International Journal of Applied Engineering Research, ISSN 0973-4562 Vol.9 No.27 (2014)

Analysis of web mining types and weblogs

S.Kamalakkannan
Research Scholar
Vels University
Chennai, India
Kamalindia81@yahoo.com

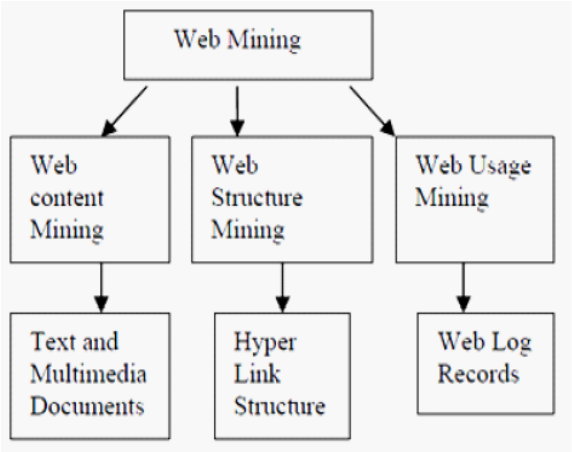
DR.S.Prasanna
AssociateProfessor
Vels University
Chennai, India
prasanna@velsuniv.org

***Abstract*-The main purpose of this paper is to analysis of Web mining types and Weblogs. 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. Web mining can be classified into web content mining, web structure mining, and web usage mining. Log files contain information about User Name, IP Address, Time Stamp, Access Request, number of Bytes Transferred, Result Status, URL that Referred and User Agent. The log files are maintained by the web servers. By analyzing these log files gives a neat idea about the user. This paper gives a detailed discussion about these log files.**

Keywords- Web content mining, Web structure mining, Web usage mining and Web Log file

I.INTRODUCTION

Web mining is useful to extract the information, image, text, audio, video, documents and multimedia. By using web mining easily extract all features and information about multimedia before this web mining difficult to extract information in proper way from web. We search the any topic from web difficult to get accurate topic information but now a days it is easy to get the proper and relevant information. Web mining is based on data mining technique by using data mining technique discover the hidden data in web log. Web mining can be classified into web content mining, web structure mining, and web usage mining as shown in Fig.



II. WEB CONTENT MINING

Web Mining is basically extracts the information on the web. Which process is happen to access the information on the web. It is web content mining. Many pages are open to access the information on the web. These pages are content of web. Searching the information and open search pages is also content of web. Last accurate result is defined the result pages content mining.

The various contents of Web Content Mining are

- Web page
- Search page
- Result page

Web Page: A Web page typically contains a mixture of many kinds of information, e.g., main content, advertisements, navigation panels, copyright notices, etc. For a particular application only some part of the information is useful and the rest are noises.

Search Page: A search page is typically used to search a particular Web page of the site, to be accessed numerous times in relevance to search queries. The clustering and organization of Web content in a content database enables effective navigation of the pages by the customer and search engines.

Result page: A result page typically contains the results, the web pages visited and the definition of last accurate result in the result pages of content mining.

III. WEB STRUCTURE MINING

We can define web structure mining in terms of graph. The web pages are representing as nodes and Hyperlinks represent as edges. Basically it's shown the relationship between user & web. The motive of web structure mining is generating structured summaries about information on web pages. It is shown the link one web page to another web page.

The various contents of Web structure mining are

• Links Structure Mining

• Internal Structure Mining

• URL Mining

Links Structure: 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 have resulted in a newly emerging research area called Link Mining. It consists Link-based Classification, Link-based Cluster Analysis, Link Type, Link Strength and Link Cardinality.

Internal Structure Mining: It can provide information about page ranking or authority and enhance search results through filtering i.e., tries to discover the model underlying the link structures of the web. This model is used to analyze the similarity and relationship between different web sites.

URL Mining: It gives a hyperlink which is a structural unit that connects a web page to different location, either within the same web page or to a different web page hyperlink.

IV. WEB USAGE MINING

It is discovery of meaningful pattern from data generated by client server transaction on one or more web localities. A web is a collection of inter related files on one or more web servers. It is automatically generated the data stored in server access logs, refers logs, agent logs, client sides cookies, user profile, meta data, page attribute, page content & site structure. Web mining usage aims at utilize data mining techniques to discover the usage patterns from web based application. It is technique to predict user behavior when it is interact with the web.

Web usage mining itself can be classified further depending on the kind of usage data considered:

Web Server Data
User logs are collected by the web server and typically include IP address, page reference and access time.

Application Server Data
Commercial application servers such as Web logic, Story Server, have significant features to enable E-commerce applications to be built on top of them with little effort. A key feature is the ability to track various kinds of business events and log them in application server logs.

Application Level Data
New kinds of events can be defined in an application, and logging can be turned on for them generating histories of these events. It must be noted, however, that many end applications require a combination of one or more of the techniques applied in the above the categories.

V CONTENTS OF A LOG FILE

The Log files in different web servers maintain different types of information. The basic information present in the log file are

• User name: This identifies who had visited the web site. The identification of the user mostly would be the IP address that is assigned by the Internet Service provider (ISP).This may be a temporary address that has been assigned. Therefore here the unique identification of the user is lagging. In some web sites the user identification is made by getting the user profile and allows them to access the web site by using a user name and password. In this kind of access the user is being identified uniquely so that the revisit of the user can also be identified.

• Visiting Path: The path taken by the user while visiting the web site. This may be by using the URL directly or by clicking on a link or trough a search engine.

• Path Traversed: This identifies the path taken by the user within the web site using the various links.

• Time stamp: The time spent by the user in each web page while surfing through the web site. This is identified as the session.

• Page last visited: The page that was visited by the user before he or she leaves the web site.

• Success rate: The success rate of the web site can be determined by the number of downloads made and the number copying activity under gone by the user. If any purchase of things or software made, this would also add up the success rate.

• User Agent: This is nothing but the browser from where the user sends the request to the web server. It's just a string describing the type and version of browser software being used.

• URL: The resource accessed by the user. It may be an HTML page, a CGI program, or a script.

• Request type: The method used for information transfer is noted. The methods like GET, POST. These are the contents present in the log file. This log file details are used in case of web usage mining process. According to web usage mining it mines the highly utilized web site. The utilization would be the frequently visited web site or the web site being utilized for longer time duration. Therefore the quantitative usage of the web site can be analyzed if the log file is analyzed.

VI. LOCATION OF A LOG FILE

A Web log is a file to which the Web server writes information each time a user requests a web site from that particular server. A log file can be located in three different places:

• Web Servers
• Web proxy Servers
• Client browsers

Web Server Log files

The log file that resides in the web server notes the activity of the client who accesses the web server for a web site through the browser.. In the server which collects the personal information of the user must have a secured transfer.

Web Proxy Server Log files

A Proxy server is said to be an intermediate server that exist between the client and the Web server. Therefore if the Web server gets a request of the client via the proxy server then the entries to the log file will be the information of the proxy server and not of the original user. These web proxy servers maintain a separate log file for gathering the information of the user.

Client Browsers Log files

This kind of log files can be made to reside in the client's browser window itself. Special types of software exist which can be downloaded by the user to their browser window. Even though the log file is present in the client's browser window the entries to the log file is done only by the Web server.

VII CONCLUSION

Designing and maintaining web based information system such as web sites is a real challenge. A huge amount of data is continuously increasing on the web day by day. So it is much easier to find the inconsistent information than the well structured information so the study of web mining help a lot to analyze this huge collection of information that is available on web and it is also used to predict the behavior of user using various techniques. Web data is growing at a significant rate. Web Mining is fertile area of research. Many Successful applications exist. We also suggest the subtask of web mining. The Paper gives a detailed look about the web mining types and web log file, its contents, its location. Web mining enhances user's ability to access information hence the capacity and potentials of enterprise information resources can be fully reflected. it is expected that more applications of web mining will be developed.

REFERENCES

[1] Dushyant Rathod, "A Review on Web Mining," IJERT, vol. 1, Issue 2, April – 2012.

[2] Tamanna Bhatia, "Link Analysis Algorithms For Web Mining," in IJCST Vol. 2, Issue 2, June 2011.

[3] Qingyu Zhang and Richard s. Segall," Web mining: a survey of current research,Techniques, and software", in the International Journal of Information Technology & Decision Making Vol. 7, No. 4 (2008) 683–720.

[4] D. Jayalatchumy, and P.Thambidurai, "Web Mining Research Issues and Future Directions – A Survey," IOSR-JCE, Vol 14, Issue 3 ,Sep. - Oct. 2013, PP 20-27.

[5] S.Vijayalakshmi V.Mohan, S.Suresh Raja, (2009) "Mining Constraint-based Multidimensional Frequent Sequential Pattern in Web Logs," European Journal of Scientific Research., Vol.36, pp 480-490.

[6] Ratnesh Kumar Jain , Dr. R. S. Kasana1, Dr. Suresh Jain, (July 2009)"Efficient Web Log Mining using Doubly Linked Tree," International Journal of Computer Science and Information Security, IJCSIS, vol. 3.

[7] K. R. Suneetha, and R. Krishnamoorthi,(April 2009)"Identifying User Behavior by Analyzing Web Server Access Log File," IJCSNS International Journal of Computer Science and Network Security, vol. 9, pp. 327-332.

[8] Kobra Etminani, Mohammad-R. Akbarzadeh-T, and Noorali Raeeji Yanehsari,(2009) "Web Usage Mining: users' navigational patterns extraction from web logs using Ant-based Clustering Method," in Proc. IFSA-EUSFLAT '09.

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