Advances in Computer Science Research, volume 80
International Conference on Computer Science, Electronics and Communication Engineering (CSECE 2018)

MOOC Course Evaluation Based on Big Data Analysis

Yong Luo[1,2,*], Jianping Li[1], Zheng Xie[1], Guochang Zhou[1] and Xiao Xiao[3]
[1]College of Science, National University Of Defense Technology, Changsha, China
[2]Hunan Provincial Key Laboratory of Network Investigational Technology, China
[3]Higher Education Press, Beijing, China
[*]Corresponding author

***Abstract*—The rapid development of MOOC, more and more same courses appear on the MOOC platform. For learners without the guidance of course selection, a lot of time wasted in the course to browse and try to learn. At the same time, lack of evaluation led to a decline in the quality of the course. In this paper, MOOC learning behavior data is used to construct a evaluation algorithm based on data distribution. Through theoretical analysis and data experiment, a standard model based on normal distribution is constructed. The evaluation algorithm is based entirely on learning data. Objective and dynamic real-time gives the standard points for each course. Not only help learner select the right courses, but also be able to promote the builders improve the lesson.**

Keywords—MOOC; big data analysis; course evaluation; normal distribution

I INTRODUCTION

The term MOOC (Massive Open Online Course) proposed by Dave Cormier and Brian Alexander. The earliest MOOC course was Artificial Intelligence by Stanford University professor Sebastian in 2011. The number of learners reached 160,000, and eventually 23,000 people completed the course. In 2012, Professor Schlumberger founded Udacity, a for-profit MOOC platform. Udacity, Coursera and Edx are the three most influential MOOC platforms in the world. MOOC brings a whole new paradigm to the sharing of higher education and resources. More and more universities participate in course construction. The number of courses offered is also rapidly increasing. Take the Chinese University MOOC as an example. By December 2017, the platform has 148 cooperative universities. A total of 1042 courses were opened. The number of similar or related courses is also on the rise. There are 325 courses related to higher mathematics in Chinese University MOOC.

After the learner comes into contact with MOOC, how to choose the course is the first step, and it is also the most crucial step. Because many same courses are hold on the same platform, learners often do not know which course to choose. But, the depth of these courses and teaching characteristics are different. In fact, students find it hard to find a course that suits their needs. They can only see the course profile. But these are not enough. MOOC course learners waste a great deal of time browsing and experimenting with the lessons. We found that a large number of students did not insist on learning, due not to choose the right courses. On the other hand, if the course does not have an objective and fair evaluation system, the builders did not promote it.

Therefore, how to recommend corresponding courses accurately and how to evaluate the courses objectively is a very important issue for MOOC. Today with big data support, we can use data processing techniques to analyze and evaluate lessons. Through the analysis of learning behavior data, to describe the characteristics and level of the course. This will improve the accuracy of student electives, improve student learning efficiency.

Currently all platforms are lack of a scientific course evaluation system. This article will establish a rating system based on big data statistical analysis. Through this, to provide learners with a comprehensive course information, as well as curriculum level indicators. Provide a reference for learners, and also promote builders to continuously improve the level of lessons. In summary, the evaluation system has a high application prospect. But also has some academic research value.

At present, the research on MOOC evaluation at home and abroad has just started. The National Center for Higher Education initiated the MOOC research project in 2017. Research and evaluate the courses in Chinese University MOOC. Expect to be able to make a scientific and reasonable evaluation of the course. To achieve the goal of raising MOOC level.

The United States, Britain, Japan and other economically developed countries have accumulated a great deal of experience in practice. In MOOC learning behavior data analysis, Anderson [3] found through data mining technology some of the factors that affect the learning effect of MOOC course. Adamopoulos [4] studied the factors that affect the retention rate of MOOC students. Gillani [5] analyzed "Business Strategy Fundamentals" course 87,000 trainees in the course forums and the relationship with grades. Guo [6] studied the relationship between MOOC video mode and learning effect through data analysis, and proposed a scientific and reasonable course video recording mode. Domestic MOOC teaching and research focused on MOOC in teaching practice and quality control. Jiang [7] classify learners according to the Chinese MOOC learning behavior characteristics. He studied in

depth the relationship between learning behavior and outcomes. Deng [8] paid attention to MOOC's problems in quality assurance and evaluation mechanism.

In summary, MOOC research has made a lot of achievements in both theory and application. But existing research rarely involves the assessment of MOOC course quality. In particular, there is a lack of quantitative evaluation of MOOC course quality.

II LEARNING BEHAVIOR DATA DISTRIBUTION

All set up similar courses MOOC behavior data for the overall ξ . Overall ξ follows a normal distribution. Each MOOC course learning behavior data as a sub-sample $\xi_1, \xi_2, ..., \xi_n$. By the sampling distribution theorem:

Let the overall ξ obey the normal distribution $N(a, \sigma)$, $\xi_1, \xi_2, ..., \xi_n$ is its sub sample, The mean and variance of the subsamples are denoted as $\overline{\xi}$ and S^2 respectively. Then $\overline{\xi}$ obeys normal distribution $N(a, \frac{\sigma}{\sqrt{n}})$, $\frac{nS^2}{\sigma^2}$ obeys the χ^2 distribution with degree of freedom $n-1$. Briefly written as: $\frac{nS^2}{\sigma^2}$: $\chi^2_{(n-1)}$, $\overline{\xi}$ and S^2 are independent of each other.

Orthogonal linear transformation of the sample is as follows:

$$\begin{cases} \eta_1 = \frac{1}{\sqrt{1 \cdot 2}}[\xi_1 - \xi_2] \\ \eta_2 = \frac{1}{\sqrt{2 \cdot 3}}[\xi_1 + \xi_2 - 2\xi_3] \\ \text{M} \\ \eta_{n-1} = \frac{1}{\sqrt{(n-1) \cdot n}}[\xi_1 + \xi_2 + ... + \xi_{n-1} - (n-1)\xi_n] \\ \eta_n = \frac{1}{\sqrt{n}}[\xi_1 + \xi_2 + ... + \xi_{n-1} + \xi_n] \end{cases}$$

The following shows that $\eta_1, ..., \eta_n$ is independent of each other and has the same normal distribution $N(a, \sigma)$. Due to orthogonal transform to maintain the same length, so

$$\sum_{i=1}^{n} \eta_i^2 = \sum_{i=1}^{n} \xi_i^2$$

Notice that (ξ_1 , ξ_2 , …., ξ_n) is a joint distributed n-dimensional normal. Its joint density function is

$$f(x_1, ..., x_n) = (\frac{1}{\sigma\sqrt{2\pi}})^n \prod_{i=1}^{n} e^{-\frac{x_i^2}{2\sigma^2}}$$

$$= (\frac{1}{\sigma\sqrt{2\pi}})^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2}$$

Since the absolute value of the determinant is equal to 1, the union density function of ($\eta_1, ..., \eta_n$) is

$$g(y_1, ..., y_n) = (\frac{1}{\sigma\sqrt{2\pi}})^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n} y_i^2}$$

Then the joint distribution of ($\eta_1, ..., \eta_n$) is an n-dimensional normal distribution. And in the normal case, irrelevance and independence are equivalent. Thus $\eta_1, ..., \eta_n$ is independent of each other and has the same normal $N(0, \sigma)$ random variables.

When $\eta_n = \sqrt{n}\overline{\xi}$, $\eta_n^2 = n\overline{\xi}^2$, that is

$$nS^2 = \sum_{i=1}^{n} (\xi_i - \overline{\xi})^2 = \sum_{i=1}^{n} \xi_i^2 - n\overline{\xi}^2$$

$$= \sum_{i=1}^{n} \eta_i^2 - \eta_n^2 = \sum_{i=1}^{n-1} \eta_i^2$$

Then

$$nS^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n-1} \eta_i^2 \sim \chi^2_{(n-1)}$$

And nS^2 and η_n are independent of each other. That is, nS^2 and $\overline{\xi}$ are independent of each other.

From the above analysis, when there are more courses to join the MOOC platform, new courses can be regarded as a sub-sample ξ_1 , ξ_2 , …., ξ_n of all MOOC courses ξ . Because of the overall MOOC learning behavior data is subject to normal, the data of new course also obey the normal distribution. Thus, according to the sampling distribution theorem, when the number of similar MOOC courses is small, its learning behavior data also obeys normal distribution.

III COURSE SCORE MATHEMATICAL MODELING

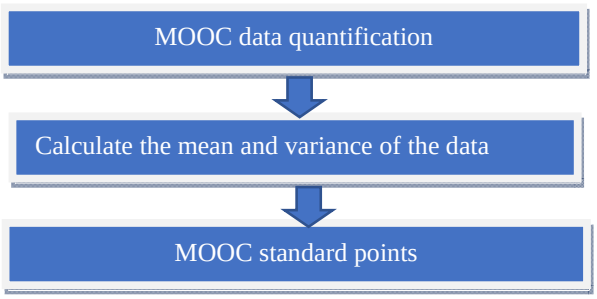

FIGURE I. DATA PROCESSING FLOW CHART

Data processing as shown above. First, quantify learning behaviors and course data. Calculate the mean and variance of the data for the same course. Get the 0normal distribution of parameters. Finally calculate the standard points.

There is an objective difference between each type of course and each type of learning behavior. And as the annual course joins, the size of the data is constantly changing. Affected by these factors, the same behavior data, at different times and different types of grading results are different. This reflects the objective and dynamic score. The algorithm can achieve real-time data calculation. And realize the equivalent calculation between different data. Based on the data approximation to normal distribution conditions, the following standard algorithm is proposed.

Convert each quantized MOOC data score to a standard normal distribution. Then use the same method to expand and translate to complete the equivalence of different learning behaviors. The specific method is described as follows:

There are n identical courses, choose a behavior data X. It obeys normal distribution. That is $X: N(a,\sigma^2)$. $a=\dfrac{\sum_{i=1}^{n} X_i}{n}$ is the average of all courses data. $\sigma=\sqrt{\dfrac{\sum_{i=1}^{n}(X_i-a)^2}{n}}$ is the standard deviation. If the data for this course is X. The standard formula is:

$$T=100\cdot\frac{X-a}{\sigma}+500$$

To improve the resolution of the data, the data is magnified and panned. 100 here to enlarge parameters, 500 for the translation parameters. Through this formula can be seen, equivalent points can be decimal, such a score more accurate.

In addition, if $T>900$, referred to as $T=900$; if $T<100$, referred to as $T=100$; Therefore $T\in[100,900]$. This equivalent process is actually based on a certain approximation. In fact, the standard normal distribution of the density function is:

$$p(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-a)^2}{2\sigma^2}}$$

Due to $X: N(a,\sigma^2)$, $\dfrac{X-a}{\sigma}: N(0,1)$. And $T\in[100,900]$, then $\dfrac{X-a}{\sigma}\in[-4,4]$.

$$P=\int_{-4}^{4}p(x)dx=0.9999$$

This shows that $\dfrac{X-a}{\sigma}$ falls in the interval $[-4,4]$ probability of 99.99%. Therefore, the above approximation is reasonable.

IV DATA EXPERIMENT AND ANALYSIS

We will test the distribution of MOOC learning behavior data. We choose the Chinese University of MOOC platform learning behavior data to conduct research. The data type is Advanced Mathematics learning behavior and course data. A total of four types of data were studied. Respectively, the number of elective courses, video traffic, the number of forum posting and the number of people who obtained the certificate. We normalize the data to [0,900] this data interval. Then calculate the mathematical expectation and standard deviation for each class of data for all classes. As shown in Table I:

TABLE I. EXPECTATIONS AND STANDARD DEVIATIONS OF LEARNING BEHAVIOR DATA

Data category	Mathematical Expectation	Standard Deviation
Elective number	407.1587	104.7397
Video traffic	406.0047	106.9768
Forum posting	398.6446	114.2184
Certificate number	396.3465	105.4982

The data in the above table is the parameter for calculating standard points. In order to analyze the distribution of data, it is necessary to draw a distribution image of the behavior data. This article is achieved by grouping the data. Calculate the frequency of each group of data. Describe the distribution of MOOC learning behavior data. In the experiment, our segmentation interval is the standard score of 20. The frequency is the number of courses falling within this range of scores.

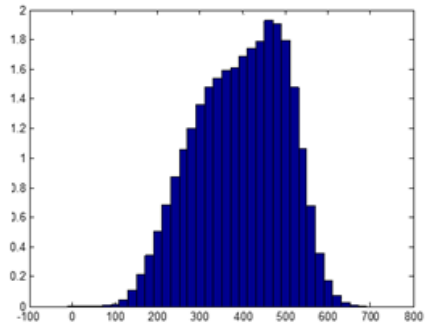

FIGURE II. MOOC FORUM POST DATA DISTRIBUTION

The figure is the equivalent distribution of MOOC behavioral data. From the learning behavior data image, the distribution image is very close to the normal distribution. It shows that MOOC behavior data is normally distributed. Our standard method of division is scientific and rational.

V SUMMARY

In this paper, we build a dynamic assessment system. The entire evaluation system is not subject to man-made subjective factors. Through the theoretical analysis and experimental research, the law of MOOC learning behavior data was obtained. These rules promote the construction of MOOC curriculum. Evaluation can motivate builders of MOOC courses to improve the quality of the course. The ultimate realization of the healthy development of MOOC.

REFERENCES

[1] Guidelines on the Quality Assurance of Distance Learning. Quality Assurance Agency for Higher Education.1999:88-101

[2] Phipps & Merisotis. Quality on the Line:Benchmarks for success In Internet –Based Distance Education. Research report of The Institute for Higher Education Policy. 2000.46-53.

[3] Anderson, D. Huttenlocher, J. Kleinberg, et al. Engaging with massive online courses, 2014:687-698.

[4] P. Adamopoulos, What makes a great MOOC? An interdisciplinary analysis of student retention in online courses. In Proceedings of the 34th International Conference on Information Systems, 2013, ICIS'13.

[5] N. Gillani. Learner communications in massively open online course. OxCHEPS Occasional Paper, 2013:53.

[6] P. J. Guo, J. Kim, R. Rubin. How video production affects student engagement: an empirical study of MOOC videos ACM Conference on Learning @ Scale Conference. ACM, 2014:41-50.

[7] Z. X. Jiang, Y. Zhang, X. M. Li, Learning Behavior Analysis and Prediction Based on MOOC Data, Journal of Computer Research and Development, 2015, 52(3):614-628. (In Chinese)

[8] H. Deng, M. Li, Y. Chi. Courses and knowledge system network modeling in MOOC. Course Education Research, 2013 (7), 5-7. (In Chinese)