Data Preprocessing: A Pre requisite for Web Log Files

Nehal G. Karelia
Dept. of Computer Science
Rajasthan College of Engineering for Women, Jaipur
Rajasthan Technological University, Kota, India

Prof. Shweta Shukla
Assistant Professor, Dept of Computer Science
Rajasthan College of Engineering for Women, Jaipur
Rajasthan Technological University, Kota, India

***Abstract*— Web mining techniques are used to extract knowledge from Web data. Web mining can be broadly defined as the search and measure of useful information from the World Wide Web. Researchers have identified three broad categories of Web mining out of which Web Usage mining technique is used to analyze the results of user's behavior and activity. The Web log files are used to store the click streams of the users. But as the data in the log files are not preprocessed, Web usage mining techniques are used to preprocess the data and organize them into some structured data. Data preprocessing and the storage is an important and challenging research topic in web usage mining. In order to identify which user is visiting which sites and through which browser, requires examining the raw web log files created by the web server. Data preprocessing is an important task used to mine the data from the log. This paper presents an algorithm used to clean the data which is a part of the data preprocessing techniques and an algorithm for its storage for the future usage.**

***Index Terms*—Web usage mining, data preprocessing, data cleaning, user identification.**

I. INTRODUCTION: WEB USAGE MINING

During the olden days that is last 70's and 80's to extract information it was time consuming where in you need to take help of computer as well as books and so on. But now in the present scenario to extract any information it can be done within a fraction of time that is through WWW. World Wide Web is a huge repository of web pages and links [6]. The growth of web is tremendous as approximately one million pages are added daily. Web applications are increasing at an enormous speed and its users are increasing at exponential speed. Users' accesses are recorded in the web log files. In today's era it has become important to know the user access mode. Because of the tremendous usage of the web, the web log files are growing at a faster rate and the size is becoming huge. So to have a relevant data being resulted or analyzed we can take help of the concept which is known as Web Mining. Web mining involves analysis of web server logs of a website whereas data mining involves using techniques to find relationships in large amount of data [13]. Web mining that discovers and extracts interesting knowledge/patterns from web is classified into three types as Web structure mining, Web content mining, and Web usage mining [19]. Web content mining is related to data mining and text mining. Figure1 highlights the three areas.

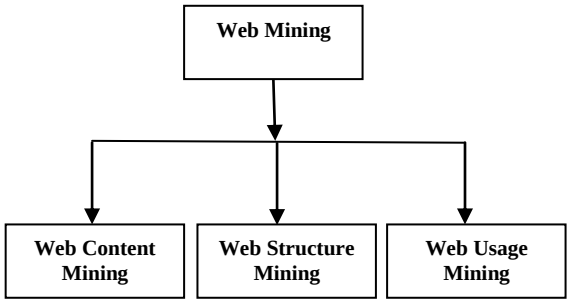

Figure1: Web Mining Structure

Web usage mining is the application of data mining techniques. It is used to discover interesting usage patterns from web data, in order to analyze, understand and better serve the needs of web-based applications. It is the task of discovering the activities of the users while they are browsing and navigating through the Web [8]. It tries to make sense of the data generated by the web surfer's sessions/behaviors. The web content mining and web structure mining utilize the primary data on the web, while web usage mining mines the secondary data derived between web pages and user groups, identification of potential customers for ecommerce, enhance the quality of data and delivery of Internet information services to the end user. It also improves web server system performance, site design and mainly facilitates personalization.

II. WEB LOG INFORMATION

These are logs which maintain a history of page requests. There are many formats available for the log files but the W3C maintains a standard format for web server log files. Most recent entries are typically appended to the end of the file. In these log files the information about the request, including client IP address, page requested, request date/time, HTTP code, bytes served, bytes received, user agent, and referrer are typically added. These data entries can be combined into a single file, or separated into distinct logs, namely access log, error log, or referrer log. However, all user specific information is not collected by the server logs. These files are accessible only to the administrative person or the web-master; it is not accessible to general Internet users. The following is a fragment from the server logs:

```
2011-09-28 02:23:41 W3SVC1 172.172.100.101 GET
/JavaScript/DateTime.js - 80 -
172.172.100.63Mozilla/5.0+(Windows+NT+5.1;+rv:2.
0b6)+Gecko/20100101+Firefox/4.0b6 404 0 3
```

Figure 2: Sample Log file entry

The above entry reflects the information as follows:

- Date used to record the hit done by the user.
- Time used to record the transaction time.
- Client IP address stores the number of the user who accesses the web site.
- Server IP address stores the number of the server.
- Server Port used for the port for data transmission.
- Server Method (HTTP Request) refers to an image, movie, sound, txt, pdf, HTML file and more.
- URI-stem is the path from the host and represents the structure of the web site.
- Status represents the status code returned from the server. 3 digit code indicating the following:
  - 200 series represents Success
  - 300 series represents Redirect
  - 400 series represents Failures
  - 500 series represents Server errors
- Bytes sent used to note down the amount of data revisited by the server.
- Bytes received represent the amount of data sent by the user to the server.

There are also various log file formats like NCSA common log format, IIS log format etc. Additional to the access log we also have the agent log, referrer log and error log to be utilized for various purposes. The remainder of this paper is organized as follows: Section III presents the related work; data preprocessing concepts are discussed in section IV. Data cleaning algorithm and individual user's data storage algorithm is discussed in section V. The experimental results are presented in section VI. Section VII concludes the paper along with future scope.

III. RELATED WORK

In this section, we introduce some related work done in data preprocessing:

Ramya C[1], has explained the concept of data preprocessing technique, which is basically used to collect data from the Web services for further usage. It briefly describes the various steps involved in the data preprocessing like merging, data cleaning, user/session identification and data formatting and summarization.

Theint Aye [2], Mohd Helmy[3], describes the various techniques used to extract information from Web document. It also highlights the structure of web logs and server logs which are basically of four different types: transfer log, agent log, error log, and referrer log. With the help of the structure of the logs we are able to derive useful information needed to identify user's access patterns in the Web. Algorithms are also described which are basically used for the data cleaning and field extraction and transferring the server logs to database.

DeMin Dong[4],have explore a visual Web usage mining method. He has proposed a SQL Server2000 based Web Usage mining solution, which concentrates on the concept of how to use data transfer service and other tools to do work related to data pre-processing and also highlights how to use Online Analysis and Process (OLAP) and data mining to realize mode discovery and mode analysis.

V.Chitraa, Dr.Antony Selvadoss Thanamani[5] starts with the different stages of Web Log Mining- data collection, pre-processing, pattern discovery and pattern analysis. Further taking the concept of pre-processing which is actually used to convert the raw web log into transactions, where in transactions are used to group user's behavior for personalization. This paper suggests another technique for identifying sessions, which will help for extraction of user patterns.

IV. DATA PREPROCESSING

Data preprocessing plays an important role in Web usage mining. It is very complex process and takes 80% of total mining process. Preprocessing is necessary, because log file contain noisy, irrelevant and unambiguous data which may affect result of the mining process. It is an important step to filter and organize appropriate data before applying any web mining algorithm. The aim of data preprocessing is to improve the data quality and increase the accuracy in the mining process.

Preprocessing consists of various phases like data cleaning, field extraction, user identification, session identification. In this paper the main tasks is data cleaning and extract individual user's behavior. So we can summarize the above task in 2 steps:

1) Clean the web log and remove the unnecessary data.
2) Store the individual user's behavior.

A. Data Cleaning: This step cleans, that is it consists of removing all the data tracked in web logs those are useless for mining purposes. The task of the data cleaning is to remove the irrelevant and redundant log entries for the mining process. Basically, there are three kinds of log entries that are irrelevant or redundant which has to be removed. They are as follows:

- •Image entries: A user's request to view a particular page often results in several log entries because that page includes other graphics, while we are only interested in what the users explicitly request, which are usually text files.
- •Status Code entries: HTTP status codes are used to indicate the success or failure of a requested event, and we only consider successful entries with codes between 200 and 299.
- • Entries, with request methods except POST and GET.

B. User Identification: This step identify individual user by using their IP address. The goal of user identification is to identify who access web site and which pages are accessed. If new IP address then there is a new user. If IP address is same but browser version or operating system is different then it

represents different users. Thus as a result in this step the different users are identified and the unique users are distinguished.

C. Session Identification: The goal of session identification is to divide the page accesses of each user at a time into individual sessions. A session can be said as a series of web pages user browse in a single access. As the server logs do not always contain all the information needed it is very complex to identify the sessions from the raw data. A referrer based method is used for identifying sessions that is the referrer information should be taken into care if the IP address, operating system and browsers are same.

V. IMPLEMENTATION

A. Proposed Framework:

Based on the researches that were already conducted in this domain, the data preprocessing step can be divided in two main parts: the classical data preprocessing, where we group the methods commonly used in the literature for preprocessing data, and the advanced data preprocessing containing new ideas for data enhancement for the data mining steps that follow. The role of this preprocessing is to considerably reduce the large quantity of Web usage data available and, at the same time, to increase its quality by structuring it and providing additional aggregated variables for the data mining analysis that will follow.

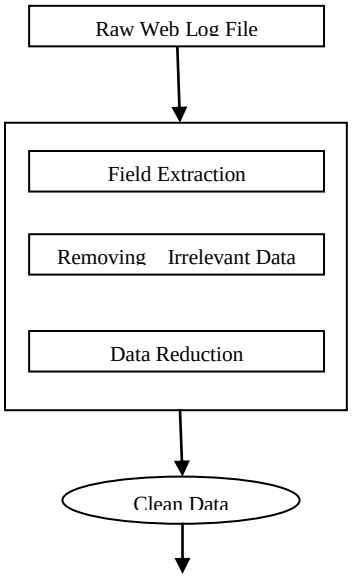

Figure 3: Framework for proposed algorithm

B. Proposed Algorithm:

The technique proposed here for classical data preprocessing involves two steps: data cleaning and data structuring. The solution for WUM preprocessing also adds an advance data preprocessing step which will allow the analyst to select only the information in fewer selected transactions. In the classical data preprocessing we have the steps for data cleaning and in the advance data preprocessing we have the step for data summarizing.

According to above presented framework for data cleaning and structuring the proposed algorithm is as follows:

Data cleaning algorithm:

Input: Web Server Log File
Output: Reduced Log file
Step 1: Read the log record from Web Server Log File
Step 2: If (logrecord.url-stem(css, js, gif, jpeg,jpg,png,) and
(logrecord.method = 'GET') and
(logrecord.sc-status >200 and
logrecord.sc-staus <300)
Then Insert logrecord in to Logfile.
End of If condition.
Step 3: Repeat the above two steps until eof(Web Server Log File)
Step 4: Stop the Process.

User Identification Algorithm:

Input: Log Database file
Output: Unique Users Database text file

Step 1: Initialize
IPList = 0; UserList = 0; BrowserList = 0;
Step 2: Read user.IP from the log file
Step 3: If User.IP not in the file
Display record not found
Else
Insert the data into the text file
End of if condition
Step 5 Stop the process

VI. EXPERIMENTAL RESULTS

We have conducted several experiments on log files collected from a University during September 2011. Our experiments were performed on an IBM ThinkPad Intel(R) core 2 duo(TM) processor 1.77 GHz with 3 GB RAM along with JDK 6.1. Through these experiments, we show that our preprocessing methodology reduces significantly the size of the initial log files by eliminating unnecessary requests and increases their quality through better structuring. The following table shows the actual reduction done to the log files.

Table 1: Experimental Results

Sr. No.	Web Server Log File	Original Size	Size after preprocessing is applied	Reduction in (%)
1	ex110927.log	82 KB	13KB	84.15%
2	ex110928.log	18,831 KB	4564 KB	75.76%
3	ex110929.log	26,809 KB	6131 KB	77.13%
4	ex110930.log	22,594 KB	4871 KB	78.44%
5	ex111001.log	25,830 KB	6810 KB	73.64%
6	ex111002.log	31,723 KB	6458 KB	79.64%
7	ex111003.log	29,784 KB	6044 KB	79.71%

VII. CONCLUSION

Thus we can conclude that data preprocessing is an important task in web mining. With the help of above mentioned algorithms, unnecessary and unwanted data request were cleaned from the log file. Only valid and important

request were regrouped and finally the result were saved in the file. The experimental result shows that by applying the data cleaning technique the actual raw web log file size has been reduced to 80% and the quality of the data also increases.

REFERENCES

[1] Ramya C, Dr. Shreedhara K S and Kavitha G, Preprocessing: A Prerequisite for Discovering Patterns in Web Usage Mining Process, 2011 International Conference on Communication and Electronics Information (ICCEI 2011), 978-1-4244-9481-1/11/$26.00 C 2011 IEEE.

[2] Theint Theint Aye , Web log Cleaning for mining of web usage patterns, 978-1-61284-840-2/11/2011 IEEE.

[3] Mohd Helmy Abd, Mohd Norzali, Data Preprocessing on Web Server log for Generalized Association Rule Mining Algorithm. World Academy of Science, Engineering and technology,48 2008.

[4] DeMin Dong, Exploration on Web Usage Mining and its Application, 5th world Congress on Intelligent Control and Automation, June 15-19,2004,China.

[5] V.Chitraa , Dr.Antony Selvadoss Thanamani A Novel Technique for Sessions Identification in Web Usage Mining Preprocessing , International Journal of Computer Applications (0975 – 8887) Volume 34– No.9, November 2011.

[6] V.Chitraa , Dr.Antony Selvadoss Thanamani An Efficient Path Completion Technique for Web Log Mining, IEEE International Conference on Computational Intelligence and Computing Research 2010. ISBN: 97881 8371 362 7.

[7] Mr. Sanjay Bapu Thakare, Prof. Sangram. Z. Gawali A Effective and Complete Preprocessing for Web Usage Mining , (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 03, 2010, 848-851.

[8] Renáta Iváncsy, and Sándor Juhász, Analysis of Web User Identification Methods, World Academy of Science, Engineering and Technology 34 2007.

[9] Rohit Agarwal, K.V.Arya, Shashi Shekhar, Rakesh Kumar An Efficient Weighted Algorithm for Web Information Retrieval System 2011 International Conference on Computational Intelligence and Communication System.

[10] Doru Tanasa and Brigitte Trousse, Advanced Data Preprocessing for Intersites Web Usage Mining 1094-7167/04 2004 IEEE.

[11] Pranam Kolari and Anupam Joshi, Web Mining: Research and Practice 1521-9615 July-August 2004 Copublished by IEEE CS and the AIP.

[12] P.V.G.S.Mudiraj, B.Jabber,David Raju, Web Mining: An Overview International Journal of Electronics Communication and Computer Engineering Volume 2, Issue 2, ISSN (Online): 2249-071X, ISSN (Print): 2278-4209.

[13] Sonia Sharama Web Mining International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, Volume 2, Issue 4, April 2012)

[14] Priyanka Patil and Ujwala Patil, Preprocessing of Web Server log file for web mining, Proceedings of "National Conference on Emerging Trends in Computer Technology (NCETCT-2012)

[15] Marathe Dagadu Mitharam, Preprocessing in Web Usage Mining, International Journal of Scientific & Engineering Research, Volume 3, Issus 2, February-2012 ISSN 2229-5518

[16] Vijayashri Losarwar, Dr. Madhuri Joshi, Data Preprocessing in Web Usage Mining, International Conference on Artificial Intelligence and Embedded Systems (ICAIES' 2012) July 15-16, 2012 Singapore

[17] A.Pappu rajan and S.P.Victor, Research Article, Features and Challenges of Web Mining Systems in Emerging Technology, International Journal of Current Research, Volume 4, Issue, 07, pp 066-70, July, 2012 ISSN: 0975-833X.

[18] Nirali Honest, Bankim Patel, and Atul Patel, Applying Web Usage Mining to a University Website Access Domain, International Journal of Applied Information Systems (IJAIS) – ISSN: 2249-0868, Foundation of Computer Science, New York,Volume 2-No. 9 June 2012

[19] Suneetha K.R, Dr. R. Krishnamoorthi, Data Preprocessing and Easy Access retrieval of Data through Data Ware House Proceedings of the World Congress on Engineering and Computer Science 2009 vol I WCECS 2009, October 20-22, 2009 San Francisco, USA.

[20] Sheetal A. Raiyani, Shailendra Jain, Enhance Preprocessing Technique Distinct User Identification using Web Log Usage Data.