A Survey on Web Mining From Web Server Log

Ripal Patel[1], Mr. Krunal Panchal[2], Mr. Dushyantsinh Rathod[3]
[1]M.E., [2,3]Assistant Professor, [1,2,3]computer Engineering Department,
[1,2]L J Institute of Engineering and Technology Ahmedabad, Gujarat, India

Abstract— In this, Web Usage Mining is used to discover interesting patterns in accesses to various Web pages within the Web space associated with a particular server. The Web Usage Mining architecture is divided into two main parts- the first part includes preprocessing and data integration components. The second part includes the largely domain independent application of finally Cleaned data and database file generation. After that Mining process will be done using different web mining techniques.

Keywords: Log Files , Web Log Pre-processing, Data Mining, Web Mining, Web Content Mining, Web Usage Mining, Web Structure Mining

I. INTRODUCTION

In dynamic systems such as the Internet, it is a common practice to periodically record samples of activity. Those samples are then used to characterize the activity in the system and to evaluate new mechanisms to be used in this system. It is called Log files. Web usage mining is the application of data mining techniques to discover interesting usage patterns from web usage data, in order to understand and better serve the needs of web-based applications. Usage data captures the identity or origin of web users along with their browsing behaviour at a web site.

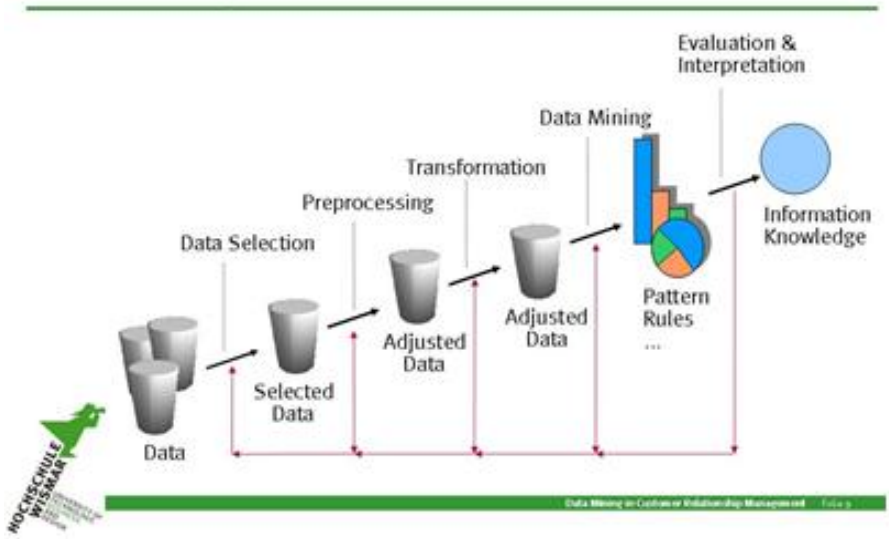

Fig 1: Phases of Data Mining[6]

Data Mining is an iterative process which consist the following list of processes:

- Data cleaning
- Data integration
- Data selection
- Data transformation
- Data mining
- Pattern evaluation
- Knowledge presentation

II. OBJECTIVE

Input raw web accessing log file. Take user choice for normal, multimedia, graphics or e-commerce applications. Merge Mining data from web log file. Read raw web log file and remove logs according to user selection to make intermediate file. Generate more structured and easily readable database file. Use clustering technique of data mining on pre-processed data and do the analysis.

III. TYPES OF WEB MINING

Web mining is a technique to analyze the online Web contents, navigate between various Web sites and perform transaction of data across the Web. Web mining can be broadly divided into three distinct categories, according to the kinds of data to be mined. Figure 2 shows the taxonomy.

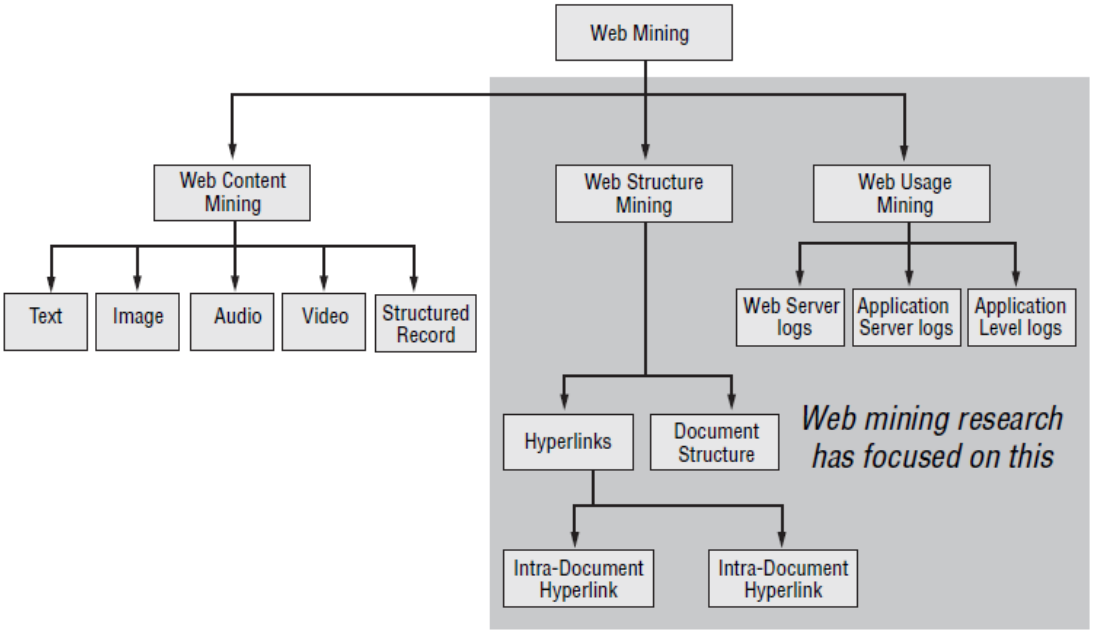

Fig 2: Types of Web Mining [7]

Web Content Mining

Web content mining discovers information or knowledge from millions of sources across the Web. In web content mining, patterns are extracted from online sources such as HTML files, text documents, images, e-books or e-mail messages. The concept of Web content mining is far wider than searching for any specific term or only keyword extraction or some simple statistics of words and phrases in documents. For example, a tool that performs web content mining can summarize a web page so that you need not read the complete document and save your time and energy.

Web Structure Mining

The structure of a typical web graph consists of web pages as nodes, and hyperlinks as edges connecting related pages. Web structure mining is the process of discovering structure information from the web. This can be further divided into two kinds based on the kind of structure information used.

Web Usage Mining [7]

Web usage mining is the application of data mining techniques to discover interesting usage patterns from web usage data, in order to understand and better serve the needs of web-based applications (Srivastava, Cooley, Deshpande, and Tan 2000). Usage data captures the identity or

origin of web users along with their browsing behaviour at a web site. Web usage mining itself can be classified further depending on the kind of usage data considered:

A. Web Server Data

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

B. Application Server Data

Commercial application servers such as Weblogic,1,2 StoryServer,3 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.

C. 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.

IV. LITERATURE REVIEW

Extracting Knowledge from Web Server Logs Using Web Usage Mining[1]:

Use of Internet is increasing day by day. So websites usage growth is also increasing. To maintain the usage data different log files are used with different formats.
Web usage mining was used to discover useful knowledge from web server logs. WUM (Web Usage Mining) worked only for single log file format which is W3C format.

In this paper, they were use Web Usage Mining technique to extract knowledge from web server logs. The process of WUM was divided in four parts, which were: Data collection, Data Preprocessing, Pattern discovery, Pattern analysis.

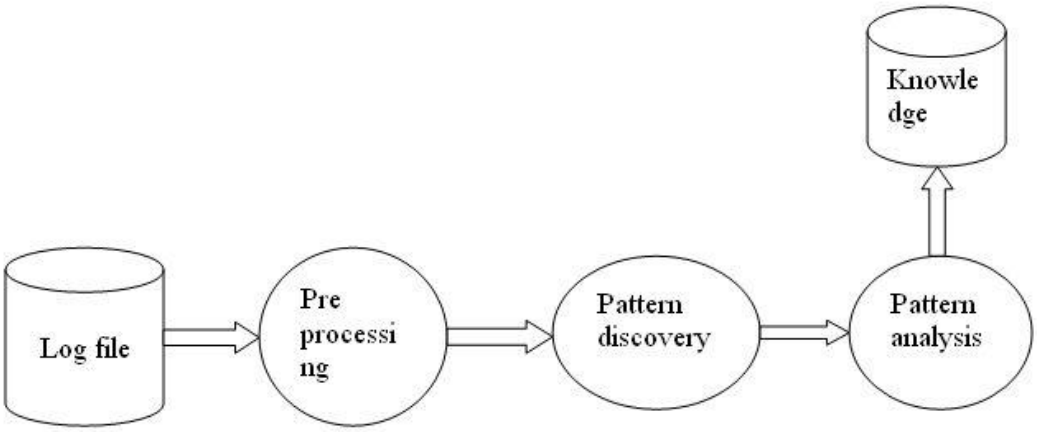

Figure 3: Web usage mining process[1]

WUM worked on unstructured data which is very useful. Figure 2.1 shows whole process of WUM.

WUM worked on single log file format and extracted knowledge was saved in database. It used SQL (Structured Query Language) and OLAP (Online Analytical Processing) for pattern analysis.

An Efficient Web Mining Algorithm for Web Log Analysis: E-Web Miner[2]:

This paper introduces an efficient web mining algorithm for web log analysis. The results obtained from the web log analysis may be applied to a class of problems; from search engines in order to identify the context on the basis of association to web site design of an ecommerce web portal that demands security. The algorithm is compared with its other earlier incarnation called Improved AprioriAll Algorithm.

E-web log miner provides better performance for time and space complexity compared to earlier techniques which was shown through performance analysis of proposed web mining algorithm.
The proposed algorithm, Efficient Web Miner or E-Web Miner can be traced for its valid results and can be verified by computational comparative performance analysis. E-Web Miner reduced the number of database scanning and the candidate sets are found to be much smaller in stage wise comparison with Improved AprioriAll Algorithm of Tong and Pi-lian. E-Web Miner, thus, is successful to be applied in any web log analysis, including information centric network design.

As it was effective in analysis of web data, it used only single log file format and also used database transaction.

Design and Implementation of WEB Log Mining [3]:

Web log mining technology has wide applications in e-commerce and institutes, as the working of internet is increasing.

The design and implementation process of the WEB mining system based on XML are introduced in detail. It uses XML with the relational database. It uses some pre-process methods to analyze Web log, which can recognize the user and conversation accurately.

When logic modelling, it adopts star logic modelling, and designs using a lot of redundant dimension data to improve the information index function.

Optimized Data Pre-processing Technology for Web Log Mining [4]:

In order to solve some existing problems in traditional data pre-processing technology for web log mining, an improved data pre-processing technology is used in this paper.

The identification strategy based on the referred web page is adopted at the stage of user identification, which is more effective than the traditional one based on web site topology.

At stage of Session Identification, the strategy based on fixed priori threshold combined with session reconstruction is introduced. First, the initial session set is developed by the method of fixed priori threshold, and then the initial session set is optimized by using session reconstruction.

Experiments have proved that advanced data pre-processing technology can enhance the quality of data pre-processing results.

Data Pre-processing on Web Server Logs for Generalized Association Rules Mining Algorithm [5]

For ensuring adequate bandwidth and server capacity on IT administrators' organizations website, web log file analysis is used. Data of log file can offer valuable insight into web site usage. Actual usage of web site data in natural working condition, compared with the artificial setting of a usability lab. Activity of users over long period of time can be compared to limited users worked for an hour or two.

Before mining process can perform, first the pre-processing is done on IIS Web Server Logs ranging from the row log file. Pre-processing is depending on the algorithm and purposes of the applications because it is tedious process.

Table 1: Pre-processed Log File [5]

T	ClientIP	Datetime	Method	ServerIP	Port	URI Stem
0	202.185.122.151	11/23/2003 4:00:01 PM	GET	202.190.126.85	80	/index.asp
1	202.185.122.151	11/23/2003 4:00:08 PM	GET	202.190.126.85	80	/index.asp
2	210.186.180.199	11/23/2003 4:00:10 PM	GET	202.190.126.85	80	/index.asp
3	210.186.180.199	11/23/2003 4:00:13 PM	GET	202.190.126.85	80	/tutor/include/style03.css
4	210.186.180.199	11/23/2003 4:00:13 PM	GET	202.190.126.85	80	/tutor/include/detectBrowser_cookie.js

Table 2: Table after data is transferred to database [5]

DataExtracted : Table

TransID	ClientIP	URI_Stem	Status	DateTime	Method
1	192.34.125.98	/tutor/bpg/index	200	2003-11-24 16:0	GET
2	189.23.204.23	/bank/ups/bm/E	200	2003-11-24 19:3	GET

Comparison Table

Table 3: Comparison of Literature Survey

Sr No.	Description	Approach	Pros	Cons
1	Extracting Knowledge from Web Server Logs Using Web Usage Mining	Web Usage Mining process , W3C log file format	Uses unstructured data	Worked on single log file format W3C
2	An Efficient Web Mining Algorithm for Web Log Analysis: E-Web Miner	Improved AprioriAll algorithm	Better log analysis, number of database scanning get	Uses single log file format

			reduced	
3	Design and Implementation of WEB Log Mining	XML and relational database	Simple and practical approach, helps to realize the further mining for Web log data from specific time quantum and specific user	Worked on single log file format, also increases transaction
4	Optimized Data Pre-processing Technology for Web Log Mining	User identification based on referred page, session restructuring algorithm	Improvement of the technology betters the quality of data pre-processing results	Used only IIS log file format and more database transaction
5	Data Pre-processing on Web Server Logs for Generalized Association Rules Mining Algorithm	Association rule mining algorithm	Improves performance of mining system	Uses only one log file format for mining process

V. CONCLUSION

In this survey paper various approaches of Web Mining techniques has been overviewed. There is also a brief discussion about algorithms for efficient mining process. From this survey paper we can conclude that by using Pre-processing we can process the unstructured data. We can use different log files and combine them in one file and then we use the mining on the integrated file. It will decrease the time and increase the efficiency.

References

[1] Eltahir. M.A. , Dafa-Alla, "Extracting knowledge from web server logs using web usage mining", IEEE Aug, 2013, pp.413-417, ISBN: 978-1-4673-6231-3

[2] Yadav, M.P. , Keserwani, P.K. , "An efficient web mining algorithm for Web Log analysis: E-Web Miner", IEEE March, 2012, pp.607-613, ISBN: 978-1-4577-0694-3

[3] Xianjun Ni , "Design and Implementation of WEB Log Mining System", IEEE Jan, 2009, pp.425-427, ISBN: 978-1-4244-3334-6

[4] Ling Zheng, Hui Gui , "Optimized data preprocessing technology for web log mining", IEEE June, 2010, pp.V1-19 – V1-22, ISBN: 978-1-4244-7164-5

[5] Mohd Helmy Abd Wahab, Mohd Norzali Haji Mohd, Hafizul Fahri Hanafi, Mohamad Farhan, Mohamad Mohsin, "Data Pre-processing on Web Server Logs for Generalized Association Rules Mining Algorithm", World Academy of Science, Engineering and Technology 48 2008, pp.190-197, DOI: 10.1.1.140.5102

[6] Tamanna Bhatia, "Link Analysis Algorithm for Web Mining," IJCST Vol. 2, Issue 2, June 2011, ISSN: 2319-5940

[7] Cooley, R, Mobasher, B., Srivastava, J., "Web Mining: Information and pattern discovery on the World Wide Web", IEEE 1997, pp.558-569, ISSN: 1082-3409

[8] Tsuyoshi, M and Saito, K., "Extracting User's Interest for Web Log Data", IEEE 2006, pp.343-346, ISBN: 0-7695-2747-7