A New Approach for Rectifying Erroneous Data in Web Usage Mining using Preprocessing

Durgadevi. D[1], Yamini. G[2]
[1, 2] Department of Computer Science and Engineering
[1, 2] Bharathidasan University, Trichy

Abstract- *Web mining is one of the technique of Data mining. Its applying to determine and collect the data from different web sites. Web usage mining (WUM) pre-processing involving to the process of automatic detection in addition to analysis of the patterns in click stream and associated with the collections of data or generate as a result of user communications in Web resources presence on more than one web sites. But these data were stored in web log files and its not avail in an exact picture of the users' access to the Web site.*

A Web usage mining preprocessing technique is the important practice to converting the unrefined data into the abstraction of data . It is necessary for the further process. In this paper focus a first stage of web usage mining preprocessing . An overview of WUM preprocessing techniques is fully aiming to rectify the erroneous log files(or) data ,to improve the data accuracy.

Keywords- WUM-Web Usage Mining, Preprocessing, Erroneous data, Web log files

I. INTRODUCTION

With the probable to explode development of information sources avail on, the WWW and the rapidly increasing pace of taking on to Internet commerce the Internet has evolve into a profitable idea that contains or dynamically generates the WWW and the rapidly increasing pace of taking on to Internet commerce the Internet has evolve into a profitable idea that contains or dynamically generates the information ,that is beneficial to E-businesses the WWW and the rapidly increasing pace of taking on to Internet commerce the Internet has evolve into a profitable idea that contains or dynamically generates the information ,that is beneficial to E-businesses. A web is the most direct link to a company has to its present and possible customers. The companies can study visitor's activities through analysis of web, and discover the patterns in visitor's behavior. The web analysis gives the well to results , when together with company data warehouses, offering great opportunities for the near future.

Web mining is one of the technique of Data mining. Its applying to determine and collect the data from different web sites. Web usage mining (WUM) pre-processing involving to the process of automatic detection in addition to analysis of the patterns in click stream and associated with the collections of data or generate as a result of user communications in Web resources presence on more than one web sites. But these data were stored in web log files and it did not avail in an exact picture of the users' access to the Web site.

Apart from structural information and content information of web site, server logs are also considered as valuable source of information. Every time when a server of a website receives a request from web user, an entry is recorded in log file which is automatically stored and maintained by web server. Web usage mining is a field of study where these log files are analyzed and mined to generate useful patterns.

WUM-PREPROCESSING:

Web usage mining is performing in three steps –

a) Data preprocessing,
b) Pattern Discovery and
c) Pattern Analysis.

Results of the pattern discovery directly influenced the quality of the data processing. Good data sources discover quality patterns and also improve the WUM algorithm. Hence, data preprocessing is an important activity for the complete web usage mining processes and vital in deciding the quality of patterns. In data preprocessing, the collection of various types of data differs not only on type of data available but also the data source site, the data source size and the way it is being implemented. These steps see in detail: The data preprocessing of Web usage mining is usually complex. Purpose of data preprocessing is to offer reliable, structural and integrated data source to pattern discovery.

II. WEB SERVER LOG FILE ANALYSIS

A **server log files** are known as log files .It automatically formed and maintain by a server consisting a list of behavior to be performed.

Erroneous data (or) Logs:

A web server log files which are preserve in page history requests. The W3C maintains a standard format. It s called as the Common Log Format used for web server log files, but other proprietary formats exist and more latest entries are typically appended to the file's end . The Information concerning to the requests that includes client IPaddress,request date/time, page requested, HTTP code, bytes to be served, user-agent, and referrer are classically added. These raw data can be pooled into a single file called as abstraction of data. This files could be separated into distinct logs, such as an access log, error log, or referrer log. It called as erroneous data or log.

Web log file locations:

Web Server : The log files provide information on the most accurate and complete data on the web server use. The log file does not save visited pages in cache. The log file data is sensitive information, thus the web server keeps them closed. [12]

- The client browser: The log file can reside in the client browser. HTTP cookies are used for the client browser, cookies are information generated by a Web server and stored in the computer of the user, for use in future access. [12]

Types of file Logs

All operations performed by the server are stored in log files that provide a detailed record of server activity. The logs can be stored in different file types [12]:

- Access logs:

Data of all incoming requests, and information on the server clients. Access log files; record all requests that are processed by the server.

- Error logs:

It keeps track of incidents in the dialogue with the server (eg wrong URL, interrupted transfer ...);

- Referential logs:

It indicates the site and the page of origin and arrival;

- Agent logs:

It caches information about user equipment (eg. characteristics of the browser, operating system ... etc.).

The log file is a simple text file that stores information about each user.

Log files can be in three different formats:

- W3C extended format (Extended log file format)
- NCSA common format
- IIS log file format

The data recorded in both formats of the NCSA log file and IIS are fixed while the W3C format allows the user to select the properties.

Web Log Files Format:

The format of the W3C log file is the format by default for IIS log file. This is most the format commonly used because it is flexible and allows you to store more information than other formats (that is to say that you can specify the information that you want to record). [12]

Figure 3 shows an excerpt of the log file that contains the following fields:

# Software - version of IIS running
# Version - the format of the log file
# Date - date and time of the first log record.
# Fields (explained in Table 1):: date time c-ip cs-username s-ip cs-method cs-uristem cs-uri-query sc-status sc-bytes cs-bytes time-taken csversion cs (User-Agent) cs (Cookie) cs (Referent).

```
#Software: Microsoft Internet Information Services 7.5
#Version: 1.0
#Date: 2012-11-19 04:36:21
#Fields: date time s-sitename s-computername s-ip cs-
method cs-uri-stem cs-uri-query s-port cs-username c-ip cs-
version cs(User-Agent) cs(Cookie) cs(Referer) cs-host sc-
status sc-substatus sc-win32-status sc-bytes cs-bytes time-
taken
2012-11-19 04:36:21 W3SVC1 DARSHAK 172.16.1.252
GET / - 80 - 172.16.1.247 HTTP/1.1
Mozilla/5.0+(Windows+NT+6.1)+AppleWebKit/537.11+(K
HTML, )+Chrome/23.0.1271.64+Safari/537.11 - -
172.16.1.252 200 0 0 1324 367 6334
2012-11-19 04:36:21 W3SVC1 DARSHAK 172.16.1.252
GET /itInfo/Images/login.jpg - 80 -
172.16.1.247HTTP/1.1Mozilla/5.0+(Windows+NT+6.1)+A
ppleWeb
Kit/537.11+(KHTML,)+Chrome/23.0.1271.64+Safari/5
37.11 - http://172.16.1.252/ 172.16.1.252 200 0 0 20819
361 79
```

Fig: Fields of log files

B. The NCSA Common Log File Format

NCSA format [9] [10] [12] is based on a fixed ASCII text format, so you cannot customize it. This is a smaller version of W3C, In this type if no directive specifies a different format, access logs are recorded in CLF (Common Log Format). The NCSA Common format is available for websites and SMTP and NNTP services, but it is not available for FTP sites. The NCSA common log file format records the following data:

1. The domain name or Internet Protocol address (IP) of the calling machine
2. The name and the HTTP user login (in case of access with a password)
3. The date and time of the request,
4. The method used in the request (GET, POST,…) and the name of the requested resource (the URL of the requested page).
5. The status of the request ie the query result (success, failure, error,...)
6. The size in bytes of the requested page.
7. The browser and operating system used by the client.

A line from a log CLF is presented below: 116.203.228.15 - - [01/Dec/2013:18:02:24 +0000] "GET /webmail/?_task=mail&_action =keep-alive&_remote=1&_unlock=0&_=1381425498 988 HTTP/1.1" 200 33 "http://www.fstbm .ac.ma/webmail/?_task=mail&_id=1877860343 529b4ac71ed12&_action=compose" "Mozilla/5.0 (Windows NT 6.1; rv:24.0) /20100101 Firefox/24.0"

Fig:3.4 NCSA common log file format

HTTP / 1.1: is the protocol used.
200: data about status of requested page (200 to "available", 404 for "not found" ...).
Indeed, the status code, integer coded on three numbers, has a specific meaning in classes depends on the first digit:

- 1xx indicates only an informal message.
- 2xx indicates success.
- 3xx redirects the client to another URL.
- 4xx indicates a client-side error.
- 5xx indicates a server-side error.
- 33: is the charged size.

http://www.fstbm.ac.ma/webmail/: the reference page, the page from which the query is run.

Mozilla / 5.0 (Windows NT 6.1; rv: 24.0) / 20100101 Firefox / 24.0: The last data block shows information about user configuration. Here the visitor uses the Mozilla browser on a Windows NT 5.0 environment.

III. EXISTING PROBLEM

Problem exist in log file

The extensive use of web caches also presented a problem for log file analysis. If a person revisits a page, the second request will often be retrieved from the browser's cache, and so no request will be received by the web server. This means that the person's path through the site is lost. Caching can be defeated by configuring the web server, but this can result in degraded performance for the visitor and bigger load on the servers.

IV.PROBLEM SPECIFICATION

Problems specific to file data LOGS

Although the data provided by the Logs files are useful, it is important to consider the inherent limitations of these data in their analysis and interpretation. Some of the challenges that can occur [9] [10]:

1) Useless requests

Each time the server receives a request, it records a line in the log file. Thus, to load a page, there will be as many lines in the file as the numbers of objects contained on this page (graphics). Pretreatment is essential to remove unnecessary queries.

2) Firewalls

These network access protections mask the IP addresses of users. Any connection request from a server with such protection will have the same address, regardless of which user is. So it is impossible, in this case, identify and distinguish visitors from the network.

3) Web Caching

To ease traffic on the Web, a copy of some pages is saved in the local browser of the user or at the proxy server in order to not download them every time a user requests them. In this case, a page can be accessed several times without any access to the server. As a result, the corresponding requests are not recorded in the log file [9] [10].

4) Use of robots

Web yearbooks, known as search engines use robots that travel all the websites to update their search index. In doing so, they trigger queries that are stored in all log files for different sites, distorting their statistics

5) Users Identification

The identification of users from the log file is not a simple task. In fact, using the log file, the unique identifier available is the IP address and the "agent" of the user. This identifier has several limitations:

- Single IP Address / Multiple server sessions:
  Same IP address can be assigned to multiple users accessing Web services through a single proxy server.
- Multiple IP addresses / single User:
  A user can access the web from multiple machines.
- Several agents / single User:

A user, who uses more than one browser, even if the machine is unique, realized as multiple users.

6) Sessions identification

All requests from a user identified constitute its session. The beginning of a session is defined by the provenance of user to the site. However, no signal indicates the disconnection from the site and consequently the end of a session.

7) Lack of information

Log file does not bring anything about the behavior of the user between queries: Is it really reading the page displayed? In addition, the number of visits of a page does not necessarily reflect the interest of it. In fact a high number of visits may simply be attributed to the organization of a site and the forced passage of a visitor in others.

Classical Methods of WUM Preprocessing

It consists of four processes:

- Data cleansing(or)Scrubbing,
- User identification,
- Session identification,
- Path completion .

Pattern discovery is the key process of the Web mining, which covers the algorithms and techniques from several research areas, such as data mining, machine learning, statistics and pattern recognition. The techniques such as statistical analysis, association rules, clustering classification, sequential pattern and dependency modeling are used to discover rules and patterns. The knowledge that can be discovered is represented in the form of rules, tables, charts, graphs, and other visual presentation forms for characterizing, comparing, predicting, or classifying data from the Web access 1og. The final stage of the Web usage mining is pattern analysis. The aim of this process is to extract the interesting rules or patterns from the output of the pattern discovery process by eliminating the irrelative rules or patterns. Here we focus on data preprocessing method of WUM.

V. PROPOSED WORK

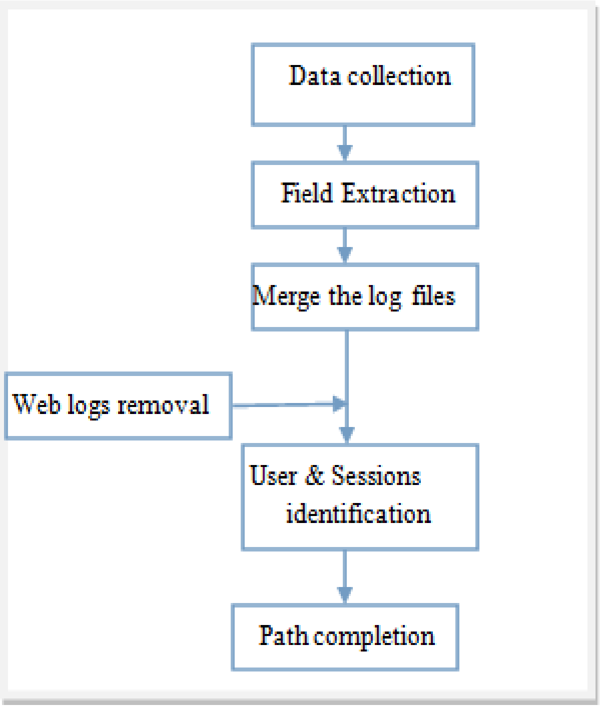

Fig: Proposed work flow

Collection of Data:

Collecting the data from the various sites. That log files contains the following information

- website visitors,
- IP-address,
- host name,
- Username,
- timestamp,
- method,
- path,
- protocol,
- status code and
- agent information.

3.6.2 Extracting the field:

A server log file consists of various data fields that should be separated before applying cleaning procedure. This process of separating out different data fields from single server log entry is identified as data field extraction. A server uses different characters such as a comma or a space character which works as separators. The algorithm proposed below for data field extraction uses the space character as a separator to separate the fields of the log file.

Merge the log files: (Data Fusion)

To reduce the load on a particular server, many servers are used; a user can come from multiple servers or Web application.

Even before starting the cleaning process, it must merge the different log files. [11] The fusion of log files of several web servers refers to the fusion of all the data they contain the queries of all log files were put together in one file.

Data cleansing:

Data cleansing is mostly use to removing extraneous references to embedded objects that may not be important for the purpose o f analysis, including references to style files, graphics, or sound files . Some information should not provide useful information in analysis or data mining tasks then Data cleaning is used. Remove erroneous references.

The first steps of data preprocessing to remove log files are useless requests. Typically, the process concerning non-images, multimedia files, page style files, JavaScript files, etc. Data cleanup and displays the web robots off their requests as requests to delete resources analyzed.

By filtering out useless data, we use log files to reduce storage space to facilitate the coming actions can reduce the size. For example, by filtering out the image requests, we at 50% of its original size to cut down the web server log files.

Algorithm:

Web log file removal algorithm

```
Input: unrefined web log file.
Output: removing erroneous log file.
1. for each lines in web log file do
2. if length of line is more then one character then #Avoid
   Blank Lines
3. if line does not start with '#' then #Avoid Comments
4. if link name contains domain name then #Consider
   Application specific links only
5. if page extension is aspx or html then #Eliminate non-
   page links like images, pdfs ,etc.,
```

Identification of user and session:

A user session identified one or more sessions over the web servers. The target user clicks (click stream) means a delimited set of individual sessions each user accesses the page divide. The methods to identify user session include timeout mechanism. The following is the rules we use to identify user session in our experiment:

- If there is a new user, there is a new session;
- If the time between page requests exceeds a certain limit (30 or 25.5mintes), it is assumed that the user is starting a new session

Algorithm : (For New Version)

```
Algorithm: Identification user && sessions
Inputs: preprocessed web log file
Output: identified users and sessions
For each record in dataset

1. If currentIP is not in ListOfIP Then
   add currentIP in ListOfIP
   mark it as a new user and new session
   assign a new userID and a new sessionID
2. Else if currentOS if not in ListOfOS Then
   add currentOS in ListOfOS
   mark it as a new user and a new session
   assign a new userID and a new session
3. Else if currentBrowser if not in ListOfBrowser Then
   add currentBrowser in ListOfBrowser
   mark it as a new user and a new session
   assign a new userID and a new sessionID
4. Else
mark the current record with its existing userID and sessionID
End if
End for
END
```

Improved algorithm for user , visit, session :

```
Algorithm: User, Session & Visit Identification
Input: processed weblog file
Output: Identified User, Session & Visit.
BEGIN
For each record in dataset do
```

```
1. If currentIP is not in ListOfIP Then
       add currentIP in ListOfIP
       mark whole record as a new user and session
       assign a new sessionID and userID
2. Else if currentOS is not in ListOfOS Then
       add currentOS in ListOfOS
       mark whole record as a new user and session
       assign a new sessionID and userID
3. Else if currentBrowser is not in ListOfBrowser Then
       add currentBrowser in ListOfBrowser
       mark whole record as a new user and session assign a
       new sessionID and userID
4. Else mark current record with existing sessionID and
       userID
   End If
5. If User and Session are well identified (userID,
   sessionID)
   if current record timestamp is more than 1800
   seconds #30minutes * 60 seconds
       mark whole record as a new visit
       assign a new visitID
   Else mark current record with existing visitID
   End If
   End If
   End For
   END
```

Path Completion:

Another critical step in data preprocessing is path completion. There is some reason that the path that results in incompletion, for example, local cache, the cache agent, "post" technique and the browser's "back" button to access some important log file accesses can result in not entered, and There are number of Uniform Resource Locators (URLs) entered in the log can be less than the real one. Use of local caching and proxy server path to meet the production difficulties because users use the server logs record without leaving any local caching or proxy server caching the pages can use. As a result, user access paths using incomplete web logs are preserved. The user to search the travel patterns, user access route missing pages should be attached. Path for the purpose of completion of this task is complete. Better results of data pre-processing, we mined patterns to improve the quality of the algorithm running and will save time. This is especially important for Web log files, in relation to the structure of the Web log files of database or data warehouse as data are not the same. They are not structured and because of various causations are completed. So this particular web usage mining Web log files to the former process is necessary. Data pre-processing, through web logs and data structure, which is easy as it can be transformed mining.

VI. EXPERIMENTAL RESULTS

To effectiveness and efficiency of methodology mentioned above, with valid we have to use web server logs. October, 2004 Initial data source for this experiment to October 2004, which size is around 150 KB. After data cleaning, the number of requests declined from 800 to 300 and the file size is 50 KB.

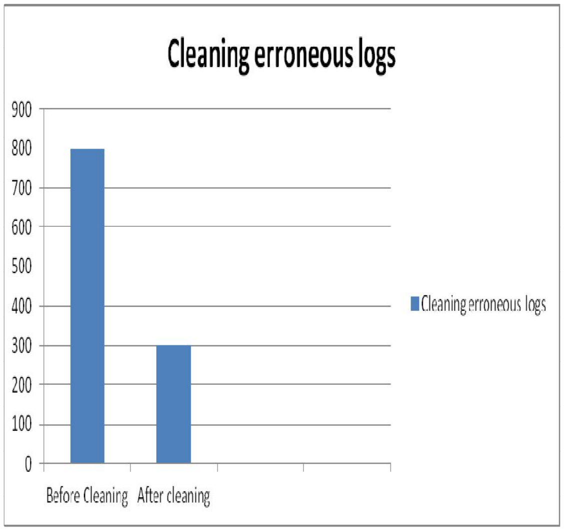

Fig: Bar chart for erroneous logs cleaning

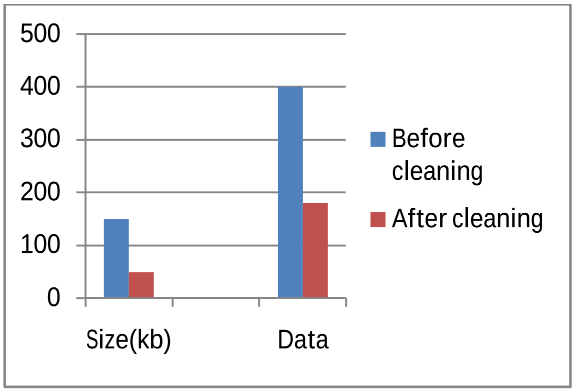

Fig: Bar chart for reducing the data size

This Web Usage Mining preprocessing in order to design and apply them easily at every stage of data to give some rules. An experiments are important to us and the practices effectiveness data preprocessing estimates. **This not only reduces log file size, but also increases the quality of data available.**

VII. CONCLUSION

Web Usage Mining (WUM) Preprocessing is important stage to the results' quality of data preprocessing influences the results of patterns discovery directly. So, data preprocessing is particularly important for the whole Web

Usage Mining processes and the key of the Web Usage Mining's quality .In the present work, an attempt would be made to improve the quality of data. In this present approach done by algorithms .It removes the erroneous and noise data.

Once preprocessing stage is performed, we can apply data mining techniques like clustering, association, classification .The applications of web usage mining such as business intelligence, e-commerce, e-learning, personalization and so on.

REFERENCES

[1] Bamshad Mobasher, "A Web Usage Mining", http://maya.cs.depaul.edu/~mobasher/webminer/survey/node6.html. 1997.

[2] Li Chaofeng , "Research and Development of Data Preprocessing in Web Usage Mining ,"

[3] Rajni Pamnani, Pramila Chawan , " Web Usage Mining: A Research Area in Web Mining "

[4] Andrew Shen , "Http User Agent List", http://www.httpuseragent.org/list/

[5] Andreas Staeding , "User-Agents (Spiders, Robots, Crawler, Browser)", http://www.user-agents.org/

[6] "Robots Ip Address", http://chceme.info/ips/

[7] Aye, T. T. (2011, March). Web log cleaning for mining of web usage patterns. InComputer Research and Development (ICCRD), 2011 3rd International Conference on (Vol. 2, pp. 490-494). IEEE.

[8] Pamutha, T., Chimphlee, S., Kimpan, C., & Sanguansat, P. (2012). Data Preprocessing on Web Server Log Files for Mining Users Access Patterns.International Journal of Research and Reviews in Wireless Communications (IJRRWC) Vol, 2.

[9] Merzoug, N., & Bessa, H. Application du processus de fouille de donnees d'usage du web sur les fichiers logs du site cubba.

[10] Charrad, M. (2005). Techniques sd'extraction de connaissances appliquees aux donnees du Web. Transformation, 56, 5-2.

[11] Tanasa, D., & Trousse, B. (2003). Le prétraitement des fichiers logs web dans le "Web Usage Mining" multi-sites. Journées Francophones de la Toile (JFT'2003), 113-122.

[12] Langhnoja, S., Barot, M., & Mehta, D. (2012). Pre-Processing: Procedure on Web Log File for Web Usage Mining. International Journal for Emerging Technology and advanced enfineering, 2(12)

[13] "Volatile Graphix, Inc.",http://www.iplists.com/nw/.