Research Article

Online Behavior Analysis-Based Student Profile for Intelligent E-Learning

Kun Liang,1 Yiying Zhang,1 Yeshen He,2 Yilin Zhou,3 Wei Tan,2 and Xiaoxia Li2

1College of Computer Science and Information Engineering, Tianjin University of Science & Technology, Tianjin 300222, China
2China GRIDCOM Co., Ltd., Shenzhen 518031, China
3Xiamen Great Power Geo Information Technology Co. Ltd., Xiamen, Fujian 361000, China

Correspondence should be addressed to Yiying Zhang; yiyingzhang@tust.edu.cn

Received 15 November 2016; Accepted 23 February 2017; Published 13 March 2017

Copyright © 2017 Kun Liang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

With the development of mobile platform, such as smart cellphone and pad, the E-Learning model has been rapidly developed. However, due to the low completion rate for E-Learning platform, it is very necessary to analyze the behavior characteristics of online learners to intelligently adjust online education strategy and enhance the quality of learning. In this paper, we analyzed the relation indicators of E-Learning to build the student profile and gave countermeasures. Adopting the similarity computation and Jaccard coefficient algorithm, we designed a system model to clean and dig into the educational data and also the students’ learning attitude and the duration of learning behavior to establish student profile. According to the E-Learning resources and learner behaviors, we also present the intelligent guide model to guide both E-Learning platform and learners to improve learning things. The study on student profile can help the E-Learning platform to meet and guide the students’ learning behavior deeply and also to provide personalized learning situation and promote the optimization of the E-Learning.

1. Introduction

As an effective way for education, the E-Learning supported more knowledge and skills than the traditional education and also is beyond the restriction of time and space based on new information and communication technologies [1]. MOOC (massive open online courses) is a representative online education platform. And the Coursera is the largest MOOC platform in the world, established by the USA top universities network learning platform. At present, there are about 1563 courses and more than 17 million registered students on the platform. Based on edX, the largest MOOC platform (http://www.xuetangx.com) is developed by Tsinghua University, in China. It has about 3 million members which are from more than 200 countries and regions [2].

E-Learning education has had a rapid development. Figure 1 shows that the number of Chinese E-Learning users reached 90.992 million in 2015, which is with an annual growth rate of 56%; it will grow to 120 million people by 2017. However, although more and more people are concerned about the E-Learning platform, there are only 7%–9% learners who completed MOOC’s course according to Coursera statistics data [3]. Therefore, it is very necessary to improve the quality of learning and optimize the teaching mechanism to push the course accurately. The student profile is a novel method to analyze the basic information and learn the behavior of online learners. Through the establishment of student profile, it is to achieve personalized situation construction and learn process guidance, which plays a positive role in promoting online learning.

The student profile is a figure portrait analysis based on the big data and labeling. We collect, process, and analyze the data generated in the learners’ behavior, for an information description of individual students or groups. According to the theory of behavioral psychology, use of the student profile to analyze the data on student behavior can reflect the students’ behavior characteristics and psychodynamics. For example, the Education Big Data Research Institute of
UESTC (University of Electronic Science and Technology of China) cooperate with other departments in developing the Student Profile System, which can give an early warning about failing the exam [4]. Similarly, Southwest Jiaotong University collected and analyzed the big data drawn students on campus “behavior track” model to predict students’ future development [5]. In foreign countries, researchers have proved that using the big data to analyze the students’ learning behavior such as reading course information online, submitting homework, and exchanging detected the warning information in poor learning performance. According to this warning information, the teachers made recommendations for improvement and given some guidance, to ensure that the students learn effectively. However, “student profile” has been applied in the education field; the big data profile technology combined with education has a very important practical significance.

In view of the E-Learning data, we use the big data technology to analyze the E-Learning characteristic, and the main research contents of this paper are as follows: (1) analysis of factors affecting the student profile. The students are classified according to age, and then we define the relations between student behavior and duration; (2) building a student profile model. We collected and preprocessed data and dug out the connection on the students’ learning behavior attributes by the Jaccard algorithm, to form the student profile finally; (3) analysis on the student profile. It is to contribute to the E-Learning platform to better understand the learning behavior of students.

2. The Definition of the Student Profile

The student profile described the learning characteristics from multidimensions and multiangle. It includes the analysis indicators and influencing factors, such as student behavior, data collection, data cleaning, and student profile building and analyzing [6].

2.1. Student Definition. The main research on the student profile is the students in the school or E-Learning platform. Assume the student set as follows:

\[ S = \{ s^j \mid s_1^j, s_2^j, \ldots, s^n_j \}, \tag{1} \]

where \( s^j \) indicates the students classified by age; \( i \) means the individual; \( j \) denotes the student age level (for convenience, the superscript is usually omitted as \( s_j \)), and there are 5 types of students classified in different age, which includes less than 17 years old, 18–24 years old, 25–34 years old, 35–54 years old, and over 55 years old, as shown in Table 1.

According to age, we can predict learner profile information and further dig into the characteristics of students learning.

2.2. Definition of Learning Behavior. The online learning behavior is the kinds of learning behavior under the network environment. We focus on digging out the characteristics of learners from online learning behavior after analysis, in order to understand the student’s performance. The core of learning behavior is the operation of online learning behaviors [7, 8]. For example, learner clicks on a course, browses the page, plays the video, and downloads the relevant courseware. The “click” and “download” are the two operations about learning behaviors. The behavior set (behavior) in the student profile is defined as

\[ B = \{ b_1, b_2, \ldots, b_n \}, \tag{2} \]

where \( b_j \) indicates varieties of behavior and includes 12 kinds of learning behaviors, such as learning goal, text learning, online practice, and making notes, as shown in Table 2.

Since the online learning is the period of time process with online learning behavior, it is an important parameter to evaluate the quality of online learning. In particularly, it
reflects the degree of focus on learning. The duration set (timeslot) in student profile is defined as follows:

\[ T = \{ t_1, t_2, \ldots, t_n \}, \]  

(3)

where \( t_i \) indicates various of durations; it is divided into periods as 1–10 minutes, 10–20 minutes, 20–30 minutes, 30–40 minutes, 40–50 minutes, 50–60 minutes, and so forth.

According to the above definitions, \( s_1 \cdot b \) means behavior \( b \) of the student \( s_1 \). We suppose behavioral differences \( \Delta b = s_1 \cdot b - s_2 \cdot b \) it represents the students’ differences in a certain behavior. For example, suppose college learner \( s_1 \), job learner \( s_2 \), and online training \( b \). We adopt \( \Delta b = s_1 \cdot b - s_2 \cdot b \) to analyze the differences between college learner and job learner in the behavior of the training online.

3. Student Profile Model

The student profile has a complete model to guide us to analyzing the students’ online learning process. The student profile model (Figure 2) includes data collection, data cleaning, and portrait analysis. Firstly, data collection can obtain the original data by means of E-Learning platform or questionnaire survey. Secondly, we utilize the attribute reduction to clean the original data and then employ the Jaccard coefficient algorithm for data analysis and data mining. Finally, we label the students according to results of the analysis in order to form the student profile. Simultaneously, we build the knowledge base (KB) for storage of knowledge sheet about the student profile [9]. In student profile model, KB is parallel to the data mining level and interaction. From the knowledge base, we take some of the student profile set to dig into and analyze and store the results in the KB. Therefore, the KB of the student profile has the self-growth and self-perfection ability.

3.1. Data Acquisition. Data acquisition includes four categories, such as student user registration data, web log data, learning behavior data, and learning content preference data. The student user registration data is mainly analysis on the characteristics of the learners, including user name, sex, date of birth, geography, occupation, and hobbies. The web log data reflects the operation of E-Learning platform, including active number, page views, access time, activation rate, and learning path. The learning behavior data is helpful for statistics analysis of online learning performance, including learning time, learning activities, learning resources, and examination results. The learning content preference data can be used to analyze the preference of courses or teachers, including browse/collection content, review content, and interactive content. It can be helpful for pushing the course accurately.

3.2. Data Cleaning. Data cleaning preprocesses the original data, removes redundant data, retains the useful data for the analysis, and organizes the data into a standard format. Because the interference of abnormal values often results in data mining distortion [10–12], data cleaning improves the accuracy of data analysis and ensures the reliability of data mining.

Attribute induction is the most important process of collecting the data source pretreatment. Suppose the original data field to \( \{ G_1, G_2, \ldots, G_{N_G} \} \), where \( N_G \) is the dimension of the original data field. Set the vector \( A \), where \( A \in \{ G_1, G_2, \ldots, G_{N_A} \} \) and \( A \) means desirable property. By the property statute of data preprocessing to give all the desired properties, the attribute induction method is defined as sig, cleaning the data to get the following results:

\[ \{ A_1, A_2, \ldots, A_{N_A} \} = \text{sig} \{ G_1, G_2, \ldots, G_{N_G} \}, \]

(4)

in which \( N_A \) is an important property for the dimension data field. In our solution, we calculate the importance of the property and select the same behavior analysis related to the desired attributes. Our solution does not deal with the concrete implementation of the attribute induction about sig.

4. Student Profile Analysis

In this section, we calculate the similarity in the behavior set of different students, through the Jaccard coefficient similarity algorithm compared with the online behavior characteristics and duration of learners, similar properties classified as a class, and the difference properties classified to different classes.

4.1. Student Behavior Feature Similarity Calculation. Similarity among the behavioral characteristics of different students objects belongs to nonnumeric objects; we adopt Jaccard coefficient calculated similarity [13, 14]. The similarity formula is as follows:

\[ r_{ijk} = \text{jaccard}(s_j \cdot b_k, s_j \cdot b_k) = \frac{|s_j \cdot b_k \cap s_j \cdot b_k|}{|s_j \cdot b_k \cup s_j \cdot b_k|}, \]

(5)

where \( s_j \cdot b_k \) and \( s_j \cdot b_k \) represent the behavior \( b_k \) of students \( s_j \) and \( s_j \). Suppose the student \( s_j \) belongs to KB; we compare
4.3. Duration Analysis of Online Learning Behavior. The learning behavior of online learners is diversity. To a certain

student could be labeled “depth learning type” if learning performance takes more than 60 minutes. Similarly, the student could be labeled “tasted type” if learning performance takes less than 10 minutes. Additionally, based on the frequency of online question and online training, we labeled “inquisitive type,” “application type,” and “perseverance type.” The specific labeling method does not repeat them here. We only proposed Jaccard coefficient algorithm and labeling idea for readers.

Algorithm 1: Jaccard coefficient algorithm.

(1) **Input:** \(ss //ss is a set of students**

(2) **Output:** \(R[i][j]//similarity matrix R\)

(3) 

\[ \text{"similarity of behavior characteristics between student } i \text{ and } j."/ \]

(4) **Dim R As Float[][]; \similarity matrix R**

(5) **Dim i, j, n As int;**

(6) **Dim k As float;**

(7) **Begin**

(8) \(n = ss.length;\)

(9) \// get the number of Student set

(10) **For** \(i = 0 \text{ To } n \{\)

(11) **For** \(j = i + 1 \text{ To } n \{\)

(12) **For** \(k = 0 \text{ To } |B|\)

(13) \//|B| is the number of behavior types

(14) \{

(15) \(r[i][j][k] = \text{Jaccard}(ss[i][j][k]);\)

(16) \}

(17) \(R[i][j] = \text{Sim}(ss[i], ss[j])\)

(18) \}\)

(19) **Return R**

(20) **End**

\[ R_{ij} = \text{Sim}(s_i, s_j) = \frac{1}{M} \sum_{k=1}^{M} r_{ijk}, \]  

\[ \text{in which } \text{Sim}(s_i, s_j) \text{ indicates the similarity between students } s_i \text{ and } s_j; \text{ } M \text{ indicates the behavior dimension attributes of student set } S; r_{ijk} \text{ represents the similarity about property } k \text{ between students } s_i \text{ and } s_j, \text{ and } i \neq j. \]

According to similarity calculation, we obtain Matrix*\_sim* as follows:

\[
\text{Matrix}_{\text{sim}} = \begin{bmatrix}
1 & R_{12} & R_{13} & \cdots & \cdots & R_{1N} \\
0 & 1 & R_{22} & \cdots & \cdots & R_{2N} \\
0 & 0 & 1 & \cdots & \cdots & R_{3N} \\
\vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\
1 & \vdots & \vdots & \vdots & \vdots & \ddots \\
0 & 0 & 0 & \cdots & \cdots & R_{N-1N} \\
0 & 0 & 0 & 0 & \cdots & 1
\end{bmatrix}
\]

It is an upper triangular matrix, where \(0 < R_{ij} < 1, R_{ij}\) denoted line \(i\) and row \(j\), and it is the similarity of behavior characteristics between student \(i\) and student \(j\).

Jaccard coefficient algorithm is described as shown in Algorithm 1.

According to the result of Jaccard coefficient algorithm, we can label the learners. Suppose that the calculation is from two dimensions about learning behavior and duration. The

\[ R_{ij} = \text{Sim}(s_i, s_j) = \frac{1}{M} \sum_{k=1}^{M} r_{ijk}, \]  

\[ \text{in which } \text{Sim}(s_i, s_j) \text{ indicates the similarity between students } s_i \text{ and } s_j; \text{ } M \text{ indicates the behavior dimension attributes of student set } S; r_{ijk} \text{ represents the similarity about property } k \text{ between students } s_i \text{ and } s_j, \text{ and } i \neq j. \]

According to similarity calculation, we obtain Matrix*\_sim* as follows:

\[
\text{Matrix}_{\text{sim}} = \begin{bmatrix}
1 & R_{12} & R_{13} & \cdots & \cdots & R_{1N} \\
0 & 1 & R_{22} & \cdots & \cdots & R_{2N} \\
0 & 0 & 1 & \cdots & \cdots & R_{3N} \\
\vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\
1 & \vdots & \vdots & \vdots & \vdots & \ddots \\
0 & 0 & 0 & \cdots & \cdots & R_{N-1N} \\
0 & 0 & 0 & 0 & \cdots & 1
\end{bmatrix}
\]

It is an upper triangular matrix, where \(0 < R_{ij} < 1, R_{ij}\) denoted line \(i\) and row \(j\), and it is the similarity of behavior characteristics between student \(i\) and student \(j\).

According to the result of Jaccard coefficient algorithm, we can label the learners. Suppose that the calculation is from two dimensions about learning behavior and duration. The

\[ R_{ij} = \text{Sim}(s_i, s_j) = \frac{1}{M} \sum_{k=1}^{M} r_{ijk}, \]  

\[ \text{in which } \text{Sim}(s_i, s_j) \text{ indicates the similarity between students } s_i \text{ and } s_j; \text{ } M \text{ indicates the behavior dimension attributes of student set } S; r_{ijk} \text{ represents the similarity about property } k \text{ between students } s_i \text{ and } s_j, \text{ and } i \neq j. \]

According to similarity calculation, we obtain Matrix*\_sim* as follows:

\[
\text{Matrix}_{\text{sim}} = \begin{bmatrix}
1 & R_{12} & R_{13} & \cdots & \cdots & R_{1N} \\
0 & 1 & R_{22} & \cdots & \cdots & R_{2N} \\
0 & 0 & 1 & \cdots & \cdots & R_{3N} \\
\vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\
1 & \vdots & \vdots & \vdots & \vdots & \ddots \\
0 & 0 & 0 & \cdots & \cdots & R_{N-1N} \\
0 & 0 & 0 & 0 & \cdots & 1
\end{bmatrix}
\]

It is an upper triangular matrix, where \(0 < R_{ij} < 1, R_{ij}\) denoted line \(i\) and row \(j\), and it is the similarity of behavior characteristics between student \(i\) and student \(j\).

According to the result of Jaccard coefficient algorithm, we can label the learners. Suppose that the calculation is from two dimensions about learning behavior and duration. The
extent, the frequency of learning behavior reflects the attention of learners to the learning resources. According to the frequency statistics [15], we analyzed which course resources are more likely to be accepted by the learners as shown in Figure 4.

In Figure 4, behaviors 1–7 are an independent learning behavior. Behaviors 8–11 are an interactive learning behavior. Behaviors 11–12 have nothing to do with learning. Most learners browse the text and make notes frequently; therefore, the text resource is the most popular type of resources. About 90% learners will first browse multimedia resources. Only 50% of the learners will participate in online exercises and only 1–5 times. About 60% learners will choose to view the learning objectives before studying the course, and the number of views is generally less than 5 times. 80% learners will rest; listen to music; or QQ chat 1–5 times during the learning process. This shows that most learners have a sense of learning strategies. They are interested in multimedia resources, but they are more used to learning through reading text resources. Online interactive learning behavior is low [16, 17]. Learning in the network easily causes fatigue and is affected by chat and other factors.

**4.4. Intelligent Learning Guide**

**4.4.1. Achievement Predicting.** MOOC is a popular E-Learning platform, whose importance is reflected in the pass rate of the course. In view of the low pass rate on MOOC [15, 18], it is supposed that a course is expected to be predicting whether the learner will eventually get the certificate according to the characteristics of the learner’s behavior. And we also try to verify the previous analysis and conclusions.

We suppose that the behavioral data was collected from the registered learners on Data Structures and Algorithm Analysis (DSAA) course for the first 5–7 weeks. After filtering the behavior of unregistered learners in the data record, the sample statistics are shown in Table 3.

**Table 3: Number of samples.**

<table>
<thead>
<tr>
<th>Course</th>
<th>1–5 weeks</th>
<th>1–6 weeks</th>
<th>1–7 weeks</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSAA</td>
<td>9401</td>
<td>9543</td>
<td>9990</td>
</tr>
</tbody>
</table>

Define each course having i learners and each learner having n characteristic values

\[ C = \{ C_1, \ldots, C_n \} \in R^{i \times n}. \quad (8) \]

Predictive value is

\[ P = f(C) \in R^i, \quad p \in [0, 1], \quad (9) \]

in which \( C_n \) denotes a course; \( p = 0 \) means it is unlikely to get a certificate, and \( p = 1 \) means you might get a certificate; and \( f(C) \) are a predictive function.

We have chosen the characteristic values of the courses, and they have an impact on result about learner’s study. From Table 2, they are the number of text learning behaviors \((b_2)\), the number of multimedia learning behaviors \((b_3)\), the number of online practice behaviors \((b_4)\), the number of download courseware behaviors \((b_7)\), and the number of online questions behaviors \((b_8)\).

According to this course, the data set is divided into training set, validation set, and test set randomly; the ratio is 3:1:1. To use the training parameters with the training set for each experiment, select the optimal parameters for the validation set, and then use the test set to calculate the indicators. We used three classification models: linear discriminant analysis (LDA), logistic regression (LR), and linear support vector machine (LSVM). They are used to predict the course, and the experimental results are shown in Table 4.

The experimental results show that the three classifiers show consistent performance, and the accuracy is higher. Figure 5 shows the time series changes of the predicted F-core by DSAA course. According to the learners’ learning behaviors in the first half of the semester, we accurately predicted the final study results, whether it can obtain the certificate [2]. In fact, if a learner performs well in the first half of the semester, it is shown that he is firm and capable. He is more likely than others to get the certificate finally.

**4.4.2. Emotional Guide Analysis.** Achievement prediction can help E-Learning platform to discover the abnormal situation, so as to timely intervention and guidance for students. Because online learners are mainly independent learners, they are in isolation and lack emotional communication, which makes them lack emotional support and have difficulty in maintaining long-term learning enthusiasm [19]. It is an effective way to solve the emotional deficiency by constructing an intelligent learning guiding mechanism in the student profile and providing some emotional help and support services to the learners during the learning process [20].

According to the E-Learning resources and learners’ behaviors, we can present an evaluation model supported by the duration, frequency of access, concentration, and
Table 4: Comparison of the forecasting results.

<table>
<thead>
<tr>
<th>Course</th>
<th>Classifier</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSAA</td>
<td>LDA</td>
<td>99.6</td>
<td>50.0</td>
<td>88.9</td>
<td>64.0</td>
</tr>
<tr>
<td></td>
<td>LR</td>
<td>99.5</td>
<td>65.0</td>
<td>76.5</td>
<td>70.3</td>
</tr>
<tr>
<td></td>
<td>LSVM</td>
<td>99.7</td>
<td>64.3</td>
<td>81.8</td>
<td>72.0</td>
</tr>
</tbody>
</table>

Finally, the student profile has been established as well as visual analysis. We confirm that E-Learning course requires definite objective, inner motive, synchronous feedback, and independence of the learners. The student profile helps the student to understand their learning situation, to find their own problems, and to improve the completion rate of online courses. With the continuous accumulation of education data and in-depth development, the student profile is bound to promote the healthy development of E-Learning. In the future, we will conduct in-depth study on the fragmentation of knowledge aggregation online.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the Tianjin University of Science and Technology Youth Innovation Foundation (no. 2016LG28).

References


