Causal Association Analysis Algorithm

For MOOC Learning Behavior And Learning Effect

Hongyan Wang, Xingwei Hao, Wenjiang Jiao
School of Computer Science and Technology
Shandong University
Jinan, China
lovewhjjourney@163.com

Xingwu Jia
Shandong Provincial Public Security Department
Traffic Management Bureau
Jinan, China
Shan-nan@163.com

Abstract—From 2012, Massive Open Online Course (MOOC), which combines the advantages of traditional teaching mode and the common network teaching, develops rapidly and has set off an education revolution. In the process of MOOC learning, a great deal of data related to the learning behavior are generated, but the data hasn't got efficient use. Embarking from the definition of learners' learning behaviors in MOOC learning environment, we analyzed the components of learning behavior, and proposed a formalized representation of learning behavior. By improving Apriori algorithm, we proposed a causal association analysis (CAA) algorithm to analyze the association between learning behavior and effect, and an application direction of daily inspection system based on the learning behaviors is also provided.

Keywords—MOOC; learning behavior; causal association analysis

I. INTRODUCTION

Learning is a process of accumulating experience. Behavior is changing due to changing experience, and this change should be relatively consistent. In this paper, "learning" refers to the process of learners’ acquiring knowledge in MOOC environment. Learning behavior refers to the various learning activities in the learning process, such as watching videos, downloading materials, completing tests and so on.

As for research on the MOOC learning behavior, the researches usually focuses on the learning behavior itself. Ebru used Naive Bayes tree and bivariate correlations classifier to achieve the establishment of Felder-Silverman Learning Style Model, and classified all learners according to their interests in learning [1]. Patricio Garcia established learning behavior model using Bayesian network, which could predict learners' learning behaviors in practical scenarios [2]. However, few people study the association between behavior and learning effects.

Although MOOC can make studying autonomic and personalized [3], learning effect is still showed by considering unit testing, final-exam score and participation in discussion forums, which is very similar to traditional evaluation method. Both MOOC and the traditional classroom teaching consider more about objective scores, and less about students’ performance in the learning process, so it is significant to analyze the association between different learning behaviors and learning effects. We proposed a casual association analysis (CAA) algorithm, which regarded different learning behaviors as reason and learning effect as result, to analyze the causal association between them.

II. RELATED WORK

Association analysis [4] is a task to find interesting relations in big data, these relationships can take two forms: frequent item sets or association rules. Frequent item sets are a set of data items often appearing together, and association rules may imply the existence of a strong relationship between the two or more items.

A. Apriori Principles

With the increase of data entry, the amounts of items will grow exponentially, that is, if the data set contains N data items, they can produce a total 2^N-1 set of data items, making traversal times dramatic growth, which is not conducive to efficiency of data processing. Therefore, in order to reduce the computation time required, we use the Apriori principle to reduce the term set.

Apriori Principle [5]: If an item set is frequent, all its subsets are also frequent, which can be described as follows: if a frequent item set is not frequent, all its superset are not frequent.

B. Apriori Algorithms

For Apriori algorithm, two input parameters are minimum support and data collection. The algorithm generates a list of all the individual items set. Then compute which items meet the minimum support requirements, and those set collections that do not satisfy the minimum support will be removed. Next, the remaining collection items are combined to generate set containing two elements. Then re-scan, remove the item that does not meet the set minimum support. This process is repeated until all sets are removed. Algorithm generating candidate set is as the following:
For each record in the data set as set:
   For each candidate set as candidate:
      check whether candidate is a subset of set:
         If so, increase the count value of candidate
   For each candidate set as candidate:
      If its support value isn’t lower than minimum, reserve it
Return a list of all frequent item sets

III. CAA Algorithm Design

Learners’ learning process is composed of learning behaviors, and their learning effects is objectively indicated by their total score. Whether learning behavior is effective will directly affect the quality of learning, so there is a clear causal relationship between them, but the influence of different learning behavior is different, even the same learning behavior, different times of behavior will generate different effects.

In this paper, we analyze a lot of learning behaviors which may affect the learning effect, then make a further study on the casual association between learning behaviors and effect by experiment.

A. Learning Behavior Components

We describe the components of MOOC learners’ behaviors with 5W1M model [5], as the followings:

(1) Learning actors (WHO). Actors refer to all learners in MOOC environment.

(2) Learning motivation (WHY). Motivation is the original purpose of learning, and learning behavior motivation can be divided into the following categories: active learning, for self-improvement, self-learning certified drivers and others.

(3) Patterns of learning behavior (WHAT). In MOOC, patterns is the way to learn, including watching teaching videos, downloading materials, participating in discussion forums, completing unit test and so on.

(4) Learning time (WHEN). Time of learning behavior includes login time, start and end time of watching a video, the time of browsing webpages and so on.

(5) Learning space (WHERE). MOOC learners’ learning space is MOOC platform.

(6) Learning behavior media (MEDIA). In MOOC learning process, the media including videos, pictures, PowerPoint and other rich-media types, which are more intuitive and specific.

In MOOC learning process, effective learning behavior of learners is broadly divided into the following five categories:

(1) Selecting learning courses. Before learning MOOC, learners need to register in a course, which means course selection is a prerequisite for learning.

(2) Watching videos. In MOOC platform, video is the most direct way to show the teaching content of a course, it is also the majority that learners would like to learn, so it is the most important learning behavior of learners.

(3) Downloading materials. Learning materials are supplemental content of video, which include PowerPoint and some additional materials of the course. Materials have been used to expand knowledge for learners, therefore it is also an important learning way.

(4) Forum. In MOOC platform, a discussion forum system is usually provided for learners and teachers, which is a space to put question and reply. The times and contents of posting messages may reflect whether a learner is superficial understanding or not.

(5) Homework and test Evaluation. It is an effective way to test learning outcomes, timely completion of the job evaluation can help learners to find their shortage.

Invalid learning behaviors include pausing a video, frequent page refreshing, publishing irrelevant posts, and browsing irrelevant pages, which significantly affect the focus of learners in the learning process and are harmful to improve learning efficiency.

B. Formal Representation of Learning Behaviors

(1) Time to choose course. MOOC is set accordance with terms. When releasing courses, the teacher set the beginning and ending time of his or her course. Also, deadline of units test is also set in advance. If a learner enrolls too late, it means the learner miss part of units testing and learning time is short, which is difficult to ensure good learning effect.

Formal definition of time to choose a course:

\[
\text{Choose} = \{\text{User}, \text{Course}, \text{Choose}, \text{Before}\}
\]

Wherein, \text{User} represents the learner’s account, \text{Course} expresses Course ID, \text{Choose} represents time to choose course, \text{Before} indicates whether choose the course before the start date.

(2) The length of time online. The length of time involved in the learning process of learners reflects whether learners have a good learning plan and the time spent for learning. With the extension of time, online time of learners is significantly decreased, so the length of online time affects learning effect to some extent.

Formal definition of time online:

\[
\text{Learning} = \{\text{User}, \text{Course}, \text{Start}, \text{Stop}\}
\]

Wherein, \text{Start} represents the time to start and \text{Stop} means exit this course of time.

(3) The time of watching videos. In MOOC learning process, watching videos is the most direct way to acquire knowledge, so video watching time is guarantee for gaining knowledge. Growth of watching videos time can significantly improve learning effects.
Formal definition of video watching time:

\[
Video = \{ \text{User, Course, Video, Length, Time, Pause, Watching} \}
\]

Wherein, Video means video number, Length refers to the video length, Time refers to the actual watching time length, Pause refers to the times of pausing video, Watching refers to the times of watching this video.

(4) Times of downloading materials. Learning materials complement of videos, which can be used to expand knowledge, so the behavior of downloading, to a certain extent, reflects whether the students gain knowledge actively. Therefore, the times of downloading learning materials can also affect student learning.

Formal definition of downloading materials:

\[
Materials = \{ \text{User, Material, Time, Times} \}
\]

Wherein, Material represents the number of learning materials, Time represents the time to download firstly, Times represents times to download, which strongly reflects learners’ attention to this material.

(5) Post in discussion area. Discussion area is a space for learners to communicate, questing and answering questions, so explaining learners’ own problems encountered in the learning process at the forum, showing that the learners master the knowledge not well, and can help consolidate their own knowledge at the same time. We think posting in discussion area can promote the upgrading of learning effect.

Formal definition of posting:

\[
Post = \{ \text{User, ID, Time, Length, Starter, Voted} \}
\]

Wherein, ID represents posting number, Time represents post time, Length represents posting content length, Starter represents whether is the landlord, Voted represents the times to be awesome.

(6) Homework and unit test. Homework and quizzes can help learners grasp their learning situation in time and make up for deficiencies, so timely completion of homework and quizzes can affect learning effects obviously.

Formal definition of homework and unit test:

\[
Test = \{ \text{User, Type, ID, Time, Score} \}
\]

Wherein, Type indicates the type of testing, such as unit testing, homework or the final exam, ID represents quiz question number, Time indicates that the time of test, Score indicates that the test results.

C. Equal-frequency binning to discrete data

Before analyzing the association between each learning behavior and learning effects, we need to discrete data firstly, which means scattering data into smaller sub classes. Each subclass is called a learning behavior, that is, each subclass learning behavior is a set of data items. In this paper, we use equal-frequency binning [6] to discrete data.

The basic idea of Equal-Frequency binning is ordering all data from small to large in accordance, and putting the data into K parts, each part is called a sub-box, that is, each sub-tank approximately accounts for 1/K of data [7]. Using equal-frequency binning assures that the number of sub-box remains substantially same, so we get the data items applied for CAA algorithm. All data for each learning behavior are discrete as follows:

- **Choose.** The number of days from the start date, discrete as (-15, 28], (28, 35], (35, inf), the negative value means enrollment the course before the start date, recorded as Choose1, Choose2 and Choose3 respectively.
- **Learning.** The sum of learning time length, discrete as (0, 5], (5, 18], (18, 43], (43, inf) recorded as Learning1, Learning2, Learning3 and Learning4 respectively. If total length of studying time in (0, 5] interval, we think the learning time is too short, we think they are not real learners. Therefore, it is not part of the data for the analysis.
- **Video.** The sum of video watching time, discrete as (0, 11], (11, 23], (23, inf), recorded as Video1, Video2 and Video3 respectively.
- **Materials.** Based on the sum of downloading learning materials, discrete as [0, 7], (7, 19], (19, inf), recorded as Materials1, Materials2 and Materials3 respectively.
- **Post.** The number of posts, discrete as [0, 9], (9, 14], (14, inf), recorded as Post1, Post2 and Post3 respectively.
- **Test.** The sum of times completing the unit testing and homework, discrete as [0, 9], (9, 11], (11, inf), recorded as Test1, Test2, and Test3 respectively.
- **Learning effects.** Expressed by the scores of courses, using equal-frequency binning method, discrete as [0, 34], (34, 68], (68, 82], (82, 89], (89, 100], recorded as Score1, Score2, Score3, Score4 and Score5 respectively.

D. Causal Association Analysis Algorithm

The collection of data items applied to CAA algorithm:

\{Choose1, Choose2, Choose3, Learning2, Learning3, Learning4, Video1, Video2, Video3 Materials1, Materials2, Materials3, Post1, Post2, Post3, Test1, Test2, Test3, Score1, Score2, Score3, Score4, Score5\}

However, according to the actual situation analysis, for each learner behavior, learning behavior data for each learner is only one subclass, for example, if total length of time watching video is 23 hours, then the value of the data item in Video learning behavior is Video3.

Apriori algorithm is used to find all relationships among all data items, but research here is to study the association between behavior and effect, so the discovery process to find frequent item sets can be simplified: all frequent item sets contains one
item of \( \text{Score1}, \text{Score2}, \text{Score3}, \text{Score4}, \text{Score5} \), that is, we do not learning analysis relationship between behaviors.

CAA algorithm analyze the causal association between a variety of learning behaviors and effects, so frequent item sets is a collection of different learning behavior and one learning effect; at the same time, each learner’s learning behavior data collection includes \( \text{Choose}, \text{Learning}, \text{Video}, \text{Materials}, \text{Post}, \text{Test} \) and a learning effect. Therefor the data structure is consistent.

According to the features described above, the CAA algorithm is described as follows:

**Causal Association Analysis (CAA) Algorithm**

**input:** data of learners’ behavior and learning effects

**Step1:** Discrete learning behavior and learning effect data

- **step1.1:** Discrete \( \text{Choose} \) as \( \text{Choose1}, \text{Choose2} \) and \( \text{Choose3} \)
- **step1.2:** Discrete \( \text{Learning} \) as \( \text{Learning1}, \text{Learning2} \), \( \text{Learning3} \) and \( \text{Learning4} \), delete the record of \( \text{Learning1} \)
- **step1.3:** Discrete \( \text{Video} \) as \( \text{Video1}, \text{Video2} \) and \( \text{Video3} \)
- **step1.4:** Discrete \( \text{Materials} \) as \( \text{Materials1}, \text{Materials2} \) and \( \text{Materials3} \)
- **step1.5:** Discrete \( \text{Post} \) as \( \text{Post1}, \text{Post2} \) and \( \text{Post3} \)
- **step1.6:** Discrete \( \text{Test} \) as \( \text{Test1}, \text{Test2} \) and \( \text{Test3} \)
- **step1.7:** Discrete \( \text{Score} \) as \( \text{Score1}, \text{Score2}, \text{Score3}, \text{Score4} \) and \( \text{Score5} \)

**Step2:** Generates candidate item set containing a causal learning behavior and a learning effect, for each record in the dataset record:

- **step2.1:** Generates a candidate set containing a learning behavior and a learning effect, candidate:
- **step2.2:** Check whether candidate is a subset of the record, if so, the count of candidate plus one

**Step3:** Calculation Rules confidence for each candidate set candidate

- **step3.1:** Get the count of candidate as \( \text{Count} \)
- **step3.2:** Calculate the frequency value of Score corresponding candidate \( \text{Count2} \)
- **step3.3:** Calculate rules confidence, \( \text{confidence} = \frac{\text{Count}}{\text{Count2}} \)
- **step3.4:** If confidence is not less than a predetermined minimum support, remain the item set

**Step4:** Returns a list of all frequent item sets

**output:** the casual association between learning behavior and learning effects

IV. EXPERIMENTAL PROCESS AND RESULTS

A. Data acquisition and process

The experimental data used for CAA algorithm are from the Chinese University of MOOC platforms course “College Computer Fundamental-a perspective on Computational Thinking”. Table 1 is a data sample of the learning behavior, and table 2 shows the discrete data samples.

### Table 1 Data Sample of the Learning Behavior

<table>
<thead>
<tr>
<th>User</th>
<th>Choose</th>
<th>Learning</th>
<th>Video</th>
<th>Materials</th>
<th>Post</th>
<th>Test</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>sdu_00100</td>
<td>026</td>
<td>31</td>
<td>26</td>
<td>18</td>
<td>12</td>
<td>12</td>
<td>89</td>
</tr>
<tr>
<td>sdu_05090</td>
<td>122</td>
<td>53</td>
<td>39</td>
<td>23</td>
<td>5</td>
<td>12</td>
<td>96</td>
</tr>
<tr>
<td>sdu_17090</td>
<td>010</td>
<td>12</td>
<td>10</td>
<td>2</td>
<td>17</td>
<td>4</td>
<td>64</td>
</tr>
</tbody>
</table>

### Table 2 Discrete Data Samples of the Learning Behavior

<table>
<thead>
<tr>
<th>User</th>
<th>Choose</th>
<th>Learning</th>
<th>Video</th>
<th>Materials</th>
<th>Post</th>
<th>Test</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>sdu_00100</td>
<td>0026</td>
<td>Choose1</td>
<td>Learning3</td>
<td>Video3</td>
<td>Materials2</td>
<td>Post2</td>
<td>Test3</td>
</tr>
<tr>
<td>sdu_05090</td>
<td>0122</td>
<td>Choose1</td>
<td>Learning4</td>
<td>Video3</td>
<td>Materials3</td>
<td>Post1</td>
<td>Test3</td>
</tr>
<tr>
<td>sdu_17090</td>
<td>0010</td>
<td>Choose1</td>
<td>Learning2</td>
<td>Video1</td>
<td>Materials1</td>
<td>Post3</td>
<td>Test1</td>
</tr>
</tbody>
</table>

### Table 3 Experiment Results

<table>
<thead>
<tr>
<th>Effect</th>
<th>behavior</th>
<th>support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score1</td>
<td>Choose3</td>
<td>0.62</td>
</tr>
<tr>
<td></td>
<td>Learning2</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>Video1</td>
<td>0.79</td>
</tr>
<tr>
<td></td>
<td>Post1</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>Test1</td>
<td>0.97</td>
</tr>
<tr>
<td>Score2</td>
<td>Choose3</td>
<td>0.51</td>
</tr>
<tr>
<td></td>
<td>Learning2</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td>Materials1</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>Test1</td>
<td>0.91</td>
</tr>
<tr>
<td>Score3</td>
<td>Choose3</td>
<td>0.51</td>
</tr>
<tr>
<td></td>
<td>Learning3</td>
<td>0.51</td>
</tr>
<tr>
<td></td>
<td>Materials3</td>
<td>0.52</td>
</tr>
<tr>
<td></td>
<td>Test3</td>
<td>0.76</td>
</tr>
<tr>
<td>Score4</td>
<td>Choose1</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td>Materials3</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td>Test3</td>
<td>0.79</td>
</tr>
<tr>
<td>Score5</td>
<td>Choose3</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>Learning4</td>
<td>0.52</td>
</tr>
<tr>
<td></td>
<td>Video3</td>
<td>0.57</td>
</tr>
<tr>
<td></td>
<td>Materials3</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td>Post3</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td>Test3</td>
<td>0.99</td>
</tr>
</tbody>
</table>
By analyzing the above results, we can obtain the following relationship between learning behavior and effects:

(1) If the enrollment time of the courses is in the last half of the course, it is difficult to achieve good grades, there may be two reasons, first, if the learning time is short, learners don’t have enough time to learn all content; the second is missing part of the assessment, which is a part of the final grade.

(2) There is a positive correlation, within a certain range, between the lengths of online time and watching video with learning results, that is the longer the time, the better the effects, but beyond a certain range, the time of studying does not enhance the effect obviously. The reason may be that learner’s attention is not easy to concentrate if learning time is too long.

(3) The times of posting does not influence learning effect greatly.

(4) Since the scores of tests and homework is added to the learning effect, the students with good effect basically completed the tests and homework on time.

C. Behaviors Not Increasing Effects

From Table 3, we can find the length of time online and watching video time grow to a certain extent, even if the time have a very significant growth, learning effects of learners do not get more significant growth. After deep research, we find some other learning behaviors not conducive to improve learning effect:

(1) Page refresh. When watching the video, refreshing the page frequently illustrates an unstable learning environment of the network, learners’ other behaviors, that is, focus on learning is not enough.

(2) Video pause. Time length of each MOOC video is generally within 5 to 20 minutes, and the learner pause frequently in a relatively short period of time, shows both the learner is lack of concentration, which not guarantee better learning results.

By using CAA algorithm to analyze the association these behaviors and learning effects, the behavior of frequent page refresh and video pause will affect the promotion of learning effect adversely.

V. APPLICATION

A. Improve Learners’ Learning Habits

MOOC provides learners a very large autonomy, but also require higher self-discipline. From the above experimental results, the higher grades need enough time of video watching and completing test and homework in time, so it is good that MOOC system makes suggestions for learner’s learning plan, according to the learner’s current learning progress, such as accelerating the progress of video watching.

B. Daily Inspect System Based on Learning Behaviors

In traditional classroom teaching, teachers can give the daily inspect score by learners' performance in class, which will be added to the final score. However, in the current MOOC platform, learner's final score is derived from homework, unit test and final exam. Although this score can reflect the learners' learning effects objectively, there may be some fortuity, so take full advantage of MOOC platform, the deep data mining of learning behavior can give daily inspect score more objectively, this part of the results can better reflect the attitude of learners and learning methods. Daily inspect score joining the final grade in a certain proportion of learners can reflect the learning effects better and more fully.

REFERENCES


