International Journal of Technology Research and Management
ISSN (Online): 2348-9006
Vol 1 Issue 3 June 2014

# An Efficient Way for Web Log Pre-Processing Towards Web Usages Mining

Monti Babulal Pal
Department of Computer Science & Engineering, Shri Vaishnav Institute of Technology and Science Indore, India
monti_pal@yahoo.co.in

***Abstract--****Now this day WWW is a strong communication media for sharing globally distributed information .In order to enhanced service response, web usages mining play an important role. Web log file is an important source of information for web uses mining but raw web log file contain extraneous object which lead inefficient mining result .So every web log file need to pre-process before use as input for web mining result .This paper gives an brief view over log file preprocessing and proposed an efficient preprocessing mechanism that lead better mining result towards web personalization.*
***Key words:-*** *Web usages mining , web log file , log file pre processing*

## 1. INTRODUCTION

Over last decade World Wide Web has grow to be an imperative medium for communication namely for storing, sharing and distribution of information globally.WWW is treated as global information service centre for widely distributed information like news ,e-commerce, advertisement. Thrombus enhancement in employment of WWW with time leads to make difficulty in information retrieval process via WWW which tends to make user bore and impatience while waiting for information due to network delay. This situation would be resolve efficiently by Web mining.

Web mining is the extension of data mining research [1] in the Web environment. It aims to automatically discover and extract information from Web documents and services [2]. However, Web mining is not merely a straightforward application of data mining. Web mining [2] that discovers and extracts interesting knowledge/patterns from Web is classified into three types as shows in figure:1 on the bases of knowledge /pattern that's its discover as Web Structure Mining that focuses on hyperlink structure, Web Contents Mining that focuses on page contents as well as Web Usage Mining that discover the activities of the users while they are browsing and navigating through the Web.

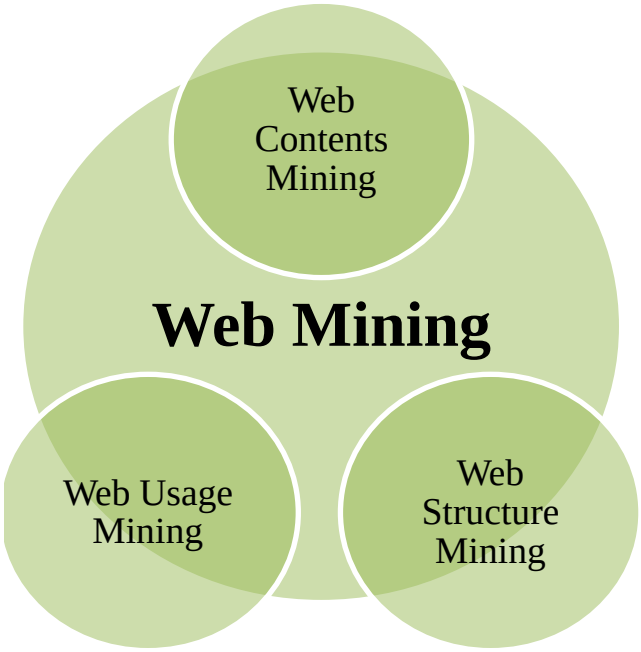


Figure 1: Classification of Web Mining

This paper, concern about Web Usage Mining (WUM) that focuses on Web log as source data and try to categorize navigation preference of end user in order to enhance quality of services and personalization of web portals. Procedure of Web usages mining contain three phases shows in figure: 2 data preprocessing, pattern detection, and pattern analysis. Data pre-processing process encompasses log unification, data cleaning, user and session identification and transaction identification. Pattern detection deals with extraction of knowledge from

preprocessed data by using pattern discovery technique namely Association rules, Classification, Clustering etc. Pattern Analysis filters out uninteresting rules or patterns from the set found in the pattern discovery phase.

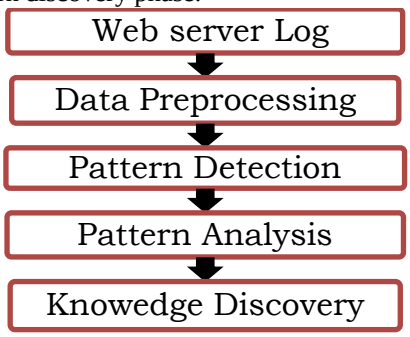


Figure 2: Web Usages Mining Phases

## 2. WEB LOG FILE

Web Log files are excellent sources for determining the client behavior over web portal. Logs are a collection of log entries and each entry contains information related to a specific event that has taken place over the network [3]. Log files are stored on the server side, on the client side and on the proxy servers. Web server logs record the user's information of accessing the site. The typical Web server logs contain the following information: IP address, request time, method (e.g. GET) , URL of the requested files , HTTP version , return codes , the number of bytes transferred , the Referrer's URL and agents. However, the data in Web logs isn't precise because of the existence of local cache, proxy servers and firewalls, along with that web log file also safer by null value, noise and unwanted data that leads bad result. Because of that web log file need to be pre-processes in such way that output of the conversion can be used as the input of the web mining algorithms.

## 3. LOG FILE PRE PROCESSING

Now this day WWW have multiple and heterogeneous server and log file generated at each server is affected by noise, null value and inconsistent data therefore there is need of refinement of log data to improve web mining result and web personalization [8]. In order to refine Web log data to get better data mining result, log pre processing suggest a series of operations on the original Web log files such as the log centralization and data cleaning, user and session identification, data integration and so on as show in figure 3

**Log centralization**

With the enrichment and extension of Web content use over the world, leads to make difficulty in information retrieval process via WWW to overcome this problem web services have employed the automated multi-server load balancing architecture. Because of that logs scattered and stored in different servers and need to periodically synchronize to the log server through the certain means [4].

**Data cleaning**

Web log use to gather information regarding each HTTP request whether some of information is not important and unwontedly increase the data responsible for web mining decision. Data cleaning is use to eliminate such irrelevant information. Irrelevant information is site specific generally contain extraneous reference for embedded object like style, graphics, and video, CSS, JS, picture and sound file[6].

Data cleaning process also involves removal of unwanted failed HTTP status code. Status code is three digit code returns by server. Server serve status code in four different classes namely as explain in table.

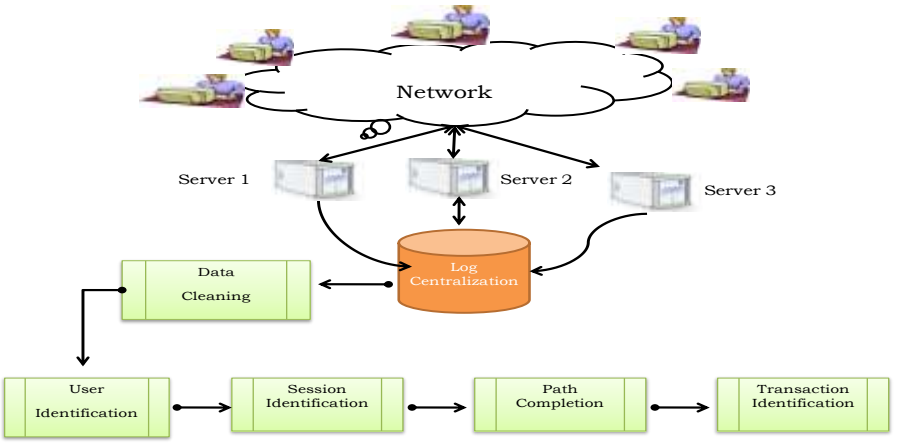


Figure 3: Architecture of Log Pre-Processing

**Table 1: Server Status Code**

| Class | | Meaning | |
|---|---|---|---|
| 200 Series | | Success | |
| 300 Series | | Redirect | |
| 400 Series | 401 | Failed | Failed authentication |
| | 403 | | Forbidden request for restricted subdirectory |
| | 404 | | File not found |
| SOD Series | | Server Error | |

Data cleaning contains also responsible to remove null value, noise and inconsistent data. The inconsistencies of data lead to the reduction of credibility of the data mining results. The data cleaning removes the noise or irrelevant data, and also processes the missing data field in the data.

**User identification**

Task of indivisible user identification over web is very important towards web personalization and categorization of end user. Generally user identification is done through IP address but some time its face due to presence of caching, firewall and proxy server for example due to presence of proxy server number of different user that communicate via that proxy server return same IP address , to overcome this situation user identification is carried out by using some more attribute like browses information ,operating system and Refer URI field as per concern[9].

**Session identification**

The session identification is use to identify user's access records in a single session which is carried out by Refer URI (cs_referer), a new user session is identified if the URL in the ReferURI - field hasn't been accessed since log duration A timeout mechanism tends to be used to divide the session. If the difference between the two-page request times exceeds certain limits, then the user needs to start a new session. Many web applications use 30 minutes as the default timeout.

**Path completion**

Path Completion should be used acquiring the complete user access path during any predefine session. Path added or path completion is the process of arranging the page accesses in one session, if there is a request from the page X, then, the page X is added as the source of this request. If a user uses a number of pages to reach to the final page, then the last page before the final page becomes the source page and it is referred to as the Referrer domain.

**Transaction identification**

Each user session can be seen as composed of multiple transactions, a transaction is a group of a certain semantic history data. The task of transaction identification is to break a large transaction down into several smaller ones or combine the small transactions into a large one. So the main methods of transaction identification

contain segmentation and consolidation[10]. A major transaction identification method is to find the session prior to the path (maximal forward references, MFP), each MFP is a transaction. MFP is defined as a group prior to the browsed page. The request page is not the visited page, "back" refers to the accessed page in the history of the user session prior to the visit. There will be new pages added to the traversal path, while "back" does not extend to the user's access records.

## 4. RELATED WORK

In recent year research towards web personalization is done by web uses mining [1,2,3] and give an idea of web pre-fetching. It also discuss about the web object pre-fetching. Some authors describe web mining techniques. They explain how the web mining works. What is the taxonomy of web mining? As the web mining uses in order to pre-fetch the web page. Web mining is the integration of information gathered by traditional data mining methodologies and techniques with information gathered over the World Wide Web. Web mining allows you to look for patterns in data through content mining, structure mining, and usage mining. There are other papers presents the idea to use the web log. How the web log can use with other method in different purposes. Here we see that web log is a very important file to investigate the crime. Similarly it can useful to pre-fetch the web page in efficient manner. A server log is a log file (or several files) automatically created and maintained by a server of activity performed by it. But unfortunately log pre processing received very less attention [4,5]. Ravindra Gupta et al [6] present an Methods for user identification, session identification, and path completion based on F-P tree. In another work T. Revathi et al present and describe different pre processing phases [7].

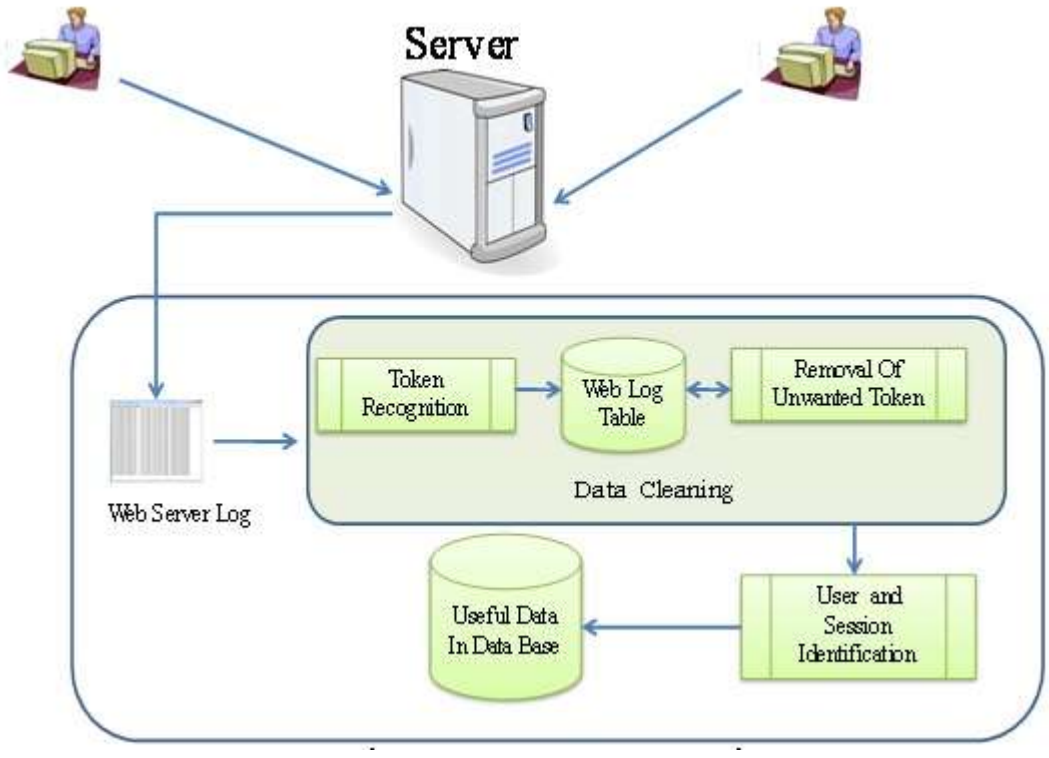


Figure 4: Web Log Pre-Processing

## 5. PROPOSED MODEL FOR LOG PREPROCESSING

Proposed model for data pre processing having two different phases namely data cleaning and user and session identification which is describe below.

Step 1. Data Cleaning

Data Cleaning is responsible for elimination of irrelevant data from web server log file to enhance the durability of web mining result. Every line of web log is consider as stream of alphanumeric and special character where alphanumeric character abstract the information regarding web accessing and special symbol(separator) like comma( ; ),bracket ( [ ] ),backward ( / ) and forward line ( \ ) use to maintain an border between attribute of log file. In data cleaning phase of log preprocessing in proposed model have two sub phase first one is token recognition in which character stream between any two separators in each line is consider as a token file and stored in web log table of data base. Whereas in second phase ie token elimination use to remove unwanted row form web log table. Row with file name suffixes such as gif, jpeg, GIF, JPEG, jpg and JPG or having status code anything else then 200(success) is taking under the decision of elimination.

*Assumption*

*N=total number of line in log file*

*M= total number of attribute in log file*

*WL=web log file*

*T= token in log file*

*For (i=1;I<=n:i++)*

*{*

*For (j=1;j<=m:j++)*

*{*

$T_j^i = \frac{dW_{L_i}}{dj} =$
$character\ stream\ between\ two\ separtor$ ( ; ,/,\ , −, [, ])

*Insert* $T_j^i$ *in web log table in database as Ith touple and Jth attribute*

*{*

*}*

*For (i=1;I<=n:i++)*

*{*

*For (j=1;j<=m:j++)*

*{*

*If (* $T_j^i$ *contain style, graphics, and video, CSS, JS, picture and sound file extension or contain status code above 200 )*

*{*

*Delete Ti record*

*Exit ()*

*}*

*}*

*}*

## Step 2. User and Session Identification

Once web log file store in database then user and session identification is one of major task toward log pre processing .In user and session identification of log pre procession phase, proposed model user a table unique user that maintain distinct user list. Every row of web log table is evaluate one by one and here initially IP address of every is consider as distinct user and insert into unique user table but if there is conflict it is break on the basis of combined attribute (IP and OS) then after (IP ,OS and browser ) and at last (IP , OS , Browser and Referral URI) for every 30 minute of time and remove all duplicate irrelevant entries.

*For (i=n;i<=n;i++)*

*{*

$if\ (T_{IP}^i\ not\ in\ list(\ distinct\ User))$

$Add\ T^i in\ list\ (distinct\ user)$

$else\ if\ (T_{IP\ and\ \ os}^i\ not\ in\ list(\ distinct\ User))$

$Add\ T^i in\ list\ (distinct\ user)$

$else\ if(T^{i}_{IP\ ,os\ and\ browwser}\ \ not\ in\ list(\ distinct\ User$

$))$

$Add\ T^{i} in\ list\ (distinct\ user)$

$else\ if(T^{i}_{IP\ ,os,browwser\ \ and\ referal\ uri}\ not\ in\ list(\ distinct\ U$

$))$

$Add\ T^{i} in\ list\ (distinct\ user)$

$Else$

$delete\ T^{i}\ from\ weblog\ table$

$\}$

$\}$

$\}$

## 6. Results Analysis

For experimental verification of proposed model a scenario of 20 clients and one server has been created and server generate web log file on the basis of client and server communications. Where raw log file contain irrelevant data like style, graphics, and video, CSS, JS, picture and sound file as show in figure 5. Whereas pre processed log file don't have such type of detail as show in figure 6.

```
127.0.0.1 - - [08/Jul/2013:00:36:47+0530] GET /sand/main/about.php HTTP/1.1" 200 523
127.0.0.1 - - [08/Jul/2013:00:36:49+0530] GET /sand/index.php HTTP/1.1" 200 914
127.0.0.1 - - [08/Jul/2013:00:36:55+0530] GET /sand/main/about.php HTTP/1.1" 200 523
127.0.0.1 - - [08/Jul/2013:00:36:56+0530] GET /sand/index.php HTTP/1.1" 200 914
127.0.0.1 - - [08/Jul/2013:00:36:58+0530] GET /sand/main/home.php HTTP/1.1" 200 513
127.0.0.1 - - [08/Jul/2013:00:36:58+0530] GET /sand/main/home.php/logo.jpg HTTP/1.1" 200 513
127.0.0.1 - - [08/Jul/2013:00:36:59+0530] GET /sand/index.php HTTP/1.1" 200 914
127.0.0.1 - - [08/Jul/2013:00:37:00+0530] GET /sand/main/institute.php HTTP/1.1" 200 1001
127.0.0.1 - - [08/Jul/2013:00:37:03+0530] GET /sand/main/home.php HTTP/1.1" 200 513
127.0.0.1 - - [08/Jul/2013:00:37:03+0530] GET /sand/main/home.php/logo.jpg HTTP/1.1" 200 513
127.0.0.1 - - [08/Jul/2013:00:37:04+0530] Fail /sand/index.php HTTP/1.1" 300 914
127.0.0.1 - - [08/Jul/2013:00:37:07+0530] GET /sand/main/institute.php HTTP/1.1" 200 1001
127.0.0.1 - - [08/Jul/2013:00:37:08+0530] GET /sand/main/administration.php HTTP/1.1" 200 632
127.0.0.1 - - [08/Jul/2013:00:37:10+0530] GET /sand/main/institute.php HTTP/1.1" 200 1001
127.0.0.1 - - [08/Jul/2013:00:37:12+0530] GET /sand/main/administration.php HTTP/1.1" 200 632
```

Figure 5: Raw log file with irrelevant data

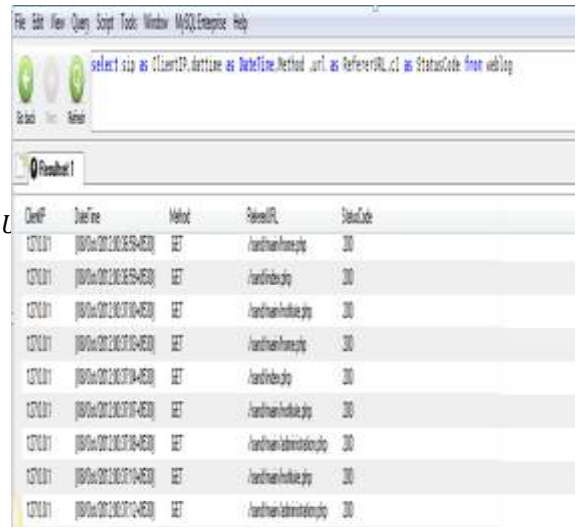

Figure 6:-Pre Processed Data

After eliminating irrelevant data from raw web log file proposed log pre processing technique lead to reduce the size of input data for web uses mining. Original raw log file size is 378 KB whereas pre processed log file size is 19 KB. Along with that log pre processing also return some interesting result as show in table.

Table 1:- Result Parameter of Log Pre Processing

| Parameter | Pre processed log file | Original log file |
|---|---|---|
| **Number Of Line** | 247 | 4214 |
| **Total Number Of Session** | 50 | ---- |
| **Total Number Of Client** | 20 | ---- |
| **Size** | 25KB | 324KB |

## 7. Conclusion

This paper gives an bird eye over web mining, Web Usage Mining (WUM) that focuses on Web log as source data and try to categorize navigation preference of end user in order to enhance quality of services and personalization of web portals describe. how log file pre processing is important for good mining result.

## REFERENCES

[1] Toufiq Hossain Kazi, Wenying Feng and Gongzhu Hu, "Web Object Prefetching: Approaches and a New Algorithm", IEEE 2010, pp 115-120.

[2] Brijendra Singh and Hemant Kumar Singh, "Web Data Mining Research: A Survey", IEEE 2010.

[3] Kavita Sharma, Gulshan Shrivastava and Vikas Kumar, "Web Mining: Today and Tomorrow", IEEE 2011, pp 399-403.

[4] WANG Yong-gui and JIA Zhen, "Research on Semantic Web Mining" IEEE 2010, pp 67-70.

[5] P. Sampath, C. Ramesh, T. Kalaiyarasi, S. Sumaiya Banu and G. Arul Selvan, "An Efficient Weighted Rule Mining for Web Logs Using Systolic Tree", IEEE 2012, pp 432-436.

[6] Nizar R. Mabroukeh and C. I. Ezeife, "Semantic-rich Markov Models for Web Prefetching", IEEE 2009, pp 465-470.

[7] A.B.M.Rezbaul Islam and Tae-Sun Chung, "An Improved Frequent Pattern Tree Based Association Rule Mining Technique", IEEE 2011

[8] Shaily Langhnoja, Mehul Barot, Darshak Mehta , "Pre-Processing: Procedure on Web Log File for Web Usage Mining." International Journal of Emerging Technology and Advanced Engineering , Volume 2, Issue 12, December 2012

[9] 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

[10] Ms. Dipa Dixit , Fr.CRIT, Vashi ,Ms. M Kiruthika "Preprocessing Of Web Logs" International Journal on Computer Science and Engineering Vol. 02, No. 07, 2010,