

Trends in student behavior in online courses

Rianne Conijn^a and Menno van Zaanen^a

^aDepartment of Communication and Information Sciences, Tilburg Center for Cognition and Communication, Tilburg University, The Netherlands.

Abstract

Learning management systems provide an easy and effective means of access to learning materials. Students' access to course material is logged and the amount of interaction is assumed to be a measure of student engagement within the course. In previous research, typically frequencies of student activities have been used, but this disregards any temporal information. Here, we analyze the amount of student activity over time during courses. Based on activity data over 11 online courses, we cluster students who show similar behavior over time. This results in three different groups: a large group of students who are mostly inactive; another group of students who are very active throughout the course; and a group of students who start out being active, but their activity diminishes throughout the course. These groups of students show different performance. Overall, more active students yield better results. In addition to these general trends, we identified courses in which alternative trends can be found, such as a group of students who become more active during the course. This shows that student behavior is more complex than can be identified from an individual course and more research into patterns of learning activities in multiple courses is essential.

Keywords: *Learning management systems, student engagement, temporal information, clustering, student performance*

1. Introduction

Learning management systems (LMSs) are widely used to provide course content and other learning materials online in a structured way. Students' activities within these systems are often logged. These data may be used as a source of measuring student behavior in the fields of educational data mining and learning analytics, to investigate improvements of learning and teaching. Typically, researchers analyze frequencies of activities in the LMS (e.g., Romero et al., 2013; Zacharis, 2015). However, these metrics provide no information about the timing and spread of learning activities. As students may vary in the amount of activity during the course, in this study we analyze patterns in sequences of learning activities in 11 open online courses. The sequences of learning activities are clustered to identify trends in learning behavior. Although having insight in the different trends of student behavior is already interesting, patterns in learning activities may provide a more accurate representation of learner engagement compared to aggregated frequencies of activities (Hadwin et al., 2007), and hence can be more useful for performance prediction. Therefore, we also investigate the relationship between clusters of learning behavior and student performance.

2. Background

2.1. Topics in educational data mining and learning analytics

The fields of educational data mining and learning analytics focus on the use of educational data to gain insight in learning processes and to improve learning and teaching. Several tasks can be distinguished, such as student modeling, prediction of student performance, visualization of student behavior, and social network analysis (Romero & Ventura, 2010). For these tasks, typically aggregated counts of activities in the LMS are used (e.g., Romero et al., 2013; Zacharis, 2015). In this study, we focus on information that can be derived from the sequences or the order of activities (without aggregation over time) in the LMS.

2.2. Analysis of sequences of learner activities

To identify patterns in the sequential learning behavior, sequences of activities that display similar trends over time may be clustered. Clustering sequences of learner activities has been used to identify patterns in various learning contexts, such as group work (Perera et al., 2009), mathematical exercises (Desmarais & Lemieux, 2013), educational games (Bergner et al., 2014), and discussion forums (Cobo et al., 2010). Clustering is also used in intelligent tutoring systems to determine differences in event sequences over time (Klingler et al., 2016) or to identify patterns with interesting temporal behavior (Kinnebrew et al., 2013). Most studies analyzing sequences of learner activities look at a single session per

student, instead of all sessions within a course. This is, for instance, common in web mining, as it is mostly impossible to identify users across different sessions. However, in online learning environments, users often have to login and hence can be followed across multiple sessions.

Cobo et al. (2010) clustered student activity in the discussion forum of an online course across multiple sessions. Three different activity profiles were found: inactive profiles, profiles with regular activity throughout the course, and profiles with limited amount of activities in different periods. In the current study, we also cluster sequences of activities across multiple sessions. Contrary to Cobo et al. (2010), we analyze sequences of activities in all parts of the LMS, using multiple (11) courses instead of one. Additionally, the relation between student performance and patterns of learning behavior are analyzed.

2.3. Relation between sequences of learner activities and student performance

In learning analytics and educational data mining, the analysis of learner behavior is often used to predict student performance. Studies on frequencies of learning activities in LMSs generally find that more activity typically leads to higher grades (e.g., Zacharis, 2015). However, it is also shown that the effects of frequencies of activities on student performance differ across courses (Conijn et al., 2016; Gašević et al., 2016). This might be because frequencies are not concrete measurements of theoretical concepts, such as motivation or engagement, which are established predictors of student performance.

Patterns of learner activities are argued to provide a more accurate representations of learner engagement with respect to frequencies (Hadwin et al., 2007). Hence, they can be more useful for performance prediction. Moreover, they might provide insight in the reason behind (un)successful behavior, which can be used for interventions and help. For example, Perera and colleagues (2009) identified patterns leading to (un)successful group work, which in turn could be used by the facilitators to help the students. Accordingly, we analyze the relation between student performance and patterns of learner activities in open online courses.

3. Method

3.1. Data

Data were collected from the restricted open source dataset Canvas Network Courses, Activities, and Users (Canvas Network, 2016). This dataset consists of anonymized Canvas data from open online courses taught between March 2014 and September 2015. The data consist of a main table with all page requests per user and tables describing the course items per course, such as assignments, quizzes, forum, and wiki. In total, there are 359 courses

and 464,602 cases (enrollments) in the dataset. The 302,134 unique students (students could follow multiple courses) accounted for more than 258 million page requests. There is no detailed information (e.g., age, background) available about the students.

3.2. Data pre-processing

Data pre-processing and analysis was done using R. First, all data related to activity outside the course period were removed. Ten courses which missed a course start or course completion date were removed. We selected courses with on average at least 20 page requests per user and student performance data available. Student performance was calculated by the normalized average quiz grade. For each quiz submission, the grades were linearly transformed with respect to the minimum and maximum grade obtained for that specific quiz on a range from 0 to 100. Only quizzes were included where at least 50 students finished the quiz, with a maximum grade higher than zero, and at least some variation in the grades (*S.D.* normalized grade ≥ 0.2). Based on these quizzes, the average grade per course per student was calculated. Grades were set to missing if the student did not finish a quiz in that course.

The 147 remaining courses lasted between 13 and 703 days ($M = 81$, $S.D. = 85$). To compare the sequences between the courses, a subsample of courses of similar length was chosen. The most common course length of 43 days was found in 12 courses. One additional course was removed, because all students only showed activity in 31 of the 43 days. Hence, 11 courses with 4,429 unique students ($M = 425$, $S.D. = 116$ per course) were analyzed. The courses were in the domains of Education (4x), Social Sciences (2x), Humanities (2x), Physical Sciences (1x), Professions and Applied Sciences (1x), and Computer Science (1x).

3.3. Data analysis

Clustering was used to identify patterns in the sequences of learner activities within the 11 courses. Since analyzing single page requests leads to too fine-grained information, the number of page requests per student were aggregated per day. This resulted in sequences of 43 numbers per student representing the number of page requests on each day of the course. Due to the highly-skewed distribution of the number of page requests per day ($M = 153$, $S.D. = 306$), the page requests were binned into: no activity, low activity (< 3 page requests), medium activity (3 to 100 page requests), and high activity (> 100 page requests).

The sequences are clustered for all courses combined as well as for each course separately, according to the procedure described by Gabadinho and colleagues (2011) with the R packages ‘TraMineR’ and ‘cluster’. To cluster the sequences, the differences between the sequences within each cluster need to be minimized, while the differences between the

clusters need to be maximized. The distances between the sequences are computed with pairwise optimal matching (OM). The obtained distance matrix agglomerative hierarchical clustering (AHC) is used to cluster the sequences with Levenshtein distance ('ward' in R). The obtained clusters are visualized with state distribution plots per cluster. A series of one-way ANOVAs with Tukey post-hoc tests were conducted on the normalized mean grade, to determine whether student performance differed significantly between the clusters.

4. Results

First, the clusters of the sequences of all 11 courses combined were analyzed. The sequences of activities of students were found to cluster into three different groups (see Figure 1). The first and largest cluster consists of students who barely showed activity ($n = 4,212$) and their activity diminishes even more over time. The second cluster consists of students who were highly active on most of the days during the whole course ($n = 203$). The students in the last cluster showed some activity in the beginning of the course, but the activity decreased during the course ($n = 265$). Clustering into more clusters did not result in new patterns, but merely in clusters with gradations between clusters 2 and 3.

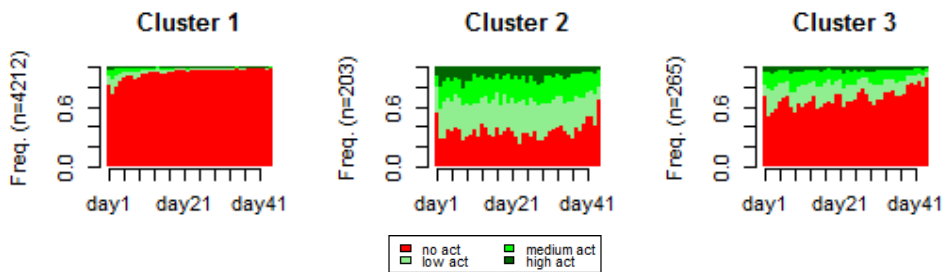


Figure 1. State distribution plots learning behavior in Canvas for all 11 courses.

A one-way ANOVA showed significant differences in student performance between the three clusters ($F(2,1977) = 164, p < .001$). As expected, students in cluster 1 received significantly lower grades ($M = 42, S.D. = 41$) than students in clusters 2 and 3. Students in cluster 2 received a higher grade ($M = 86, S.D. = 18$) compared to cluster 3 ($M = 73, S.D. = 27$). Additionally, in cluster 1 fewer students took a quiz (37%), compared to clusters 2 (94%) and 3 (92%), which indicates that these students dropped out during the course.

Second, the clusters of student behavior were analyzed for all 11 courses individually. In one course, students showed almost no activity, which resulted in less meaningful clusters. In all other courses a cluster with students who show almost no activity (similar to cluster 1 in Figure 1) and a cluster with students who show high activity during the whole course (similar to cluster 2 in Figure 1) was found. Additionally, some courses showed clusters with different patterns, such as clusters where students show high activity during the whole course, but activity drops considerably in the last two weeks of the course (3 courses). A series of one-way ANOVAs were used to determine the differences in performance

between the clusters. In 7 of the 11 courses significant differences were found in average quiz grade between the clusters. Tukey post-hoc tests were used to determine which specific clusters differed.

Two courses showed somewhat different clusters of student behavior. Four clusters were extracted in both courses (Figure 2). In the Education course (top), cluster 3 shows a different pattern compared to other courses: the students show little activity in the beginning of the course, but there is an increase of activity at the end of the course. This might indicate that these students are trying to catch up with the courses. There are no significant differences found between the clusters and student performance ($F(3,56) = 0.57, p = .64$). Thus, in this course there is little effect on final grade if you show no activity, mostly activity in the beginning of the course, activity in the end of the course, or activity throughout the entire course. However, this could also be due to the small sample sizes of the clusters.

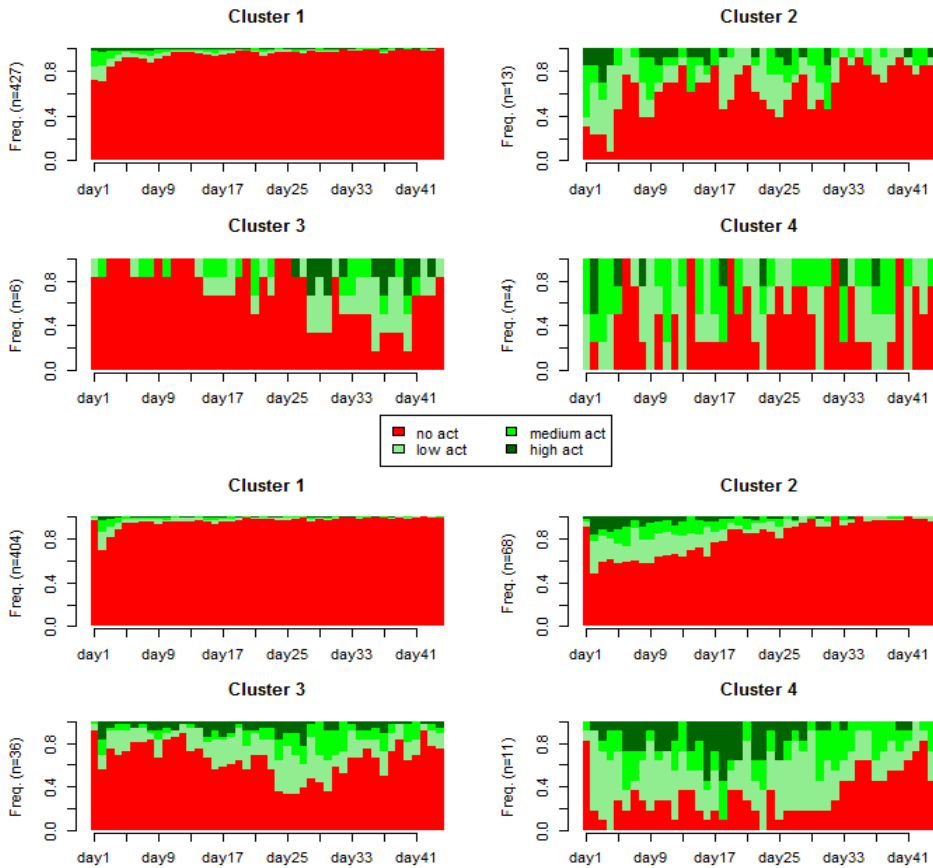


Figure 2. State distribution plots learning behavior for two Canvas courses: a) Education, b) Social Sciences.

In the Social Sciences course (bottom), cluster 3 shows a different pattern compared to other courses: the students show higher activity in the middle of the course. A significant difference is found between the clusters and student performance ($F(3,142) = 17, p < .001$). Students who show almost no activity (cluster 1; $M = 42, S.D. = 39$) have significantly lower grades than all other students. Students who show more activity in the middle of the course (cluster 3; $M = 90, S.D. = 21$) have significantly higher grades than students in cluster 2 ($M = 70, S.D. = 34$). Interestingly, no difference is found in student performance between students who show high activity during the whole course (cluster 4; $M = 88, S.D. = 16$) and those in clusters 2 and 3.

5. Discussion and Conclusion

We analyzed the patterns in sequences of learning activities and the relationship between these patterns and student performance in 11 open online courses. The results based on all courses combined showed three clusters of learning activities: students who showed almost no activity, students who showed activity mostly in the beginning of the course, and students who showed regular activity during the course. These patterns are in line with the patterns found by Cobo et al. (2010) in a course forum. However, when looking at the courses separately, more interesting patterns emerge. For instance, some courses show patterns where students are active mostly in the middle or in the last part of the course. Thus, student behavior seems to be more complex than could be identified in multiple courses combined.

The different patterns within a course and across courses can be explained by the theory of self-regulated learning. According to this theory, learning is influenced by task conditions, such as time, course design, social context, and cognitive conditions such as beliefs, motivation, and knowledge (Winne & Hadwin, 1998). Indeed, several cognitive conditions are identified to influence students' persistence in online learning (Hart, 2012), and hence might result in different activity patterns. However, no additional data is available about the students to verify this in the current context. The different patterns across courses may partly be explained by the smaller sample sizes in individual courses, but differences in task conditions could also have played a role. Lockyer et al. (2013) argued that patterns of learning activities are influenced by course design. For instance, students may show more activity in weeks with a compulsory quiz compared to weeks where no (new) course content is provided. Unfortunately, the current dataset also did not include information on course design and context. Therefore, future work should include qualitative as well as quantitative data about cognitive and task conditions to examine why the different patterns were found.

The patterns of learning activities were found related to student performance. Students who show regular activity throughout the entire course receive higher grades compared to

students who show almost no or limited activity. This corroborates with studies analyzing frequencies of activities, which generally found that more activity results in higher performance (e.g., Zacharis, 2015). Yet, these findings do not always hold when we look at individual courses. In some courses, no differences were found in performance between more active and less active clusters. This is in line with work that showed that the effect of frequencies of activities on student performance differs across courses (Conijn et al., 2016; Gašević et al., 2016).

For educational practice, the current findings imply that sequences of learning activities can provide additional insights next to frequencies of activities. This can be especially useful for improvements in learning and teaching, for example, to guide temporal course design or the design of interventions. Yet, future empirical studies are needed to verify whether the proposed improvement indeed leads to different patterns and increased student performance.

References

- Bergner, Y., Shu, Z. & von Davier, A. (2014). Visualization and confirmatory clustering of sequence data from a simulation-based assessment task. In *Educational Data Mining 2014*.
- Canvas Network (2016). “Canvas Network Courses, Activities, and Users (4/2014 - 9/2015) Restricted Dataset”, Harvard Dataverse, V1
- Cobo, G., García-Solórzano, D., Santamaría, E., Morán, J. A., Melenchón, J., & Monzo, C. (2010). Modeling students’ activity in online discussion forums: a strategy based on time series and agglomerative hierarchical clustering. In *Educational Data Mining 2011*.
- Conijn, R., Snijders, C., Kleingeld, A., & Matzat, U. (2016). Predicting student performance from LMS data: A comparison of 17 blended courses using Moodle LMS. *IEEE Transactions on Learning Technologies*.
- Desmarais, M., & Lemieux, F. (2013). Clustering and visualizing study state sequences. In *Educational Data Mining 2013*.
- Gabardinho, A., Ritschard, G., Mueller, N. S., & Studer, M. (2011). Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software*, 40(4), 1–37.
- Gašević, D., Dawson, S., Rogers, T., & Gasevic, D. (2016). Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *The Internet and Higher Education*, 28, 68–84.
- Hadwin, A. F., Nesbit, J. C., Jamieson-Noel, D., Code, J., & Winne, P. H. (2007). Examining trace data to explore self-regulated learning. *Metacognition and Learning*, 2, 107–124.
- Hart, C. (2012). Factors associated with student persistence in an online program of study: A review of the literature. *Journal of Interactive Online Learning*, 11(1), 19-42.

- Kinnebrew, J., Mack, D., & Biswas, G. (2013). Mining temporally-interesting learning behavior patterns. In *Educational Data Mining 2013*.
- Klingler, S., Käser, T., Solenthaler, B., & Gross, M. (2016). Temporally coherent clustering of student data. In *Proceedings of the 9th International Conference on Educational Data Mining, International Educational Data Mining Society*, 102–109.
- Lockyer, L., Heathcote, E., & Dawson, S. (2013). Informing pedagogical action: Aligning learning analytics with learning design. *American Behavioral Scientist*.
- Perera, D., Kay, J., Koprinska, I., Yacef, K., & Zaïane, O. R. (2009). Clustering and sequential pattern mining of online collaborative learning data. *IEEE Transactions on Knowledge and Data Engineering*, 21(6), 759–772.
- Romero, C., Espejo, P. G., Zafra, A., Romero, J. R., & Ventura, S. (2013). Web usage mining for predicting final marks of students that use Moodle courses. *Computer Applications in Engineering Education*, 21(1), 135–146.
- Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics*, 40(6), 601–618.
- Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. *Metacognition in Educational Theory and Practice*, 93, 27–30.
- Zacharis, N. Z. (2015). A multivariate approach to predicting student outcomes in web-enabled blended learning courses. *The Internet and Higher Education*, 27, 44–53.