Data mining of students’ behavior in E-learning system

O M Gushchina* and A V Ochepovsky
Institute of Mathematics, Physics and Information Technology, Togliatti State University, 14 Belorussskaya Street, Togliatti, 445020, Russian Federation

*E-mail: g_o_m@tltsu.ru

Abstract. The article deals with educational data mining techniques aimed at increasing effectiveness of E-learning process as well as the idea of adaptive feedback, individual assessment and more personalized attention to student’s profile due to dynamic monitoring and tracking of students’ behavior in the E-learning system. The following techniques are identified: cluster analysis to determine the most popular time threshold for the task per session; analysis and visualization of data to highlight the main options that contribute to the effective completion of courses, and the most popular educational resources; V-fold cross-checking with the use of statistical processing aimed at students by their main indicators of activity to determine the correlation between high percentage of activity and academic performance. The proposed educational data mining techniques allow to assess student’s behavior in the E-learning system for understanding student’s interest in studying the learning materials and assessing the quality of educational content.

1. Introduction
Currently, there is a rapid development of e-education, in which any educational activity of the student is tracked and recorded in numerous databases, log files, personal profiles, etc. [1]. E-learning constantly generates large volumes of data describing continuous interaction between training and teaching, electronic systems and students. Moreover, the abundance of information about the student and his behavior formed in the E-learning system is an increasingly growing problem as uncontrolled information can give wrong ideas and misperceptions without providing any clear knowledge [2]. Modern techniques and technologies, including educational data mining (EDM) techniques, are needed to avoid this. In the field of E-learning, EDM are used to increase the level of students, to analyze student’s profile, to predict and improve student performance, etc. [3].

Several studies propose different approaches to the use of educational data mining techniques in the E-learning system. The article [4] shows cluster approach for dividing students into different groups based on their learning behavior. It also presents a personalized architecture of the E-learning system, which detects and responds to the educational content in accordance with the education opportunities of students. There is an application of data mining techniques to ensure the security of the E-learning process while working with a large volumes of available educational data [5]. The paper [6] shows the importance of personalization in providing effective methods of E-learning that gives the content based on the student’s performance. The authors of [3] believe that data mining techniques are used in E-learning to analyze the profile of students, to predict and improve the level of students, providing training in accordance with students’ requirements. The article [7] demonstrates experimental testing of data mining techniques aimed at identifying the compliance of the received answers from students and their behavior in the E-learning system. The authors of [8] analyze behavior patterns of a student to adjust
online learning strategies and have determined the ratio of E-learning indicators to create a student profile and gave countermeasures. The paper [9] solves the problem of creating electronic educational content that is interesting and customized to user needs. Analytics of student’s behavior in programming practice is used to provide timely and high-quality feedback using business technologies, such as click stream [10]. The authors of [11] develop a functional solution for classifying and understanding student behavior in Scratch-based programming. Through the data mining generated by students’ clicks at Scratch, they receive predictive analytics to identify patterns in student behavior when developing task solutions.

The analysis of the studies showed that data mining techniques, actively used in the E-learning system, are aimed at different characteristics of the student’s profile. But it is also important in the E-learning system to assess educational behavior in order to understand the interest of the students to the educational content and its realizability. Therefore, a new paradigm is needed to assess student behavior in the E-learning system based on data mining techniques. It is important to understand that large volumes of data are developed in the educational process. They are dynamically changed over the time, but it allows to provide results monitoring. Timely data mining and real time visualization of the results will allow students and teachers to adjust their approach to the educational process.

The purpose of the study is to build visual models of educational data mining techniques’ application for assessment the behavior of students in the E-learning system. These models will be focused on solving the problem of big data structuring to gain new knowledge that helps to accept the needs of students and to assess the educational content used in the E-learning system.

2. Methodology

In the E-learning system, there are many different types of data, both structured and unstructured, which are difficult to process using traditional statistical methods. Therefore, it is necessary to use new technologies that would allow to collect and process the large data flows, as well as the output of the results in a visual form [12]. The main data that needs to be processed are: student user registration data (provide analysis of student characteristics, including ID, first name, last name, gender, date of birth, etc.), web log data (reflect the operation of the E-learning platform, including the number of active users, number of page views, access time, activation speed and learning path), learning behavior data (useful for statistical analysis of online learning performance, including training time, training activities, learning resources and exam results) and educational content data (can be used to analyze teacher or teacher preferences, including viewing/collecting content, reviewing content, and interactive content).

For processing big data and output the obtained educational results, data mining (EDM) techniques are used that allow to extract and visualize the framework for determination of the learning behavior of students in E-learning system [13].

In this study, the cluster analysis was used for studying the educational behavior of students in the E-learning system. K-means algorithm was used for clustering, so data was combined according to the formula (1):

\[ J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 \]

where, \( \left\| x_i^{(j)} - c_j \right\|^2 \) – the square of the distance between the point \( x_i^{(j)} \) and the cluster center \( c_j \) [14].

The result of the clustering was the division of students into groups of similar objects, which are characterized by close values of some numerical or qualitative indicators. The data for the analysis were obtained in the framework of the implementation of online courses: "Transformation of information and educational environment by means of cloud technologies", "Electronic portfolio as an effective tool for assessing digital literacy" and "Formation of information competence in the development of digital educational resources". All of these courses were organized as informal education in a form of MOOC. Each course has several modules that contain theoretical material, practical tasks and testing. More than 1000 respondents took part in online learning, they completed tasks with different time of understanding the material and showed different final grades (figures 1-3).
For the purity of the experiment on assessing students’ behavior in the E-learning system, their performance was measured by several parameters as attendance of the course: the number of users on the course and the meantime spent by them for studying (figures 4a, 4b); statistical indicators of tasks’ performance (figure 5); attendance of the tasks to view theoretical data and the results (figures 6a, 6b); and the rating score of course assignments (table 1) that provides the dynamics of assignments and displays the percentage of correctly completed tasks.
Table 1. Statistics of students’ assignment completion in electronic course.

| Average response time | Correct answers | Incorrect answers | Average mark |
|------------------------|----------------|-------------------|--------------|
| 00:01:17               | 126 (79%)      | 33 (21%)          | 0.79 (79%)   |
| 00:00:54               | 124 (79%)      | 32 (21%)          | 0.79 (79%)   |
| 00:00:30               | 127 (79%)      | 32 (21%)          | 0.80 (80%)   |
| 00:00:58               | 125 (77%)      | 37 (23%)          | 0.77 (77%)   |
| 00:00:45               | 136 (86%)      | 22 (14%)          | 0.86 (86%)   |
| 00:00:45               | 34 (21%)       | 124 (79%)         | 0.22 (22%)   |

Figure 4. Graph containing students’ attendance of electronic course: (a) the total number of students in the course on the current date (b) the meantime spent by students for completing the course on the current date.

Figure 5. Summary for tests built into the electronic course on the current date.
The obtained activity indicators of students in online courses were used for processing them with educational data mining techniques.

3. Results and discussion
In the E-learning system the student has been marked depending on the execution time of the task for 1 session time as a "type 1" (if the educational process for 1 session is 60 min), ..., type "5" (if academic work is less than 10 minutes). Figure 7 shows selected clusters of student learning outcomes grouped according to the time spent in the E-learning system. Based on the obtained results, it can be concluded that most students do not spend more than 25 minutes to complete the tasks.

In addition, the analysis of the students’ actions in the E-learning system was carried out and the students’ attitude to the learning process was assessed to determine the level of their impact on the received learning effect.

Analysis and visualization of educational data (the presence of a clear learning goal or its absence, as well as the presence of a training plan or not, etc.) allowed to reflect the subjective initiative of students [15] and their educational behavior in the E-learning system. All of these actions lead to the analysis of E-learning interference factors and enhancing its effectiveness (table 2).
Figure 7. Analysis of the dependence of learning outcomes on time for 1 session of work in the E-learning system.

Table 2. Analysis of attitudes to learning.

| Actions in the E-learning system                  | Number of students (%) |
|--------------------------------------------------|------------------------|
| Study with maximum effort                        | 94.30%                 |
| Study without interest                           | 5.70%                  |
| Have clear learning goals                        | 18.70%                 |
| Understand learning goals                        | 58.60%                 |
| Make learning route                               | 6.70%                  |
| Do not make learning route                        | 45.90%                 |
| Focus on training tasks                          | 17.51%                 |
| Put a greater emphasis on training               | 55.73%                 |
| Pass the time uselessly in the E-learning system | 16.88%                 |
| Drop out studying                                | 9.87%                  |

The analysis showed that about 94.3% of students believe that distant learning is beneficial for them. 18.7% have clear learning goals and 58.6% accept learning goals and understand the need for education. These indicators clearly demonstrate that among students there are those who have a clear idea of the final learning outcomes. However, among the students there are those who do not pay attention to the process of studying learning materials: 45.9% - do not build the path of the educational route of the training plan, 16.88% - in the process of learning are engaged in secondary matters. According to figure 8, it can be concluded that effective learning using distant learning technologies requires specific goals, internal motivation, synchronous feedback and the ability of students to plan the educational process.

During the online courses, many reviews were received from students. On the basis of the reviews, a subjective perceptions about the ways of perceiving the learning material and the quality of its understanding, as well as other students’ actions in the E-learning system, were formed. The obtained data structuring allowed to identify the main actions of students when studying online courses and to determine the level of their attention to educational resources. According to the statistics of frequency [16], the main actions in the E-learning system were identified, which are more likely to be taken by students (figure 8).
The given statistics shows that the behavior associated with the execution of actions 1-7, related to independent learning of students; behavior action with 8-10 focused on the organization of interaction, and the actions 11-12 – do not apply to the educational process. Most students view text and often take notes; therefore, a text resource is the most popular type of resource. About 70% of the students will first look multimedia resources, and then perform the job; 50% - perform online practical exercises; 60% choose to view learning objectives before starting the course; more than 80% of students start the learning process slowly, spending a lot of time on actions unrelated to the educational process.

The analysis showed that most students are interested in multimedia resources, but they also use text resources; they try to find answers to the solution of practical tasks through learning environments, paying little attention to interactive interaction. In addition, many students in the learning process use distracting sources, such as chats, listening to music and more.

To analyze the time spent by students to work with different tools of the electronic environment, a V-fold cross-check was carried out [17]. It allowed the educational data to be grouped into 3 classes according to the main indicators of students’ activity. In accordance with the selected clusters were defined learning styles and learning behavior of students in the E-learning system (figure 9). The final assessment of students for each cluster was calculated as 1 – 4,564,568182, 2 – 4,836111, and 3 – 4,604839.

As it can be seen from the figure 10 and received calculations of the average value of the final assessment, all students from cluster 1 are the most active in the E-learning system, but have the lowest final grades. For a more in-depth analysis, cluster 1 was considered separately using additional criteria for evaluating the performance of students. Each student in the E-learning system is identified by identification number (e.g. id_123129). In assessing the educational behavior, many parameters are taken into account, such as the time spent on task completion, the time spent on other resources, the number of clicks when re-visiting the online course, etc. were taken into account. Students’ actions in the E-learning system highlighted in the figure 10 were divided into 3 groups: important, not very important, no matter. The total time for the given groups was summed up to assess the time spent on tasks. Figure 10 shows the statistics of the distribution of time spent by students from the first cluster in the E-learning system, to solve educational problems (figure 9). The analysis showed that in the learning process, students not only spent time on completing educational tasks, but also spent it an ineffective
way on checking personal achievements, searching for active participants in the educational process, discussing, etc.

Figure 9. Clustering of students’ activities using $k$-means: 1 – average score on graduation; 2 – number of course connections; and time spent 3 – in the system; 4 – on tools; 5 – on special tools; 6 – on calendar tools; 7 – on chat tools; 8 – on learning material; 9 – on discussion tools; 10 – on file manager tools; 11 – on mail tools; 12 – on media; 13 – on gradebook tools; 14 – on online workshop tools; 15 – on notebook tools; 16 – on educational content; 17 – on course outline; 18 – on student tools; 19 – on hyperlink tools (hyperlink to other educational content or web pages); 20 – on other online users with search tools.

Thus, a high percentage of the use of work in the E-learning system is not a guarantee of good performance. Since the purposeful, logical use of important tools is an important condition for the effectiveness of the training course, having information about this condition can improve the study of the course and the time spent on tasks.

Figure 10. Analysis of students’ activities from first cluster according to efficiency of their activities.
4. Conclusion
The use of cluster analysis to study the educational behavior of students in the E-learning system allowed to identify the main groups of students depending on the time of implementation of educational tasks. It was determined the most popular time threshold that a student spends on training for 1 session of work.

Using the methods of analysis and visualization of educational data, the attitude of students to the use of distant learning technologies was determined, and the main options that contribute to the effective passage of courses in the E-learning system were identified. Also, the analysis and visualization of educational data allowed to determine the level of students’ attention to educational resources, among which the most popular are text and multimedia. In addition, it was determined that some students in the learning process use distracting resources, such as chats, listening to music and more.

The use of V-fold cross-checking allowed students to be grouped into 3 classes according to the main indicators of their activity to determine educational behavior, as well as to conduct additional statistical processing, during which it was determined that a high percentage of activity in the E-learning system is not a guarantee of good performance. Since the purposeful, logical use of important tools is an important condition for the effectiveness of the training course, having information about this condition can improve the study of the course and the time spent on tasks.

Thus, the educational data mining techniques can be used for studying the educational behavior of students in the E-learning system in order to understand the interest of the students to the educational contents and its realizability. The results of the study showed that different students with different levels of academic performance differently distribute their activities in the online environment that is, they have different educational behavior.

References
[1] Estacio R and Raga Jr R C 2017 Analyzing students online learning behavior in blended courses using Moodle. Asian Association of Open Universities Journal 12(1) 52 https://doi.org/10.1108/AAOUJ-01-2017-0016
[2] Brajkovic E, Rakic K and Kraljevic G 2018 Application of data mining in E-learning systems. Proc. 17th Int. Symp. Infoteh-Jahorina (INFOTEH) (New Jersey: Piscataway) p 146 doi: 10.1109/INFOTEH.2018.8345536
[3] Mustamid A El, En-Naimi E and Boughidji J El 2017 Integration of data mining techniques in E-learning systems: Clustering Profiles of Learners and Recommender Course System. BDC’A’17: Proc. 2nd Int. Conf. on Big Data, Cloud and Applications (Tetouan, Morocco: Association for Computing Machinery) p 1 https://doi.org/10.1145/3090354.3090453
[4] Kausar S, Xu H, Hussain I, Zhu W and Zahid M 2018 Integration of data mining clustering approach with the personalized E-learning system. Preprints 6 72724 doi: 10.20944/preprints201808.0350.v2
[5] Agarwal A, Patel A V and Saxena A 2018 Secure E-learning using data mining techniques. Proc. 3rd Int. Conf. on Internet of Things and Connected Technologies (ICIoTCT) (New York: Springer) p 26 http://dx.doi.org/10.2139/ssrn.3167309
[6] Ramyasree P, Bhuvaneswari T, Vamsi V, Reddy S and Jonnalagadda S K 2019 Personalized E-learning system based on user's performance and knowledge: an adaptive technique. International Journal of Recent Technology and Engineering 8(4) 8695 doi:10.35940/ijrte.D8899.118419
[7] Martin C, Mesarosova M and Balogh Z 2011 Analysis of students' behaviour in E-learning system. Proc. 22nd EAEIE Annual Conf. (New Jersey: Piscataway) p 1
[8] Liang K, Zhang Y, He Y, Zhou, Tan W, Li X 2017 Online behavior analysis-based student profile for intelligent E-learning. Journal of Electrical and Computer Engineering 2017 1 https://doi.org/10.1155/2017/9720396
[9] Baharudin A, Sahabudin N and Kamaludin A 2017 Behavioral tracking in E-learning by using learning styles approach. Indonesian Journal of Electrical Engineering and Computer Science 8(1) 17 doi:10.11591/ijeeec.v8.i1.pp17-26
[10] Amo Filva D, Alier Forment M, García-Penalvo F, Fonseca D and Casan Guerrero M 2018 Learning analytics to assess students’ behavior with scratch through clickstream. Learning Analytics Summer Institute Spain – LASI 2188 74 http://ceur-ws.org/Vol-2188/Paper8.pdf

[11] Filva D, Forment M, García-Penalvo F, Eccudero D and Casan M 2018 Clickstream for learning analytics to assess students’ behavior with Scratch. Future Generation Computer Systems 93 673 doi:10.1016/j.future.2018.10.057

[12] Vaitsis C, Hervatis V and Zary N 2016 Introduction to Big Data in education and its contribution to the quality improvement processes. Big Data on Real-World Applications 113 58 http://dx.doi.org/10.5772/63896

[13] Guleria P and Sood M 2014 Data mining in education: a review on the knowledge discovery perspective. International Journal of Data Mining & Knowledge Management Process 4(5) 47 https://doi.org/10.5121/ijdkp.2014.4504

[14] Krone T, Albers C J, Kuppens P and Timmerman M E 2018 A multivariate statistical model for emotion dynamics. Emotion 18(5) 739 https://doi.org/10.1037/emo0000384

[15] Papamitsiou Z and Economides A A 2015 Temporal learning analytics visualizations for increasing awareness during assessment. International Journal of Educational Technology in Higher Education 12(3) 129 doi:10.7238/rusc.v12i3.2519

[16] Wong B 2017 Learning analytics in higher education: an analysis of case studies. Asian Association of Open Universities Journal 12(1) 21 https://doi.org/10.1108/AAOUJ-01-2017-0009

[17] Krpan D and Stankov S 2012 Educational data mining for grouping students in E-learning system. Proc. 34th Int. Conf. on Information Technology Interfaces (Piscataway: IEEE Operations Center) p 207 doi:10.2498/iti.2012.0470