THE USE OF DATA MINING METHODS FOR ANALYZING AND EVALUATING COURSE QUALITY IN THE MOODLE SYSTEM

Gabrijela Dimić
Electrical Engineering and Computing Science College, Vojvode Stepe 283, Belgrade, Serbia

Dragana Prokin
Electrical Engineering and Computing Science College, Vojvode Stepe 283, Belgrade, Serbia

Kristijan Kuk
Electrical Engineering and Computing Science College, Vojvode Stepe 283, Belgrade, Serbia

Petar Spalević
Faculty of Technical Sciences Kneza Miloša 7, 38000 K.Mitovica, Serbia

Abstract

The work describes investigation of data mining methods and techniques for evaluating quality and successfulness of a course realized in the e-learning system Moodle. We showed what kind of data can be collected, how to perform the previous data preparation, how to apply appropriate methods on the data, and how to use the discovered knowledge. The focus is on the use of data mining methods and techniques as a tool for discovering those parts of activities than cannot be obtained from polls filled in by users. For the research presented in this work we chose and analyzed data on the course Programmable logic devices for the school year 2008/2009. The course was realized under the open source platform Moodle, which represents one of e-learning systems and environments for distant learning. We analyzed data on students' activities, generated from the Moodle system database. The basis for the described research is theoretic knowledge in the field of data mining, educational data mining and practical knowledge in the field of usage of educational models in electronic education.

Keywords: e-learning, Moodle, data mining, educational data mining, data mining methods,clustering, clasification.

INTRODUCTION

The Learning Management System (LMS) is a software application used for creation, organization, realization, administration, and preparation of reports and documentation of one or several e-learning courses, training programs for one or several users in a classroom or virtual environment. They enable user self-authentification, access to learning contents, easy upload of documents, leading discussions, testing and polling of users, saving marks, and so on. LMS systems accumulate large quantity of data on registered users and all subscriptions to the system [1]. They can generate daily reports with a large quantity of data, and the analyzing procedure of generated reports and extraction of useful information by the course creators is thus extremely difficult and long-lasting. Therefore, they need tools that would help them in performing this task. Although some platforms offer specific tools for preparation of reports, when the number of students is large the process of extracting useful information is still complicated. The course creator (teacher) has to select a content that will be presented, to organize the course content, and to identify appropriate content elements intended for various types of potential course users. Generally speaking, the course design, regardless of whether it refers to the structure, contents, navigation or user interface, is in most cases subject to constant improvement and modification based on students' user information, especially by following a continuous empirical development approach [2].

The traditional data analysis in e-learning systems is a hypothesis or is conducted by assumptions in the sense that users start from a question and research data in order to confirm their observation and intuition [3]. Data mining, unlike the traditional analysis, is based on discovering knowledge in the sense that an assumption is automatically derived from data, i.e. that it is driven by data rather than based on research or driven by a man [4].

COURSE IN MOODLE

A typical example of an open-source e-learning system that is becoming more and

more popular is Moodle, which enables course creators to efficiently organize online learning environment [5]. The modular design enables creation of courses that can support the learning model called social constructionist pedagogy [6]. This way of learning implies that the learning process is most efficient if a student interacts with the learning material and other course participants (associates, students). Moodle keeps detailed records on all activities and accesses of participants, regardless of whether they are students or teachers. Teachers can easily obtain a complete report on all activities for each individual participant, as well as a complete report on all participants for certain activities.

For the purpose of our study we have collected and analyzed data on students in the course Programmable logic devices, held in the summer semester of the school year 2008/2009 at the Electrical Engineering and Computing Science College in Belgrade. The course was chosen by 130 students from different study programs. The blended-learning model was applied, i.e. a mixture of electronic material flexibility and direct teacher's participation. Students were obliged to attend the exercises, and the material for preparation of tasks for laboratory exercises was available to students in the paper form and in the form of multimedia material [21] included in the online course.

The course user environment was organized in the following way:

- Forum: Notifications by which students were delivered information related to lectures by the course teacher.
- Learning material: PPT presentations, Lessons in the form of HTML pages, Self-check tests, Interactive tasks [22,23].
- Knowledge check tests for each laboratory exercise.
- Final knowledge check tests.

EDUCATIONAL DATA MINING

Data mining is a process of extracting samples or models from observed data. It is defined as a process of identifying valid, new, potentially useful and, finally, understandable patterns in data [7]. The use of data mining in education implies integration of data discovery processes and methods for knowledge discovery in educational environment. This new developing field, known as Educational Data Mining. Knowledge discovery techniques have a significant role in extraction of interesting and useful data on students' activities, on the basis of data kept in the learning system database. The discovered knowledge can be used for better understanding of students' behavior, for improving lectures, course performance, for evaluation and improvement of e-learning systems, improvement of teaching plans, and many other benefits. The data needed for research in this field are collected from operational data that are kept in databases of educational institutions and represent personal and academic data on students, and from the systems that support electronic learning and have a huge quantity of information.[8].

Data mining uses advanced techniques for discovering behavior patterns from existing data, and they can be interpreted and adjusted to checking efficiency of the e-learning system [9]. Some of the most frequently used methods and techniques of data mining are the following:

- Clustering – grouping of users or events without previous knowledge of categories.
- Classification – Division of events or users into previously defined categories.
- Association – discovering patterns in which one event is connected with another event.
- Sequential analysis – discovering patterns in which one event follows another in a sequence.
- Prediction – discovering samples in data that can lead to reasonable prediction of future events.

A data mining process comprises the following stages: collection of data to be analyzed, preparation and pre-procession of data, usage of data mining algorithms, interpretation and evaluation of results.

DATA COLLECTION

One of the main objectives of this study was grouping students on the basis of analyzing of their activities in the course. The purpose of clustering is to achieve a better course personalization, checking quality of learning the materials within the course; determining the connection between frequency

of using of certain learning materials by students and the final score achieved at the exam.

For the purpose of our reasearch we collected personal and academic data on students in the course Programmable logic devices. From Moodle we first collected information on students' access to the course. We generated information on the use of learning material, solving self-check tests, and success in final knowledge check tests. Moodle keeps a record on accesses to the system in the relation database. MySQL and PostgreSQL are the best supported databases, but Oracle, Access, Interbase and others can be used as well [10] .

DATA PREPROCESSING AND DISCRETIZATION

In order to obtain better data for further analysis and use of data mining techniques we had to carry out preparation and pre-procession of data. This stage implies previous procession and cleaning of data, removal of inapplicable and unimportant information, collection of information necessary for modelling, selection of methods for handling the missing data fields, and finally discretization. Data pre-procession allows original data to be transformed in an appropriate form for the use of data mining algorithm. We should mention that information on access to the Moodle system is not only directly recorded in the database, but also in the log file on the server where Moodle is installed. Anyway, the database is a much more reliable and powerful source of data on all activities in the system than the log file. Data pre-procession stage [11] in educational data mining, when it comes to a course in the Moodle system, includes the following activities:

- Creation of a summary table that will record in one entry all the information referring to activities of each student of the analyzed course.
- Performing of the best possible discretization of collected and cleaned data.
- Transformation of discretized data into a proper format for the use of data mining algorithms.

In [11] are proposed the following attributes of the summary table for the analyzed course: id_student, no_lessons (number of read lessons), no_int_tasks (number of solved interactive tasks), no_files (number of used PPT presentations), no_tests_for_learning (number of taken self-check tests), mark_exam (points obtained in the exam). The data obtained as the result of the SQL inquiries for the said attributes are written in the table mdl_summary. As its name says, the attribute values in the table mdl_summary are of numeric type. The use of most algorithms of the data mining method functions is more efficiently if data have discreet and category values. We therefore took the following, necessary step in data analysis, which is discretization of the attributes values in the table mdl_summary so that we could apply appropriate algorithms. Through discretization procedure numerical data are divided into appropriate classes that will be more understandable, but there is a possibility to lose a piece of information during this procedure. It is therefore very important to perform the discretization procedure in the best possible way [12].

In this research we chose to use Weka open-source tools for the use of data mining methods, because they are easy to install and easy to use [13]. Having in mind that the selected tool Weka works only with certain file formats, we exported data in the table mdl_summary from the database in supported CSV format and imported them to the Weka system.We applied two discretization methods. With the method equal-width we divided the range of the attribute numerical values into three equally long intervals, which we marked as: low, medium, high, while with the manual discretization method we determined the cut-off points on the basis of the maximum number of points and grouped them into five groups: failed, passed, good, very good, excellent. After the use of dicretization filter, the resulting file with data has all nominal attributes, and it is transformed into a text file of the ARFF format (Attribute-Relation File Format), which is an ASCII text file and describes the list of instances assigned within the collection of attributes [14] and is ready for the use of data mining methods.

THE USE OF DATA MINING METHODS

Clustering is a procedure of grouping items into classes, groups, and clusters made of similar items [15] . The main task of this method is to put items into groups so that the similarity of items within a group is maximized, while similarity of items in various groups is minimized [16]. With the use of clustering in evaluation of courses of e-learning systems the implemented algorithms can find clusters of students with similar learning characteristics, group students in order to obtain various environments based on their capabilities and other characteristics [17]. The teacher can group active students with inactive students in order to improve activeness and learning characteristics of inactive students.

In our case, the clustering method, we investigated the use of Expectation-Maximization Algorithm (EM-clustering). With the use of this algorithm we grouped students according to their activities in the course, i.e. frequency of the use of learning material. The EM algorithm [18] is a mixture of the basic algorithm that finds maximum probabilities of parameter evaluation in a possible model. Table 1 shows the results of grouping. The EM algorithm grouped students into five clusters on the basis of frequency of the use of various learning materials and the mark obtained in the exam, which students took via Moodle. The algorithm assigned to attributes the maximum probability of affiliation with each cluster. Observing maximum affiliation probabilities of the attribute mark_exam, Cluster 1 includes students with the mark good, Cluster 2 includes students with the mark failed, Cluster 3 includes students with the mark good, Cluster 4 includes students with the mark very good, and Cluster 5 includes students with the mark excellent. Cluster 1 includes 22% of students that used lessons, interactive tasks and PPT presentations for the exam preparation to a small extent, but they used self-check tests to a medium extent. The final mark of students grouped in this cluster is good. Cluster 2 comprises 13% of students that used lessons, interactive tasks, PPT presentations and self-check tests for the exam preparation to a small extent. The final mark of students grouped in this cluster is failed. Cluster 3 groups 18% of students that used lessons for the exam preparation to a medium extent, and interactive tasks, PPT presentations and self-check tests to a small extent. The final mark of students grouped in this cluster is good. Cluster 4 comprises 30% of students who used lessons and self-check tests for the exam preparation to a great extent, and interactive tasks and PPT presentations to a small extent. The final mark of students grouped in this cluster is very good. Cluster 5 includes 17% of students who used lessons, interactive tasks, PPT presentations and self-check tests for the exam preparation to a great extent. The final mark of students grouped in this cluster is excellent.

Table 1. Results of grouping with the use of EM algorithm

Attrib.	*Cl.1*	*Cl.2*	*Cl.3*	*Cl.4*	*Cl.5*
no_lessons					
low	**15.1**	**14.78**	1.54	2.45	1.04
medium	9.97	1.12	**16.43**	4.71	2.75
high	2.37	1.56	4.21	**24.95**	**19.91**
no_int_tasks					
low	**19.6**	**15.39**	**16.59**	**20.30**	2.09
medium	3.93	1.02	3.74	7.89	7.39
high	3.95	1.07	1.84	3.91	**14.21**
no_files					
low	**20.3**	**14.38**	**15.76**	**24.19**	5.35
medium	4.97	1.07	3.86	3.23	5.84
high	2.16	2.01	2.54	4.77	**12.49**
no_tests_for_learning					
low	2.48	**14.53**	**15.37**	1.48	1.14
medium	**18.3**	1.63	1.71	2.25	1.03
high	6.76	1.26	5.09	**28.37**	**21.51**
mark_exam					
failed	1.66	**14.31**	1.03	1.01	1.00
passed	2.07	1.12	5.68	1.10	1.00
good	**20.1**	1.37	**14.89**	2.17	1.37
very good	4.33	1.08	1.32	**17.34**	3.89
excellent	1.33	1.52	1.24	12.48	**18.41**

Accompanying the activities of students in Moodlu, during the semester, the teacher can notice those students who do not use any of the material to learn and to warn them in time.

Students who are interested for the offered materials, the teacher can provide a different model of learning which can be more efficient

for them. Students grouped in cluster 1 and cluster 3 achieved good score at the exam using in medium scale tests or lessons to learn. For these students the teacher can provide the lessons and tests to verify the knowledge to be in the same units that are compact module of the course content. So these students can successfully prepare and achieve even better passing the exam. Looking at the probability of attributes for cluster 4 and cluster 5 the teacher can organize learning material in the form of lessons with interactive assignments and tests for self-checking knowledge and thus increase the number of students grouped in the cluster, which is characterized with excellent score at the exam.

Classification

Classification is a data mining task that implies affiliation of an instance with certain group. Pre-classified patterns are used to classify classes used for finding and marking new patterns. On the basis of a collection of formerly defined categories, mapping of existing data is performed [16]. In e-learning, classification can be used for discovering potential study groups that have similar characteristics and reactions to certain pedagogical strategies [19], for predicting students' performance and their final marks [20].

In our research we used the J48 decision tree for presenting logical rules of the final mark in the exam. We perform the J48 algorithm with implied parameters over a discretized file with attributes transformed into nominal ones. When the algorithm is performed we obtain a big collection of rules that present interesting information on students' classification. We separated only the rules with the highest probability of prediction evaluation.

If no_tests_for_learning = low and no_lessons = low then mark_exam = failed
If no_int_tasks = low and no_files = high then mark_exam = failed
If no_tests_for_learning = middle and no_lessons = low then mark_exam = good
If no_lessons =middle and no_int_tasks = low then mark_exam = good
If no_lessons =middle and no_int_tasks =middle then mark_exam = good
If no_tests_for_learning = middle and no_lessons = low then mark_exam = good
If no_lessons = high and no_tests_for_learning = low then mark_exam = good
If no_lessons = high and no_int_tasks = low then mark_exam = very_good
If no_tests_for_learning = high and no_int_tasks = low then mark_exam = very_good
If no_int_tasks = high and no_lessons = high then mark_exam = excellent
If no_tests_for_learning = high and no_int_tasks = high then mark_exam = excellent

By analyzing the obtained rules we can carry out classification of students into the following categories:

- Students who used only PPT presentations to a great extent, and didn't use lessons, self-check tests and interactive tasks at all, or used them only to a small extent, are classified in the group failed.
- Students classified in the group good used a combination of various learning materials, but it is evident that they used lessons, self-check tests and interactive tasks at least to a medium extent.
- Students who used lessons or self-check tests to a greater extent, regardless of the weak level of use of interactive tasks as additional learning material, are classified in the group very good.
- Students who used self-check tests and interactive tasks, or lessons and interactive tasks to a great extent are classified in the group excellent.

In the process of analyzing the use of learning material we found a connection between learning material and the test for the final exam. Namely, the final exam is based on questions that, for the way of posing and organization, are similar to questions from self-check tests that were available to students after a corresponding lesson. Students could solve these tests for an unlimited number of times, and the number of questions, the question point structure and maximum time were adjusted in the same way as in the exam test. An analysis of these data showed that students tested themselves several times in order to see as many correct answers as possible. Students who used self-check tests, lessons and solved interactive tasks several times passed the exam test in 85% of cases with some of the values of the attribute mark_exam: good, very good or excellent.

CONCLUSION

In this work we presented research related to the use of some data mining methods for evaluation of a course realized in the Moodle learning system. An advantage of these methods lies in the fact that a student's low mark can be predicted on time. For instance, the teacher can predict what students have a tendency to fail the exam and can work with them on improvement of their characteristics and achievements prior to the end of semester and the final exam.

Based on the detected information from the student activities, the teacher can create a new kind of activities and learning materials from which the students receive higher grades, decide to eliminate some of the activities related to the low grade or to offer a new way of learning, which will improve passing the exam. The teacher can detect a group of students with problem solving when reviewing the test, which is related to the time for review. Also, the teacher can use decision tree models (logical rules), to classify new students more effectively and discover on time those students that will have troubles in learning (students classified as failed).

Nowadays, data mining tools are usually designed with more power and flexibility than simplicity. Therefore, these tools should have a more intuitive user interface, with parameters of data mining algorithms, simplified configuration and performance, and good visualization of items, in order to make the results of data mining algorithms useful for teachers and e-learning system designers. It is also necessary for data mining tools to be integrated in the e-learning system environment similar to other author tools. In this way all data of data mining procession could be performed in one applications, and feedback information and obtained results could be directly applied to the e-learning environment.

REFERENCE

[1] Mostow J. and Bck J.: Some useful tactics to modify, map and mine data from intelligent tutors. Natural Language Engineering 12(2), 195-208. (2006).

[2] Ortigosa, A., Carro, R.M.: The Continuous Empirical Evaluation Approach: Evaluating Adaptive Web-based Courses. In Int. Conf. User Modeling, Canada (pp. 163-167). (2003).

[3] Gaudioso E., Talavera L.: Data mining to support tutoring in virtual learning communities: experiences and challenges. (2006).

[4] Tsantis, L. and Castellani, J.: Enhancing learning environments through solution-based knowledge discovery tools: Forecasting for self-perpetuating systemic reform. Journal of Special Education Technology, 16(4), 39–52. (2001).

[5] Moodle, a free open source course management system for on line learning. <http://moodle.org/> (2006).

[6] Rice, W.H.: Moodle E-learning Course Development. A complete guide to succeful learning using Moodle. (2006).

[7] Fayyad, U.M, Piatetsky-Shapiro, G.and Smyth, P.: From data mining to knowledge discovery in databases. AI Magazine, 17(3), pp. 37-54. (1996).

[8] Hom, J.:The Usability Methods ToolBox. (1998)

[9] Zaiane O., Han J.: Mining for E-Learning. (2001)

[10] Cole, J.: Using Moodle. (2005).

[11] Romero, C., Ventura, S. and Garcia, E.: Data mining in course management systems: Moodle case study and tutorial. Computers & Education, Vol. 51, No. 1. pp. 368-384. (2008).

[12] Dougherty, J. Kohavi and M Sahami: Supervised and unsupervised discretization of continuous features. In Int. Conf. Machine Learning Tahoe City, CA (pp.194–202). (1995).

[13] Weka. http://www.cs.waikato.ac.nz/ml/weka/

[14] Witten, I.H. and Frank, E.: Data mining: practical machine learning tools and techniques.(2005).

[15] Jain, A.K., Murty M.N.and Flynn P.J.: Data Clustering: A Review, ACM Computing Surveys, 31(3), 264-323. (1999).

[16] Han,J. and Kamber, M.: Data Mining: Concepts and Techniques, 2nd edition. The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor. (2006).

[17] Hamalainen, W., Suhonen, J., Sutinen, E. and Toivonen, H.: Data mining in personalizing distance education courses. World Conference on Open Learning and Distance Education, Hong Kong (pp. 1-11). (2004).

[18] Frank D.: The Expectation Maximization Algorithm. (2002).

[19] Chen, G., Liu, C., Ou, K. and Liu, B.: 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. (2000).

[20] Minaei-Bidgoli B., Kashy, D. Kortemeyer G. and Punch W.: Predicting Student Performance: An Application of Data Mining Methods with an Educational Web-Based System. In the Processing of 33rd ASEE/IEEE conference of Frontiers in Education. (2003)
[21] D.Prokin, K.Kuk and G.Dimić: Multimedijalne laboratorijske vežbe iz programabilnih logičkih kola, ETRAN 2009, V.Banja.
[22] K. Kuk, D. Prokin, G. Dimic and P. Spalevic: Interactive Tasks in Support to Modern Pedagogical Approach in Implementation of Teaching in the Subject Programmable Logic Devices, UNITECH 2009, Gabrovo, Bulagria.
[23] K. Kuk, D. Prokin, G. Dimic and B. Stanojevic: Interactive Tasks as a Supplement to Educational Material in the Field of Programmable Logic Devices, ISSN 1392–1215, Journal "Electronics and Electrical Engineering", 2010. No. 2(98)