Web Log Cleaning for Mining of Web Usage Patterns

Theint Theint Aye
University of Computer Studies, Mandalay
theinttheintaye.cu@gmail.com

Abstract—**Web usage mining (WUM) is a type of web mining, which exploits data mining techniques to extract valuable information from navigation behavior of World Wide Web users. The data should be preprocessed to improve the efficiency and ease of the mining process. So it is important to define before applying data mining techniques to discover user access patterns from web log. The main task of data preprocessing is to prune noisy and irrelevant data, and to reduce data volume for the pattern discovery phase. This paper mainly focus on data preprocessing stage of the first phase of web usage mining with activities like field extraction and data cleaning algorithms. Field extraction algorithm performs the process of separating fields from the single line of the log file. Data cleaning algorithm eliminates inconsistent or unnecessary items in the analyzed data.**

Keywords: Web Usage Mining; Data Preprocessing; Log File Analysis.

I. INTRODUCTION

Web mining refers to the use of data mining techniques to automatically retrieve, extract and analyze information for knowledge discovery from Web documents and services. The expansion of the World Wide Web (WWW) has resulted in a large amount of data 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 different users can access them efficiently. Several data mining methods are used to discover the hidden information in the Web. Therefore, the application of data mining techniques on the Web is now the focus of an increasing number of researchers.

Generally, Web Usage Mining consists of three processes: data preprocessing, patterns discovery and patterns analysis.

The paper presents two algorithms for field extraction and data cleaning and then discusses the goals of web usage mining and the necessary steps involved in developing an effective web usage mining system. A brief description of web usage mining is given as well as the data that is required for this analysis.

This paper is organized as follows. Section 2 and 3 discuss some related work and an overview of web usage mining. In section 4 Web Log File Structure is described. Section 5 presents proposed system and data storage. In section 6 data preprocessing activities like field extraction and data cleaning algorithms are presented and conclusion is given in section 7.

978-1-61284-840-2/11/$26.00 ©2011 IEEE

II. RELATED WORK

R.Cooley et al. 99 have clarified the preprocessing 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 preprocess 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 [2].

Navin Kumar Tyagi observed some data preprocessing 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, 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 itemsets 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 itemsets 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 preprocessing, 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) is the application of data mining techniques to discover usage patterns from Web data. In a general process of WUM, distinguish three main steps: data preprocessing, pattern discovery and pattern analysis.

During preprocessing phase, raw Web logs need to be cleaned, analyzed and converted before further pattern mining. The data recorded in server logs, such as the user IP address, browser, viewing time, etc, are available to identify users and sessions. However, because some page views may be cached by the user browser or by a proxy server, we should know that the data collected by server logs are not entirely reliable. This problem can be partly solved by using

some other kinds of usage information such as cookies. 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 following are some preprocessing tasks

(a) Data Cleaning: The server log is examined to remove irrelevant items.
(b) User Identification: To identify different users by overcoming the difficulty produced by the presence of proxy servers and cache.
(c) Session Identification: The page accesses must be divided into individual sessions according to different Web users.

The Pattern Discovery Phase is the key 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 phase in the web usage mining process is pattern analysis. This process involves the user evaluating each of the patterns identified in the pattern discovery phase and deriving conclusions from them. The miner is generally concerned in finding patterns that provide useful information regarding the users' navigation [7].

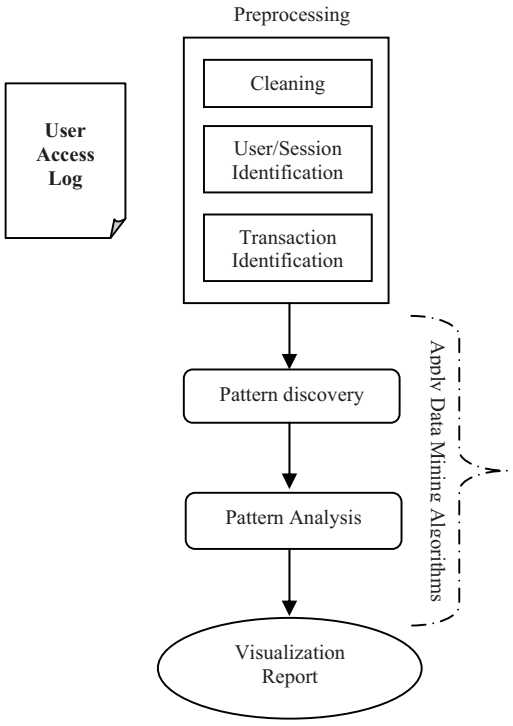

Figure 1. Web usage mining system structure

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. A web server can record user accesses in one of two log formats. Various web servers generate different formatted logs: CERF Net, Cisco PIX, Gauntlet, IIS standard/Extended, NCSA Common/Combined, Netscape Flexible, Open Market Extended, Raptor Eagle [5]. 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

Every log entry records the traversal from one page to 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.

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.

V. PROPOSED SYSTEM

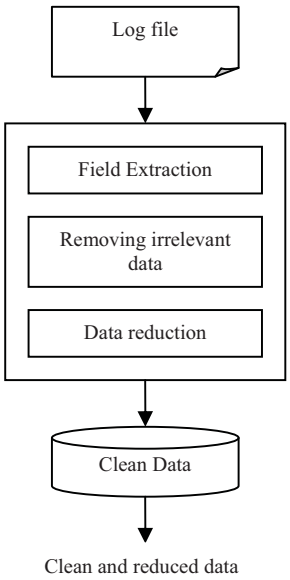

Figure 3. Web log data pre-processing

A. Data Storage

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

VI. DATA PREPROCESSING

Preprocessing converts the raw data into the data abstractions necessary for pattern discovery. The purpose of data preprocessing is to improve data quality and increase mining accuracy. Preprocessing consists of field extraction, data cleansing. This phase is probably the most complex and ungrateful step of the overall process.

This system only 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 to apply the data mining techniques in the second phase of the process.

So the main steps of this phase are:

1) Extract the web logs that collect the data in the web server.

2) Clean the web logs and remove the redundant information.

3) Parse the data and put it in a relational database or a data warehouse and data is reduced to be used in frequency analysis to create summary reports.

B . 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 used different characters which work as separators. The most used separator character is ',' or 'space ' character. The FieldExtract algorithm is given below.

Input: Log File
Output: DB
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

Figure 4. Algorithm for field extraction

C. Data Cleaning

Data cleaning eliminates irrelevant or unnecessary items in the analyzed data. A web site can be accessed by millions of users. The records with failed HTTP status codes also may involve in log data. Data cleaning is usually site-specific, and involves extraneous references to embedded objects that may not be important for purpose of analysis, including references to style files, graphics or sound files. 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 [8].An algorithm for cleaning the entries of server logs is presented below –

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

Figure 5. Algorithm for data cleaning

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 existence. 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. TOTAL OF CONSIDERED FILES

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

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 preprocessing is an important task of WUM application. Therefore, data must be processed before applying data mining techniques to discover user access patterns from web log. The data preparation process is often the most time consuming. This paper presents two algorithms for field extraction and data cleaning. Not every access to the content should be taken into consideration. So this system removes accesses to irrelevant items and failed requests in data cleaning. After that necessary items remain for purpose of analysis. 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.

ACKNOWLEDGMENT

I thank Department of Engineering from University of Computer Studies, Mandalay for providing the Web server user access log files to me.

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

[3] Navin Kumar Tyagi, A.K. Solanki and Sanjay Tyagi: “An Algorithmic 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". Journal 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] Martinez E. Karamcheti V. “A Prediction Model for User Access Sequence" .In WEBKDD Workshop: Web Mining for usage Patterns and user Profile, July 2002.