International Journal of Innovation and Research in Educational Sciences
Volume 6, Issue 1, ISSN (Online) : 2349–5219

MOOC Dropout Factor Discovery Study Based on Learning Behavior

Yong Luo[a,b], Zhihao Xiao[a], Jianping Li[a,*] and Zheng Xie[a]
[a]College of Science, National University of Defense Technology, Changsha, Hunan, China
[b]Warwick University, United Kingdom.
Corresponding author email id: yngluo@163.com (Yong Luo)

Date of publication (dd/mm/yyyy): 05/03/2019

***Abstract* – MOOC is currently facing the challenge of high dropout rates. This article will mine the MOOC learning behavior data and explore the rules of online learning. Through studying the differences in online learning behaviors of various learners, we divided the MOOC learners into authenticator, practitioners, observer, visitor. Naive Bayes classification algorithm was introduced, and the classification of learners provides the basis for personalized learning. The Bayesian formula is used to mine the main factors of MOOC dropout from learning behavior data. Statistics show that watching course videos and doing exercises are the most important factors that causes students to give up. So, video recording and exercises design are still the most important works to improve the course retention rate. On the other hand, since most of the learners being not in order to obtain MOOC certificates, we need to design modules for them. Students can customize their own learning modules, which enhance the effectiveness of MOOC.**

***Keywords* – Computer Assisted Learning, Communication Technology, Medical Education, Information Technology and Systems Analysis.**

I. INTRODUCTION

MOOC is the massive open online courses. These courses are provided by internationally renowned university, and shared through the network [16]. Coursera, Udacity, edX are currently the three largest providers in the world [13]. The Chinese University MOOC [23] is currently the most popular platform in China.

Its philosophy is to enable global learners to share quality resources through information technology and web technologies. MOOC [10] have a complete set of teaching modules, which include registration, watching video lectures, quizzes, assignments, discussions, exams, graduation, certificates and other processes. Massive open online courses bring learners a new experience, and give chance to get involved in high education. On the other hand, universities have improved their own level by building MOOC courses. Therefore, the MOOC can be considered as a new exchange place between universities.

Although it has been rapid development and recognition, but there are still many controversies. From the history of online education, changes in learning resources promote the revolution in learning styles. How to use online course resources scientifically is a common concern. MOOC is currently facing such problems as high dropout rates, low resource utilization and lack of an effective profit model. MOOC courses often enroll tens of thousands of students, but often the last certificate students only 1% [18]. This requires careful analysis of the online learning behavior, from the perspective of the effectiveness to objectively evaluate the learning behavior. According to the characteristics of learning behavior to scientific and reasonable classification, so that we can more comprehensive evaluation of the value of curriculum resources.

At present, there are few researches on the statistical algorithms for the dropout. These studies are rarely mathematical modeling and algorithm design. In MOOC learning behavior data analysis, Adamopoulos (2013) [1] studied the factors that affect the retention rate of MOOC students. Judy Kay (2013) [10] explained what is

MOOC in paper and how MOOC is classified. Formanek (2018) [5] analyzed and studied motivations of college students studying MOOC courses. Sarah (2017) [15] analyzed the mutual evaluation model in MOOC. And studied the performance of students in the process of mutual assessment and their impact on learning. Alyssa (2017) [22] explored the discussion patterns between cross-disciplinary and interdisciplinary MOOC courses through data mining. Lloyd et al. (2017) [18] calculated and analyzed MOOC course learning behavior and learning performance data. Bronwen (2017) [20] studied and analyzed the combination of MOOC and classroom teaching in medical anatomy. Andrew (2017) [21] based on the characteristics of chemistry courses, studies how to enable more effective communication among learners and improve their learning in MOOC. Alario-Hoyos (2016) [2] analyzed MOOC learning behavior and its effectiveness as a means of assessing MOOC. Barak (2016) [3] analyzed motivation to learn MOOC from the perspective of student exchange. Hone et al. (2016) [7] analyzed the factors affecting the retention of MOOC learning through the questionnaire. Jiang et al. (2014) [9] studied MOOC learning behavior data for the first week of study. Use logistic regression as a classifier to predict the probability of a learner gaining a course certificate. Phan (2016) [17] analyzed students' performance in participating in the MOOC course, and did a comparative study of classroom learning and online learning. Stich (2017) [19] analyzed the MOOC in the United States and found that the platform blindly pursued the number of courses but neglected the service to learners.

This paper uses data mining [6] technology to analyze MOOC learning behavior data. According to the characteristics of learning behavior, the learners were scientifically classified. This classification is based on effective learning theory. Because end-of-course rates cannot be used as the only indicator of course evaluation. We also propose a drop-in factor influence algorithm. Through analysis of the data, it has been found that watching video and completing exercises are the biggest obstacles to students completing the course.By comparison, the impact of participation in the discussion has little effect on students' certification. Therefore, high quality course videos and exercises are the key to MOOC's success.

II. Effective Learning Theory

MOOC's core teaching principle is master learning [14]. The central point is: the learning ability of a student does not directly determine his learning effectiveness. It only determines the amount of time it takes him to master the content. Students just invest the time needed to learn the knowledge. With the help of the teacher, 90% of the students can master the knowledge imparted.

$$\text{Effectiveness} = f\left(\frac{Study\ time}{Necessary\ time}\right)$$

We can define validity indicators. Learning effectiveness [4] is a function of the amount of time and necessary time of learning behavior. Based on this basic theory, it is scientific to measure the validity of learning based on the learning time of students. According to the result of data analysis, most of the learning objectives are for some independent knowledge points. Obviously, the completion rate is not enough to assess the effective learning.

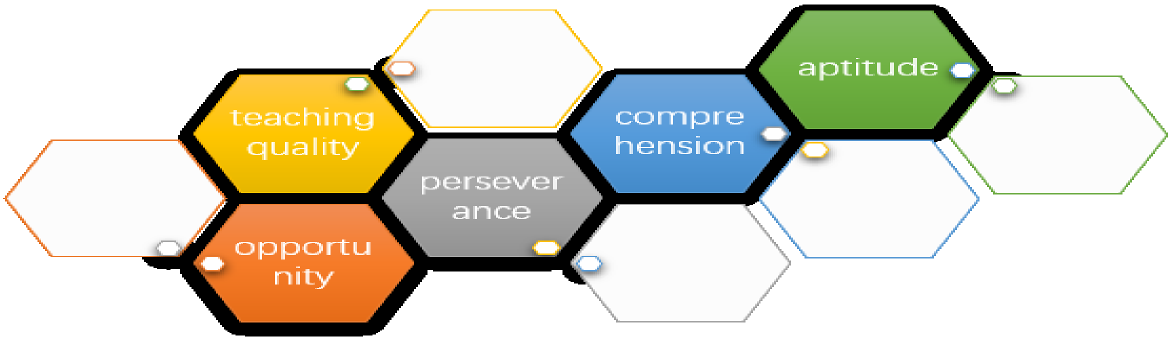

Fig. 1. Five factors of learning efficiency.

The time needed to understand knowledge is related to learning efficiency [4]. Learning efficiency is affected by the interaction of five factors. The five factors are: opportunity, perseverance, teaching quality, comprehension and aptitude, as shown in Figure 1. These factors affect the learning habits. Enables us to classify learners. By analyzing MOOC behavioral data, effective learning behaviors should include the following types.

Complete all course contents, pass the test, get the certificate.

Complete all course content, failed the test.

Learning a single point of knowledge, and complete the practice.

Continuous browsing the entire knowledge point video.

Completed the course test, and the result reached a certain standard.

Forum posts, discussions with teachers or learners to master certain knowledge.

Of course, we could not list all the validated learning behaviors. There are still some behaviors that can achieve the learning effect. And the impact of each learning behavior on the individual is not the same. More in-depth study of learning behaviors and their impact on learners are needed.

Therefore, it is reasonable to predict the learning effect based on learner's learning behavior data. In order to establish a learner's classification algorithm, we need to translate these effective learning behaviors into learner's attribute parameters. Assuming that there are m attributes per sample, these attributes can be denoted as $\{A_i\}$ $(i = 1,2,...,m)$. When using the naive Bayes method [12] for classification, we also need to use the training sample to get the prior probability. These training samples come from the Chinese University MOOC Advanced Mathematics course (https://www.icourse163.org/ learn/ 100 preview/ NUDT-9004? tid = 10002).

III. The Classification of MOOC Learners

Data show that the most important motivation for MOOC learner comes from obtaining knowledge and ability. The course certificate is more of an honor attribute. But to gain the course certificate is not the main target for them. There are many learners who do not even have the goal of learning. They are attracted to the new teaching mode. Most of the time spent in aimless browsing.

From the figure 2, we can see that more than half of learners (55.4%) learned for knowledge and skills. Another quarter of people look for challenges for themself. Obviously, we should design scientifically to meet the mainly needs of learners. It is the direction of development. Not all learners concentrate on the chosen course. It is possible that the chosen course was dropped because the course did not meet expectations or did not have a basic knowledge. Through the study of learning behavior, MOOC learners can be divided into four categories.

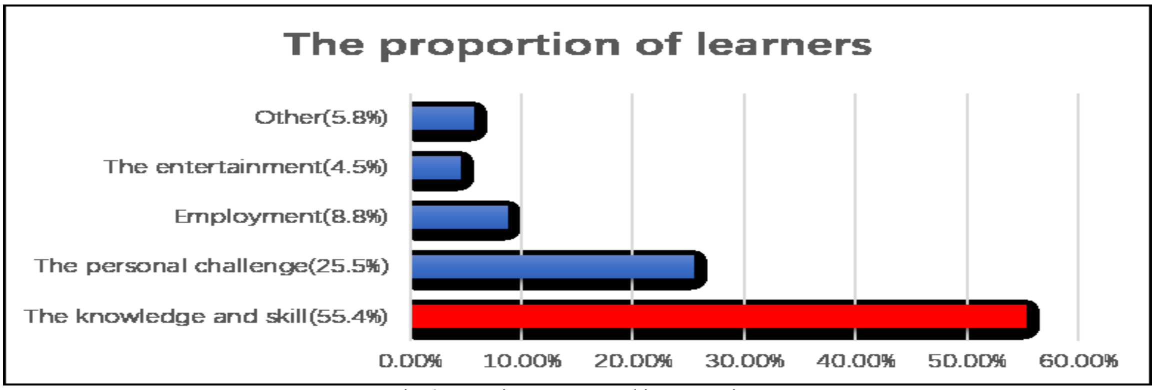

Fig. 2. Learning purpose and its proportion.

- Authenticator.

It is complete MOOC learner. Learning courses have a high interest and motivation. According to the requirements of the course to complete the corresponding learning objectives. It is divided into two types of active learners and passive learners. For active learners, they often have knowledge needs. They improve themselves by studying MOOC. Most passive learners are college students. Many colleges and universities now require students to register on the MOOC and learn some courses. The purpose of such learners is to earn course credits. So, they put in a lot of energy and time. In all learners, the learning effect is also the best.

- Practitioners

These learners are passionate about completing coursework and exercises. Such learners tend to have partial or complete mastery of the course content. Practice is to deepen the understanding of the knowledge. Many of these students come from flip classes. The teacher asks the student to complete the course related exercises. This is a typical pattern that MOOC resources are used in the classroom. This type of learner's goal is not the certificate. But they are easy to get it for having finished the exercise.

- Observers

Such learners do not have a clear purpose of learning. They just entered the course with curiosity. Their length of most study does not meet the necessary time standards. Based on effective learning theory, they cannot grasp the knowledge. These learning behaviors are invalid. The number of such learners is also very large. Of course, through the experience of MOOC, many visitor learners finally converted into other types. Therefore, it is very significant to make MOOC attractive.

- Visitors

They are interested in a part of the course. But there is no long-term learning behavior. Such learners are more purposeful. They often choose to learn specific knowledge through MOOC. The retrieval of knowledge resources is their greatest difficulty. The current MOOC tends to use chapters as video segmentation units. And lack of efficient knowledge retrieval ability. Such learners spend a lot of time in resource browsing. The current resource structure is not conducive to their learning. However, the number of such learners is huge. So need to improve MOOC learning mode and meet the needs of such learners.

This paper uses Naive Bayes method [12] to classify learners. The learner's classification set is denoted as $\{C_1, C_2, C_3, C_4\}$. The elements correspond to the learner's four types Authenticator, Practitioners, Observers, and Visitors, respectively. The samples to be classified are $X = \{x_1, x_2, \ldots x_n\}$. Each sample has 4 attributes, denoted as $\{A_1, A_2, A_3, A_4\}$. In addition, when classifying with naive Bayes, we need to assume that each attribute is independent. They can be represented using the following four sets.

A_1 = {Watch video rate}, A_2 = {Complete course work rate}, A_3 = {Participate in discussion times}, A_4 = {Tests core}

Attribute A 1 refers to the ratio of the number of videos the learner sees to the total number of videos. The meaning of attribute A 2 is the ratio of the number of learners completing coursework to the total number. A 3 and A 4 are the number of participating discussions and test scores, respectively.

The conditional probabilities $P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} (i = 1,2,3,4)$ are respectively calculated, in which the maximum probability is the corresponding class of the sample. Because the denominators of the four conditional probabilities are the same, we only need to calculate the numerator for comparison. In the numerator $P(X|C_i)P(C_i)$ used for comparison, $P(C_i)$ and $P(X|C_i)$ need to be calculated separately. These prior probabilities can be statistically obtained from the training sample set. In order to ensure the accuracy of the classification, our training sample set is the learning behavior data of the previous semester of the same course.

IV. Learner Characteristics

By analyzing MOOC data, we found some learning rules. Watching videos is the most important form of online learning. So, we analyzed the watch video data. This article chooses the Higher Mathematics (1) video viewing data on the Chinese University MOOC. Courses were opened a total of 5 times. There are 17557 video records. Each of them contains the learner's ID, video ID, and watch duration. After statistics, a total of 1851 learners watched the video. Accounting for 6.7% of the total enrollment. Among learners with learning behaviors, 42.3% watched the video. Watch video records can reflect the utilization of video resources.

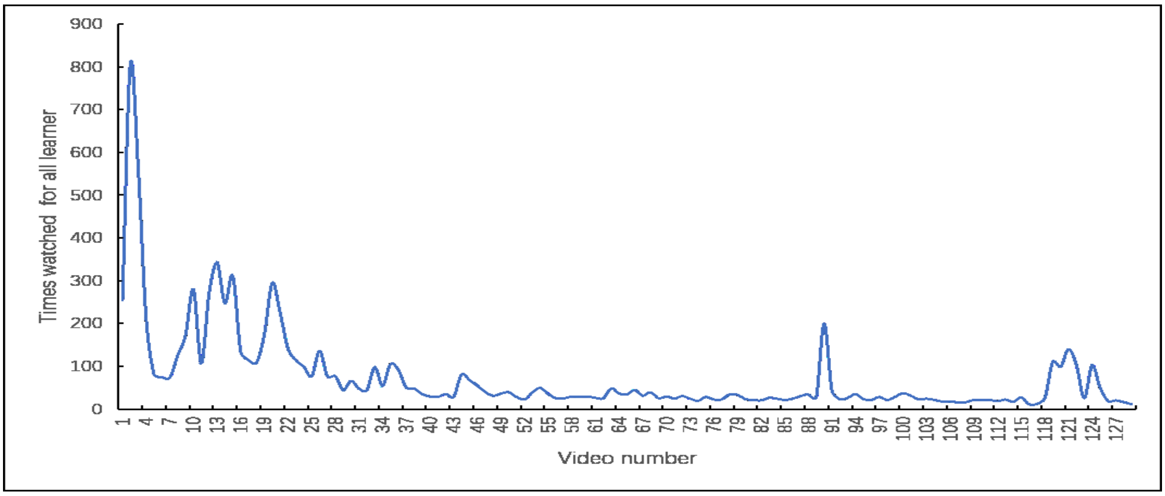

Fig. 3. Video watch record distribution.

As the Figure 3 shows, as the course progressed, the number of learners who watched the video dropped drastically. Select the first week of video (Just in front of 22 videos) learning records, display the watch video distribution in Figure 4. This phenomenon was found to be particularly noticeable.

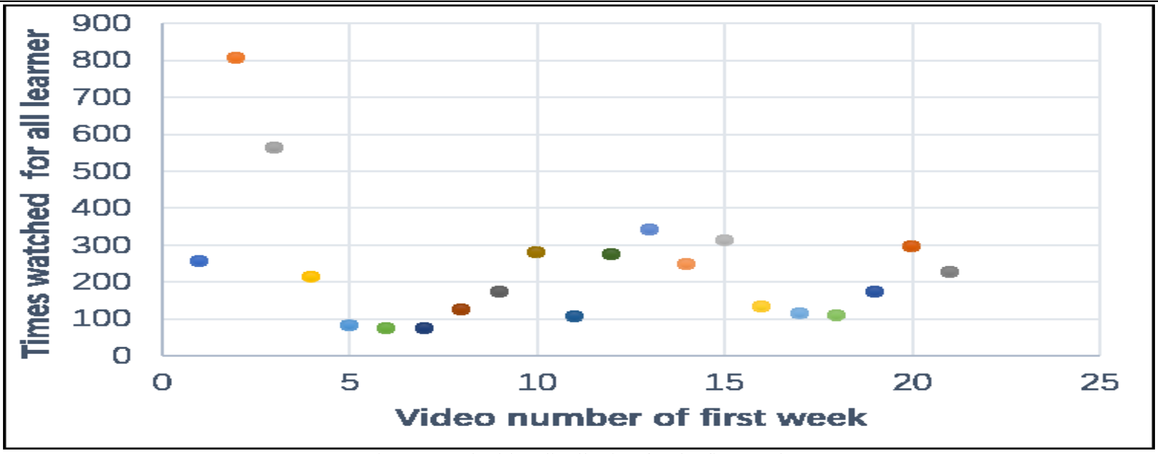

Fig. 4. Watch video distribution for the first week.

This phenomenon shows that most learners have only brief curiosity about online courses. Their enthusiasm for learning is also very fragile. Few learners insist on watching video. In a recent advanced mathematics course, data show that in the advanced mathematics course, only 18 people watched 50% of the course video. This is an unexpectedly low amount of data. Even lower statistical standards, only 408 people watched 10% of the course video. Data analysis shows that, few learners are willing to spend time watching the video. In this respect, we should make videos more appealing.

Later in the lesson, two videos were observed with a sudden increase in viewing from Figure 3. In Figure 4 there is also a corresponding performance. Analysis found that they are exercises counseling courses. It shows that students are more willing to watch tutoring videos. This phenomenon is consistent with the result of the paper [2]. From this phenomenon can also find that most learners are talking MOOC as a counselor. This is why Khan Academy's tutorial classes are so popular. Many MOOC producers do not change the teaching model for the rules of online learning. They focus on the recording environment and video clarity. Spent a lot of production costs, but no one is willing to watch.

In addition, we also analyze the situation of students taking the course exams. The subjects were learners who watched at least 10% of the video. Of the 408 students, 156 attend the final exam, accounting for 38.2% of the total. Of the 156 students who took the final exam, 57 people made outstanding achievements. And, 141 achieved academic credentials, the proportion of up to 90.4%. Accounting for 60.3% of all certified students. This data shows that as long as watching the course video, is likely to obtain a certificate.

Table 1. Different types of learners get the certificate ratio.

Category	Total	Examination	Passing	Certificate
Authenticator	67	100%	70.15%	95.52%
Practitioners	124	99.19%	68.55%	87.10%
Observer	196	17.86%	8.16%	9.69%
Visitors	3929	2.70%	1.09%	1.02%

We also research the percentage of different learner types getting a certificate. The purpose of each type of learning is different, so the proportion of certificates is different. However, it is unscientific to evaluate a course based solely on the acquisition of a certificate.

As can be seen from the table 1, the number of Visitors is the largest, accounting for 91% of all learners. But they get the lowest percentage of certificates. In contrast, the number of Authenticator and Practitioners is very small. However, they have a high percentage of certificates. For C, they also have effective learning. For Observer, they also have effective learning. The problem they face is the difficulty of using online open course resources efficiently.

V. Rop Out Factor Analysis Algorithm

Assume that $B_1, B_2, \cdots, B_n$ is a variety of causing dropout factors. First, statistical methods are applied to determine the prior probability. The prior probability is denoted as $P(B_i)$ $(i = 1,2,\cdots,n)$. Analyze the probability of occurrence of various dropout factors.

Let A be the event

$$A = \{\text{MOOC learners drop out}\}.$$

Table 2. Analysis of dropout factors.

Behavier	Total	Dropouts
Did not Watch the video	4199	3937
Did not Participate in the discussion	3837	3682
Did not complete assignments	3961	3914

Next, calculate the learners dropout probability for these factors in big data and determine the cause probability $P(A|B_i)$ $(i = 1,2,\cdots,n)$. Finally, apply the Bayes formula [8] to calculate the posterior probability $P(B_i|A)$ $(i = 1,2,\cdots,n)$.

$$P(B_i|A)=\frac{P(A|B_i)P(B_i)}{\sum_{j=1}^{n} P(B_j|A)P(B_j)}\ (i = 1,2,\ldots,n)$$

The drop-out factors that correspond to the large posterior probability are available to MOOC builders. These factors can be used as the focus of MOOC platform needs improvement. This method can also quantitatively present the problems with MOOCs.

For the same course, we did a Bayes factor analysis [11]. The number of students with learning behavior in this course is 4317. This article examines the impact of watching videos, participating in discussions, and completing assignments on the completion rate of the course. The number of students who did not have these learning behaviors and the corresponding number of dropouts were counted during the MOOC course.

We define the unfinished work as less than 30% of the total. Not participating in discussions, and not completing assignments were recorded as B_1, B_2, B_3. From the statistical data in Table 2, the probability of occurrence of the three factors is $P(B_1) = 0.973$, $P(B_2) = 0.889$, $P(B_3) = 0.918$. The prior probability is $P(A|B_1) = 0.938$, $P(A|B_2) = 0.960$, $P(A|B_3) = 0.988$. Not participating in the discussion parameter means that there is less than one posting. Participate in the discussion Because there is no total number setting, we define it as one. Failure to complete the assignment means that less than 7 assignments were completed in the entire course. The three factors of not watching videos, The Bayesian formula can be used to calculate the corresponding posterior probability.

$$P(B_1|A) = 0.341, ,P(B_2|A) = 0.319, P(B_3|A) = 0.340$$

From this result, it can be found that the most important factor affecting the dropout rate are watching videos and doing homework. However, because the learning behavior is linked together, these three factors have a similar influence on MOOC learning. But in comparison, participating in discussion has the least impact on students ability to persist in learning. And we found in the data that there are some outstanding students who do not even participate in any discussion. Therefore, course videos and exercises are the key to improving MOOC adoption rates.

VI. Summary

The rise of MOOC is to meet learners' needs for knowledge. After its popular teaching resources are shared, institutions of higher learning have aroused the interest of learners. However, with the further development of MOOC, the extremely high dropout rate of MOOC has become more and more obvious. On the other hand, MOOC's existing knowledge structure and learning model do not apply to all types of learners. Learners waste a lot of time browsing invalid resources. In order to improve effective learning, it is imperative to change the MOOC course knowledge organization model. Different learners design different types of learning modules.

The effective learning theory [4] ensures that the online learning time and the number of participants can reflect the learning effect. Then we list effective online learning behaviors and abstract them as four learner attributes. In this way of classification we consider the independence between the various attributes. Because this is a prerequisite for applying Naive Bayes [12] classification. In Table 4, we can see that there is a very significant difference in the pass rates of the four types of learners. This also shows that the classification algorithm is scientific from the perspective of probability.

In the drop-out factor analysis algorithm, we can find that watching videos and completing homework are the biggest factors causing dropouts. But from the final result, the difference between the three factors we analyzed is notsignificant. This shows that the lack of any kind of learning behavior will cause learners to give up learning. However, we believe that dropping out of school does not mean wasting time. The purpose of the vast majority of learners is not to pass the course exam. We should design corresponding learning modules for them to improve learning efficiency. Our classification algorithm can predict the type based on learner's early behavior, which will help improve the MOOC's service function.

In future work, we will make a more detailed analysis of the classification criteria for each type of learner. And we will also study whether the language used in the course affects the classification parameters.

References

[1] Panagiotis Adamopoulos. What makes a great mooc? an interdisciplinary analysis of online course student retention. In Proceedings of the 34th International Conference on Information Systems (ICIS13), 2013.

[2] Carlos Alario-Hoyos, PJ Mun˜oz-Merino, Mar P´erez - Sanagust´ın, C Delgado Kloos, G Parada, et al. Who are the top contributors in a mooc? relating participants' performance and contributions. *Journal of Computer Assisted Learning*, 32(3):232–243, 2016.

[3] Miri Barak, Abeer Watted, and Hossam Haick. Motivation to learn in massive open online courses: Examining aspects of language and social engagement. *Computers & Education*, 94:49–60, 2016.

[4] David A Cook, Anthony J Levinson, and Sarah Garside. Time and learning efficiency in internet-based learning: a systematic review and meta-analysis. *Advances in health sciences education*, 15(5):755–770, 2010.

[5] Martin Formanek, Matthew Wenger, Sanlyn Buxner, and Chris David Impey. Motivational differences between mooc and undergradu--ate astronomy students. In *American Astronomical Society Meeting Abstracts*, volume 231, 2018.

[6] John D Holt and Soon M Chung. Mining association rules using inverted hashing and pruning. *Information Processing Letters*, 83(4):211–220, 2002.

[7] Kate S Hone and Ghada R El Said. Exploring the factors affecting mooc retention: A survey study. *Computers & Education*, 98:157–168, 2016.

[8] WANG Hong-chun. Research on bayesian formula and bayesian statistics[j]. *Journal of Chongqing University of Science and Technology (Natural Sciences Edition)*, 3:065, 2010.

[9] Suhang Jiang, Sean M Fitzhugh, and Mark Warschauer. Social positioning and performance in moocs. In *Workshop on Graph-Based Educational Data Mining*, volume 14, 2014.

[10] Judy Kay, Peter Reimann, Elliot Diebold, and Bob Kummerfeld. Moocs: So many learners, so much potential... *IEEE Intelligent systems*, 28(3): 70–77, 2013.

[11] Igor Kononenko. Semi-naive bayesian classifier. In *European Working Session on Learning*, pages 206–219. Springer, 1991.

[12] David D Lewis. Naive (bayes) at forty: The independence assumption in information retrieval. In *European conference on machine learning*, pages4–15. Springer, 1998.

[13] Robert McGuire. The best mooc provider: A review of coursera, udacity, and edx. *Skilledup. com, June*, 19, 2014.

[14] Irving Pressley McPhail, Donna McKusick, and Al Starr. Access with success: The master learning community model. *Community College Journal of Research & Practice*, 30(2):145–146, 2006.

[15] Sarah EM Meek, Louise Blakemore, and Leah Marks. Is peer review an appropriate form of assessment in a mooc? student participation and performance in formative peer review. *Assessment & Evaluation in Higher Education*, 42(6):1000–1013, 2017.

[16] Laura Pappano. The year of the mooc. *The New York Times*, 2(12):2012, 2012.

[17] Trang Phan, Sara G McNeil, and Bernard R Robin. Students patterns of engagement and course performance in a massive open online course. *Computers & Education*, 95:36–44, 2016.

[18] Lloyd P Rieber. Participation patterns in a massive open online course (mooc) about statistics. *British Journal of Educational Technology*, 48(6):1295–1304, 2017.

[19] Amy E Stich and Todd D Reeves. Massive open online courses and underserved students in the United States. *The Internet and Higher Education*, 32:58–71, 2017.

[20] Bronwen J Swinnerton, Neil P Morris, Stephanie Hotchkiss, and James D Pickering. The integration of an anatomy massive open online course (mooc) into a medical anatomy curriculum. *Anatomical sciences education*, 10(1):53–67, 2017.

[21] Andrew A Tawfik, Todd D Reeves, Amy E Stich, Anila Gill, Chenda Hong, Joseph McDade, Venkata Sai Pillutla, Xiaoshu Zhou, and Philippe J Giabbanelli. The nature and level of learner–learner interaction in a chemistry massive open online course (mooc). *Journal of Computing in Higher Education*, 29(3):411–431, 2017.

[22] Alyssa Friend Wise, Yi Cui, Wanqi Jin, and Jovita Vytasek. Mining for gold: Identifying content-related mooc discussion threads across domains through linguistic modeling. *The Internet and Higher Education*, 32:11–28, 2017.

[23] Mingming Zhou. Chinese university students' acceptance of moocs: A self365 determination perspective. *Computers & Education*, 92:194–203, 2016.

AUTHORS PROFILE'

First A. Yong Luo Associate professor of National University of Defense Technology, received his PhD in 2004. He entered the postdoctoral mobile station of computer science at Central South University in 2010. Starting from 2018, he was invited by Warwick University to participate in the joint research of big data analysis in UK. The main research areas include online learning and big data analysis. At present, he has participated in building four MOOC courses in China University MOOC, and published 12 papers about MOOC. The Advanced Mathematics MOOC course was evaluated as a Chinese National Quality Online Course in 2017, and the Mathematics and Military MOOC course was evaluated as a Chinese National Quality Online Course in 2018.

Second B. Zhihao Xiao Who received his bachelor of science degree from the School of Mathematics of Shan Dong University in June 2017, is currently studying in the mathematics department of National University of Defense Technology. Won the national scholarship in 2016, and applied for the postgraduate research innovation project in Hunan province in October 2018. Research direction is network optimization. Currently, engage in the analysis of online learning behavior data in the context of big data.

Third C. Jianping Li Ph.D., professor of Mathematics Department, National University of Defense Technology, Member of the teaching guidance committee for mathematics majors in institutions of higher learning, ministry of education, 2018-2022, and executive director of national team of mathematics public course teaching. Currently, he is the vice President of hunan mathematics societ. He is the main lecturer of mathematical modeling and mathematical experiment for national high-quality courses and high-quality resource-sharing courses. Four courses were launched on the MOOC platform. Eight textbooks were published. Mainly engaged in network optimization and game research, hosted or participated in a number of natural science foundation, national defense pre-research projects, published more than 50 academic papers, including 3 SCI/SSCI zone I papers.

Fourth D Zheng Xie Research Interests: Geometric graph theory and its applications. Education experiences: Ph.D. Applied Mathematics, Chinese Academy of Sciences,2008; M.A. Computational Mathematics, Dalian University of Technology, 2005; B.S. Applied Mathematics, Dalian University of Technology, 2003. Employment: National University of Defense Technology, Department of Mathematics, Associate Professor, 2014- present; University of California at Los Angeles, Department of Mathematics, Visiting Scholar, 2018- present; University of Bath, Department of Mathematics, Visiting Scholar, 2014-2015; National University of Defense Technology, Department of Mathematics and Systems Science, Lecturer, 2010-2013; Zhejiang University, Center of Mathematical Sciences, Post-doctoral, 2008-2009.