

An Association Rule Model of Course Recommendation in MOOCs: Based on edX Platform

Min Tan

Maoguo Wu

SHU-UTS SILC Business School, Shanghai University,
Jiading District, Shanghai, China

Doi:10.19044/esj.2018.v14n25p284 [URL:http://dx.doi.org/10.19044/esj.2018.v14n25p284](http://dx.doi.org/10.19044/esj.2018.v14n25p284)

Abstract

The rapid development of the Internet has promoted online education as a new method of education. This paper analyzes the edX platform from various aspects, such as business environment and industry development, technology maturity, and recommendation algorithms and their feasibility. This paper mainly focuses on the curriculum recommendation mechanism under association rules on the edX platform. With the fast development of online education platform, curriculum recommendation technology has matured to a high extent. The recommendation mechanism based on association rules chosen by edX also proves to be feasible at data acquisition level and technical level. However, some defects of this mechanism per se have brought certain risks and limitations to edX's future sustainable development.

Keywords: Online Education, Course Recommendation, edX, Association Rule

I. Introduction

With the rapid development of technology and the urgent need to build a learning society, online education has become a new model of education and one of the most important ways to learn professional knowledge. Compared with traditional learning methods, online education has many advantages. On the one hand, users can engage in repetitive learning of the course; on the other hand, users can get rid of time constraints in traditional learning and enjoy strong learning flexibility.

Massive open online courses (MOOCs) have experienced an exponential growth after it was created. A large number of institutions have started offering courses free of charge. MOOCs enable institutions to expand their influencing by reaching a much wider range of students from different backgrounds rather than using a traditional mode of teaching.

However, due to its short history of development, uncertainties for institutions about offering MOOCs have gradually appeared. An increasingly important question that has been frequently raised is how MOOCs can effectively deliver courses.

Coursera, Udacity and edX are some representative platforms of MOOCs, the acronym of massive open online courses. Each course of MOOCs has a systematic curriculum system, and it integrates multiple forms of teaching methods such as online-work, online-test and group discussion. MOOCs has its own unique industrial value chain, which is based on those MOOCs operating agencies, and dynamically associated and co-evolved by entities such as basic equipment providers, course and service providers, and course assistance providers.

The profit of MOOCs comes from diversified paths, mainly from learners and companies. Table 1 compares profit methods of 3 famous MOOCs. MOOCs' appearance is a kind of value creation, which not only greatly reduces the learning cost of learners, but also helps companies to develop training courses for their employees, evaluates and recruits talents, and directly reduces time-consuming and expenditure problems to accelerate the pace of talent excavation. The whole education industry has realized MOOC's potential business value, and taken steps to help improve those MOOCs platforms.

TABLE 1 COMPARISON OF PROFIT METHODS BETWEEN 3 FAMOUS MOOCs

edX	Fees of taking exams and obtaining a course certificate from students
Coursera	<ul style="list-style-type: none"> ● Fees of taking exams and obtaining a course certificate from students ● Expenses paid by employers to identify excellent students through the platform ● Sponsorship ● Tuition fees
Udacity	<ul style="list-style-type: none"> ● Fees of taking exams and obtaining a course certificate from students ● Expenses paid by employers to identify excellent students through the platform ● Charges of the matching service between students' resumes and job opportunities

It is therefore important for institutions engaged in MOOCs to find how courses can be best delivered in this new learning environment for effective student learning. This paper attempts to answer this question by proposing an association rule model of course recommendation in MOOCs based on edX platform. The remaining part of this paper is organized as follows. Section 2 provides technology maturity analysis. Section 3 introduces method design. Section 4 assesses the feasibility of association rules algorithm. Section 5 concludes the paper.

II. Technology Maturity Analysis

Online education in the era of big data was not able to display the advantages of massive data at the beginning, because users are doing more data searching tasks like searching for needles in the sea. The pattern of ‘people looking for information’ is no doubt much difficult than ‘information to find people’. If the data of user’s access log and registration information are accumulated in those online education platforms and made the most use of, the user’s browsing preferences and learning habits can be analyzed to form a personalized customization. It can greatly increase the efficiency of online education and solve the problem of ‘information overload’.

From a technical point of view, data mining can be defined as an operation that extracts useful and valuable information from those vague, noisy, massive, random, and incomplete data. From a business perspective, data mining is a business information processing technology that can assist in the formulation of business decisions. The discovery of sales associations between beer and diapers is a typical example of data mining under the association rules. Thus, while massive educational resources are placed on the online education platform, the data mining technology also needs to be put into practice.

Research on educational data mining has become a trend in research on online education. There has been a great deal of achievement in data mining after digging into this field and putting the theory into practice. In 2012, the U.S. National Ministry of Education released the report ‘Educating and Learning through Educational Data Mining and Learning Analysis’, which provides a detailed introduction to the fields and cases of the application of big data education in the United States and the challenges it faces.

For improving personalization of the online education platform, the most popular method is to use Bayesian network to establish a student learning model and return the students’ learning evidence encountered in the platform to the system. And then, apply the probabilistic reasoning and make predictions of students’ learning behavior. In China, the application of data mining technology has moved from pure research to products development. With the continuous improvement of the market economy of China, the market demand for data mining is growing at a high speed. Besides, there has been positive attitude towards the development and application of data mining technology in online education.

When accessing the online education system, the user generates corresponding ‘access traces’, which are actively recorded and stored in the system. Also, records of viewing the course video forms log information that is stored in the database. After mining the association rules, the relationship between these access and learning behaviors and their preferences and needs will be discovered, and after that the initiative course recommendation will be operated.

The use of association rules in data mining of users' access logs to discover the user's learning preference can get a lot of valuable information. In addition to recommend related courses for users to learn, it also has a reference value for teachers to arrange online courses and adjust content. At the same time, it can also make full use of the results of data mining to optimize the path structure, organization and layout of the online education platforms to achieve the purpose of improving the user's frequency of visiting the platform and their satisfaction of using it.

In addition to the association rules, there are other data mining technologies that can also be applied to curriculum recommendation. Among them, collaborative filtering is the most successful and widely used technology in various recommendation systems. It recommends information based on the preferences of groups of users who share the same interests and common experiences. In addition, cluster analysis is also a common recommendation algorithm. It distinguishes different groups of learners and summarizes the characteristics of each group; then it applies cluster analysis to individualized services to tailor different educational resources and services for different groups.

III. Method Design

Collaborative filtering algorithms, content-based recommendation algorithms, and association rules recommendations are commonly used algorithms in online course platforms. However, due to the disadvantages of these algorithms per se in the practice of online curriculum recommendation, those network teaching platforms usually adopt a hybrid algorithm, which often has better quality. edX also uses a combination of different recommendation algorithms, and this paper focuses on the recommendation algorithm of association rules.

The association rules method is to give two sets of items to analyze and derive the correlation between them. The whole recommendation model under this method consists of three parts: user, associated recommendation algorithm, and recommendation result.

a. The core module of the recommendation model is the recommendation algorithm module, which completes most of the computational tasks that are mainly on the calculation of users' historical behavior information, support degree and confidence of association rules. The completed project set will be available for project filter module to get the final recommendation.

b. When the user is a new user, this model will give the recommendation after the cluster analysis and the filter process according to the user's standardized interests and hobby information.

c. The association recommendation algorithm module is supported by the data formed in the online education platform.

The historical behavior data of new users are all blank. At this time, it is difficult to recommend courses to users based on collaborative filtering recommendation. Therefore, the cold start problem is an inevitable problem for the recommendation system. Therefore, for such users, curriculum recommendation should base on curriculum selection of old users, who have similar interest with the new one. After the user’s interest preference is obtained, and the user’s interest model is used to perform similar user cluster analysis on the user’s interest model, and then data analysis is performed.

When a new user registers his personal account, he will submit his full name, location, preferences and other information initiatively, which forms a collection of his attributes and preferences. In the process of implementing user similarity matching, a clustering algorithm is used to perform the maximum degree of similarity analysis for the user, which aims to find the old user with the greatest degree of similarity to the new user. And then, based on the course selection of the old user, this mechanism will recommend the courses to the new user.

Defining and assigning personal attributes of all users within the platform to form a set of user groups. Assume that there are n users in the platform, and each one has t attributes variables. Then the set of user groups can be presented as follow:

$$X = \{x_i \mid i=1, 2, \dots, n\}, x_i = \{x_{im} \mid x_{i1}, x_{i2}, x_{i3}, \dots, x_{it}\}, m=1, 2, \dots, t$$

Then, use the K-Means algorithm to select k samples from the group and dim them as the cluster centers.

$$K = \{x_j \mid j=1, 2, \dots, k\}$$

Calculate the distance of attributes m between the samples in X and K , and dim it as d .

$$d_{im,j} = \sqrt{(x_{1m} - x_{jm})^2 + (x_2 - x_{jm})^2 + \dots + (x_{im} - x_{jm})^2} \quad (1)$$

Based on the principle that the distance from the cluster center is the shortest, all n samples are divided into k sets. In these sets, the median value is selected as the new clustering center to do the iterative calculation until the clustering result no longer changes.

After the users are divided into k classes, courses also need to be divided according to the user class. In order to make the clustering result more accurate, the calculation of complex similarity degree is introduced. At the same time, based on the clustering of previous users’ attribute information, making the user-provided hobby information as a supporting factor to analyze the course similarity $Dem(x_1, x_2)$. And then, correct and fit the clustering results as follow:

$$CSD(x_1, x_2) = Sim(x_1, x_2) \times [1 + Dem(x_1, x_2)] \quad (2)$$

Finally, analyze the association rules of the generated data set and make a course recommendation.

For the old user, the course that he has already selected will be recorded in the system. Based on these data, the associated algorithm is used to calculate their relevance to other courses on the platform, so as to make course recommendation.

Association rule mining relies on Apriori algorithm to find the association rule that satisfies the constraints of minimum support degree and confidence level. The Apriori algorithm is divided into two steps. The first step is to generate all frequent itemsets in whole courses. The so-called frequent itemset refers to the item set whose support is higher than the minimum support degree. The second step is to generate a trusted association rule. The so-called trusted association rule is a rule with confidence level higher than the minimum confidence level.

The association rules between courses can be described as: all courses in edX platform form a set of items, $I = \{i_1, i_2, \dots, i_p, \dots, i_n\}$, i_p is called a course item; all users constitute the set D , $D = \{t_1, t_2, \dots, t_k, \dots, t_n\}$, t_k is called transaction, where each transaction is a collection of items, of which each item has a unique identifier, denoted as TID.

An association rule between courses can be expressed as a normal form $A \rightarrow B$, where both courses A and B are subsets of I and $A \cap B = \Phi$. This rule can be explained as that the user who choose A will prefer to choose B in the following phase. Support degree and confidence level are two indicators that measure the strength of the association rules.

The support degree of the rule $A \rightarrow B$ refers to the percentage of the transactions that contain courses A and B in D . This degree indicates the frequency of the rules in the transaction set D . The formula of the support degree X is:

$$X = \frac{A \cup B \text{ count}}{n}, \text{ Where } n \text{ is the total number of transactions in } D$$

The confidence level of rule $A \rightarrow B$ refers to the percentage of transactions that include both course A and course B in the transaction that contains course A . The formula of confidence level Y is:

$$Y = \frac{A \cup B \text{ count}}{n}, \text{ Where } n \text{ is the total number of transactions in } D$$

Assume that the minimum support degree is set as x and the minimum confidence level is y . If $X \geq x$ and $Y \geq y$, then recommend B to the user if he has already selected A .

Through the above algorithm, after analyzing and mining a large amount of data in the database of edX, all association rules that pass the constraints of minimum support degree and confidence level are obtained, and then a curriculum recommendation with higher credibility will be given.

An example of association algorithm course recommendation with hot start is illustrated in detail. Each time the user leaves ‘trace’ in edX system, it will be logged into the user’s log file and recorded in the database, just as shown in Table 2. Since the Apriori algorithm’s transaction database contains only TID and item, the user’s historical behavior record will be converted into transaction data format, as shown in Table 3.

TABLE 2 USERS’ LOG FILE

UID	Item	Login Time
13121111	A1	2016-5-8
13122222	A1	2016-5-8
13123333	A1	2016-5-8
13121111	B1	2016-5-8
13122222	B2	2016-5-8
13121111	C3	2016-5-8
13123333	B2	2016-5-8

TABLE 3 TRANSACTION DATA FORMAT

TID	Item
13121111	A1; B1; C3
13122222	A1; B2
13123333	A1; B2

Set a minimum support degree of 40% and a minimum confidence level of 60%. Based on the association algorithm and the minimum threshold, the frequent item set is {A1, B2} and an association rule ‘B2→A1’ is obtained. In this case, the support degree is 100% and the confidence level is 66.7%, both of which are larger than the threshold constraints. Then the rule can be interpreted as that more than 60% of the users who select course B2 have studied the course A1 before, and this relationship seems to be a reliable relation. Thus, if a user chooses B2, he will be recommended to study A1.

IV. The Feasibility of Association Rules Algorithm

Whether it is a cold start or a hot start, edX users need to register relevant information, and due to the compulsory requirements of this registration process, the information filled in by the user is generally true and effective. When there is a cold start, the recommendation will be given based on similar users’ association clustering and the result is credible. Each time the user logs in to edX, his or her behavior will be recorded and stored on the server. The resulting database is very objective. Therefore, when using the correlation algorithm recommendation, data acquisition and reliability are all feasible.

The algorithm of association recommendation relies on Apriori algorithm. Apriori algorithm adopts a layer-by-layer search iteration method, and there is no complicated theoretical derivation, so it is easy to be implemented on the technical level. It is also the basis of most association rule algorithms.

However, Apriori searches for too much iteration, which in turn generates a large number of intermediate itemsets. At the same time, the support degree adopted is unique, so the calculated results have a narrow adaptation range. Although there are some deficiencies in the Apriori algorithm, the association recommendation is still feasible.

V. Conclusion

As one of the three major MOOCs platforms in the United States, edX adopts a hybrid recommendation technology, of which association rule recommendation is the basis. Association rule mining is to mine association rules between courses through data in the database to recommend relevant courses to users. Under this excavation technology, edX can provide learners with more personalized services, and this technology is also feasible to some extent. However, this recommendation mechanism still has some shortcomings and defects.

From a technical perspective, firstly, although this recommendation mechanism solves the problem of ‘cold start’ from the perspective of the user, it still faces the problem of ‘cold start’ of the curriculums: When the new curriculum is introduced, there is almost no data between the curriculum and other courses, which means that it is impossible to know the support degree and confidence level between it and other courses and then, have no evidence to do the course recommendation. Secondly, in the face of a huge database of user information, it is necessary to continuously scan the database to determine the support degree and confidence level and judge the relationship between them and the thresholds. This operation is complicated to a high extent. Finally, the more itemsets and larger database there are, the more accurate the association rule is. In fact, the set of itemsets that can be collected in the system background which satisfy the constraints of minimum support degree and confidence level is limited. Therefore, the available data of association rules under the course recommendation will be very sparse.

From an operation perspective, the personalized recommendation mechanism is divided into explicit one and implicit one. The explicit one recommends courses based on the user’s own preferences which they initiatively provide when register into the platform. And the implicit one obtains the user’s online information mainly from the website cookies, which include the browsing history, IP address, name, and even the passwords of the users. Even the users will get more targeted recommendations on the MOOCs platforms based on this mining process but accessing to those undisclosed data without permission actually violates the user’s privacy and brings concerns. Thus, this is also a problem that all MOOCs platforms must face and to solve in the future.

References:

1. Han J, Pei J, and Yin Y, “Mining frequent patterns without candidate generation,” *Data Mining and Knowledge Discovery*, vol. 8, pp. 53-87, 2004
2. Lijuan Zhou, Mingsheng Xu, Yanyan Zhang, and Zhang Zhang, “Curriculum recommendation model based on collaborative filtering,” *Computer Application Research*, vol. 4, pp. 1-4, 2010
3. Moon S, Russell G. J, “Predicting product purchase from inferred customer similarity: An autologistic model approach,” *Management Science*, vol. 45, no. 1, pp. 71-82, 2008
4. Xuechan Li. “Research and application of association rules in curriculum relevance,” *Computer and Digital Engineering*, vol. 9, pp. 173-176, 2006
5. Xiaoxia Dong, Jianwei Li. “The operation model of MOOC”, *China’s Audio-visual Education*, vol. 7, pp.112-120, 2014
6. Yi Zhao, Yan Zhou. “The Apriori algorithm in the application of university curriculum,” *Teaching in Forest*, vol. 6, pp. 121-122, 2008
7. Yukselturk E, Ozekes S, and Türel Y K. “Predicting dropout student: An application of data mining methods in an online education program,” *European Journal of Open, Distance and e-Learning*, vol. 17, no. 1, pp. 118-133, 2014
8. Ye Sun, Jie Sun. “Cluster analysis of commercial users based on improved K-Means algorithm,” *Journal of Hebei United University (NATURAL SCIENCE EDITION)*, vol.1, pp.67-68, 2016
9. Zhenxia Gao, Xuejun Wang, and Junhua Wu. “Application of association rule mining in recommendation of online courses,” *Journal of Chengde Petroleum University*, vol.12, pp. 52-53, 2011
10. Zhenxia Gao. “Application research of curriculum information recommendation based on association rules,” *Journal of Liaocheng University (NATURAL SCIENCE EDITION)*, vol.2, pp.79-83, 2015