Web Usage Mining Project for Improving Web-Based Learning Sites

M.E. Zorrilla[1], E. Menasalvas[2], D. Marín[1], E. Mora[1], and J. Segovia[2]

[1] Department of Applied Mathematics and Computer Sciences, University of Cantabria,
Avda. de los Castros s/n 39005 Santander. Spain
{zorrilm, morae}@unican.es, dmoujo@yahoo.es
[2] Dpto de Lenguajes y Sistemas Informáticos. Facultad de Informática,
Universidad Politécnica de Madrid. Boadilla del Monte. Spain
{fsegovia, emenasalvas}@fi.upm.es

Abstract. Despite the great success of data mining being applied for personalization in web environments, it has not yet been massively applied in the e-learning domains. In this paper, we outline a web usage mining project which has been initiated in University of Cantabria. The aim of this project is to develop tools which let us improve its Web-based learning environment in two main aspects: the first that the teacher obtains information which allows him to evaluate the learning process and the second that the student feels supported in this task.

1 Introduction and Background

It can be said that, the World Wide Web is today the most important media for collecting, sharing and distributing information. Higher Education (HE) is one of the fields where web-based technology has been quickly and successfully adopted. The great proposal of online courses that, nowadays, is offered by universities is one proof of that. Even more, completely virtual universities are appearing.

Managing and tracking students, designing courses, making evaluations, etc. requires specific systems which are called Learning Management Systems (LMS). These systems can be organized in 3 subsets according to Jackson [5]: Course Management Systems (CMS), Enterprise Learning Management (ELMS) and Learning Content Management Systems (LCMS).

CMS facilitate web delivery and management for instructor-led topics and include conferencing systems, polling and quiz modules, virtual workspaces and other tools for measuring outcomes and reporting progress for individual or groups of students. They tend to be very textual and template oriented to provide ease of use, but limiting flexibility. These systems are the most popular in HE institutions (85% according to Gartner). Examples include Blackboard, Virtual-U or WebCT. ELMS and LCMS are more expensive and require significant customization. They typically add strong integrated authoring tools and components to connect to database systems.

Many CMS have been developed and are in use around the world. However they do not support tools which allow educators to thoroughly track and assess all the

R. Moreno Díaz et al. (Eds.): EUROCAST 2005, LNCS 3643, pp. 205 – 210, 2005.

activities performed by all learners, nor to evaluate the structure of the course content and its effectiveness on the learning process. In fact, these environments provide the educator with access summary information such as more visited pages, favourite communication method, and other statistics. Nevertheless this information is not enough to analyze the behaviour of each student and his evolution.

The problem is that E-learning environments lack a closer student-educator relationship. The lack of this relation is manifested in facts such as: teacher does not really control the evolution of his students, and students cannot express their problems and deficiencies in a natural way.

This problem has yet been tackled in marketing environments using web mining techniques. In [1], an architecture that successfully integrated data mining with an e-commerce system is shown. In [6] a methodology for evaluating and improving the "success" of a commercial web site based on the exploitation of navigation pattern discovery is proposed. In [3] an overview of approaches for incorporating semantic knowledge into Web Usage Mining and personalization processes is provided.

Results of the application of Web Mining in e-commerce have not been massively applied in e-learning environments while web-based learning systems can profit from them [4]. In this direction, our project tries to solve the presented problems by pre-processing and analysing the web log files, which provide a raw trace of the learners' navigation and activities on the site, using OLAP and data mining techniques [8] in order to extract valuable patterns that will be used to enhance the learning system and help in the learning evaluation.

Thus in this paper we present advances of an e-learning project in which OLAP techniques are applied to obtain data that later will be used to improve the relationship between professor and student.

The rest of this paper is organized as follows: Section 2 briefly presents the objectives and main tasks of the E-learning project which is being developed in University of Cantabria (UC) with the collaboration of Universidad Politécnica de Madrid (UPM). Section 3 presents obtained results in the OLAP analysis. Finally, Section 4 provides conclusions and future work.

2 E-Learning Project

This project initially springs up with the aim to give concrete answers to professors who compromised with these new methods of learning based on new technologies do not get the appropriate feedback compared to the feedback you get from students with traditional teaching methods. Besides, we have in mind other goals which will help administrators and academic responsible to do better their task.

2.1 Objectives

On the one hand, professors will have information that provides them with tracking information to assess the learning process of their students. The system will also provide professor with the most common navigation patterns in their courses that will help them to evaluate their courses structure effectiveness.

On the other hand, learners will obtain a personalized environment that in near future, will recommend them activities and/or resources that would favour and improve their learning

An added value of these tools will be that the site administrator will have parameters to improve the site efficiency and adapt it to the behaviour of their users.

Academic responsible will have information which allows them to know their student profile. It will provide them with measures to better organize their resources, both human and material, and their educational offer.

2.2 Scheduled Tasks

We briefly describe the main five tasks we have considered:

1. Data pre-processing: clean and prepare the web server log file and load the clean data into a relational database.
2. OLAP analysis: design a multidimensional structure in which the main factors under analysis: sessions, courses, pages, time, demographical user data, user behaviour (content or navigational) will be taken as dimensions and later build OLAP cubes in order to analyze the recorded data.
3. Pattern discovery: application of data mining algorithms. Firstly descriptive algorithms (clustering and association) will be applied to obtain typologies of users. Then classification techniques will be used in a later step to classify behaviours according to historical patterns.
4. Pattern evaluation: All the patterns obtained will be valuated to distinguish the patterns that really help to better achieve the site goals.
5. Recommendation engine: integrate the discovered patterns in the online component to provide personalized environment to learners.

In the next section results obtained with the OLAP components are summarized.

3 OLAP Analysis

Although the project main aim is not only to generate OLAP reports but also data mining analysis, in this paper we focus on the OLAP analysis. Consequently results obtained so far are presented in what follows.

3.1 Pre-processing

Web server logs are the primary source of data in which the activities of Web users are captured, although often can be enriched with external information obtained from corporative database systems. These files, due to their huge size and their lack of structure, require to be processed, this means, to be read and recorded in a relational database to be easily managed.

Pre-processing [2] is, in fact, the first part of Web Usage Mining (WUM) which includes the domain dependent tasks of data cleaning, user identification, page identification, session identification and path completion. Although there are free tools such us [7] which allow cleaning and making sessions, they have not been suitable for us because they work with general logs (commercial environments) and they do not allow us to configure specific e-learning characteristics, for this reason we have developed our own tool.

The example we use for illustration in this paper comes from a WebCT web log with records of students' on-line activities in all courses from October 1st, 2004 to December 31st, 2004. There are near 10.000.000 entries in this web log file of a size of 1.1 Gigabytes. After cleaning process, the number of entries was reduced to 2.206.024. Next, 3.235 learners and 322 courses were identified and finally, 48.691 sessions were built. In our case, a new session was considered when a change in a user-course happened or when the time interval between two successive inter-transaction clicks upped 30 minutes. In this last step, the duration of the visit and the number of visited pages were calculated. Besides, the visited pages during each session were registered to be used in the incoming navigational behaviour analysis.

3.2 OLAP Analysis

The multi-dimensional structure of the data cube provides remarkable flexibility to manipulate the data and view it from different perspectives. Building a web log data cube allows the application of OLAP (On-Line Analytical Processing) operations to view and analyze the web log data from different angles, derive ratios (average stay-session time, etc) and compute measures across many dimensions.

The first step towards analysing user behaviour in an e-learning system is to be able to answer questions as the ones that follow:

- How long is our learner connected (by course, degree, month, etc)?
- What is the connected learner distribution over time (hour of the day, day of the week, month and year)?
- How many learning sessions do our learners establish over time?
- Which courses are the most frequent acceded?
- Which is the percentage of connections done inside university campus?
- What is the distribution of network traffic over time?

Consequently after the data was cleaned and transformed, a multi-dimensional structure was designed. From this, a MOLAP data cube was built to aggregate the number of visited pages, the number of error pages and the visit duration according to the following five dimensions: date, time, courses, sessions and learners.

Not only the questions above but any question involving the dimensions of the OLAP structure will be answered. Results of some of these questions are shown.

Example 1. Usage pattern analysis

The session analysis lets us understand how the system was used, how the course structure was designed and how the learners' behaviour evolved over time.

Fig. 1a and Fig. 1b show sessions per learner, average session time and pages per session measures. The first one shows the results taking into account all courses in this period (in weeks) and the second one, the same values but for one of the most frequent acceded and well-designed course.

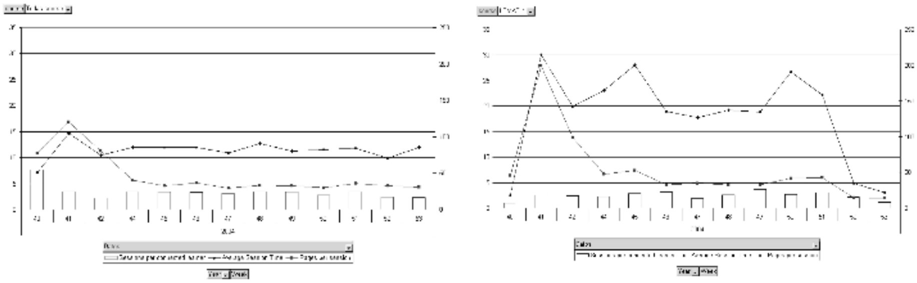

Fig. 1. (a) Usage pattern analysis of the full system, and (b) for a well-designed course

One phenomenon we have discovered is that at the beginning of a course the students tended to explore many different system features (more visited pages per session). However, they became more and more focused over time.

In Fig. 1a, it can be observed how the average connection time by student in this term is low, less than 15 minutes, nevertheless in Figure 1b, this time ups to 20 or 30 minutes. This suggests us that most of the courses have been designed as repositories of contents, i.e., professors have designed html pages with several links to PDF or zip files, so that, students only need connect to the system to download these files. On the other hand, this reduced time indicates us that students do not use very often collaborative tools (chats, mail …) because the interaction requires longer sessions.

Example 2. Learner distribution over time

Another interesting information consists on analysing the distribution over time of the number of connected students depending on the week day and time of the day. It is easy to see in Fig. 2 how, in this study, students were highly connected from outside UC (value IN) in working days (value 1) and from 12 to 20 hours.

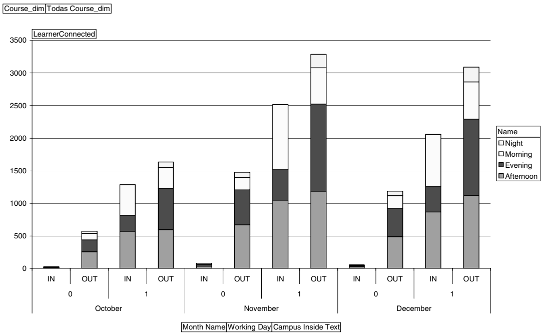

Fig. 2. Learner distribution over time

In relation to these results we can also say that analysed courses are chiefly exploited as a complementary tool as the main activity begins around November. This shows that students need have minimum knowledge about the subject before accessing to the WebCT platform for further knowledge.

Besides, we have observed that the number of connected students versus registered students is around 50% in the case of not completely virtual courses. This can be interpreted as students not connecting to the system unless the activity is completed integrated in the course syllabus.

4 Conclusions and Future Work

In this paper we have presented an e-Learning WUM project which is being developed in UC using its own data. This project is divided in five stages, where two of them have been yet done. Likewise, the first relevant results have been shown.

Our experience shows us that pre-processing step although being time consuming is crucial for the success of the discovery process. A good data cleaning and filtered process needs metadata provided by web site designers and a good knowledge about how the LMS works because generally are environments based on scripts.

Also we can say that, the multi-dimensional structure of the data cube provides remarkable flexibility to manipulate the data and view it from different perspectives. Besides, as it can be managed with MS Excel, teachers and system administrators can evaluate easily the system use.

Our next step will be to select a course whose design allow us to analyse learner navigational behaviour and compare it with professor intention. For that rule association and sequential algorithms will be used. Also its effectiveness in the learning process will be evaluated. On the other hand, we will try to complete learner information and, applying descriptive algorithms, obtain typologies of students. Further development and experiments will be reported in the future.

References

1. Ansari, S., Kohavi, R., Mason, L., Zheng, Z. Integrating E-Commerce and Data Mining: Architecture and Challenges. Workshop on Web Mining for E-Commerce-Challenges and Opportunities Working Notes (KDD 2000), Boston, MA.
2. Cooley, R., Mobasher, B., Srivastava, J. Data Preparation for Mining World Wide Web Browsing Patterns. Jounal of Knowledge an Information Systems, 1(1). 1999.
3. Dai, H., Mobasher, B. A road map to more effective web personalization: Integrating domain knowledge with web usage mining. In Proc. of the International Conference on Internet Computing 2003, Las Vegas, Nevada, June 2003.
4. Hanna, M. Data Mining in the e-learning domain. Campus Wide Information Systems. 16 Jaunary 2004. vol. 21 no. 1 pp. 29-34(6). Research article ISSN: 1065-0741
5. Jackson, R. An Overview of Web-Based Learning. 2004. http://www.knowledgeability.biz/weblearning/
6. Spiliopoulou, M., Pohle, C. Data mining for measuring and improving the success of web sites. Journal of Data Mining and Knowledge Discovery, Special Issue on E-commerce, volume 5(1-2), pages 85-114. Kluwer Academic Publishers, Jan.-Apr. 2001.
7. Web Utilization Miner WUM 7.0 Beta.
8. Zaïane, O. Web Usage Mining for a Better Web-Based Learning Environment. Proc. of Conference on Advantage Technology for Education. Alberta, Canada. 2001.