Weekly Analysis of Moodle Log Data in RStudio for Prediction

Neslihan Ademi
Computer Engineering Department
International Balkan University
Skopje, North Macedonia
neslihan@ibu.edu.mk

Suzana Loshkovska
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
suzana.loshkovska@finki.ukim.edu.mk

***Abstract*— This paper presents the results of the study to find in how many weeks the prediction/ classification can be done through a learning management system. The results are obtained by analyzing the effects of learners' online presence and activities in Moodle on their grades at the early stages. Analyses are done by partitioning log data of a course in three years by time in terms of weeks. For this purpose we used RStudio and developed script to automatize the analysis. We found that starting from the third week of the lecture period online presence of the students becomes stable and classification can be done starting from that time and accordingly different material and assessment methods can be offered to the students in LMS.**

Index Terms—*Adaptive Learning Systems, Moodle logs, Learning Analytics, Data Mining, Online Presence*

I. INTRODUCTION

Learning management systems (LMSs) such as Moodle are commonly used in the modern education settings as they offer many opportunities to the students and to the tutors. On the other side they open a door for Educational Data Mining (EDM). Huge amount of data stored in LMS databases and log files give the possibility of gaining knowledge about the students following a course. The focus of this paper is the knowledge which can be achieved through the log files by analyzing weekly presence of the students.

Log data collected through LMSs provide descriptive overview of users' online behavior. According to [1] observing behavior provides insights about how people interact with existing systems and services and reveals surprises.

Observational Log Studies contain two common ways to partition log data; by time and by user. Partitioning by time is interesting because log data often contains significant temporal features, such as periodicities (including daily, weekly, and yearly patterns) and spikes in behavior during important events. It is often possible to get an up-to-the-minute picture of how people are behaving with a system from log data by comparing past and current behavior. It is also interesting to partition log data by user characteristics [1].

Recently there are many studies in the literature about log analysis in e-learning environments [2]–[6]. In [2], authors profile the students' learning approaches through the Moodle logs. Some of them are for only visualization purposes [3], some are for the prediction like [4] and [5].

Data mining and Learning Analytics can be useful to explore, visualize, and analyze in order to understand students' learning behavior and apply the gained knowledge for the adaptation of the learning contents.

The e-learning data mining process consists of four steps like in the general data mining process. Collect data, Pre-process the data, Apply data mining, Interpret, evaluate and deploy the results. [7]

In a previous study [8] we examined the relationship between variables which are obtained from Moodle logs and students' grade. We found positive correlations which show the linear relationship between all Moodle activities and the course grade of the students. In [9], we applied three Machine Learning algorithms; namely Decision Tree(DT), Bayesian Network(BN) and Support Vector Machine(SVM) to make prediction of the grades based on online activities of the students. In [10], we found the possibility of early predictions of grade through Moodle log data considering 7 weeks out of 14 weeks of lecturing time and we suggested early predictions to be used for preventing drop outs.

The purpose of this study is to find the relation between the online presence and the grades in an early stage. The research questions which we want to find answer for are: "1. How many weeks would be enough for an adaptive system to categorize the learners according to their online activities?" and "2. Which online activities can be taken into consideration in that time?"

This paper presents the possibility of making predictions in earlier weeks of studies by making weekly analysis of online activities. The second section defines the used methodology for the study, while the third section gives results and discussion; finally the last section is the conclusion.

II. METHODOLOGY

In this section, the used methodology is explained. For the study, log files are taken from Moodle which is installed and used at the Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Skopje, North Macedonia. Recorded log files of Moodle LMS are downloaded in the form of .csv and they contain all online activities of the students for a one semester bachelor's degree course of User Interfaces. The course is in the third year of the studies. For the analysis to find the trends in online presence three academic years in a row are taken: 2016-2017, 2017-2018 and 2018-2019, where the teaching process was designed as blended learning. Moodle was used to support classroom teaching to distribute course material,

lectures, homework, laboratory exercises and to provide discussion through the forums within the course.

The standard retrieved fields in the log files are: Time, User full name, Affected user, Event context, Component, Event name, Description, Origin, IP address. Log files were composed of approximately 150K rows for each. Number of regular students was 255 for the first year, 205 for the second year and 260 for the third year.

A. *Data Pre-processing*

Applied pre-processing steps to the data were, data cleaning, transformation and integration. We used sqldf package which allows complex database queries in Rstudio. All pre-processing steps in details are listed below:

1. Clean the log files from the events performed by instructors and administrators as the focus is the students' actions.
2. Remove log data produced by the system by filtering the data where component field is system.
3. Remove duplicate records.
4. Extract userID from the description text and generate new columns for userID to be used instead of user full name to provide anonymity.
5. Transform the file into a data frame with the name of "events.csv" which will contain data attributes given in Table I.
6. Integrate the new file (events.csv) to the scores file (scores.csv) which gives the final grades of the students.

TABLE I. ATTRIBUTES AFTER DATA TRANSFORMATION

Name	Description
UserID	ID number of the student
Visits	Total number of visits by the student
Quizzes	Number of quizzes taken by the student
Assignments	Number of submitted assignments by the student
ForumCreated	Number of forum creations by the student
ForumView	Number of forum views by the student
CourseView	Number of course views by the student
FileSubmission	Number of file submissions by the student
GradeView	Number of grade views by the student

B. *Data Exploration and Statistical Analysis*

Data exploration and statistical analysis consist of following steps; visualizing online presence of the students for each week, correlations of total visits, quizzes, assignments, forum creations, forum views, file submissions, and grade views with the course grade.

Correlations are found by using Pearson correlation test. Equation 1 gives the formula for Pearson correlation coefficient. The value of r is always between –1 and +1. r = –1 or +1 indicates a perfect linear relationship, where sign indicates the direction and r = 0 indicates no linear relationship.

$$r = \frac{\sum (x-\bar{X})(y-\bar{Y})}{\sqrt{\left[\sum (x-\bar{X})^2\right]\left[\sum (y-\bar{Y})^2\right]}} \quad (1)$$

We have used Pearson correlation coefficient to discover whether there is a linear relationship between the attributes given in Table I and the grade.

III. RESULTS AND DISCUSSION

A. *Frequencies of student activities*

Upon the analysis of the log data; we noticed that, for the first year in the first week of the semester only 143 students visited the course page on Moodle. This means 44% of the students did not have any online presence. In the second week online presence is increasing to 91%, and in the third week to 96%. Fig.1 gives the online presence in the course in terms of number of active students on Moodle for the first seven weeks of the semester for the academic year 2016-2017

Fig.2 gives the online presence of the students for the academic year 2017-2018. As it can be calculated from the figure, in the first week online presence is 68%, in the second week 90% and in the third week 99%.

Fig.3 shows the same trend as in Fig.1 and 2; starting from the third week online presence is regular.

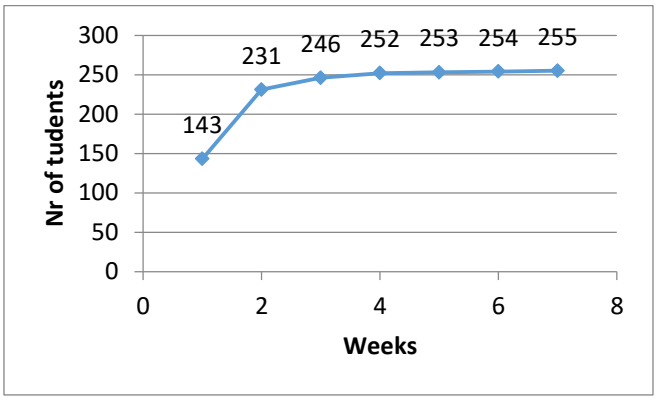

Fig. 1. Number of active students on Moodle per week in 2016-2017

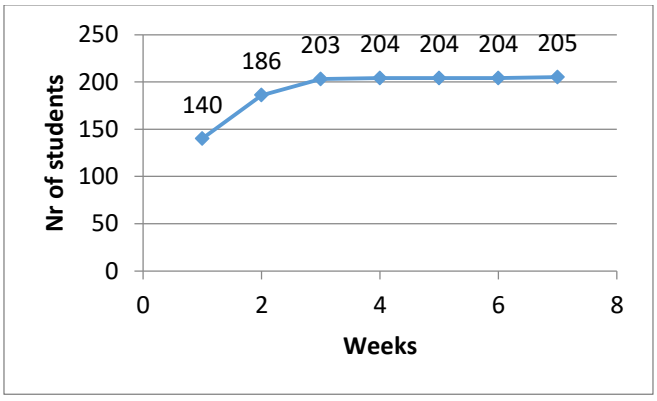

Fig. 2. Number of active students on Moodle per week in 2017-2018

TABLE II. CORRELATIONS OF TOTAL VISITS AND COURSE VIEWS WITH THE GRADE BY WEEKS (2016-2017)

	1 week	2 weeks	3 weeks	4 weeks	5 weeks	6 weeks	7 weeks
Total Visits	0.17	0.26	0.34	0.33	0.37	0.35	0.35
Course Views	0.13	0.25	0.31	0.30	0.29	0.27	0.28

TABLE III. CORRELATIONS OF TOTAL VISITS AND COURSE VIEWS WITH THE GRADE BY WEEKS (2017-2018)

	1 week	2 weeks	3 weeks	4 weeks	5 weeks	6 weeks	7 weeks
Total Visits	0.12	0.22	0.30	0.28	0.30	0.25	0.25
Course Views	0.11	0.21	0.26	0.21	0.22	0.21	0.21

TABLE IV. CORRELATIONS OF TOTAL VISITS AND COURSE VIEWS WITH THE GRADE BY WEEKS (2018-2019)

	1 week	2 weeks	3 weeks	4 weeks	5 weeks	6 weeks	7 weeks
Total Visits	0.01	0.11	0.22	0.26	0.27	0.26	0.27
Course Views	-0.04	0.05	0.16	0.2	0.23	0.24	0.24

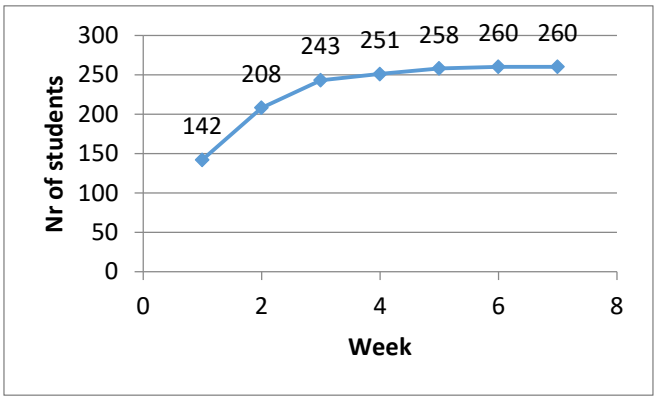

Fig. 3. Number of active students on Moodle per week in 2018-2019

Overall picture of three academic years show that starting from the third week of the lectures online presence becomes more stable.

B. Correlations

Table II, III and IV give the correlations of total visits and course views with grade by weeks for the academic years 2016-2017, 2017-2018 and 2018-2019 respectively. As it can be seen from these tables, after the 3rd week of the lectures students' online activities have similar correlations with their grade, they are following the same trend (also see Fig.4, 5 and 6).

In Table II, III and IV, only first seven weeks of the lectures are given as our focus is to see how much earlier the correlations start to be stable for early predictions. In the study, correlations of other online activities extracted from log files (given in Table I) with the grade are calculated but they are not shown in Table II, III and IV. Some of those correlations are not available in early stages such as grade views, forum views and forum creations or they are not significant with very low correlations such as quizzes, file submissions and assignments.

When the total course period of 14 weeks was taken into consideration as in our previous study [8], correlation coefficient of the total visits was 0.55 and correlation coefficient for the course views was 0.43 for the academic year 2016-2018. For the seventh week it is calculated here for the same year as 0.35 for total visits and 0.28 for course views.

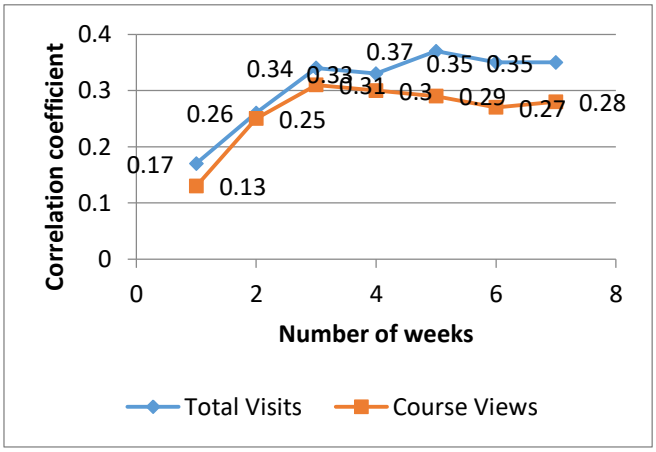

Fig. 4. Correlation coefficients of total visits and course views per week in 2016-2017

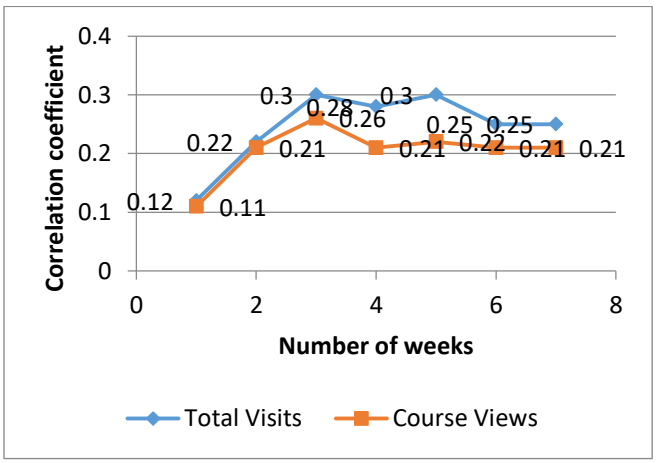

Fig. 5. Correlation coefficients of total visits and course views per week in 2017-2018

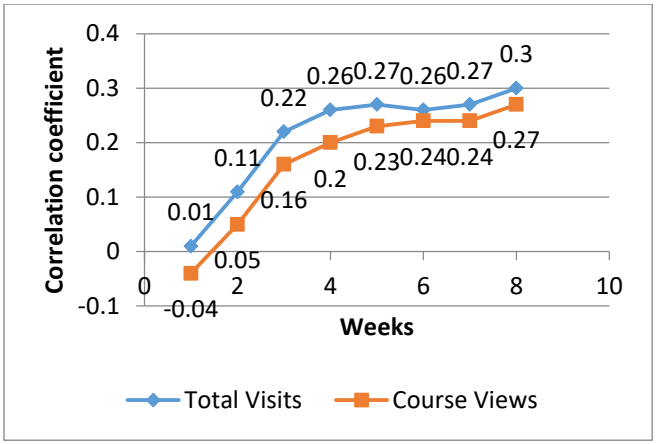

Fig. 6. Correlation coefficients of total visits and course views per week in 2018-2019

IV. Conclusion

In this paper we used some learning analytics techniques such as pre-processing and statistical analysis; to discover the effect of students' online presence on their grades for the early stages of the learning period. This knowledge will be a base of our future data mining process. We also developed an R script to automatize these analyses for the future use in our more detailed examinations and predictions.

Present study showed the relationship between variables which are obtained from Moodle logs and students' grade for the first half of the learning period. The correlations for the attributes such as forum views, forum creations, file submissions, assignments and quizzes are not as significant as for the complete period of the learning process. Some of those correlations are not available in early stages such as grade views, forum views and forum creations or they are not significant with very low correlations such as quizzes, file submissions and assignments. We found correlations between total visits and course views with grade are higher and these correlations tend to be stable starting from the third week of the lectures. So for the future adaptive learning system architecture total visits and course views can be taken as two of the criteria in terms of online presence for classification purpose in the early weeks, starting from the third week. And after the second half of the course period the other attributes such as forum views, file submissions, quizzes and assignments can be included.

Future direction of our research will be analyzing also preferences of the students in terms of learning materials and assessment methods for applying data mining techniques to provide adaptivity in learning management systems.

References

[1] S. Dumais, R. Jeffries, D. M. Russell, D. Tang, and J. Teevan, "Understanding User Behavior Through Log Data and Analysis," in *Ways of Knowing in HCI*, New York, NY: Springer New York, 2014, pp. 349–372.

[2] G. Akçapınar, "Profiling Students' Approaches to Learning through Moodle Logs," *Proceedings of Multidisciplinary Academic Conference on Education, Teaching and E-learning in Prague 2015, Czech Republic (MAC-ETeL 2015)*, no. December, p. 7, 2015.

[3] A. Konstantinidis and C. Grafton, "Using Excel Macros to Analyse Moodle Logs," *UK Research.Moodle.Net*, no. September, pp. 4–6, 2013.

[4] Á. Figueira and Álvaro, "Mining Moodle Logs for Grade Prediction," in *Proceedings of the 5th International Conference on Technological Ecosystems for Enhancing Multiculturality - TEEM 2017*, 2017, pp. 1–8.

[5] T. Käser, N. R. Hallinen, and D. L. Schwartz, "Modeling exploration strategies to predict student performance within a learning environment and beyond," in *Proceedings of the Seventh International Learning Analytics & Knowledge Conference on - LAK '17*, 2017.

[6] M. Cocea and S. Weibelzahl, "Log file analysis for disengagement detection in e-Learning environments," *User Modeling and User-Adapted Interaction*, 2009.

[7] C. Romero, S. Ventura, and E. García, "Data mining in course management systems: Moodle case study and tutorial," *Computers and Education*, vol. 51, no. 1, pp. 368–384, 2008.

[8] N. Ademi and S. Loshkovska, "Exploratory Analysis of Student Activities and Success Based On Moodle Log Data," in *16th International Conference on Informatics and Information Technologies*, 2019.

[9] N. Ademi, S. Loshkovska, and S. Kalajdziski, "Prediction of Student Success Through Analysis of Moodle Logs: Case Study," in *11th International Conference, ICT Innovations 2019, Ohrid, North Macedonia, October 17–19, 2019, Proceedings*, S. Gievska and G. Madzarov, Eds. Springer, Cham, 2019, pp. 27–40.

[10] N. Ademi and S. Loshkovska, "Early Detection of Drop Outs in E-Learning Systems," *Academic Perspective Procedia*, vol. 2, no. 3, pp. 1008–1015, Nov. 2019.