# A Novel Technique to Pre-Process Web Log Data Using SQL Server Management Studio

S.Kalaivani[1], Dr.K.Shyamala[2]

1 Research Scholar, PG & Research Department of Computer Science, Dr.Ambedkar Government Arts College, Chennai, India

2Associate Professor, PG & Research Department of Computer Science, Dr.Ambedkar Government Arts College, Chennai, India

***Abstract—*** *Web log data available at server side helps in identifying user access pattern. Analysis of Web log data poses challenges as it consists of plentiful information of a Web page. Log file contains information about User name, IP address, Access Request, Number of Bytes Transferred, Result Status, Uniform Resource Locator (URL), User Agent and Time stamp. Analysing the log file gives clear idea about the user. Data Pre-Processing is an important step in mining process. Web log data contains irrelevant data so it has to be Pre-Processed. If the collected Web log data is Pre-Processed, then it becomes easy to find the desire information about visitors and also retrieve other information from Web log data. This paper proposes a novel technique to Pre-Process the Web log data and given detailed discussion about the content of Web log data. Each Uniform Resource Locator (URL) in the Web log data is parsed into tokens based on the Web structure and then it is implemented using SQL server management studio.*

***Keywords—Web log data, Data Pre-Processing, User access patterns, URL, Mining.***

## I. INTRODUCTION

Web mining is used to discover useful information from Web hyperlink structure, page content and usage data [1]. Web mining uses many data mining techniques which include supervised learning or classification, unsupervised learning or clustering, association rule mining, and sequential pattern mining. Web mining is a kind of data mining process. We can find difference in data collection. In traditional data mining, the data is already collected and stored in the data warehouse. For Web mining, data collection is an important task especially for Web structure and content mining which involves crawling large number of target Web pages.

Web usage mining [9] is partitioned into three widespread phases known as Pre-Processing, pattern discovery, and pattern analysis. Web log data [1] Pre-Processing aims to reformat the original Web logs to identify all Web access sessions. The Web server usually registers all the users' access activities through the Web server logs. This paper is started with the detailed discussion about the log files, then pre-treatment methods were presented which is used to clean the Web robots queries and also discussed about removing queries relating to scripts (".js", ".css", ".swf"), image files etc.,

## II. RELATED WORK

C.P.Sumathi et al. [1] present different steps involved in the Pre-Processing stage. Various heuristics are employed in each step so as to remove irrelevant data and identify users and sessions along with the browsing information. The output of this phase results in the creation of a user session file. Nevertheless, the user session file may not exist in a suitable format as input data for mining tasks to be performed.

There are number of data Pre-Processing techniques. Dipa Dixit et al. [2] they discussed two different approaches for data Pre-Processing: first method based on XML and then the second method based on text file. But the basic algorithm and steps involved in Pre-Processing are considered same for both the approaches.

S.Prince Mary et al. [3] described the importance of Pre-Processing methods and steps involved in retrieving the required information effectively. To use the Web usage mining efficiently, it is important to use the Pre-Processing steps. Steps of Pre-Processing are analysed and tested successfully with sample Web server log files.

## III. WEB USAGE MINING

Web usage mining is used to extract interesting patterns from the Web log data. Web log is an interaction between the user and the Website that automatically recorded in the Web server [5]. Web Usage Mining process is divided into three phases Pre-Processing, Pattern Discovery and Pattern Analysis as shown in figure 1.

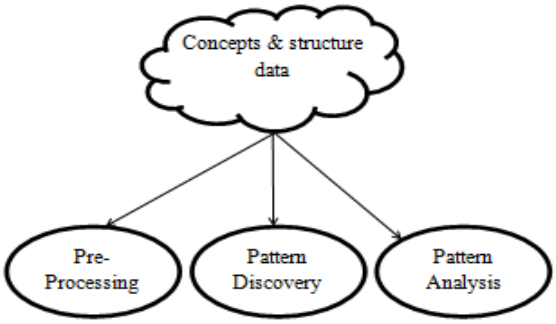


*Fig.1: Phases of Web Usage Mining Process*

**Phase 1 Data Pre-Processing**

Data Pre-Processing [10] is a complex task it takes around 80% of time to do Pre-Process. Data mining techniques cannot be directly applied on the data sets. So, the data Pre-Processing is done to remove inconsistent data, redundant data, and noise data. The steps for data Pre-Processing are Data Cleaning, User Identification, Session Identification and Path completion.

**Phase 2 Pattern Discovery**

Pattern discovery [12] is used to find patterns using data mining techniques like Path analysis, Association Rule, Classification and Clustering. Many different types of graphs can be formed from path analysis. The most obvious is a graph representing the physical layout of a Website where Web pages are nodes and hypertext links between pages are directed edges. Association rules are used for prediction of next event or discovery of associated event. In the Web data set, the transaction consists of the number of Uniform Resource Locator (URL) visits by the client, to the Web site. Applying different association rule mining algorithm, we can predict which are Web pages frequently accessed together by users of Website. Classification is the technique to map a data item into one of several predefined classes. The classifications can be done by using supervised inductive learning algorithms such as Decision tree classifiers, Naïve Bayesian classifiers, k-nearest neighbour classifier, Support Vector Machines etc. Clustering analysis is a technique to group together users or data items (pages) with the similar characteristics.

**Phase 3 Pattern Analysis**

The pattern analysis stage is to analyse the patterns found during the pattern discovery step. For analysing multidimensional data OLAP cube or any visualization tool is used. Knowledge Query management or Intelligent Agents are also used for Pattern Analysis.

## IV. DATA PRE-PROCESSING

Data Pre-Processing [15] is very important task in mining to find efficient patterns and to get efficient result. Data Pre-Processing use log data as input then process the log data and produce the reliable data. The goal of Data Pre-Processing is to remove irrelevant information from the log data.

### 4.1 Collect the Web log data

Web log file contain information about the Website visitors activity. Log files are created by Web servers automatically. Each time when a visitor requests any file (page, image, etc.) from the site information on his request is added to a current log file. There are different forms of Web log file like W3C, NASA and IIS log file. Log files range from 1KB to 100MB [8].

### 4.2 Contents of a Log File

Web log file is a simple plain text file which record information about each user [6]. The basic information present in the log files are:

*User Name*

It helps to identify who had visited the Website. The identification of the user mostly would be the IP address that is assigned by the internet service provider (ISP).

*Visiting Path*

The path chosen by the user while visiting the Website. This may be done using the Uniform Resource Locator (URL) directly or by checking the link.

*Path Traversed*

This identifies the path chosen by the user within the Website using various link.

*Time Stamp*

The time spent by the user in each page while surfing through the Website.

*Page Last Visited*

The page that was visited by the user before he/she leaves the Website.

*Success Rate*

The Success rate of the Website can be determined by the number of downloads made and the number of copying activities done by the user.

*User Agent*

The browser from where the user sends the request to the Web server.

*URL*

The resource accessed by the user. It may be an HTML (Hypertext Mark-up Language) page, a CGI program or a script.

*Request Type*

The method used for information transfer is noted. The methods like GET, POST etc. GET method is the standard request type for a document or program. POST method tells the server that the data is following. The specific level of HTTP protocol is also recorded.

These are the contents present in log file.

### 4.3 Sample Raw Web log

Sample Raw Web log dataset Collected from Makoto Uchida School of Engineering, the University of Tokyo website [14].

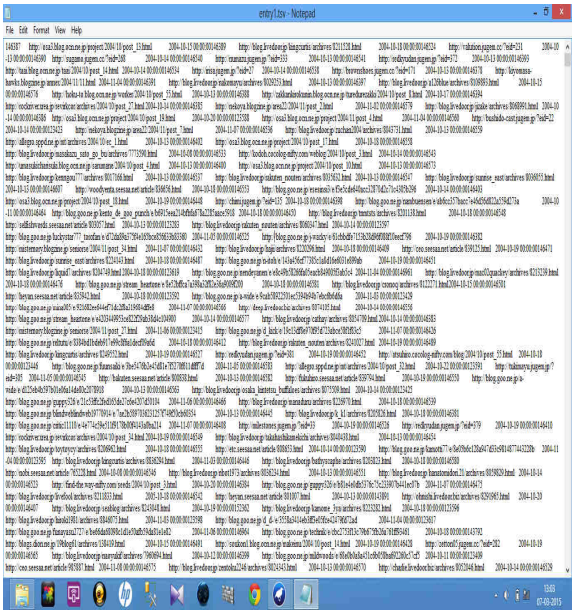

*Fig.2: Sample Web log data*

**An Example from the collected Web log data which is shown in Figure 2.**

146607 http://woodyenta.seesaa.net/article/836656.html 2004-10-18 00:00:00

**4.4 Data Cleaning**

Data Cleaning is the first step in Pre-Processing Web log data. Data cleaning technique is used to find irrelevant, inconsistency, noise data to improve the quality of data [4]. Web server log file contains raw data and it is important to extract the field from the file to remove inconsistent data. Usually log file data are separated using (,) or (“”). Field extraction plays vital role in Pre-Processing where the data will extracted from different fields. This can also be done using Excel or other software which will extract fields and place it in a tabular column. The main objective of Web usage mining is to improve the efficiency of the websites by providing novel methods [7].

***4.4.1 Elimination of local and global Noise:***

Local Noise: This is also known as inter-page noise, which includes unrelated data in the Web page [3]. Local noise includes Decoration pictures, navigational guides, banner etc. Local noise can be removed for efficient result.

Global Noise: Irrelevant objects with high granularities which are larger than the Web page are belongs to global noise. This noise includes replicated Web pages, mirror Web sites and previous version Web pages.

***4.4.2 The Records graphics, video and the format information:***

JPEG, GIF, CSS file name extension is found in the every record on URI field, this can be eliminated from the log file. The files with these extensions are the documents embedded in the Web page. So it is not necessary to include these files in identifying the user interested Web pages [3]. This process support to identify user interested patterns.

***4.4.3 Failed HTTP- status code:***

This cleaning process will reduce the evaluation time for finding the user's interested patterns. In this process, the status field of every record in the Web access log is checked and the status codes over 299 or below 200 are removed.

***4.4.4 Method- field:***

Records which contain methods like POST or HEAD are used to get complete referrer information.

***4.4.5 Robots- Cleaning:***

Robots-cleaning is also known as spider. It is a software tool that scans a Website periodically to mine the content [13]. All the hyperlinks from a Web page are automatically followed by Web Robot. The uninterested session from the log file is removed automatically when the Web Robot is removed.

**4.5 Algorithm for Data cleaning**

*Input: Raw Web log Data*
*Output: Pre-Processed Web log Data*
*Begin*
*Read Web log data from log file*
*If Web log data.url=”*.jpg,*.gif,*”*
*Then*
*Remove records*
*Else*
*Save Records*
*Repeat until last record*
*End*

This algorithm not only cleans the irrelevant data but can also remove the inconsistent and incomplete data. Error request are not in use of mining technique.

## V. IMPLEMENTATION OF DATA CLEANING ALGORITHM

First of all to clean the Web log data, read the Web log file and count all the records. The logic behind that the procedure is to read character by character from a file and compare the character from ASCII value of space and enter key and count all the records from Web log file [11]. The output returns the number of records from a file. Number of entries in raw Web log before Pre-Processing is 2, 01,824.

After counting the total number of records, we have to Pre-Process i.e. clean the collected raw Web log data. In this procedure, first we need to remove the entire suffixes like *.jpg,*.css,*.gif etc. These suffixes are not necessary in a file. The file size is also reduced after cleaning the data. Data cleaning is done using Microsoft SQL Server management studio 2008. First image files, multimedia files and incomplete URL are also removed using SQL query shown in Figure 3. Then the number of entries in Web log data after Pre-Processing is 25,671. After the cleaning process has been done, the Table 1 shows the log data.

*Table.1: Evolution of Log data*

| Web server log file | Result |
|---|---|
| Original data | 2,01,824 |
| Pre-Processed data | 25,671 |
| Noise data | 1,76,153 |

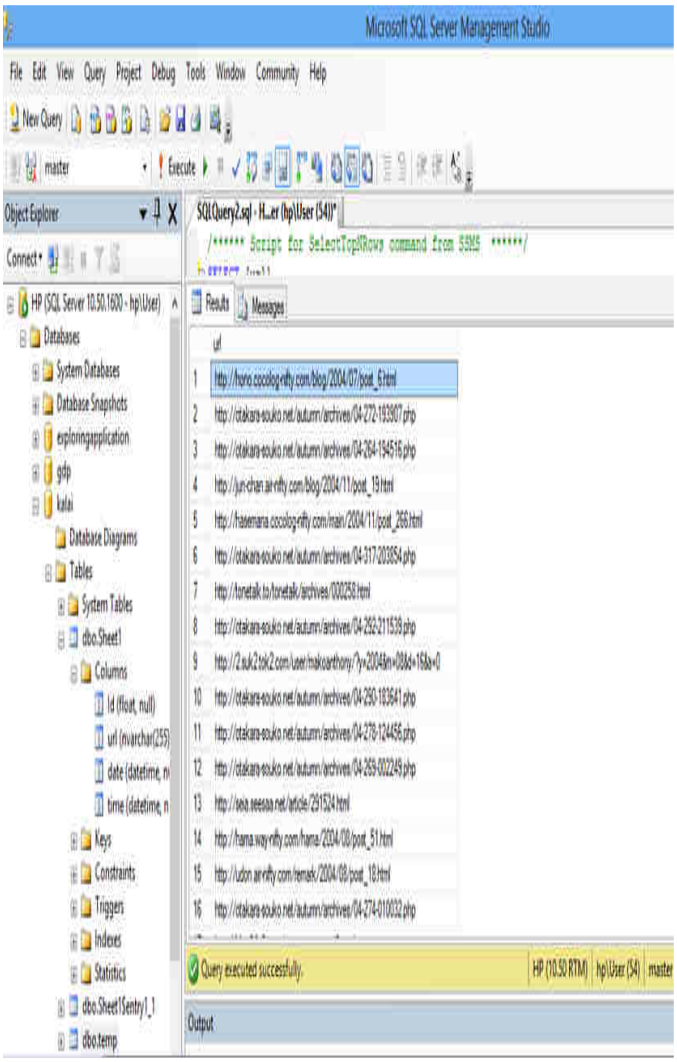


*Fig.3:Pre-Processed Web log data using Microsoft SQL Server Management Studio*

The raw Web log data is Pre-Processed using SQL server queries. The data is reduced and cleaned and it is ready for pattern discovery. Figure 4 represents the data cleaning process.

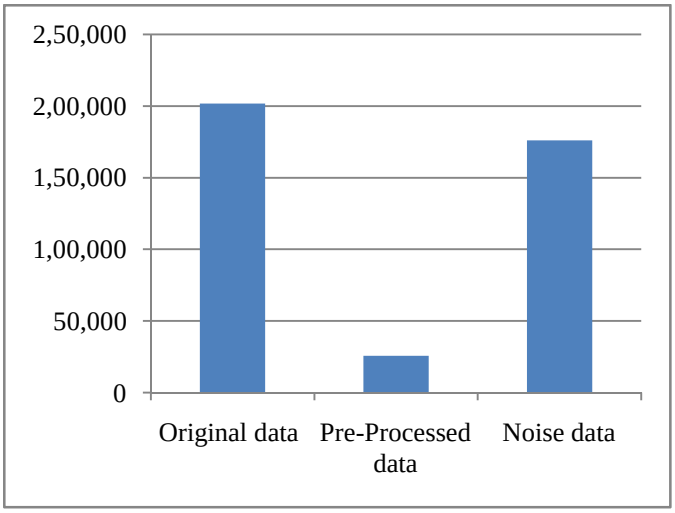


*Fig.4: Process of Data Cleaning*

The graph makes it clear that there is a severe change in the number of records after data cleaning. In General Pre-Processing can take up to 60-80% of the time spending in analysing the data. Incomplete Pre-Processing task can easily result in invalid pattern and wrong conclusions.

## VI. CONCLUSION

Data Pre-Processing is an important step to filter and organize appropriate information before using data mining algorithm. Once Pre-Processing is performed on Web server log, then the patterns are discovered using data mining techniques such as Statistical Analysis, Association, Clustering and Pattern matching on Pre-Processed data. In this research paper raw Web log data is Pre-Processed efficiently using Microsoft SQL server management studio. Web log data size reduced.

## REFERENCES

[1] C.P.Sumathi, R.Padmaja Valli and T. Santhanam, "An Overview Of Pre-Processing Of Web Log Files For Web Usage Mining", Journal Of Theoretical And Applied Information Technology, 31st December 2011. Vol. 34 No.2.

[2] Ms. Dipa Dixit and Ms. M Kiruthika, "Pre-Processing of Web Logs", (IJCSE) International Journal on Computer Science and Engineering, Vol. 02, No. 07, 2010, 2447-2452.

[3] S. Prince Mary and E. Baburaj, "An Efficient Approach to Perform Pre-Processing", Indian Journal of Computer Science and Engineering (IJCSE), ISSN : 0976-5166, Vol. 4 No.5 ,Oct-Nov 2013

[4] Shaily G. Langhnoja, Mehul P. Barot and Darshak B. Mehta , "Web Usage Mining to Discover Visitor Group with Common Behavior Using DBSCAN Clustering Algorithm", International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 7, January 2013.

[5] Mehak, Mukesh Kumar and Naveen Aggarwal, "Web Usage Mining: An Analysis", Journal Of Emerging Technologies In Web Intelligence, Vol. 5, No. 3, August 2013.

[6] L.K. Joshila Grace, V.Maheswari and Dhinaharan Nagamalai, "Analysis of Web Logs and Web User in Web Mining", International Journal of Network Security & Its Applications (IJNSA), Vol.3, No.1, January 2011.

[7] R. Suguna and D.Sharmila, "An Overview of Web Usage Mining", International Journal of Computer Applications (0975 – 8887) Vol.39– No.13, February 2012.

[8] Tyagi, N.K & Solanki, A. K. "An Algorithmic approach to Data Pre-processing in Web Mining", International Journal of Information Technology and Knowledge Management, 2010, Volume 2, No. 2, pp. 279-283.

[9] Shaily Langhnoja, Mehul Barot and Darshak Mehta, "Pre-Processing: Procedure On Web Log File For Web Usage Mining" , International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 2, Issue 12, December 2012.

[10] Aye, TT. "Web log cleaning for mining of web usage patterns", Computer Research and Development (ICCRD), 2011.

[11] Neha Goel, Sonia Gupta and C.K. Jha, "Analyzing Web Logs of an Astrological Website Using Key Influencers", International Research Journal, Vol. 05 No. 01 2015.

[12] Yew Chuan Ong & Zuraini Ismai. "Enhanced Web Log Cleaning Algorithm for Web Intrusion Detection", Recent Advances in Information and Communication Technology Advances in Intelligent Systems and Computing Volume 265, 2014, pp 315-324.

[13] Ankit R Kharwar, Chandni A Naik and Niyanta K Desai, "A Complete Pre Processing Method for Web Usage Mining", International Journal of Emerging Technology and Advanced Engineering.

[14] http://archive.ics.uci.edu/ml/datasets.html.

[15] NeetuAnand and Prof(Dr.)SabaHilal, "Identifying the User Access Pattern in Web Log Data", International Journal of Computer Science and Information Technologies, Vol. 3 (2) , 2012,3536-3539.