ANALYSIS OF E-EXAM DATA USING DATA MINING TECHNIQUES

Jelena Mamcenko, Irma Sileikiene, Jurgita Lieponiene, Regina Kulvietiene

Vilnius Gediminas Technical University, Department of Information Technologies, Sauletekio str. 11, Vilnius, Lithuania, jelena@ gama.vtu.lt, irma@ gama.vtu.lt, jurgita-lieponiene@panko.lt, regina.kulvietiene@gama.vtu.lt

Abstract. The assessment of knowledge and competences is an integral part of the learning process. It summarizes the level of student success in achieving the corresponding outcomes and objectives. This paper focuses on analyzing electronic exam data using clustering and association rules in data mining. The aim of this work is to obtain and interpret the results and propose solutions to improve the e-examination system using a descriptive model.

Keywords: data mining, e-examination, clustering, association rules.

1 Introduction

Currently, Lithuania's higher education system is reformed, and the basic idea of the reform is to create a free market of higher education, which would create a natural competition between higher education providers. Nowadays higher education providers might be state universities, independent universities, state colleges and independent colleges. The essence of the reform is the student's basket (committed money for a student's studies paid by the state to a higher education institution, where the student is enrolled). The promotion of this basket encouraged higher education institutions to compete for students, who would study in a chosen higher academy and thereby would ensure the sponsorship of this high school. The resulting competition led the higher education community to seriously look into these aspects of the study process, which most affects students' satisfaction with the learning process.

During discussions with students about the quality of the courses, students, defining their expectations and desires, often point out that they want and expect: interesting and professionally useful lessons, an objective, fair and transparent assessment, modern laboratories for acquiring practical skills, study materials and additional literature available for all.

The purpose of this article is to choose right data mining techniques for analysis of students' e-examining data in order to explore students' behavioral characteristics whereas having the exam electronic way and in accordance with the results to offer recommendations for a higher quality of exam arrangement and organization of examination.

Research methods: the analysis of scientific literature, data analysis using data mining techniques.

2 Literature review

Data mining means searching for certain patterns within large sets of data, which creates a lot of possibilities for decision makers. By analyzing those patterns, better decisions can be made in order to improve learning and assessment process. The research interest in using data mining in e-learning is constantly increasing. According to L. Shen, M. Wang, R. Shen the database of learning management system includes much useful information which can be effectively used for the improvement of e-learning process [11]. E.Garcia, C. Romero, S. Ventura, T. Calders emphasise that learning management systems accumulate a vast amount of information which is very valuable for analyzing the students' behavior and could create a gold mine of educational data [4]. Authors emphasise that due to the vast quantities of data these systems can generate daily, it is very difficult to analyze this data manually and a very promising approach towards this analysis objective is the use of data mining techniques [4].

Using data mining methods many kinds of knowledge can be discovered [5]. The discovered knowledge can be used to better understand students' behaviour, to access student's learning style [7], to adapt a course content according to student's knowledge and abilities [6], to assist instructors, to improve learning and teaching process [5], [1]. Literature describes a number of scientific research works which use data mining methods on e-learning data. A. Merceron, K. Yacef present a case study that uses data mining methods to identify behaviour of failing students to warn students at risk before final exam [8]. V. Namdeo, A. Singh, D. Singh, R. C. Jain compare different classification algorithms and check which algorithm is optimal for classifying students' based on their final grade [10]. Other authors use neural networks for predicting student's marks [2].

Different data mining methods are used for e-learning data analyze, the most common ones are: association, classification, clustering and outlier detections [5]. Literature describes various algorithms of these methods. The choice of data mining method and its realizable algorithm depends on available data, set research goals and intended results [10]. According to A. Merceron, K. Yacef association rules are very useful in

educational data mining since they extract associations between educational items and present the results in an intuitive form to the teachers [9]. Authors emphasise that association rules require less extensive expertise in data mining than other methods [9].

3 E-examination system of Vilnius Gediminas Technical University

The assessment of knowledge and competences is an integral part of the learning process. In order to ensure the anonymity and objectivity of assessment and decrease the possibility of cheating, to simplify the analysis of exams results Vilnius Gediminas Technical University uses the electronic testing of students. For the creating of electronic tests IBM Authoring Tool is used. IBM Authoring Tool creates electronic tests as SCORM-based course packages.

The process by which electronic tests are delivered to learning management system IBM Workplace Collaborative Learning includes some stages. At the first stage the course administrator imports SCORM-based course package to the FTP server; the Learning server imports the course package from the FTP server. At the second stage the Learning server creates a course master and enters information about the master into the database; course administrator registers the master and creates offerings in the course catalog. Finally, course structure is sent to the Delivery server which registers course structure in the database, sends request to the Learning server, and the Learning server sends the course content to the Content server [12]. The process by which electronic tests are delivered to learning management system IBM Workplace Collaborative Learning is showed in the Fig. 1.

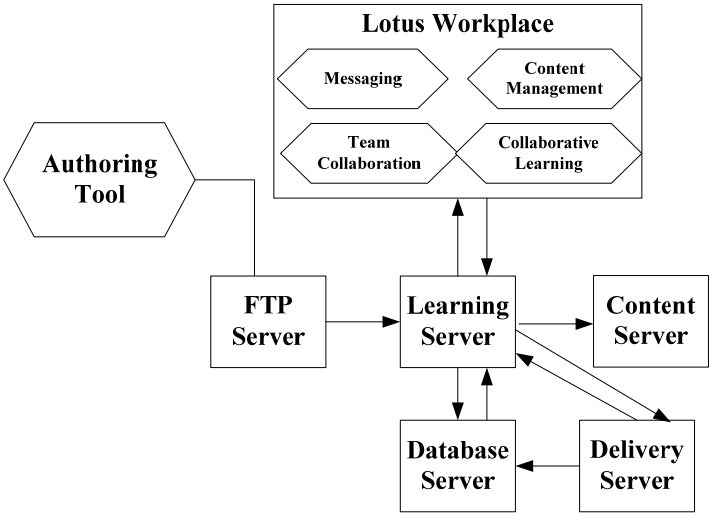

Figure 1. The process by which electronic tests are delivered to learning management system IBM Workplace Collaborative Learning

The learning management system IBM Workplace Collaborative Learning uses a relational database to store information about settings needed to run the system, courses, their structure, users, their accomplished actions, answers to the tests, time spent for separate course activities, tests results and other interesting tracking data. The database consists of 124 tables.

4 Research data and methods

In this research, the data of C++ Programming language exam are analyzed. For the electronic exam the test of 25 questions generated from 100 questions set was prepared with IBM Authoring Tool. The group of students passing the exam consisted of 100 students of full-time and evening courses from seven groups. For every group of students the exam took place at different times, so in the learning management system, a separate electronic exam offering was registered for every group and exam retaking. The students who did not pass the exam could retake it two times.

All data necessary for this research were transferred to DB2 database in the local server for further processing. IBM DWE Design Studio tool was used to perform necessary collected data pre-processing. The data set was gathered using data selection, filtering and consolidation operations. There were created the procedures to perform calculations. The data set gathering process is showed in the Fig. 2.

Selected data set consists of 19 fields and 3167 records. The sources of selected data are analyzable exam records: questions given to a student, chosen answers, the times spent to answer the questions and so on.

For data analysis are chosen descriptive data mining model's techniques: clustering and association rules. A descriptive model which is used in this work identifies patterns or relationship in data [3]. Clustering is a division of data into groups of similar objects. Each group, called cluster, consists of objects that are similar

between themselves and dissimilar to objects of other groups [5]. For the clustering in this work is used Kohonen algorithm. Using Kohonen algorithm the data are represented as five-dimensional vectors because the set for clustering is defined by five indicators and are divided into three clusters.

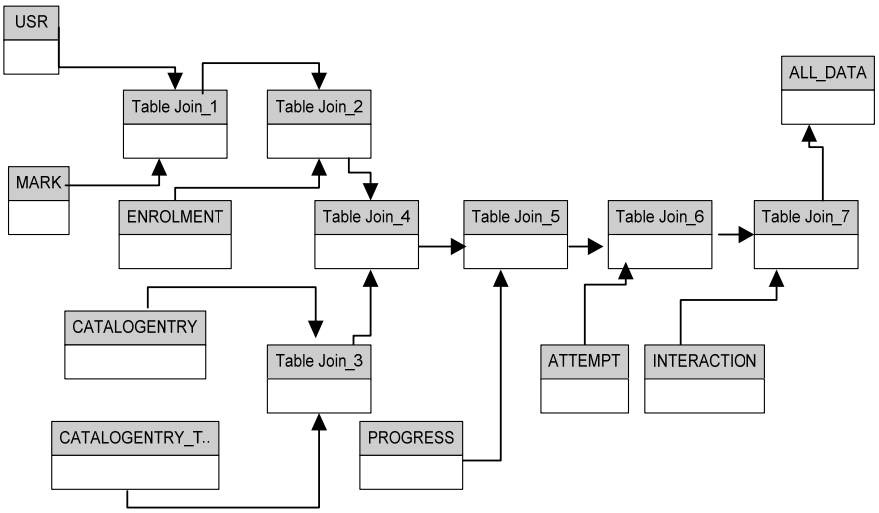

Figure 2. The data set gathering process

Another technique that describes the results is association rules. It is used to find hidden relationships. We use SIDE (Simultaneous Depth-first Expansion) algorithm that works in four steps: data examination, preparation, training and results presentation. On the initial step, the data statistics about pending data element pairs of how often it's iterant is placed together. In the next preparation step, the most iterant data element pairs are transformed into binary format. The training step includes available rules set generation and each one rule is evaluated iteratively and chosen under suitable criteria. During next iteration excluded rules are attached to new circumstances and evaluation process is repeated again. Using this algorithm we looked for associations between e-examination data attributes. This algorithm was developed by IBM.

5 The results of data analysis

During clustering three clusters were generated. First one includes the first exam pass, i.e even 71.54% of all records; second includes first exam's repass records – 21.95% of all; third – second exam's retake – 6.50% of all records. Data statistics is presented in Table 1

Table 1. Data statistics

Cluster	**1**			**2**			**3**		
Examination No.	**1**			**3**			**2**		
	Min	**Max**	**Mean**	**Min**	**Max**	**Mean**	**Min**	**Max**	**Mean**
Incorrect answers	2	21	12.75	7	17	12.13	3	19	11.59
Correct answers	4	23	12.22	8	18	12.88	6	22	13.41
Time spent	700	2556	1636.49	671	2.47	1411.25	304	2182	1174.6
Incorrect answers time	198	1980	978.47	316	1.75	892.75	202	1908	683.74
Correct answers time	233	1486	658.02	299	1.13	518.5	88	954	490.85

Analyzing clusters statistical data we noticed that there is deference between clustered records exam pass duration values. It is evident that students who passed the first time spent more time compared with the next passes. Although comparing with first passing time the second time was more successful, and the percent of correct answers was increased from 12.2 to 13.4, and students spent less time (about 25%).

To find out what kind of exam answers were the most difficult during first pass and first retake the association rules technique was deployed. In Table 2 the most interesting rules which were received using association are presented.

The obtained results show that students were the most serious about OOCo1q146, OOCo1q143, OOCo1q106, OOCo1q142, OOCo1q145 of questions. For example, the group of OOCo1q146 questions were incorrectly answered by even 77.80% learners. The analysis of study results highlighted the more difficult part of the course absorbed by students, refers to the course development policies. The present analysis provides

detailed feedback to the course instructor, releases should be emphasized, is difficult to understand the course is taught.

Table 2. The most difficult exam question groups

Rules	Frequency	Confidence	Lift
[QUESTION_ID=OOCo1q146]==>[RESULT=0]	3.10%	77.80%	1.464
[QUESTION_ID=OOCo1q143]==>[RESULT=1]	2.60%	64.80%	1.220
[QUESTION_ID=OOCo1q106]==>[RESULT=0]	2.50%	63.00%	1.185
[QUESTION_ID=OOCo1q142]==>[RESULT=0]	2.40%	61.10%	1.151
...	...	...	...

Continuing the investigation, we searched for the association rules of responses to difficult enough test questions between first pass and exam retake. These relationships are relevant to the analysis of student behavior during the overshoot. The selected result findings are presented in Table 3.

Table 3. The first examination and retaken results of more difficult exam question

Rules	Frequency	Confidence	Lift
[PASS=1]+[QUESTION_ID=OOCo1q146]==>[RESULT=0]	1.63%	84.62%	1.5932
[PASS=2]+[QUESTION_ID=OOCo1q146]==>[RESULT=0]	1.48%	71.43%	1.3449
[PASS=1]+[QUESTION_ID=OOCo1q143]==>[RESULT=0]	1.41%	73.08%	1.3759
[PASS=2]+[QUESTION_ID=OOCo1q143]==>[RESULT=0]	1.19%	57.14%	1.0759
...	...	...	...

The obtained results show that students during the exam retaking answered tougher test questions more accurately. Interpreting the findings suggest that the students examined were better prepared to retake or knew the right answers to the questions provided in the retake.

Continuing the analysis we searched for association between time spent and the right answer. First of all, we analysed the first test data. To find associations exam time spent is broken down into 12 intervals. The obtained results are presented in Table 4.

Table 4. The relationship between time spent and correct answer during first pass

Rules	Frequency	Confidence	Lift
[DURATION>=5 AND <22.5]==>[RESULT=0]	8.15%	68.62%	0.8167
[DURATION>=5 AND <22.5]==>[RESULT=1]	8.62%	31.38%	1.2698
[DURATION>=22.5 AND <40]==>[RESULT=0]	16.15%	58.99%	0.9908
[DURATION>=22.5 AND <40]==>[RESULT=1]	11.23%	41.01%	1.0136
[DURATION>=40 AND <57.5]==>[RESULT=0]	12.31%	61.54%	1.0136
[DURATION>=40 AND <57.5]==>[RESULT=1]	7.69%	38.46%	0.9506
[DURATION>=92.5 AND <110]==>[RESULT=0]	4.92%	59.26%	0.9953
[DURATION>=92.5 AND <110]==>[RESULT=1]	3.38%	40.74%	1.0069
...	...	...	...
[DURATION>=180]==>[RESULT=0]	4.92%	84.21%	1.4144

The table shows that the correction of a given answer depends on time. The students who were spending on the response from 5 to 22.5 seconds answered test questions correctly most often. If the candidates answering the question spent more than 180 seconds, the answers are often not correct.

In finding associations between the time spent to answer a question and correct answering during first-time retaking very similar results were obtained. Students, who respond within 40 seconds, usually answer the question correctly while those who take more than 180 seconds often were wrong.

Analyzing the reasons influencing these results, we can accurately identify and formulate the incidence of ambiguous exam questions. Therefore, the exam questions answering by students with difficulty and taking the most time should be further reviewed and revised.

6 Conclusions and further work

For analysis of electronic exam data were chosen descriptive data mining model's techniques: clustering and association rules. They help to identify patterns and relationship in electronic exam data.

After clustering technique we obtained statistical data which showed that the behavior of each student is different - each test retaking decreases the mean of exam time duration and increases the mean of the number of correct answers.

Using the association technique we examined questions that were given during the exam and student responses to them, received a set of rules, representing the most complex sets of questions. Due to the association technique the rules which showed students knowledge have been found.

The obtained results show the behavior of students in each test, as well as the exams, their reliability and the benefits of the study process.

In this work we suggest to apply data mining techniques to find the appropriate e-learning patterns, related to student evaluations and time spent for answers. This work will be used to improve VGTU electronic examination system.

The small-scale research of e-examination results was carried out, using data extraction techniques, and it presented an interesting and useful knowledge, according to which it is possible to establish a higher quality e-tests and to improve the examination process. However, this study used relatively limited data, the number of questions to which answers we had hoped for has not been great.

Further we plan to analyze the distance learning information and e-examination results of Master students. Currently we have the data for e-learning and e-examination data of two subjects (Virtual learning environment and Multimedia) for three years. Each year, about 20-30 graduate students study the courses. In e-environment they are not only examined, but are also supplied with all the course materials; they can get consultations, practical exercises, and self tests are provided. During this study, we will seek to identify more factors influencing the successful learning practice.

References

[1] **Ayesha A., Mustafa T., Khan M. I.** Data Mining Model for Higher Education System. *European Journal of Scientific Research.* 2010, volume 43, no.1, pp.24-29.

[2] **Delgado Calvo-Flores M., Gibaja Galindo E**. Predicting students‘ marks from Moodle logs using neural network models. *Current developments in Technology-Assisted Education*. 2006, pp.586-590.

[3] **Dunham H. M**. Data Mining: Introductory and Advanced Topics. *Prentice Hall*. 2007.

[4] **Garcia E., Romero C., Ventura S., Calders T.** Drawbacks and solutions of applying association rule mining in learning management systems. *Proceedings of the International Workshop on Applying Data Mining in e-Learning*, 2007, pp 13-23.

[5] **Halees A.** Mining Students Data to Analyze e-Learning Behavior: A Case Study. 2006.

[6] **Jun-Ming Su, Shian-Shyong Tseng, Wei Wang, Jui-Feng Weng, Jin-Tan David Yang, Wen-Nung Tsai**. Learning Portfolio Analysis and Mining for SCORM Compliant Environment. *Educational Technology & Society*. 2006, volume 9, no 1, pp. 262-275.

[7] **Liu F.-J., Shih B.-J**. Learning Activitybased E-learning Material Recommendation System. *Ninth IEEE International Symposium on Multimedia*. 2007, pp. 343–348.

[8] **Merceron A., Yacef K**. Educational Data Mining: a Case Study. *Proceedings of the 12th International Conference on Artificial Intelligence in Education AIED, IOS Press*. 2005, pp.20-26.

[9] **Merceron A., Yacef K**. Interestingness Measures for Association Rules in Educational Data. *Proceedings of the first International Conference on Educational Data Mining.* 2008, Canada, pp. 35-41.

[10] **Namdeo V**., **Singh A.**, **Singh D., Jain R.C.** Result Analysis Using Classification Techniques. *2010 International Journal of Computer Applications* (0975 - 8887), vol. 1, no. 22.

[11] **Shen L., Wang M., Shen R.** Affective e-Learning: using “emotional” data to improve learning in pervasive learning environment. *Educational Technology & Society*. 2009, volume 12, no. 2, pp. 176–189.

[12] Workplace Collaborative Learning Authoring Tool Guide. [interactive 2011.03.31]. Access via the Internet: http://sdo.miemp.ru/lms-help/guides/AuthoringToolGuide.pdf.