DATA PRE-PROCESSING - PRIMARY PHASE OF WUM

[1]Mr. Kirti Kumar Joshi,[2]Mr. Tarun A. Saluja

[1]Assistant Professor,[2]Assistant Professor.
[1]Computer Engineering Department.
[1]Sardar Patel College of Engineering, Anand,India

Abstract : Web has been increasing as a prevailing platform for retrieving information and discovering knowledge from web data. Web data is stored in web server log files. Web usage mining is the process of extracting useful knowledge from web server logs in order to analyze web user's behavior. Web usage analysis requires data abstraction for pattern discovery. This data abstraction can be achieved through data preprocessing. This paper presents how web server log data is preprocesses for web usage analysis.

IndexTerms - Web Mining, Web Usage Mining (WUM), Web Server Log File, Data Pre-Processing, Pattern discovery, Pattern analysis.

I. Introduction

A Web mining has become very vital for effective web site personalization and management. It is crucial for network traffic Flow analysis, creating business services, business support, etc. [1]. Also a Web Mining is the one of the most application of data mining techniques to discover patterns from the web. Web Mining can be classified into three different types, which are Web Content mining, Web structure mining, and Web Usage Mining[2][3].

a)Web Content Mining (WCM): WCM is the method of extracting significant information from the inside of web documents such as text, audio, video, and image thus it is also known as Text Mining [3][4].

b) Web Structure Mining (WSM): WSM is the method of extracting significant information from Web hyperlink structure. HITS (Hyperlink Induced topic search) and Page Rank Algorithms are used in WSM. [3][4].

c) Web Usage Mining (WUM): WUM is the method of extracting usage pattern from Web Log Files. It is also known as web Log Mining [3][4].

II. Web Usage Mining

A Web Usage Mining (WUM) is the method of extracting usage pattern from Web Log Files. It is also known as Web Log Mining. Three basic phases of Web Usage Mining are Data Pre-processing, Pattern Discovery and Pattern Analysis [3] [5].

a) Data Preprocessing:In data pre-processing, sequence of processing tasks are applied on Web log file such as data cleaning, user identification, session identification, and path completion [5] [6].

b) Pattern Discovery:Pattern discovery deals with extracting information from preprocessed data. In this stage, techniques from several research areas, such as data mining, machine learning, statistics, and pattern recognition are examined that are deduced from different fields such as data mining, statistics, machine learning and pattern recognition are applied to web usage data to discover user access patterns of the web. [5] [6].

c) Pattern Analysis:A Pattern Analysis process would remove the unrelated patterns that were generated. They tend to take out the interesting patterns from the output of the pattern discovery phase. There are two most common approaches for the pattern analysis: SQL query mechanism and constructing multi-dimensional data cube to perform OLAP operations [5] [6].

III. PROPOSED ALGORITHM

The Preprocessing Phase includes the Data Collection, Data Extraction, Data Cleaning, User Identification and Session identification which are briefly described below,

a) Data Collection:In Data Pre-processing phase, we have collected data from web usage mining (WUM) processes. The web log file (WLF) is the main source of data for WUM processes. Web log file (WLF) contains information about website visitors, IP-address, host name, Username, timestamp, method, path, protocol, status code and agent information.

b) Log File Extraction:A Log file consists of various data fields that should be separated before applying cleaning process. This process of separating different data fields from single server log entry is identified as data field extraction. This algorithm extracted only those records which length are not greater than 1500 character and only extract until OS information, rest information are ignored in web log file.

c) Web Log File Cleaning:A Web Log file cleaning is the main step in Pre-Processing. This algorithm retains only those data entries in the log file whose status code is 200 and method is GET or POST and file extension is except from js, xml, txt, gif, jpg, pdf, docx and css or retains only those records whose extension is html, php, aspx and jsp. Our algorithm follows some rules for cleaning log file. These rules are as follows:

✓ Symbol like “ , ; ,[,] ,(,) are removed from records.

✓ Records whose suffix (extension) is js, txt, xml, docx, pdf, css and js are removed from log file.

✓ Records with status code 200 are retained and others are removed.

d) User Identification:The cleaning process is followed by finding users process by using user identification algorithm. The users are identified by their IP (Internet Protocol) Address. Our algorithm following the rules to identify users:

✓ If there is new IP address then it is represent new user/client.

✓ If IP address is same but OS (Operating System) is different than its represent new user.

e) Session Identification: In Session Identification, Webpage access of each user divided into individual session. Timeout mechanism is used for identify user session. This algorithm following the rules to identify user session:

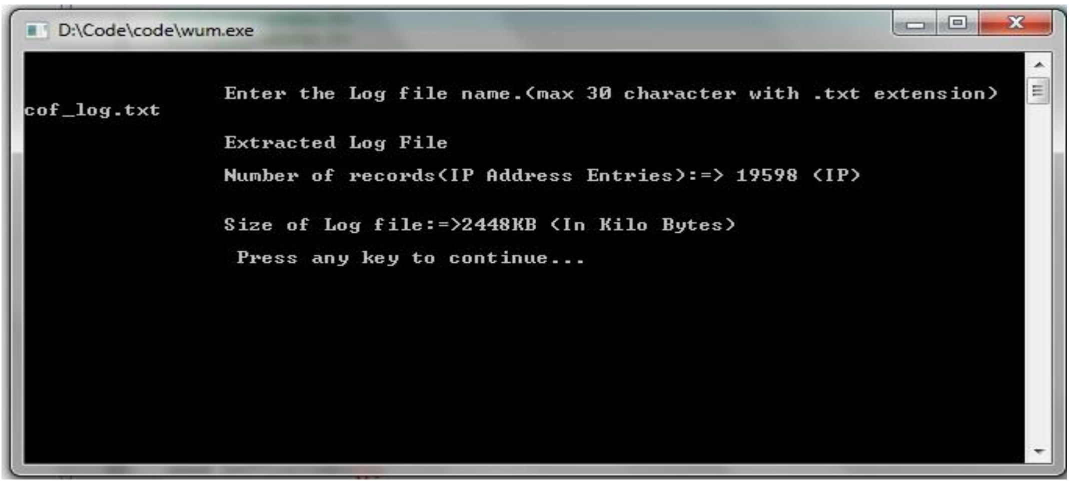
✓ If there is a new IP address (User), there is a new session.

✓ Default 30 minutes timeout taken.

✓ If users access website or web page above 30 minutes then its new session started.

IV. EXPERIMENTAL RESULTS

The size of extracted log file is 2448 KB and consists of 19,598 entries. The result is shown through the snapshot in figure-1. Original log file size is 3863 KB and 19611 numbers of entries, when we have downloaded from server. Apply extraction algorithm then found 2448 KB size of file and 19,598 numbers of entries.

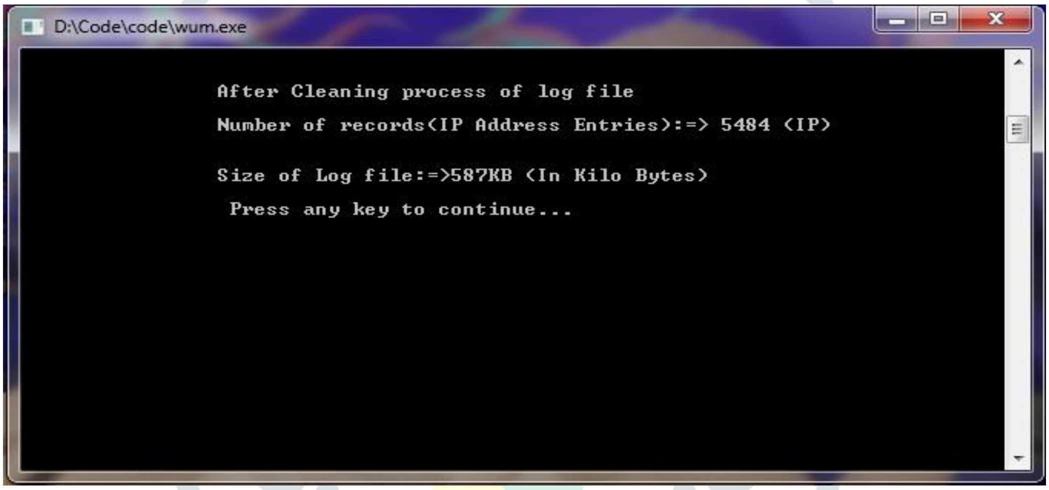


```
                 Enter the Log file name.(max 30 character with .txt extension)
cof_log.txt
                 Extracted Log File
                 Number of records(IP Address Entries):=> 19598 (IP)

                 Size of Log file:=>2448KB (In Kilo Bytes)
                  Press any key to continue...
```

Fig-1. Result of extraction log file algorithm

After cleaning process the size of log file is 587 KB and 5484 entries are left in the cleaned log file. The results are shown through the snapshot in figure-2.

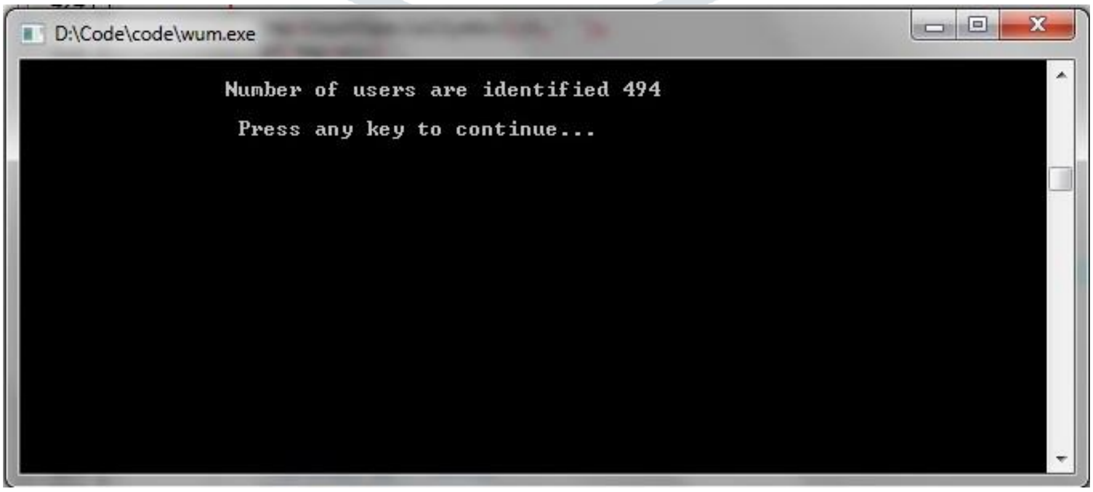


```
                 After Cleaning process of log file
                 Number of records(IP Address Entries):=> 5484 (IP)

                 Size of Log file:=>587KB (In Kilo Bytes)
                  Press any key to continue...
```

Fig-2. Result after cleaning process

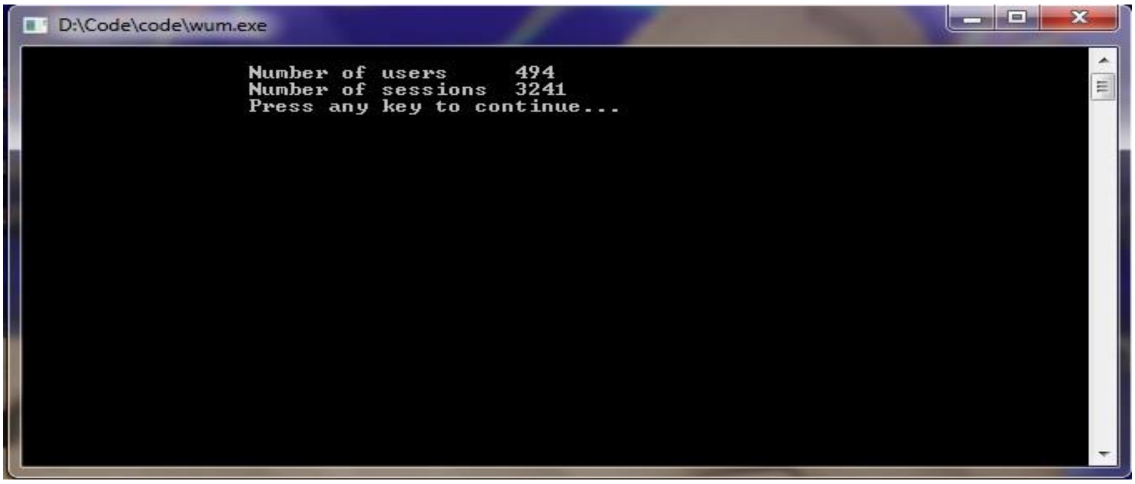
The cleaning process is followed by finding users process by using user identification algorithm. The users are identified by their IP (Internet Protocol) Address. The result is shown through the snapshot in figure-3.



```
                 Number of users are identified 494
                  Press any key to continue...
```

Fig.-3. User Identification

After cleaning process, we are finding user and session of website using user identification and session identification algorithm. We have found 494 users and 3241 sessions in cof_log file. The result is shown through the snapshot in figure-4.



```
Number of users      494
Number of sessions   3241
Press any key to continue...
```

Fig.-4: User and Session Identification

V. RESULT ANALYSIS

There were 19,598 records with 2448 KB file size were present in log file before cleaning process and after cleaning process only 5484 records with 587 KB file size were found in log file. This was shown in Table-1 and graphically present in figure 5.

	No. of Records	Size of file (in KB)
Before Cleaning	19,598	2448
After Cleaning	5484	587
Reduction (in %)	72.02	76.03

Table-1: Comparison of log file before & after cleaning file

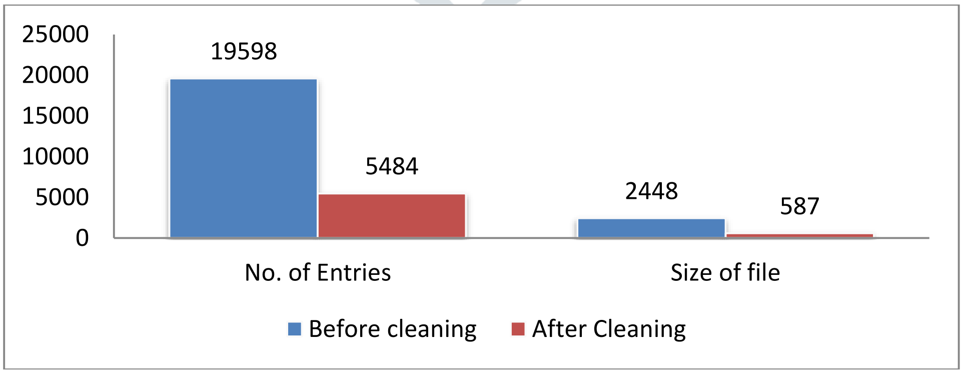


Fig.-5: Comparison of log file before & after cleaning.

We have performed one by one our algorithm and got results.We have applied algorithms on log file and results are tabulated in Table-2.

	Cofmpuat.org
Before cleaning size(KB)	2448
After cleaning size(KB)	587
Reduction in size (%)	76.03
Before cleaning records	19,598
After cleaning records	5484
Reduction in records (%)	72.02
Number of users	494
Number of sessions	3241

Table-2: Result

VI. Conclusion

A Data Cleaning is a significant task in WUM process. The results which were obtained after cleaning contained valuable information about the log files and after cleaning step number of records or size of file are reduced hence increases the worth or quality of the log file and reduced time required for pattern discovery process.

References

[1] MonaS.Kamat, J.W.Bakal, MadhuNashipudi, “Comparative Study of Techniques to Discover Frequent Patterns of Web Usage Mining”, Volume-2, Issue-3, 2013.

[2] SurbhiAnand, Rinkle Rani Aggarwal, “An Efficient Algorithm for Data Cleaning of Log File using File Extensions”, International Journal of Computer Applications, Volume 48– No.8, June 2012.

[3] K.S.R. Paven Kumar, V.V. Sreedhar, ManojChowdary, “A Critique on Web Usage Mining”, International Journal of Computer Science and Information Technologies, Volume-3, Issue-5, May 2012.

[4] ShailyG.Langhnoja,Mehul P. Barot,Darshak B. Mehta, “Web Usage Mining Using Association Rule Mining on Clustered Data for Pattern Discovery”, International Journal of Data Mining Techniques and Applications, Vol 02, Issue 01,June 2013.

[5] Naga Lakshmi, Raja Sekhara Rao, SaiSatyanarayana Reddy, “An Overview of Preprocessing on Web Log Data for Web Usage Analysis”, International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-2, Issue-4, March 2013.

[6] L.K. Joshila Grace, DhinaharanNagamalai, V.Maheswari, “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] Kavita Sharma and Gulshanshrivastva, “Web mining: Today and Tomorrow”, third international conference of Electronics Computer Technology,IEEE, 2011.

[8]Smita Gupta, AishwaryaRastogi, Srishti Agarwal, Nimisha Agarwal, “Web Mining: A Comparative Study”, International Journal of Computational Engineering Research IJCER, Vol.2, Issue No.2, Mar-Apr 2012.