Visualizing MOOC User Behaviors: A Case Study on XuetangX

Tiantian Zhang and Bo Yuan(✉)

Intelligent Computing Lab, Division of Informatics, Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, People's Republic of China
2573546543@qq.com, yuanb@sz.tsinghua.edu.cn

Abstract. The target of KDD CUP 2015 is to use the MOOC (Massive Open Online Course) user dataset provided by XuetangX to predict whether a user will drop a course. However, despite of the encouraging performance achieved, the dataset itself is largely not well investigated. To gain an in-depth understanding of MOOC user behaviors, we conduct two case studies on the dataset containing the information of 79,186 users and 39 courses. In the first case study, we use visualization techniques to show that some courses are more likely to be simultaneously enrolled than others. Furthermore, a set of association rules among courses are discovered using the Apriori algorithm, confirming the practicability of using historical enrollment data to recommend courses for users. Meanwhile, clustering analysis reveals the existence of clear grouping patterns. In the second case study, we examine the influence of two user factors on the dropout rate using visualization, providing valuable guidance for maintaining student learning activities.

Keywords: User behavior · Visualization · Association rule · Clustering · MOOC

1 Introduction

The MOOC (Massive Open Online Course) refers to a new type of online courses aimed at unlimited participation and open access via the web [1]. It is the result of a recent development in distance education [2], which was first introduced in 2008 and emerged as a popular learning mode in 2012 [3]. Most importantly, MOOCs build on the active engagement of millions of students who self-organize their participation according to their individual learning goals, prior knowledge and skills as well as common interests [4]. MOOCs provide not only traditional course materials such as lecture slides, question sets and reading materials but also course videos, online self-testing, and interactive forums to support community interactions among students, instructors, and teaching assistants [1]. Meanwhile, as a brand new education mode based on the concept of "Internet + Education", MOOCs break the limitation of space and time and users can conduct individual learning anytime and anywhere.

User behavior analysis is an important factor for designing and implementing an effective MOOC, which can make learning a convenient, rewarding and personalized experience. Meanwhile, the ability of MOOCs to generate a tremendous amount of data opens up unprecedented opportunities for educational research [5]. Researchers can get

H. Yin et al. (Eds.): IDEAL 2016, LNCS 9937, pp. 89–98, 2016.
DOI: 10.1007/978-3-319-46257-8_10

better acquainted with users' learning behavior and learning outcomes by taking advantage of data analytics methods and propose insightful suggestions on curriculum improvement. Educators can also provide personalized guidance and recommend appropriate materials to learners to improve the quality of learning.

As a new education mode, existing research on MOOCs is relatively limited. In 2012, an article in *Science* made an introduction of MOOCs and predicted that they will change the future of education [6]. Afterwards, an article in *Nature* discussed the development and trend of MOOCs in 2013 [7]. In the last two years, MOOCs have received more and more attentions from the research community. Some examples are: evaluating the geographic data in MOOCs [8]; analyzing student submissions to help instructors understand the problem solving process [9]; using the records of learning activity to develop a conceptual framework for understanding how users engage with MOOCs [10]; investigating and improving the peer assessment mechanism [11, 12]; using timely interventions to improve online learning [13]. Researchers have also proposed a framework to classify posts in discussion forums [14] and established a large scale collaborative data analytics platform named MoocViz to analyze the data from different courses and MOOC platforms [15]. There are also studies on user behavior to speculate which MOOC platforms are easy to use [16] as well as on dropout prediction and user retention [17, 18].

In the above, there are very few studies on courses themselves. However, it is not surprising that students often do not know which courses are the most beneficial ones for them and which other courses they also need to enroll to make the most of their study. Our work focuses on the data of user enrollment on MOOCs to investigate the relationships among courses and make proper course recommendation for students. Another critical issue is the high dropout rate. For all MOOC platforms, there is a phenomenon that a large number of users may enroll in a course but most of them will drop the course somewhere before the end. In this paper, we also study the connection between learning behavior and course completion rate, which will assist in designing and implementing effective MOOCs to maintain and encourage learning activities.

Section 2 provides a detailed description of the data source. Section 3 contains the case study on course analysis and presents the association rules among courses and shows the clustering pattern of courses. Section 4 presents another case study on factors affecting the dropout rate. This paper is concluded in Sect. 5 with some discussions and directions for future work.

2 Preliminaries

The high dropout rate has been a major issue in MOOC platforms and some reports show that the certification rates may be less than 10 %. Due to the inherent characters of online learning, users take far less obligation than in the normal classroom. Also, there are many distracting factors that may prevent users from consistent learning. After all, users may enroll in a course with totally different intentions and motivations (e.g., browsing vs. earning certification). In order to maintain and encourage students' learning activities, it is important to predict their likelihood of dropout so that retention measures can

be taken as necessary. In KDD Cup 2015, an anonymous dataset was provided by XuetangX, the largest MOOC platform in China. The target of the competition is to predict whether a user will drop a course based on his or her prior activities within a time period of 30 days. If a user leaves no records for a specific course in the log during the next 10 days, the case is claimed as a dropout. Despite of the promising performance achieved by competition participants, there is still a lack of clear understanding of the dataset itself. In this paper, we conduct in-depth analysis of the dataset using visualization and other data analytics techniques to identify relationships among courses, which can help students make right decisions on enrollment. We also examine some factors associated with the dropout rate, providing valuable information to course instructors and platform operators.

The dataset contains the information of 79,186 users and 39 courses, with 120,542 enrollment records in total. Each record has a binary label indicating its dropout status with "1" indicating that the user will drop the course. The dataset is unbalanced with 79.29 % enrollment records labeled as "1" and 20.71 % labeled as "0". It also provides 8,157,277 user behavior logs within 30 days after the course starts, which contain the detailed user activities such as solving problems, watching videos or engaging in discussion. Figure 1 (left) shows the statistics of all courses where many courses have high enrollment numbers but all courses also feature high dropout rates ranging from 66.09 % to 92.93 %. Figure 1 (right) visualizes the starting time of each course where all courses roughly start from one of two time points, possibly corresponding to the two semesters in practice.

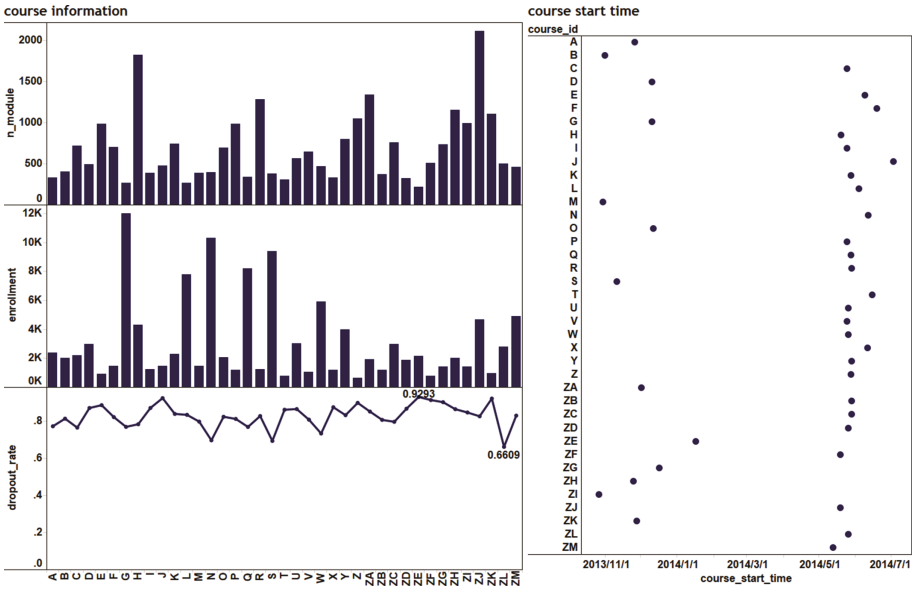

Fig. 1. The statistics of all courses: the number of modules, enrollment numbers and the dropout rate of each course (left) and the starting time of each course (right).

3 Course Analysis

From the enrollment records, we extracted two attributes username and course ID to explore the relationships among courses. There were many users enrolling in only a single course and around 27.81 % users enrolled in at least two courses. So we finally selected the enrollment records of these 22,021 users as the data source.

The relationship between courses and users was visualized using the network diagram. For the sake of clarity and readability, we further selected the records of users who enrolled in at least 6 courses and visualized them using *Gephi*. In Fig. 2, nodes with label represent courses while nodes without label represent users and the enrollment status is indicated by arcs. The size of each course node is measured by its indegree (enrollment numbers). It is clear that: (1) courses had different levels of enrollment (popularity); (2) course nodes were roughly grouped into some clusters, indicating that certain courses were often enrolled simultaneously by the same user. Based on this preliminary observation, we will conduct further investigation using association rules learning and clustering analysis to reveal more insightful information.

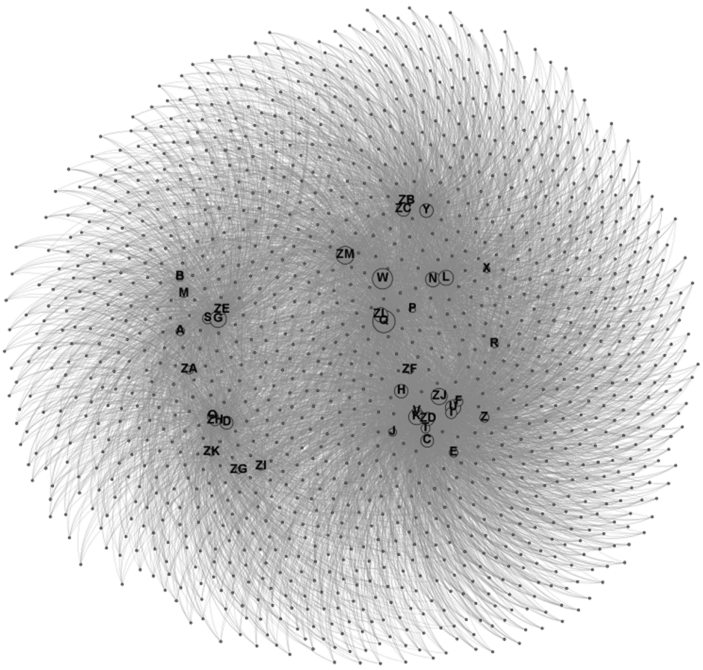

Fig. 2. Complex network graph visualization showing the relationship between courses and users who enrolled in at least 6 courses. Courses are represented by nodes with label.

We used the Apriori algorithm to find the frequent course sets and association rules among courses. The minimum support value and confidence value were set to 1.5 % and 20 %, respectively. Table 1 shows association rules with confidence greater than 30 %.

The maximum confidence value was up to 43.6 %, which means that if a user enrolls in course *ZA*, there is a possibility of 43.6 % that he/she will also enroll in course *G*. This finding is likely to be valuable in practice and these association rules can be employed by MOOC platforms to provide course recommendation services to students based on their enrollment records. By doing so, MOOC platforms are expected to become more personalized and user-friendly. Figure 3 shows all association rules among courses and the arrow from *X* to *Y* represents a rule $X \rightarrow Y$. Furthermore, the thickness of the arrow represents the confidence value of the rule.

Table 1. Association rules among courses(confidence value > 30 %)

Rule	Confidence	Rule	Confidence
('ZC') → ('Y')	0.301	('S') → ('G')	0.370
('L') →('N')	0.308	('O') → ('G')	0.375
('O') → ('D')	0.308	('A') → ('G')	0.382
('ZH') → ('G')	0.309	('D') → ('G')	0.402
('ZC') → ('W')	0.310	('ZE') → ('G')	0.406
('B') → ('S')	0.340	('C') → ('ZJ')	0.431
('ZD') → ('U')	0.357	('ZA') → ('G')	0.436

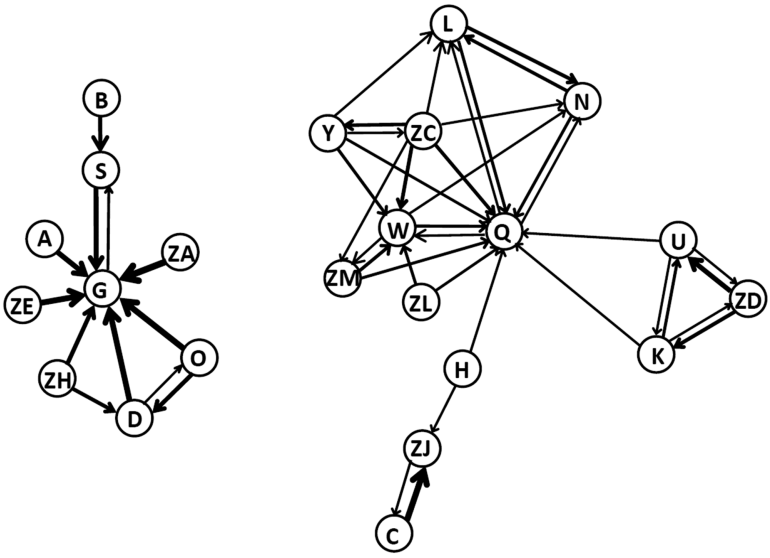

Fig. 3. Association rules among courses. The arrow from *X* to *Y* represents the rule *X*→*Y*. The thickness of the arrow represents the confidence of the corresponding rule.

We also conducted course clustering using the k-means and hierarchical clustering. The Jaccard distance (Eq. 1) was used to measure the dissimilarity between courses *M* and *N*, depending on their enrollment sets.

$$d_j(M, N) = 1 - \frac{|M \bigcap N|}{|M \bigcup N|} = 1 - \frac{|M \bigcap N|}{|M| + |N| - |M \bigcap N|} \quad (1)$$

Figure 4 shows the graphical results of course clustering using different techniques. For the k-means clustering, we used multi-dimensional scaling to transform the distance matrix of courses into a coordinate matrix. As shown in Fig. 4(a), courses were grouped into three clusters, marked by different colors. Figure 4(b) shows the dendrogram of the hierarchical clustering with the average link criterion. Although the number of clusters can be varied as necessary, it indicates the existence of three major clusters. These two clustering results are largely consistent, apart from a few discrepancies. Furthermore, if the real course names/contents are available, we can expect to reveal more underlying patterns by matching the clustering results with the properties of courses and make more principled course recommendations.

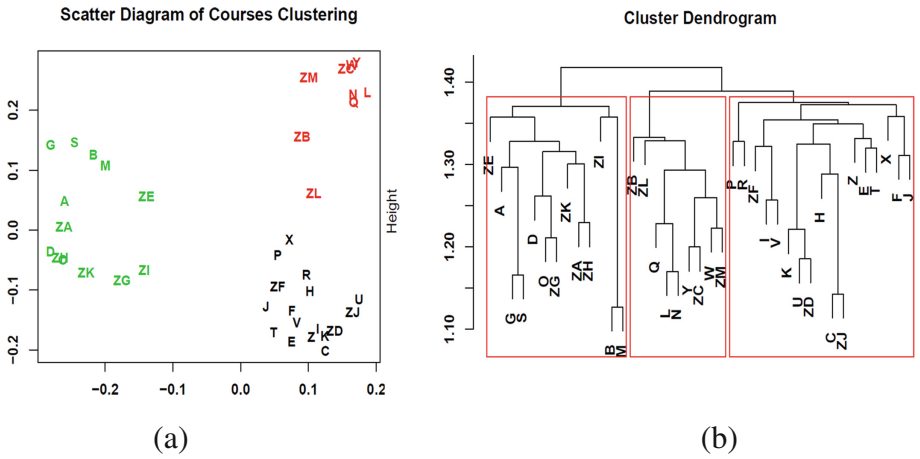

Fig. 4. The results of course clustering analysis: (a) the clustering result by k-means; (b) the dendrogram by hierarchical clustering with the average link criterion. (Color figure online)

4 Dropout Analysis

In this section, we present some analysis on the dropout pattern. We extracted the first and the last time records that users logged into the courses and their dropout label as well as the numbers of courses that they enrolled and dropped from user logs.

Based on the date of last login, we calculated the distribution of dropout students. Figure 5(top) shows the dropout rate curves of selected courses within 30 days after the starting of courses. Each curve represents a course and the dropout rate is defined as the number of dropout students who stopped learning the corresponding course at a specific date, divided by the total number of students enrolled in that course. Overall, the dropout rates of most courses were at peaks within 3 days after the course started. After that, the dropout rates declined sharply and finally reached a relatively steady state, which is

consistent with the tendency of all courses, as shown in Fig. 5(bottom) where the dropout rate is defined as the number of enrollment cancellations with a specific last login date, divided by the total enrollment numbers. This shows that most users who finally dropped the course actually stopped learning after only a few days. However, there were some courses showing different patterns. For example, the black curve increased abruptly at the 16th day and the red curve maintained a low dropout rate within the first half period but increased to its peak at the 22nd day.

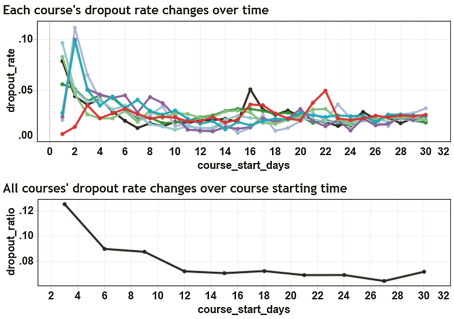

Fig. 5. Distribution of enrollment cancellations based on the last date of login

Similarly, based on the date of first login, the relationship between dropout rate and the starting time of learning is shown in Fig. 6. Within the first week, the dropout rate increased monotonically, indicating that if a user started leaning the course later, it will be more likely that he or she would drop the course. Note that the dropout rate here is defined as the number of enrollment cancellations with a specific first login date, divided by the total enrollment numbers with the same first login date. The relatively low dropout rate at the beginning may be contributed to the common sense that well-organized and self-motivated users were often get themselves ready for studying long before the course started. During the middle period (between the 10th day and the 20th day), the dropout rate was stable and fluctuated slightly but at the end of the curve, the dropout rate dropped rather unexpectedly. A possible explanation is that some late starting users still needed extra times to figure out whether the specific course was suitable for them and continued to engage in learning activities within the next 10 days after the 30th day.

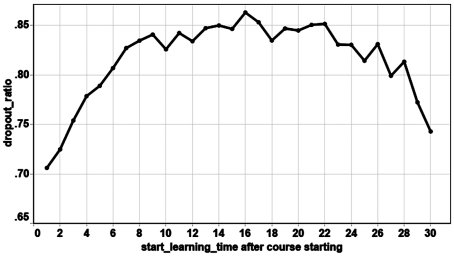

Fig. 6. Relationship between dropout rate and the first date of login

Another factor related to dropout is the number of courses enrolled. In Fig. 7, the horizontal axis represents the total number of courses enrolled and the vertical axis represents the average number of courses dropped. The simple linear regression was used to identify possible relationships. The red line is based on students who only enrolled in courses in the same semester while the black line is based on students who enrolled in courses in both two semesters. Firstly, it is evident that there was a linear relationship between the number of courses dropped and the number of courses enrolled. In other words, the probability of a user dropping a course was not directly influenced by the total number of courses enrolled.

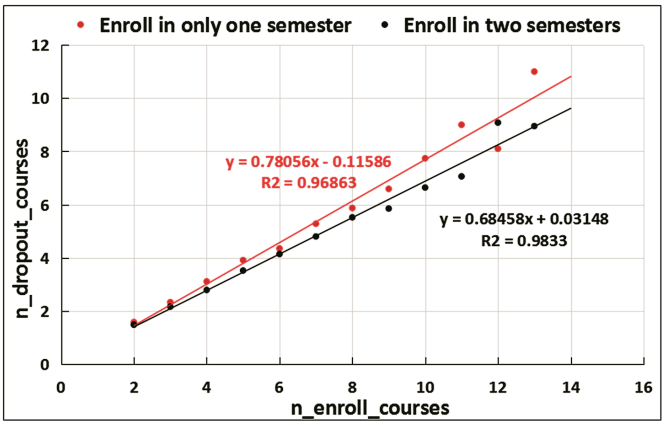

Fig. 7. The results of linear regression on the average number of courses dropped and the number of courses enrolled. It shows the correlation coefficients and the R^2 values.

Furthermore, the correlation coefficients were 78.06 % and 68.46 % for the red line and the black line, respectively, both with large R^2 values. This shows that, given the same number of courses enrolled, students enrolling in two semesters were less likely to drop courses than those enrolling in only a single semester. Consequently, we can conclude that reduced stress in study can help improve the course completion rate.

5 Conclusion

In this work, we conducted a series of MOOC user behavior analysis based on the dataset from KDD CUP 2015. In the course analysis, we found that some courses were likely to be enrolled simultaneously. Based on this finding, we discovered a set of association rules among courses so that it is possible to use the enrollment data to recommend courses for users. Furthermore, we also used the k-means and hierarchical clustering techniques to group all courses into three clusters. In the dropout analysis, we identified

some factors related to the dropout rate such as the first and last login dates as well as the distribution of courses over semesters. In general, early starting users and users distributing courses in different semesters are both less likely to drop the course compared to other students.

As more and more universities are contributing high-quality courses to MOOC platforms and even offer online degree programs, people from all over the world will benefit greatly from the popularity of MOOCs. As a new education mode, MOOCs provide unprecedented opportunities for educators and researchers but also face many challenges such as the high dropout rate and low completion rate. In the future, we will conduct further study on personalized recommendation and user behavior prediction to help build more effective MOOC platforms.

Acknowledgement. This work was partially supported by the research foundation (QTone Education) of the Research Center for Online Education, Ministry of Education, P.R. China.

References

1. Massive Open Online Course. https://en.wikipedia.org/wiki/Massive_open_online_course
2. Bozkurt, A., Akgun-Ozbek, E., Yilmazer, S., Erdogdu, E., Ucar, H., Guler, E., Sezgin, S., Karadeniz, A., Sen-Ersoy, N., Goksel-Canbek, N., Dincer, G.D., Ari, S., Aydin, C.H.: Trends in distance education research: a content analysis of journals 2009-2013. Int. Rev. Res. Open Distrib. Learn. **16**(1), 330–363 (2015)
3. Pappano, L.: The year of the MOOC. N. Y. Times **2**(12), 2012 (2012)
4. McAuley, A., Stewart, B., Siemens, G., Cormier, D.: The MOOC model for digital practice (2010). http://www.elearnspace.org/Articles/MOOC_Final.pdf
5. Breslow, L., Pritchard, D.E., DeBoer, J., Stump, G.S., Ho, A.D., Seaton, D.T.: Studying learning in the worldwide classroom: research into edX's first MOOC. Res. Pract. Assess. **8**, 13–25 (2013)
6. Stein, L.A.: Casting a wider net. Science **338**(6113), 1422–1423 (2012)
7. Waldrop, M.M.: Campus 2.0. Nature **495**(7440), 160–163 (2013)
8. Nesterko, S.O., Dotsenko, S., Hu, Q., Seaton, D., Reich, J., Chuang, I., Ho, A.: Evaluating geographic data in MOOCs. In: NIPS Workshop on Data Driven Education (2013)
9. Han, F., Veeramachaneni, K., O'Reilly, U.M.: Analyzing millions of submissions to help MOOC instructors understand problem solving. In: NIPS Workshop on Data Driven Education (2013)
10. Anderson, A., Huttnocher, D., Kleinberg, J., Leskovec, J.: Engaging with massive online courses. In: Proceedings of the 23rd International Conference on World Wide Web, pp. 687–698. ACM (2014)
11. DíezPeláez, J., Rodríguez, Ó.L., Betanzos, A.A., Troncoso, A., Rionda, A.B.: Peer assessment in MOOCs using preference learning via matrix factorization. In: NIPS Workshop on Data Driven Education (2013)
12. Shah, N.B., Bradley, J.K., Parekh, A., Wainwright, M., Ramchandran, K.: A case for ordinal peer-evaluation in MOOCs. In: NIPS Workshop on Data Driven Education (2013)
13. Williams, J.J., Williams, B.: Using interventions to improve online learning. In: NIPS Workshop on Data Driven Education (2013)

14. Stump, G.S., DeBoer, J., Whittinghill, J., Breslow, L.: Development of a framework to classify MOOC discussion forum posts: methodology and challenges. In: NIPS Workshop on Data Driven Education (2013)
15. Dernoncourt, F., Halawa, S., O'Reilly, U.: MoocViz: a large scale, open access, collaborative, data analytics platform for MOOCs. In: NIPS Workshop on Data Driven Education (2013)
16. Pireva, K., Imran, A.S., Dalipi, F.: User behavior analysis on LMS and MOOC. In: IEEE Conference on e-learning, e-Management and e-Services, pp. 21–26 (2015)
17. Yang, D., Sinha, T., Adamson, D., Rose, C.P.: Turn on, Tune in, Drop out: anticipating student dropouts in massive open online courses. In: NIPS Workshop on Data Driven Education (2013)
18. Balakrishnan, G.K.: Predicting student retention in massive open online courses using Hidden Markov Models. Technical report No. UCB/EECS-2013-109, University of California, Berkeley (2013)