Web Logs Conversion to Improve the Analysis of Web Usage Data for Web Usage mining in E-Learning

Maneet Kaur
LSECA –CSE/IT Department
Lovely Professional University, Punjab, India
maneet_meet@yahoo.co.in

AnupNandy
LIECA - CSE/IT Department
Lovely Professional University, Punjab, India
nandy.anup @gmail.com

***Abstract*—This paper explicitly describes the underlying concept of web usage mining in accordance with the inherent enhancement of web usage mining by applying multi dimensional schema. The improvement policy is to be carried out by the analysis of the Web usage data and performing the web usage mining. It is to be done by taking the flat files as an input and convert those flat log files into multidimensional data. Those data are to be used in analysis tools which are commonly known as Online Analytical Processing or data mining tools. These advanced OLAP tools are used in e-learning environments in order to analyze the students learning behavior. The entire mining process is based on the web logs of server to provide amendable success in this application. It includes an e-learning platform which could be used to analyze the student's learning behavior. A complete test data set has been taken from Lovely Professional University for mining purposes.**

***Keywords*—web usage mining, web logs, e-learning, OLAP, data mining, pentaho**

I. Introduction

In the recent arena, numerous amounts of network technologies are applied in education field. The internet based learning makes it possible the long life education with the help of e-learning. The internet based learning has exponentially changed the mode of learning. Today, number of educational institutes and universities are using e-learning platforms to improve their education system. All the learning material or data related to that course will be available through the internet. It provides the availability of the learning material to students at every time [2].

The web server's web logs have the information of user such as the number of visited pages ,date and time of connection, time spent on each page, the browser and operating system type of visitors etc [5]. There is huge amount of data in web log files but we need only the required data for the analysis and mining purposes. This information is not sufficient to analyze the behavior of the learners. When the number of students is potentially to be increased it tends to create a problem to extract the useful information.

To overcome the above problem, the web usage mining is to be used to extract the meaningful and hidden information from huge amount of data [9, 7]. For this purpose the OLAP tool is used to accomplish the prerequisites [4]. It immensely investigates the use of business intelligence and OLAP (online analytical processing is an approach to answer multi-dimensional analytical queries) tools in e-learning environments. In this paper, to improve the analysis of web usage mining need to formulate the huge amount of data in multidimensional schema. In the multidimensional schema [6] the data is being represented into facts and dimension tables. This multidimensional data is used in OLAP tool to analyze the behavior of learner [3].

This paper is organized into seven sections. Section 2 briefly presents about the web usage mining. Section 3 represents about the history of related work. In section 4 illustrate about the source of data taken with the description of the important attributes which needs to be filtered from the server web logs. A description of the main idea to convert the web logs into multidimensional data and the proposed model are represented in section 5, moreover, excerpt of the code for a schema is also presented. Section 6 presents the analysis of data in perspective of time efficiency and handling multiple queries at once. Section 7 contains the conclusions and the future work.

II. Overview of web usage mining

Web mining is the use of data mining techniques to automatically discover and extract the hidden or undiscovered information from Web documents/services [13].web mining is divided into three categories, one is web content mining second is web structure mining and third one is web usage mining.Fig. 1 illustrates the three categories of web mining.Web usage mining understands the user's behaviors by analyzing the web logs [8].

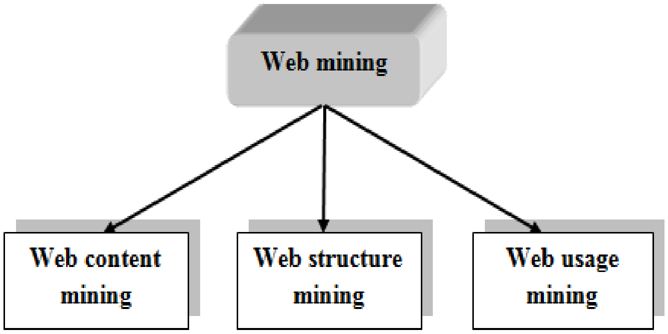


Figure1. Types of web mining

Web logs are usually text files recording all the transactions between the web server and the users. Each transaction consists of identifiable information [1]. The type of information we get from web logs is, number of hits , average hits per time period, what are the popular pages in your site, who is visiting your site, what keywords are users searching for to get to you, what is being downloaded. Web logs may records for months or even for years, so that they are huge in size. Fig. 2 shows the step by step process of mining. Description of the process web usage mining.

- Firstly collect the web logs data from the web log server.
- Data cleansing remove the useless transactions from the web logs, such as requests for non-existing pages. This step can be done through the cleansing tools but in this paper the data cleansing done manually.
- Identification is done manually. Each user having a unique IP address, with the help of these address users can be differentiated. The requests from the same IP address consider as from the same user.
- Create the dataset of above web logs data. Only the data needed for mining is stored in dataset.All the preprocessing steps are completed now.
- After that need to perform the mining task on the dataset.
- The Summarize analysis report on the mining results.

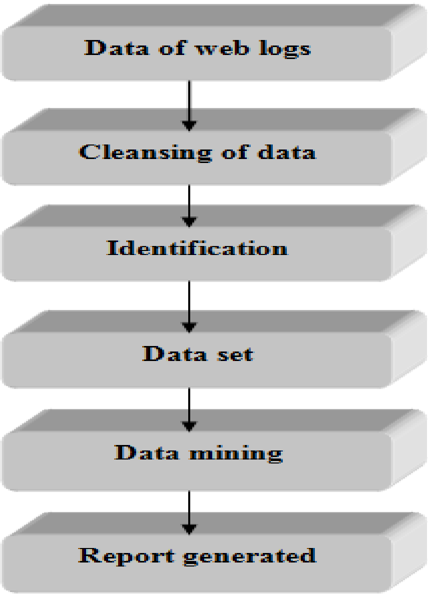


Figure2. Flow chart of Mining process

III. related work

There are some tools for web log data analysis such as analog is a good tool for web log file analysis. There is a limitation of using these types of analysis tools. Analog provides simple statistical results without performing the advanced analytical tasks such as OLAP and data mining [2]. These commercial tools of analysis have the limited capability of drill down and decision making and the data is not stored in an efficient way.Fig. 3 lists the countries of the computers which requested files. Table1 shows theListing domains, sorted by the amount of traffic.

To overcome the above limitations of commercial analysis tools in this paper using the advanced OLAP tools for data mining. OLAP (online analytical processing, or OLAP is an approach to swiftly answer multi-dimensional analytical queries.) tools are much capable for drill down and have good decision making power [10]. Improve the analysis of web usage mining need to structuring the huge amount of data in multidimensional schema. Once the data is structured in multidimensional schema, it is possible to use the OLAP and data mining tools to analyze the web logs data for the purpose of analyzing the behavior of learner. In the multidimensional schema the data is available into facts and dimension tables.

Finally the most recent tool is PENTAHO used to generate the report of the data. This analyzed data will help the course instructor to discovering the student behavior pattern.The experimental results represent the impact of proposed work in improving web based teaching and learning effectiveness.

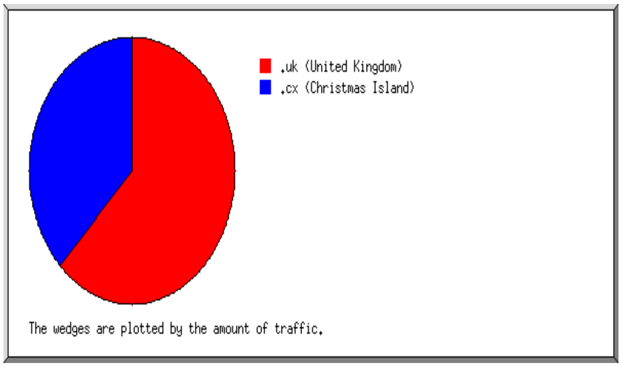


Figure3.Result of Analog

TABLE I. AMOUNT OF TRAFFIC

Requests	%Bytes	Domain
30	62.24%	.uk (United Kingdom)

Requests	%Bytes	Domain
16	37.76%	.cx (Christmas Island)

IV. Source of data

For this paper the data is taken fromlovely professional university. The data of web logs is of one semester started from 16 august to 20 January 2010. The data of web logs is confidential that’s why I am getting the analyzed data of web logs and creating the synthetic data for analysis on the basis of analyzed web logs data of lovely professional university. Format of web logs data is flat files. Web logs data is huge collection of data. Each transaction has near about 15 fields such as, IP address, user name, timestamp, access request, result status code, bytes transferred, referrer URL, user agent etc [12]. Filtered the only that fields which are needed for the analysis of web usage data: IP address, web site URL, user ID, date and time, referrer URL.

There is huge amount of data in web log files but for this paper need only the required data for the analysis and mining purposes. This information is not sufficient to analyze the behavior of the learners.

V. Proposed apporoach

The flat files provide the statistical results without performing the advanced analytical tasks such as OLAP and data mining [11]. These types of flat files are capable to handle only the one dimensional queries. To overcome this problem the flat files are firstly converted into the multidimensional schema, which is much of capable to handle the multidimensional queries of the user. The question like "how many students access the web site at the particular time and from the particular location” and “how long they spend viewing the learning materials of which type”.

In the multidimensional schema the data is available into facts and dimension tables. Fact table holds the main data such as IP address, web site URL, user ID, date and time, referrer URL. Dimension tables which are usually smaller than fact tables, include the attributes that describe the facts. This schema is known as star schema. Fig. 4 shows one fact table sessions and five dimension tables time, page, location, referrer. Fig. 5 and 6 shows an excerpt of the SQL code for a star schema. Fig. 5 shows the code for creating the dimension table, while Fig. 6 shows the code for creating the fact table. Table 2 describes fact and dimensions.

Fig. 7 illustrates the model of the web usage mining for e-learning. All the enrolled students are provided user id for internet access. Through this id the student can logins to the e-learning platform. The flat web log data files are converted into the multidimensional data through the star schema. After that the report is generated through the web usage mining.

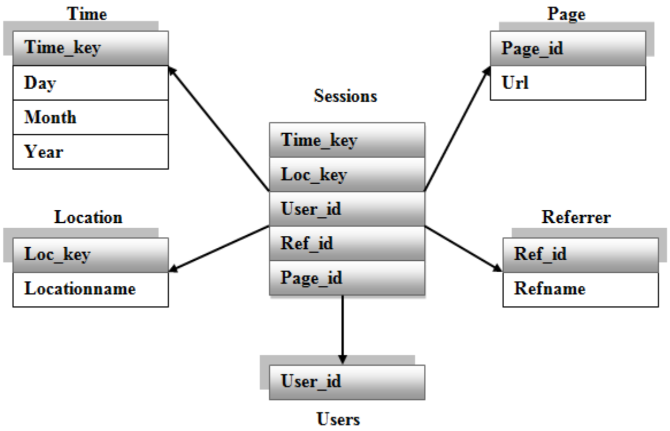
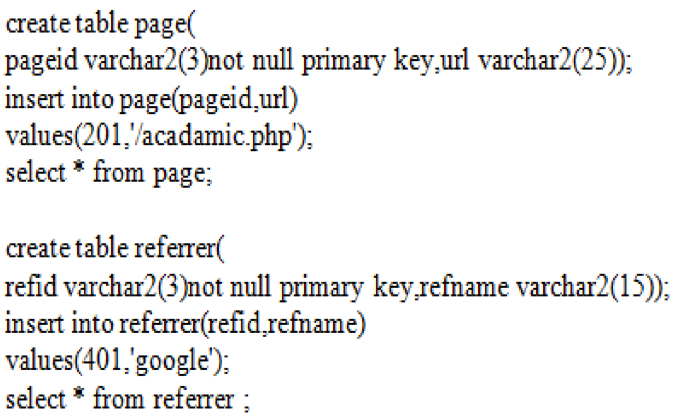


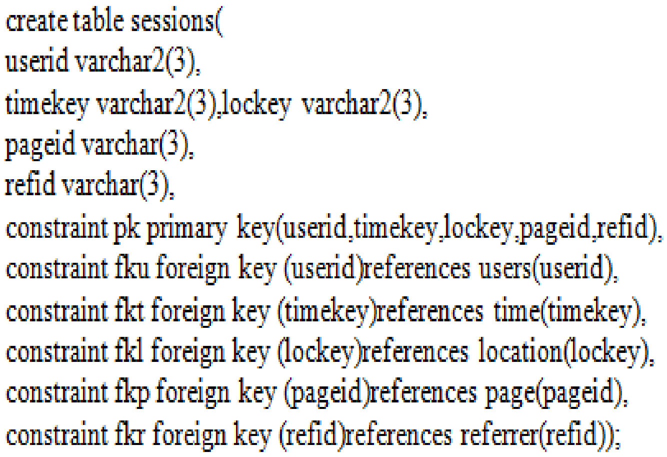
Figure4. Representation of Star schema



```
create table page(
pageid varchar2(3)not null primary key,url varchar2(25));
insert into page(pageid,url)
values(201,'/acadamic.php');
select * from page;

create table referrer(
refid varchar2(3)not null primary key,refname varchar2(15));
insert into referrer(refid,refname)
values(401,'google');
select * from referrer ;
```

Figure5. SQL code for dimension table



```
create table sessions(
userid varchar2(3),
timekey varchar2(3),lockey varchar2(3),
pageid varchar(3),
refid varchar(3),
constraint pk primary key(userid,timekey,lockey,pageid,refid),
constraint fku foreign key (userid)references users(userid),
constraint fkt foreign key (timekey)references time(timekey),
constraint fkl foreign key (lockey)references location(lockey),
constraint fkp foreign key (pageid)references page(pageid),
constraint fkr foreign key (refid)references referrer(refid));
```

Figure6. SQL code for fact table

TABLE II. FACT AND DIMENSIONS

Time	Dimension
Location	Dimension
User	Dimension
Sessions	Fact

Page	Dimension
Referrer	Dimension

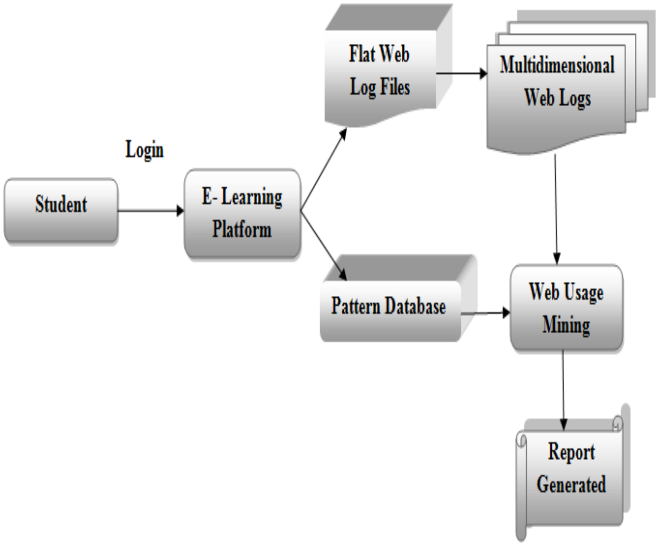


Figure7. Model of e-learning

After the creation of the multidimensional data need of the product which can be used to generate the business intelligent solution. There are ample of products which can support the data warehousing and OLAP technology. In this paper we have presented PENTAHO open source platform for the purpose of business intelligence.

Fig. 8 illustrates the Flow of process for proposed model. It has two ways to do the work:-

A. *First way*

Take the data of student web logs from server then convert it into the Comma Separated Value data file. The input of the CSV is taken by pentaho tool which converts the flat web logs into multidimensional data. Now this data is used for the analysis and web usage mining purposes.

B. *Second way*

Use the oracle to generate the star schema, and then make the connectivity between the oracle database and pentaho. Now we can mine the data from the database to generate the report.

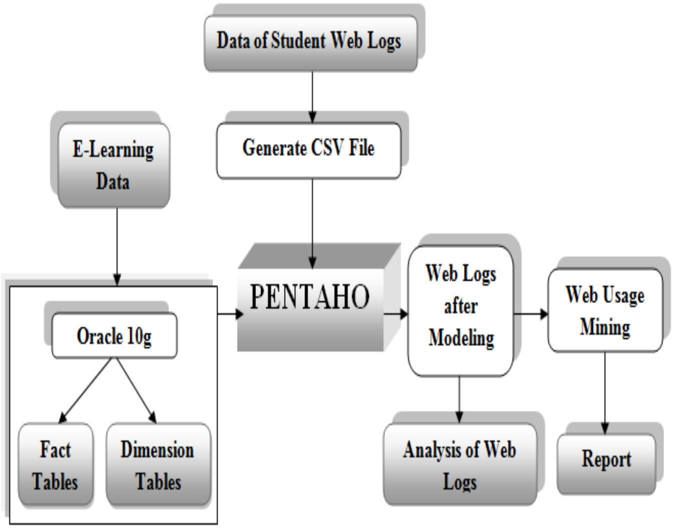


Figure8. Proposed model

VI. Result Analysis

The experimental results have been generated using PENTAHO tool. Fig. 9 shows the link from where the LPU web site getting more hits. There are four main pages for LPU web sitelike /academic.php, /jobs_at_lpu.php, /programserch.aspx, /about_lpu.php are being accessed more. The result shows that More traffic on the LPU web site is generated by the /academic.php from 16 august to 20 January 2010. Second highest is /programsearch.aspx. X-axis presents the year and pages. Y-axis presents the number of users. Table

3 presents the related information about the percentage of traffic pages of the above Fig. 9.

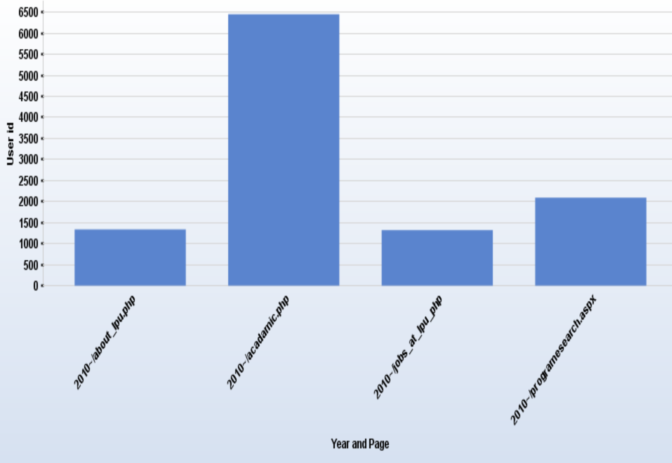


Figure9. Traffic pages

TABLE III. PERCENTAGEOF TRAFFIC PAGES

Pages	%Page views
/academic.php	60%
/jobs_at_lpu.php	30%
/programsearch.aspx	40%
/about_lpu.php	25%

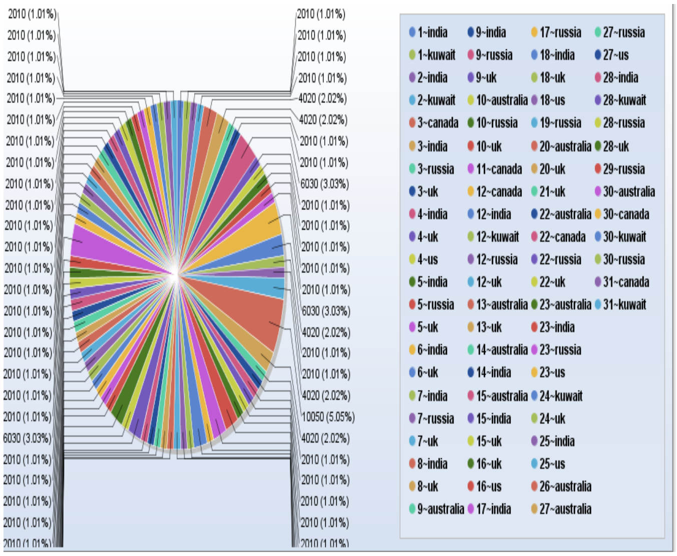


Figure10. Location of hits

Fig. 10 shows the percentage of hits the web site getting from particular location with respect to the particular days. Locations are represented through the different colors. For example in first day the web site got hits from India (1.01%) and Kuwait (1.01%), at twentieth day the web site got hits from Australia (5.05%) and UK (2.02%).Table 4 presents the related information about the percentage of location hits of the above Fig. 10.

TABLE IV. PERCENTAGEOF LOCATION HITS

Country	Pages/Visit
India	55%
US	42%
Kuwait	35%
UK	50%
Russia	50%
Canada	45%
Australia	40%

Fig. 11 shows the source through which the web site getting hits. There are four types of sources like:- Google, Mail.lpu.co.in, Search, Bing. It also takes the location of access into consideration. Result shows, web site got more of the hits through the Google source and Mail.lpu.co.in. X-axis represents the location and source. Y-axis represents the user id.Table 5 presents the related information about the percentage of traffic sources of the above Fig. 11.

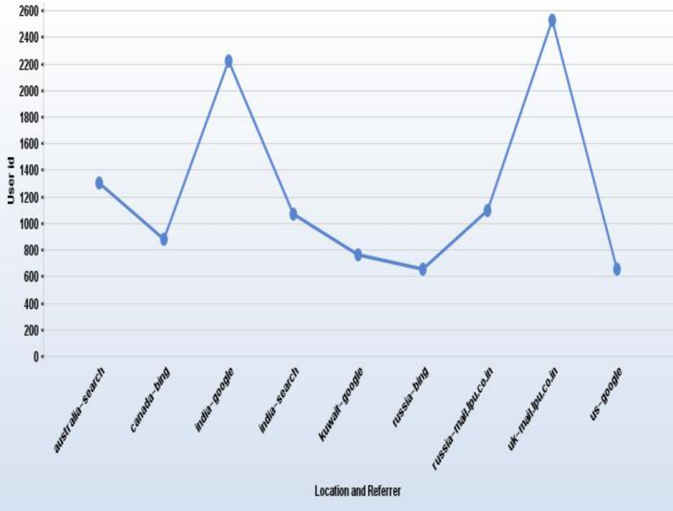


Figure11. Traffic sources

TABLE V. PERCENTAGEOF TRAFFIC SOURCES

Sources	% visits
Google	60%
Mail.lpu.co.in	62%
Search	45%
Bing	40%

In which months a particular user gives hits to the web site shown in this Fig. 12. Different colors represent the user ids. For example user with 109.235.49.143:80 id gives hits to the web site in 1st, 2nd, 4th, 11th and 12th month. User with109.75.163.12:8080id gives hits to the web site in 2nd and 9th month.

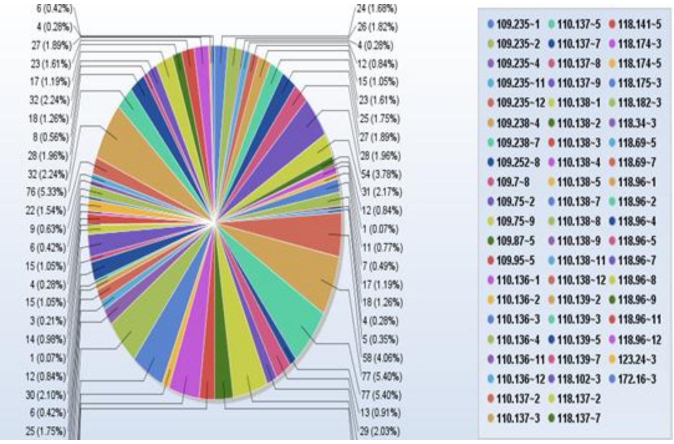


Figure12. Number of hits

After examining the web logs of lovely professional university the analyzed results shows that the 75%-80% of students participated in internet learning. They read there course related material through the /academic.php page of web site. The course material is uploaded time to time by the course instructor in the account of student. Assigning the four assignments during the semester and forcing the students to upload their assignments at the time considered as an appropriate way to increase the communication between the student and course instructor. In activity of assignment uploading 90% of students participated.

VII. Conclusion and future work

A generic model has been proposed to overcome the difficulties occurred in conventional e-learning environments. The concept of data warehousing and business intelligence has been represented in the field of e-learning in order to make proper decisions. It overcomes the limited capability of drill down and decision making. This proposed work produces the results in the form of time and storage efficiency. It handles the multiple queries at once to make the business intelligent.The experimental results represent the impact of conversion in improving web based teaching and learning effectiveness.

The future work includes applying clustering and classification algorithms to predict the behavior of students on the basis of performance. Moreover, it is proposed to use a snowflake schema to deal with the huge amount of data presented in the hierarchical structure with more than one fact tables.

Acknowledgment

I would like to thank to Info Tech Department of Lovely Professional University (Punjab) for giving me the web logs repository for my research work. I would also like to extend my thanks to Mr.ParulKhurana, faculty of LPU for his esteemed support towards my research work.

References

[1] Chih-Hung Wul, Yen-Liang Wu2, Yuan-Ming Chang 3 , and Ming-Hung Hungl, “Web Usage Mining on the Sequences of Clicking Patterns in a Grid Computing Environment,” Proceedings of the Ninth International Conference on Machine Learning and Cybernetics, Qingdao, pp. 2909-2914, July 2010.

[2] Xinjin Li Sujing Zhang, “Application of Web Usage Mining in e-learning Platform,” E-Business and E-Government (ICEE), 2010 International Conference , pp. 1391-1394,September 2010.

[3] Paul Hern´andez, Irene Garrig´os, and Jose-Norberto Maz´on Lucentia Research Group, “modeling web logs to enhance the analysis of Web usage data,”Workshops on Database and Expert Systems Applications,pp.297-301,2010.

[4] Blanco, C. Fernandez-Medina, E. Trujillo, J. Piattini, M. Escuela Super, “Implementing Multidimensional Security into OLAP Tools,”Availability, Reliability and Security, 2008. ARES 08. Third International Conference ,pp. 1248-1253,2008.

[5] Tasawar Hussain, Dr. Sohail Asghar, Dr. Nayyer Masood, “Web Usage Mining: A Survey on Preprocessing of Web Log File,”IEEE transaction,pp.466-500,2010.

[6] Wang Fu-shan, Dezhou, “Application Research of Data Warehouse and Its Model Design,”Information Science and Engineering (ICISE), 2009 1st International Conference ,pp. 798-801,2010.

[7] Han Qiyun, “ERP-based Data Warehouse Model Design,”Computer Engineering and Technology (ICCET), 2010 2nd International Conference , pp. 698-702 ,2010.

[8] Rao, V.V.R.M. Kumari, “An Efficient Hybrid Predictive Model to Analyze the Visiting Characteristics of Web User using Web Usage Mining,”Advances in Recent Technologies in Communication and Computing (ARTCom), 2010 International Conference, pp. 225-230,2010.

[9] Levchenko, M. Gopeyenko, “The Modern Approach to Planning and Implementation of Enterprise Data Warehouse,”Application of Information and Communication Technologies (AICT), 2010 4th International Conference ,2010.

[10] unjie Chen Wei Liu,“Research for Web Usage Mining Model,”Computational Intelligence for Modelling, Control and Automation, 2006 and International Conference on Intelligent Agents, Web Technologies and Internet Commerce, International Conference , 2007.

[11] Zhidan Wu Yue Yang, Shenyang, “Research and Design of Decision Support System based on Data Mining and Web Technology,” Management and Service Science (MASS), 2010 International Conference ,2010.

[12] Oskouei, R.J. Chaudhary, B.D., “Internet Usage Pattern by Female Students: A Case Study,” Information Technology: New Generations (ITNG), 2010 Seventh International Conference , pp. 1247-1250,July 2010.

[13] Ramakrishna, M.T. Gowdar, L.K. Havanur, M.S. Swamy, “Web Mining: Key Accomplishments, Applications and Future Directions,”Data Storage and Data Engineering (DSDE), 2010 International Conference ,pp. 187-191, 2010.