An Automatic Classification and Clustering Algorithm for Online Learning Goals Based on Cognitive Thinking

https://doi.org/10.3991/ijet.v13i11.9587

Ying Wang(✉), Weifeng Jiang
Tianjin Maritime College, Tianjin, China
wangying90182@126.com

Abstract—To improve the learning effect of online learning, an online learning target automatic classification and clustering analysis algorithm based on cognitive thinking was proposed. It was applied to a multi-dimensional learning community. A new form of virtual learning community concept was proposed. The design ideas of its multi-dimensional learning environment were elaborated. Ontology technology was used to collect student learning process data. A cognitive diagnostic model for assessing student learning status was generated. Finally, through the cluster analysis technology, the registered students in the curriculum center were automatically divided into different levels of community groups. The results showed that the proposed algorithm for automatic classification and clustering of online learning targets had a good application effect in the learning community. Therefore, this method has practical application value.

Keywords—online learning, cluster analysis, virtual learning community, cognitive diagnosis model

1 Introduction

With the development of computer technology and networks, the number of students who have received academic education or formal training through the Internet has increased dramatically. According to statistics from the Ministry of Education website, online learners have increased 4 times over the past 5 years. Online learning has become the main learning method of continuing education and lifelong learning. To realize collaborative learning, knowledge building and smart development, the virtual learning community, an ideal learning environment, is being concerned by many educational technology researchers and Internet developers.

E-learning is a very broad concept that represents a large learning system. From the point of view of the composition system, it includes learning management systems, content management systems, learning object libraries, and virtual learning communities. Among them, the virtual learning community is an important application mode in the E-learning environment. In general, the virtual learning community is also referred to as the E-learning community on the Internet. Compared with the traditional e-learning system, the virtual learning community emphasizes the concept of interaction. It belongs to the category of collaborative learning. In the virtual learning community, the distribution of learners and the transfer of knowledge are separated in time and space. Through collaboration among learners and effective guidance of teachers, problems in distance learning are overcome to achieve better learning results.

In the face of large-scale student data information, the traditional E-learning environment is difficult to provide teachers with accurate and useful information. It cannot provide a personalized and adaptable learning environment for large-scale distance learners. Therefore, these systems have some defects in smart guidance. It is one of teachers' responsibility to guide learners in collaborative learning. Teachers must be responsible for the process of implementing and controlling collaborative learning activities. Cluster analysis technology is applied in the intelligent learning system, which can acquire student's learning process data from the background system. Meaningful features were extracted to generate models that correctly assessed their learning status. Based on the model, cluster analysis can effectively classify students.

2 State of the art

Network is widely used. Information technology has developed rapidly. The society is very concerned about education. The field of personalized intelligent online learning platform has attracted a large number of domestic and foreign scholars to study it. One of the research topics is still to categorize users so as to achieve personalized guidance. Shaw et al. [1] introduced personalized learning into traditional learners and non-traditional learners through an asynchronous learning platform, enabling learners to advance to individual learning goals at their own pace. The platform was built using a guided learning path (GLP) to provide thread user scenarios for each module. It can help readers to achieve visualization. Andruseac et al. [2] proposed an e-learning platform for personalized treatment and online monitoring of learning disabilities patients. By developing new methods that support ICT, patients' cognitive functions are improved. Xi et al. [3] established a personalized adaptive learning behavior analysis model. Based on this model, a personalized MOOC platform was designed. By analyzing the learning behavior on the MOOC platform, the model mines the pattern of learning behavior. This lays the foundation for personalized intervention in the learning process. Xiao et al. [4] proposed a personalized recommendation system for online learners. By using a combination of association rules, content filtering, and collaborative filtering, the system recommends learning resources. It can improve the utilization of educational resources and improve students' learning autonomy and efficiency. By considering the personalized activity recognition problem as a multi-task learning problem, Sun et al. [5] establish models for different human activities. A new online multitasking learning method is proposed to identify large-scale personalized activities. Compared with the existing multitasking learning hypothesis task relationship work, it can automatically discover task relationships from real-world data. Lee [6] argues that the formalization of folk taxonomy in annotation learning resources as well as in the perspective of demographics to rating resources can not only support personalization, but also enable learners to participate more effectively in technology-enhanced learning. Shih [7] studied the human factor. Gender differences and cognitive styles influence learners' responses. Based on a personalized and non-personalized learning system, learner's prior knowledge is studied. Women and serializers respond positively to personalized learning systems. Men and holists showed similar reactions to personalized learning systems and non-personalized learning systems. These results have an important influence on the design of the personalized learning system. Cavus and Zabadi [8] studied the comparative communication tools of the six open source learning management systems (LMS). By selecting a learning management system with the best communication tools, teachers are more likely to make the best choice when choosing a learning management system.

To sum up, most of the researches are based on the establishment of platform and the realization of personalized modules. Few scholars have applied the cluster analysis method to learner's learning activity analysis and target classification. In view of the above situation, the cluster analysis algorithm is applied to the online learning target classification to classify the students' cognitive diagnosis model. Students with different levels of learning are divided into different community learning groups to provide learners and teachers with more personalized and intelligent services.

3 Methodology

3.1 Data acquisition of learning process based on Ontology technology

The ontology concept refers to the course ontology. Course ontology is also called the concept hierarchy. It represents the outline and framework of the course content. A course ontology is defined as a tree R. Among them, TS denotes a set of terms (term) TS = {term1, term2, ..., termn}, and R denotes a binary relationship on the TS set. In the course ontology, the keywords may represent chapters, sections, or key concepts in the course content. The dimensions of the TS set are determined by the teacher.

In an Ontology-based multi-mode interactive network learning environment, real-time, dynamic, and rapid acquisition of learner's learning process data is needed. A model that can feed back learner learning status in real time is generated. Based on this, learners are analyzed. In a general E-learning environment, in order to collect learners' learning process data, the system framework of the acquisition process is generally designed as shown in Figure 1.

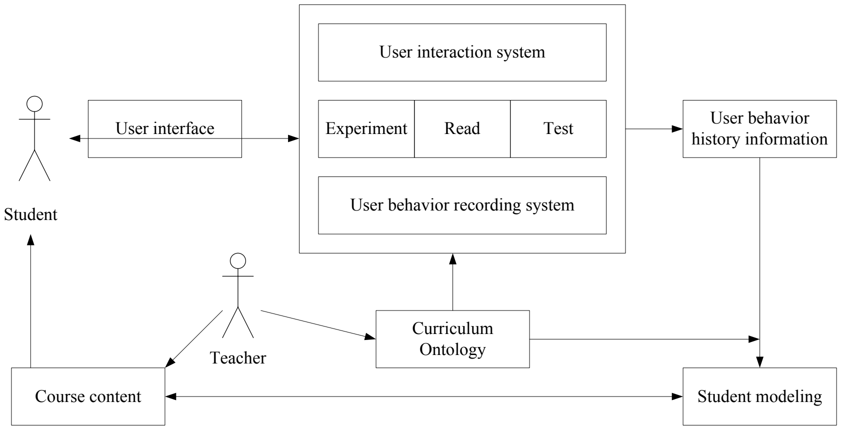


Fig. 1. System framework for acquiring learning process data

The above system architecture can be embedded in any E-learning system to achieve user-adaptive or personalized tutoring. The main components of the system framework include course ontology, user learning process data, and student modeling. Course ontology is also called a concept hierarchy or a knowledge structure. It represents the appropriate granularity and scale of course content predefined by teachers (or expert instructors). It is the basis for generating the learner's knowledge state model. The learning process data mainly includes behavior records and test results generated by students during the learning, reading and testing processes. These process data are generated by the student behavior record system based on the pre-defined curriculum ontology of the teacher. These students' behavior records will be recorded and accumulated as historical data to generate student learning process data models. Based on the historical data of student behavior records, a student learning process model was established. Through the model, students' knowledge needs for curriculum ontology are analyzed.

3.2 Generation of cognitive diagnostic models

In recent years, some cognitive diagnosis models have become the mainstream method for assessing the knowledge of students. The relevant content of the theoretical model of the Q-matrix is described in detail. Based on the Q-matrix and student learning process data, a corresponding cognitive diagnosis model is generated.

The student's test question responses reflect the knowledge state, which is the original intention of the Q-matrix. The Q-matrix is a binary (0/1) matrix that represents the relationship between a test item and a potential attribute or concept. Students will be divided into different knowledge states based on their test answers and the Q-matrix developed by the experts. The Q-matrix is a two-dimensional matrix. One dimension represents a question, which is represented by Q_n, and the other dimension represents a concept, which is represented by Con_n. If a problem is related to a concept, the value in the corresponding Q-matrix is 1, and otherwise it is 0. Table 1 is an example of a binary Q-matrix. Both Q1 and Q6 are related to Concept 1 (Con_1). The value in the corresponding matrix is 1. They are not related to Concept 2 (Con_2). Therefore, the value in the corresponding matrix is 0.

Table 1. Q- matrix example

Concept	**Question**					
	Q₁	***Q₂***	***Q₃***	***Q₄***	***Q₅***	***Q₆***
Con_1	1	0	0	1	0	1
Con_2	0	1	0	1	0	0
Con_3	1	1	1	0	1	0
Con_4	1	0	1	0	0	0

In some cases, the value of the Q-matrix is 0 or 1. However, the range of values in the matrix can also be any value between [0,1]. It represents the probability that a student answers a question wrong without knowing a concept.

Several new cognitive diagnostic models will be introduced based on the Q-matrix basis. Suppose there is a Q-matrix of size J×K, J represents the total number of questions, and K represents the total number of concepts. If the problem j involves the concept k, the value of q_{jk} in the corresponding Q-matrix is 1, and otherwise the value is 0.

$$Q=\begin{bmatrix} q_{1,1} & q_{1,2} & \cdots & q_{1,\mathrm{J}} \\ & . \; . \; . \; . & & \\ q_{\mathrm{K},1} & q_{\mathrm{K},2} & \cdots & q_{\mathrm{K},\mathrm{J}} \end{bmatrix} \tag{1}$$

Q-matrix has two different ways. The first one depends on the number of concepts involved in each problem. Some problems may only involve a single concept, and some problems involve multiple concepts. Some Q-matrices include all issues involving a single concept. Some Q-matrices include all issues involving multiple concepts, and others include issues involving a single concept and multiple concepts. The second difference depends on the balance of the Q-matrix. The so-called balance means that in the matrix, the problem types of a single concept appear the same number of times, or the problem types of multiple concepts appear the same number of times.

According to the status of the answer question, the student is constructed as a matrix Y of size N×J. N indicates the number of students and J indicates the number of questions. Student n answers question j. If the answer is correct, the corresponding y_{nj} value in matrix Y is 1; otherwise, it is 0. If student n does not answer question j, then y_{nj}=NA.

$$Y=\begin{bmatrix} y_{1,1} & y_{1,2} & \cdots & y_{1,\mathrm{J}} \\ & . \; . \; . \; . & & \\ y_{\mathrm{N},1} & y_{\mathrm{N},2} & \cdots & y_{\mathrm{N},\mathrm{J}} \end{bmatrix} \tag{2}$$

The DINA model is a simple common conjunctive cognitive diagnosis that can be used to assess a student's understanding of knowledge. The concrete expression of its model is as follows:

$$P\left(Y_{ij}=1\,|\,\eta_{ij},s_j,g_j\right)=\left(1-s_j\right)^{\eta_{ij}}g_j^{1-\eta_{ij}} \tag{3}$$

In the formula, η_{ij} indicates whether student i has mastered all the concepts involved in question j. If the student has mastered all the concepts, the value is 1; otherwise, the value is 0. s_j is a smoothing parameter, which indicates the probability of students being wrong in mastering the relevant concepts. g_j is a guess parameter, which indicates that students have the correct probability without grasping relevant concepts.

The main idea of the Sum-scores model is to use a vector $W_i=(w_{i1},w_{i2},...,w_{ik})$ to represent the "total score" of a student answering a question. Its model expression is as follows:

$$w_{ik}=\sum_{i=1}^{J}y_{ij}q_{jk} \tag{4}$$

In the formula, y_{ij} and q_{jk} are the corresponding elements in the matrix Y of the student's answer question and the Q-matrix developed by the expert. W_i indicates the number of questions correctly answered by student i for each concept k.

The third kind of cognitive diagnostic model for assessing students' knowledge points is the capability matrix. The capability matrix B is a matrix of size N×K. N is the number of students and K is the number of concepts in the Q-matrix. The element B_{ik} in matrix B represents the proportion of students i correctly answering the question contained in concept k. The expression of its model is as follows:

$$B_{ik}=\frac{\sum_{j=1}^{J}I_{y_{ij\neq NA}}\cdot y_{ij}\cdot q_{jk}}{\sum_{j=1}^{J}I_{y_{ij\neq NA}}\cdot q_{jk}} \tag{5}$$

In the formula, y_{ij} and q_{jk} are the corresponding elements in the matrix Y of the student's answer question and the Q-matrix developed by the expert. $I_{y_{ij\neq NA}}$ indicates that student i answered question j. If the student does not answer any questions concerning the concept k, then B_{ik}=NA.

3.3 Cluster analysis of the model

Cluster analysis refers to the process of grouping collections of physical or abstract objects into multiple classes consisting of similar objects. It is an important human behavior. The goal of cluster analysis is to collect data on a similar basis for classification. Taking into account the relationship of the entire system structure, the K-means algorithm is used as a mining module to implement the classification algorithm.

The k-means algorithm is a centroid-based algorithm. It takes k as a parameter and divides n objects into k clusters, which makes the clusters have higher similarity and the similarity among clusters is the lowest. The calculation of similarity is based on the average of the objects in a cluster (which is considered as the center of gravity of the cluster). The specific process of the K-means clustering algorithm is as follows: First, the k centroids C_1, C_2,..., C_K are selected as the initial cluster center from the data set.

Then, each object is assigned to the cluster where the closest cluster center is. After all the objects have been allocated, the centroid C_K of each cluster is recalculated. The above steps are executed cyclically until the division of the data no longer changes.

In order to achieve a cluster analysis based on cognitive diagnostic models for students, students in different learning states are divided into different levels of learning community groups in the virtual learning community. Data analysis tool of Microsoft's SSAS was used to design organizational solutions. The data mining analysis module is part of the data service module. Its relationship with the curriculum center platform and the virtual learning community as well as the specific program flow for clustering grouping is shown in Figure 2.

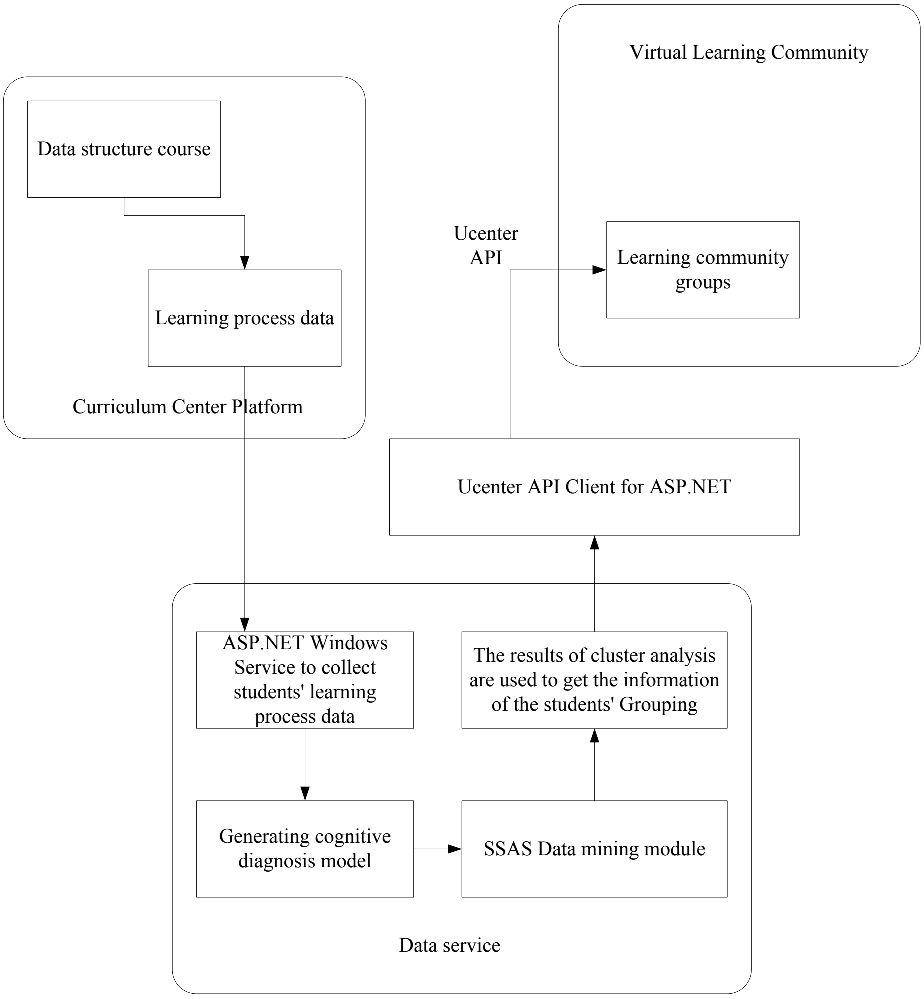

Virtual Learning Community
Data structure course
Ucenter API
Learning community groups
Learning process data
Curriculum Center Platform
Ucenter API Client for ASP.NET
ASP.NET Windows Service to collect students' learning process data
The results of cluster analysis are used to get the information of the students' Grouping
Generating cognitive diagnosis model
SSAS Data mining module
Data service

Fig. 2. The process of clustering grouping

3.4 The design of virtual learning community

Based on the platform of the Shanghai Jiao Tong University Network School Curriculum Center, a comprehensive analysis of the back-end data of the curriculum center was conducted on the data structure of the network exquisite course. Two aspects of a hierarchical virtual learning community are constructed. A conceptual model of a learning community that is hierarchically organized into learning communities based on the evaluation of student knowledge points is proposed. The traditional virtual learning community is mainly to provide learners with an informal communication environment. Learners have their own independent personal space in this environment. When students encounter problems in their studies or work, or have a certain interest, they can communicate with people with similar interests to solve problems. With reference to various virtual learning communities on the network at this stage, a typical virtual learning community to be built should have the following main functions. First, after entering the community environment, learners should have their own personal space, which is similar to the personal homepage functions of common SNS communities (such as Renren.com and Kaixin.com). In this space, learners can freely record their own learning and thinking process, communicate with other learners, check their own trajectory in the community, and understand the status of other members of the community. Second, learners can search for friends in the community to facilitate learning and communication. Different from the traditional virtual learning community, learning communities are mainly formed through the way that learners search for friends. The innovative point of this virtual learning community is to cluster and divide learning community groups according to students' knowledge points. According to the clustering results of student learning status assessment, the community will divide students with different levels of learning and ability into different learning community groups to form a learning group. In each learning group, learners with a close learning level will carry out various forms of learning activities. Finally, in the learning community, teachers or students can initiate a form of learning task for the entire community. Other students in the community can choose to follow or participate in the learning task and discuss and publish resources within the learning task section.

The proposed virtual community solution is a system that uses E-learning to provide learners with personalized, intelligent, and systematic E-learning services. In this virtual learning community solution, it mainly consists of three parts: the curriculum center platform, the data mining module for assessing student learning status, and the virtual learning community. Its system structure is shown in Figure 3.

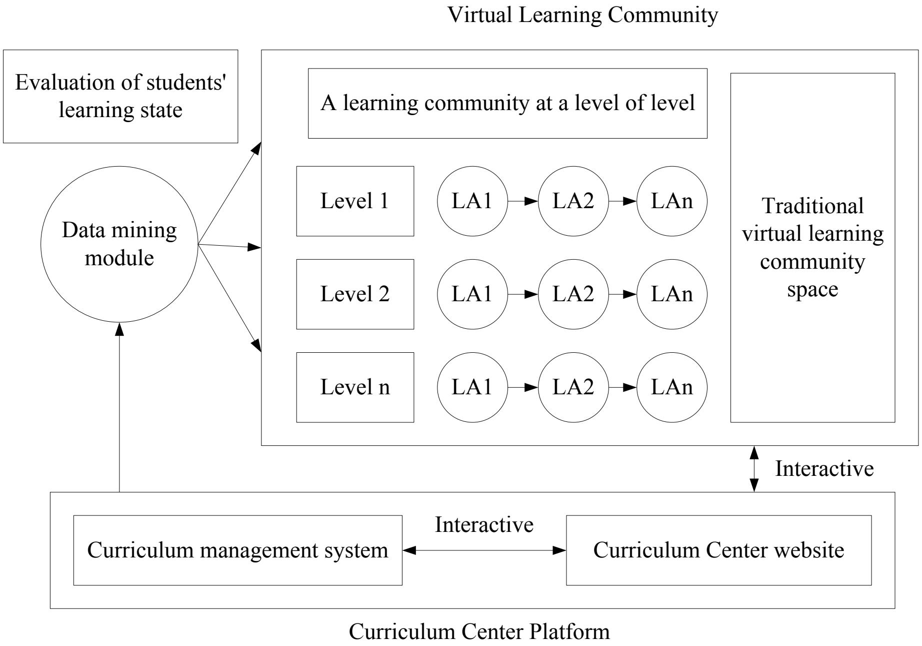

Virtual Learning Community
Evaluation of students' learning state
A learning community at a level of level
Data mining module
Level 1
Level 2
Level n
LA1
LA2
LAn
Traditional virtual learning community space
Interactive
Curriculum management system
Interactive
Curriculum Center website
Curriculum Center Platform

Fig. 3. The scheme design of the virtual learning community system

As shown in Figure 3, the curriculum center platform consists of a curriculum management system and a curriculum center website. Teachers and administrators design and maintain the curriculum center through the management system. Students conduct online learning of a range of resources (such as reading, experimenting, testing, etc.) through the website system. The data mining module mainly analyzes the test scores of students' corresponding knowledge points and completes the assessment of their learning status. Based on the assessment of clustering, students are divided into different levels of learning community groups in the virtual learning community, so that students with similar abilities freely use various community mechanisms to conduct a series of interactive learning activities in the learning community. Teachers can also give different personalized guidance according to their situation.

By acquiring student's learning process data (such as the result of knowledge testing), a corresponding result model is generated. In combination with the Q-Matrix developed by experts, a cognitive diagnostic model that can assess student learning status is generated. Based on this, a series of cluster analysis was performed. The entire assessment process is shown in Figure 4.

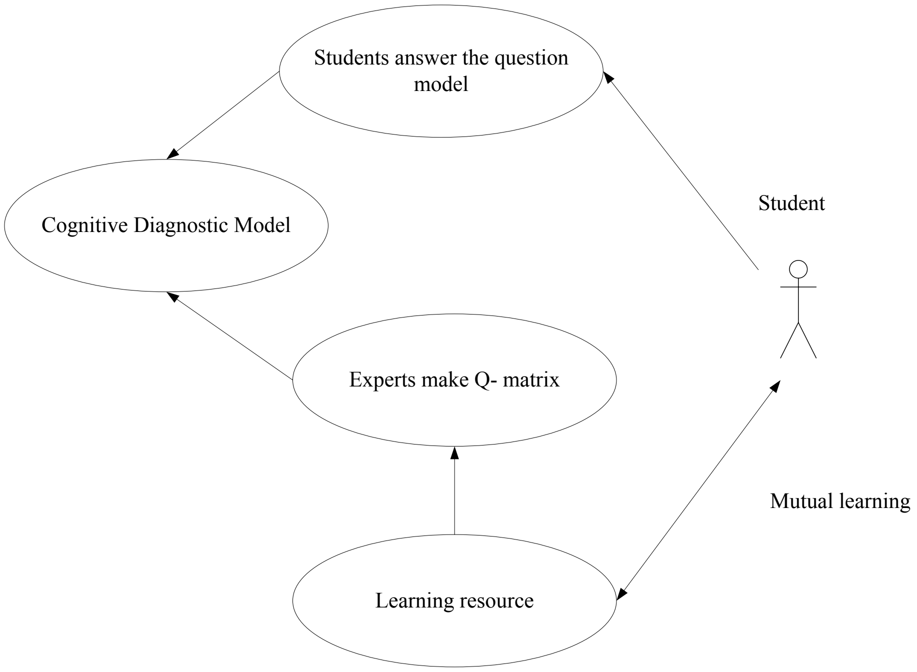


Fig. 4. The evaluation process of the students' learning state

According to the evaluation results of students' learning status, students of different ability levels are systematically clustered into different learning community groups, and different groups are provided with different learning activities for them. Learning activity refers to the sum of various operations that students and teachers need to complete to achieve a specific learning goal. It represents the interaction between the student and the learning environment. At present, the main forms of e-learning activities include six major components such as learning tasks, learning processes, supervision rules, learning support, evaluation rules, and learning resources. The community learning environment is the support of the online learning environment. It is the technical foundation for students to carry out a series of learning activities. Students complete learning activities through interaction with the learning environment. The learning environment mainly includes two levels of elements: One is the material condition that supports the learning process, that is, the hardware environment. It mainly refers to the learning facilities and other aspects. The other is the non-material condition, that is, the software environment. It mainly includes teaching mode, teaching strategy, learning atmosphere, interpersonal relationship and other factors.

At present, the distributed application software architecture based on network environment mainly includes two kinds of C/S structure and B/S structure. The B/S structure is the browser and server structure. It is a change or an improved structure of the C/S structure with the rise of Internet technology. Under this kind of structure, the user working interface is realized through the WWW browser, and a very small part of the transaction logic is implemented in the browser. However, the main transaction logic is implemented on the server to form a three-tier structure. This can simplify the load on the client computer, reduce the cost and workload of system maintenance and upgrade, and reduce the overall cost of the user. The proposed virtual learning community uses a three-tier architecture based on B/S. The three layers are the presentation layer, business logic layer, and data access layer. The overall framework of its system is shown in Figure 5.

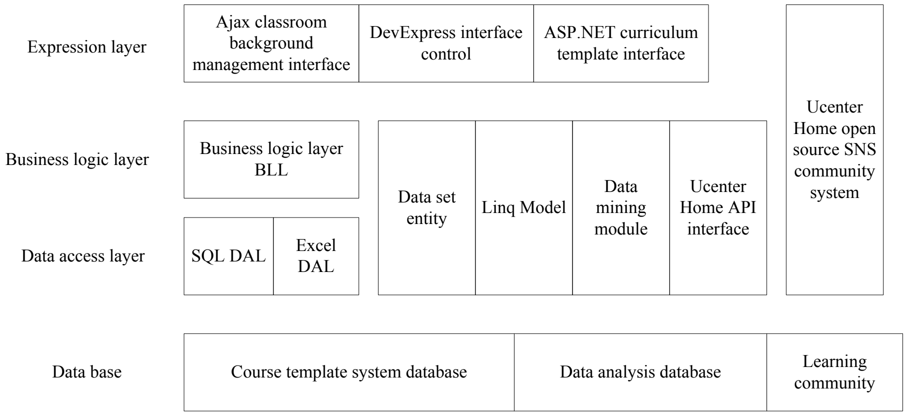

Expression layer
Ajax classroom background management interface
DevExpress interface control
ASP.NET curriculum template interface
Ucenter Home open source SNS community system
Business logic layer
Business logic layer BLL
Data set entity
Linq Model
Data mining module
Ucenter Home API interface
Data access layer
SQL DAL
Excel DAL
Data base
Course template system database
Data analysis database
Learning community

Fig. 5. Framework diagram of the overall system

There are two major components of the presentation layer. One part is the curriculum center webpage developed based on the .Net platform, and the other part is the virtual learning community webpage developed based on PHP. Their main function is to accept learners' requests. After being analyzed and processed by the system, the content is displayed to the learner as a web page. At the same time, the presentation layer needs to provide learners with high-quality human-computer interaction interfaces and provide specific application functions. When designing the presentation layer, professional and aesthetics need to be taken into consideration, and the user experience is enhanced. Third-party controls such as Devexpress9.1.5 and ExtJs can be considered. The business logic layer centralizes the system's main network services and analysis and processing functions. It acts as an intermediary between the client and the database. Its role is to accept the client's request. Then, the results of the database are returned to the Web server. Finally, the result is presented to the client. The data access layer provides the entire virtual learning community with data resources. It includes basic databases such as learners' basic data bases, community activity record data bases, and virtual learning community knowledge bases. It is responsible for querying, modifying, and storing information of the entire community to ensure data security.

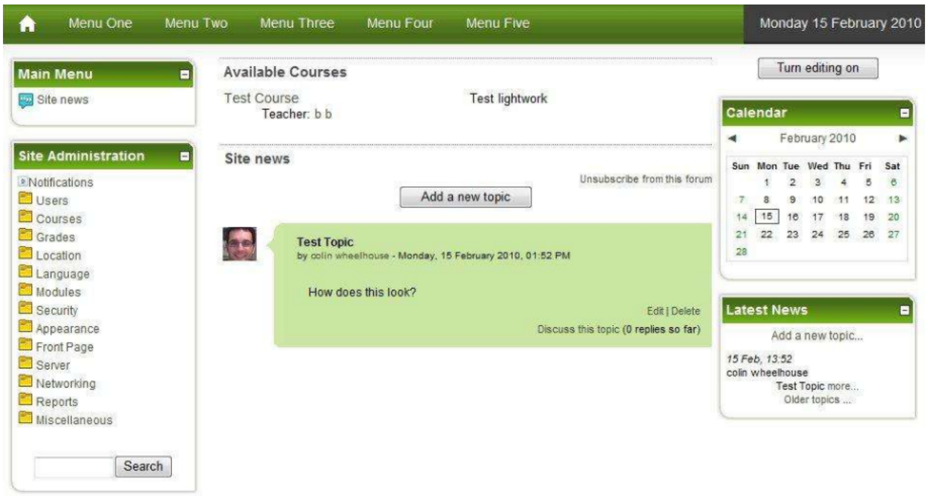
4 Result analysis and discussion

The foundation of the proposed new virtual learning community is Ucenter Home, an open source SNS community software provided by the domestic Kangsheng

Imagination Company. It is a set of social networking software built using PHP+MySQL. An exchange network centered on friends can be easily constructed. Users of the site can easily share information with other friends and discuss topics of interest so that they can quickly learn about the latest developments of their friends. Ucenter Home provides developers with rich application development interfaces such as user interfaces, short message interfaces, friend interfaces, event interfaces, and application interfaces. Developers can easily perform secondary development through these interfaces.

The realization of data management and acquisition technology is mainly achieved through Microsoft's ASP.NET and SQL Server. Visual Studio 2008 provides a wealth of WEB controls and a complete class library. SQL Server 2008 provides a secure, reliable, and efficient platform for data management and business intelligence applications. The integration service and analysis service module can support integrated extraction, transformation and data mining of large-scale, multi-dimensional and complex data.

Figure 6 is an implementation diagram of a constructed virtual learning community.



Test Topic
by colin wheelhouse - Monday, 15 February 2010, 01:52 PM

How does this look?

Discuss this topic (0 replies so far)

Sun	Mon	Tue	Wed	Thu	Fri	Sat
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28						

15 Feb, 13:52
colin wheelhouse
Test Topic more...
Older topics ...

Fig. 6. Implementation diagram

5 Conclusions

A new design scheme using educational data mining technology is proposed. Different competence level learning community groups in the virtual learning community are constructed. The theory of cognitive diagnostic model for assessing student learning status is introduced in detail. Several major clustering algorithms are applied. It provides enough support for educational theory and technology implementation methods. A new form of virtual learning community concept is proposed. The design ideas of its multi-dimensional learning environment are elaborated. Ontology technology is used to collect students' learning process data and generate a cognitive diagnosis model to assess student learning status. Finally, the clustering analysis technology is used to automatically classify registered students in the curriculum center into community groups at different levels. The results show that the proposed algorithm for automatic classification and clustering of online learning targets has a good application effect in the learning community. Therefore, it has practical application value.

6 References

[1] Shaw, C., Larson, R., Sibdari, S. (2014). An Asynchronous, Personalized Learning Platform—Guided Learning Pathways (GLP): Creative Education, 5(13): 1189-1204 https://doi.org/10.4236/ce.2014.513135

[2] Andruseac, GG., Rotariu, D., Rotariu, C., Costin, H. (2013). eLearning Platform for Personalized Therapy of Learning Disabilities: Procedia - Social and Behavioral Sci-ences, 83(7): 706-710 https://doi.org/10.1016/j.sbspro.2013.06.133

[3] Xi, J., Chen, Y., Wang, G. (2018). Design of a Personalized Massive Open Online Course Platform: International Journal of Emerging Technologies in Learning, 13(4): 58 https://doi.org/10.3991/ijet.v13i04.8470

[4] Xiao, J., Wang, M., Jiang, B., Li, J. (2017). A personalized recommendation system with combinational algorithm for online learning: Journal of Ambient Intelligence & Humanized Computing, 2017(1): 1-11

[5] Sun, X., Kashima, H., Ueda, N. (2013). Large-Scale Personalized Human Activity Recognition Using Online Multitask Learning: IEEE Transactions on Knowledge & Data Engineering, 25(11): 2551-2563 https://doi.org/10.1109/TKDE.2012.246

[6] Lee, CS. (2015). A folksonomy-based lightweight resource annotation metadata schema for personalized hypermedia learning resource delivery: Interactive Learning Environments, 23(1): 79-105 https://doi.org/10.1080/10494820.2012.745429

[7] Shih, YC. (2016). Investigation of multiple human factors in personalized learning: Interactive Learning Environments, 24(1): 119-141 https://doi.org/10.1080/10494820.2013.825809

[8] Cavus, N., Zabadi, T. (2014). A Comparison of Open Source Learning Management Systems: Procedia - Social and Behavioral Sciences, 143: 521-526 https://doi.org/10.1016/j.sbspro.2014.07.430

7 Authors

Ying Wang and **Weifeng Jiang** are with the Department of Information Engineering of Tianjin Maritime College, Tianjin 300350 , China (wangying90182@126.com)

Article submitted 19 September 2018. Resubmitted 23 October 2018. Final acceptance 13 November 2018. Final version published as submitted by the authors.