Clustering and Visualizing Study State Sequences

Michel C. Desmarais
Polytechnique Montréal
michel.desmarais@polymtl.ca

François Lemieux
Polytechnique Montréal
francois.lemieux@polymtl.ca

ABSTRACT

This paper investigates means to visualize and classify patterns of study of a college math learning environment. We gathered logs of learner interactions with a drill and practice learning environment in college mathematics. Detailed logs of student usage was gathered for four months. Student activity sessions are extracted from the logs and clustered in three categories. Visualization of clusters allows a clear and intuitive interpretation of the activities within the clustered sessions. The three clusters are further used to visualize the global activity of the 69 participating students, which would otherwise be difficult to grasp without such means to extract patterns of use. The results reveal highly distinct patterns. In particular, they reveal an unexpected and substantial amount of navigation through exercises and notes without students actually trying the exercises themselves. This combination of clustering and visualization can prove useful to learning environments designers who need to better understand how their application software are used in practice by learners.

1. INTRODUCTION

The human eye is a powerful means to extract patterns from data, given the proper visualization tools. We borrow visualization techniques from social sciences to display state sequence diagrams and demonstrate their use in Educational Data Mining (EDM).

We combine the visualization tools with clustering techniques to better understand the patterns of use of a learning environment. Detailed user sessions of interaction with a drill and practice environment for college math are encoded as sequences of activities.

2. VISUALIZATION OF TEMPORAL SEQUENCES

Visualization of student interactions is one of the core topics of educational data mining and a few studies have introduced innovative visualization tools in the last decade [11; 10; 9].

This paper focuses on the visualization of temporal sequences of student activity. This type of data can be represented in two different forms:

(1) *Event sequences.* A given event occurs at a specific time. Events can be considered as having no duration, and the focus is more on the transition from one event to another. A majority of studies in EDM have studied such transition data as we see below.

(2) *State sequences.* Each student is engaged in a given activity, or state, for a specific time duration. For example, a student can be consulting notes, involved in problem solving, reading a scaffolded hint, etc.

Student state changes are triggered by events, as a transition from one state to another occurs after some event. Therefore, the two concepts are tightly related. But the type of analysis differ whether we focus on state sequences or the event sequences. Event sequences will often be represented as graphs, emphasizing the path between events and transition frequencies, whereas state sequences are often represented as a flow of states (activities) on a time line. The emphasis for state sequences is on the types and the duration of activities over a given period, instead of transition between states.

2.1 Event Sequences

Event sequences have been studied by a few researchers who were aiming to find patterns of student learning. Beal and Cohen have surveyed a number of these techniques [2]. In one of their study, they used Hidden Markov Models (HMM) to predict sequences of answer types of a math tutor. They showed that modeling the level of student engagement as a hidden factor of the HMM helped improve predictions [3]. Jeong et al. also showed the use of HMM to characterize student behavior in a “learning by teaching” paradigm [7].

Hadwin et al. [6] have also investigated activity event sequences to find patterns of study behavior. They used transition graph and graph theoretic statistics to characterize the student study patterns.

Köck and Paramythis [8] did an extensive study of event sequence analysis over the ANDES Tutor data [12]. They combined k-means clustering techniques with Discrete Markov Models to successfully extract and identify student problem-solving styles.

2.2 State Sequences

Instead of emphasizing the transition between states, temporal state sequences of student activity emphasize the time line perspective and the duration of activities. This type of representation has been used in sociology [1], but has not received much attention in Educational Data Mining.

The time line perspective representation is well suited for visualization of activities as a function of time. Each sequence of activity is represented as a single horizontal bar, and each activity is displayed by a segment on the bar with a given color. We will refer to this as *state sequences*. This type of visualization is shown in figure 1. Both types of diagrams will be explained in more details later. In the current study, we use the TraMineR package [5] available on the R statistical analysis platform.

3. LEARNER ACTIVITY SEQUENCES AND THEIR CLUSTERING

A prerequisite to effective visualization is that the data must be arranged in a meaningful and organized manner such that the patterns emerge naturally to the human eye.

Our solution to effective visualization of usage patterns by a large number of student is two staged: (1) we use a clustering technique to extract the main patterns and use state sequence visualization to characterize them intuitively (section 3.5), and (2) we display student usage as a function of these patterns (section 3.6). Let us first explain sequence data, the clustering algorithm and the results obtained, and finally show the global student picture.

3.1 The Drill and Practice Learning Environment

The web base drill and practice learning environment records detailed logs of user interactions with the application in the form of events that are processed to visualize activities.

The application was made available for four months in the summer of 2012 to newly enrolled university students who wanted to refresh, or enhance their knowledge of prerequisite mathematical concepts for all engineering programs at Polytechnique Montreal. The decision to do the exercises were entirely left at the discretion of the students and no marks or bonus were given for those who used the application. Furthermore, the exercises are presented with a button next to each of them, that, once clicked, immediately shows the right answer and lets the student assess for himself or herself if his/her answer is right. If deemed right, the exercise is marked as completed in the Results section of the application. The student can thereby assess his progress in terms of the proportion of exercises completed.

A total of 1030 exercises are available and they span over 10 mathematics topics such as basic algebra, logarithms and exponentials, trigonometry, calculus, and linear algebra. The notes section represents the equivalent of about 150 pages of a textbook.

3.2 Event and Activity Data

Detailed log data, such as answers to exercises, clicking on a hyperlink, and even scrolling with the mouse is logged as events with a time stamp. This allows to record almost as much as can be recorded on a web browser to assess the level of activity of a user.

These events can be considered as having no time duration, and need to be transformed into sequences of student states, which involves some pre-processing. This process involves the creation of pause events if no event occurs for more than 5 minutes, such as scrolling or navigation. An exception to this rule is if the following event is an answer to an exercise, since problem solving during exercises can last longer than 5 minutes. It also involves pre-processing to ensure that answers to exercises will not go unseen and allow for the distinction between active answer periods, and time spent on problem solving.

Once these adjustments are made to the event sequence, the next step consists in projecting the sequence of events over a time line. A time line represents a series of equal segments of time for which a state is given. If the granularity of the time line is 15 sec., for example, then the state at each segment of the time line is set to the most recent event in this 15 sec. interval. This may result in events that never get displayed if the time segment is longer then the time between events.

3.3 From Events to Activity States Sequences

As explained above, the projection of the events sequence to a state sequence is based on labelling a time segment based on the last event of the current sequence, or the last event of previous sequences if none occurred during the time segment.

A student sequence of activities is broken down per session. A session is defined as all activities contiguous in time that are no more than 1.5 hour apart. A time difference between events of over 1.5 hour creates a new sequence (session).

Seven types of activities are derived from the log of user events:

1. **Answer Ex.**: A click over the answer button ("Answer ex." event) occurred during the time step and is represented as the activity of that time step.
2. **Nav. Exerc.**: Student is browsing through the exercises but has not answered an exercise during the time segment.
3. **Nav. Notes**: Student is browsing through sections of the notes module.
4. **Pause**: No event occurred in the last 5 minutes. Pauses can last up to 1.5 hour (the maximum time after which a new session is created).
5. **Prblm. solv.**: Last event was an answer to an exercise, but no event was recorded during the time step and therefore we assume the student is in problem solving mode over the exercises shown on the page.
6. **Result**: Browsing a page that summarizes statistics on the exercises completed and the number of remaining exercises per main section.
7. **Start**: Activity on the login page.

For the purpose of this study, we ignore sessions that are shorter than 5 minutes. This leaves a total of 454 sessions of activity sequences by 69 students. Mean and median session duration are respectively 42 and 20 minutes with a minimum duration of 5 minutes and a maximum duration of 6.3 hours. Mean and median number of sessions per student are respectively 2 and 6.5, with a minimum of 1 and a maximum of 93 sessions. 24 students completed no exercise whereas one student completed all 1030 of them. Median number of attempts to exercises is 12 and mean is 174. An exercise can be attempted more than once if the student answers that he did not get the answer right. In this case, the exercise is shown with a "validate my answer" button. If the student answers his response is correct, the solution is displayed in place of the button.

3.4 Clustering Algorithm

As mentioned, there are 454 sessions by 69 students. To build an synthetic view of their usage patterns, we first extract types of state sequences from the data with a clustering algorithm.

Based on the well known Levenshtein distance, an agglomerative (bottom up) hierarchical method is used to aggregate the most similar sequences (Ward method of the R cluster package [4]). In short, the algorithm consists in pairing the most similar individual sequences, and in further pairing groups of sequences, ensuring that the mean distance between two clusters is minimal at every level.

We chose to create 3 clusters based on exploratory visualization of the different results.

3.5 Clusters of Activity Sequences

The 3 clusters of state sequences are shown in figure 1. Each horizontal line in a graph shows an activity state sequence (session). The time segments are 15 sec. each. The figures show 240 segments, which corresponds to a total of one hour. Longer sessions are truncated. There are 7 types of activities as described in section 3.2. For the sake of visibility, we have randomly sampled only 30 of the sequences of each type. The actual numbers of sessions per type are shown below.

The clustered sequences obtained can be characterized as follow:

- **Type 1 (N=135)**: Exploratory behavior. The students engage in a mixture of browsing through exercises and notes.
- **Type 2 (N=196)**: Short sessions comprising a variety of behaviors. Shorter sessions are aggregated due to the fact that the Levenshtein distance penalizes deletion and addition of sequence elements.
- **Type 3 (N=123)**: Exercise intensive sessions.

3.6 Activities per Student

Figure 2's top diagram shows the proportion of session types for each of the 69 students. The students are ordered according to the time they spent with the application. The y axis corresponds to the three session types. Darker cells indicate the dominant session(s) for that student.

The bottom diagram is a frequency plot of the number of sessions per student.

We can see that the more engaged students, defined as those who spend the most time with the applications, have sessions of all types (students 60 to 69). However, we notice that students 40 to 55, who have a usage time over the median, are not engaged in the same manner as the ones who had a predominance of type 3 sessions. They do not engage much in doing exercises, if at all, but do spend a substantial amount of time browsing through the application content.

Also noteworthy is that the student with very short sessions (type 2) have diverse patterns of activities. Most combine browsing and answer exercises, and browsing through notes, in a relatively short period of time compared to the others. They are essentially exploring the application. Unsurprisingly, almost all of the students who have only a single session of interaction with the learning environment have this type of behavior.

4. DISCUSSION

This paper introduces a technique to visualize student learning activities with a self-regulated drill and practice environment, and reports on an experiment that combines the visualization method with clustering and classification techniques to obtain a global view of student activities.

The clustering clearly reveals that very distinct study patterns emerge per session. Without the visualization of clusters as state diagrams, cluster interpretation would remain a difficult task. This is particularly the case because a single student often adopts different patterns of study sessions. Therefore, session study patterns do not necessarily discriminate between student themselves. However, we do notice a predominance of certain session types on a per student basis, in good part due to the correlation of session length with session patterns.

How can such visualization and clustering method be of use to the design of learning environments, or to the teachers? It might help them to better understand how students use the learning environment and, in return, make adjustments to the features of the learning environment to better suit the actual patterns of usage.

At least in our case, this study revealed that, although the most engaged students did use the application as intended, with a predominance of Type 3 sessions, many students did not use it as intended, or even as expected. Both authors were involved in the design, and we had expected much more back and forth between the study notes and exercises. Except for a few students, this did not happen very much. Instead, many sessions consisted of relatively long pauses with browsing periods through the notes and exercises, lasting sometimes over an hour without actually doing practice exercise. We intend to conduct interviews in later studies to better understand this behavior, but this current study helps us reveal the patterns to investigate further.

Acknowledgements

This study was funded by MATI (Maison des technologies de formation et d'apprentissage Roland-Giguère) under the project "Services d'apprentissage personnalisés".

5. REFERENCES

[1] A. Abbott and A. Tsay. Sequence analysis and optimal matching methods in sociology review and prospect. *Sociological Methods & Research*, 29(1):3–33, 2000.

[2] C. Beal and P. Cohen. Temporal data mining for educational applications. *PRICAI 2008: Trends in Artificial Intelligence*, pages 66–77, 2008.

[3] C. Beal, S. Mitra, and P. R. Cohen. Modeling learning patterns of students with a tutoring system using Hidden Markov Models. *Artificial intelligence in education: Building technology rich learning contexts that work*, 158:238, 2007.

[4] B. Everitt, S. Landau, and M. Leese. Cluster analysis. 4th. *Arnold, London*, 2001.

[5] A. Gabadinho, G. Ritschard, M. Studer, and N. S. Müller. Mining sequence data in R with the TraMineR package: A users guide for version 1.2. *Geneva: University of Geneva*, 2009.

[6] A. Hadwin, J. Nesbit, D. Jamieson-Noel, J. Code, and P. Winne. Examining trace data to explore self-regulated learning. *Metacognition and Learning*, 2:107–124, 2007.

[7] H. Jeong, A. Gupta, R. Roscoe, J. Wagster, G. Biswas, and D. Schwartz. Using hidden Markov models to characterize student behaviors in learning-by-teaching environments. In *Intelligent Tutoring Systems*, pages 614–625. Springer, 2008.

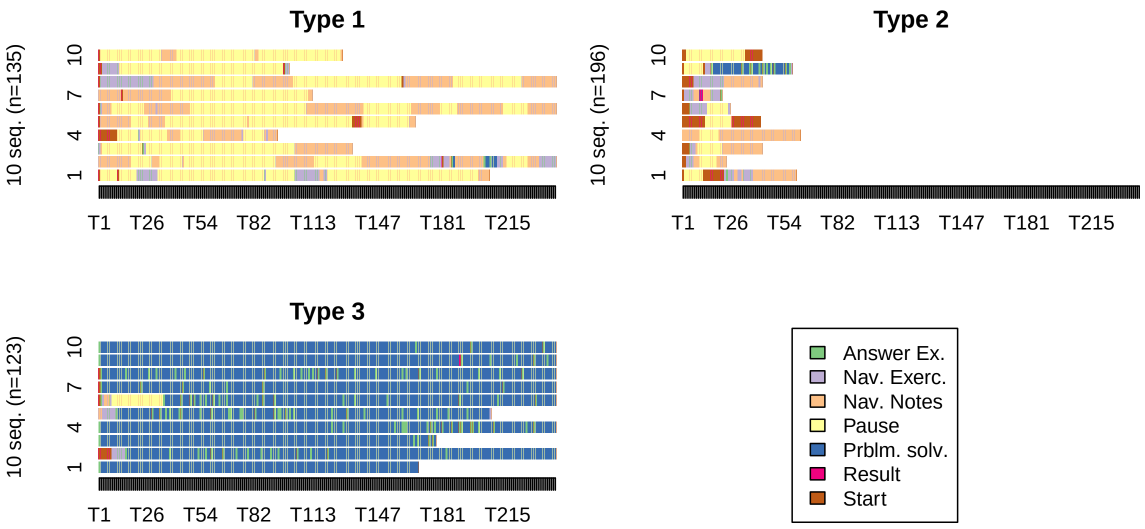

Figure 1: Random samples of 10 sessions for each type of cluster from various students. Type 1 corresponds to browsing through notes and exercises with frequent pauses and very little problem solving. Type 2 corresponds to short sessions of various behaviour. Type 3 are sessions focused on problem solving and answers to exercises.

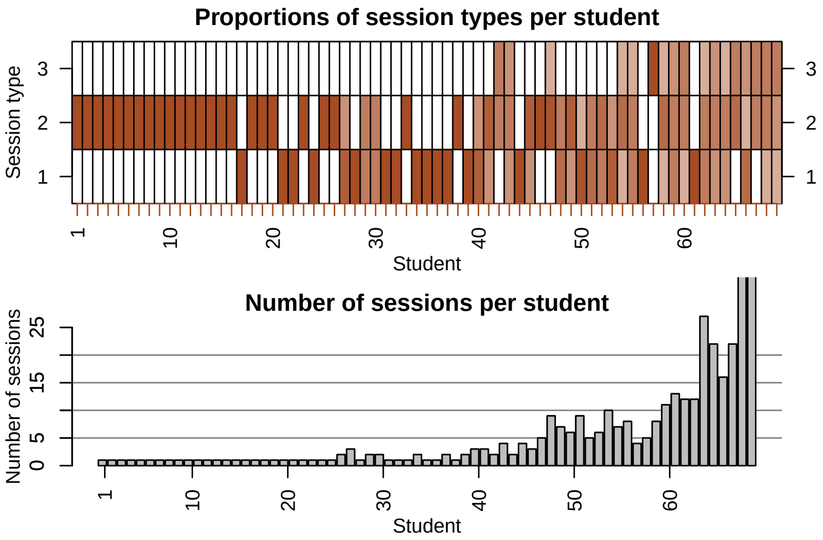

Figure 2: Type of sessions per student. Students are on the x-axis and ordered from shortest time of use (302 sec.) to the longest time (120 hrs.)—median time is 1.2 hr. and mean time is 6.1 hrs.

[8] M. Köck and A. Paramythis. Activity sequence modelling and dynamic clustering for personalized e-learning. *User Modeling and User-Adapted Interaction*, 21:51–97, 2011. 10.1007/s11257-010-9087-z.

[9] A. Merceron and K. Yacef. Tada-ed for educational data mining. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, 7(1):267–287, 2005.

[10] J. Mostow, J. Beck, H. Cen, A. Cuneo, E. Gouvea, and C. Heiner. An educational data mining tool to browse tutor-student interactions: Time will tell. In *Proceedings of the Workshop on Educational Data Mining, National Conference on Artificial Intelligence*, pages 15–22, 2005.

[11] C. Romero, S. Gutiérrez, M. Freire, and S. Ventura. Mining and visualizing visited trails in web-based educational systems. *Educational Data Mining 2008*, page 182, 2008.

[12] K. Vanlehn, C. Lynch, K. Schulze, J. A. Shapiro, R. Shelby, L. Taylor, D. Treacy, A. Weinstein, and M. Wintersgill. The ANDES physics tutoring system: Lessons learned. *Int. J. Artif. Intell. Ed.*, 15(3):147–204, 2005.