International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue12, December-2014

Algorithms for Web Log Data: WUM Pre-Processing phase

Mansi Yadav
M.Tech (S.E.) Scholar

Pankaj Dalal
Professor

Shrinathji Institute of Technology & Engineering, Nathdwara-313301

***Abstarct* - The age of information is Web cantered. The World Wide Web (WWW) servers are the source of spreading information in world using websites. The users find information through pages of hyper text mark-up language (HTML), PHP or ASP. The arrangement of pages is also one factor for improving accessibility of information. The maximum user hits same pattern of pages or maximum hit pattern arrangements can improve the accessibility. The pattern of pages can be extracted from log file by Web Usage (Log) mining (WUM). WUM extracts pages hit users and pattern information by WLM process. The paper WLM implements through algorithm for data extract, data cleaning process, user information (user's identification and session identification etc.) extract algorithms.**

Keywords* - *Web Mining, Web Usage Mining (WUM), Server Log File, Data Pre-Processing, web page pattern discovery, web page pattern analysis.

I. INTRODUCTION :

Web mining is the application of Data Mining. Data mining is the process of extracting meaningful data from the Warehouse. It is classified into three categories: Web Content Mining (WCM), Web Structure Mining (WSM) and Web Usage Mining (WUM). [1][6]

Web Content Mining (WCM):

WCM is the process of extracting meaningful information from the contents of web documents such as text, audio, video, and image thus it is also known as Text Mining. [2][6]

Web Structure Mining (WSM):

WSM is the process of mining useful information from Web hyperlink structure. HITS (Hyperlink Induced topic search) and Page Rank Algorithms are used in WSM. [2][6]

Web Usage Mining (WUM):

WUM is the process of extracting usage pattern from Web Log Files. It is also known as web Log Mining. [2][6]

II. WEB USAGE MINING:

WUM is the process of extracting meaningful user's access information from Web data. It is used to understand and improved Web based applications to users. WUM is a powerful tool to analyzing, designing and modifying a websites according to user's access patterns. WUM main three phases are Data Pre-Processing, Pattern Discovery and Pattern Analysis. [1]

Figure 1 shows the process of WUM. Data Pre-Processing is the first step to involve Data Cleaning, User Identification, Session Identification and Path completion steps. [1][3]

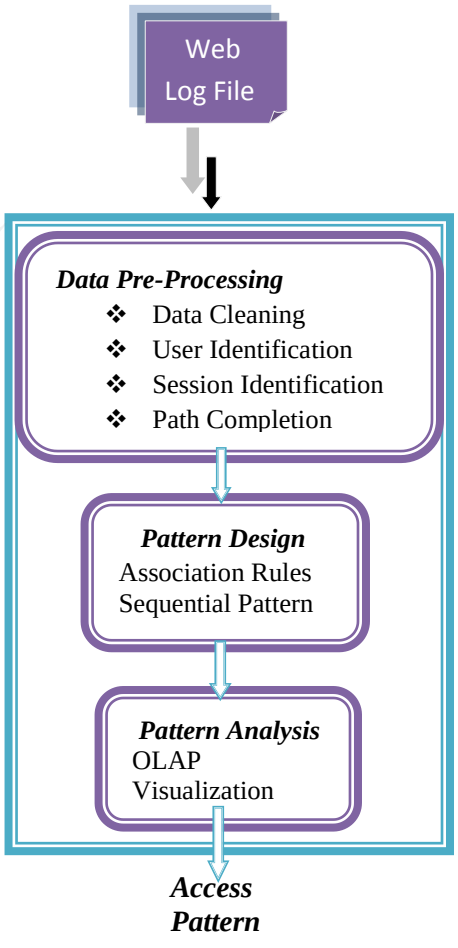

Fig.1. WUM Process

Data Cleaning removes unwanted and irrelevant records which are not used in pattern design process like records that have suffix of css, js, jpeg, png etc and status code except 200. Result of cleaning phase is reduced Log file size and increased the accuracy of log file.[4] User Identification is second steps of Pre-Processing, find website users. New IP-Address represents new User.

Session Identification finds session of particular users default timeout for single session is 30 minutes. Path Completion is last step of Pre-processing, finding complete user access paths and the missing paths are added. [7]

Pattern Discovery is the second phase of WUM process, this phase find out users' access patterns from cleaned log files using different techniques like Sequential Patterns, Association Rules, Clustering and Classification rules. [3][8][9][10]

Pattern Analysis is the final phase of WUM process, this phase removes uninteresting patterns form pattern design and mined most frequent pattern using knowledge query mechanism such as Structure Query Language (SQL) and Visualization Techniques. [3][8][9][10]

III. PROPOSED ALGORITHM

1) Data Collection:

We have collected live six month log data of www.drgoyal.co.in website server. Web log file (WLF) contains information about website visitors, IP-address, host name, Username, timestamp, method, path, protocol, status code and agent information.

Example of Log File Entry:
188.143.232.211 - - [31/Jul/2014:04:54:15 -0500] "GET /msg.php HTTP/1.1" 200 10743 "-" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1)"

2) Web Log File Extraction:

A WLF consists of various data fields. Data field extraction separates data fields before cleaning process. This process of separating different data fields from single server log entry.

The implementation of the Data field extraction algorithm is in C language. The separated fields of the log file are saved into a file. We have calculated log file size and counted number of records.

Algorithm for Log File Extraction

Input: - Web Log File (In text format)
Output: - Extracted Log File (In text format)
Step 1: Open WLF in read mode
Step 2: Open Extracted Log File in write mode.
Step 3: While {
Read WLF until EOF (end of file);
If next line {
Read one line;
Write line in Extracted Log File with #end;
}
}
Step 4: Calculate size of Extracted Log File and number of records.
Step 5: Close both files.

3) Log file Cleaning:

Log file cleaning algorithm retains only those data entries in the log file whose status code is only 200, method is GET or POST and file suffix is except from js, xml, txt, gif, jpg,png and css.

Algorithm for Log File Cleaning
Input: Extracted Log File (In text format)
Output: Cleaned Log File (In text format)

Step 1: Open Extracted Log File in read mode
Step 2: Open Cleaned Log File in write mode.
Step 3: While(read until EOF) {
Read Extracted Log File until EOF (end of file);
If (status code=200 && method GET||POST && suffix != css || xml|| js|| png||jpeg||gif) {
Write record in Cleaned Log File;
}
Else remove records;
}
Step 4: Calculate size of Cleaned Log File and number of records in that.
Step 5: Close both files.

4) User Identification:

Our algorithm follow the below rules to identify users:

If there is new Internet Protocol (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.

Algorithm for User Identification:

Input: - Cleaned log File (In text format)

Output: - Number of users, LogUser File

Step 1: Open WLF //in read mode
Step 2: Open LogUser File //in write mode
Step 3: initialize int ucount=1;
Char OldIP[20],NewIP[20];
Step 4: if OldIP != NewIP
Then
Increment in ucount by one
Write records in LogUser File
Else
Write records in LogUser File
End if condition
Step 5: repeat above step 4 until EOF
Step 6: return number of users
Step 7: Close both files.

5) Session Identification:

Webpage access of each user divided into individual session. Timeout mechanism is used for identify user session.

The rules for identify user session algorithm:

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.

Algorithm for Session Identification

Input: - LogUser File (In text format)

Output: - LogSession File (In text format), number of sessions

Step1: Open LogUser File //in read mode
Step 2: Open LogSession File // in write mode
Step 3: initialize isession, usertime
Step 4: Usertime=TimeDiff(time1,time2)
//TimeDiff function used for finding difference between 2 times(prev and next)
Step 5: if usertime> 30 min
Then
Increment in isession+1
Write in LogSession file
End if
Step 6: repeat above 4 & 5 steps until end of file
Step 7: return isession value
Step 8: Close both files.

IV. EXPERIMENTAL RESULTS

The 1825 KB size of extracted log file having 14225 entries. The figure 2 shows result snapshot.

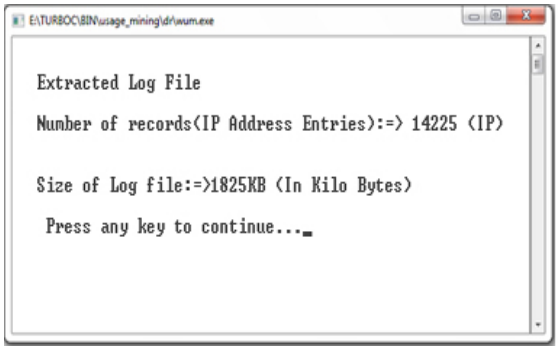

Fig.2. Result after extraction log data

After cleaning the size of log file is 340 KB with 3061 entries are in the cleaned log file. The figure 3 shows result snapshot.

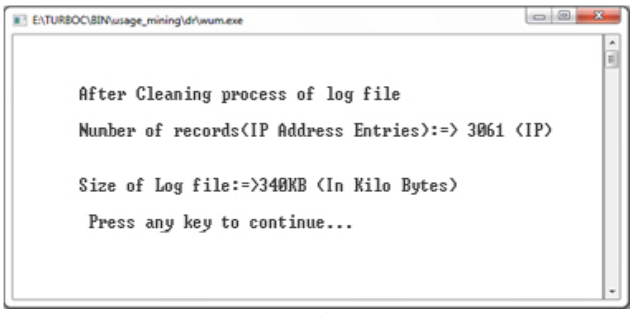

Fig.3. Result after Cleaning

Next, the users and sessions of users are found 325 users and 1831 sessions in drgoyal website log data. The figure 4 shows results snapshot.

Fig.4. User and session identification

V. RESULT ANALYSIS

There were 14,225 records and size of file was 1825 KB before cleaning, after cleaning process only 3,061 records and size of file was 340 KB. This was shown in table 1 and graphically present in figure 5.

Table 1: Comparison in no. of records or Size before and after cleaning log

	No. of Records	Size of file (in KB)
Before Cleaning	14,225	1825
After Cleaning	3,061	340
Reduction (in %)	78.49	81.37

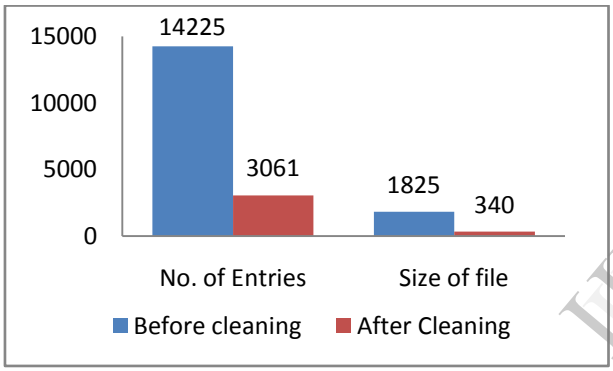

Fig.5. Graph between size and number of records before and after cleaning log

The results of algorithms of drgoyal.co.in log are tabulated in table 2.

Table 2: Result comparison

	Drgoyal.co.in
Duration of log data records	April 1, 2014 to October 1, 2014
Before cleaning size(KB)	1825
After cleaning size(KB)	340
Reduction in size (%)	81.37
Before cleaning records	14225
After cleaning records	3061
Reduction in records (%)	78.49
Number of users	325
Number of sessions	1831

VI. CONCLUSION

This paper, presents a brief overview of WUM. Data Cleaning is an important task in WUM process. We have implemented proposed algorithms for log file extraction, log file cleaning, user identification and session identification. 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 quality of the log file and reduced time required for pattern discovery process.

VII. FUTURE WORK

The access patterns are expected and reduce accessibility time of websites using most frequent pattern algorithm and maximum hits pattern algorithm. Also give our suggested pattern algorithm for improve accessibility time of websites.

REFERENCES

[1] Vijay Kumar Padala, Sayeed Yasin, Durga Bhavani Alanka, " A Novel Method for Data Cleaning and User- Session Identification for Web Mining", Vol. 3, Issue. 5, Sep - Oct. 2013.
[2] Shaily G.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.
[3] Mona S.Kamat, J.W.Bakal, Madhu Nashipudi,"Comparative Study of Techniques to Discover Frequent Patterns of Web Usage Mining", Volume-2, Issue-3, 2013.
[4] Mrs.R.Kousalya, Ms.K.Suguna, Dr.V. Saravanan, "Improving the Efficiency of Web Usage Mining Using K-Apriori and FP-Growth Algorithm", International Journal of Scientific & Engineering Research Volume 4, Issue3, March-2013.
[5] S.Gowri Shanthi, Dr. Antony Selvadoss Thanamani, "Web Page Categorization Using Web Mining" International Journal of Advanced Research in Computer Engineering & Technology, Volume-1, Issue-7, September 2012
[6] Devinder Kaur, Ravneet Kaur, "Minimizing the Repeated Database Scan Using an Efficient Frequent Pattern Mining Algorithm in Web Usage Mining", International Journal of Research in Advent Technology, Vol.2, No.6, June 2014
[7] V.Chitraa, Dr.Antony Selvadoss Thanamani, "A Novel Technique for Sessions Identification in Web Usage Mining Preprocessing", International Journal of Computer Applications, Volume 34– No.9, November 2011
[8] Monika Verma, Shikha Pandey, "An Efficient Algorithm for Frequent Pattern
Mining using Web Analysis Approach", IJCSET, Vol 2, Issue 7, July 2012
[9] K.S.R. Pavan Kumar, L. Manoj Chowdary, V.V. Sreedhar, "A Critique on Web Usage Mining", International Journal of Computer Science and Information Technologies, Vol. 3
[10] L.K. Joshila Grace, Dhinaharan Nagamalai, 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.