Research on MOOC Learners’ Behavior in Discussion Area Based on Text Mining Technology

Jiong GAO, Jia-mei YU and Lu WANG
School of Mathematics, Southwest Jiaotong University, Chengdu, Sichuan, China

Keywords: Learners’ behavior, MOOC discussion forum, Text mining, Hidden Markov segmentation model, Cosine similarity of high dimensional space, K-Means Clustering.

Abstract. In recent years, MOOC, as a kind of online learning, is becoming increasingly popular among students. Learners communicate mainly through the discussion area while learning in MOOCs. Therefore, MOOC forums play an important role in learning. In this paper, the text mining is adopted to search for the speech data in the discussion area of 39 online learning course, the segmentation algorithm is used to study the emotion, frequency and quality of the speech, furthermore, K means clustering algorithm is designed to classify users by their performance. The results show that users with different identities have a wide range of roles in the discussion forum. Therefore, some users have higher quality of speeches, but the overall speeches is of poor quality. There is also a big gap between the numbers of speeches in various courses.

Introduction

The rapid development of technology in the field of Internet has greatly promoted the combination of education and the Internet. Massive Open Online Courses (MOOC) is a combination of the two. Online course website provides learners with a variety of learning resources, but also provides more statistics for the study of course learning. Internet technology provides an endless stream of power for the development of online courses.

Teachers dominate class in traditional education, while they are no longer the dominant role in MOOCs but learners' own learning ability [1]. In addition to the course video itself, forums serve as an important platform for learners and teachers to interact with each other and are of great value in terms of curriculum activity and study of learning behaviors. However, there are few researches on MOOC discussion areas at present, all based on the statistical data to study and make suggestions on the improvement of their construction, degree of interaction, participation[2], repetition rate of the theme posts[3], etc., but lack of text data research. Because forums are presented as textual information, research on discussion forums should not be limited to statistics. The text data should be studied under the premise of ensuring the activeness of the discussion area so that the value of the discussion area can be transformed from quantitative change to qualitative change.

Based on this, this paper makes a statistical analysis of the discussion area data based on the results of text mining. The DFS web crawler, the HMM based text tokenizer, the branch and bound method are used to mine the discussion area text data and convert the text data into keywords for statistical analysis. There are significant differences in the role of forum users and their knowledge reserves, subjective awareness and so on. Furthermore, the article study the number and quality of students' speech with the above algorithms, and then uses the K-Means cluster to divide the discussion area participants into four types. This article presents opinions for different types of participants so that issues can be more focused and resolved more clearly. Compared with previous studies, the article has the following highlights. First, the article selects the popular text data of the discussion area and transforms the text data into numerical statistics, which is more convincing and reasonable than the single numerical statistics. Second, the article uses the cosine similarity of high-dimensional space to replace the Euclidean distance, and reduces the index dimension to simplify the classification difficulty. Thus, learner speeches can be characterized more intuitively and the large sample classification can be more conveniently performed by K-means clustering.
Theoretical Review

In 1959, the famous sociologist Katz put forward the "use and satisfaction theory". In the history of research on the effect of mass communication, previous studies mainly focused on the communicators or the media to see whether the media has achieved its intended purpose or have an impact on the audience. In contrast, "use and satisfaction" research is based on the audience's perspective. By analyzing the audience's motivation for media contact and what they meet, these contacts examine the psychological and behavioral effects of mass communication and emphasize the audience's initiative. Therefore, this theory is often used to study emerging media such as the Internet. MOOC education is based on Internet communication technology. In forums, the number of conversations between students and teachers and the quality of the topics all has an impact on the MOOC's practice and development. Based on the above analysis, the article classifies students from the above two perspectives to study the role of different categories of learners.

The Number of Speaking

Learners and curriculum teams (teachers, teaching assistants, technical consultants) can communicate freely with each other in forum of MOOC platforms. Therefore, different actors play different roles, such as initiating a conversation (initiating a conversation), problem proponents (asking questions), and respondents (answering related questions or posting other comments). Different people, starting from their own needs, play different roles and communicate with each other in discussion forums, which in turn promote the activeness of discussion forums.

Analyzing the topic posts in the discussion area, we can find that the topic initiators are mainly divided into curriculum teams and learners. Teachers or teaching assistants initiate a conversation to explain teaching plan settings and course-related topics, thus initiating discussions for learners. During the discussion, teachers or TA will also pay attention to the performance of students and guide them, thus improving the learning effect. Learners often start their conversation after a period of study, and post some puzzles or feelings based on what they have learned. They will further resolve their doubts and deepen their understanding of knowledge through mutual discussion. During this process, the teacher or assistant also comments on the post and reply content. MOOC forum topics mainly related to learning, education management, MOOC platform, daily communication, chatting and other content [2]. Therefore, from the perspective of topic analysis, we can see that discussants tend to ask questions based on learning, educational management, and the use of MOOC platforms.

In the discussion area, participants exchange ideas and sublimate each other's ideas in different topic posts according to their own need, which enhances the learning effect and improves the activeness of the discussion area.

The Quality of Speaking

Due to the convenience and binding power of the network communication platform, participants in the discussion area have a variety of topics. In the MOOC forum, the number of topic posts accounts for the vast majority of the number of posts. The content of the topic involves learning, daily communication, chatting and other aspects. Therefore, the topic of the content is not the same. In the analysis of the quality of dialogues and speeches, the article divides it into discussion related to learning content, discussions about the arrangement of curriculums and discussion unrelated to learning curriculums based on the topic of the posts.

Divide the quality of the topic, we found that the content with higher quality concentrate mainly on learning puzzles, and other topics around the course. Among them, the topic of the extension of the course content helped learners most. In this kind of speech, learners can deepen their understanding of knowledge, develop their ability to learn independently, stimulate their interest in learning, and even gradually form a learning group to play the role of mutual supervision.

In contrast, the discussion of course settings is less valuable. At the same time, such speeches are prone to duplication of similar speeches, which has a certain negative impact on the participation of the curriculum team. In addition to the above two categories, other non-course-related speeches are...
of low quality, and such speeches will have a negative effect on the standard management, promotion and optimization of forums. Differences in the content of different speeches affect their quality, and thus have different effects on the learning outcomes of learners.

Data Source

This passage selects 39 courses and 1910429 discussion area data of 42521 learners in discussion forums from MOOCs in China. Firstly, we select 39 randomly selected subjects from the completed curriculums. Secondly, we use Web crawler technology to obtain forum page by DFS method, parsing user's speech record and store it in database. Finally, the user participates in the participle, constructs the sentence-word adjacency matrix according to the participle result, extracts the characteristic keyword by PCA method and marks the emotion label, and then calculates the number of times each user expresses the three emotion words. In addition, the number of speeches given by each student in each course forum is also counted.

In the text mining model design, the article chooses POS tagging method to split the adjacent sentence components. The parameters of Hidden Markov Model were calibrated by performing the part-of-speech tagging on the small amount of cleaned text data and using the method of maximum likelihood estimation. Compared with the maximum entropy algorithm, this method has higher requirements on the quality of manual marking, but the solution procedure is more efficient and concise. Based on the analysis of 42521 linearly independent 42-dimensional vectors \( P_1, P_2, \ldots, P_n \) \((n=42521)\) with known data structure, it can be seen that the average number of students in each course is more than 10 times and less than 100 times and the students are active. On the one hand, this reflects the convenience of online communication. On the other hand, the speaking level of all the members showed a right-deviation distribution. Individuals who are particularly active in their speeches will, to a certain extent, raise their average level. In addition, there is a clear gap between the average number of speeches in each course, which may be affected by the course attributes. However, the role of instructors in teaching cannot be ignored.

Mathematical Models

Hidden Markov Segmentation Model

The coherent statement is consist of different parts of speech words, a word is made up of one or more characters and in the words within the character position is fixed. The paper gets the rules as follows: first, as for the part of speech of any word, it only related to the part of speech of the previous character, therefore, the law of speech obeys Markov process; secondly, according to the specific character of CI with the words part of speech change, maintain independence. The hidden Markov model (HMM) can be used to deal with the word segmentation in the MOOC discussion area because of the above rules.

Based on the Bayes formula and the word law, to improve the quality of participle processing, the parameter being estimated by using maximum likelihood estimation (MLE), the paper uses the vector of words \( S=(w_0, w_1, \ldots, w_n) \) and the vector of their characters \( Y=(y_0, y_1, \ldots, y_n) \) in one sentence to establish the likelihood function as follows,

\[
L(\theta | D) = \prod_{j=1}^{N} \prod_{i=o}^{y_j} p(y'_i | y_{i-1}) p(w'_i | y'_i)
\]

According to the thought of maximum value, the paper uses the objective function of maximizing the logarithmic likelihood function.

Cosine Similarity of High Dimensional Space

For any two n-dimensional vectors \( Pi, Pj \), the paper first computes Cosine Similarity as a measure of the difference between the two vectors. The mathematical expression is as follows,
\[
\cos \theta_{ij} = \left( \sum x_{i1} x_{j1} \right) / \sqrt{\sum x_{i1}^2 \sum x_{j1}^2}
\] (2)

Among them, \(x_{i1}, x_{j1}\) are the ‘l’ elements in 1-dimensional vectors \(P_i, P_j\). The matrix \(P\) is constructed because of the linear independence of each n-dimensional vector.

\[
P = \text{Gram}(P_1, P_2, \ldots, P_n) = \begin{bmatrix} R_1 & P_1 P_2 & \cdots & P_1 P_n \\ \vdots & \vdots & \ddots & \vdots \\ P_n R_1 & P_n P_2 & \cdots & P_n P_n \end{bmatrix}
\] (3)

The angle cosine can also be written as,

\[
\cos \theta_{ij} = \left| P_{ij} \right| / \sqrt{P_{ii} P_{jj}}
\] (4)

Among them, \(P_{ij}\) is the algebraic cofactor of \(P_i, P_j\)

This paper studies the number of times and the quality of speech that students participate in the discussion in all courses, so we calculate the cosine value between each student's individual vector and the reference vector. Moreover, in order to express the distribution of all learners' angle \(\theta\) more intuitively, the maximum minimization is processed to the value in \([0,1]\), and the same expression is maintained with cosine value consistency.

\[
\rho_i = (\max \theta_i - \theta_i) / (\max \theta_i - \min \theta_i)
\] (5)

K-Means

Considering the relevant data of all the learners in the selected MOOC course in this study, the sample size is huge. In order to ensure the efficiency of computing, the K-Means clustering algorithm is selected, the main flow of the K-Means algorithm is as follows:

Step 1. Determine the \(K\) initial class center points. In order to ensure that the selected sample points are not influenced by the subjective idea, and to reduce the number of iterations in the later text as much as possible, therefore, the paper uses the minimum maximum method.

Step 2. Calculate the Euclidean distance from each sample point to the selected \(k\) center points of \(k\) classes, and according to the shortest distance principle, each sample point is allocated to the nearest class and all the sample points are divided into \(K\) clusters.

Step 3. Recalculate the mean of all samples in each cluster, as the new class center.

Step 4. Repeat the above 2,3 steps until all the sample points are not changed.

Usually, due to the huge amount of data, it is difficult to ensure that all samples completely offset, so when clustering points below a certain range, it is regarded that the clustering has accomplished. In addition, if the number of current iterations is too large and exceeds the specified maximum number of iterations, the algorithm will stop the update of the cluster and produce the final result.

Empirical Results

Based on \(\rho_1, \rho_2\), the paper analyses the number and the quality of speech, it’s obviously that, there are great differences between different students on both number and quality of speech. Therefore, based on \(\rho_1, \rho_2\) the paper uses K-Means clustering method to cluster the students, the students are divided into four types and we draw a two-dimensional diagram as follows. Among them, the abscissa and the ordinate represents \(\rho_1, \rho_2\) respectively, and each point represents a student.
Figure 1. Scatter diagram of clustering result.

About ‘Clustering-Learner-1’, the difference between the data and the reference sample on the number and quality of speech is large. It can be deduced that such kind of learners rarely participate in classroom discussions and after class discussions, and a few of discussions are quite different from the course contents. This indicates that such learners’ learning attitude is relatively negative and their learning quality is poor. Therefore, this kind of users are called Non-Participants.

Compared to Non-Participants, the \( \rho_1, \rho_2 \) of ‘Clustering-Learner-2’ values with the magnitude of the significant progress, they have some discussion in the discussion forum of their chosen courses, but not much, and the quality is common level. Although this kind of users are more positive in their learning attitude, but they seldom express their opinions voluntarily. Therefore, this kind of users are called Low-Level-Participants.

The ‘Clustering-Learner-3’ are more active in participating in the curriculum, and have the willingness of the curriculum, so this kind of users are called Core-Participants. Although their speaking quality is high, the proportion of these users is relatively small, so it is also difficult to play a significant role and they can’t greatly improve the status of the discussion area.

The percentage of ‘Clustering-Learner-4’ accounts for less than 1% of all participants in the discussion area. They choose more courses and maintain high quality speeches, which is also considered as a relatively professional or even authoritative person in related fields. These people hope to learn from other teaching experience, show their research results, and they help to answer other students' questions on curriculum content, which is called Professional-Participants.

**Conclusions and Suggestions**

Based on the study of the number and quality of users in the MOOC discussion area, the paper divides the users into four kinds: Non-Participants, Low-Level-Participants, Core-Participants, Professional-Participants. The specific recommendations are as follows.

First of all, MOOC is a platform for network learners to learn and communicate, therefore, its function construction is very important. Because some courses are highly professional and require software support, for the purpose of convenient communication, MOOC platform should attach importance. In addition, because of the high repetition rate of the discussion posts in many topics, the problem has brought a lot of inconvenience to the recovery work and the problem searching. Therefore, the MOOC platform should classify the topic of discussion more carefully, add the automatic filtering function of repeated topic and so on, so as to achieve better learning effect.

Secondly, teachers should play their guiding role and establish a proper system of reward and punishment in order to enhance the activity of the curriculum and improve the quality of the activities. The teachers can promptly reply the problem, and praise high quality and professional posts. At the same time, the discussion can be part of the curriculum assessment to encourage students to actively participate in discussions, and participate in it with high quality.

Finally, the learners should also actively share and communicate with each other. The significance of online teaching is not only to provide learners with opportunities to learn high-quality courses all over the world, but also to promote communication and interaction among
learners of different regions, different concepts and foundations. Online teaching also facilitates the combination of autonomous learning and cooperative learning.

Acknowledgement
This research was financially supported by the National Science Foundation, Projects in Sichuan Education Development Research Center of Key Research Base for Humanities and Social Sciences of Sichuan Provincial Education Department (Grant NO. CJF14014), Student Research Training Program in Southwest Jiaotong University (Grant NO. 201704008).

References