Tracking Behavioral Patterns among Students in an Online Educational System

Stephan Lorenzen
University of Copenhagen
lorenzen@di.ku.dk

Niklas Hjuler
University of Copenhagen
hjuler@di.ku.dk

Stephen Alstrup
University of Copenhagen
s.alstrup@di.ku.dk

ABSTRACT

Analysis of log data generated by online educational systems is an essential task to better the educational systems and increase our understanding of how students learn. In this study we investigate previously unseen data from Clio Online, the largest provider of digital learning content for primary schools in Denmark. We consider data for 14,810 students with 3 million sessions in the period 2015-2017. We analyze student activity in periods of one week. By using non-negative matrix factorization techniques, we obtain soft clusterings, revealing dependencies among time of day, subject, activity type, activity complexity (measured by Bloom's taxonomy), and performance. Furthermore, our method allows for tracking behavioral changes of individual students over time, as well as general behavioral changes in the educational system. Based on the results, we give suggestions for behavioral changes, in order to optimize the learning experience and improve performance.

Keywords

Student clustering, Non-negative matrix factorization, Educational Systems

1. INTRODUCTION + RELATED WORK

How students behave in educational systems is an important topic in educational data mining. Knowledge of this behavior in an educational system can help us understand how students learn, and help guide the development for optimal learning based on actual use. This behaviour can be understood both through an explicit study [5], or as in this paper through the automatically generated log data of the system.

The analysis of log data is usually done as an unsupervised clustering of students [2, 3, 4, 7]. A popular approach is to extract action sequences and transform them into an aggregated representation using Markov models [4, 7]. The Markov chains can then be clustered by different methods.

Klingler et al. did student modeling with the use of explicit Markov chains and the clustering with different distance measures defined on the Markov chains [7]. Hansen et al. assumed the actions sequences to be generated by a mixture of Markov chains and used an heuristic algorithm to find the generating Markov chains [4]. Gelman et al. used non-negative matrix factorization to find clusters for different measures of activity aggregated in weekly periods during a MOOC course. These clusters are then matched from week to week by cosine similarity.

Our work is similar to Gelman et al. [3] in that we also use *Non-negative Matrix Factorization* (NMF) to make a soft clustering at the student level in a given time period, however our clustering is only made once, and we are looking at primary school data over a vastly longer period of time, (2 years compared to 14 weeks).

Our soft clustering by non-negative matrix factorization is based on log data from Clio Online.[1] Clio Online is the largest provider of digital learning for all subjects in the Danish primary school (except mathematics), having 90% of all primary schools in Denmark as customers.

Using NMF, we assume that the set of features chosen can be represented by a set of fewer underlying behaviors. These underlying behaviours would each be represented by a cluster in the non-negative matrix factorization. Each student will then get a number for each cluster in each time period representing how much of that underlying behavior he has shown in the given time period. Non-negativity gives the behaviors an additive structure, which is more natural than showing a negative amount of a given behavior. We reason that the soft clustering will show both the behaviors of individual students, as well as how the behaviors change over time, both individually and on a system-wide level.

In this paper, we will consider two main questions: a) how does student activity in the system affect performance, and b) how does student activity distribute between different levels of Bloom's taxonomy in different subjects. Both questions are important in regards to optimizing learning; the first in relation to performance, the latter in relation to utilization of all taxonomy levels.

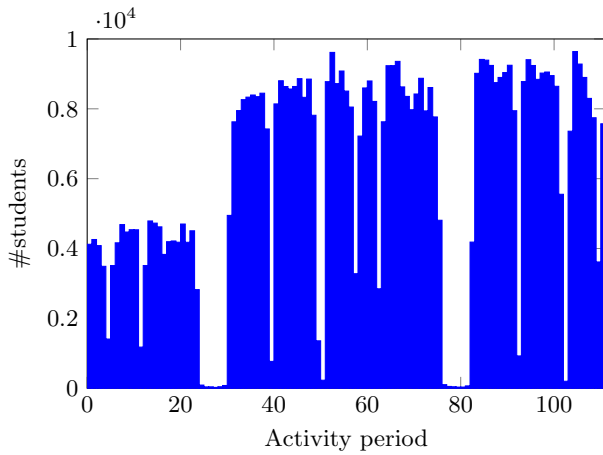[1]This data is proprietary and not publicly available.

Figure 1: Number of students active in each period. Note that period 0 starts on 2015-01-08, while period 111 ends on 2017-03-01. The drops in activity occur due to vacation in Danish primary school, with the two large drops around periods 25 and 79 being due to the summer vacation.

2. EXPERIMENTAL SETUP

This section describes our experimental setup and methods. We start by describing our data and how it is preprocessed, and then move on to describing our clustering method.

2.1 Data Preprocessing

As mentioned, we consider log data generated in the Danish online educational system Clio Online. The system is used in Danish primary schools and contains learning objects across all Danish subjects (except mathematics), for instance texts, videos, sound clips and exercises. Furthermore, the system includes a large number of quizzes, used for evaluating students. Students may use the system for self study, but they may also be assigned homework by their teacher. Our data covers 14,810 students.

The raw data consists of logs detailing page accesses for individual students in the system. For quizzes, the final score (between 0 and 1) and total time spent for the quiz is also available. In our preprocessing, we combine these log entries to *sessions*. Two consecutive entries are considered in the same session, if they have the same subject, and their timestamps differ by less than some threshold. For our study, we choose this threshold to be 600 seconds, based on recommendations from Clio Online, who have a deeper knowledge of the content and flow of the system (e.g. expected time per page). Furthermore, quizzes are considered separate sessions. A total of 3 million sessions is obtained in this way.

With the sessions defined, we consider student activity in *activity periods*, with a length of one week. The data spans a total of 112 activity periods, starting January 2015 and ending in March 2017. For each activity period, we add an entry for a student, if the student is active (accesses the system) within that period. The entry for the given student contains all sessions for that student, which starts within the activity period. We end up with approximately 677,000 student entries across the 112 periods. Figure 1 shows the

active number of students in each period. Note the drop in active students around periods 25 and 79; these drops in activity occur due to summer vacation.

The final step of data preprocessing is the feature extraction. For each activity period, a set of activity/performance related features are extracted. The features are chosen so as to answer the questions posed in the previous section. A complete overview of all features considered in our experiments is given in Table 1, including the maximum, mean and variance across all active students in all periods. Not all features are used for each experiment, see section 3.

All features are aggregates over the activity period. Below follows a detailed description:

- f_1 describes the activity during the period of day, where Danish students are normally in school, while f_2 describes the activity during non-school hours.
- f_3, f_4 and f_5 describe time spent doing exercises, reading texts and taking quizzes respectively.
- f_6, f_7 and f_8 describe time spent working with different topics: languages (Danish, English, German), societal (social studies, history, etc.) and science (physics, biology, etc.), respectively.
- f_9 is the average session length during the activity period.
- f_{10} is the average quiz score; this feature may be missing, if a student takes no quizzes during an activity period, but our analysis methods can handle this, see section 2.2.
- f_{11}, f_{12}, f_{13} and f_{14} describe the time spent doing exercises of different complexity, measured by their level in Bloom's taxonomy. We regroup the levels of Bloom's taxonomy into 4 levels:

  f_{11} **Remember/Understand**: Exercises involving reading and describing, e.g. "Read a map".
  f_{12} **Apply**: Exercises involving application of previously learned concepts, e.g. "Practice adjectives".
  f_{13} **Analyze/Evaluate**: Exercises involving discussion, analysis and experimenting, e.g. "Work with the poem", "Analyze the game".
  f_{14} **Create**: Exercises involving creation of a product, e.g. "Create a cartoon", "Write a story".

Having extracted m features for each student in each period, we construct the matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$, where each of the n rows consists of the feature vector for an active student in a given activity period. Thus each student occurs several times in $\mathbf{X}$; once for each period, where they are active.

2.2 Soft Clustering using Non-negative Matrix Factorization

We will utilize non-negative matrix factorization for our soft clustering. The use of NMF as a soft clustering technique has become popular in recent times [10], with applications within several fields, such as clustering of images and documents [8, 13]. NMF has also seen success in the educational data mining community, for clustering tasks, as well as other tasks such as performance prediction [3, 12].

| i	f_i	Max	Mean	Variance
1	Hours between 8AM and 4PM	31.85	0.940	0.862
2	Hours before 8AM and after 4PM	71.84	0.174	0.283
3	Hours doing exercises	3.61	0.048	0.019
4	Hours reading texts	7.73	0.344	0.148
5	Hours taking quizzes	23.76	0.231	0.297
6	Hours working with language subjects	58.28	0.531	0.693
7	Hours working with societal subjects	45.96	0.294	0.285
8	Hours working with science subjects	103.69	0.277	0.326
9	Average session length in hours	7.91	0.268	0.027
10	Average quiz score (in $[0, 1]$)	1.00	0.733	0.034
11	Hours working with Bloom level 1	2.83	0.016	0.006
12	Hours working with Bloom level 2	1.64	0.008	0.002
13	Hours working with Bloom level 3	1.51	0.014	0.003
14	Hours working with Bloom level 4	2.04	0.009	0.003

Table 1: Overview of features.
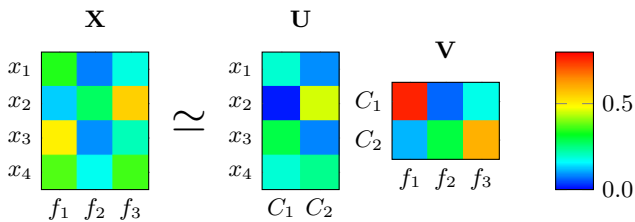
Figure 2: The soft clustering given by NMF.

NMF is a dimensionality reduction method, in which we are given a non-negative matrix $\mathbf{X} \in \mathbb{R}_+^{n \times m}$ and $k \in \mathbb{N}$, and wish to determine $\mathbf{U} \in \mathbb{R}_+^{n \times k}, \mathbf{V} \in \mathbb{R}_+^{k \times m}$, such that $\mathbf{X} \simeq \mathbf{UV}$. More specifically, we search for $\mathbf{U}$ and $\mathbf{V}$, such that the error $||\mathbf{X} - \mathbf{UV}||_F$ is minimized, where $|| \cdot ||_F$ is the Frobenious norm. For our analysis, we need to be able to handle missing values in $\mathbf{X}$. In this case the NMF problem is reformulated as the *weighted non-negative matrix factorization*, in which we are also given a binary weight matrix $\mathbf{W} \in \{0, 1\}^{n \times m}$, where a 0 indicates missing data. Now, we wish to find $\mathbf{U}, \mathbf{V}$ such that $||\mathbf{W} \odot (\mathbf{X} - \mathbf{UV})||_F$ is minimized[2].

$\mathbf{U}$ and $\mathbf{V}$ admits a soft k-clustering as shown in Figure 2; $\mathbf{V}$ describes the importance of each feature for each cluster (for instance, f_1 has high importance in C_1), while $\mathbf{U}$ describes the membership of each data point to the different clusters (for instance, x_3 is mostly in C_1, while x_4 is in both clusters).

Note, that for NMF, we have $\mathbf{X} \simeq \mathbf{UV} = \mathbf{UIV} = \mathbf{UA}^{-1}\mathbf{AV}$, where $\mathbf{I}$ is the $k \times k$ identity matrix and $\mathbf{A}$ is a $k \times k$ invertible matrix. This means that we may rescale $\mathbf{U}$ and $\mathbf{V}$ by this matrix, $\mathbf{A}$, and its inverse. In our analysis, we use this to rescale $\mathbf{V}$, such that all rows of $\mathbf{V}$ (the clusters) sum to one, thus making the clusters comparable, and membership of the different clusters easier interpretable.

There exist several algorithms for obtaining the non-negative matrix factorization of $\mathbf{X}$, for instance basic gradient de-

scent, multiplicative update rules and alternating least squares; [1] gives a good overview in the non-weighted setting. Several of these algorithms have been adapted for the WNMF case, while approaches based on *expectation maximization* have also been proposed, see [6]. For our analysis, we will use the weighted version of the multiplicative update method, proposed by Lee and Seung [9].

The NMF algorithm given in [9], adopted to WNMF [6], is as follows:

1. Initialize $\mathbf{U}$ and $\mathbf{V}$.
2. Repeatedly update $\mathbf{U}$ and $\mathbf{V}$ by the following rules:

$$\mathbf{U} \leftarrow \mathbf{U} \odot \frac{(\mathbf{W} \odot \mathbf{X}) \mathbf{V}^T}{(\mathbf{W} \odot (\mathbf{UV})) \mathbf{V}^T}$$

$$\mathbf{V} \leftarrow \mathbf{V} \odot \frac{\mathbf{U}^T (\mathbf{W} \odot \mathbf{X})}{\mathbf{U}^T (\mathbf{W} \odot (\mathbf{UV}))}$$

where division is done element-wise.

The literature explores several ways of initializing $\mathbf{U}$ and $\mathbf{V}$; in our case, we will simply use random initialization. The alternating optimization steps are applied until the decrease in error reaches below a set threshold. Finally, Lin has noted that the procedure described above may not converge to a stationary point, hence we modify the update rules as proposed by them [11]. Furthermore, since we in our case know all missing values of $\mathbf{X}$ to be bounded by a constant c, we modify the above procedure such that 0-weight values of $\mathbf{UV}$ that deviate above c are penalized, i.e. whenever a value $(\mathbf{UV})_{ij}$ with $\mathbf{W}_{ij} = 0$ gets larger than c, we set $\mathbf{X}_{ij} = c$ and $\mathbf{W}_{ij} = 1$, before the next update step. If $(\mathbf{UV})_{ij}$ decreases below c again, the weight is reset to 0.

It remains to be seen, how we select the number of clusters, k. For each experiment, we construct clusterings with $k = 1, 2, ...,$ and stop when the decrease in error going from k clusters to $k + 1$ clusters is below some threshold, which depends on the initial error. As a consequence clusters will be uncorrelated on a student level, since otherwise we would pick a lower k.
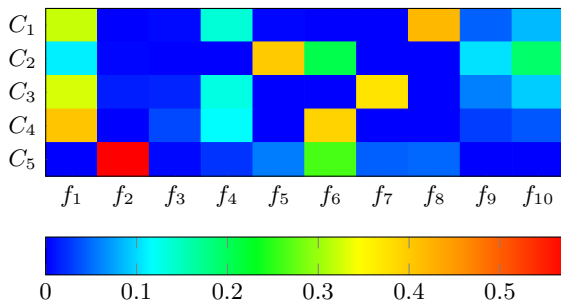
---

[2]$\odot$ denotes the Hadamard product (element-wise multiplication).

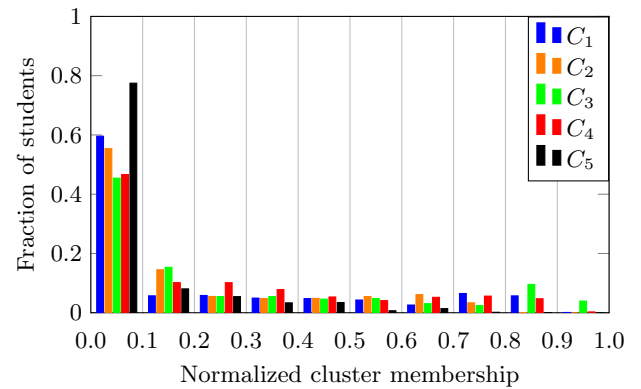**Figure 3: The cluster matrix for the first experiment.**

Figure 4: The distribution of cluster membership for the first experiment.

Figure 5: The average cluster membership in each activity period for the first experiment.

3. EXPERIMENTS AND RESULTS

In this section, we present two different experiments using the setup described above. In the first experiment, we investigate the relation between activity, activity type, subject, time of day, average session length and performance. In the second experiment, we investigate the relation between complexities of exercises and subjects.

3.1 Performance and Optimal Behavior

In the first experiment, we investigate the relation between activity, activity type, subject, time of day, average session length and performance, i.e. we consider features $f_1, ..., f_{10}$. The features are extracted and $k = 5$ is selected, as described in section 2. We run the WNMF algorithm, and obtain the cluster matrix V as shown in Figure 3. From the figure, we can make several observations about the clusters:

C_1 In this cluster, we find students mostly working with the science subjects (f_8). These students seem to work mostly during school hours (f_1). The students also seem to spent a lot of time reading (f_4).

C_2 Students in this cluster spend a lot of time taking quizzes (f_5). They will spend some time during school hours (f_1) and some time working with language subjects (f_6). Furthermore, students in this cluster seem to both have fairly long average session length and high performance (f_9 and f_{10}).

C_3 In cluster C_3, we see students working with societal subjects (f_7). They work during school hours (f_1) and spend time reading texts in the system (f_4).

C_4 This cluster shows a relationship between being active in school (f_1) and spending time in the language subjects (f_6). Students in this cluster also spend time reading texts (f_4) and doing some exercises (f_3).

C_5 The most important feature for C_5 is f_2, i.e. the students in this cluster spend most time using the system during non-school hours. These students spent time in all subjects, but mostly languages (f_6), and they spent time taking quizzes (f_5).

From the clusters, we can see that the impact on performance from different behaviors depends on the subject. From cluster C_2, we see that students working mostly with language subjects gain most performance from spending time taking quizzes and working during school hours, whereas students working mostly with societal (cluster C_3) and science (cluster C_1) subjects gain most from reading texts,
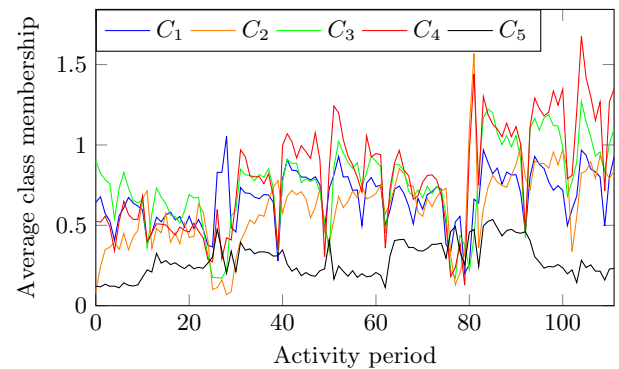
while working mostly during school hours. Note that cluster C_4 indicates that students working with languages may also improve performance by reading texts, but to a lesser degree than students working in other subjects. Finally, C_5 indicates that working mostly from home and primarily taking quizzes, does not improve performance. While C_5 indicates this for all subjects, the high importance of f_4 indicates that this most often occur for students working with languages, confirming the observations from C_2. Finally, it is also worth noticing, that there is a strong relation between performance and average session length (clusters C_1, C_2 and C_3), indicating that students, who perform well, also have longer sessions on average.

From the above discussion, it appears that the behavior in clusters C_4 and C_5 are sub-optimal, when considering performance, while students gain more from being in C_1, C_2 or C_3, i.e. by working during school hours, having longer sessions and taking quizzes (in the case of languages) or reading texts (in the case of societal or science subjects).

Figure 4 describes the distribution of cluster membership across all students and all activity periods , i.e. the columns of the first interval $[0, 0.1)$ gives for each cluster the fraction of students with 0%-10% membership. We see, that we do indeed get a soft clustering, with students often belonging to more than one cluster. Only C_3 seems to be the sin-
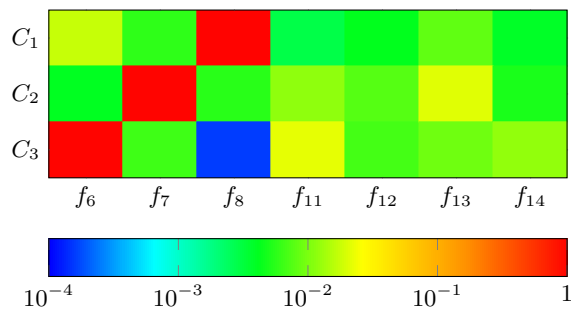
Figure 6: The cluster matrix for the second experiment. Note, that a logarithmic scale is used for this plot.

Figure 7: The distribution of cluster membership for the second experiment.

Figure 8: The average cluster membership in each activity period for the second experiment.

gle dominant cluster of some students. From the figure, we also see that students are typically never exclusively in C_5, which is positive, as the behavior observed in that cluster was not very productive in terms of performance. Other than that, we generally observe that students seem to distribute fairly well between the top four clusters, indicating most time spent during school hours and a varied use of both quizzes and texts across all subjects.

Next, we analyze how the membership of different clusters change over time. Figure 5 plots the average membership for each period, i.e. the average of rows from **U** belonging to the given period. The first observation we make from Figure 5, is that clusters C_1, C_2, C_3 and C_4 appear correlated at the system-wide level. This is due to these clusters being dependent on the general activity in the online system; most of the sudden drops occur at the same time as Danish school vacations, most notably the two larger drops around activity periods 25 and 79 (see Figure 1). C_5 seems to be relatively unaffected by the general activity, but this makes sense, as C_5 contains mostly students, who work outside school hours, and thus a lower membership is expected in that cluster in general, which is also the pattern we see in periods with no vacation.

Looking at the general distribution between the different clusters, C_3 and C_4 seem to be the most dominant, indicating that most students are working with language and societal subjects and reading texts. Cluster C_1 (science subjects) is fairly constant in the non-vacation periods, and C_2 seems to increase starting period 80, indicating that more students spend time taking quizzes. Finally, as mentioned, C_5 is the least active cluster across most periods. One general trend for the top four clusters seem to be an increase in activity during the 112 periods, indicating that students are spending more time in the system on average.

3.2 Subject and Exercise Complexity

In the second experiment we look at the relation between subjects and exercises grouped by Bloom's taxonomy level, i.e. we consider features $f_6, f_7, f_8, f_{11}, f_{12}, f_{13}, f_{14}$

We expect three clusters, one for each of the subject classes, which will tell us how much each Bloom level is used within each subject class. Figure 6 shows the cluster matrix found. From Figure 6, we make the following observations:
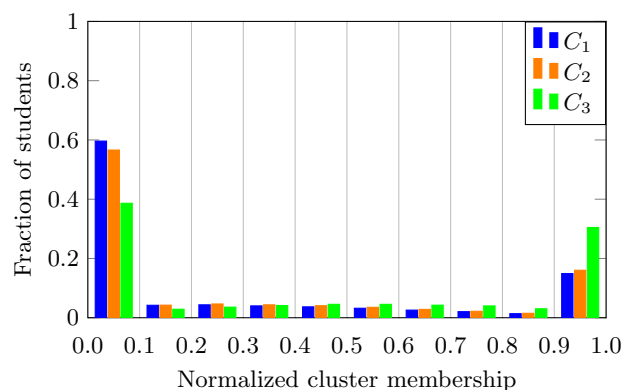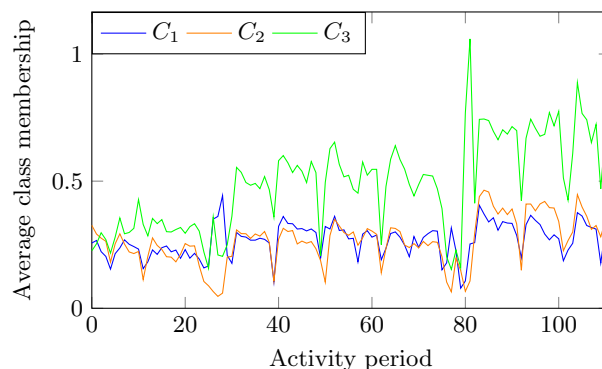
C_1 In the science subjects, only very little of the 3 higher levels are used, and almost none of reading and understanding.

C_2 For societal subjects, students have only little activity in the first 2 levels, a lot in analyzing and evaluating, and very little activity in creation.

C_3 In languages, students have a tendency to read and understand a lot, and then distribute almost evenly on the 3 higher levels.

This implies that if we want to attract students to use an online educational system for languages, focus should be on exercises with Bloom's taxonomy level read and understand. For societal subjects the focus should be on exercises with analyzing and evaluating. For science we see no preference.

From Figure 7, we see that the clustering has many high values which is most likely explained by having a teacher who uses the system exclusively in only one of the subjects, which we can see happens most often for languages.

As we can see in Figure 8 all three clusters share similar curvature, which is partly explained by holidays. Especially the science and societal clusters behave seem highly correlated on a general level. We also see that in all three subjects, the average time spent during a week has gone from 15 minutes,

to 45 minutes for languages and 25 minutes for both societal subjects and sciences. A clear indication that teachers and students in Denmark are using online educational systems more, especially for languages.

4. CONCLUSIONS AND FUTURE WORK

Several points can be taken from our analysis. We have identified three optimal and two sub-optimal behaviors in relation to subject and performance. One notably conclusion is that students using the Clio Online system during non-school hours (at home) do not seem to gain any significant boost to performance. We also saw how taking quizzes seems to increase the performance of students in languages, more so than in other subjects, where reading texts are of more importance. This fits the intuition that skills such as grammar need to be trained, in order to be learned. We inform how exercises are used depending both on their subject and their level in Bloom's taxonomy. And lastly we see that the average amount of time spent in the system is increasing both generally and for the individual students in all subjects, but especially for students working with languages. Furthermore, both experiments show how behaviors can have high correlation on a system-wide level, despite being uncorrelated on the individual student level. While the change of behavior for individual students was not directly analyzed in this paper (due to privacy concerns), our method allows for tracking such individual changes, hopefully helping teachers encourage optimal student behavior, e.g. by recommending training quizzes for students working with languages, or making sure that students are allowed more time to use the system in school.

In our setting, the number of clusters is fixed. It may be interesting to use an adaptive clustering strategy instead, as done in [7], as one might expect clusters to change over time. In the future, it might also be interesting to include other features, that were not available to us at this time, for instance whether a text (or quiz) have been assigned by a teacher, or whether the student reads it by themselves. For this study, we also only had access to a limited amount of data; better and more reliable results might be obtained by including more data.

5. ACKNOWLEDGMENTS

6. REFERENCES

[1] Michael W. Berry, Murray Browne, Amy N. Langville, V. Paul Pauca, and Robert J. Plemmons. Algorithms and Applications for Approximate Nonnegative Matrix Factorization. *Computational Statistics & Data Analysis*, 52(1):155 – 173, 2007.

[2] Louis Faucon, Lukasz Kidzinski, and Pierre Dillenbourg. Semi-Markov model for simulating MOOC students. In *Proceedings of the 9th International Conference on Educational Data Mining (EDM)*, pages 358–363. International Educational Data Mining Society (IEDMS), 2016.

[3] Ben U. Gelman, Matt Revelle, Carlotta Domeniconi, Kalyan Veeramachaneni, and Aditya Johri. Acting the Same Differently: A Cross-Course Comparison of User Behavior in MOOCs. In *Proceedings of the 9th International Conference on Educational Data Mining (EDM)*, pages 376–381. International Educational Data Mining Society (IEDMS), 2016.

[4] Christian Hansen, Casper Hansen, Niklas Hjuler, Stephen Alstrup, and Christina Lioma. Sequence modelling for analysing student interaction with educational systems. In *Proceedings of the 10th International Conference on Educational Data Mining (EDM)*, pages 232–237. International Educational Data Mining Society (IEDMS), 2017.

[5] Stephen Hutt, Caitlin Mills, Shelby White, Patrick J. Donnelly, and Sidney K. D'Mello. The Eyes Have It: Gaze-based Detection of Mind Wandering during Learning with an Intelligent Tutoring System. In *Proceedings of the 9th International Conference on Educational Data Mining (EDM)*, pages 86–93. International Educational Data Mining Society (IEDMS), 2016.

[6] Yong-Deok Kim and Seungjin Choi. Weighted Nonnegative Matrix Factorization. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1541–1544, 2009.

[7] Severin Klingler, Tanja Käser, Barbara Solenthaler, and Markus Gross. Temporally Coherent Clustering of Student Data. In *Proceedings of the 9th International Conference on Educational Data Mining (EDM)*, pages 102–109. International Educational Data Mining Society (IEDMS), 2016.

[8] Cosmin Lazar and Andrei Doncescu. Non Negative Matrix Factorization Clustering Capabilities; Application on Multivariate Image Segmentation. In *Proceedings of the 3rd International Conference on Complex, Intelligent and Software Intensive Systems (CISIS)*, pages 924–929, 2009.

[9] Daniel D. Lee and H. Sebastian Seung. Algorithms for Non-negative Matrix Factorization. In *Advances in Neural Information Processing Systems 13, Papers from Neural Information Processing Systems (NIPS) 2000, Denver, CO, USA*, pages 556–562, 2000.

[10] Tao Li and Chris Ding. Non-negative matrix factorization for clustering: A survey. In *Data Clustering: Algorithms and Applications*, pages 149–176. Chapman & Hall/CRC, January 2013.

[11] Chih-Jen Lin. On the Convergence of Multiplicative Update Algorithms for Non-negative Matrix Factorization. *Trans. Neur. Netw.*, 18(6):1589–1596, 2007.

[12] Stephan Lorenzen, Ninh Pham, and Stephen Alstrup. On Predicting Student Performance Using Low-rank Matrix Factorization Techniques. In *Proceedings of the 16th European Conference on e-Learning (ECEL)*, pages 326–334. Academic Conferences and Publishing International, 2017.

[13] Farial Shahnaz, Michael W. Berry, Victor P. Pauca, and Robert J. Plemmons. Document clustering using nonnegative matrix factorization. *Information Processing & Management*, 42(2):373 – 386, 2006.