Investigating Subpopulation of Students in Digital Textbook Reading Logs by Clustering

Christopher C.Y. Yang[1], Brendan Flanagan[1], Gökhan Akçapınar[1,2], Hiroaki Ogata[1]
Kyoto University[1], Hacettepe University[2]
yang.yuan.57e@st.kyoto-u.ac.jp, flanagan.brendanjohn.4n@kyoto-u.ac.jp,
akcapinar.gokhan.2m@kyoto-u.ac.jp, ogata.hiroaki.3e@kyoto-u.ac.jp

ABSTRACT: The increasing volume of student reading logs from virtual learning environment (VLE) provides opportunities for mining student' engagement pattern in digital textbook reading. In order to mine and measure students' engagement pattern, in this paper, we extract several students' reading interaction variables from the digital textbook as metrics for the measurement of reading engagement. Moreover, in order to explore the presence of subpopulation of students that can be differentiated based on their engagement patterns and academic performances, we cluster students into different groups. Students are clustered based on their reading interactions such as total session of reading, total notes adding, etc. Accordingly, we identify students' engagement patterns from different groups based on the clustering analysis results. Several student subpopulations such as low engagement high academic performances and low engagement low academic performances are identified based on students' reading interaction characteristics by clustering analysis. The obtained results can be used to provide researchers with opportunities to intervene in the specific group of students and also an optimal choice for student grouping.

Keywords: Student engagement pattern, academic performance, clustering, digital textbook

1 INTRODUCTION

1.1 Student Engagement Pattern

Student engagement can be considered as the extent of students' involvement and active participation in learning activities (Cole & Chan, 1994). In addition, student engagement through active classroom participation is an important ingredient for learning that has many educational benefits for students (Berman, 2014; Lippmann, 2013; Kuh, 2009). Hence educational data mining (EDM) techniques help researchers with the extraction of students' behavioral features in various domains including e-book reading, MOOCs learning, etc. Moreover, reading interaction variables representing student engagement have been used to prove the relation to self-regulated learning theory (Yamada, Oi, & Konomi, 2017). Therefore, in this paper, we extract several reading interaction variables as metrics for the measurement of reading engagement in digital textbook. We then analyze students reading engagement pattern and the corresponding academic performance (test scores given by the lecturers during lecture time).

1.2 Student Grouping

In terms of exploring subpopulation of students in higher educational domains, researchers often face problem on how to properly, comprehensively group students according to different demands based on tracking logs or self-report assessments. In context of learning analytics, the combination of

students with different learning styles in specific groups may have in the final results of the tasks accomplished by them collaboratively (Alfonseca et al., 2006). Therefore, many of the researchers applied clustering algorithms for optimal student grouping such as k-means (Kizilcec, Piech, & Schneider, 2013) or Ward's method (Pardo, Han, & Ellis, 2017) in order to explore a subgroup of learners with specific learning pattern in the context of digital textbook reading, MOOCs learning, and Self-Regulated Learning (SRL) theory. Data clustering is a process of extracting previously unknown, valid, positional useful and hidden patterns from large data sets (Connolly, 1999). The goal of clustering is to identify structure in dataset by objectively organizing data into homogeneous groups where the within-group-object similarity is minimized, and the between-group-object dissimilarity is maximized (Liao, 2005).

In this paper, we group students by a standard centroid-based clustering algorithm k-means method, to explore the presence of subpopulation of students that can be differentiated based on their interaction characteristics and academic performances in digital textbook reading. Moreover, we identify subpopulation of students based on their engagement pattern observed in clustering analysis results.

1.3 Digital Textbook System

BookRoll is a digital textbook reading system which is able to offer many kinds of interaction between users and system, including adding memos and highlighting text, etc (Flanagan & Ogata, 2017; Ogata et al., 2015). In BookRoll, student reading behaviors can be tracked and recorded into the learning analytics system (Flanagan & Ogata, 2017). By analyzing students' reading interactions recorded in BookRoll, in this paper, we expect to answer the following two research questions:

1. How many subpopulations of students can be identified based on reading interactions?

2. How do students' academic performances differ in different subpopulations of students?

2 METHOD

2.1 Data Collection and Variable Extraction

In this paper, we cluster and explore students' engagement pattern in digital textbook reading based on their reading interaction variables and identify the subpopulation of students based on reading characteristics. We used KU dataset[1] which is one of the given datasets that contains around 1.9 million students' click-steam reading events from ten classrooms with totally 1326 students. All classrooms used the same learning materials and quizzes. Students' reading events are collected by BookRoll system. In KU dataset, students from ten classrooms were provided the same learning contents with the same curriculum designs during the semester. Therefore, we combined ten classrooms into one then compared students reading interactions. Moreover, in order to analyze students' engagement pattern and the corresponding academic performance in digital textbook, we extracted seven variables from reading events collected in BookRoll as shown in Table 1. We also included students' test scores (academic performance) as one of the variables for clustering.

[1] https://sites.google.com/view/lak19datachallenge

Furthermore, since we wanted to obtain a better distribution of population for the following clustering analysis, a two-stage approach for the outlier removal when using k-means (Hautamäki et al., 2005) was performed. The first stage consist of purely k-means process, while the second stage iteratively removes vectors which are far away from the cluster centroid, resulting of 9 outliers were removed from 1326 students.

Table 1: Description of digital textbook reading variables (N=1317).

Variable	Description of Variable	Average	SD
SESSION	Total number of reading session	16.30	7.60
NEXT	Total times students turn to next page	856.26	468.00
PREV	Total times students turn to previous page	425.84	320.66
PREV/NEXT	Clicking ratio of PREV and NEXT	0.46	0.17
NOTE	Total times students add notes	78.08	120.14
SEARCH	Total times students search for contents	1.27	3.82
JUMP	Total times students jump to another page	36.69	3.82
SCORE	Students' test score given by lecturers	83.70	7.82

2.2 Clustering Analysis

In this paper, k-means method (MacQueen, 1967) from Python packages was applied to cluster 1317 students into different groups based on their digital textbook reading variables as shown in Table 1. Reading variables from 1317 students were normalized in advance by using Z-score normalization. We determined the optimal number of clusters for k-means method by applying Elbow method which is one of the most popular method for determining the optimal number of clusters in a data set (Ng, 2012). The Elbow method maps the within-cluster sum of squares onto the number of possible clusters. The location of the elbow in the resulting plot suggests an optimal number of clusters objectively. We then computed the average score for each individual cluster for the representation of the corresponding academic performance. The optimal number of clusters by Elbow method and the results of clustering analysis are shown in Figure 1 and Table 2 and explained in the next section.

3 RESULTS AND DISCUSSIONS

In this section we present the optimal number of clusters determined by Elbow method and results of clustering analysis. By applying Elbow method for the optimal number of clusters, we obtained several possible optimal numbers of clusters which were 2, 5, and 8 as shown in Figure 1. We then clustered students' reading interaction variables based on those obtained number of clusters accordingly. We finally chose 5 as the optimal number of clusters since we observed the most explainable results of students' engagement pattern and corresponding academic performance. Based on the optimal number of clusters, we clustered 1317 students into 5 groups, the average value and standard deviation of each variable for each group are shown in Table 2. The number of students from cluster 1 to cluster 5 are 512 (38.9%), 177 (13.4%), 256 (19.4%), 338 (25.7%), 34 (2.6%), respectively. As shown in Table 2, we identified 5 student subpopulations based on the engagement patterns in digital textbook reading and the characteristics of each subpopulation of students are described below.

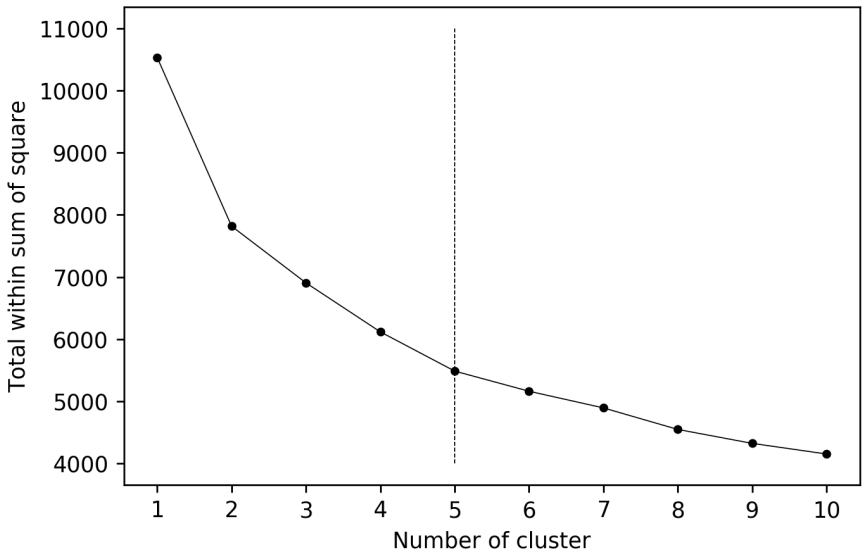

Figure 1: Optimal number of clusters by Elbow method

Table 2: Students' reading engagements and academic performances in different cluster (N=1317).

		Average (SD)							
Cluster	**n**	**SESSION**	**NEXT**	**PREV**	**PREV/NEXT**	**NOTE**	**SEARCH**	**JUMP**	**SCORE**
1	512	12.02 (4.45)	548.97 (233.51)	236.20 (129.84)	0.41 (0.15)	35.36 (58.94)	0.38 (1.14)	22.89 (15.74)	87.78 (4.58)
2	177	25.48 (8.93)	1333.75 (471.62)	586.34 (305.92)	0.43 (0.12)	271.53 (174.28)	1.49 (2.44)	88.77 (48.85)	85.73 (6.00)
3	256	14.74 (5.95)	625.61 (253.32)	231.15 (136.90)	0.36 (0.15)	27.54 (51.94)	0.46 (1.33)	27.70 (20.74)	72.97 (6.52)
4	338	18.46 (5.81)	1225.40 (382.17)	772.90 (297.66)	0.63 (0.10)	74.20 (82.17)	1.18 (2.14)	32.48 (18.35)	84.34 (6.11)
5	34	23.09 (10.16)	1064.91 (418.82)	461.65 (255.66)	0.43 (0.15)	133.47 (133.43)	20.50 (8.44)	83.00 (43.66)	86.21 (5.68)

Cluster 1: Students in cluster 1 engaged the least on digital textbook reading compared to other four groups such as session reading, contents searching, etc. Surprisingly, students in this group obtained the highest academic performances as shown in Table 2. For now, we do not know the reason of it, still, the observation in this cluster showed us the subpopulation of Low Engagement High Academic Performance.

Cluster 2: Students in cluster 2 engaged more on session reading, NEXT events, note adding, and page jumping compared to other groups. Students in this group obtained similar academic performances to cluster 1. The observation in this cluster showed us the subpopulation of High Engagement (SESSION, NEXT, NOTE and JUMP) High Academic Performance.

Cluster 3: Students in cluster 3 also engaged very few on digital textbook reading compared to cluster 2, 4, and 5, such as sessions of reading, note adding, etc. Unsurprisingly, students in this group obtained the worst academic performances as shown in Table 2. To mention an interesting finding in this paper, the clicking ratio of PREV event and NEXT event (PREV/NEXT) in this group is significantly lower than other groups as shown in Table 2, indicating that students in this group tended to turn to next page frequently but rarely turned back to previous pages for review while reading. The observation in this cluster showed us the subpopulation of Low Engagement Low Academic Performance.

Cluster 4: Students in cluster 4 engaged more on NEXT event PREV events and clicking ratio of PREV events and NEXT events, indicating that students in this group tended to turn to next page frequently and also turned back to previous pages frequently for review while reading. Although students in this group engaged not as much as cluster 2 and cluster 5 on session reading, note adding, and page jumping, they engaged more comprehensive than cluster 1 and cluster 3 and the academic performances are similar to cluster 1, cluster 2 and cluster 5. The observation in this cluster showed us the subpopulation of High Engagement (NEXT, PREV and PREV/NEXT) High Academic Performance.

Cluster 5: Students in cluster 5 engaged more on sessions of reading, note adding, contents searching, and page jumping compared to other groups. Students in cluster 5 obtained similar academic performances to cluster 1, cluster 2, and cluster 4. The observation in this cluster showed us the subpopulation of High Engagement (SESSION, NOTE, SEARCH and JUMP) High Academic Performance.

4 CONCLUSION

In this paper, we investigated subpopulation of students in digital textbook reading. Students' engagement pattern and the corresponding academic performance in digital textbook reading are analyzed by applying k-means algorithm for clustering. We clustered 1317 students into 5 different groups based on reading variables extracted from BookRoll. To answer two research questions above, we identified 5 students' reading characteristics to represent different subpopulation of students in digital textbook reading which are Low Engagement High Academic Performance, High Engagement (SESSION, NEXT, NOTE and JUMP) High Academic Performance, Low Engagement Low Academic Performance, High Engagement (NEXT, PREV and PREV/NEXT) High Academic Performance, and High Engagement (SESSION, NOTE, SEARCH and JUMP) High Academic Performance. The results showed us that subpopulation of students in digital textbook reading can be identified by clustering students into different groups as students' engagement patterns and academic performances differ while learning. Lastly, the obtained results provide researchers opportunities to find homogeneous groups for collaborative group activities and also demonstrated the importance of student grouping with respect to learning analytics. As an implication, we hope that the results provide chances for instructors to consider different kinds of intervention for the improvement of engagement for different subpopulation of students in digital textbook reading.

ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant Number 16H06304.

REFERENCES

Alfonseca, E., Carro, R. M., Martín, E., Ortigosa, A., & Paredes, P. (2006). The impact of learning styles on student grouping for collaborative learning: a case study. *User Modeling and User-Adapted Interaction*, *16*(3-4), 377-401.

Berman, R. A. (2014). Engaging students requires a renewed focus on teaching. *Chronicle of Higher Education*, *61*(3), 28-30.

Cole, P. G., & Chan, L. (1994). *Teaching principles and practice*. Prentice Hall.

Connolly T., C. Begg and A. Strachan (1999) Database Systems: A Practical Approach to Design, Implementation, and Management (3rd Ed.). Harlow: Addison-Wesley.687

Flanagan, B., Ogata, H. (2017). Integration of Learning Analytics Research and Production Systems While Protecting Privacy. In International Conference on Computers in Education (ICCE2017) (pp.333-338).

Hautamäki, V., Cherednichenko, S., Kärkkäinen, I., Kinnunen, T., & Fränti, P. (2005, June). Improving k-means by outlier removal. In *Scandinavian Conference on Image Analysis* (pp. 978-987). Springer, Berlin, Heidelberg.

Kizilcec, R. F., Piech, C., & Schneider, E. (2013, April). Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In *Proceedings of the third international conference on learning analytics and knowledge* (pp. 170-179). ACM.

Kuh, G. D. (2009). The national survey of student engagement: Conceptual and empirical foundations. *New directions for institutional research*, *2009*(141), 5-20.

Liao, T. W. (2005). Clustering of time series data—a survey. *Pattern recognition*, *38*(11), 1857-1874.

Lippmann, S. (2013). Facilitating Class Sessions for Ego-Piercing Engagement. *New Directions for Teaching and Learning*, *2013*(135), 43-48.

MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, No. 14, pp. 281-297).

Ng, A. (2012). Clustering with the k-means algorithm. *Machine Learning*.

Ogata, H., Yin, C., Oi, M., Okubo, F., Shimada, A., Kojima, K., & Yamada, M. (2015). E-Book-based learning analytics in university education. In International Conference on Computer in Education (ICCE 2015) (pp. 401-406).

Pardo, A., Han, F., & Ellis, R. A. (2017). Combining university student self-regulated learning indicators and engagement with online learning events to predict academic performance. *IEEE Transactions on Learning Technologies*, *10*(1), 82-92.

Yamada, M., Oi, M., & Konomi, S. I. (2017). Are Learning Logs Related to Procrastination? From the Viewpoint of Self-Regulated Learning. *International Association for Development of the Information Society*.