Volume 2, Issue 3, March 2012 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

An Enhanced Pre-Processing Research Framework for Web Log Data

T. Revathi (Asst. Prof), M. Mohana Rao, Ch. S. Sasanka
IST & KLCE
sashankchaturuvedula@gmail.com
K. Jayanth Kumar, B. Uday Kiran
IST & KLCE

Abstract*-The information on the web is growing dramatically and it is well known that over 80% of the time required to carry out any real world data mining project is usually spent on data pre-processing. Data pre-processing lays the groundwork for data mining. Before the discovery of useful information/knowledge, the target data set must be properly prepared. But it is unfortunately ignored by most researchers on data mining due to its perceived difficulty. This paper describes an efficient approach for data pre-processing for mining Web based user data in order to speed up the data preparation process. It not only provides flexibility for data pre-processing but also reduce complexity and difficulty of preparation for mining user data. However, the Web log data doesn't perform the data mining directly in most cases because of the messy and redundant content and other reasons. So, this paper analyzes the data pre- processing on Web log in order to meet the needs of data mining.

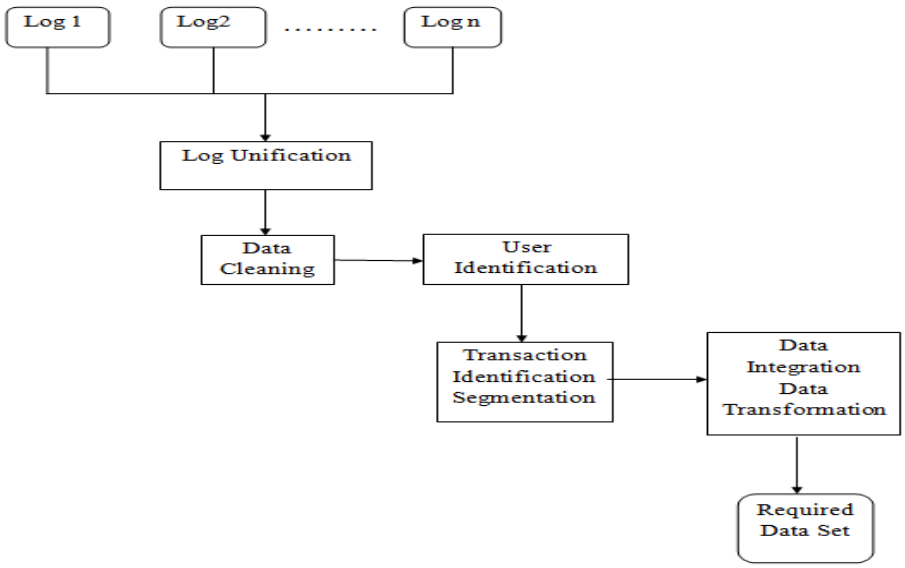
Keywords- Pre-processing, Cleaning, Null Values, Webmining, logs

I. Introduction

Over the last decade, with the continued increase in the usage of the WWW, web mining has been established as an important area of research. Whenever, the web users visit the WWW, they leave abundant information in web log, which is structurally complex, heterogeneous, high dimensional and incremental in nature. Analyzing such data can help to determine the browsing interest of web user. To do this, web usage mining focuses on investigating the potential knowledge from browsing patterns of the users and to find the correlation between the pages on analysis. The main goal of web usage mining is to Capture, Model and Analyze the web log data in such a way that it automatically discovers the usage behaviour of web user. Following the successful application of the data mining techniques in the traditional database fields, people have begun to study the Web-based data mining technology (Abbr. Web Mining). The Web log mining is a technology which applies the data mining technologies to the log files in Web server in order to find the browsing patterns of users and analyze the site usage. And it can be also used to assist the site managers to optimize the sites. The 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. So it is difficult to make a mining directly on it and we may get some wrong results. An amount of researches focused on personalized service to achieve the specific technology, such as the recommended technology, information retrieval, user clustering technology, but user modelling techniques are rarely mentioned. However, with the development and in-depth study of personalized service, researchers gradually realize that the quality of personalized service not only depends on the specific recommendation technology, search technology, but also relies on user's preferences and other characteristics of interest, description of its computable, while the latter is particularly important.

II. WEB LOG DATA PREPROCESSING

Web Log Data is a kind of data that records user's web- watching behaviours (such as visited URL, time of the visit and so on).



An example of Web Log Data is shown in TABLE.

Host	User Id	Time	URL
115.248.116.137	1	01-01-2012 20:28:28	/img-sys/bg.jpg
202.62.86.58	2	02-01-2012 12:38:55	/img-sys/bg.jpg
65.255.37.250	3	02-01-2012 15:01:00	/img-sys/bg.jpg
112.79.40.103	4	02-01-2012 16:04:39	/image/data/logo.png

A user ID is the unique name you use to identify . user ID is displayed when you buy or sell on website or any other means. However, other users won't be able to see your real name and other personal information. Each row of Web Log Data represents the URLs that the user visits. Attributes of the data include Visit Time, Host, URL, and other miscellaneous information about users' actions. Visited URLs of Web Log Data are only records of users' web-watching behaviours. In order to get user's interest categories, we should know the categories of web pages that the user visits.

In order to get the suitable Web log data to perform the data mining, we must undertake a series of operations on the original Web log files such as the log consolidation and data cleaning, user and transaction identification, data integration and so on. The basic process of Web log mining pre-processing is shown below:

III. LOG UNIFICATION

In the previous studies, the unification of Web logs is relatively simple. This is because under the circumstance that the Web service content is limited, as for a single Web server, the log files are just generated according to some certain naming rules and the different time. The log unification is just a simple timing accumulation of these files. With the enrichment and expansion of Web application content, many large-scale Web services include more and more contents. Then most of Web services have employed the automated multi-server load balancing architecture. At this time, the logs which are served by the same Web are usually scattered and stored in different servers. In light of these circumstances, we need to periodically synchronize to the background or a special log server through the certain means. After the unification, we should use the log analysis tools to analyze.

IV. DATA CLEANING

Data cleaning is usually site-specific, and involves tasks such as, removing extraneous references to embedded objects that may not be important for the purpose of analysis, including references to style files, graphics, or sound files. The cleaning process also may involve the removal of at least some of the data fields. The status code return by the server is three digit number. There are four class of status code: Success (200 Series), Redirect (300 Series), Failure (400 Series), Server Error (SOD Series). The most common failure codes are 401 (failed authentication), 403 (Forbidden request to a restrict subdirectory, and the dreaded 404 (file not found) messages. Such entries are useless for analysis process and therefore they are cleaned form the log files. Data cleaning contains the null value noise and data processing the inconsistent data processing and some others. 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.

Being mainly against the problems of the irregularity of data in multiple data sources, ambiguity, duplication, non-integrity and some others, the data cleaning accordingly performs the cleaning operations for the error data. Data cleaning can improve the quality of data, thus enhance the accuracy and performance of the subsequent data mining process. Because the high-quality analysis and decision must depend on the good-quality data, data cleaning is an important step in the data mining.

V. USER AND TRANSACTION IDENTIFICATION

A. Identification of the Users

Identification of individual users who access a web site is an important step in web usage mining. Various methods are to be followed for identification of users. The simplest method is to assign different user id to different IP address. But in Proxy servers many users are sharing the same address and same user uses many browsers. An Extended Log Format overcomes this problem by referrer information, and a user agent. If the IP address of a user is same as previous entry and user agent is different then the user is assumed as a new user. If both IP address and user agent are same then referrer URL and site topology is checked. If the requested page is not directly reachable from any of the pages visited by the user, then the user is identified as a new user in the same address. Caching problem can be rectified by assigning a short expiration time to HTML pages enforcing the browser to retrieve every page from the server.

B. Transaction Identification Segmentation.

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. In the Web log mining, the user session is the only object with the characteristics of natural services. However, the granularity of it is too coarse for the mining association rules and some other methods. It is necessary to use the segmentation algorithm to translate them into smaller transactions. Each user session in a user session file can be thought of in two ways; either as a single transaction of many page references, or a set of many transactions each consisting of a single page reference. The goal of transaction identification is to create meaningful clusters of references for each user. Therefore, the task of identifying transactions is one of either dividing a large transaction into multiple smaller ones or merging small transactions into fewer larger ones. This process can be extended into multiple steps of merge or divide in order to create transactions appropriate for a given data mining task. Both types of approaches take a transaction list and possibly some parameters as input, and output a transaction list that has been operated on by the function in the approach in the same format as the input. Three different ways of identifying transactions i.e., based on: Reference Length (time spent when visiting a page), Maximal Forward Reference (set of pages in the path from the first page in a user session up to the page before a backward reference is made) and Time Window.

By analyzing this information, a Web Usage Mining system can determine temporal relationships among data items such as the following Cricinfo Web site examples:

– 7.60% of the site visitors accessed the EspnCricinfo home page followed by the chrome Extension.

– 32.48% of the site visitors accessed the India main page followed by the cricinfo main page.

VI. DATA INTEGRATION AND TRANSFORMATION

As for the access sequences of the user which are obtained by data cleaning, they may go beyond a long time period. So it is possible that the user has visited the site more than once in this period. The purpose of transaction identification is to divide all the visit sequences of user into multiple separate user visit sequence. The simplest method to obtain this division is to define a time period. If the access time interval that the user requests any two adjacent pages goes beyond this time period, then it will consider that the user has started a new session. And in general, this time period is selected as 30 minutes.

At the same time, the attribute value of the same real world entity from different data sources may be different. This difference may be due to the different representation, ratio or coding. For example, the representation of time is inconsistent in the server logs and agent logs or the error referenced logs. Even the representation of the same page is also inconsistent in the server logs. At this time, we should transform them into the same representation in order to improve the accuracy and speed of the subsequent mining. Data transformation is to convert the data into the form which is suitable for mining. The common methods contain smooth, aggregation, standardization, data generalization, attribute constructors and some others. After the data transformation, the corresponding data mining could be performed for the formative data, such as: association rules mining, sequential pattern mining and so on.

VII. OBTAINED RESULTS

Initially, we considered a web log data to perform preprocessing. For a perfect website, it is difficult observe missing values noisy and redundant data. So, we considered an offline website and we created some missing values in the data. Now we clean this data using different efficient methods like replacing null values with most probable value and so on. These can be viewed in diagrammatic representations below.

FNAME	LNAME	UNAME	EMAIL	PWD	GENDER	CITY	PHNO	PROFESSION	SALARY	REG_DATE
immu	gow	immugow	immugow@hotmail.com	immu	Male	vijayawada	9219939399	Un-Employed	908789	2/19/2012
uday	prabha	prabhasuday	prabhasuday@live.com	prabha	Male	-	9882884882	Un-Employed	-	2/19/2012
deepak	mamma	deepmam	deepakmam@live.com	deepu	Male	guntur	9030708248	Employed	90	2/19/2012
deepak	dcruz	ilianadeep	ilianadeep@yahoo.com	deepu	Male	-	9030708248	Employed	909909	2/19/2012
ravindra	jadeja	ravindra.j	ravindra.j@yahoo.com	ravindra	Male	bangalore	7373728384	Un-Employed	30000	2/19/2012
ashwin	r	ashwin	ashwin.r@yahoo.com	ashwin	Male	bangalore	8348848848	Employed	20000	2/19/2012
irfan	pathan	irfan	irfan.pathan@yahoo.com	irfan	Male	-	8747748282	Employed	20000	2/19/2012
ashok	dinda	ashok.d	ashok.dinda@yahoo.com	ashok	Male	bangalore	9883873733	Un-Employed	-	2/19/2012
manoj	tiwary	manoj	manoj.t@yahoo.com	manoj	Male	bangalore	8199198328	Employed	20000	2/19/2012
yusuf	pathan	yusuf	yusuf.p@yahoo.com	yusuf	Male	-	8899491919	Un-Employed	20000	2/19/2012
vinay	kumar	vinay	vinay.k@gmail.com	vinay	Male	mumbai	8319747472	Un-Employed	-	2/19/2012
sherlyn	chopra	sherlyn	sherlyn.c@yahoo.com	sherlyn	Female	bangalore	6363782881	Un-Employed	200000	2/19/2012
aishwarya	rai	aish220	aish220@gmail.com	aishwarya	Female	-	7768383883	Un-Employed	-	2/19/2012
shriya	sharan	shriya	shriya.saran@gmail.com	shriya	Female	mumbai	7277282882	Un-Employed	-	2/19/2012
abishek	bachan	abhishek	abhi.shek@yahoo.com	abhishek	Male	mumbai	8881883882	Un-Employed	100000	2/19/2012

row(s) 91 - 105 of 114

Figure 1: Data with Missing Values

Figure 2: Null Values Count of the data

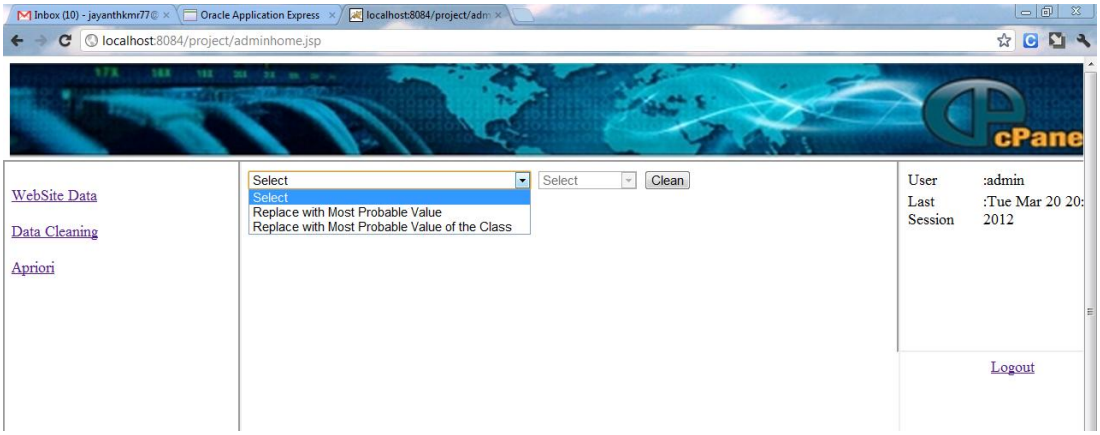


Figure 3: Cleaning Mechanisms

FNAME	LNAME	UNAME	EMAIL	PWD	GENDER	CITY	PHNO	PROFESSION	SALARY	REG_DATE
immu	gow	immugow	immugow@hotmail.com	immu	Male	vijayawada	9219939399	Un-Employed	908789	2/19/2012
uday	prabha	prabhasuday	prabhasuday@live.com	prabha	Male	mumbai	9882884882	Un-Employed	20000	2/19/2012
deepak	mamma	deepmam	deepakmam@live.com	deepu	Male	guntur	9030708248	Employed	90	2/19/2012
deepak	dcruz	ilianadeep	ilianadeep@yahoo.com	deepu	Male	mumbai	9030708248	Employed	909909	2/19/2012
ravindra	jadeja	ravindra.j	ravindra.j@yahoo.com	ravindra	Male	bangalore	7373728384	Un-Employed	30000	2/19/2012
ashwin	r	ashwin	ashwin.r@yahoo.com	ashwin	Male	bangalore	8348848848	Employed	20000	2/19/2012
irfan	pathan	irfan	irfan.pathan@yahoo.com	irfan	Male	mumbai	8747748282	Employed	20000	2/19/2012
ashok	dinda	ashok.d	ashok.dinda@yahoo.com	ashok	Male	bangalore	9883873733	Un-Employed	20000	2/19/2012
manoj	tiwary	manoj	manoj.t@yahoo.com	manoj	Male	bangalore	8199198328	Employed	20000	2/19/2012
yusuf	pathan	yusuf	yusuf.p@yahoo.com	yusuf	Male	mumbai	8899491919	Un-Employed	20000	2/19/2012
vinay	kumar	vinay	vinay.k@gmail.com	vinay	Male	mumbai	8319747472	Un-Employed	20000	2/19/2012
sherlyn	chopra	sherlyn	sherlyn.c@yahoo.com	sherlyn	Female	bangalore	6363782881	Un-Employed	200000	2/19/2012
aishwarya	rai	aish220	aish220@gmail.com	aishwarya	Female	mumbai	7768383883	Un-Employed	20000	2/19/2012
shriya	sharan	shriya	shriya.saran@gmail.com	shriya	Female	mumbai	7277282882	Un-Employed	20000	2/19/2012
abishek	bachan	abhishek	abhi.shek@yahoo.com	abhishek	Male	mumbai	8881883882	Un-Employed	100000	2/19/2012

row(s) 91 - 105 of 114

Figure 4: Cleaned Data

NOTE: Observe the values circled in Figure 4, those are the valued replaced by the chosen mechanism. Observe the difference by comparing Figure 1 and Figure 4

In this way, we performed cleaning operations on raw data . Similarly, we can apply those operations on raw log data too.

VIII. CONCLUSION

Web sites are one of the most important tools for advertisements in international area for universities and other foundation. The quality of a website can be evaluated by analyzing user accesses of the website by web usage mining. The results of mining can be used to improve the website design and increase satisfaction which helps in various applications. Log files are the best source to know user behavior. But the raw log files contains unnecessary details like image access, failed entries etc., which will affect the accuracy of pattern discovery and analysis. So pre processing stage is an important work in mining to make efficient pattern analysis. Therefore, the pr eprocessing before the data mining in Web logs should become a more important research.

IX. REFERENCES

[1]Borges J, Levene M. 1998. Mining association rules in hypertext databases. Proc. 1998 Int'l Conf. on Data Mining and Knowledge Discovery (KDD'98), 149–153.

[2]Buchner A, Mulvenna. 1998. Discovering internet marketing intelligence through online analyticalweb usage mining. SIGMOD Record 27.

[3]Cooley R. 2000. Web Usage Mining: Discovery and Application of Interesting Patterns from Web Data. PhDdissertation, Univ. of Minnesota.

[4]Cooley R, Mobasher B, Srivastava J. 1997. Webmining: information and pattern discovery on the World Wide Web. Proc. Int'l Conf. on Tools with Artificial Intelligence, 558–567, Newport Beach, CA.

[5]Cooley R, Mobasher B, Srivastava J. 1999. Data preparation for mining World Wide Web browsing patterns. Journal of Knowledge and Information Systems 1.

[6]Etzioni O. 1996. The World Wide Web: quagmire or gold mine. Communications of theACM36: number 11 (November), 65–68.

[7]Hallam-Baker PM, Behlendorf B. Extended Log File Format, http://www.w3.org/pub/WWW/TR/WD-logfile.html.

[8]Luotonen A. 1995. The Common Log File Format.http://www.w3.org/pub/WWW/Daemon/User / Config/Logging.html.

[9]Madria S, Bhowmick S, Ng WK, Lim EP. 1999.Research issues in web data mining. DAWAK'99,Florence, Italy, Sept.

[10]Mobasher B, Dai H, Luo T, Nakagawa M, Sun Y,Wiltshire J. 2000. Discovery of aggregate usage profiles for web personalization. Proceedings of the Web Mining for E-Commerce Workshop(WebKDD'2000), Boston,August.

