Analysis of Students' E-learning Behavior Based on Bik-Means Clustering Algorithm

Li-xian ZHAO*, Rong LI, Jun-min YE, Zhi-feng WANG, Xun BIN, Da-xiong LUO

*152 Luoyu Road, Wuhan, Hubei, China.
Central China Normal University Computer College
E-mail: 1770946273@qq.com TEL: 13476138479

Keywords: E-Learning Behavior, Clustering, Bik - Means algorithm.

Abstract. With the development of education technology, the number of E-learning users rises dramatically. How to evaluate and classify students through their learning ability and how to provide personalized guidance to students become valuable research points. Work in this paper based on the open source e-learning behavior data via DataShop (http://www.pslcdatashop.org/help?page=citing). First of all, this data is preprocessed to get each student's correct rate of each learning time (CRELT) (The basic unit of time is one hour) and hint times of each question (HTEQ). Then the data is classified via Bik-Means algorithm to get the classification about students' learning ability.

Introduction

With the development of network technology, the prospect of the network teaching becomes more and more bright. A new generation of Internet technology based on Web2.0 provides a convenient online learning condition for both teachers and students. They can share teaching resources, strengthen the communication from the online-learning website such as MOOCs and couresa[1]. People's learning activity presents new characteristics, which are superior to the traditional education, such as virtual, collaborative and individual. Although the network offers a large number of learning resources, platforms and some learning tools, which facilitate learners to expand the scope of learning, enhance communication between learners and process the knowledge point. Questions, such as low learning efficiency, weak learning motivation and insufficient interactions between teachers and students, still exist. Besides, the phenomenon that teachers are not familiar with students' learning ability and the low utilization rate of learning resources and tools are also big problems to be solved. Therefore, how to carry out the online-course more effectively becomes inevitable focus among people [2].

This article aims to help teachers have more clear understanding of different students' learning ability. Furthermore, the Web server log provides basis for our work. Because the Web server log can record the online information automatically, and the information is easy to obtain, thus a number of researchers have developed learning behavior monitoring system through establishing the learning behavior model [3] to analyze or even predict students' behavior[4] .they also analyze the Web server log to explore the application of Web data mining in the online learning behavior analysis [5].

In this paper, we study the open source data about students' online learning behavior recorded by the Web server log. We can get the classification of students' individual learning ability by preprocessing and clustering the data. According to that, teachers can have a more clear sense about student, and then they can find more personalized teaching methods to improve the teaching quality.

Research Foundation

Data Introduction

The data used in this paper is achieved from DataShop (http://www.pslcdatashop.org/Help?page=citing), a website provides open source data. It records the data about a series of learning behavior of 221 students from a learning website, including listening to audio, watching videos, doing exercises and so on. The data has 45 attributions, and we only use 8 of them in this study, which is shown in Table 1.

Table 1: Interpretations of used attributions

Attribution's name	Interpretation of attribution
Anon Student Id	Student ID
Action	Users' action,Including 11 types of action[1]
Input	Users' action to the input of the system
Student Response Type	Student Response Type, including 4 types[2]
Attempt At Step	Users' attempts to the same question
Is Last Attempt	Whether it's uses' last attempt to the question
Outcome	Feedbacks to users via system, including 3types[3]
duration	Duration of a student's one behavior

1: the Action includes: Multiple Selection Button Pressed, Jumble Check Button Pressed, play, stop, pause, end, cue, mute, unmute, volume-change and Button Pressed.

2: the Student Response Type includes: attempt, video-action, audio-action, and HINT_REQUEST.

3: the Outcome includes: correct, incorrect and hint.

Evaluation Index of Learning Ability

Ability is the comprehensive quality of one people. It is also people's practical skills, the skill level and energy showed in activities in their real life. So it's not only an effective way to realize human value, but also a positive power to influence the development of society and human life activities[6]. Thus learning ability is a psychological phenomenon, which is complex, multi-dimensional and multi-level [7].

In this paper, according to the attributions existed in the open source data, we characterize students' learning ability with two indexes. One is the CRELTand the other is HTEQ. The number of hints one student used in exercise and the time student spends online will both effect their correct rate, therefore only rely on students' correct rate can not reflect student's learning ability. Thus CURL and HTEQ are used to characterize students' learning ability.

Clustering and Bik-Means Algorithm

Clustering is a very important technology to analysis data in data mining, machine learning and data recognition. Clustering in data analysis means data with similar features are grouped together within a particular valid cluster. Each cluster consists of data which are more similar among themselves and data with weaker similarities are clustered in different clusters[8].Bik-Means algorithm is an improvement of traditional K-means algorithm. The iterative processing flow of K-means clustering algorithm is as follows:

(1)The number of clusters is first initialized and accordingly the initial cluster centers are randomly selected.

(2) A new partition is then generated by assigning each data to the cluster that has the closest centroid.

(3) When all objects have been assigned, the positions of the K centroids are recalculated.

(4) Steps 2 and 3 are repeated until the centroids no longer move any cluster [8].

Therefore, K - Means algorithm is sensitive to "noise" and outlier data, and a small amount of such data can have a great impact on the average value. It is also easy to be affected by the initial cluster center. Therefore, Bik-Means algorithm is used in this paper. The main idea of Bik-Means algorithm is as follows: first, the first cluster consists of the whole data, then we divided it into two clusters, then choose the cluster which can maximum reduce clustering cost that the error sum of squares, then divided it into two clusters. Repeated until the number of clusters is equal to the number of initial cluster centers that is given by users.

Research Process and Results

Reasons and Thoughts for Choosing Characteristic Data to Represent Learning Ability

Firstly, we analyze the open source data that we have got. We find that there exist some data directly related to learning ability in the open source data, such as the number of times that students watch the same video and the total number of the videos students watch. But through our analysis, we find there is too much noise in these data. So many students just watch videos for a short few seconds without further action. So this data is invalid. In addition, some data has effect on judging learning ability and others has a little effect, such as feedback, play components, etc.

In view of this, we analyze the attributes of the open source data to evaluate learning ability. Finally, we identify two indicators, CRELT and HTEQ, to evaluate learning ability, which can't directly obtain from the open source data. We use every student's correct rate divided by his/her total learning time to get each student's CRELT. The total learning time includes the length of time to watch videos, listen to audios and use hints. Then we use every student's total number of hints, which he/she used during exercises, divided by the number of questions to describe each student's HTEQ. However each student's learning time varies from others. Besides, the number of questions they answered and the hints they used are also different, so we decide to use CRELT and HTEQ to judge each student'slearning ability.

Methods and Process of Data Preprocessing

Methods of Data Preprocessing

We use CRELT and HTEQ to describe learning ability. Therefore, we need to calculate students' learning time, correct rate, the number of questions they did and hints they used. Though the correct rate can't be drawn directly through the original open source data, we can calculate the number of questions that students did and count the correct numbers. In summary, firstly, we need to calculate every student's learning time, the number of questions, the correct number of questions and hint-times that were used. Then we can calculate each student's correct rate and get CRELT and HTEQ finally.

(1) Calculating learning time. The total learning time includes the time to watch videos, listen to audios and use hints. Our group finds out the video-action and audio-action in Student Response Type of each student at the first time, and then finds each student's 'play' and 'stop' and 'end' type in their action attribution column ,which is related to the 'video-action' or 'audio-action' type in the Student Response Type attribution column. Then search for the time related to the 'play' and 'stop' and 'end' type in Input attribution column,calculate the D-value between 'play' and 'stop' or 'play' and 'end'. The time of using hint is the sum of the data in duration which is related to the HINT_REQUEST in Student Response Type attribution column.

(2) Calculating the number of questions. The number of questions is the count of times that attempt behavior appears in the column of Student Response Type in the original open source data.

(3) Calculating the number of questions done correctly. The times that "correct" appears in the column of Outcome are the correct numbers.

(4) Calculating times that hints were used. The number of hints used is the number of HINT_REQUEST in the Student Response Type column.

(5) Calculating correct rate. The number of questions done correctly divided by the number of questions that were done by students, and then correct rate can be got.

(6) Calculating CRELT. Our group uses every student's correct rate divided by his/her learning time to get each student's CRELT.

(7) Calculating HTEQ. We use times that hints were used divided by the number of questions to describe each student's HTEQ.

Process of Data Preprocessing

According to the methods of data preprocessing, in this paper, we take a student whose ID is 1 as an example to explain the process of data preprocessing. In line with the original open source data and the method to calculate learning time, we get his learning time: 0.5+0.5+0.5+0.5+18+0+1+3+9+2+…+1.306+23.38+2.952+23.38 =526(s)= 0.146111111(h).Because there are 95 “attempt” records in his Student Response Type column, the number of questions that he did is 95. And there are 38 “correct” records in Outcome column, so the number of questions done correctly is 38. There are 4 “HINT_REQUEST” records in Student Response Type column. Therefore, times that he used hints are 4. His correct rate is 38/95=0.4. So his CRELT is 0.4/0.146111111=2.737642588, and his HTEQ is 4/95=0.042105263.

We give 26 students’ data about CRELT and HTEQ, as shown in Table 2.

Table 2: 26 students’ data about CRELT and HTEQ

Student ID	CRELT	HTEQ
1	2.737642586	0.042105263
2	0.025150212	0
3	0.02584763	0.018111255
4	0.217056756	0.027918782
5	1.435506241	0.007142857
6	2.372719458	0
7	0.147770429	0.031904287
8	0.041680012	0.027181688
9	6.567345676	0
10	1.625112855	0
11	0.05177032	0.009922822
12	0.117049909	0.022891566
13	0.095272429	0.065963061
14	0.926164137	0
15	7.237090072	0
16	0.979258756	0
17	2.076902211	0
18	0.000145218	0.008179959
19	0.029743978	0.066154755
20	0.058296407	0.098949919
21	1.062037091	0
22	1.431027075	0
23	0.001665769	0.227091633
24	1.333412351	0
25	1.568529813	0
26	0.000318145	0.197341513

Clustering and Visualization of Results

211 students’ data about CRELT and HTEQ is clustered in the first time, and then we find that there is a point (400, 0) away from other points. We consult the student represented by the point and find he only did 1 question; the correct number of questions is 1, too. Besides his learning time is only 9 seconds. His data is obviously

unrealistic, so we remove it and cluster the second time. In the second time, we find there are two points (24.32432, 0.75), (0.011659, 1.192941) that separate from other points. We find that the student who represented by the first point, his learning time is 7 seconds. And the other student did 850 questions, by using hints for 1014 times. The data is also unrealistic, so we delete their data. Finally we cluster 208 students' data. The result is shown in Fig.1.

The abscissa of Fig.1 represents student's CRELT and the ordinate represents student's HTEQ. Analyzing Fig.1, we find that the aim to classify students is no achieved. Due to some students whose CRELT is lower and HTEQ is less, and students whose CRELT is lower but HTEQ is higher, are grouped in the same cluster. Through analyzing, we find the different metrics between CRELT and HTEQ cause the result. Therefore, our group tries to change the two metrics, so that the data's distribution is more accurate. After several attempts, we finally decide to multiply the HTEQ by 24.2 times for each student. The clustering result is shown in Fig.2.

Analyzing the clustering result in Fig.2, students are accurately classified into four categories that CRELT is lower, HTEQ is less; CRELT is lower, HTEQ is more; CRELT is medium, HTEQ is less; CRELT is higher, HTEQ is less.

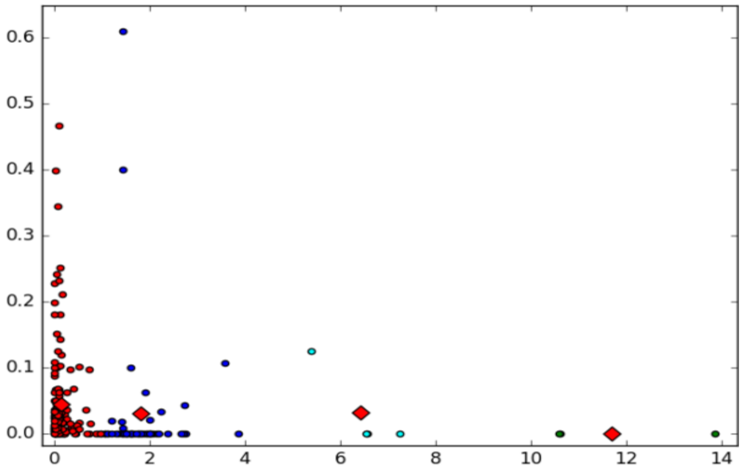

Fig.1 the clustering result of 208 students' data

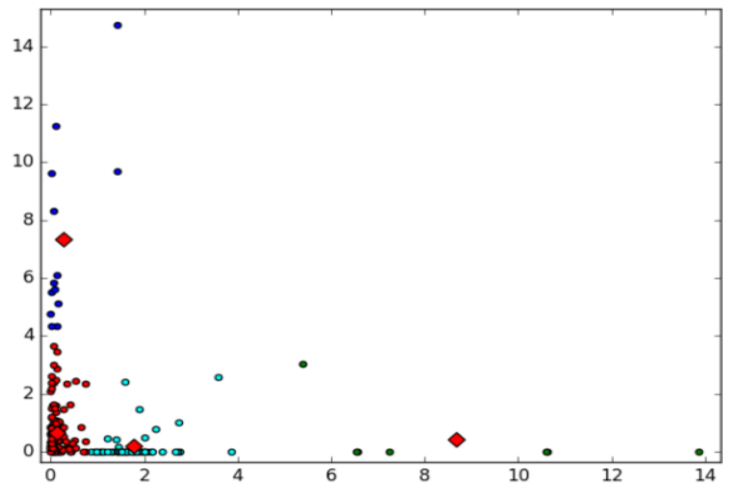

Fig.2 the final clustering result

Conclusions and Prospect

In this paper, we classify students by clustering students'CRELT and students' HTEQ, so as to distinguish the learning ability of different students. According to the classification results, teachers can master the learning ability of each student, and then provide students with personalized recommendations.

Next, we will be able to further study the different categories of students, based on the prediction algorithm to obtain different categories of students' "learning time – correct rate" models. Based on these models, we can have a more accurately prediction on whether the student will fail the subject according to the class of the student and student's learning time. They can also help teachers to stipulate the basic learning time of students, and know whether the student may fail the subject. Then teachers can implement interventions, so as to reduce students' rate of failing the subjects and improve teaching quality.

Acknowledgement

This work was supported by the Projects of the Social Science and Humanity Fund of the Ministry of Education (No. 15YJA880095).

References

[1] Jonathan M. Kevan, Michael P, Menchaca, Ellen S. Hoffman, (2016) Designing MOOCs for Success: A Student Motivation-Oriented framework .LAK

Information on http://dx.doi.org/10.1145/2883851.2883941

[2] Carly Robinson, Michael Yeomans, Justin Reich, Chris Hulleman, Hunter Gehlbach, (2016) Forcasting Student Achievement in MOOCs with Natural Language Processing, LAK

Information on http://dx.doi.org/10.1145/2883851.2883932

[3] ZacharoulaPapamitsiou, EiriniKarapistoli, AnastasiosA, Economides (2016)

Applying classification techniques on temporal trace data for shaping student behavior models. LAK

Information on http://dx.doi.org/10.1145/2883851.2883926

[4] Eric Van Inwegen, Seth Adjel, Yan Wang, Neil Heffeman (2015).An Analysis of the Impact of Action Order on Future Performance: the Fine-Grain Action Model. LAK

Information on http://dx.doi.org/10.1145/2723576.2723616

[5] Laura K.Allen, Caitlin Mills, Matthew E.Jacovina, Scott Crossley, Sidney D’Mello, Danielle S.McNamara (2016) Investigating Boredom and Engagement during Writing Using Multiple Sources of Information: The Essay, The Writer, and Keystrokes. LAK

Information on http://dx.doi.org/10.1145/2883851.2883939

[6] Han Qing-Xiang. Competence-based [M].Beijing: China Development Press, 1999. (In Chinese)

[7] Yang Su-Juan. The essence and composition of online learning ability [J]. Chinese Distance Education, 2009, 05: 43-48+80. (In Chinese)

[8]BikramKeshari Mishra, AmiyaRath, NiharRanjanNayak, Sagarika Swain(2012)Far efficient K-means clustering algorithm. ICACCI’12 Proceedings of the International Conference on Advances in Computing, Communications and Informatics.106-110