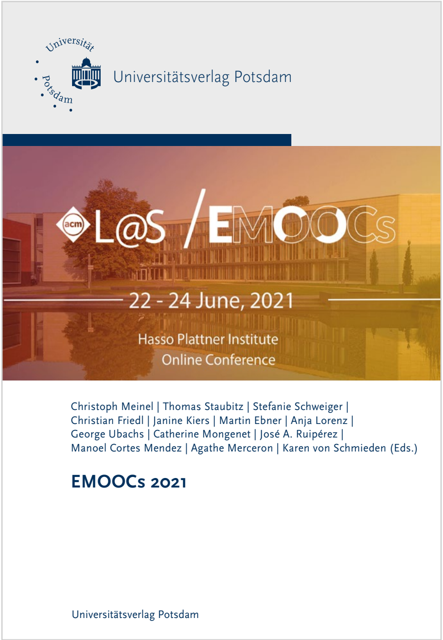
Christoph Meinel, Thomas Staubitz, Stefanie Schweiger, Christian Friedl, Janine Kiers, Martin Ebner, Anja Lorenz, George Ubachs, Catherine Mongenet, José A. Ruipérez-Valiente, Manoel Cortes Mendez (Eds.)

EMOOCs 2021

2021 – xii, 295 p.
ISBN 978-3-86956-512-5
DOI https://doi.org/10.25932/publishup-51030



Universitätsverlag Potsdam

L@S / EMOOCs

22 - 24 June, 2021

Hasso Plattner Institute
Online Conference

Christoph Meinel | Thomas Staubitz | Stefanie Schweiger |
Christian Friedl | Janine Kiers | Martin Ebner | Anja Lorenz |
George Ubachs | Catherine Mongenet | José A. Ruipérez |
Manoel Cortes Mendez | Agathe Merceron | Karen von Schmieden (Eds.)

EMOOCs 2021

Universitätsverlag Potsdam

Who Are the Students of MOOCs?
Experience from Learning Analytics Clustering Techniques

Mohammad Khalil

Centre for the Science of Learning & Technology (SLATE)
University of Bergen, Norway
mohammad.khalil@uib.no

Clustering in education is important in identifying groups of objects in order to find linked patterns of correlations in educational datasets. As such, MOOCs provide a rich source of educational datasets which enable a wide selection of options to carry out clustering and an opportunity for cohort analyses. In this experience paper, five research studies on clustering in MOOCs are reviewed, drawing out several reasonings, methods, and students' clusters that reflect certain kinds of learning behaviours. The collection of the varied clusters shows that each study identifies and defines clusters according to distinctive engagement patterns. Implications and a summary are provided at the end of the paper.

1 Introduction

Massive open online courses (MOOCs) attract hundreds and thousands of students who barely reach the end of the courses. Only a small fraction of students successfully finishes MOOCs [6]. At the current state of MOOCs in 2021, understanding the low percentage of students who succeed in MOOCs has been thoroughly researched in the literature. Among the reasons, we may link that to student motives to get a specific portion of knowledge from courses, the lack of interest in courses because of the general infrastructure of the MOOC platforms or the courses, MOOC design issues, or other reasons from student social lives that prevent them from continuing. Many researchers and practitioners focused on student engagement to keep the students' retention high to the extent that it may keep them motivated.

Because tracking student interactions in MOOCs in real-time is challenging [13], at least to a large extent of the MOOC platforms including off-the-shelf and custom learning analytics tools, engagement cannot be easily incited. However, engagement can be understood by interpreting their behaviours in MOOCs and by then define their type of involvement [12]. Still, MOOCs provide excellent scope for analysing large-scale online interaction and behavioural data to explore, understand and improve student engagement, and the overall experience. We have also seen how the educational research of MOOCs, including but not limited to Learning Analytics and Educational Data Mining, has led to a surge in explanation, predicting, and optimizing students' behaviour not only for MOOCs but also for other online learning environments (e.g. learning management systems, virtual labs).

Data-driven methods in education (i.e. Learning Analytics methods) such as clustering are popular techniques to categorize student engagement in MOOCs. The goal of clustering is to discover a new set of categories. In the case of MOOCs, clustering brings potential in understanding students' behaviour and their degree of engagement. Howbeit, the knowledge that exists on profiling students in MOOCs is limited and is just restricted to the ratio of presence and general profile description [3]. In fact, this has opened the line for this work to highlight a selected number of papers from the literature that bring in a good overview on MOOC student profiles based on their level of engagement and using clustering techniques.

The main body of the paper is organised as follows: 1) a list of the five selected papers, 2) the reasoning behind using clustering from the papers, 3) methods used in clustering, 4) variables used from the MOOCs to carry out the clustering, 5) the identified clusters by the researchers from the five papers, and 6) finally a summary of grouping the clusters from the selected works.

1.1 Clustering of MOOC Datasets

Recently, researchers have been exploring clustering methods of educational datasets and comparing them aiming for creating cohorts of mutual objects. Despite that clustering techniques are broadly used in other fields, they are not thoroughly explored as much as those in computing in educationa [1]. Applying clustering in MOOCs is powerful. Barthakur, Kovanovic, Joksimovic, Siemens, Richey, and Dawson [1] were able to track the assessment of student learning strategies over four MOOCs. The implication of their work has been reflected in identifying those with poor self-regulation skills and suggest interventions. Another study by Li, You, and Sun [11] shows that their clustering has helped them to divide groups of students which in return make them to understand students in real-time. Li, You, and Sun [11] see that their contribution will yield into enhancing MOOCs teaching.

2 Five Studies, Diverse Student Categories

This work is based on our experience with the "Better Learning Experience (BLE)" project, of which we worked on designing a customized clustering of students used in OXALIC, an Open Edx Advanced Learning Analytics tool [8]. Our selection of the five papers in this study has: 1) focused on engagement as a reason for clustering, 2) classify cohorts based on degrees of engagement (e.g. low engaged, high engaged), and 3) used Learning Analytics as a keyword to do the study and reveal a variety of student clusters. We scanned the first appearing 50 papers in Google Scholar, provided "MOOCs", "Learning Analytics", "Clustering" as search terms and applied the inclusion criteria. The selected five papers are: "Deconstructing Disengagement: Analysing Learner Subpopulations in Massive Open Online Courses" [10]; "Moving through MOOCS: Pedagogy, learning design and patterns of engagement" by Ferguson, Clow, Beale, Cooper, Morris, Bayne, and Woodgate [4]; "Clustering patterns of engagement in Massive Open Online Courses (MOOCs): the use of learning analytics to reveal student categories" by Khalil and Ebner [9]; "Who will pass? Analysing learner behaviours in MOOCs" by Tseng, Tsao, Yu, Chan, and Lai [14]; and "Research on Clustering Mining and Feature Analysis of Online Learning Behavioural Data Based on SPOC" by Zhang, Zhang, and Ran [15].

2.1 Reasoning for Using Clustering

We first start by listing the reasons behind carrying out the clustering in the selected papers. While the focus of the studies relied on classifying the students based on their engagement, we wanted to explore the motivation behind using clustering in the five papers (see Table 1). The reviewed works show that exploring learning behaviour and defining common learning characteristics have been the authors' greatest motive to do the clustering in the MOOCs.

2.2 Method(s) Used for Clustering

Table 2 shows the used clustering techniques in the five selected papers. It is clear that the authors preferred to use k-means clustering in the first place. While it is essential to validate clustering [5], only two studies validated the outcome. Validation of clustering and engagement can show that the number of clusters of students provides an authentic generalization of the data.

2.3 MOOC Variables Used for Clustering

Clustering depends on sourcing out data variables. In MOOCs, variables vary depending on the platform, Learning Analytics data collection systems, and the objectives of the clustering. In the case of the five studied papers, MOOC variables used to cluster the students (see Table 3) were a series of at least two events, such as the study by Kizilcec, Piech, and Schneider [10], and five variables like Zhang, Zhang, and Ran's [15] study.

Table 1: Reasons for using clustering in MOOCs

Study	Reasons
Kizilcec, Piech, & Schneider (2013)	Characterise learning engagement, define learning trajectories of patterns of engagement, analyse learning behaviour
Ferguson et al. (2015)	Investigate engagement; create engagement profiles; engagement with content, with assessment and with discussion
Khalil & Ebner (2016)	Discover characteristics of student profiles, portray engagement and behaviour of learners, assign common learning styles to groups
Tseng et al. (2016)	Understand students' engagement in MOOCs and offer insight to what keeps the student engaged in MOOCs
Zhang et al. (2018)	Explore learning behaviour characteristics, combine learner behavioural data with clustering algorithms to group learners into relatively homogenous groups

Table 2: Clustering methods

Study	Clustering method	Clustering validation
Kizilcec, Piech, and Schneider [10]	custom longitudinal distribution and k-means clustering	Yes, silhouette cluster validation
Ferguson, Clow, Beale, Cooper, Morris, Bayne, and Woodgate [4]	k-means clustering	No
Khalil and Ebner [9]	k-means clustering	Yes, elbow method
Tseng, Tsao, Yu, Chan, and Lai [14]	Ward's hierarchical and k-means non-hierarchical clustering	No
Zhang, Zhang, and Ran [15]	k-means clustering + hierarchical clustering	No

Table 3: MOOC variables used for clustering

Study	**MOOC variables**
Kizilcec, Piech, and Schneider [10]	course–video lectures and assessments
Ferguson, Clow, Beale, Cooper, Morris, Bayne, and Woodgate [4]	course visits, forum posts, assessment submission
Khalil and Ebner [9]	discussion forum reading frequency, discussion forum writing frequency, videos watched, self-assessment attempts
Tseng, Tsao, Yu, Chan, and Lai [14]	logging in system, watching lecture videos, submitting assignments, posts in the discussion forums
Zhang, Zhang, and Ran [15]	discussion forum posts, discussion forum replies, final scores, total duration of watched videos, number of videos viewed

2.4 Clusters Identified

Kizilcec, Piech, and Schneider [10] found four categories in three MOOCs on the Coursera platform based on their engagement behaviour as the following:

- On track: if students submitted assessment in the week it was set
- Behind: if students completed an assessment after the week in which it was set
- Auditing: if students engaged with content but not with the assessment
- Out: if students did not participate in a course week

Furthermore, Kizilcec, Piech, and Schnieder further employed cluster analysis and categorized the students into the four following categories:

- Completing: students who complete most of the assessments
- Auditing: students who watch videos but complete assessments infrequently
- Disengaging: students who complete some assessments but then withdraw from the MOOC
- Sampling: those who explore the MOOCs through simple engagement with videos

Ferguson, Clow, Beale, Cooper, Morris, Bayne, and Woodgate's [4] study applied k-mean clustering on five MOOCs, two long and three short. The number of clusters identified in the longer MOOCs which are two months long, were as the following:

- Samplers: visited a course briefly
- Strong Starters: left after the first week's assessment
- Returners: completed assessments in the first two weeks, then left
- Mid-way Dropouts: completed 3–4 assessments before leaving
- The Nearly There: cluster completed most assessments but left early
- Late Completers: completed most assessments but were either late in submitting these or missed some
- Keen Completers: engaged actively throughout the course

Ferguson, Clow, Beale, Cooper, Morris, Bayne, and Woodgate [4] applied the k-mean where k = 7 to shorter MOOCs, however the results were not meaningful due to either smaller groups identified and extremely biased toward one of the variables (see Table 3). The authors then applied different value for the k-means for the other three MOOCs and found the following clusters:

- Very Weak Starters: who show low level of engagement in the first weeks
- Improvers: students whose their activity rise along the MOOC
- Surgers: who visit more than two-third of a short MOOC but stop before the MOOC finishes
- Saggers: engaged actively throughout the course but not as high as the keen completers

Khalil and Ebner [9] study applied k-means study on two MOOCs. The first one is a compulsory MOOC provided to undergraduates, and the second MOOC was a free open one. The number of clusters founded by Khalil and Ebner are four and three clusters respectively as the following:

- Dropout: low engaged students with high attrition
- Perfect students: very active students who are engaged in forums, video watchers, and pass all the self-assessment tests
- Gaming the system: students who pass the exams but have several attempts with barely watched videos

- Social: students who are engaged only in forums.

The authors when replicating the study on the open and free MOOC found three out of four groups: "Dropout", "Perfect students", and "Gaming the system". "Social" students were not detected by the k-means value applied before.

Zhang, Zhang, and Ran's [15] work explored a small private online course (SPOC) using k-means and hierarchical analysis and found the following four groups of learners:

- Weak-cognitive learners: those with high video viewing rates, long duration but low final scores.
- Self-conscious learners: the excellent learners who have completed the indicators that do not count toward achievement.
- Short-cut learners: those with a higher final score, but who have a low completion rate of indicators that do not count towards achievement
- Lazy learners: the learners who do not have high-scored indicators.

Tseng, Tsao, Yu, Chan, and Lai [14] classified learning behavior in three MOOCs provided by the Yuan Ze University in Taiwan and came up with three types of students based on their interaction in the MOOCs as follows:

- Active learner: who submitted assignments on time and frequently watched lecture videos with high completion and engagement ratio.
- Passive learner: who frequently watch MOOC videos, show limited participation in course forums, and attempted few assignments and quizzes
- Bystander: learners who register and their activity is way below a low threshold.

3 Clusters – Combined

The total number of clusters identified by the researchers from the five papers is (N = 30) clusters. Each of which describes distinctive learning behaviour according to each paper. Nevertheless, combining all the clusters together draws a line of common learning attitude that is shared in the alluvial illustration in Figure 1 below. Unfortunately, grouping clusters alike is subjective to what the authors of the papers describe. For instance, Zhang, Zhang, and Ran's [15] four clusters were identified based on students' cognitive ability and engagement. In my attempt to group the 30 clusters into general themes, three are identified: Motivation (N = 7), Chronology (N = 6), and Commitment (N = 17). The motivation includes students by whom their engagement present motives in studying in MOOCs according to their engagement. Chronology includes students by whom their engagement is defined within a time frame. Commitment incorporates hierarchies that belong to the engagement of dropping out or completing.

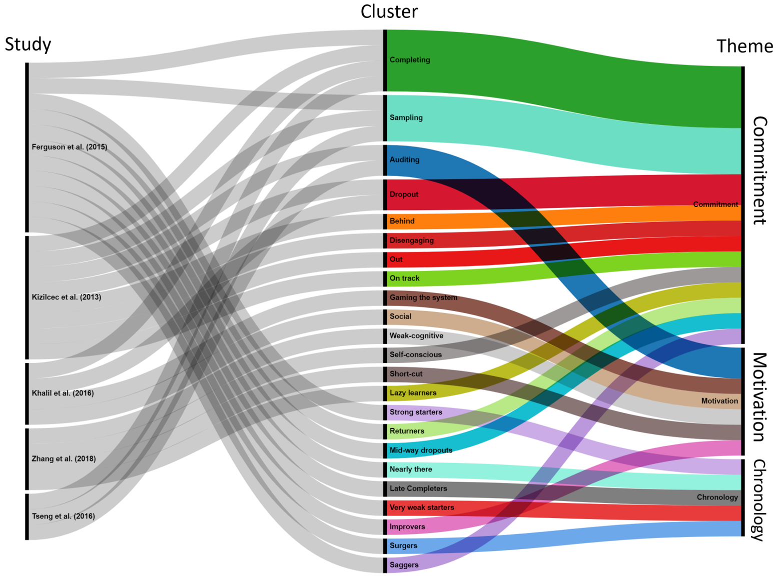


Figure 1: Grouping the student clusters from the five selected papers (N = 30 clusters)

4 Implications and Conclusions

This work is a contribution to provide a brief and quick overview on how clustering is used in MOOCs. Sharing experiences in the field of MOOCs is a key functioning wheel for research. Through reviewing five papers from the literature on clustering in MOOCs, the following remarks are highlighted:

- *Clustering in MOOCs is connected to Learning Analytics with focus on engagement*: research that uses datasets from MOOCs is strongly linked to Learning Analytics. The goal of which is to optimise the learning experiences of the learners.

The reviewed papers looked deeper into engagement, trying to understand the learning behaviour through clustering. However, clustering MOOCs is not limited to engagement, but other forms of grouping and classification such as text and discourse clustering.

- *K-means is the popular technique for clustering in MOOCs*: Table 2 shows that k-means is the primary method used to cluster engagement. Bharara, Sabitha, and Bansal [2] had also concluded the same results including platforms other than MOOCs. Through a bigger review of 15 papers, the authors found that k-means was used the most among the papers. Barthakur, Kovanovic, Joksimovic, Siemens, Richey, and Dawson [1] agreed that k-means is the most used algorithm for clustering in MOOCs.
- *There is a positive correlation between engagement and completion ratio*: the reviewed papers show that active engagement through fulfilling MOOCs' assignments and tasks are associated to an increased completion rate.
- *Cohorts in clustering are named subjectively to the authors and study objectives*: Even though researchers came to similarities of learners' behaviour in MOOCs, their naming of the identified cohorts was different. For example, "Perfect student" in one study was named "Completing" in another.
- *Clustering has not yet been used to optimize student learning*: The papers reviewed are clickstream (i.e. clicks of activity) dependent. That is, clustering is focused on exploring and explaining learning behaviour based on data-driven approaches. There is a gap in linking traces of students' learning behaviour in MOOCs and the actual learning processes. Perhaps including self-reporting information from the students helps improve student learning.
- *Absent learning theories*: It becomes quite common in data-driven approaches that learning theories are overlooked [7]. This has been evident in the reviewed studies. Data might not only suggest that theory is unnecessary but that it could make sense of that data.

5 Limitations

The study has some limitations. The search to retrieve and select the papers were conducted only using Google Scholar, the number of selected papers was restricted to five, the aggregation of all the clusters from the five studies is subjective, and even though the time of the work publication of this work is 2021, the most recent paper of the five reviewed is 2018.

References

[1] A. Barthakur, V. Kovanovic, S. Joksimovic, G. Siemens, M. Richey, and S. Dawson. *Assessing program-level learning strategies in MOOCs*. 2021.

[2] S. Bharara, S. Sabitha, and A. Bansal. "Application of learning analytics using clustering data Mining for Students disposition analysis". In: *Education and Information Technologies* 23.2 (2018), pages 957–984.

[3] R. Cabedo, T. C. Edmundo, and M. Castro. "A Benchmarking Study of Clustering Techniques Applied to a Set of Characteristics of MOOC Participants". In: 2016 ASEE Annual Conference & Exposition (New Orleans, Louisiana). 2016, page 26247.

[4] R. Ferguson, D. Clow, R. Beale, A. J. Cooper, N. Morris, S. Bayne, and A. Woodgate. "Moving through MOOCS: Pedagogy, learning design and patterns of engagement". In: *Design for teaching and learning in a networked world*. Cham: Springer, 2015, pages 70–84.

[5] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. *On clustering validation techniques. Journal of intelligent information*. 2001.

[6] K. S. Hone and G. R. El Said. "Exploring the factors affecting MOOC retention. A survey study". In: *Computers & Education* 98 (2016), pages 157–168.

[7] I. Jivet, M. Scheffel, M. Specht, and H. Drachsler. "License to evaluate: Preparing learning analytics dashboards for educational practice". In: Proceedings of the 8th international conference on learning analytics and knowledge. 2018, pages 31–40.

[8] M. Khalil and G. Belokrys. "OXALIC: an Open edX Advanced Learning Analytics Tool". In: *2020 IEEE Learning With MOOCS*. IEEE, 2020, pages 185–190.

[9] M. Khalil and M. Ebner. *Clustering patterns of engagement in Massive Open Online Courses (MOOCs). the use of learning analytics to reveal student categories*. 2017. DOI: 10.1007/s12528-016-9126-9.

[10] R. F. Kizilcec, C. Piech, and E. Schneider. "Deconstructing disengagement: analyzing learner subpopulations in massive open online courses". In: Proceedings of the third international conference on learning analytics and knowledge. 2013, pages 170–179.

[11] Z. Li, F. You, and D. Sun. "Research on the evaluation of learning behavior on MOOCs based on cluster analysis". In: Proceedings of the 2020 4th International Conference on Electronic Information Technology and Computer Engineering. 2020, pages 1055–1059.

[12] “Modeling learner engagement in MOOCs using probabilistic soft logic”. In: *NIPS workshop on data driven education*. Volume 21. 2013, page 62.

[13] C. Schumacher and D. Ifenthaler. *Features students really expect from learning analytics. Computers in human behavior*. 2018.

[14] S. F. Tseng, Y. W. Tsao, L. C. Yu, C. L. Chan, and K. R. Lai. “Who will pass? Analyzing learner behaviors in MOOCs”. In: *Research and Practice in Technology Enhanced Learning* 11.1 (2016), pages 1–11.

[15] G. Zhang, Y. Zhang, and J. Ran. “Research on Clustering Mining and Feature Analysis of Online Learning Behavioral Data Based on SPOC”. In: 2018 13th International Conference on Computer Science & Education. IEEE, Aug. 2018, pages 1–6.