Analysis of learning behaviour based on edX open data

Jian Feng and Yupeng Zhao

Computer Science & Technology, Xi’an University Science and Technology, Xi’an, Shaanxi, 710600, China
fengjian@xust.edu.cn

Abstract. Since 2012, the continuous development of MOOC has triggered new thinking and research in many aspects, and the massive data it generates provides a foundation for learning behaviour analysis and education data research. edX is a new field of Web education jointly developed by Harvard University and MIT integrating educational resources. Based on the first MOOC open data set released by edX, through the use of data analysis methods, the learners in edX are studied from the three aspects of learner type, learner characteristics, and learner behaviour to show the actual situation of MOOC learning.

1. Introduction

With the rapid progress of modern information technology, network education as a new type of education has been recognized by more and more people. In 2012, MOOC (Massive Open Online Course), as a large-scale online open classroom for the masses, had a huge impact on higher education. More and more domestic and foreign universities began to launch their own MOOC platforms, among which edX, Udacity, and Coursera are the three most famous MOOC platforms.

The edX platform is jointly built by the Massachusetts Institute of Technology and Harvard University based on the MITx project of the Massachusetts Institute of Technology [1]. It provides a platform for the public to open cross-disciplinary, high-standard free courses, and also for other institutions of higher learning. He Education Organization provides a platform for educational reform and innovation, and continuously promotes the development of world education.

Nowadays, more and more Chinese learners have joined MOOC courses. Therefore, based on the open data of the edX platform, this article will analyze and study the scholars of the edX platform to explore the learning characteristics and behaviour of MOOC learners.

2. EdX open data overview

In 2014, MIT and Harvard University jointly released a public data set of 16 courses they opened on the edX platform. This data set is the first publicly released MOOC study in the world. Data, and then lay a solid foundation for the analysis and research of learning behaviour.

2.1. Basic data

The entire data set includes the learning data and related documents of 16 courses opened on the edX platform by two schools in the fall of 2012, spring of 2013 and summer of 2013 [2]. There are 641138 records in total, of which 302915 are from Massachusetts Institute of Technology and 338223 are from Harvard University. Each record refers to the learning data of a learner in a MOOC course, which includes up to 20 items such as the course name, learner ID, gender, nationality, date of birth, and educational background.
2.2. Data processing
In order to ensure that the data has higher accuracy, Excel2010 and other tools are used to perform operations such as blanking, missing values, and outlier filtering on the data, which provides a solid foundation for further research and exploration.

At the same time, in the information age with the rapid development of science and technology, the data on the Internet contains a large amount of privacy for users. Therefore, the protection of privacy data has begun to receive attention and received legal protection. All open data must undergo a De-identification process and can be released publicly after privacy protection processing.

MIT and Harvard University’s public edX platform learning data is through the process of de-identification, and uses K-Anonymity, standard identification attributes, Quasi-Identifies, and L-diversification. Diversity) and other technologies process the data to protect user privacy without affecting the overall research and analysis of the data.

3. Data analysis
This article is based on the edX open data set, through the three aspects of curriculum design, learner type, learner behaviour, research and analysis of all aspects of learning behaviour of MOOC learners.

3.1. Course design analysis
EdX is one of the three major MOOC platforms in the world, and its courses cover all aspects. From the 7 courses at the time of creation, to 150 courses in 2014, to 475 courses in 2015, the course system of the edX platform has been continuously expanded and improved. The edX platform categorizes the platform’s course system from five aspects: course language, course type, course subject, course status, and course level. It is convenient for users to retrieve and view, and it is also convenient for data statistics and analysis.

3.1.1. Course language. In terms of course language, the edX platform divides the MOOC course into 6 parts: English, Spanish, Chinese, French, Turkish, and Hindi [2].

It should be noted that although the courses on the edX platform include teaching methods in multiple languages, it can be seen that most of the courses still use English as the main teaching method, even if they are courses in other languages, they still contain a large number of English elements. Although some English courses are accompanied by subtitles in other languages, excellent English listening and writing skills are a must for every learner who is learning on the edX platform.

3.1.2. Course type. In terms of course types, the edX platform divides MOOC into 5 parts: verification courses, college courses, Xseries courses, professional education courses, and credit courses [3].

3.1.3. Course status. In terms of course status, the edX platform divides the MOOC course into history, under opening, future opening, upcoming course, and self-controlled speed course [4].

3.1.4. Course level. In terms of course level, the edX platform divides the MOOC course into three parts: elementary, intermediate, and advanced [5]. Most of the courses are elementary, and this is more in line with the online teaching model, because most learners read the instructional videos by themselves for self-study, so everyone’s knowledge is not at the same level, so the difficulty of the elementary courses is more suitable for most learners.

3.1.5. Course subjects. The courses and disciplines included in the edX platform are quite comprehensive, including 29 aspects such as technology, economy, culture, and society. The types of disciplines are rich and varied, and the fields involved are also very wide. However, because the MOOC courses on the edX platform are intersected by different disciplines during the design process, it is impossible to accurately classify them, so the table design is not carried out here.
The public data set released by the edX platform this time includes 16 courses, including the choices of learners of different genders and different learning subjects, and the proportion of learners of different subjects to obtain certificates, see Figure 1.

Figure 1. Course selection and certificate acquisition.

From Figure 1, among the number of learners in different courses, the number of male learners is significantly greater than the number of female learners. Among them, male learners are more inclined to choose science and engineering courses such as “Computer Science Overview” and “Introduction to Computer Science and Programming”, while female learners prefer to choose “Ancient Greek Heroes”, “Human Health and Global Environmental Changes”, etc. Liberal arts courses. This phenomenon may be related to the different hobbies and thinking styles of men and women.

At the same time, it can be seen that the certificate acquisition ratio for most courses is between 3% and 5%. Among them, the number of learners in “Computer Science Overview” is the largest, but the certificate acquisition rate is only less than 1%. “Challenge” course, although the number of learners is not very large, the certificate acquisition rate is as high as 8%. The reason for this phenomenon may be related to the difficulty and popularity of the courses.

3.2. Learner type analysis
The data set jointly released this time contains 641,138 learning records of learners, including records of people of different nationalities, different genders, different educational backgrounds, and other differences, and this section will also proceed from these aspects. Research and analysis.

3.2.1. Nationality analysis. Due to the differences in the development level and education level of different countries, the education level of learners is different, which in turn leads to differences in their learning abilities and levels. The statistics and analysis of the number of learners in 14 countries are used to obtain the distribution of learners worldwide, see Figure 2.

Figure 2. Distribution of learners.
From Figure 2, the ratio of male to female learners in different courses in different countries is basically similar. The number of male learners is far greater than the number of female learners. The reason for this phenomenon is related to the proportion of male and female populations in different countries. There is a close relationship between the different education levels of men and women.

The average age of learners in Western developed countries such as the United Kingdom and France is over 27 years old. The average age of learners in the United States is even as high as 32 years old. In developing countries such as China and India, the average age of learners is mostly 23 years old. And 25 years old. It can be seen that, for most people in China, the act of learning is only carried out at the student stage, and self-learning is rarely carried out after work. For people in developed countries, learning may be more of a single person. A daily behavior. Of course, this phenomenon may be related to the age ratio of the population in different countries, but this is still a question we need to think about: do we just treat learning as a task, not as an interest or even a kind of By nature, what impact will this have on the development of our country and the development of individuals?

3.2.2. Gender analysis. Through research and analysis of the released data set, it can be concluded that in almost all countries, the number of male learners is greater than the number of female learners [7], see Figure 3.

![Figure 3. The proportion of female learners.](image)

It can be clearly seen from the figure that the number of female learners studying on the edX platform in almost all countries accounts for less than 50% of all learners. Among them, the proportion of female learners in China is 28%, which is lower than the average. The value is basically flat. However, in other countries, such as Greece, the Philippines, and the United States, the number of female learners is as high as 47%, 41%, and 35%, which are more than 1/3. Here we combine the global gender gap index to compare, analyse the reasons for this phenomenon from the aspects of education, economy, politics, etc., and conclude that the number of female learners is related to the reform of the country’s overall female education procedures.

Through this conclusion, we need to arouse our thinking. What is the reason that the number of female learners in our country is only 28%, and whether our traditional ideas need to be improved urgently.

3.2.3. Educational background analysis. Through the statistics of the data set, among all the learners who have obtained the certificate, there are mainly 5 parts with lower education level, high school education level, undergraduate education level, undergraduate education level, master’s degree, and doctoral, see Figure 4.
Figure 4. Percentage of certificate-obtained learners.

Through the analysis of the data on the certificates obtained by learners in 16 courses, it is concluded that in social disciplines, most of the learners who obtain certificates have bachelor’s or master’s degrees. In practical science and engineering courses such as computer, mechanics, and chemistry, most learners who obtain certificates have only high school degrees. Based on this, the reason for this phenomenon is that after students have finished their high school studies, they can learn more on the campus according to their hobbies, so choose the edX platform to study the courses they want to learn in advance carry out self-study to increase your knowledge reserve and improve your learning ability.

3.3. Learner behavior analysis

The edX platform saves and conducts research on the learning records of each learner, including course visits, video playback, and forum posts. After screening, sorting, and analysing these behaviour data, they will learn the behaviour of students is divided into the following four types: behaviourism, humanism, motivation theory, and master learning theory.

3.3.1. Behaviourism. Behaviourism believes that the learning process of learners is the behaviour generated by external learning materials, while ignoring the role of the brain’s thinking process. The model of this learning behaviour is “deliver-receive”. The specific process is: review-stimulate learning-learn new lessons-practice-check and consolidate-regular review. The following principles must also be observed:

- Small steps: refers to the decomposition of the curriculum into small parts, each part is step by step, closely linked, doing so will help learners to learn more effectively.
- Positive response: when learners react to the learning content, they will receive a certain reward. The purpose of this is to strengthen the learner’s memory of the knowledge point.
- Immediate feedback: let learners know whether their ideas are correct in the first time.
- Self-paced: allow students to arrange their own study plan according to their actual situation.

3.3.2. Humanism. Humanism believes that every learner is a unique individual. In the process of learning, the potential of each learner should be fully stimulated, and learning should be regarded as an organic process. The mode of this learning behaviour is “self-study-tutoring”. The specific process is: self-study-discussion-enlightenment-summary-practice-consolidation.

- Advantages: The learning behaviour in this mode can fully stimulate the learner’s subjective initiative, cultivate the learner’s ability to learn independently, cultivate the learner’s exploration spirit, and improve the learner’s communication and cooperation ability.
- Disadvantages: when the learner’s autonomy and self-control ability are insufficient, it will affect the learning efficiency. At the same time, it also has higher requirements for the teacher’s teaching ability. If these two points dissatisfactions at any point will have a deep impact on the final learning result.
3.3.3. Motivation theory. Motivation theory believes that when learners choose the edX platform to learn, they are for a certain purpose, such as: subject certificate, curiosity, satisfaction and so on. For learners supported by this theory, as long as they can achieve their ultimate goal, sometimes the process is not important. For example, for learners who want to learn knowledge, it may not be important to obtain a certificate in the end, even for a large part. For learners, choosing MOOC to study is just to experience the freshness of online teaching.

3.3.4. Master the learning theory. Mastering the learning theory believes that when enough time and the best conditions, the learner’s academic performance can reach the best. Adequate study time and good teaching methods are considered necessary conditions that can determine academic performance:

- Learning time: When the teaching method is fixed, the learning time is the main reason to control the learner’s performance. As long as the learners are given sufficient learning time and practice time, more than 90% of the learners can achieve good results.
- After the learners have studied, whether they are asking questions or conducting tests, they should receive timely feedback. Teachers should give corresponding guidance to students’ problems and give feedback to students’ test results.

Only by satisfying the above points at the same time, can the learners achieve the best performance.

4. Summary and outlook

Since its establishment, MOOC has a history of 8 years of development. Regardless of the course content, teaching system, number and scope of learners, it has been greatly improved and perfected. This article analyses the data set disclosed by the edX platform in three aspects: curriculum design, learner type, and learner behaviour, to study the meaning, advantages and disadvantages of MOOC teaching methods, and to explore the relationship between “teaching” and “learning”. Relevance, and put forward questions that need to be considered, so as to promote the development and improvement of China’s MOOC teaching model and overall teaching quality level, so that MOOC gradually becomes a learning method that Chinese learners really adapt to, and then learners have a positive learning concept, Correct learning behaviour and excellent learning ability.

Acknowledgments

This work was financially supported partly by the Educational Reform and Research Project of Xi’an University of Science and Technology (No. JG18056), the Education and Teaching Reform and Quality Engineering Cultivation Project of College of Computer Science and Technology (No. 2020-KCSZ-02 and 2020-MOOC-01), Online and Offline Hybrid Teaching Reform Project of Computer Education MOOC Alliance of China Universities (No. B190176), and First-class Offline Undergraduate Courses (No. 10).

References

[1] HarvardX and MITx. (2014) The First Year of Open Online Courses. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2381263.
[2] HarvardX-MITx Person-Course Academic Year 2013 De-Identified dataset. (2019) http://thedata.harvard.edu/dvn/dv/mxhx.
[3] Jiao, J.L. (2013) MOOC Global Action. China Education Network, 9: 32.
[4] Li, J.Y. (2013) The idea behind MOOCs. China Education Network, 4: 39-41.
[5] Wang, Z.L. (2013) MOOC: Is an Education Strom Coming. China Education Network, 4: 11-15.
[6] Cai, W.X., Wang, Q. (2013) 2012: The First Year of MOOC. China Education Network, 4: 16-18.
[7] Emma Pierson., Chuong(Tom) Do. (2014) What About the Women, http://tech.coursera.org/blog/2014/03/08/what-about-the-women.