International Journal of Review and Research in Applied Sciences and Engineering (IJRRASE)
Vol3 No.1. PP 28- 33 March 2013 ISSN: 2231 – 0061X

Pre-processing of Web Logs for Mining World Wide Web Browsing Patterns

Yogish H K[#1] **Dr. G T Raju**[*2]

[#] *Department of Computer Science and Engineering*
Bharathiar University Coimbatore, 641046, Tamilnadu India
[1]yogishhk@gmail.com, [2]gtraju1990@yahoo.com

***Abstract: W*eb usage mining is a type of web mining, which exploits data mining techniques to extract required information from navigational behaviour of WWW users. Hence the data must be pre-processed to improve the efficiency and ease of the mining process. So it is important to pre-process before applying data mining techniques to discover user access patterns from web logs. The main task of data pre-processing is to remove noisy and irrelevant data, and to reduce data size for the pattern discovery phase. This paper mainly focus on the first phase of web usage mining i.e data pre-processing with activities like field extraction and data cleaning algorithms. Field extraction algorithm used for separating the single line of the web log file into fields. Data cleaning algorithm eliminates the inconsistent and unnecessary items in the analyzed data.**

Keywords: Web Mining, Web Usage Mining, Data Pre-processing, Log File.

I. INTRODUCTION

Web mining is mining of data related to the World Wide Web. This may be the data actually present in web pages or data related to web activity [9]. Web data can be classified into following classes:
Contents of actual Web pages, Intrapage structure includes the HTML or XML code for the page i.e linkage structure between web pages, usage data that describe how web pages are accessed by visitors and user profiles include demographic and registration information of user.

The expansion of the World Wide Web has resulted in a largest database that is now in general freely available for user access. The different types of data have to be managed and organized in such a way that various users can access them effectively and efficiently. Hence several data mining methods are used to discover the hidden information from the WWW. Generally, Web Usage Mining consists of three phases: data pre-processing, patterns discovery and analysis. This paper presents two algorithms for extracting fields from line of a log file, data cleaning and finally discusses the goals of web usage mining and the necessary steps involved in developing an efficient and effective web usage mining system.

II. RELATED WORK

R.Cooley et al. 99 have clarified the pre-processing tasks necessary for Web usage mining. Their approach basically follows their steps to prepare Web log data for mining [1].

Mohammad Ala'a Al- Hamami et al described an efficient web usage mining framework. The key ideas were to pre-process the web log files and then classify this log file into number of files each one represent a class, this classification done by a decision tree classifier. After the web mining processed on each of classified files and extracted the hidden pattern they didn't need to analyze these discovered patterns because it would be very clear and understood in the visualization level [8].

Navin Kumar Tyagi observed some data pre-processing activities like data cleaning and data reduction. They proposed the two algorithms for data cleaning and data reduction. It is important to note that before applying data mining techniques to discover user access patterns from web log, the data must be processed because quality of results was based on data to be mined [3]. The paper [4] proposed a new approach to find frequent item sets employing rough set theory that can extract association rules for each homogenous cluster of transaction data records and relationships between different clusters. The paper conducts an algorithm to reduce a large number of item sets to find valid association rules. They used the most suitable binary reduction for log data from web database. G. Castellano et al. presented log data pre-processing, the first step of a common Web Usage

Mining process. In the working scheme of LODAP four main modules are involved namely data cleaning, data structuring, data filtering and data summarization [5].

III. WEB USAGE MINING

Web Usage Mining (WUM) performs mining on web usage data or web logs. A web log is a listing of page reference data; sometime it is referred to as click stream data because each entry corresponds to a mouse click. These logs can be examined from either a client perspective or a server perspective. When evaluated from a server perspective, mining uncovers information about the sites where the service resides. It can be used to design improve the design of the sites. By evaluating a client's sequence of clicks, information about a user or group of users is detected. This could be used to perform prefetching and cashing of pages.

In general WUM consists of three main steps:

i. Data pre-processing
ii. Pattern discovery
iii. Pattern analysis.

During pre-processing step, the raw Web logs need to be reformatted and cleansed. The data recorded in server logs, such as the user IP address, browser, viewing time, etc, are used to identify users and sessions.

After each user has been identified, the entry for each user must be divided into sessions. A timeout is often used to break the entry into sessions.

The tasks performing in the pre-processing step are:

- Data Cleaning: The web log is examined to remove irrelevant information for example the log entries with figures (jpg, gif, etc.) can be removed.
- User Identification: The User identification plays a significant role to identify the distinct and unique users of website. Although users alone play no role in web session clustering, they provide significant information about who the distinctive website users are.
- Session Identification: A session is a set of page references from one source site during one logical period. Historically a session would be identified by a user logging into a computer, performing work and then logging off. Identifying the user sessions from the log file is complex task due to proxy servers, dynamic addresses, and cases where multiple users access the same computer (at a library, Internet cafe, etc.) or one user uses multiple browsers or computers – Raju and Satyanarayana, (2008) [10].

In most of the session identification techniques, 30 minute timeout was taken and transactions made by user with web site are in 30 minutes are grouped as session. Stermsek., et al., (2007) [12] and Raju, and Satyanarayana, (2008) [10] followed the same strategy to identify the sessions as proposed by Yuan, et al., (2003) [13].

The Pattern Discovery step is an important component of the Web mining. Pattern discovery converges the algorithms and techniques from several research areas, such as data mining, machine learning, statistics, and pattern recognition, etc. applied to the Web domain and to the available data [6].

The last step in the web usage mining process is pattern analysis. This process involves the user evaluating each of the patterns identified in the pattern discovery step and deriving conclusions from them.

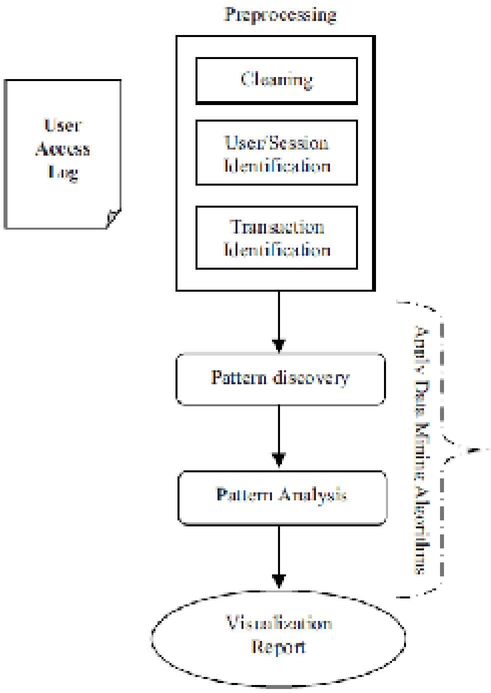


Figure 1. Web usage mining system structure

IV. LOG FILE STRUCTURE

During a user's navigation session, all activity on the web site is recorded in a log file by the web server. The typical Web server logs contains the following information: IP address, request time, method(eg GET),URL of the requested files, HTTP version, return codes, the number of bytes transferred, the Referrer's URL and agents.

The Web logs are often available in two formats: CLF(Common Log Format) and ECLF (Extended Common Log Format). The most basic data fields of ECLF format data contain the client IP, User , Time, Request, Status, BytesRecvd, BytesSent, Process Time, Reference, Agent. Among them, the User has the data only when the request files need to be certificated. Time records the time of issuing files that the server responses to the user request. Request records the method of user request, URL and the used protocol. Status is recorded by the server which shows the response to a request. BytesRecvd records the number of bytes that the users send to the server when they make a request. BytesSent records the number of bytes that the server which processes the request has sent. Reference records the URL which has sent the requests, and when the users enter an address or utilize the bookmark to access it, the reference is empty. Finally, Agent records the operating system and the browser type of users.

The common log file is shown as the following.

```
1007949021.553 3089 192.168.201.11 TCP_HIT/200
12044 GET http://www.computer.org Graeme
DIRECT/64.58.76.99 text/html

1292703446.102 2750 10.100.29.22 TCP_MISS/200
7676 GET http://livescore.com/
DEFAULT_PARENT/2001:d30:101:1::5 text/html
1293006348.196 1156 10.100.29.78 TCP_MISS/200
1003 GET
http://websms.starhub.com/websmsn/usr/chec
kNewMsg.do?
DEFAULT_PARENT/2001:d30:101:1::5
text/html
```

Figure 2. Sample web log data

Each log data record in this format consists of 12 attributes such as Timestamp/ Elapsed /Client /Action/Code/Size Method /URI/Ident/ Hierarchy/From /Content.

Another, storing user IP number or domain name, time and type of access method (GET, POST, etc.) and address of the page being accessed [7]. The fields that have been identified as necessary for the analysis of web usage patterns.

V. PROPOSED SYSTEM

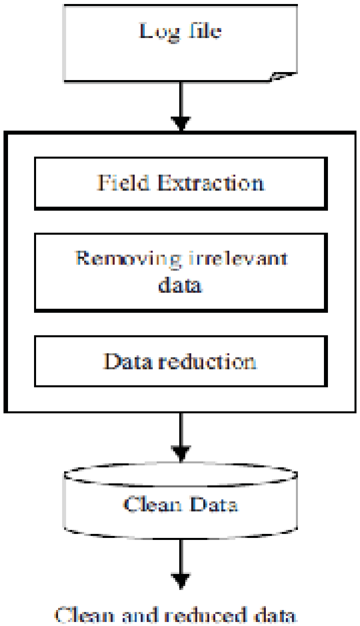


Figure 3. Web log data pre-processing

The results of pre-processing the web server logs are stored in a relational database to facilitate easy retrieval and analysis.

VI. DATA PREPROCESSING

Pre-processing converts the raw data into the data necessary for pattern discovery. The purpose of data pre-processing is to improve data quality and increase mining accuracy. Pre-processing consists of field extraction and data cleansing. This stage is probably the most complex and ungrateful step of the overall process of web usage mining.

This paper describe it shortly and say that its main task is to "clean" the raw web log files and insert the processed data into a relational database, in order to make it appropriate for applying the data mining techniques in the second stage of the web usage mining process.
So the main steps of this phase are:

1. Extract the web logs which are stored in the web server.
2. Clean the web logs and remove the redundant and irrelevant information.
3. Parse the data and put it in a relational database or in the data warehouse and data is used in analysis to create summary reports.

A. Field Extraction

The log entry contains various fields which need to be separate out for the processing. The process of separating field from the single line of the log file is known as field extraction. The server uses different characters as separators. The mostly used separator character is ‘,’ or 'space.'

```
Algorithm:        Field Extraction
Input: Web Log File
Output: Data Base

Begin
    1. Open a DB connection
    2. Create a table to store log data
    3. Open Log File
    4. Read all fields contain in Log File
    5. Separate out the Attribute in the string log
    6. Extract all fields and Add into the Log Table (LT)
    7. Close a DB connection and Log File
End
```

B. Data Cleaning

Data cleaning eliminates irrelevant or unnecessary items in the data base. A web site can be accessed by millions of users. The records with failed HTTP status codes, graphics for example the log entries with figures (jpg, gif, etc.) or sound files can be removed.

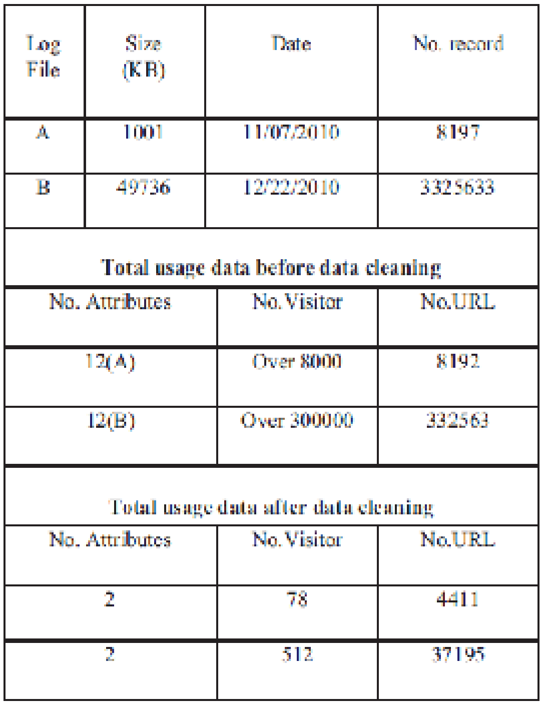
Therefore some of entries are useless for analysis process that is cleaned from the log files. By Data cleaning, errors and inconsistencies will be detected and removed to improve the quality of data [2].

```
Algorithm: For cleaning server logs:
Input: Log Table (LT)
Output: Summarized Log Table (SLT)
‘*’: access pages consist of embedded objects (i.e .jpg, .gif, etc)
‘**’: successful status codes and requested
      methods (i.e 200, GET etc)
Begin
    1. Read records in LT
    2. For each record in LT
    3. Read fields (Status code, method)
    4. If Status code=‘**’and method= ‘**’
    Then
        5. Get IP_address and URL_link
        6. If suffix.URL_Link=
             {*.gif,*.jpg,*.css}
           Then
              7. Remove suffix.URL_link
              8. Save IP_sddress and URL_Link
        End if
      Else
           9. Next record
   End if
End
```

By detecting successful series and method, this algorithm had not only cleaned noisy data but also reduced incomplete, inconsistent and irrelevant requests according to step 4 and 5. Error requests are useless for the process of mining. These requests can be removed by checking the status of request.

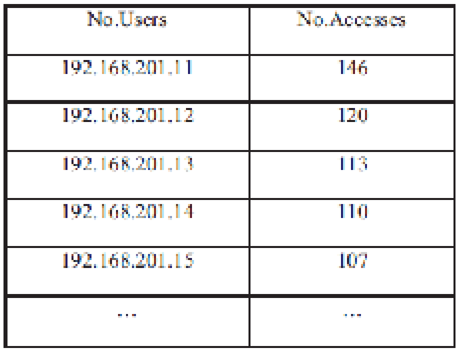
For example, if the status is 404, it is shown that the requested resource is not found or exists. Then, this log entry in log files is removed. Moreover, unnecessary log data is also eliminated URL name suffix, such as *.gif, *.jpg and so on in step 6 and 7. Finally, usefulness and consistent records remain in SLT of database after data cleaning.

TABLE I. OF CONSIDERED FILES



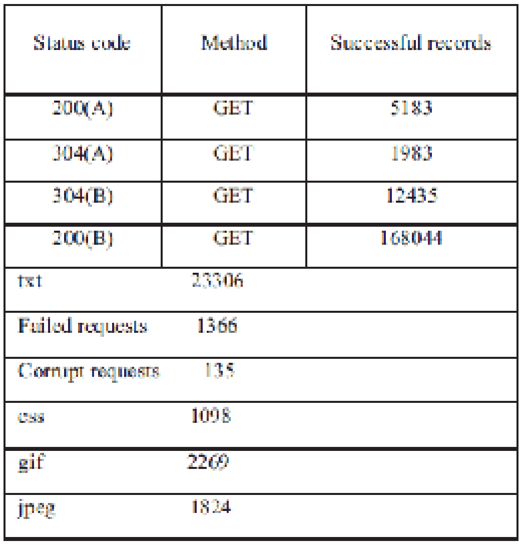
<table>
<tr><th>Log File</th><th>Size (KB)</th><th>Date</th><th>No. record</th></tr>
<tr><td>A</td><td>1001</td><td>11/07/2010</td><td>8197</td></tr>
<tr><td>B</td><td>49736</td><td>12/22/2010</td><td>3325633</td></tr>
<tr><th colspan="4">Total usage data before data cleaning</th></tr>
<tr><th colspan="2">No. Attributes</th><th>No.Visitor</th><th>No.URL</th></tr>
<tr><td colspan="2">12(A)</td><td>Over 8000</td><td>8192</td></tr>
<tr><td colspan="2">12(B)</td><td>Over 300000</td><td>332563</td></tr>
<tr><th colspan="4">Total usage data after data cleaning</th></tr>
<tr><th colspan="2">No. Attributes</th><th>No.Visitor</th><th>No.URL</th></tr>
<tr><td colspan="2">2</td><td>78</td><td>4411</td></tr>
<tr><td colspan="2">2</td><td>512</td><td>37195</td></tr>
</table>

TABLE II. NUMBER OF ACCESSES OF UNIQUE USERS



No.Users	No.Accesses
192.168.201.11	146
192.168.201.12	120
192.168.201.13	113
192.168.201.14	110
192.168.201.15	107
…	…

TABLE III. SUMMARY STATISTICAL REPORT



Status code	Method	Successful records
200(A)	GET	5183
304(A)	GET	1983
304(B)	GET	12435
200(B)	GET	168044
txt	23306	
Failed requests	1366	
Corrupt requests	135	
css	1098	
gif	2269	
jpeg	1824	

VII. CONCLUSION

Data pre-processing is an important task of WUM application. Therefore, data must be processed before applying data mining techniques to discover user access patterns from web logs. The data preparation process is often the most time consuming. This paper presents two algorithms for field extraction and data cleaning. So this system removes irrelevant items and failed requests in data cleaning. After that analysis is performed on the items remaining. Speed up extraction time when users' interested information is retrieved and users' accessed pages is discovered from log data. The information in these records is sufficient to obtain session information.

REFERENCES

[1]. R. Cooley, B. Mobasher, and J. Srivastava, "Data preparation for mining World Wide Web browsing patterns," Knowledge and Information Systems,Vol.1,No.1,1999,pp. 5-32.

[2]. Martinez E. Karamcheti V. "A Prediction Model for User Access Sequence" .In WEBKDD Workshop: Web Mining for usage Patterns and user Profile, July 2002.

[3]. Navin Kumar Tyagi, A.K. Solanki and Sanjay Tyagi: "AnAlgorithmic Approach to Data Preprocessing in Web Usage Mining". International Journal of Information technology and Knowledge Management, Volume 2, No. 2, July-December 2010, pp. 279-283.

[4]. Youquan He, "Decentralized Association Rule Mining on Web Using Rough Set Theory". J ournal of Communication and Computer,Volume 2, No.7, Jul. 2005, (Serial No.8) ISSN1548-7709, USA.

[5]. G. Castellano, A. M. Fanelli, M. A. Torsello. "Log Data Preparation for Mining Web Usage Patterns". IADIS International Conference Applied Computing 2007, pg 371-378.

[6]. José Roberto de Freitas Boullosa. "An Architecture for Web Usage Mining".

[7]. Yan Wang." Web Mining and Knowledge Discovery of Usage Patterns". CS 748T Project. February, 2000.

[8]. Mohammad Ala'a Al- Hamami et al: "Adding New Level in KDD to Make the Web Usage Mining More Efficient".

[9]. Data Mining – Introductory and advanced Topics –Margarate H Dhunham

[10]. Raju. G. T. and Satyanarayana. P. S.,"Knowledge Discovery from Web Usage Data: Complete Preprocessing Methodology", IJCSNS International Journal of Computer Science and Network Security, VOL. 8 No. 1, January 2008.

[11]. Theint Theint Aye aju, "Web Log Cleaning for Mining of Web Usage Patterns", IEEE International Conference on Computer Research and Development(ICCRD 2011)

[12]. Stermsek, G., M. Strembeck, et al. 2007 A User Profile Derivation Approach based on Log-File Analysis. IKE 2007: 258-264.

[13]. Yuan, F., L.-J. Wang, et al. (2003). Study on Data Preprocessing Algorithm in Web Log Mining. Proceedings of the Second International Conference on Machine Learning and Cybernetics, Wan, 2-5 November 2003.

[14]. R. Cooley, B. Mobasher, and J. Srivastava, "Data preparation for mining World Wide Web browsing patterns," Knowledge and Information Systems,Vol.1,No.1,1999,pp. 5-32.

[15]. Martinez E. Karamcheti V. "A Prediction Model for User Access Sequence" .In WEBKDD Workshop: Web Mining for usage Patterns and user Profile, July 2002.

[16]. Navin Kumar Tyagi, A.K. Solanki and Sanjay Tyagi: "AnAlgorithmic Approach to Data Preprocessing in Web Usage Mining". International Journal of Information technology and Knowledge Management, Volume 2, No. 2,July-December 2010, pp. 279-283.

[17]. Youquan He, "Decentralized Association Rule Mining on Web Using Rough Set Theory". J ournal of Communication and Computer,Volume 2, No.7, Jul. 2005, (SerialNo.8) ISSN1548-7709, USA

[18]. G. Castellano, A. M. Fanelli, M. A. Torsello. "Log Data Preparation for Mining Web Usage Patterns". IADIS International Conference Applied Computing 2007, pg 371-378.

[19]. José Roberto de Freitas Boullosa. "An Architecture for Web Usage Mining".

[20]. Yan Wang." Web Mining and Knowledge Discovery of Usage Patterns". CS 748T Project. February, 2000.

[21]. Mohammad Ala'a Al- Hamami et al: "Adding New Level in KDD to Make the Web Usage Mining More Efficient".

[22]. Data Mining – Introductory and advanced Topics Margarate H Dhunham

[23]. Raju. G. T. and Satyanarayana. P. S.,"Knowledge Discovery from Web Usage Data: Complete Preprocessing Methodology", IJCSNS International Journal of Computer Science and Network Security, VOL. 8 No. 1, January 2008.