ISBN: 978-972-8924-97-3 © 2009 IADIS

PROPOSED FRAMEWORK FOR DATA MINING IN E-LEARNING: THE CASE OF OPEN E-CLASS

Ioannis Kazanidis, Stavros Valsamidis, Theodosios Theodosiou
TEI of Kavala

Sotirios Kontogiannis
Democritus University of Thrace

ABSTRACT

Web-based learning environments are extensively used nowadays. These environments maintain and produce vast amounts of data. Such vastness lead to the application of data mining techniques so as to further improve usability and qualify e-learning platform courses content. This paper proposes a platform dependant framework for recording, processing and analyzing data from Learning Management Systems (LMS). Its purpose and functionality is to facilitate instructors towards achievement of course efficiency. In depth, the improvements of course content are based on feedback information from data mining techniques over LMS data. In addition, a case study is discussed and an automated data mining tool, under development, is presented.

KEYWORDS

E-Learning, Learning Management Systems, data mining, Web usage mining, Open eClass.

1. INTRODUCTION

Learning Management Systems (LMSs) offer several methods for distribution of information and knowledge among participants of an on-line course. They allow instructors to deliver assignments to students, produce and publish course educational material, prepare tests, etc. (Romero et al., 2008).

LMS systems usually produce statistic reports. These reports however do not assist instructors in drawing out useful conclusions either for the course potential or student abilities and are useful only for platform administrative purposes. Moreover, existing e-learning platforms do not offer concrete tools for the assessment of user actions and educational courses content.

We propose a framework that will lead to the creation of a software tool incorporated in e - learning platforms. This tool will perform platform usage analysis in a manner that will pinpoint specific ways-actions to instructors in order to improve their course content, course usability, and follow up the needs of their courses in accordance to student capabilities. Furthermore, instructors will also benefit from framework's course evaluation metrics that will be incorporated in the proposed software and will use data mining techniques in order to deliver course ranking. Such ranking will lead instructors to improve their course content towards course efficiency, following suggestions derived from comparison with other instructors courses. Efficiency of educational content, will give to students the opportunity of asynchronous study of courses with actualised and optimal educational material.

2. RELATED WORK

Data mining is the search for relationships and patterns that exist in large data sets, but are hidden among vast amounts of data. Web mining (Srivastava et al., 2000) is a sub-category of data mining. It is divided to three sub categories: Web content mining, Web structure mining and Web usage mining (Spiliopoulou, 1999; Kosala and Blockeel, 2000; Bing, 2007). This paper focuses in Web usage mining. Web usage mining is the

application of data mining techniques in order to analyze and discover interesting patterns at user's data on the web. Usage data are the records of user's behaviour as he/she browses or makes transactions on a web site. Web analysis tools simply provide mechanisms that report user activity recorded by web servers (Web Usage Mining – WUM). Using such tools, it is possible to extract and determine information as the number of accesses to the server, the times or time intervals of visits, as well as the domain names and users requests at the web server (Srivastava et al., 2000).

There are several free tools that do WUM; Analog, SUGGEST, MATEP et al. The Analog system (Yan et al., 1996) consists of two main components, performing online and offline data processing with respect to the web server activity. Past users activity is recorded in server log files which are processed to form clusters of user sessions. The online component builds active user sessions which are then classified into one of the clusters found by the offline component. This idea inspired our research team in a different manner, discussed latter in the paper. The SUGGEST WUM (Baraglia and Palmerini, 2002) was implemented as module to Apache web server and produces links to pages of potentional user interest. It provides useful feedback to make easier the web user navigation and thus optimising web server performance. A specialized WUM tool used in e-learning platforms is MATEP. MATEP (Zorrilla and Alvarez, 2008) acts in two levels. First, it creates a mixture of data from different sources suitably processed and integrated. These data originate from e-learning platform log files, virtual courses, academic and demographic data. Second, it feeds those data to a data warehouse which in turn, with the use of MATEP provides static and dynamic reports.

The afforementioned work lead to the application of data mining methods in the e-learning (Romero and Ventura, 2007). However, there is no standardized framework in applying data mining techniques in this area. Data mining methods in the WWW have been systematically used in e-commerce applications. Nevertheless, their utilisation impact is lower than in LMS systems (Zaiane, 2001). As a result, we propose a framework for the evaluation of LMS platforms based on usage of data mining techniques over traffic analysis data.

3. PROPOSED FRAMEWORK

The proposed framework structure includes three steps. That is, logging, data pre-processing and data mining step. More specifically the three steps are the following:

Step 1 - Logging the data

This step involves the logging of specific data from e-learning platforms. In depth, the development of a data recording module, which will be embedded in the web server of the e-learning platform and will record specific e-learning platform fields such as SessionID, UserID, request, time duration between consecutive requests etc. The development of such a module has the following advantages: i) rapid storage of user information, since it is executed straight from the server API and not by the e-learning application and ii) the produced data are independent of specific formulation used by the e-learning platform.

Step 2 - Data pre-processing

The data of the log file contain noise such as missing values, outliers etc. These values have to be preprocessed in order to prepare them for data mining analysis. Specifically, this step will filter the recorded data delivered from step 1. It will use statistical methods such as outlier detection, filling out of the missing values etc. This step should not be performed by the e-learning platform and thus can be embedded into a variety of LMS systems. Also it will facilitate data mining analysis methods construction of robust results.

Step 3 - Data mining

In this step, several data mining techniques may be applicable to the filtered logged file data. The main data mining techniques that will be used are clustering, classification, association rules and regression. The purpose of clustering will involve user separation according to usage patterns extracted from e-learning platform data. It can also be used for similarity page clustering according to the content of course web pages. Classification methods will be used so as to classify users according to predefined e-learning stereotypes. The association rule mining techniques will lead to the discovery of web pages viewed together by user clusters. Eventually, the regression analysis will discover linear relations between data and prediction models.

4. CASE STUDY

Technological Education Institute (TEI) of Kavala uses the Open eClass e-learning platform (GUNet, 2009) which is an extended version of the Claroline LMS (Claroline, 2009). We are under the development of an automated data mining tool which will incorporate the proposed framework and assist the instructors of the institute by grading and comparing course content with other LMS courses. Consequently the proposed framework was applied to data collected during the last semester (spring 2009) from specific courses. The framework steps are presented bellow.

4.1 Logging the Data

An Apache module was developed in Perl programming language as a first step of the proposed framework for logging 11 fields (request_time_event, remote_host, request_uri, remote_logname, remote_user, request_method, request_time, request_protocol, status, bytes_sent, referer, agent) and recorded use requests from 78 different courses.

4.2 Preprocessing

The log file produced from the previous step has been filtered so as to include only the following fields : i) courseID, which is the identification string of each course; ii) sessionID, which is the identification string of each session; iii) page Uniform Resource Locator (URL), which contains the requests of each page of the platform that the user visited. Although these fields contain information about the e-learning process, more index and metrics (Table 1) were proposed in order to adequately facilitate the evaluation of course usage.

Table 1. Indexes name and description

Index name	Description of the index
Sessions	The number of sessions per course viewed by users.
Pages	The number of pages per course viewed by users.
Unique pages	The number of unique pages per course viewed by users.
Enrichment	The enrichment of courses (Pages/Unique pages)
Homogeny	The homogeny of unique pages per session (Unique Pages/Sessions)
Unique Pages per CourseID per Session (UniquePCSession)	The number of unique pages per course viewed by users per session

At first the number of the sessions and the number of the pages were counted in order to calculate course activity. The *unique pages* metric tries to measure the total number of pages per course. *Enrichment* is another metric which is proposed in order to express the "enrichment" of each course defined as the number of pages divided by the number of unique pages.

Since users may visit just few pages of each course, sessions alone may lead to unreliable results. Similarly, visited pages as a metric alone is not reliable enough to confirm course activity. Thus, a new index was invented, named *homogeny*, that combines sessions and pages viewed by users allowing us to evaluate course activity. Finally, *UniquePCSession* metric expresses the unique user visits per course and per session in order to calculate activity in a more objective manner. For example, some novice users may navigate in a course visiting a page more than once. *UniquePCSession* eliminates duplicate page visits, considering the visits of the same user in a session only once that could not be clarified by enrichment metric.

A sample of the first 5 courses by ranking is presented in the Table 2.

Table 2. Processed e-learning data

Course (1)	Sessions (2)	Pages (3)	Unique pages (4)	Enrichment=Pages /Unique pages (5)	Homogeny =Unique Pages/Sessions (6)	Unique Pages per CourseID per Session (UPCP) (7)	Total(8)= (2)+(3)+(7)
IMD105	91	297	11	27,00	0,12	216	604

IMD35	87	338	8	42,25	0,09	179	604
IMD132	152	230	7	32,86	0,05	184	566
IMD36	72	217	7	31,00	0,10	134	423
IMD129	75	209	6	34,83	0,08	131	415

4.3 Data Mining

The last step in the current case study according to the proposed framework involves data clustering. The clustering was performed using the open source data mining tool Weka (Witten and Eibe, 2005). The metrics and index described in the previous step were used with the SimpleKmeans for clustering platform courses. The properties of SimpleKmeans were Euclidean distance with 2 predefined clusters. The produced results show that 22 (28%) of the courses had high activity and 56 (72%) of the courses had low activity.

Visualization of the results using *UniquePCSession* is shown in Figure 1. Black points in the left indicate high activity courses whereas points in grey show low activity courses. Clustering results were used in order to propose appropriate improvements to the educational material of the low activity courses.

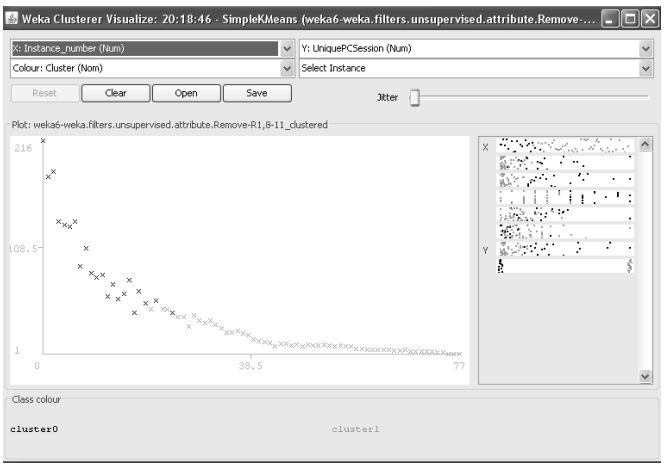

Figure 1. Cluster visualization using Unique per course per session

5. DATA MINING TOOL

The previously presented framework may lead to useful course evaluation results and thus lead to course efficiency. However, this process is time consuming and cumbersome. Furthermore, it requires the knowledge of both data mining and statistics in order to be used by tutors. For this reason we have already designed and we are developing a tool which will automate the whole process of the presented framework. The proposed tool will be plugged in the Open eClass platform and it will be easy to use by tutors.

The proposed tool will act in 2 levels, offering: (a) On-line process: total statistical information such as number of visits per course (pages and sessions), user trends and activities at their visits, as well as detailed information per student (user duration per course and activity, user preferences and activities for all courses), and (b) Off-line process: with the use of data mining techniques such as pre-process, visualization, clustering, classification, regression and association, the tool will eventually discover hidden data patterns.

6. CONCLUSION

This paper proposes a framework for analyzing data from LMS. The main advantages of the framework are that: i) It uses data mining techniques for user and course evaluation; ii) it proposes new indexes and metrics to be used with data mining algorithms; iii) it can be easily adapted to any LMS.

Its main disadvantage is that it is time consuming and cumbersome, due to the usage and application of advanced data mining techniques. Furthermore, it may be difficult to use it and it may lead an instructor without the knowledge of basic data mining to non accurate conclusions. To overcome such limitation we are developing a plug-in tool to automate the data mining process. The application of the proposed framework and tool will assist the instructors to improve the quality of their courses with the use of appropriate wizard GUI.

ACKNOWLEDGEMENT

The authors would like to thank teaching personnel of the Department of Information Management at TEI Kavalas, for their gentle permission in accessing the log data of Open eclass (LMS). We would also like to thank associate prof. Mr. Mardirios V. for the technical assistance and provisions for the installation of appropriate software used by our tool for logging purposes at the web server of Dept. of I.M., TEI Kavalas.

REFERENCES

Baraglia R. and Palmerini P., 2002. SUGGEST : A Web Usage Mining System, *Proceedings of IEEE International Conference on Information Technology: Coding and Computing*.

Bing L., 2007. Web Data Mining: Exploring Hyperlinks, Contents and Usage Data, Springer, ISBN 3540378812.

Claroline, 2009. Available at http://www.claroline.net (10/08/2009).

GUNet, 2009. OPENeCLASS. Available at http://www.openeclass.org/ (10/08/2009)

Kosala R. and Blockeel H., 2000. Web Mining Research: A Survey, *SIGKDD Explorations*, Vol 2, No 1, pp 1-15.

Romero C. and Ventura S. (2007), Educational data mining: A survey from 1995 to 2005, *Elsevier journal of Expert Systems with Applications*.

Romero C., et al. 2008. Data Mining in course management systems: Moodle case study and tutorial. *Computers & Education*, Vol. 51, No. 1, pp. 368-384.

Spiliopoulou M., 1999. Data mining for the web, In Principles of Data Mining and Knowledge Discovery, *Second European Symposium, PKDD '99*, pp. 588-589.

Srivastava J., et al., 2000. Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, *ACM SIGKDD Explorations Newsletter*, Vol. 1, No. 2, pp. 12-23.

Witten I. and Eibe F. (2000), *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, 2nd Edition, Morgan Kaufmann, San Francisco,

Yan T. W., et al., 1996. From user access patterns to dynamic hypertext linking. *Proceedings of the Fifth International World Wide Web Conference*.

Zaiane, O.R., 2001. Web Usage mining for a better web-based learning environment. *Proceedings of Conference on Advanced Technology for Education*.

Zorrilla, M. E., and Álvarez, E., 2008. MATEP: Monitoring and Analysis Tool for e-Learning Platforms, *Proceedings of the 2008 Eighth IEEE International Conference on Advanced Learning Technologies*, pp. 611-613.