Available online at www.sciencedirect.com
ScienceDirect
Computers & Education xxx (2007) xxx–xxx
COMPUTERS & EDUCATION
www.elsevier.com/locate/compedu

Data mining in course management systems: Moodle case study and tutorial

Cristóbal Romero *, Sebastián Ventura, Enrique García

Department of Computer Sciences and Numerical Analisys, University of Córdoba, 14071 Córdoba, Spain

Received 5 March 2007; received in revised form 19 May 2007; accepted 25 May 2007

Abstract

Educational data mining is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from the educational context. This work is a survey of the specific application of data mining in learning management systems and a case study tutorial with the Moodle system. Our objective is to introduce it both theoretically and practically to all users interested in this new research area, and in particular to online instructors and e-learning administrators. We describe the full process for mining e-learning data step by step as well as how to apply the main data mining techniques used, such as statistics, visualization, classification, clustering and association rule mining of Moodle data. We have used free data mining tools so that any user can immediately begin to apply data mining without having to purchase a commercial tool or program a specific personalized tool.

Keywords: Distance education and telelearning; E-learning; Evaluation of CAL systems; Data mining; Web mining

1. Introduction

Course management systems (CMSs) can offer a great variety of channels and workspaces to facilitate information sharing and communication among participants in a course. They let educators distribute information to students, produce content material, prepare assignments and tests, engage in discussions, manage distance classes and enable collaborative learning with forums, chats, file storage areas, news services, etc. Some examples of commercial systems are Blackboard (BlackBoard, 2007), WebCT (WebCT, 2007) and TopClass (TopClass, 2007) while some examples of free systems are Moodle (Moodle, 2007), Ilias (Ilias, 2007) and Claroline (Claroline, 2007). Nowadays, one of the most commonly used is Moodle (modular object oriented developmental learning environment), a free learning management system enabling the creation of powerful, flexible and engaging online courses and experiences (Rice, 2006).

These e-learning systems accumulate a vast amount of information which is very valuable for analyzing students' behaviour and could create a gold mine of educational data (Mostow & Beck, 2006). They can record

* Corresponding author. Fax: +34 957 218630.
E-mail address: cromero@uco.es (C. Romero).

0360-1315/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compedu.2007.05.016

any student activities involved, such as reading, writing, taking tests, performing various tasks, and even communicating with peers (Mostow et al., 2005). They normally also provide a database that stores all the system's information: personal information about the users (profile), academic results and users' interaction data. However, due to the vast quantities of data these systems can generate daily, it is very difficult to manage manually. Instructors and course authors demand tools to assist them in this task, preferably on a continual basis. Although some platforms offer some reporting tools, it becomes hard for a tutor to extract useful information when there are a great number of students, (Dringus & Ellis, 2005). They do not provide specific tools allowing educators to thoroughly track and assess all learners' activities while evaluating the structure and contents of the course and its effectiveness for the learning process (Zorrilla, Menasalvas, Marin, Mora, & Segovia, 2005). A very promising area for attaining this objective is the use of data mining (Zaïane & Luo, 2001).

In the last few years, researchers have begun to investigate various data mining methods to help instructors and administrators to improve e-learning systems (Romero & Ventura, 2006). Data mining or knowledge discovery in databases (KDD) is the automatic extraction of implicit and interesting patterns from large data collections (Klosgen & Zytkow, 2002). Data mining is a multidisciplinary area in which several computing paradigms converge: decision tree construction, rule induction, artificial neural networks, instance-based learning, bayesian learning, logic programming, statistical algorithms, etc. And some of the most useful data mining tasks and methods are statistics, visualization, clustering, classification and association rule mining. These methods uncover new, interesting and useful knowledge based on students' usage data. Some of the mains e-learning problems or subjects to which data mining techniques have been applied (Castro, Vellido, Nebot, & Mugica, in press) are dealing with the assessment of student's learning performance, provide course adaptation and learning recommendations based on the students' learning behaviour, dealing with the evaluation of learning material and educational web-based courses, provide feedback to both teachers and students of e-learning courses, and detection of atypical student's learning behaviour.

Nowadays, data mining tools are normally designed more for power and flexibility than for simplicity. Most of the current data mining tools are too complex for educators to use and their features go well beyond the scope of what an educator might require. As a result, the CMS administrator is more likely to apply data mining techniques in order to produce reports for instructors who then use these reports to make decisions about how to improve the student's learning and the online courses.

This knowledge, however, can be useful not only to the providers (educators) but also to the users themselves (students), as it can be oriented towards different ends for different partakers in the process (Zorrilla et al., 2005). It could be oriented towards students in order to recommend learners' activities, resources, suggest path pruning and shortening or simply links that would favor and improve their learning or to educators in order to get more objective feedback for instruction. Instructors can evaluate the structure of course content and its effectiveness in the learning process and also classify learners into groups based on their needs for guidance and monitoring. Learners' regular and irregular patterns could be determined allowing the most frequently made mistakes to be identified and more effective activities to be elaborated. There could be more orientation towards obtaining parameters and measures to improve site efficiency and adapt it to the behaviour of the users (optimal server size, network traffic distribution, etc.) and to organize better institutional resources (human and material) and educational offer.

Data mining has been applied to data coming from different types of educational systems. On one hand, there are traditional face-to-face classroom environments such as special education (Tsantis & Castellani, 2001) and higher education (Luan, 2002). On the other, there is computer-based education as well as web-based education such as well-known learning management systems (Pahl & Donnellan, 2003), web-based adaptive hypermedia systems (Koutri, Avouris, & Daskalaki, 2005) and intelligent tutoring systems (Mostow & Beck, 2006). The main difference between one and the other is the data available in each system. Traditional classrooms only have information about student attendance, course information, curriculum goals and individualized plan data. However, computer and web-based education has much more information available because these systems can record all the information about students' actions and interactions onto log files and databases.

This paper is oriented to the specific application of data mining in computer-based and web-based educational systems (in particular, course management systems). It is arranged in the following way: Section 2 describes the general process of applying data mining to e-learning data, especially to Moodle usage

Please cite this article in press as: Romero, C. et al, Data mining in course management systems: Moodle case ..., *Computers & Education* (2007), doi:10.1016/j.compedu.2007.05.016

information. Section 3 details the preprocessing step necessary for adapting the data to the appropriate format. Section 4 describes the application of the main data mining techniques in e-learning and an example case study with Moodle data. Finally, the conclusions and further research are outlined.

2. Process of data mining in e-learning

The traditional development of e-learning courses is a laborious activity (Herin, Sala, & Pompidor, 2002). The developer (usually the course teacher or online instructor) has to choose the contents that will be shown, decide on the structure of the contents, and determine the most appropriate content elements for each type of potential user of the course. Due to the complexity of these decisions, a one-shot design is hardly feasible, even when carefully done. Instead, most cases will probably need evaluation and possibly modification of course content, structure and navigation based on students' usage information, preferably even following a continuous empirical evaluation approach (Ortigosa & Carro, 2003). To facilitate this, data analysis methods and tools are used to observe students' behaviour and to assist instructors in detecting possible errors and shortcomings and in incorporating possible improvements. Traditional data analysis in e-learning is hypothesis or assumption driven (Gaudioso & Talavera, 2006) in the sense that the user starts from a question and explores data to confirm his intuition. While this can be useful when a moderate number of factors and data are involved, it can be very difficult for the user to find more complex patterns that relate to different aspects of the data. An alternative to this traditional data analysis is to use data mining in an inductive approach to automatically discover hidden information present in the data. Data mining, in contrast, is discovery-driven in the sense that the hypothesis is automatically extracted from the data and therefore is data-driven rather than research-based or human-driven (Tsantis & Castellani, 2001). Data mining builds analytic models that discover interesting patterns and tendencies in student's usage information.

The application of data mining in e-learning systems is an iterative cycle (Romero & Ventura, 2007). The mined knowledge should enter the loop of the system and guide, facilitate and enhance learning as a whole, not only turning data into knowledge, but also filtering mined knowledge for decision making. The e-learning data mining process consists of the same four steps in the general data mining process (see Fig. 1) as follows:

- *Collect data*: The CMS system is used by students and the usage and interaction information is stored in the database. In this paper we have used the students' usage data in the Moodle system.
- *Preprocess the data*: The data is cleaned and transformed into an appropriate format to be mined. In order to preprocess the Moodle data, we can use a database administrator tool or some specific preprocessing tool.

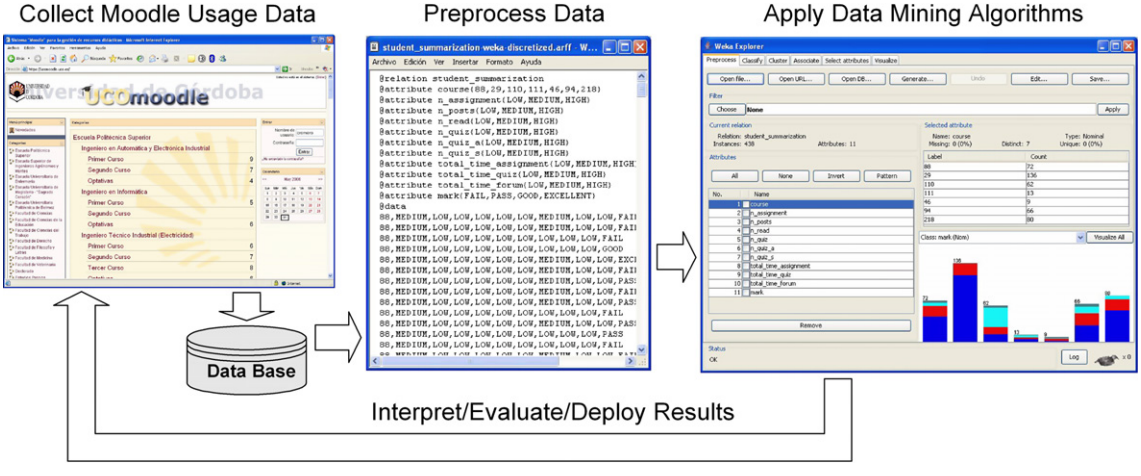


Fig. 1. Mining Moodle data.

Please cite this article in press as: Romero, C. et al, Data mining in course management systems: Moodle case ..., *Computers & Education* (2007), doi:10.1016/j.compedu.2007.05.016

– *Apply data mining*: The data mining algorithms are applied to build and execute the model that discovers and summarizes the knowledge of interest to the user (instructor, student and administrator). To do so, either a general or a specific data mining tool, or a commercial or free data mining tool can be used.
– *Interpret, evaluate and deploy the results*: The results or model obtained are interpreted and used by the instructor for further actions. The instructor can use the information discovered to make decisions about the students' and Moodle course activities to improve the students' learning.

3. Preprocessing Moodle data

Moodle (Moodle, 2007) is an open-source course management learning system to help educators create effective online learning communities. It is an alternative to proprietary commercial online learning solutions, and is distributed free under open source licensing. Moodle has been installed at universities and institutions all over the world (Cole, 2005). An organization has complete access to the source code and can make changes if need be. Its modular design makes it easy to create new courses, adding content that will engage learners and it is designed to support a style of learning called social constructionist pedagogy (Rice, 2006). This style of learning believes that students learn best when they interact with the learning material, construct new material for others, and interact with other students about the material. Moodle does not require the use of this style in the courses but this style is what it best supports, and it has a flexible array of module activities and resources to create five types of static course material (a text page, a web page, a link to anything on the Web, a view into one of the course's directories and a label that displays any text or image), as well as six types of interactive course material (assignments, choice, journal, lesson, quiz and survey) and five kinds of activities where students interact with each other (chat, forum, glossary, wiki and workshop).

Moodle keeps detailed logs of all activities that students perform (Rice, 2006). It logs every click that students make for navigational purposes and has a modest log viewing system built into it (see Fig. 2). Log files can be filtered by course, participant, day and activity. The instructor can use these logs to determine who has been active in the course, what they did, and when they did it. For activities such as quizzes, not only the score

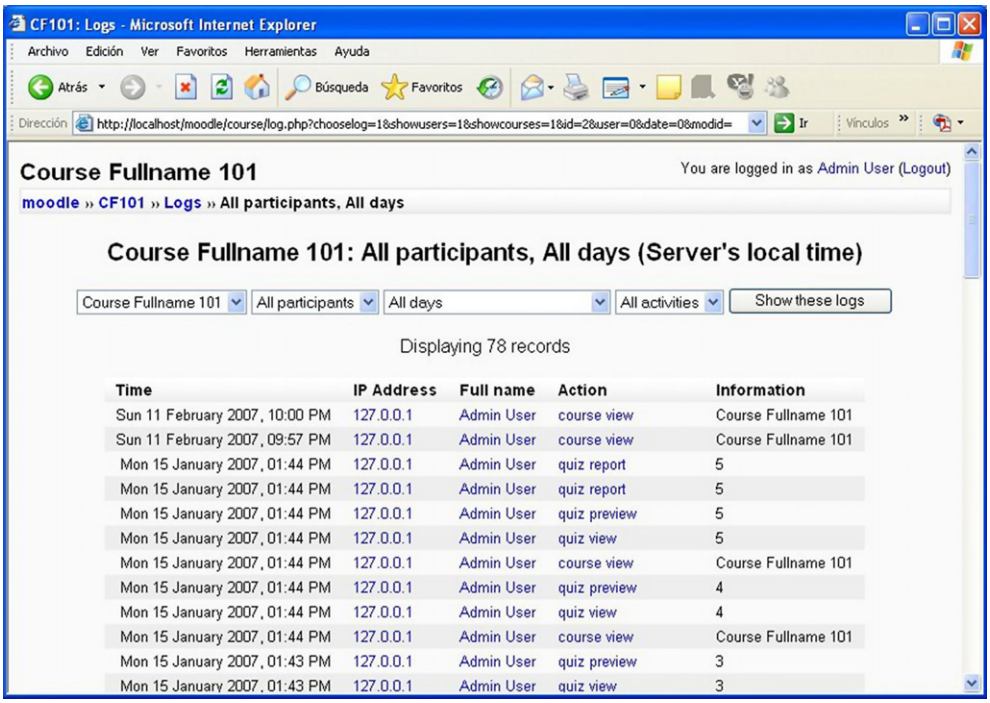


Fig. 2. Moodle log report screen.

Please cite this article in press as: Romero, C. et al, Data mining in course management systems: Moodle case ..., *Computers & Education* (2007), doi:10.1016/j.compedu.2007.05.016

and elapsed time are available, but also a detailed analysis of each student's responses and item analysis of the items themselves. Instructors can easily get full reports of the activities of individual students, or of all students for a specific activity. Activity reports for each student are available and details about each module (last access, number of times read) as well as a detailed story of each student's involvement. Logs can show the activity in the class for different days or times. This can be useful to check to see if everyone has done a certain task, or spent a required amount of time online within certain activities.

Moodle does not store logs as text files. Instead, it stores the logs in a relational database (MySQL and PostgreSQL are the best supported, but it can also be used with Oracle, Access, Interbase, and others). Databases are more powerful, flexible and bug-prone than typical log text files for gathering detailed access and high level usage information from all the services available in the CMS. The Moodle database has about 145 interrelated tables. But we do not need all this information and it is also necessary to convert it into the required format used by data mining algorithms. For this reason, we have to perform a previous step to preprocess Moodle data. Data preprocessing allows the original data to be transformed into a suitable shape to be used by a particular data mining algorithm or framework. It is important to notice that preprocessing tasks are normally done by administrators and not by instructors themselves. This is normally a manual process in which the administrator has to apply a number of general data preprocessing tasks (data cleaning, user identification, session identification, path completion, transaction identification, data transformation and enrichment, data integration, data reduction). In our case, data preprocessing of CMS is a little more simple due to Moodle and most CMS employ user authentication (password protection) in which logs have entries identified by users since users have to log-in. In this way, sessions are already identified since users may also have to log-out and this eliminates the need for typical user and session identification tasks. So, the data gathered by a CMS may require less cleaning and preprocessing than data collected by other web-based systems. Although the amount of work required in data preparation is less, the following tasks also need to be done:

- *Select data*: It is necessary to choose which courses mining can be useful for. Although information is available about 4223 students in 192 courses corresponding to different Moodle courses in the University of Cordoba, instructors normally use only assignments. So, we have chosen only 7 courses from among all these courses because they use a higher number of Moodle activities and resources (at least assignments, messages, forums and quizzes) and the final marks obtained by students only in these courses are also available. So, the total number of students that we have used in this study is 438 students.
- *Create summarization tables*: It is necessary to create a new table in the Moodle database that can summarize information at the required level (e.g. student). As student and interaction data are spread over several tables, a summary table has been created (see Table 1) which integrates the most important information for our objective. This table (*mdl_summarization*) has a summary per row about all the activities done by each student during the course and the final mark obtained by the student in the course in question.

Table 1
Attributes used for each student

Name	Description
Course	Identification number of the course
n_assignment	Number of assignments done
n_quiz	Number of quizzes done
n_quiz_a	Number of quizzes passed
n_quiz_s	Number of quizzes failed
n_messages	Number of messages sent to the chat
n_messages_ap	Number of messages sent to the teacher
n_posts	Number of messages sent to the forum
n_read	Number of messages read on the forum
total_time_assignment	Total time spent in assignment
total_time_quiz	Total time used on quizzes
total_time_forum	Total time used on forum
Mark	Final mark the student obtained in the course

Please cite this article in press as: Romero, C. et al, Data mining in course management systems: Moodle case ..., *Computers & Education* (2007), doi:10.1016/j.compedu.2007.05.016

- *Data discretization*: Performing a discretization of numerical values may be necessary to increase interpretation and comprehensibility. Discretization (Dougherty, Kohavi, & Sahami, 1995) divides the numerical data into categorical classes that are easier to understand for the instructor (categorical values are more user-friendly for the instructor than precise magnitudes and ranges). All numerical values of the summarization table have been discretized (*mdl_summarization_discretized*) except for the course identification number. The manual method (in which the user has to specify the cut-off points) is used in the mark attribute with four intervals and labels (FAIL if value is <5; PASS if value is >5 and <7; GOOD if value is >7 and <9; and EXCELLENT if value is >9) and the equal-width method (divides the range of the attribute into a fixed number of intervals of equal length) in all the other attributes with three intervals and labels (LOW, MEDIUM and HIGH).
- *Transform the data*: The data must be transformed to the required format of the data mining algorithm or framework. In our case, the *mdl_log* Moodle table and the two versions of the summary table (with numerical and categorical values) have been exported to text files with ARFF format. An ARFF (Attribute-Relation File Format) file is an ASCII text file that describes a list of instances sharing a set of attributes (Witten & Frank, 2005).

4. Applying data mining techniques to Moodle data

Data mining is the process of efficient discovery of non-obvious valuable patterns from a large collection of data (Klosgen & Zytkow, 2002). There are a lot of general and specific data mining tools and frameworks. Some examples of commercial mining tools are DBMiner (DBMiner, 2007), SPSS Clementine (Clementine, 2007) and DB2 Intelligent Miner (Miner, 2007). And some examples of public domain mining tools are Weka (Weka, 2007) and Keel (Keel, 2007). There are also some specific educational data mining tools such as the Mining tool (Zaïane & Luo, 2001) for association and pattern mining, MultiStar (Silva & Vieira, 2002) for association and classification, Tool (Chang, Hung, & Shih, 2003) for performing a quantitative analysis based on students' learning performance, EPRules (Romero, Ventura, & Bra, 2004) for association, KAON (Tane, Schmitz, & Stumme, 2004) for clustering and text mining, Synergo/ColAT (Avouris, Komis, Fiotakis, Margaritis, & Voyiatzaki, 2005) for statistics and visualization, GISMO (Mazza & Milani, 2005) for visualization, Listen tool (Mostow et al., 2005) for visualization and browsing, TADA-Ed (Merceron & Yacef, 2005) for visualizing and mining, O3R (Becker, Vanzin, & Ruiz, 2005) for sequential pattern mining, MINEL (Bellaachia, Vommina, & Berrada, 2006) for mining learning paths, CIECoF (García, Romero, Ventura, & Castro, 2006) for association rule mining, Simulog (Bravo & Ortigosa, 2006) for looking for unexpected behavioural patterns, and Sequential Mining tool (Romero et al., in press) for pattern mining.

In this paper, we used Weka and Keel systems because they have what we consider to be three important characteristics in common: they are free software systems, they have been implemented in Java language and they use the same dataset external representation format (ARFF files). So, they can easily be obtained from Internet, used without data format problems and, if necessary, modified using the same programming language. Weka (Witten & Frank, 2005) is open source software that provides a collection of machine learning and data mining algorithms for data pre-processing, classification, regression, clustering, association rules, and visualization. Keel (Alcalá et al., in press) is an open source software tool developed to build and use different data mining models such as pre-processing algorithms, decision trees, rule extraction, descriptive induction, statistical methods, neural networks and evolutionary multi-classifier systems.

Data mining has a number of representation formalisms such as probabilities, rules and trees. and a number of tasks and methods from machine learning, statistics, visualization and artificial intelligence (Klosgen & Zytkow, 2002). Next, examples are given of the specific application of data mining techniques in e-learning systems and Moodle as the case study, grouped into different categories.

4.1. Statistics

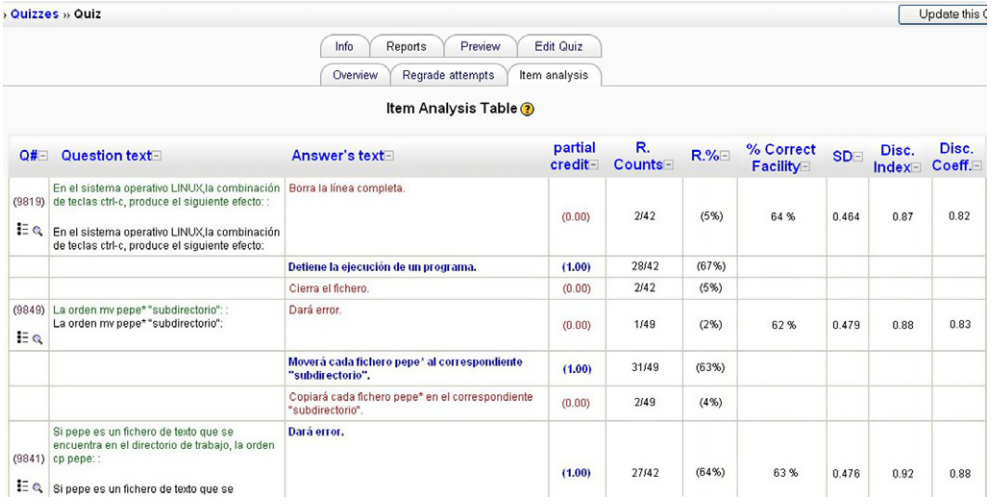
Student's usage statistics are often the starting point for evaluation in an e-learning system (Zaïane, Xin, & Han, 1998). Usage statistics may be extracted using standard tools designed to analyze web server logs such as

Please cite this article in press as: Romero, C. et al, Data mining in course management systems: Moodle case ..., *Computers & Education* (2007), doi:10.1016/j.compedu.2007.05.016

AccessWatch, Analog, Gwstat and WebStat. However, there are specific statistical tools in educational data such as Synergo/ColAT (Avouris et al., 2005). Some examples of usage statistics in e-learning systems are simple measurements, such as number of visits and visits per page (Pahl & Donnellan, 2003). Some statistics show the connected learner distribution over time and the most frequently accessed courses (Nilakant & Mitrovic, 2005), others show visits and their duration per quarter, the top referrer and top search terms (Grob, Bensberg, & Kaderali, 2004), and still others show reports about help systems (Feng & Heffernan, 2006) and reports of weekly and monthly user trends and activities (Monk, 2005). There are also specific statistics that show the average number of constraint violations and average problem complexity (Zorrilla et al., 2005). More advanced statistical methods, such as correlation analysis between variables, have been used to infer students' attitudes that affect learning (Arroyo, Murray, Woolf, & Beal, 2004), or for predicting the final exam score (Pritchard & Warnakulasooriya, 2005). At the same time, regression analyses have been used to predict a student's knowledge and which metrics help to explain the poor prediction of state exam scores (Feng, Heffernan, & Koedinger, 2005). Regression has also been applied for predicting whether the student will answer a question correctly enough (Beck & Woolf, 2000), and for predicting end-of-year accountability assessment scores (Anozie & Junker, 2006).

Moodle does not provide a statistics module in which the instructor can obtain specific reports regarding detailed statistics about every single student's performance (how many hours on the site, how much time at every activity, etc.). It would be very useful for attendance reports (total student enrollment in all courses, total student activity for a whole course for a specific instructor, student history of all courses taken, time spent in each and grades for each in chronological order) and financial reports (total enrollment income for specific periods of time, total fees collected from any student for all courses taken, total income generated per student by an instructor for specific periods of time).

Moodle only shows some statistical information in some of the modules (grades and quizzes). On one hand, the instructor can use scales to rate or grade forums, assignments, quizzes, lessons, journals and workshops in order to evaluate students' work. Moodle comes with two preexisting scales, one that is numerical (from 1 to 100) and the other to indicate if an item is connected to other knowledge in the course. But, the instructor can customize grade scales (categorize grades, assign ranges to letter grades, use weighted grades, hide/reveal grades to students) in order to have a powerful way to view the progress of the students. On the other hand, Moodle has statistical quiz reports which show item analysis (see Fig. 3). It presents processed quiz data in a way suitable for analyzing and judging the performance of each question for the function of assessment. The statistical parameters used are calculated as explained by the classical test theory: Facility Index or % Correct (F.I.), Standard Deviation (S.D.), Discrimination Index (D.I.) and Discrimination Coefficient (D.C.). The



Q#	Question text	Answer's text	partial credit	R. Counts	R.%	% Correct Facility	SD	Disc. Index	Disc. Coeff.
(9819)	En el sistema operativo LINUX,la combinación de teclas ctrl-c, produce el siguiente efecto: : En el sistema operativo LINUX,la combinación de teclas ctrl-c, produce el siguiente efecto:	Borra la línea completa.	(0.00)	2/42	(5%)	64 %	0.464	0.87	0.82
		Detiene la ejecución de un programa.	**(1.00)**	28/42	(67%)				
		Cierra el fichero.	(0.00)	2/42	(5%)				
(9849)	La orden mv pepe* "subdirectorio": : La orden mv pepe* "subdirectorio":	Dará error.	(0.00)	1/49	(2%)	62 %	0.479	0.88	0.83
		Moverá cada fichero pepe* al correspondiente "subdirectorio".	**(1.00)**	31/49	(63%)				
		Copiará cada fichero pepe* en el correspondiente "subdirectorio".	(0.00)	2/49	(4%)				
(9841)	Si pepe es un fichero de texto que se encuentra en el directorio de trabajo, la orden cp pepe: : Si pepe es un fichero de texto que se	**Dará error.**	**(1.00)**	27/42	(64%)	63 %	0.476	0.92	0.88

Fig. 3. Moodle item analysis.

Please cite this article in press as: Romero, C. et al, Data mining in course management systems: Moodle case ..., *Computers & Education* (2007), doi:10.1016/j.compedu.2007.05.016

instructor can see which questions are the most difficult and easiest for the students (low and high F.I.) as well as the most discriminating ones (high D.I. and D.C.). This information can also be downloaded in text-only or Excel format in order to use a spreadsheet to chart and analyze it.

Using this information the instructor can carry out a continuous maintenance of the quizzes. For example, the instructor can modify questions (question text or answer text). The instructor can delete/modify questions if they are much too easy or difficult and almost all the students fail them (they can have some syntax/semantic error or they are really very difficult). Or delete/modify questions if almost all the students do them perfectly (they are very easy), and if they are not discerning enough to make a distinction between good and bad students.

4.2. Visualization

Information visualization (Spence, 2001) is a branch of computer graphics and user interface which is concerned with the presentation of interactive or animated digital images so that users can understand data. These techniques facilitate analysis of large amounts of information by representing the data in some visual display. Normally large quantities of raw instance data are represented or plotted as spreadsheet charts, scatter plots and 3D representations. Information visualization can be used to graphically render complex, multidimensional student tracking data collected by web-based educational systems. The information visualized in e-learning can be about complementary assignments, admitted questions, exam scores, etc. (Shen, Yang, & Han, 2002). There are some specific visualization tools in educational data. CourseVis (Mazza & Dimitrova, 2004) visualizes data from a java on-line distance course inside WebCT. GISMO (Mazza & Milani, 2005) uses Moodle students' tracking data as source data, and generates graphical representations that can be explored by course instructors. Listen tool (Mostow et al., 2005) browses vast student–tutor interaction logs from Project LISTEN's automated Reading Tutor. Using these tools, instructors can manipulate the graphical representations generated, which allow them to gain an understanding of their learners and become aware of what is happening in distance classes.

Moodle does not provide visualization tools of student usage data; it only provides text information (log reports, items analysis, etc.). But we can download and install GISMO (Gismo, 2007) into our Moodle system. GISMO is a graphical interactive student monitoring and tracking system tool that extracts tracking data from Moodle. It generates graphical representations that can be explored by course instructors to examine various aspects about distance students. GISMO provides different types of graphical representations and reports, such as graphs reporting the student's access to the course (see Fig. 4), graphs reporting all students' accesses to course resources, graphical representations of discussions pertaining to a course and graphs reporting data from the evaluation tools.

The image in Fig. 4 represents the global number of accesses made by students (in *X*-axis) to all the resources of the course (*Y*-axis). If the user clicks with the right mouse button on one of the bars of the histogram and selects the item "Details", he can see the details for a specific student (the resource sequence order inside the course). Using this graph, the instructor has an overview of the global access made by students to the course with a clear identification of patterns and trends, as well as information about the attendance of a specific student in the course. Starting from this information, the instructor can more easily detect students with some learning problems. For example, students with a very low number of accesses (the first student and the last but one student in the graph in Fig. 4), a very low number of assignments and a very low number of quizzes can be detected quickly.

4.3. Clustering

Clustering is a process of grouping objects into classes of similar objects (Jain, Murty, & Flynn, 1999). It is an unsupervised classification or partitioning of patterns (observations, data items, or feature vectors) into groups or subsets (clusters) based on their locality and connectivity within an n-dimensional space. In e-learning, clustering has been used for: finding clusters of students with similar learning characteristics and to promote group-based collaborative learning as well as to provide incremental learner diagnosis (Tang & McCalla, 2005); discovering patterns reflecting user behaviours and for collaboration management to characterize

Please cite this article in press as: Romero, C. et al, Data mining in course management systems: Moodle case ..., *Computers & Education* (2007), doi:10.1016/j.compedu.2007.05.016

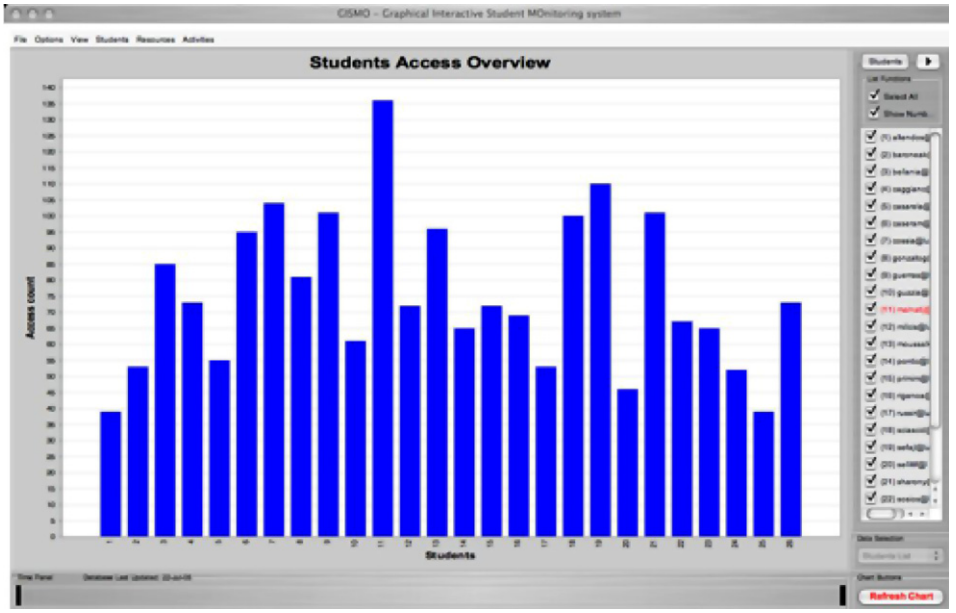


Fig. 4. Students access overview on resources.

similar behaviour groups in unstructured collaboration spaces (Talavera & Gaudioso, 2004); grouping students and personalized itineraries for courses based on learning objects (Mor & Minguillon, 2004); grouping students in order to give them differentiated guiding according to their skills and other characteristics (Hamalainen, Suhonen, Sutinen, & Toivonen, 2004); grouping tests and questions into related groups based on the data in the score matrix (Spacco, Winters, & Payne, 2006); grouping users based on the time-framed navigation sessions (Wang & Shao, 2004).

The Weka system has several clustering algorithms available. The KMeans (MacQueen, 1967), one of the simplest and most popular clustering algorithms, has been used here and it is an algorithm that clusters objects based on attributes in k partitions. In this case, our objective is to group students from a specific course into different clusters depending on the activities done in Moodle along with the final marks. We have used the numerical summarization information about course 218 (Technical Office) omitting two attributes: course (because it is the same in all cases) and mark (because it classifies the cases). We have executed the KMeans over this file with a value of 3 to the number of clusters. Weka shows (Fig. 5) information about the cluster centroids (mean/mode and standard deviation of each attribute) of each cluster, the number and percentage of instances in each cluster.

We can see in Fig. 5 that there are 3 clusters of students. Cluster 0 is characterized by students with no assignments (0), very low messages read (0.045), very few quizzes done, passed and failed (1.54, 1.16 and 0.37) and low total times in assignment, quiz and forum (0, 834.04 and 918.45). Cluster 1 is characterized by students with more than one message sent to the forum (1.22), about 3 messages read (2.8), a high number of quizzes done and passed (7.1 and 6.4), a low number of quizzes failed (0.7), and high total times in assignment, quiz and forum (788.40, 2760.31 and 2191.13). Finally, cluster 2 is characterized by students with values somewhat smaller than cluster 1 but greater than cluster 0. We can also see in the figure that the students are grouped into 3 clusters with uniform numbers of students (24, 22 and 29).

The instructor can use this information in order to group students into three types of students: very active students (cluster 1), active students (cluster 2) and non-active students (0). Starting from this information, for example, the instructor can group students for working together in collaborative activities (each group with only students of the same cluster or each group with a similar number of students of each cluster). The instructor can also group new students into these clusters depending on their characteristics.

Please cite this article in press as: Romero, C. et al, Data mining in course management systems: Moodle case ..., *Computers & Education* (2007), doi:10.1016/j.compedu.2007.05.016

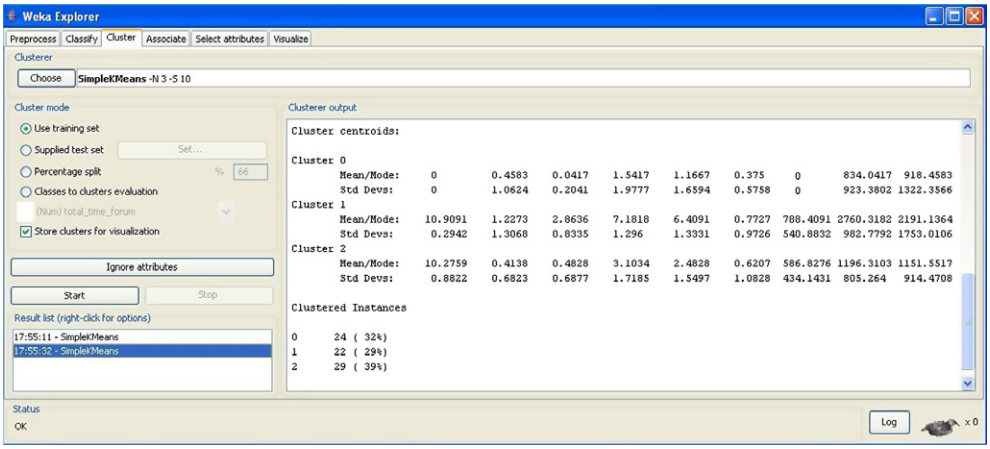


Fig. 5. Weka executing KMeans algorithm.

4.4. Classification

A classifier is a mapping from a (discrete or continuous) feature space X to a discrete set of labels Y (Duda, Hart, & Stork, 2000). Classification or discriminant analysis predicts class labels. This is supervised classification which provides a collection of labeled (preclassified) patterns, the problem being to label a newly encountered, still unlabeled, pattern. In e-learning, classification has been used for: discovering potential student groups with similar characteristics and reactions to a specific pedagogical strategy (Chen, Liu, Ou, & Liu, 2000); predicting students' performance and their final grade (Minaei-Bidgoli & Punch, 2003); detecting students' misuse or students playing around (Baker, Corbett, & Koedinger, 2004); predicting the students' performance as well as to assess the relevance of the attributes involved (Kotsiantis, Pierrakeas, & Pintelas, 2004); grouping students as hint-driven or failure-driven and finding students' common misconceptions (Yudelson, Medvedeva, Legowski, Castine, & Jukic, 2006); identifying learners with little motivation and finding remedial actions in order to lower drop-out rates (Cocea & Weibelzahl, 2006); for predicting course success (Hamalainen & Vinni, 2006).

The Keel system has several classification algorithms available. The C4.5 algorithm (Quinlan, 1993) is used to characterize students who passed or failed the course. The C4.5 is an algorithm for generating decision trees and inducing classification rules from the tree. In this case, our objective is to classify students into different groups with equal final marks depending on the activities carried out in Moodle. We have executed the C4.5 with the default parameters over the *student_summarization_discretized* file (in which the class is the last attribute) and k-fold cross validation with $k = 3$ (the original sample is partitioned into K subsamples, and of the K subsamples, a single sub-sample is retained as the validation data for testing the model, and the remaining K-1 subsamples are used as training data.). On executing the algorithm in Keel, a decision tree is obtained (see Fig. 6) as well as a summary with a number of nodes and a number of leaves on the tree, number and percentage of correctly and incorrectly classified instances.

We obtain a set of IF-THEN-ELSE rules from the decision tree that can show interesting information about the classification of the students. Summarizing the rules obtained, they classify at least three main categories of students. Students with a low number of quizzes passed are directly classified as FAIL (first IF in Fig. 6). Students with a high number of passed quizzes are directly classified as EXCELLENT (last ELSEIF in Fig. 6). And students with a medium number of passed quizzes are classified as FAIL, PASS or GOOD (the rest of IF and ELSEIF) depending on other values (total time of assignments, number of quizzes, number of quizzes failed, number of assignments, course, etc.).

The instructor can use the knowledge discovered by these rules for making decisions about Moodle course activities and for classifying new students. For example, it is very logical that the number of quizzes passed was the main discriminator of the final marks. But there are some others that can help the instructor to decide to promote the use of some types of activities to obtain higher marks. Or on the contrary, to decide to

Please cite this article in press as: Romero, C. et al, Data mining in course management systems: Moodle case ..., *Computers & Education* (2007), doi:10.1016/j.compedu.2007.05.016

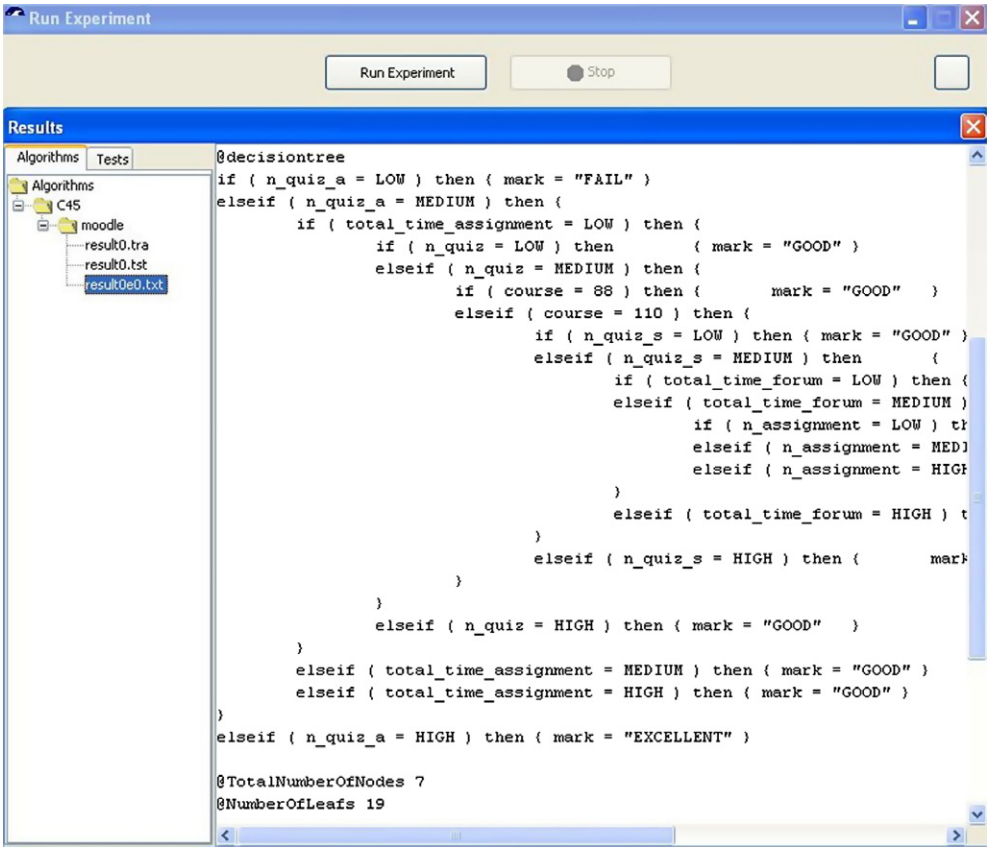


Fig. 6. Keel executing C45 algorithm.

eliminate some activities related to low marks. The instructor can also detect new students with learning problems in time (students classified as FAIL). The instructor can use the decision tree model in order to classify new students and detect in time if they will have learning problems (students classified as FAIL) or not (students classified as GOOD or EXCELLENT).

4.5. Association rule mining

Association rule mining discovers relationships among attributes in databases, producing if-then statements concerning attribute-values (Agarwal, Imielinski, & Swami, 1993). An $X \Rightarrow Y$ association rule expresses a close correlation between items (attribute-value) in a database with values of support and confidence. The confidence of the rule is the percentage of transactions that contains the consequence in transactions that contain the antecedent. The support of the rule is the percentage of transactions that contains both antecedent and consequence in all transactions in the database. Association rule mining has been applied to web-based education systems for: building recommender agents that could recommend on-line learning activities or shortcuts (Zaïane, 2002); diagnosing student learning problems and offer students advice (Hwang, Hsiao, & Tseng, 2003); guiding the learner's activities automatically and recommending learning materials (Lu, 2004); determining which learning materials are the most suitable to be recommended to the user (Markellou, Mousouroui, Spiros, & Tsakalidis, 2005); identifying attributes characterizing patterns of performance disparity between various groups of students (Minaei-Bidgoli, Tan, & Punch, 2004); discovering interesting relationships from student's usage information in order to provide feedback to course author (Romero et al., 2004); finding out relationships in learners' behaviour patterns (Yu, Own, & Lin, 2001); finding students'

Please cite this article in press as: Romero, C. et al, Data mining in course management systems: Moodle case ..., *Computers & Education* (2007), doi:10.1016/j.compedu.2007.05.016

mistakes that often accompany each other (Merceron & Yacef, 2004); guiding the search for best fitting transfer models of student learning (Freyberger, Heffernan, & Ruiz, 2004); and optimizing the content of the e-learning portal by determining what most interests the user (Ramli, 2005).

The Weka system has several association rule-discovering algorithms available. We have used the Apriori algorithm (Agarwal et al., 1993) for finding association rules over the discretized summarization table of the course 110 (Projects), executing this algorithm with a minimum support of 0.3 and a minimum confidence of 0.9 as parameters. Weka shows a list of rules (see Fig. 7) with the support of the antecedent and the consequent (total number of items), and the confidence (percentage of items in a 0 to 1 scale) of the rule.

Fig. 7 shows how a huge number of association rules can be discovered. There are also a lot of uninteresting rules, like the great number of redundant rules (rules with a generalization of relationships of several rules, like rule 3 with rules 1 and 2, and rule 6 with rule 7 and 8). There are some similar rules (rules with the same element in antecedent and consequent but interchanged, such as rules 15, 16, 17, and rules 18, 19, 20). And there are some random relationships (rules with random relations between variables, such as rules 1 and 5). But there are also rules that show relevant information for educational purposes, like those that show expected or conforming relationships (if a students does not send messages, it is logical that he/she does not read them either, such as rule 2, and, in like manner, rules 10, 11 and 13). And there are also rules that show unexpected relationships (such as rules 4, 12, 14 and 9) which can be very useful for the instructor in decision making about the activities and detecting students with learning problems, like rules 4, 12 and 9 that show that if the number of messages read and messages sent in the forum is very low, and the total time and number of passed quizzes is very low, then the final mark obtained is fail. Starting from this information, the instructor can pay more attention to these students because they are prone to failure. As a result, the instructor can motivate them in time to pass the course.

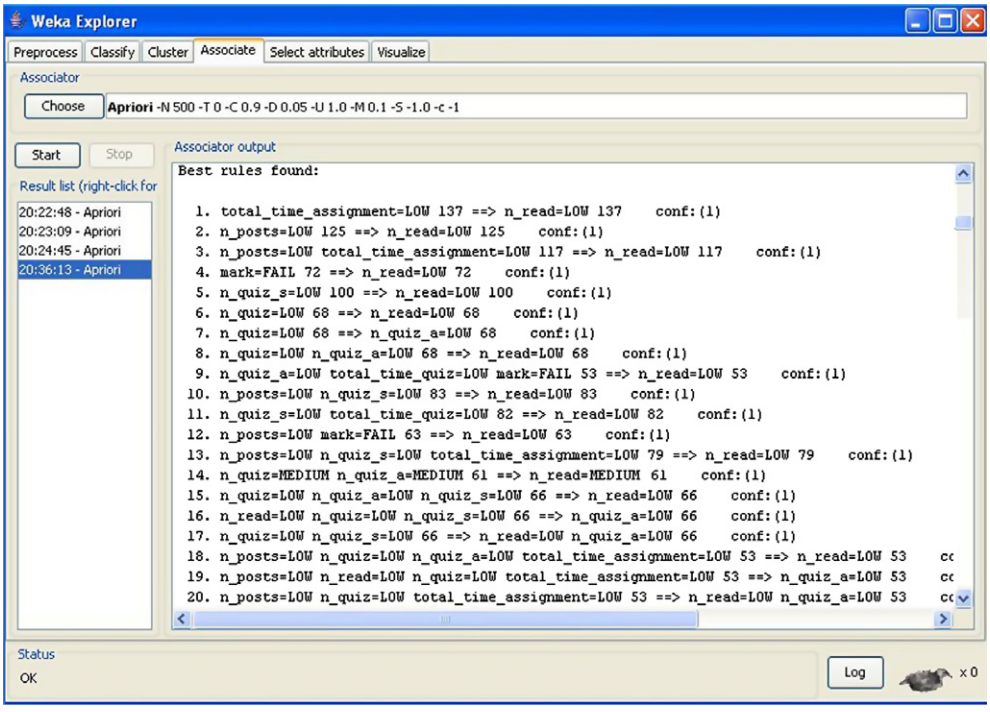


Fig. 7. Weka executing Apriori algorithm.

Please cite this article in press as: Romero, C. et al, Data mining in course management systems: Moodle case ..., *Computers & Education* (2007), doi:10.1016/j.compedu.2007.05.016

4.6. Other techniques

In the previous sections we have described the most general and well-known data mining techniques. However, there are other data mining techniques as well that are also used in e-learning, such as sequential pattern mining, text mining, outlier analysis and social network analysis.

- Sequential pattern mining (Agarwal & Srikant, 2005) is a more restrictive form of association rule mining in which the accessed items' order is taken into account. It tries to discover if the presence of a set of items is followed by another item in a time-ordered set of sessions or episodes. The extraction of sequential patterns has been used in e-learning for: evaluating learners' activities and can be used in adapting and customizing resource delivery (Zaïane & Luo, 2001); discovering and comparison with expected behavioural patterns specified by the instructor that describe an ideal learning path (Pahl & Donnellan, 2003); giving an indication of how to best organize the educational web space and be able to make suggestions to learners who share similar characteristics (Ha, Bae, & Park, 2000); generating personalized activities to different groups of learners (Wang, Weng, Su, & Tseng, 2004); supporting the evaluation and validation of learning site designs (Machado & Becker, 2003); identifying interaction sequences indicative of problems and patterns that are markers of success (Kay, Maisonneuve, Yacef, & Zaiane, 2006).
- Text mining (Feldman & Sanger, 2006) can be viewed as an extension of data mining to text data and it is closely related to web content mining. Its methods include text mining that can work with unstructured or semi-structured data sets such as full-text documents, HTML files and emails. The specific application of text mining techniques in e-learning can be used for: grouping documents according to their topics and similarities and providing summaries (Hammouda & Kamel, 2006); finding and organizing material using semantic information (Tane et al., 2004); supporting editors when gathering and preparing the materials (Grobelnik, Mladenic, & Jermol, 2002); evaluating the progress of the thread discussion to see what the contribution to the topic is (Dringus & Ellis, 2005); collaborative learning and a discussion board with evaluation between peers (Ueno, 2004a); identifying the main blocks of multimedia presentations (Bari & Benzater, 2005); selecting articles and automatically constructing e-textbooks (Chen, Li, Wang, & Jia, 2004) and personalized courseware (Tang, Lau, Yin, Li, & Kilis, 2000); detecting the conversation focus of threaded discussions, classifying topics and estimating the technical depth of contribution (Kim, Chern, Feng, Shaw, & Hovy, 2006).
- Outlier analysis (Hodge & Austin, 2004) is a type of data analysis that seeks to determine and report on records in the database that differ significantly from expectations. An outlier is an observation (or measurement) that is unusually large or small when compared to the other values in a data set. This technique is normally used for data cleansing, spotting emerging trends and recognizing unusually good or bad performers. In e-learning, outlier detection has been used for: assisting instruction in the detection of learners' irregular learning processes (Ueno, 2004b); detecting atypical behaviour in the grouping structure of the users of a virtual campus (Castro, Vellido, Nebot, & Minguillon, 2005); detecting regularities and deviations in the learner's or educator's actions with others (Muehlenbrock, 2005).
- Social network analysis (SNA) is based on the idea that a social environment can be expressed by the patterns of relations among its interacting units (Scott, 2000). The SNA uses the connections among units as data which relate them in a system. In e-learning, SNA can be used for mining group activities by analyzing the sociograms associated with a given group and the status of participants and the group cohesion of social interactions (Reyes & Tchounikine, 2005) and by interpreting and analyzing the structure and content of online educational communities (Rallo et al., 2005).

5. Conclusions

In this work we have shown how useful the application of data mining techniques in course management systems can be for online instructors. Although we have shown these techniques separately, they can also be applied together in order to obtain interesting information in a more efficient and faster way. First, instructors can use visualization techniques to obtain a general view of the student's usage data. And for example, if they find something strange or irregular in the plots, then they can obtain more detailed information about these

Please cite this article in press as: Romero, C. et al, Data mining in course management systems: Moodle case ..., *Computers & Education* (2007), doi:10.1016/j.compedu.2007.05.016

events by viewing statistical values. Or, if they find some similar groups of students in graphs, then they can apply clustering techniques in order to obtain the exact groups students can be divided into. And these groups can also be used to create a classifier in order to classify students. The classifier shows what the main characteristics of the students in each group are, and it allows new online students to be classified. Finally, the instructors can apply association rule mining to discover if there is any relationship between these characteristics and other attributes. These rules can not only help to classify students, but also to detect the sources of any incongruous values obtained by the students.

At present, we are developing a specific Moodle data mining tool oriented for use by on-line instructors which would obviate the need for CMS administrators to help these instructors to preprocess or apply mining techniques. It has an intuitive and user-friendly interface to do data mining and automatically preprocesses Moodle data, making it easier to configure and execute data mining techniques due to its parameter-free data mining algorithms. This tool is integrated into the Moodle environment itself as another Moodle author tool such as GISMO. In this way, instructors can both create/maintain courses and carry out all data mining processing in the same interface. Likewise, they can directly apply feedback and results obtained by data mining into Moodle courses.

Acknowledgement

The authors gratefully acknowledge the subsidy provided by the Spanish Department of Research under TIN2005-08386-C05-03 Projects.

References

Agarwal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD international conference on management of data, Washington DC, USA* (pp. 1–22).

Agarwal, R., & Srikant, R. (2005). Mining sequential patterns. In *Proceedings of the eleventh international conference on data engineering, Taipei, Taiwan* (pp. 3–14).

Alcalá, J., Sánchez, L., García, S., del Jesús, M. J., Ventura, S., Garrell, J. M., et al. (in press). KEEL: A data mining software tool for assessing the performance of knowledge extraction-based on evolutionary algorithms. *Soft computing: A fusion of foundations, methodologies and applications*.

Anozie, N., & Junker, B. W. (2006). Predicting end-of-year accountability assessment scores from monthly student records in an online tutoring system. In *Educational data mining AAAI workshop, California, USA* (pp. 1–6).

Arroyo, I., Murray, T., Woolf, B., & Beal, C. (2004). Inferring unobservable learning ariables from students' help seeking behavior. In *Intelligent tutoring systems, Alagoas, Brazil* (pp. 782–784).

Avouris, N., Komis, V., Fiotakis, G., Margaritis, M., & Voyiatzaki, E. (2005). Why logging of fingertip actions is not enough for analysis of learning activities. In *Workshop on usage analysis in learning systems at the 12th international conference on artificial intelligence in education, Amsterdam, Netherland* (pp. 1–8).

Baker, R., Corbett, A., & Koedinger, K. (2004). Detecting student misuse of intelligent tutoring systems. In *Intelligent tutoring systems, Alagoas, Brazil* (pp. 531–540).

Bari, M., & Benzater, B. (2005). Retrieving data from pdf interactive multimedia productions. In *International conference on human system learning: Who is in control?, Marrakech, Morroco* (pp. 321–330).

Beck, J. E., & Woolf, B. P. (2000). High-level student modeling with machine learning. In *Proceedings of Fifth International Conference on Intelligent Tutoring Systems, Alagoas, Brazil* (pp. 584–593).

Becker, K., Vanzin, M., & Ruiz, D.A. (2005). Ontology-based filtering mechanisms for web usage patterns retrieval. In *International conference on e-commerce and web technologies, München, Germany* (pp. 267–277).

Bellaachia, A., Vommina, E., & Berrada, B. (2006). Minel: A framework for mining e-learning logs. In *Proceedings of the fifth IASTED international conference on Web-based education, Mexico* (pp. 259–263).

BlackBoard (2007). <http://www.blackboard.com/>.

Bravo, J., & Ortigosa, A. (2006). Validating the evaluation of adaptive systems by user profile simulation. In *Proceedings of the workshop on user-centred design and evaluation of adaptive systems, Dublin, Germany* (pp. 52–56).

Castro, F., Vellido, A., Nebot, A., & Minguillon, J. (2005). Detecting a typical student behaviour on an e-learning system. In *Simposio Nacional de Tecnologías de la Informacin y las Comunicaciones en la Educacion* (pp. 153–160). Spain: Granada.

Castro, F., Vellido, A., Nebot, A., & Mugica, F. (in press). Applying data mining techniques to e-learning problems: A survey and state of the art. In L. C. Jain, R. Tedman, & D. Tedman (Eds.), *Evolution of Teaching and learning paradigms in intelligent environment*. Studies in Computational Intelligence (Vol. 62). Springer-Verlag.

Chang, F. C. I., Hung, L. P., & Shih, T. K. (2003). A new courseware for quantitative measurement of distance learning courses. *Journal of Information Science and Engineering, 19*, 989–1014.

Please cite this article in press as: Romero, C. et al, Data mining in course management systems: Moodle case ..., *Computers & Education* (2007), doi:10.1016/j.compedu.2007.05.016

Chen, J., Li, Q., Wang, L., & Jia, W. (2004). Automatically generating an textbook on the web. In *International conference on advances in web-based learning, Beijing, China* (pp. 35–42).

Chen, G., Liu, C., Ou, K., & Liu, B. (2000). Discovering decision knowledge from web log portfolio for managing classroom processes by applying decision tree and data cube technology. *Journal of Educational Computing Research, 23*(3), 305–332.

Claroline (2007). <http://www.claroline.net/>.

Clementine (2007). <http://www.spss.com/clementine/>.

Cocea, M., & Weibelzahl, S. (2006). Can log files analysis estimate learners' level of motivation? In *Proceedings of the workshop week Lernen - Wissensentdeckung - Adaptivität, Hildesheim* (pp. 32–35).

Cole, J. (2005). Using Moodle. O'Reilly.

DBMiner (2007). <http://www.dbminer.com>.

Dougherty, J., Kohavi, M., & Sahami, M. (1995). Supervised and unsupervised discretization of continuous features. In *International conference machine learning Tahoe City, CA* (pp. 194–202).

Dringus, L., & Ellis, T. (2005). Using data mining as a strategy for assessing asynchronous discussion forums. *Computer & Education Journal, 45*, 141–160.

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). Pattern classification. Wiley Interscience.

Feldman, R., & Sanger, J. (2006). The text mining handbook. Cambridge University Press.

Feng, M., Heffernan, N., & Koedinger, K. (2005). Looking for sources of error inpredicting student's knowledge. In *Proceedings of AAAI workshop on educational data mining, California, USA* (pp. 1–8).

Feng, M., & Heffernan, N. (2006). Informing teachers live about student learning: Reporting in the assistment system. *Technology, Instruction, Cognition, and Learning Journal, 3*, 1–8, Old City Publishing.

Freyberger, J., Heffernan, & N., Ruiz, C. (2004). Using association rules to guide a search for best fitting transfer models of student learning. In *Workshop on analyzing student–tutor interactions logs to improve educational outcomes at ITS conference, Alagoas, Brazil* (pp. 1–10).

García, E., Romero, C., Ventura, S., & Castro, C. (2006). Using rules discovery for the continuous improvement of e-learning courses. In *International conference intelligent data engineering and automated learning, Burgos, Spain* (pp. 887–895).

Gaudioso, E., & Talavera, L. (2006). Data mining to support tutoring in virtual learning communities: Experiences and challenges. In C. Romero & S. Ventura (Eds.), *Data mining in e-learning* (pp. 207–226). Southampton, UK: Wit Press.

Gismo (2007). <http://gismo.sourceforge.net/>.

Grob, H. L., Bensberg, F., & Kaderali, F. (2004). Controlling open source intermediaries – a web log mining approach. In *Proceedings of the international conference on information technology interfaces, Zagreb* (pp. 233–242).

Grobelnik, M., Mladenic, D., & Jermol, M. (2002). Exploiting text mining in publishing and education. In *Proceedings of the ICML workshop on data mining lessons learned, Sydney, Australia* (pp. 34–39).

Ha, S., Bae, S., & Park, S. (2000). Web mining for distance education. In *IEEE international conference on management of innovation and technology, Singapore* (pp. 715–719).

Hamalainen, W., Suhonen, J., Sutinen, E., & Toivonen, H. (2004). Data mining in personalizing distance education courses. In *World conference on open learning and distance education, Hong Kong* (pp. 1–11).

Hamalainen, W., & Vinni, M. (2006) Comparison of machine learning methods for intelligent tutoring systems. In *Proceedings of the eighth international conference in intelligent tutoring systems, Taiwan* (pp. 525–534).

Hammouda, K., & Kamel, M. (2006). Data mining in e-learning. In Samuel Pierre (Ed.), *E-learning networked environments and architectures: A knowledge processing perspective, Springer Book Series: Advanced information and knowledge processing* (pp. 1 – 28).

Herin, D., Sala, M., & Pompidor, P. (2002). Evaluating and revising courses from web resources educational. In *International conference on intelligent tutoring systems, Spain* (pp. 208–218).

Hodge, V., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review, 22*(2), 85–126.

Hwang, G. J., Hsiao, C. L., & Tseng, C. R. (2003). A computer-assisted approach to diagnosing student learning problems in science courses. *Journal of Information Science and Engineering, 19*, 229–248.

Ilias. (2007). <http://www.ilias.de/>.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys, 31*(3), 264–323.

Kay, J., Maisonneuve, N., Yacef, K., & Zaiane, O. R. (2006). Mining patterns of events in students' teamwork data. In *Proceedings of educational data mining workshop, Taiwan* (pp. 1–8).

Keel. (2007). <http://www.keel.es/>.

Kim, J., Chern, G., Feng, D., Shaw, E., & Hovy, E. (2006). Mining and assessing discussions on the web through speech act analysis. In *Proceedings of the AAAI workshop on web content mining with human language technologies, Athens, GA* (pp.1–8).

Klosgen, W., & Zytkow, J. (2002). Handbook of data mining and knowledge discovery. New York: Oxford University Press.

Kotsiantis, S. B., Pierrakeas, C. J., & Pintelas, P. E. (2004). Predicting students' performance in distance learning using machine learning techniques. *Applied Artificial Intelligence, 18*(5), 411–426.

Koutri, M., Avouris, N., & Daskalaki, S. (2005). A survey on web usage mining techniques for web-based adaptive hypermedia systems. *Adaptable and adaptive hypermedia systems*. IRM Press (pp. 125–149).

Luan, J. (2002). Data mining, knowledge management in higher education, potential applications. In *Workshop associate of institutional research international conference, Toronto* (pp. 1–18).

Lu, J. (2004). Personalized e-learning material recommender system. In *International conference on information technology for application, Utah, USA* (pp. 374–379).

Please cite this article in press as: Romero, C. et al, Data mining in course management systems: Moodle case ..., *Computers & Education* (2007), doi:10.1016/j.compedu.2007.05.016

Machado, L., Becker, K. (2003). Distance education: A web usage mining case study for the evaluation of learning sites. In *International conference on advanced learning technologies, Athens, Greece* (pp. 360–361).

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth berkeley symposium on mathematical statistics and probability, California, USA*. (Vol. 1, pp. 281–297).

Markellou, P., Mousourouli, I., Spiros, S., & Tsakalidis, A. (2005). Using semantic web mining technologies for personalized e-learning experiences. In *Proceedings of the web-based education, Grindelwald, Switzerland* (pp. 461–826).

Mazza, R., & Dimitrova, V. (2004). Visualising student tracking data to support instructors in web-based distance education. In *International world wide web conference, New York, USA* (pp. 154–161).

Mazza, R., & Milani, C. (2005). Exploring usage analysis in learning systems: Gaining insights from visualisations. In *Workshop on usage analysis in learning systems at 12th international conference on artificial intelligence in education, New York, USA* (pp. 1–6).

Merceron, A., & Yacef, K. (2004). Mining student data captured from a web-based tutoring tool: Initial exploration and results. *Journal of Interactive Learning Research, 15*(4), 319–346.

Merceron, A., & Yacef, K. (2005). TADA-Ed for educational data mining. *Interactive Multimedia Electronic Journal of Computer-enhanced Learning, 7*(1).

Minaei-Bidgoli, B., & Punch, W. (2003). Using genetic algorithms for data mining optimization in an educational web-based system. In *Genetic and evolutionary computation conference, Chicago, USA* (pp. 2252–2263).

Minaei-Bidgoli, B., Tan, P., & Punch, W. (2004). Mining interesting contrast rules for a web-based educational system. In *International conference on machine learning applications, Los Angeles, California* (pp. 1–8).

Miner (2007). <http://www-306.ibm.com/software/data/iminer/>.

Monk, D. (2005). Using data mining for e-learning decision making. *Electronic Journal of e-Learning, 3*(1), 41–54.

Mor, E., & Minguillon, J. (2004). E-learning personalization based on itineraries and long-term navigational behavior. In *Proceedings of the 13th international world wide web conference* (pp. 264–265).

Moodle (2007). <http://moodle.org/>.

Mostow, J., Beck, J., Cen, H., Cuneo, A., Gouvea, E., & Heiner, C. (2005). An educational data mining tool to browse tutor–student interactions: Time will tell! In *Proceedings of the workshop on educational data mining, Pittsburgh, USA* (pp. 15–22).

Mostow, J., & Beck, J. (2006). Some useful tactics to modify, map and mine data from intelligent tutors. *Natural Language Engineering, 12*(2), 195–208.

Muehlenbrock, M. (2005). Automatic action analysis in an interactive learning environment. In *Proceedings of the workshop on usage analysis in learning systems at the 12th international conference on artificial intelligence in education, Amsterdam, The Netherlands* (pp. 73–80).

Nilakant, K., & Mitrovic, A. (2005). Application of data mining in constraintbased intelligent tutoring systems. In *Proceedings artificial intelligence in education, Amsterdam, The Netherlands* (pp. 896–898).

Ortigosa, A., Carro, & R. M. (2003). The continuous empirical evaluation approach: Evaluating adaptive web-based courses. In *International conference user modeling, Canada* (pp. 163–167).

Pahl, C., & Donnellan, C. (2003). Data mining technology for the evaluation of web-based teaching and learning systems. In *Proceedings of the Congress E-learning. Montreal, Canada* (pp. 1–7).

Pritchard, D. E., & Warnakulasooriya, R. (2005). Data from a web-based homework tutor can predict student's final exam score. In *World conference on educational multimedia, hypermedia and telecommunications, Canada* (pp. 2523–2529).

Quinlan, R. (1993). *C4.5: Programs for machine learning*. Morgan Kaufman Publishers.

Ramli, A. A. (2005). Web usage mining using apriori algorithm: UUM learning care portal case. In *International conference on knowledge management, Malaysia* (pp. 1–19).

Rallo, R. Gisbert, M., & Salinas, J. (2005). Using data mining and social networks to analyze the structure and content of educative on-line communities. In *International conference on multimedia and ICTs in education, Caceres, Spain* (pp. 1–10).

Reyes, P., & Tchounikine, P. (2005). Mining learning groups' activities in forum-type tools. In *Proceedings of the 2005 conference on computer support for collaborative learning, Taiwan* (pp 509–513).

Rice, W. H. (2006). *Moodle e-learning course development. A complete guide to successful learning using Moodle*. Packt Publishing.

Romero, C., Ventura, S., & Bra, P. D. (2004). Knowledge discovery with genetic programming for providing feedback to courseware author. *User Modeling and User-Adapted Interaction: The Journal of Personalization Research, 14*(5), 425–464.

Romero, C., & Ventura, S. (2006). *Data mining in e-learning*. Southampton, UK: Wit Press.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications, 33*(1), 135–146.

Romero, C. Ventura, S. Delgado, J. A., & Bra, P. D. (in press). Personalized links recommendation based on data mining in adaptive educational hypermedia systems. In *Second European Conference on technology enhanced learning. Crete, Greece*.

Scott, J. (2000). *Social network analysis: A handbook* (2nd ed.). Newberry Park, CA: Sage.

Shen, R., Yang, F., & Han, P. (2002). Data analysis center based on e-learning platform. In *Workshop the internet challenge: Technology and applications, Berlin, Germany* (pp. 19–28).

Silva, D., & Vieira, M. (2002). Using data warehouse and data mining resources for ongoing assessment in distance learning. In *IEEE international conference on advanced learning technologies, Kazan, Russia* (pp. 40–45).

Spacco, J., Winters, T., & Payne, T. (2006). Inferring use cases from unit testing. In *AAAI workshop on educational data mining, New York* (pp. 1–7).

Spence, R. (2001). *Information visualization*. Addison-Wesley.

Please cite this article in press as: Romero, C. et al, Data mining in course management systems: Moodle case ..., *Computers & Education* (2007), doi:10.1016/j.compedu.2007.05.016

Talavera, L., & Gaudioso, E. (2004). Mining student data to characterize similar behavior groups in unstructured collaboration spaces. In *Workshop on artificial intelligence in CSCL, Valencia, Spain* (pp. 17–23).

Tane, J., Schmitz, C., & Stumme, G. (2004). Semantic resource management for the web: An elearning application. In *Proceedings of the WWW conference, New York, USA* (pp. 1–10).

Tang, C., Lau, R. W. H., Li., Q., Yin, H., Li, T., & Kilis, D. (2000). Personalized courseware construction based on web data mining. In *Proceedings of the first international conference on web information systems engineering, Hong Kong, China* (pp. 204–211).

Tang, T., & McCalla, G. (2005). Smart recommendation for an evolving e-learning system. *International Journal on E-Learning, 4*(1), 105–129.

TopClass. (2007). <http://www.topclass.nl/>.

Tsantis, L., & Castellani, J. (2001). Enhancing learning environments through solution-based knowledge discovery tools. *Journal of Special Education Technology, 16*(4), 1–35.

Ueno, M. (2004a). Data mining and text mining technologies for collaborative learning in an ilms ''samurai''. In *IEEE international conference on advanced learning technologies, Joensuu, Finland* (pp. 1052–1053).

Ueno, M. (2004b). Online outlier detection system for learning time data in e-learning and its evaluation. In *International conference on computers and advanced technology in education, Beijing, China* (pp. 248–253).

Wang, F. H., & Shao, H. M. (2004). Effective personalized recommendation based on time-framed navigation clustering and association mining. *Expert Systems with Applications, 27*(3), 365–377.

Wang, W., Weng, J., Su, J., & Tseng, S. (2004). Learning portfolio analysis and mining in scorm compliant environment. In *ASEE/IEEE frontiers in education conference, Savannah, Georgia* (pp. 17–24).

WebCT (2007). <http://www.webct.com/>.

Weka (2007). <http://www.cs.waikato.ac.nz/ml/weka/>.

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques*. Morgan Kaufman.

Yu, P., Own, C., & Lin, L. (2001). On learning behavior analysis of web based interactive environment. In *Proceedings of the implementing curricular change in engineering education, Oslo, Norway* (pp. 1–10).

Yudelson, M. V., Medvedeva, O., Legowski, E., Castine, M., Jukic, D., & Rebecca C. (2006). Mining student learning data to develop high level pedagogic strategy in a medical ITS. In *Proceedings of AAAI workshop on educational data mining, Boston* (pp. 1–8).

Zorrilla, M. E., Menasalvas, E., Marin, D., Mora, E., & Segovia, J. (2005). Web usage mining project for improving web-based learning sites. In *Web mining workshop, Cataluna* (pp. 1–22).

Zaïane, O., Xin, M., & Han, J. (1998). Discovering web access patterns and trends by applying OLAP and data mining technology on web logs. In *Advances in digital libraries* (pp. 19–29).

Zaïane, O., & Luo, J. (2001). Web usage mining for a better web-based learning environment. In *Proceedings of conference on advanced technology for education, Banff, Alberta* (pp. 60–64).

Zaïane, O. (2002). Building a recommender agent for e-learning systems. In *Proceedings of the international conference in education, Auckland, New Zealand* (pp. 55–59).

Please cite this article in press as: Romero, C. et al, Data mining in course management systems: Moodle case ..., *Computers & Education* (2007), doi:10.1016/j.compedu.2007.05.016