing the RelatedPages functionality [15]. The main drawback of the HITS algorithm is that the hubs and authorityscore must be computed iteratively from the query results, which does notmeet the real-time constraints of an on-line search engines. However, the implementationof a similar idea in the Google search engine resulted in a majorbreak-through in search engine technology. In [16] suggested theuse of the probability that a page is visited by a random surfer on the Web as akey factor for ranking search results. They approximated this probability withthe socalled PageRank, which is again computed iteratively.

HITS was used for the first time in the Clever [17]search engine from IBM and PageRank is used by Google [18]combined with other several features such as an anchor text, IRmeasures, and proximity.The notion of authoritativeness comes from the idea thatwe wish not only to locate a set of relevant pages ratherto locate the relevant pages with highest quality. However, the Webconsists not only of pages but also of links that connect onepage to another. This structure contains a large amount of informationthat should be exploited.PageRank and HITS belong to a class of ranking algorithms,where the scores can be computed as a fixed point of alinear equation.

Bianchini [19] noted that HITS and PageRankare used as starting points for new solutions and there aresome extensions of these two approaches. There are otherlink-based approaches to be applied on the Web and this is discussed in [19,20].

In [21]they have discussed all the algorithms,which use links, to below three categories.

- Relevant Linkage Principle: Links points to relevantresources.
- Topical Unity Principle: Documents often co-citedare related, as are those with extensive bibliographic overlap.This idea is previous addressed by Kesseler for bibliographicinformation retrieval in [22].
- Lexical Affinity Principle: Proximity of text and linkswithin a page is a measure of the relevance of one to another

IV. CONCLUSION

The World Wide Web contains a huge, universal, heterogeneous and unstructured data. Web mining is a broad research area which trying to solve issues that arise due to the WWW Phenomenon. This paper discusses about web Mining and their types.In addition to that it discusses about web structure mining and their terminologies. We also reviewed about web structure mining specification and it is given in table II and also provides comparison of different algorithms in table I. i.e. HITS and Page rank.

REFERENCES

[1] Etzioni, O. "The World Wide Web: Quagmire or gold mine", Communications of the ACM, 39(11):65-68, 1996.
[2] Raymond Kosala, Hendrik Blockeel, Web Mining Research: A Survey, ACM SIGKDD Explorations Newsletter, June 2000, Volume 2 Issue 1.
[3] Masand, B., Spiliopoulou, M., Srivastava, J and Zaiane, O. (2002) Proceedings of "WebKDD2002 –Web Mining for Usage Patterns and User Profiles", Edmonton, CA, 2002.
[4] Getoor, L (2003) "Link Mining: A New Data Mining Challenge", SIGKDD Explorations, vol. 4, issue 2, 2003.
[5] Kleinberg, J.M. (1998) "Authoritative sources in a hyperlinked environment", In Proceedings of ACM-SIAM Symposium on Discrete Algorithms, 1998, pages 668-677 – 1998.
[6] Page, L., Brin, S., Motwani, R and Winograd, T. (1998) "The Page rank citation ranking: Bring order to the web. Technical report", Stanford University, 1998.
[7] Srivastava, J., Cooley, R., Deshpande, M and Tan, PN. (2000)."Web Usage Mining: Discovery and Applications of usage patterns from Web Data", SIGKDD Explorations, Vol1, Issue 2, 2000
[8] Lu, Q and Getoor, L. (2003) "Link-based classification", In Proceedings of ICML-03, 2003.
[9] R. Albert, H.Jeong, and A.-L. Barabasi. Diameter of the World Wide Web, Nature 401: 130-131, Sep 1999.
[10] Broder et al, Graph Structure in the Web. In the Proc. 9th WWW Conference 2000
[11] McBryan, OA. "GENVL and WWWW: Tools for taming the Web", In Proceedings of the 1st World-Wide Web Conference (WWW-1), pages 58–67, Geneva, Switzerland, 1994. Elsevier.
[12] Spertus, E. "ParaSite: Mining structural information on the Web", Computer Networks and ISDN Systems, 29(8-13):1205–1215, September 1997. Proceedings of the 6th International World Wide Web Conference (WWW-6).
[13] Chakrabarti, B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson, and J. Kleinberg. "Automatic resource compilation by analyzing hyperlink structure and associated text. Computer Networks", 30(1–7):65–74, 1998b. Proceedings of the 7th InternationalWorldWideWeb Conference (WWW-7), Brisbane, Australia.
[14] Bharat, K and Henzinger, M. R." Improved algorithms for topic distillation in a hyperlinked environment", In Proceedings of the 21st ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-98), pages 104–111, 1998.
[15] J. Dean and M. R. Henzinger. Finding related pages in the World Wide Web. In A. Mendelzon, editor, Proceedings of the 8th International World Wide Web Conference (WWW-8), pages 389–401, Toronto, Canada, 1999.
[16] Brin and Page, L. "The anatomy of a large-scale hyper textual Web search engine. Computer Networks, 30(1–7):107–117, 1998. Proceedings of the 7th International World Wide Web Conference (WWW-7), Brisbane, Australia.
[17] http://www.research.ibm.com/topics/popups/innovate/hci/html/clever.html. Last accessed 15/04/2005.
[18] http://www.google.com/. Last accessed 15/04/2005.
[19] Bianchini, M., Gori, M and Scarselli, F "Inside Page Rank", ACM Transaction on Internet Technology (TOIT), Volume 5 Issue 1 – February, 2005.
[20] 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
[21] Ziv Bar-Yossef and Sridhar Rajagopalan. "Template Detection via Data Mining and its Applications", In: Proceedings of WWW2002, May 7-11, 2002, Honolulu, Hawaii, USA. 580-591.
[22] M. Kessler. Bibliographic coupling between scientific papers. American Documentation, 14:10-25, 196.

Mrs. B. Rajdeepa is presently working as an Assistant Professor in the Department of Computer Science, PSG College of Arts & Science (An Autonomous), Coimbatore, INDIA.She is currently pursuing her Ph.D. in Computer Science under Bharathiar University, Coimbatore, INDIA. She has received the M.Phil. Degree in Computer Science from Bharathidasan University, INDIA, in year 2006. She received MCA degree from Bharathiar University, Coimbatore, INDIA. She has published various papers in reputed international journals. She has attended and published papers in conferences. She has participated in various seminars and workshops. She has about ten years of teaching and research experience. Her area of interest is Data Mining.

Dr.P.Sumathi is presently working as an Assistant Professor inPG & Research Department of Computer Science, Government Arts College, Coimbatore She received her Ph.D., in the area of Grid Computing in Bharathiar University. She has done her M.Phil. in the area of Software Engineering from Mother Teresa University and received MCA degree from Kongu Engineering College, Perundurai. She has published a number of papers in reputed journals and conferences. She has about twenty years of teaching and research experience.She has participated in various seminars and workshops. Her research interests include Data Mining, Grid Computing and Software Engineering.

