Received ; | Revised ; | Accepted

DOI: xxx/xxxx

RESEARCH ARTICLE

SPOC online video learning clustering analysis: Identifying learners' group behavioral characteristics

Summary

As the wide spread of SPOC (*Small Private Online Courses*) in colleges and universities, course organizers provide high-quality personalized course activities needs to understand learners' learning status and characteristics, and then optimize the course teaching. However, little research has been done on different learners' group behavioral characteristics, such as which indicators can reflect learner groups' behavior, what are the typical behavioral characteristics of differ learner groups. We established the *Python Language Foundation* self-built SPOC course, and one hundred and nine undergraduates' learning behavior data were collected and analyzed. From 74-dimensional behavior log data consisting of seventy-two video-viewing, course video score and comprehensive score, PCA Dimensionality Reduction is performed, divides learners into different categories by using agglomerative hierarchical clustering, and the results show that 15 groups are clustered. According to the analysis of the four indicators for group characteristics, which the completion and viewing-stability of task-point videos, unit test excellence and comprehensive score, it is identified into five primary types, including positive type, regular type, negative type, semi-negative type and a fluctuation type. Experiments on a real data set along different academic years and courses are conducted show that the proposed approach has better clustering accuracy and practicability. It is expected that this work may be used to strengthen the personalized learning support services system in educational institutions and develop a tool which integrates exploration and analysis work onto the web platform.

KEYWORDS:

SPOC, Learners' group behavioral characteristics, Video-viewing log analytics, Learner cluster, Analysis indicators

1 | INTRODUCTION

SPOCs (*Small Private Online Courses*) are one of the forms of higher education teaching. Different from the traditional Face-to-Face Teaching in Classroom, SPOC has advantages of flexible learning style and content expansion. And the most important advantage is recording learners' learning motivation, process and learning behavior [1, 2, 3]. Recent research on SPOCs focuses on the behavioral characteristics and performance of online learners in the activities of SPOCs [4, 5, 6]. And more data mining technologies are also applied, for example, Zhang et al. [2] used SPOC's record data, and adopted K-Means clustering and hierarchical clustering analysis algorithms to mine learners' learning characteristics. Mona Al-Saleem et al. [7] focused on the previous students' academic records, and build a performance prediction model used different classification techniques. Rodriguez, Duque and

Ovalle. [8] presented a knowledge-based recommender systems for Learning Objects (*LOs*), which uses a K-Means clustering technique to find similar LOs and delivering resources suitable for a specific student. Romero et al. [9] used both clustering and sequential pattern mining algorithms together for discovering personalized recommendation links. Their endeavor mainly aims to resolve the well-known problem of low quality personalized learning support service [1, 10, 6]. SPOC is an important means of blended teaching in colleges and universities. Teaching organizers provide high-quality personalized course activities, which can effectively improve the teaching quality of colleges and universities, cultivate talents, and create economic value for the society.

This study aims to makes exploration and analysise learners' learning status and characteristics by carrying out explanation and quantitative in SPOCs. The findings are be used to support high quality personalized learning services and assist teachers to optimize teaching on specific courses. However, SPOCs contain a variety of learners, and learners' learning needs and preferences are quite different, resulting in diverse learning behaviors [11]. Furthermore, SPOC is a popular restricted online course in colleges and universities, including standardized teaching videos, classroom task points, regular tests, forum discussions and other teaching forms [2]. The data recording learners' learning behavior in SPOCs has the characteristics of large quantity, multiple types and fast updates [12]. Under the above two conditions, it is difficult to obtain learners' learning status and characteristics by analyzing learning behavioral data manually [2]. In addition, little research has been done on different learners' group behaviors and characteristics.

Therefore, the objective is present an approach to cluster previous learners' video-viewing data on SPOC's specic courses, and then according to the analysis indicators to mine and analyze learners' learning status and group behavioral characteristics. Specifically, there are three questions to be addressed in this study:

- The log data of learning behavior is diverse and high-dimensional. The more features are selected, the more learners' learning behaviors are expressed, but not better [13]. How to deal with high-dimensional data?
- What indicators are used to mine characteristics for learner groups?
- What are the typical characteristics among learner groups?

In this paper, we construct a *Python Language Foundation* self-built SPOC course, and then collect and analyze the video-viewing behavior data, the data set used is learner data from the Computer Science Department, Information and Electrical Engineering College, Heilongjiang Bayi Agricultural University (*HBAU*). In the rest of this paper, section 2 introduces the theoretical background and current research status of SPOC. Section 3 proposes a method to identify learners' group characteristics. Section 4 is the experiment results and discussion. Finally, summary and future lines are outlined in Section 5.

2 | THEORETICAL BACKGROUND

2.1 | SPOCs

SPOC represents a specific, defined form of fully online education [14]. In comparison with Massive Online Open Courses (*MOOCs*), SPOCs are characterized by the restricted number of students, which generally ranges between dozens and hundreds [15], significant instructor's guidance who is often assisted by an e-moderator, significant peer interaction, and usually fixed start and end dates. Different from MOOCs, the purpose of students starting a SPOC is not only to participate in some courses, but also to consciously complete in it [16]. The most striking feature of SPOC, says professor Robert Lou of Harvard University, is that it goes beyond trying to replicate the traditional classroom model completely. On the contrary, it attempts to produce a more flexible and effective model by combining a variety of teaching models [17]. SPOC is the perfect combination of traditional classroom teaching and MOOC, i.e., SPOC = Classroom + MOOC [18].

2.2 | Analysis of the characteristics of learners in SPOC

When learners participate in SPOC online course activities, the SPOC platform can record learners' learning and interaction data in a comprehensive and multi-angle manner [19]. The analysis of student behavior data is inseparable from educational data mining technology [20]. Educational data mining (*EDM*) as an area that uses statistical techniques, machine learning and data mining to analyse educational data [21]. Among these prediction of college student performance [22], pattern detection of students' behavior [23], a student model [24], educational data visualization [25], the recommendation system [26, 27], and adaptive system [9] are hot topics in educational data mining. Particularly, EDM technology includes special data interpretation and processing techniques, such as clustering [28], classification [29], prediction [30, 31], asso-Ciation [32], visualization [33], outlier detection [34], time series analysis [23]. In addition, many methods and algorithms have been applied, such as an intelligent agent, neural network [35], the decision tree algorithm [36], bayesian network [30], logistic regression [37], K-means [38] and fuzzy logic [39], etc.

In SPOC online teaching, the research on educational data mining of learning behavior in SPOC courses mainly focuses on model construction and utilization [2]. For example, Liu [40] constructed students' behavior model [41] by analyzing the characteristics of the Cloud Classroom and comparing the SCROM and the xAPI specification. The study in Peng, Han, Ouyang et al. [5] incorporated time, emotional and behavioral features into topic modeling, creatively proposed the time information-emotion behavior model (*TI − EBTM*). In addition, based on the structure of the classical sentence-level emotion-topic unification model, Yang [42] added time and emotion prior knowledge, and a more fine-grained temporal-emotion-topic model has been proposed to explore the themes of learners' attention and the corresponding emotion and evolutionary trend. The work in L, Wang, Zheng et al. [43] collected the learning records of 1,2517 undergraduates participating in the SPOCs during one academic year, called StarC developed by Central China Normal University. Empirical observation and lag sequential analysis method (*LSA*) [2] applied to discover the most significant several sequence patterns in each grade, and the differences between grades in behavioral patterns were also discussed. Researchers from Bois State University utilized the existing learners' behavior to build models, and adopted classification and regression tree decision methods to analyze data of curriculum learning to predict students' learning performance and recognition of curriculum and teachers [44]. Ramesh, Kumar, Foulds et al. [45] proposed a weak supervision joint model (*PSL − Joint*) to predict the emotional state of online courses. In Mubarak et al. [31], the authors used hidden Markov model [46] to predict learners' dropout behavior.

Since 2008, researchers have paid attention to the emotional information in the interaction of learner forums, and explored the relationship between emotional information and learning effects. For example, Estrada, Cabada, Bustillos and Gra. [47] presented a comparison among several sentiment analysis classifiers using three different techniques – machine learning, deep learning, and EvoMSA, proposed an emotion mining method for forum text learning. The results show that there is a significant positive correlation between positive emotions and learning effects. Liu Zhi explored natural language processing technology, e.g., Word segmentation, Latent Dirichlet Allocation (*LDA*) Text Topic Modeling Technique, text similarity calculation, bayesian probability generation model, etc. By studying the relationship between learners' discourse behavior and learning performance in SPOC forums, it is found that the number of posts is significantly positive correlated with students' learning outcomes, and confusion over the emotions and learning effects are significantly negative correlated [48, 49], furthermore, they discovered that the correlation among the topics of learners' interests and teaching content in SPOC forum has impact on learners' learning outcomes. [6]. Liu, Xian, Hercy N. H. et al. [50] put the behavior and sentimental tendency in the forum posts into the topic model, called behavior -sentiment topic mixture (*BSTM*) model, mainly from text mining and behavioral analysis, to explore the emotional tendencies and interactive behaviors of the specific topic. Uijl et al. [16] divided the forum interaction content in SPOC course into content specific, functional or technical and social. Meantime, through SPSS analysis of interactive content, it is evaluated that SPOC course plays an active role in providing students with social space for collaborative learning, promoting meaningful and task-related interactions, and creating social interaction to stimulate students' learning motivation. In addition, some researchers used data mining technology to normalize each data item in the teaching process, and then constructed the calculation model of early warning value, which was used to dynamically monitor the change of early warning value of students in the classroom implementation process, and to warn and help students [51, 52].

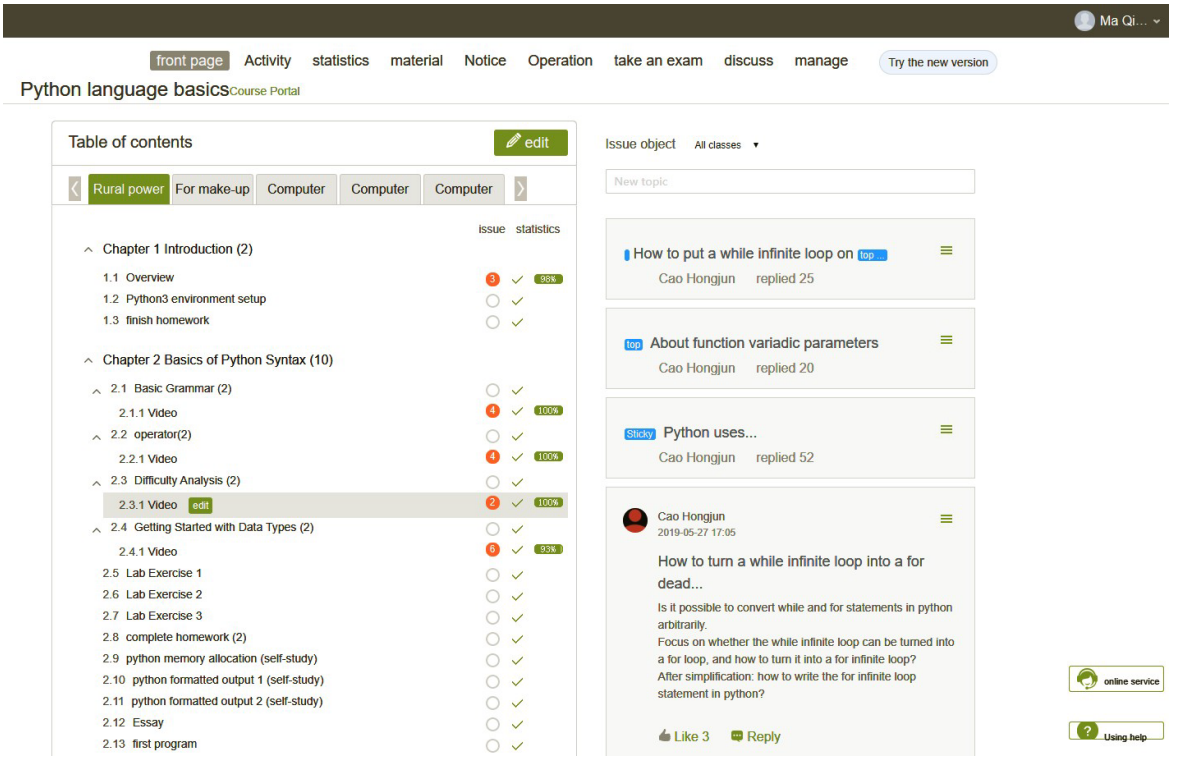
It is concluded that learners' implicit and intrinsic characteristics cannot be directly observed and measured, but they can be excavated and perceived through massive data of online learning behavior [53], so as to implement corresponding interventions to promote effective learning or better teaching. Moreover, the video in SPOC course is a preferred form of learning for students to acquire knowledge [54, 55], due to the lack of sufficient text or metadata information, there is little work on video clustering or classification [4]. We focus on discussing and analyzing students' video-viewing behavior data in the later.

3 | METHOD

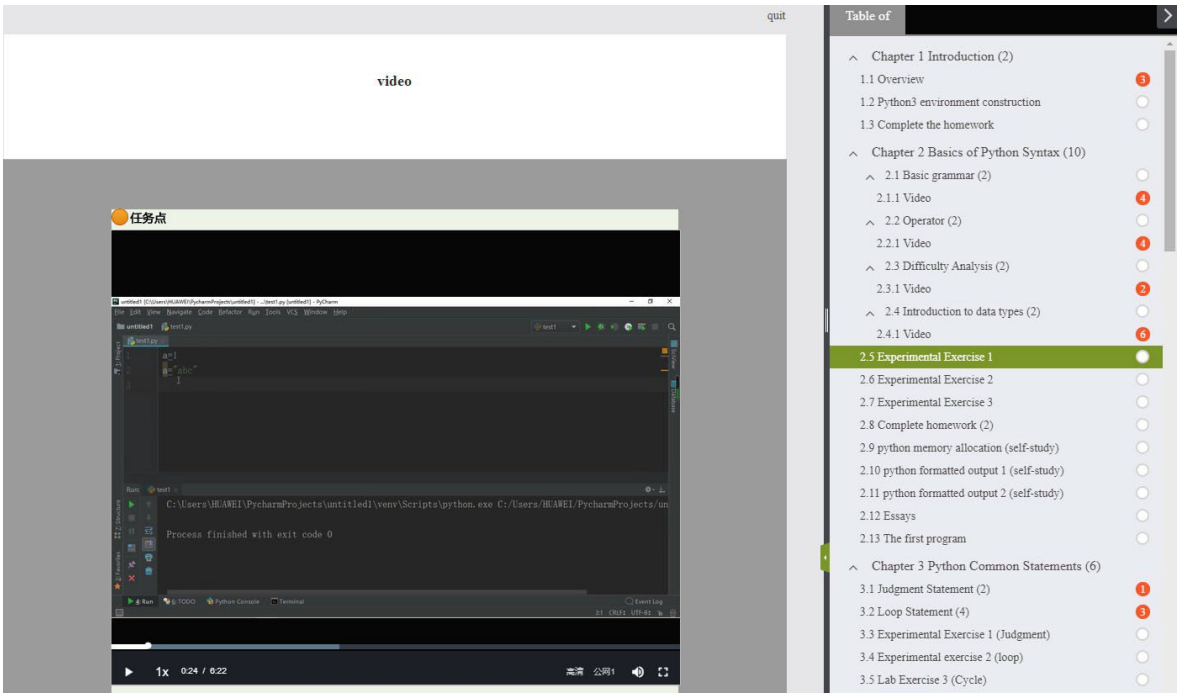
3.1 | Self-bulit *Python Language Foundation* SPOC course

Super Star Learning platform uses the advantages of MOOC to provide a MOOC system for self-built SPOC courses, which changes the traditional classroom learning. First, students' learning time is more flexible. They can learn anytime and anywhere using computers or portable devices, and can repeatedly play teaching videos and download learning materials. Second, students can see a list of all the SPOC courses that have been constructed in their school by logging into the Super Star learning account. We create a self-built *Python Language Foundation* SPOC course on the Super Star learning platform, as shown in Figure 1 (a), which includes the following nine modules:

Home page: This module provides static information settings about selected courses, including directory editing and



(a)



(b)

FIGURE 1 *Python Language Foundation* SPOC course interface: (*a*) The teacher edit page of Super Star Learning Platform. (*b*) The student learning page of Super Star Learning Platform.

preview, distribution objects, and topic forums. With this information set, students have a comprehensive understanding of the course content. They can select the corresponding chapter, and then play the course task video, as shown in Figure 1 (b).

Activities: This module provides teachers with the organization of various classroom activities. Teachers can use this module to publish dynamic information related to the course, such as check-in, in-class exercises, voting, selection, answering, topic discussion, questionnaires and group tasks. After receiving the notice, students can participate in activities and communicate with teachers, so as to stimulate students' learning enthusiasm.

Statistics: This module provides comprehensive information of curriculum learning, including class statistics such as homework, examination and early warning, and statistics of teaching resources such as videos, documents and animations, as well as statistics of curriculum reports such as comparison of curriculum scores. "Classroom Statistics" mainly include statistics on the number of semesters, teaching hours, and accumulated numbers. This module aims at helping teachers master the dynamics of the learning situation and promoting the improvement of teaching.

Curriculum material: This module provides teachers to upload various learning materials. "Adding Data" can choose local uploads, cloud links, online books, etc. "Question Bank" allows teachers to manually create or bulk import the questions. "Homework" can create multiple types of questions such as single-choice questions, multiple choice questions, written response questions, fill-in-the-blank questions and true-false questions. "Test Paper Database" allows the import and export of test papers, and strictly manages test papers. Only the teacher who have been given permission have the right to open the test paper and review it.

Notification: The main functions of this module include releasing school announcements, publishing course adjustment messages, and reminders for assignments and exams. "Notification" can set the sending and copying object, sending time, title and text content. Once the notice is released by teachers, it will push and remind students in the form of short messages.

Homework: Students can use this module to submit homework assigned by teachers, and teachers can check students' homework.

Examination: This module provides normal online examination. With strict time limit and the submission function, only allow students to log in to the "Examination" page once and ask for answers within the specified time period.

Discussion: This module provides topic discussion and interaction. All the topics can be retrieved according to keywords, and the deleted topics and content can be restored in "Recycling Station". Click on "Add" to create a new topic and set the distribution object. After topics are released, students who choose the SPOC course can reply and discuss topics. "Batch Export" can export the topic content and download the content in the download center.

Management: This module provides a unified management function for students and teacher teams. Classes can be added in "Class Management", and multiple classes can be added in one course. For different classes, batch import student ID number, can also manually add one by one. "Teacher Team's Management" can add teacher information and give specified permissions. "Course Management" is divided into students terminal setting and course portal. The "Students Terminal Setting" can set the contents displayed in the navigation box. "Course Portal" sets up notification services, course reading, chapter display, course reuse and course description. "Oper- ation Log" allows for recording all relevant information of operation and object in detail.

The learning mode of SPOC course is online and offline blended learning [56], and the classroom learning mode is not restricted. Specifically, teachers provide students with the learning videos and course materials. In the meantime, stu- dents can choose to learn before class, download learning materials and complete homework. In the classroom, teach- ers answer students' doubtful questions, and comprehensively teach or discuss, so as to promote the implementation of flipped classroom teaching mode.

3.2 Research design

Figure 2 shows the exploratory research design of this study, which consists of three stages. The three stages are the stage of selecting learning equipment, educational data mining and analysize stage, and the results associate to learning and teach- ing. Each stage is detailed as follows.

In the first stage, SPOC courses are released on the Super Star Learning Online platform. Students can use tablet computer or mobile phones, laptops and other devices for online learning. An account is only allowed to login on one device. There are two ways for students to join SPOC courses： one is that teachers introduce student ID number into the course in advance, and students can learn by logining in to the Super Star learning account (using the student ID number). Another is that teachers give students invitations, students enter the course by scanning the Quick Response Code (QR code).

In the second stage, we collect the data of learners' learning behavior in SPOC course and pre-processing. Then the hierarchical clustering algorithm based on class-average-diameter optimization is used to cluster the data properly, and the characteristics of different learner groups are understood from the macro level according to the analysis indicators. This stage is the specific realization of the motivation of this study. The critical steps are as follows：

· Learners' behavior data collection. The original data of learners' online behavior recorded by SPOC platform is obtained. The database mainly includes structureddata such as original course information, video viewing and scores, and tests.

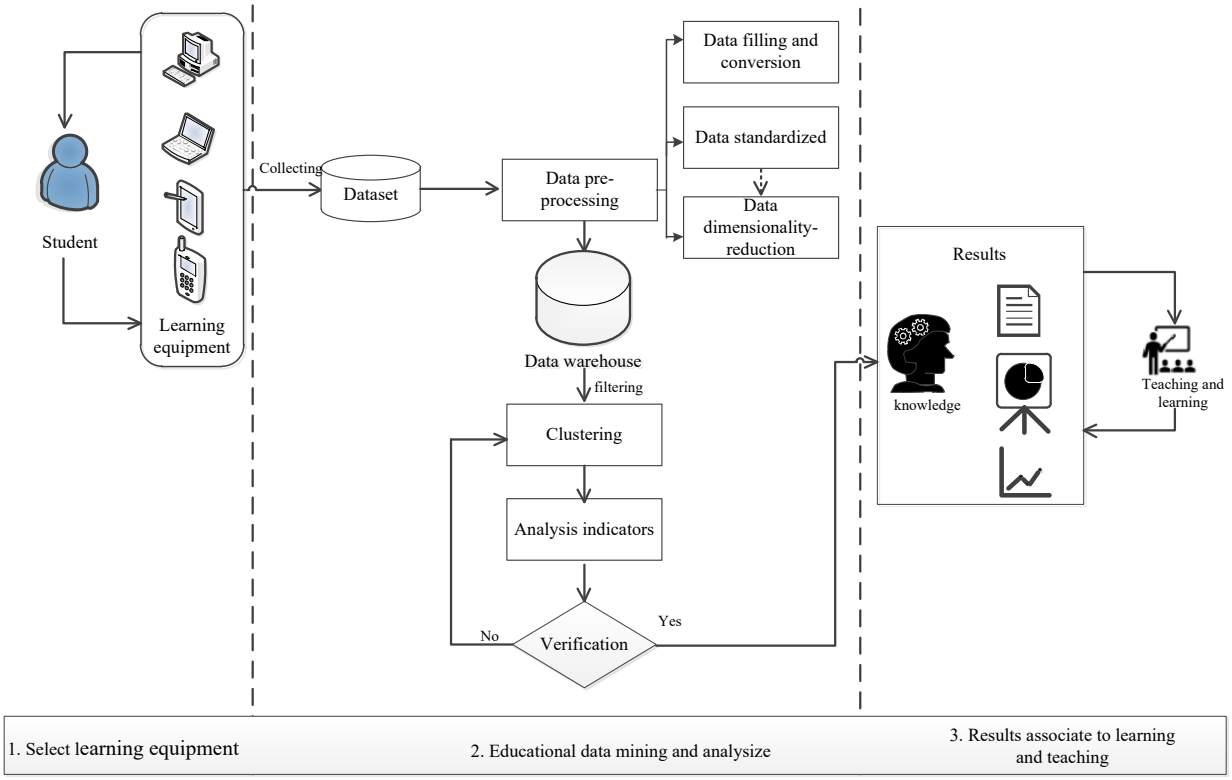


FIGURE 2 Research design of this study

· Data pre-processing. Deal with missing, repeated and inconsistent data in original data. In this process, data standardization and data dimension-reduction are mainly used. The processed results are stored in the data warehouse.

· Mining and analysis. Clustering and analysis indicators are used to mine and analyze learner groups.

· Verification. Using the data of learners' learning behavior in different academic years and different courses, the applicability of clustering and analysis indicators is verified.

In the last stage. Learners' group characteristics are identified and explained. The results can be converted into knowledge and reports, applied to learners' personalized learning support service and organization optimization of teaching process.

3.3 Participants and data collection

The study's participants are 109 undergraduate students. Statistics on gender and age distribution of learners' in this course show that women account for 36% , men account for 64%, and the median age is 21 years old. The results are shown in Figure 3 .

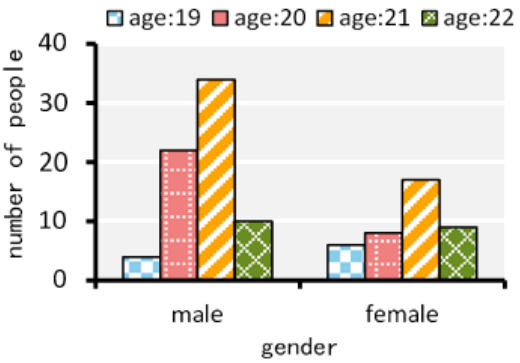


FIGURE 3 Gender and age statistics of learners

The data is derived from the *Python Language Foundation* self-built SPOC course offered on the Super Star Learning Platform in the spring of 2020, which contains learning behavior data of all participants, such as duration of learning, learning times, downloads retention progress, check-in, comprehensive score etc. Table 1 shows an example of curricula-variable records and behavior records. In order to protect students' personal privacy, the data is desensitized in some places, such as the student ID number, name and department are replaced with other data.

3.4 Data Pre-processing

Table 1 must include one row and several columns for each student. The columns will be grouped by courses for the purpose of analysis of the students' behavior. We collected and "tidying up" the data (Data preprocessing): the record format of the video-viewing duration in the source data is "hour :minute: second", which is inconvenient for subsequent analysis and processing, so it is converted into a format in seconds. The missing data formed without watching the video is filled manually with 0. Calculate the rumination ratio, which is the task-point completion degree (see Equation 5).

TABLE 1 Example of source data storage format

Course	Description	Student ID	Task Video ID	Video-Viewing Visualization	Video-Viewing Duration	Video Duration	Check-in Score	Finish Time	Course Video Score	Download Retention Progress	Comprehensive score	...
Python	Elective course	Sdn407101	2.1-1		16.6 minutes	16.2 minutes	10.0	2021-06-06 15:39	40	72/72	92.5	...
Python	Elective course	Sdn407102	2.1-1		8.1 minutes	16.2 minutes	10.0	2021-03-25 21:21	40	71/72	94.5	...
Python	Elective course	Sdn407103	2.1-1		12.6 minutes	16.2 minutes	10.0	2021-05-07 12:42	40	72/72	87.5	...
Python	Elective course	Sdn407105	2.1-1		21.1 minutes	16.2 minutes	10.0	2021-06-25 13:05	33.89	61/72	71.5	...
...	...	...	...	...	...	...	...	...	...	...	...	...
Python	Elective course	Sdn407101	2.1-2		15.9 minutes	15.3 minutes	9.5	2021-06-06 18:57	40	72/72	82	...

In SPOC-based learning, because the SPOC course is set up during the pandemic, COVID-19, students are required to complete video learning within the specified time like offline courses, so we selected 109 students and 72 task-point videos' completion degree λ(see Equation 5) and course video score, comprehensive score as feature data. Among them, the course videos' scores and comprehensive scores are standardized(see eq.(1)).

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{1}$$

where X_{max} and X_{min} represented the maximum and minimum values of the set where X is located, respectively, and calculation result is to change the number into the decimal number between [0, 1], which is convenient for subsequent clustering. A total of 74 (72 + 2) dimensions and 8066 (109* 74) space vector value [57] were selected out as feature data for analysis. Each data dimension reflects the learner's learning information in varying degrees.

3.4.1 Dimension-reduced processing

Humans intuitively perceive data in two-dimensional and three-dimensional spaces, and the 74-dimensional representation of learners' learning behavior belongs to high dimensional data. When clustering high-dimensional data, there are drawbacks, i.e., the feature dimension of clustering student objects is large, and the complete object clustering will definitely increase the running cost of clustering algorithm. In addition, the more features are selected, the more comprehensive students' learning behavior is reflected, but not all features are beneficial to clustering, and those that redundant and irrelevant features will make the class structure more blurred [58]. Based on the basic idea of manifold learning, high-dimensional data can show its essential characteristics in low-dimensional state after features' Dimensionality Reduction [59].

The Principal Component Analysis (PCA) algorithm makes the data express the most features with fewer dimensions by transforming the coordinate space [13]. The purpose of dimension reduction is to retain some important features and remove noise and unimportant features, so as to improve the data processing speed and reduce the computational complexity and analysis difficulty in the process of analyzing high - dimensional data. Since the knowledge progressiveness and relevance of the content of task-point video, the closely related content is transformed into as few new variables as possible, so that these new variables are irrelevant and can be used to represent the various types of information that exists in each variable.

3.5 Clustering and analysis

3.5.1 Agglomerative hierarchical clustering

Hierarchical clustering contains a bottom-up aggregation and a top-down splitting strategy [60]. In this paper, the condensed aggregation hierarchical clustering algorithm is selected to treat the feature space vector of each learner after dimensionality reduction as a separate cluster, and then the distance between the two clusters is calculated. Two clusters with the smallest distance between clusters are selected to merge into a new cluster, and the merging operation is repeated until the stop condition is satisfied. Finally, the hierarchical category of all learners' behavior is obtained. Among them, there are two key problems in the algorithm ： First, how to calculate the distance D_{pq} between the two clusters G_p and G_q. Second, how to determine the appropriate clustering number.

On the first key issue, we investigate the method of distance calculation such as Euclidean distance, cosine similarity measure, Mahalanobis distance and correlation coefficient [61, 62], and based on the similar problems encountered before, the experimental experience of the connection rules of singlechain, full-chain, center and group average between clusters is used [63]. D_{pq} takes the average distance within the corresponding cluster, and the distance within the cluster is measured by the cosine similarity measure. The equation of d_{ij} and D_{pq} are defined as follows.

$$d_{ij} = 1 - \frac{\sum_{v=1}^{m} x_{v_i} x_{v_j}}{[\sum_{v=1}^{m} x_{v_i}^2 \sum_{v=1}^{m} x_{v_j}^2]^{\frac{1}{2}}} \tag{2}$$

$$D_{pq} = \frac{1}{n_p n_q} \sum_{x_i \in G_a} \sum_{x_j \in G_a} d_{ij} \tag{3}$$

where d_{ij} represents the distance(1-cosine similarities) between I and j elements belonging to the same k-cluster. The larger the value, the further and dissimilar the elements are, and vice versa. m represents the number of dimensions. n_p, n_q represents the total number of elements of the cluster p and cluster q respectively. x_{v_i}, x_{v_j} represents the learning behavior feature vectors of i and j learners.

Another key problem of clustering analysis is to give a number of reasonable categories. In this paper, we use the cluster diameter to judge the closeness between elements in a cluster the intra-cluster, and determine the number of clusters. The diameter Φ_G of cluster $G = \{x_1, x_2, \ldots, x_n\}$ is calculated are as follows (see eq.(4))

$$\Phi_G = \frac{2}{n*(n-1)} \sum_{x_i \in G} \sum_{x_j \in G} d(x_i, x_j), i! = j, i > j \tag{4}$$

where G_x denotes the x element in the cluster, and $d(x_i, x_j)$ represents the distance between x_i *and* x_j

3.5.2 Learner groups characteristics analysis

The indicators of learners' characteristics analysis are diverse and not unique. For example, Guiyun Zhang [2] select the number of posts and replies, the final score, the total duration of watching videos and the number of watching videos for cluster analysis to identify learners with different learning styles and analyze the characteristics of learners on the four indicators. Shunping Wei [64] find out the characteristics of learners' online learning behaviors such as login, resource browsing, homework and test by collecting data from students' online courses and using data mining methods. Qiangping SONG [19] analyzes and researches online behaviors of SPOC learners, mainly from video viewing, learning documents, problem publishing and answering, homework and so on, and draws three characteristics, which are easy start and difficulties in persistence, emphasis on videos instead of graphs. focus on quizzes instead of process, as well as stress on replies instead of comments. It can be seen that, unlike the traditional offline courses which directly use the single indicator of students' score as the basis for the division of learner characteristics, analysis the online courses combine a variety of analytical indicators, which can comprehensively and objectively reflect the learners' learning characteristics. In this paper, we conduct cluster analysis on the video-viewing data of task points. In order to reduce the high complexity of the analysis process caused by the large number of analysis indicators, we select the completion and viewing-stability of task-point videos as the analysis indicators, and add two indicators of unit test excellence and comprehensive score to comprehensively measure the learning status of learners, so as to analyze the characteristics of learners' learning behavior.

4 RESULTS AND DISCUSSION

4.1 Question 1： Learning behavior log data is diverse, and the more features are selected, the more comprehensive information of learning behavior is reflected, so how to deal with high-dimensional (74 –dimensions) data ?

We perform PCA Dimensionality Reduction on the 74-dimensional learning behavior feature data of learners, and five indicators were statistically analyzed, and the results were shown in Table 2 .

TABLE 2 Results of statistical analysis of PCA dimensionality reduction

	PA1	PA3	PA2	PA11	PA9	PA10	PA6	PA4	PA8	PA7	PA12	PA5	PA15	PA14	PA16	P
SS loadings	15.47	5.06	4.95	4.71	4.57	4.10	3.23	3.16	3.16	2.39	2.30	2.05	1.75	1.27	0.83	(
Proportion Var	0.21	0.07	0.07	0.06	0.06	0.06	0.04	0.04	0.04	0.03	0.03	0.03	0.02	0.02	0.01	(
Cumulative Var	0.21	0.28	0.34	0.41	0.47	0.53	0.57	0.61	0.65	0.69	0.72	0.75	0.77	0.79	0.80	(
Proportion Explained	0.26	0.08	0.08	0.08	0.08	0.07	0.05	0.05	0.05	0.04	0.04	0.03	0.03	0.02	0.01	(
Cumulative Proportion	0.26	0.34	0.43	0.51	0.58	0.65	0.71	0.76	0.81	0.85	0.89	0.92	0.95	0.97	0.99	1

where SS loadings are the eigenvalues associated with principal components, representing the standardized variance values associated with specifically principal components. Proportion Var is the proportion of variance, i.e. how well each principal component interprets the entire data set. Cumulative Var is the sum of the cumulative principal component variance ratios. Proportion Explained and Cumulative Proportion were divided into principal components and cumulative percentage according to the existing total variance percentage. As can be seen from Table 2 , the Cumulative Var of 16 principal components is 0.81, indicating that these principal components can interpret the source data to an extent of 81%, so the first 16 principal components with large correlation after Dimensionality Reduction are used as the principal components for the cluster analysis.

4.2 Question 2： What indicators are used to mine characteristics for learner groups?

4.2.1 Learner groups acquisition

The experiment on the students' behavior data is clustered when K ∈ [1, 20] as different values [65], the clustering results of the average diameter $\bar{\Phi}_G$ of the changes, as shown in Figure 4 . When $\bar{\Phi}_G$ changes steadily, it indicates that the less changes of elements in the cluster, the more suitable to the number of clusters.

Figure 4 can be seen that the stable trend of the number of clusters is [4, 5] and [15, 17], and the balance of [4, 5] is followed by a relatively large range of $\bar{\Phi}_G$ value jitter (drop rapidly), which indicates that the balance at this time is temporary. When the number of clusters is [15, 17], the $\bar{\Phi}_G$ value tends to be stable and there is no significant change in the value. Therefore, we choose [15, 17] to determine the number of clusters, after the initial number of clusters in this range is 15, the $\bar{\Phi}_G$ value almost no longer changes, and the number of clusters K=15 is more appropriate.

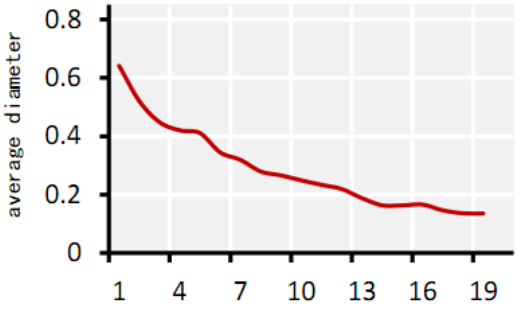


FIGURE 4 Average diameter under different number of clusters

The condensed hierarchical clustering is used to categorize students' learning behavior data in SPOC courses. The results are shown Figure 5 .

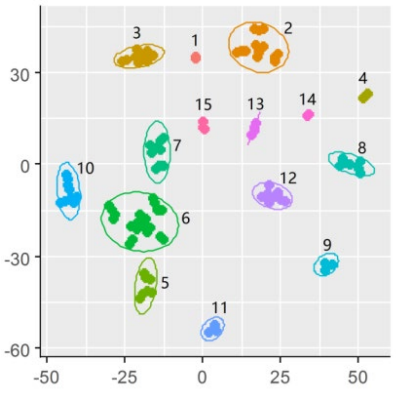


FIGURE 5 Clustering results

Figure 5 shows the two-dimensional visualization of condensed clustering results. Any solids in the figure represent the corresponding 109 learners. The outer circles that surround the solids are clusters, and each cluster represents a different group of learners. To clearly see the clustering results, the learner group is marked with decimal numbers of 1-15. According to the clustering results, learners with similar behaviors are divided into the same type of clusters, and a total of 15 learner groups are obtained. Each type is defined by the indicators of analysis proposed below. It can be seen that the number of learners in different groups is different. For example, the solid points in clusters 1,15,13,14,11,9 and 4 are sparse, and

the corresponding number of learners is small. In cluster 2, 3, 6, 7, 5, 10, 12, 8, there are more solid points and more learners.

4.2.2 The indicators of learner groups characteristics analysis

To identify the characteristics of different learner groups, we analyze the following indicators：

Definition 1 (The completion of task-point video)**.**

There are two ways to measure the completion of the SPOC task-point video. One is whether the task-point video is completed within the specified time, and the other is the total time.it takes to complete the task-point video. In this paper, the task-point video completion of our approach is denoted as λ. λ_i which the completion's degree of the i-th task-point video is as follows：

$$\lambda_i = \frac{TV_i}{TL_i} \tag{5}$$

where TV_i represents the actually viewing time of i-th task-point video, TL_i describes the total time of i-th task-point video. λ can relieve the problem of imbalance caused by the input of viewing time directly due to the unequal duration of each task-point video.

Definition 2 (The viewing-stability of the task-point video)**.**

The viewing-stability of the task-point video is denoted as α, which is a description of n task-point video-viewing during the whole period of learning. The smaller α is, indicating that students complete the learning task uniformly, and the learning state is more stable. The value of α is as follows：

$$\alpha = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\lambda_i - \overline{\lambda})^2} \tag{6}$$

where $\overline{\lambda}$ is a description of the average completion of n taskpoint videos.

Definition 3 (Unit test excellence)**.**

Define students' unit test excellence β_i, the value of β_i is as follows:

$$\beta_i = 1 - \frac{R_i}{M_i} \tag{7}$$

where β_i denotes excellence of a student in the i-th unit test, R_i represents the scores ranking of a student in unit i, M_i denotes the actual number of completions of unit i.

The overall excellence E of students' unit test is defined, and the value of E method is as follows：

$$\mathrm{E} = \frac{\sum_{i=1}^{n}\beta_i * S_i}{\sum_{i=1}^{n} S_i} \tag{8}$$

where S_i represents the total score of unit i, n denotes the total number of task units.

Definition 4 (Comprehensive score)**.**

Comprehensive score is the direct result of learning behavior, which can reflect the level of knowledge and skills acquired by learners. It is an important reference to distinguish the state of learners. There are three ways to measure learners' academic performance： one is to take the final standard test scores as academic performance. The second is to sum the test scores of each unit and the final test scores of the curriculum by weight. The third is the weight summation based on the students' unit test, assignments, completed discussions and final test scores. In order to evaluate the situation of students' learning, we define students' academic performance according to the syllabus and teaching plan, that is, video watching (40%), the scores of unit tests (20%), assignments (15%), curriculum discussion and thinking (5%), final exam grade (20%), and the total mark is a weight average to measure the student's academic performance.

In order to explore learning status and learner groups characteristics in different clusters, the proportion and number of learners in each cluster are counted, and the indicators of learner groups characteristics analysis are summed and divided by the number of people in the cluster. The average completion, average academic comprehensive score, average degree of the viewing-stability and average unit tests' excellence is obtained. The results are shown in Table 3 .

Among the 15 groups obtained by clustering, assuming that learners watch all task-point videos at the maximum speed of two times, the degree of average completion is 0.5. In terms of task completion, 86.3% of the learners can complete the task, while 13.7% of the learners could not. From the perspective of academic performance, the average score of 90 and above accounted for 22.94% of the total number of students, 80 to 89 accounted for 27.53%, 70 to 79 accounted for 13.75%, 60 to 69 accounted for 22.01%, and less than 60 accounted for approximately 13.76%.

4.3 What are the typical characteristics among learner groups?

Figure 6 shows learner groups characteristic indicators boxplot. The box in boxplot contains 50% of the data. The middle line represents the median of the data in the cluster, and the upper and lower edges represent the maximum and minimum values. The points scattered outside the box are outliers, and 25% of the data can be arbitrarily "wild" [66]. The height of the box reflects the fluctuation of the data to a certain extent, the higher the fluctuation is more obvious. Through data analysis and visualization of learner groups characteristic indicators shown in Table 3 and Figure 6 , it is identified into five learner groups types, including positive type, regular type, negative type, semi-negative type and a fluctuation type. Typical characteristics of each learner groups are as follows.

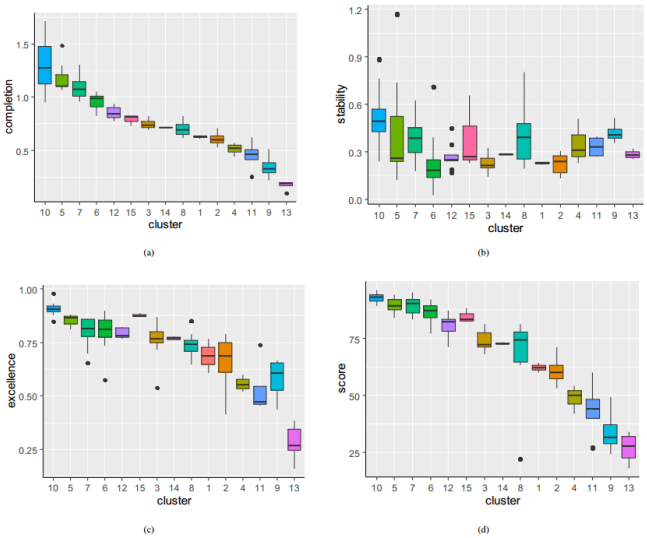


FIGURE 6 (*a*)：Example of Box-line diagram of completion degree. (*b*)：Example of Box-line diagram of stability degree. (*c*)：Example of Box-line diagram of excellent. (*d*)：Example of Box-line diagram of academic achievement.

TABLE 3 Clustering information statistics of learners' behavior data

Clusters label	Proportion(%)	Number of learners	Average completion degree	Average Academic Comprehensive Score	Average viewing stability degree	Average unit test excellence
1	1.83	2	0.63	61.7	0.22	0.69
2	12.84	14	0.60	60.4	0.22	0.67
3	9.17	10	0.75	72.5	0.21	0.76
4	2.75	3	0.51	47.9	0.35	0.56
5	6.42	7	1.18	91.2	0.44	0.85
6	19.27	21	0.97	88.6	0.22	0.80
7	8.26	9	1.09	90.5	0.38	0.80
8	7.34	8	0.70	68.3	0.42	0.74
9	3.67	4	0.35	31.6	0.39	0.58
10	8.26	9	1.29	94.3	0.52	0.91
11	3.67	4	0.45	42.8	0.33	0.53
12	8.26	9	0.86	84.3	0.27	0.79
13	3.67	4	0.17	23.7	0.28	0.28
14	1.83	2	0.72	70.1	0.28	0.77
15	2.75	3	0.79	78.4	0.38	0.88

· Positive type–clusters 10, 5, 7, is "high completion and instability". Specific performance for the completion is more than one hundred percent, and watched all the task-point videos. There are repeated video-viewing behaviors and timely review. From the perspective of stability, the 10th and 5th categories that have achieved good results are groups of the most unstable, and the average degree of stability is higher than 0.4, indicating that the learners spend unevenly on watching taskpoint videos. From the perspective of performance, the unit test has excellent performance and has finally achieved good academic performance. This positive type of students accounts for 23.8% of the total.

· Regular type–clusters 6, 12, is "moderate completion and stability". The specific performance is that the degree of completion exceeds 85%, and basically watched all the task-point videos, and there is no behavior of repeated viewing on the whole. The order of task-point video viewing is basically the same as the order recommended in the curriculum, and learning in the recommended time according to the way of "watching the task point" to "completing the unit test". This type of learning behavior is similar to the learning process in offline teaching. Video-viewing is relatively stable, meanwhile, the stability degree is less than 0.3, and the attention to each task-point video is balanced. The proportion of this kind of behavior is 27.5%, indicating that the learning style of studying in accordance with the established order of the course occupies a dominant position in the *Python Language Foundation*, and most students belong to Regular type.

· Negative type–clusters 4, 11, 9, 13, is "low completion and instability", which is specifically manifested in that the overall completion of all task-point videos below 50%. Some students who are negative type are not learning at all, but that they may selectively watch task-point videos. Even some students only watch a small number of videos or watch videos only one or two. The average stability above 0.3, indicating that this type of students are making it unstable, and most of these students lacking the ability of good self-management. In terms of academic performance, failed to complete all tests or unit test scores are not good, and lead to unsatisfactory final results. This type of learning behavior accounted for 13.7%.

· Semi-negative type–clusters 1, 2, is "low completion and stability", specific performance is just completed the task-point video on the whole, the task-point video has a choice to watch, and students' learning enthusiasm is not high, there is no regularity of repeated video-viewing behavior, individuals cannot complete all learning tasks independently, and test scores and academic performance are poor. This semi-negative type of students accounted for 14.7%.

· Fluctuation type–cluster 15,3,14,8, the specific performance is that in the overall learning process of the course, and there are types of changes in learning behavior, which are regular learning, semi-negative and passive learning in the learning of task point videos, accounting for 21.1%. And the fluctuation rule is generally at the position where the course unit changes, as shown in Figure 7 , which is the curve of completion that the center point of cluster eight. Namely, the curve of completion is that all learners in cluster eight at 72 video-viewing task points. It can be seen that cluster average unit completion degrees values appear obvious peaks or valleys when the unit changes.

In addition, according to the degree of completion, the fluctuating learning behavior can be divided into good volatility learning behaviors (cluster 15), medium volatility learning behaviors (cluster 3), and dangerous volatility learning behaviors (cluster 14,8).

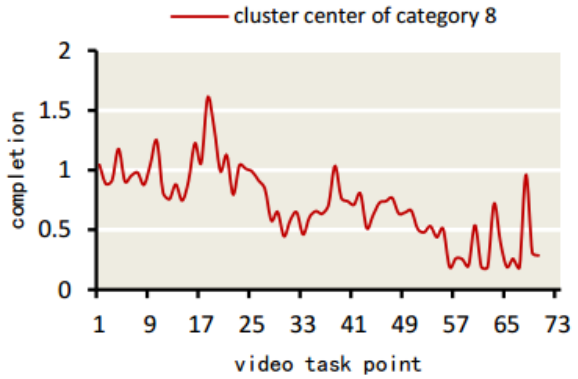


FIGURE 7 The center completion curve of cluster 8

A longitudinal study along different academic years and course are conducted, as shown in Table 4 , we find that for the same course, due to the teaching objects (learners) of different majors, the curriculum category is different. For example, *Python Language Foundation* is designed for students of different majors. The curriculum categories are professional compulsory course and professional elective course. By clustering and analyzing the data of learning behavior, the different learners' group behavioral characteristics are obtained. Among them, there are fewer learners of Agricultural Electrification, and fewer types of learners' group behavioral characteristics in the excavation. Meanwhile, for different professional compulsory courses such as *C programming Language*, this method can still mine and analyze the characteristics of learners' groups, which verifies the feasibility and practicability of this method.

TABLE 4 Different academic years and course statistics of learners' group behavior characteristics

SPOC course name	Curriculum category	Major and class	Number of people	Number of task-point videos	Optimal number of Clusters	Learner characteristics category
Python Language Foundation	Professional Compulsory course	Computer Science and Technology 19(1,2,3)	101	72	8	Regular type, Negative type, Positive type, Fluctuation type, Semi-negative type
Python Language Foundation	Professional Selective course	Agricultural Electrification 21(1)	31	72	5	Regular type, Positive type, Semi-negative type
C Programming Language	Professional Compulsory course	Data Science and Big Data Technology 21(1)	60	71	13	Regular type, Positive type, Semi-negative type, Fluctuation type, Negative type

4.3.1 Learners' group characteristics discussion

Learners' characteristics include Intellectual and Nonintellectual factors. Intellectual factors mainly include the knowledge base, cognitive ability and cognitive structure variables. Non-intellectual factors consist of interests, motivations, emotions, will and character [67].

Due to learners in SPOC online courses can set their own learning pace, and their learning methods are extremely flexible and inconsistent [68], the analysis of underlying causes of above five learner groups characteristics are

analyzed as follows.

· Learners' motivation intensity is different. Some learners expect to see the learning effect and have strong motivation to learn, so they will complete the curriculum requirements as soon as possible. Some learners are not interested in the course content, have weak motivation to learn, delay or even fail to complete learning tasks that are requested.

· Learners' learning styles are different, which can be reflected in the differences of content selection. Some learners will watch SPOC videos and completed homework assignments in the way of learning traditional offline courses. Some learners regard SPOCs as a reference materials, only watching the required videos, selectively downloading materials. Simultaneously, several learners who focus only on tests and immerse in SPOCs, rarely watch videos carefully and use uploaded learning resources effectively.

· Learners' learning behavior is not synchronized. It is asynchronous that mainly manifested in the different time periods when learners choose to learn. Some learners are accustomed to learning in the daytime, while others are more efficient at night.

In addition, all learning tasks are completed in different forms or levels, and some learners cannot complete the learning tasks within the recommended time due to various reasons such as learning environment and family factors. There are also some learners who like to concentrate on completing the course in a fixed time period. Most learners take a step-by-step approach and complete the course gradually.

To help course organizers implement targeted teaching for different students, help adjust the learning process, improve learning efficiency, and ultimately provide personalized learning support services. Above group characteristics of online learning and underlying causes were analyzed, the following measures were taken：

First, for negative and semi-negative learning groups, incentives can be taken to carry out SPOC teaching. For example, sending reminders online [21], re-pushing teaching videos and questions, combined with offline teachers with purposeful and targeted questions to broaden the range of teacher questioning to get students talking[69], and giving appropriate praise. The purpose is to improve the group's completion of task-videos and reduce the instability, so as to avoid the transformation from semi-negative into negative.

Second, for the learning groups of positive and regular types to maintain their enthusiasm, for example, the non-task videos can be appropriately increased, while increasing the number and difficulty of test questions. Teachers purposefully set some challenging topics for groups to choose and complete, and recommend them to participate in skill competitions[70]. The purpose is to remain the task-point videos' completion by such groups and give full play to students' learning enthusiasm, learning rules were used to further improve the skill level, so as to avoid the transformation from regular to fluctuating.

The third for the fluctuating groups,For instance, online learning calendars or curriculum schedules and reminders are added, and offline teachers help students to make planning and urge them to complete on time. The purpose is to improve students' learning regularity and gradually transform fluctuation into regular type.

A final explanation for this paper analyzes learners' behavior characteristics based on clustering and four indicators, which helps to master characteristics of students' group and implement targeted teaching strategies according to feedback results, this is very valuable knowledge for teachers that can be used to motivate groups of students who are less inclined to participate. On the other hand, students' behavior clustering can also be very useful for finding small groups of students with outlier behavior patterns, either because their academic performance is very low, or they have high enthusiasm and strong participation. In either case, teachers can design strategies and take steps to help these group members achieve the best outcome in each course so as to obtain greater personal satisfaction and encourage them to complete the course.

5 CONCLUSIONS AND FUTURE LINES

Our research is a cluster analysis of video learning behavior data of different students in SPOC course, under the conditions that the video-viewing data can reect students' real learning situation, teachers can understand students' situation by analyzing students' video-viewing data[4], especially those with lower abilities, and most of them are students with fluctuating and negative learning behaviors. To do this, we pursue three contributions with this study. First, the selection and processing of feature data are very important. Under the pandemic, COVID-19, video-based online learning has become commonplace in higher education settings. We select the log data that 74-dimensional behavior consisting of seventy-two video-viewing, course video score and comprehensive score, and perform PCA dimensional reduction to obtain mainly feature for high-dimensional features, and eliminate interference with cluster, and improve the clustering efficiency and realize the low-dimension visualization of the clustering results, and provide users with analysis results intuitively. Second, we use agglomerative hierarchical clustering algorithms based on class average to determine the optimal number of clusters. The agglomerative hierarchical clustering algorithm is an effective method for learner behavior analysis [2]. However, the number of clusters is determined based on experience and observation, which is one-sided and lacks a certain degree of confidence. We add the class average diameter statistics to obtain the clustering results under the optimal cluster, which is objective and interpretable. Third, different from the traditional method of single indicator analysis, we propose analysis indicators of learners' group characteristics, including the four indicators of the completion and a viewing-stability degree of task-point video, unit-test excellence, comprehensive score is evaluated comprehensively. The analysis indicators in our study are more comprehensive. Finally, according to analysis indicators, five types of learner groups were identified, and the learners' status and group behavioral characteristics were obtained through quantitative analysis.

This work is the preliminary exploration of students' behavior data clustering under the current SPOCs environments. The issues for future research listed are as follows：

· Instructors should receive organized reminders at the end of each week of the course. These alerts would provide educators with knowledge that struggles to proactively discover and learn in a large number of information

environments. Data mining techniques speed up this process, and collects, processes and analyzes important information of interest(video-viewing data in this example) to mine the knowledge and laws behind the data for teachers to take remedial action to improve the performance of underachievers.

· Students' achievement prediction. Deep Learning is widely used in educational data mining, and trains the original datasets to get the reference model. However, whether the students will or will not meet the course requirements and drop out of the course in the middle are predicted according to the historical data, this will be an interesting way for future research.

· Learner behavior clustering and analysis tool development. The next step we will explore how to integrate the results into the SPOC learning platform, so that even teachers without programming knowledge can more easily obtain each the clustering information and group characteristics of each student visually, thus optimizing the teaching process.

This study has some limitations. E.g., although the selected data is high-dimensional, after the features' dimensional reduction to remove redundancy and retain the principal components, it will still cause the loss of some useful information. As an important form of flipped classroom, SPOC videos are the main form for students to acquire knowledge, in addition to the analysis of students' video-viewing log data, forum discussions and topic discussions in SPOCs deserve further study. Meantime, the factors that affect learning behavior also include family, society and other factors, as well as the influence of students' motivation, purpose and emotion, the application of cluster analysis in teaching needs extensive experimental verification to ensure its accuracy and universality. In conclusion, the current study provides a valuable base to build on.

References

1. Jung Eulho, Kim Dongho, Yoon Meehyun, Park Sanghoon, Oakley Barbara. The influence of instructional design on learner control, sense of achievement, and perceived effectiveness in a supersize MOOC course. *Computers & Education.* 2019;128:377-388.

2. Zhang Guiyun, Zhang Yu, Ran Juan. Research on Clustering Mining and Feature Analysis of Online Learning Behavioral Data Based on SPOC. *2018 13th International Conference on Computer Science Education (ICCSE).* 2018;:1-6.

3. Romero Cristóbal, Ventura Sebastián, García Enrique. Data mining in course management systems: Moodle case study and tutorial. *Computers & Education.* 2008;51(1):368–384.

4. Feng Zhang, LIU Di, Liu Cong. Difficulty-Based SPOC Video Clustering Using Video-Watching Data. *IEICE Transactions on Information and Systems.* 2021;E104.D:430-440.

5. Peng Xian, Han Chengyang, Ouyang Fan, others . Topic tracking model for analyzing student-generated posts in SPOC discussion forums. *International Journal of Educational Technology in Higher Education.* 2020;17:35.

6. Zhi Liu, Shiqi LIU, qing LI, others . Research on Modeling of Learners' Interest Topics and Its Relationship with Learning Outcomes in SPOC Forum. *E-education Research.* 2019;40(12):87-96.

7. Al-Saleem Mona, Al-Kathiry Norah, Al-Osimi Sara, Badr Ghada. Mining Educational Data to Predict Students' Academic Performance. In: Perner Petra, ed. *Machine Learning and Data Mining in Pattern Recognition*, :403–414Springer International Publishing; 2015; Cham.

8. Rodriguez Paula, Duque N, Ovalle Demetrio A. Multiagent System for Knowledge-Based Recommendation of
Learning Objects Using Metadata Clustering. In: Bajo Javier, Hallenborg Kasper, Pawlewski Pawel, et al. , eds. *Highlights of Practical Applications of Agents, MultiAgent Systems, and Sustainability - The PAAMS Collection*, :356–364Springer International Publishing; 2015; Cham.

9. Romero Crist bal, Ventura Sebasti, Delgado Jose Antonio, De Bra Paul. Personalized Links Recommendation Based on Data Mining in Adaptive Educational Hypermedia Systems. In: Duval Erik, Klamma Ralf, Wolpers Martin, eds. *Creating New Learning Experiences on a Global Scale*, :292–306Springer Berlin Heidelberg; 2007; Berlin, Heidelberg.

10. Chen Chih-Ming, Wang Jung-Ying, Chen Yong-Ting, others . Forecasting reading anxiety for promoting English-language reading performance based on reading annotation behavior. *Interactive Learning Environments.* 2016;24(4):681-705.

11. Yang Zong, Qinhua Zheng, Li Chen. Research on the value of Chinese MOOCs learnersOnline learning behavior analysis based on RFM model. *Modern Distance Education.* 2016;02:21-28.

12. Hecking Tobias, Ziebarth Sabrina, Hoppe H. Ulrich. Analysis of Dynamic Resource Access Patterns in a Blended Learning Course. In: LAK '14:173182Association for Computing Machinery; 2014; New York, NY, USA.

13. Jolliffe Ian T, Cadima Jorge. Principal component analysis: a review and recent developments. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences.* 2016;374(2065):20150202.

14. Guo Ping. MOOC and SPOC, which one is better?. *Eurasia Journal of Mathematics, Science and Technology Education.* 2017;13:5961-5967.

15. Fox Armando. From MOOCs to SPOCs. *Commun. ACM.* 2013;56(12):3840.

16. Uijl Sabine, Filius Renée, Ten Cate Olle. Student Interaction in Small Private Online Courses. *Medical Science Educator.* 2017;27(2):237242.

17. Kang Yeqin. An Analysis on SPOC: Post-MOOC Era of Online Education. *Tsinghua Journal of Education ISSN1001-4519.* 2014;35:85-93.

18. Hoffmann R. MOOCs-Best Practices and Worst Challenges. *ACA Seminar Bins sels.* 2013;.

19. Song Qiang-Ping, Fang Hai-Guang, Teng Ying, Jiao BaoCong. Analysis and Research about Characteristics of Online Learning Behaviours of SPOC Learners. *ITM Web of Conferences.* 2016;7:04011.

20. Zhang Wei. Research on Online Learning Behavior Analysis and Its Influencing Factors Based on SPOC Data. In:
:337-344Atlantis Press; 2018.

21. Lara Juan, Lizcano David, Martínez María, Pazos Juan, Riera Teresa. A system for knowledge discovery in elearning environments within the European Higher Education Area Application to student data from Open University of Madrid, UDIMA. *Computers & Education.* 2014;72:2336.
22. Yang Tzu-Chi, Chen Sherry Y., Hwang Gwo-Jen. The influences of a two-tier test strategy on student learning: A lag sequential analysis approach. *Computers & Education.* 2015;82:366-377.
23. Cheng Hercy N. H., Liu Zhi, Sun Jianwen, Liu Sanya, Yang Zongkai. Unfolding online learning behavioral patterns and their temporal changes of college students in SPOCs. *Interactive Learning Environments.* 2017;25(2):176-188.
24. Wang Y. Beck J.. Class vs. Student in a Bayesian Network Student Model.. *AIED 2013: Artificial Intelligence in Education.* 2013;7926:151-160.
25. Vieira Camilo, Parsons Paul, Byrd Vetria. Visual learning analytics of educational data: A systematic literature review and research agenda. *Computers & Education.* 2018;122:119-135.
26. Mariappan Premalatha, Viswanathan V., Suganya Ganesan, Kaviya M., Vijaya Aparna. Educational Data Mining
and Recommender Systems Survey. *International Journal of Web Portals.* 2018;10:39-53.
27. Thai-Nghe Nguyen, Drumond Lucas, Krohn-Grimberghe Artus, Schmidt-Thieme Lars. Recommender system for predicting student performance. *Procedia Computer Science.* 2010;1(2):2811-2819. Proceedings of the 1st Workshop on Recommender Systems for Technology Enhanced Learning (RecSysTEL 2010).
28. Kriani Snjeana. Educational data mining using cluster analysis and decision tree technique: A case study. *International Journal of Engineering Business Management.* 2020;12:1847979020908675.
29. Kaur Parneet, Singh Manpreet, Josan Gurpreet Singh. Classification and Prediction Based Data Mining Algorithms to Predict Slow Learners in Education Sector. *Procedia Computer Science.* 2015;57:500-508. 3rd International Conference on Recent Trends in Computing 2015
(ICRTC-2015).
30. Mueen Ahmed, Zafar Bassam, Manzoor Umar. Modeling and Predicting Students' Academic Performance Using Data Mining Techniques. *International Journal of Modern Education and Computer Science.* 2016;11:36-42.
31. Mubarak Ahmed A., Cao Han, Zhang Weizhen. Prediction of students' early dropout based on their interaction
logs in online learning environment. *Interactive Learning Environments.* 2020;:1-20.
32. Ougiaroglou Stefanos, Paschalis Giorgos. Association Rules Mining from the Educational Data of ESOG WebBased Application. In: Iliadis Lazaros, Maglogiannis Ilias,
Papadopoulos Harris, Karatzas Kostas, Sioutas Spyros, eds. *Artificial Intelligence Applications and Innovations*, :105–114Springer Berlin Heidelberg; 2012; Berlin, Heidelberg.
33. Wang Lidong, Wang Guanghui, Alexander Cheryl Ann. Big Data and Visualization: Methods, Challenges and Technology Progress. *Digital Technologies.* 2015;1(1):33–38.
34. Ueno Maomi, Nagaoka Keizo. Learning Log Database and Data Mining system for e-Learning OnLine Statistical Outlier Detection of irregular learning processes. In: ; 2002.
35. Chen Jeng-Fung, Hsieh Ho-Nien, Do Quang Hung. Predicting Student Academic Performance: A Comparison of Two Meta-Heuristic Algorithms Inspired by Cuckoo Birds for Training Neural Networks. *Algorithms.* 2014;7(4):538–553.
36. Baradwaj Brijesh, Pal Saurabh. Mining Educational Data to Analyze Students' Performance. *International Journal of Advanced Computer Science and Applications.* 2011;2:63-69.
37. Ran Juan, Zhang GuiYun, Zheng Tong, Wang Wenhui. Logistic Regression Analysis on Learning Behavior and Learning Effect Based on SPOC Data. In: :1-5; 2018.
38. Silva Carla, Fonseca Jose. Educational Data Mining: A Literature Review. In: Rocha lvaro, Serrhini Mohammed, Felgueiras Carlos, eds. *Europe and MENA Cooperation Advances in Information and Communication Technologies*, :87–94Springer International Publishing; 2017; Cham.
39. Barlybayev Alibek, Sharipbay Altynbek, Ulyukova Gulden, Sabyrov Talgat, Kuzenbayev Batyrkhan. Student's Performance Evaluation by Fuzzy Logic. *Procedia Computer Science.* 2016;102:98-105. 12th International Conference on Application of Fuzzy Systems and Soft Computing, ICAFS 2016, 29-30 August 2016, Vienna, Austria.
40. Liu min. The Analysis of Learning Behavior Based on SPOC Data. Master's thesisCentral China Normal University2015.
41. Hou Huei-Tse. Integrating cluster and sequential analysis to explore learners flow and behavioral patterns in a simulation game with situated-learning context for science courses: A video-based process exploration. *Computers in Human Behavior.* 2015;48:424-435.
42. Yang Chongyang. Joint Modeling of Learners' Emotions and Topics in SPOC Forum. Master's thesisCentral China Normal University2019.
43. L Zhi., Wang Y., Zheng N., others . Analyses of Differences on Learners' Behavioral Sequences in a College's SPOC Environment. *China Educational Technology.* 2017;(7):8.
44. Jui long Hung, Yu chang Hsu, Kerry Rice. *Integrating Data Mining in Program Evaluation.* 2012.
45. Ramesh A, Kumar S. H, Foulds J, others . Weakly Supervised Models of Aspect-Sentiment for Online Course Discussion Forums. In: :74-83Association for Computational Linguistics; 2015.
46. Yukselturk Erman, Ozekes Serhat, Türel Yaln. Predicting Dropout Student: An Application of Data Mining Methods in an Online Education Program. *European Journal of Open, Distance and E-Learning.* 2014;17:118-133.
47. Estrada Mlb, Cabada R. Z., Bustillos R. O., Graff M.. Opinion mining and emotion recognition applied to learning environments. *Expert Systems with Applications.* 2020;150:113265.
48. Zhi Liu, Chongyang Yang, Xian Peng, others . Learners

Emotional Characteristics in SPOC Forums and Their Association with Learning Effect. *E-education Research.* 2018;04:102-110.
49. Liu Z., Zhang W., Sun J., Liu S., Xian P., Hao Z.. Analysis of Learners' Interactive Discourse Behaviors in the Cloud Classroom Forum. *E-education Research.* 2016;37(09):95-102.
50. Liu Sannyuya, Xian Peng, Hercy N. H. Cheng, others . Unfolding Sentimental and Behavioral Tendencies of Learners' Concerned Topics From Course Reviews in a MOOC. *Journal of Educational Computing Research.* 2019;57(3):670-696.
51. Dunhong Yao, Xiaowu Deng. An Learning Situation Early Warning Method Based on Linear Regression. In: :354-357; 2020.
52. Fadli A., Zulfa M. I., Ramadhani Y.. Performance Comparison of Data Mining Classification Algorithms for Early Warning System of Students Graduation Timeliness. *Jurnal Teknologi dan Sistem Komputer.* 2018;6(4):158.
53. Melendez Armenta Roberto, Huerta Pacheco Nery, Morales Rosales Luis, Rebolledo Mendez Genaro. How Do Students Behave When Using A Tutoring System? Employing Data Mining to Identify Behavioral Patterns Associated to The Learning of Mathematics. *International Journal of Emerging Technologies in Learning (iJET).* 2020;15:39.
54. Li Nan, Kidzi ski ukasz, Jermann Patrick, Dillenbourg Pierre. MOOC Video Interaction Patterns: What Do They Tell Us?. In: Conole Grinne, Klobuar Toma, Rensing Christoph, Konert Johannes, Lavou Elise, eds. *Design for Teaching and Learning in a Networked World*, :197–210Springer International Publishing; 2015; Cham.
55. Stöhr Christian, Stathakarou Natalia, Mueller Franziska, Nifakos Sokratis, McGrath Cormac. Videos as learning objects in MOOCs: A study of specialist and nonspecialist participants' video activity in MOOCs. *British Journal of Educational Technology.* 2019;50:166-176.
56. Zeng , Xianlu , Yu , et al. The construction and online/offline blended learning of small private online courses of
Principles of Chemical Engineering. *Computer Applications in Engineering Education.* 2018;.
57. Jing L., Ng M. K., Huang J. Z.. Knowledge-based vector space model for text clustering. *Knowledge & Information Systems.* 2010;25(1):35-55.
58. Bin LIAO, Jing lai HUANG, Xin WANG, na SUN, Xiao yan GE, Bing lei GUO. SCEA: A Parallel Clustering Ensemble Algorithm for High- Dimensional Massive Data. *ACTA ELECTONICA SINICA.* 2021;49(6):1077-1087.
59. Li Xuelong, Chen Mulin, Wang Qi. Quantifying and Detecting Collective Motion in Crowd Scenes. *IEEE Transactions on Image Processing.* 2020;29:5571-5583.
60. Heller Katherine A, Ghahramani Zoubin. Bayesian Hierarchical Clustering. In: ICML 05:297304Association for ´ Computing Machinery; 2005; New York, NY, USA.
61. Starczewski Artur. A New Hierarchical Clustering Algorithm. In: Rutkowski Leszek, Korytkowski Marcin, Scherer Rafał, Tadeusiewicz Ryszard, Zadeh Lotfi A., Zurada Jacek M., eds. *Artificial Intelligence and Soft Computing*, :175–180Springer Berlin Heidelberg; 2012; Berlin, Heidelberg.
62. Chun liu WANG, Yong hui YANG, Fei DENG, Hui-yuan LAI. A Review of Text Similarity Approaches. *Information Science.* 2019;37(3):158-168.
63. Qiang Ma. Research on the Standardization Method of Course Knowledge in Big Data Environment. Master's thesisNortheast Petroleum University2019.
64. Shunping WEI. An Analysis of Online Learning Behaviors and Its Influencing Factors: A Case Study of Students Learning Process in Online Course " Open Education Learning Guide" in the Open University of China. *Open Education Research.* 2012;18(4):11.
65. Yoon Meehyun, Lee Jungeun, Jo Il-Hyun. Video learning analytics: Investigating behavioral patterns and learner clusters in video-based online learning. *The Internet and Higher Education.* 2021;50:100806.
66. Kürzl Hans. Exploratory data analysis: recent advances for the interpretation of geochemical data. *Journal of Geochemical Exploration.* 1988;30(1):309-322.
67. Justice Kintu, Zhu Chang. Blended learning effectiveness: the relationship between student characteristics, design features and outcomes. *International Journal of Educational Technology in Higher Education.* 2017;14.
68. Patru Mariana, Balaji Venkataraman. *Making sense of MOOCs: a guide for policy-makers in developing countries*. The United Nations Educational, Scientific and Cultural Organization (UNESCO); 2016.
69. DeJarnette Anna F., Wilke Edana, Hord Casey. Categorizing mathematics teachers questioning: The demands and
contributions of teachers questions. *International Journal of Educational Research.* 2020;104:101690.
70. Liebenberg Leon, Mathews Edward. Integrating innovation skills in an introductory engineering design-build course. *International Journal of Technology and Design Education.* 2012;22:93-113.
71. Maldonado-Mahauad Jorge, Pérez-Sanagustín Mar, Kizilcec René F., Morales Nicolás, Munoz-Gama Jorge. Mining theory-based patterns from Big data: Identifying selfregulated learning strategies in Massive Open Online Courses. *Computers in Human Behavior.* 2018;80:179-196.
72. Kizilcec René F., Pérez-Sanagustín Mar, Maldonado Jorge J.. Self-regulated learning strategies predict learner behavior and goal attainment in Massive Open Online Courses. *Computers & Education.* 2017;104:18-33.
73. Grundspenkis Janis, Anohina Alla. Evolution of the Concept Map Based Adaptive Knowledge Assessment System: Implementation and Evaluation Results. *Scientific Journal of Riga Technical University.* 2009;38(38):1324.
74. Grundspenkis Janis. Concept Map Based Intelligent Knowledge Assessment System: Experience of Development and Practical Use. In: Ifenthaler Dirk, Spector J. Michael, Isaias Pedro, Sampson Demetrios, eds. *Multiple Perspectives on Problem Solving and Learning in the Digital Age*, :179–197Springer New York; 2011; New York, NY.
75. Novak Joseph D, Cañas Alberto J. The Origins of the Concept Mapping Tool and the Continuing Evolution of the Tool. *Information Visualization.* 2006;5(3):175-184.
76. Davis Dan, Chen Guanliang, Zee Tim, Hauff Claudia, Houben Geert-Jan. Retrieval Practice and Study Planning in MOOCs: Exploring Classroom-Based Self-regulated

Learning Strategies at Scale. In: Verbert Katrien, Sharples Mike, der Zee, Claudia Hauff, Geert Jan Houben, eds. *Adaptive and Adaptable Learning*, :57–71Springer International Publishing; 2016; Cham.
77. Martnez A., Dimitriadis Y., Rubia B., Gómez E., de la Fuente P.. Combining qualitative evaluation and social network analysis for the study of classroom social interactions. *Computers & Education.* 2003;41(4):353-368. Documenting Collaborative Interactions: Issues and Approaches.
78. Han Guihua, Lin Mingyu, Li Cuilin, Ju Jianping. Flipped classroom teaching design based on SPOC in ordinary undergraduate college. In: :617-620; 2017.
79. Romero Cristóbal, Ventura Sebastián. Educational Data Mining: A Review of the State of the Art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews).* 2010;40(6):601-618.
80. Abuteir Mohammed, El-Halees Alaa. Mining Educational Data to Improve Students Performance: A Case Study. *International Journal of Information and Communication Technology Research.* 2012;2:140-146.
81. Stefan Slater, Sreko Joksimovi, Vitomir Kovanovic, Ryan S. Baker, Dragan Gasevic. Tools for Educational Data Mining: A Review. *Journal of Educational and Behavioral Statistics.* 2017;42(1):85-106.
82. Dutt Ashish, Ismail Maizatul Akmar, Herawan Tutut. A Systematic Review on Educational Data Mining. *IEEE Access.* 2017;5:15991-16005.
83. Khalil Mohammad, Kastl Christian, Ebner Martin. Portraying MOOCs Learners: a Clustering Experience Using Learning Analytics. In: :265–278BoD; 2016. EMOOCs 2016 - European Stakeholders Summit ; Conference date: 22-02-2016 Through 24-02-2016.
84. cai Yan. My exploration and pursuit in the field of intellectual factors and non-intellectual factors. *Education Sciences in China(In Chinese and English).* 2019;02(03):3-8.
85. Siemens George, Long Phil. Penetrating the Fog: Analytics in Learning and Education.. *EDUCAUSE review.* 2011;46(5):30.
86. L Wu, Q Liu, H Huan, L Man, J Huang. The Design and Development of Educational Resources Clustering System Oriented to e-Learning. *China Educational Technology.* 2014;.
87. Chi Michelene, Wylie Ruth. The ICAP Framework: Linking Cognitive Engagement to Active Learning Outcomes. *Educational Psychologist.* 2014;49(4):219-243.
88. Ran Chen, Cheng Yang. Blended Learning for SPOC. *Distance Education In China.* 2015;(5):7.
89. Chen Yibo, Wu Chanle, Guo Xiaojun, Wu Jiyan. Semantic Learning Service Personalized. *International Journal of Computational Intelligence Systems.* 2012;5:163-172.