Research on Student Model Construction of Educational Big Data System Based on CD-CAT*

YANG ZHAO [1,2]
[1]*JiLIin University*
[2]*Institute of Computing Technology of the Chinese Academy of Sciences*
Changchun,CHINA
18844189223@163.com

YAQING FAN
JILIN UNIVERSITI
Changchun,CHINA
fanyaqin_joy@163.com

PENG WANG [1,2]
[1]*CEIEC*
[2]*Institute of Computing Technology of the Chinese Academy of Sciences*
Beijing,CHINA
wangpeng@ict.ac.cn

***Abstract*—The era of "big data" technology has penetrated into all areas of society. In the field of education, personalized teaching has gradually become a new development direction. With the widespread promotion of online education, the adaptive learning system has attracted attention because of its good course recommendation function. As an important part of the general model, the student model reflects the individual characteristics, knowledge status and cognitive ability of the students. The traditional student model judges students based solely on the student's basic information and simple test scores.This paper proposes a CD-CAT hybrid model based on the actual scene of the learner, retains the original static question bank, and dynamically updates the student model based on the test database of the learner's different knowledge structure.The knowledge field is divided into multiple categories, and the dynamic allocation of the question bank between categories is carried out. After testing, this dynamic model has a good effect.**

Keywords—Cognitive Diagnosis Theory, CAT, Adaptive Learning System, Construction of Student Model

I. INTRODUCTION

In recent years, with the popularization of Internet technology, major changes have taken place in the education field. Online learning has broken through the limitations of time and space, making knowledge acquisition more flexible. More users are no longer relying solely on offline physical classroom learning. Online education has solved the problem of sharing educational resources and greatly optimized the allocation of educational resources.

Adaptive learning system fully takes advantage of machine learning and data mining, and has a good curriculum of self-adaptive recommendation, providing adaptive learning content and learning paths for different learners. Therefore, it has received extensive attention from the education community and the computer industry. Peter Brusilovsky first proposed the concept of adaptive learning [1] in 1996 and proposed a general-purpose model for an adaptive learning system. Fig.1 shows the general model. of an adaptive learning system. Overseas has also developed InterBook[2], ELM-ART[3], AHA！[4]and other adaptive learning systems.

In the universal model, the student model is used to represent the learner's current knowledge status, cognitive level, and hobbies. Through summarizing and analyzing the existing student models, it provides an important basis for the recommendation of the engine based on the adaptive learning system recommendation. The accuracy of the statistics and analysis of the data directly affects the quality of the system recommendation service. However, in the previous studies, most of the student model construction process mainly considers the learner's basic information. The study of the learner's cognitive level often uses the static test database or the traditional Computer Adaptive Testing (CAT), and cannot use the test database dynamically.

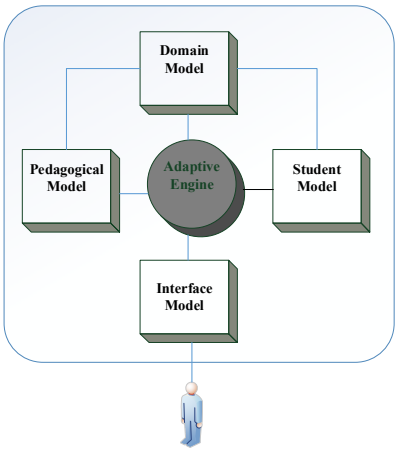

Fig.1. The general model of an adaptive learning system.

However, candidates with the same or similar scores (capabilities) may also have different cognitive states and different knowledge processing processes. Cognitive Diagnosis (CD) was therefore applied.

The difference between cognitive diagnosis theory and traditional CAT is that it does not provide a single score as the subject's overall assessment, but rather pays more attention to the subject's individual condition - its knowledge structure. The current research on cognitive diagnosis theory focuses more on the learning and improvement of algorithms and the basic research of theories, and it is less applied to the specific process of educational learning.

Therefore, this paper proposes a CAT hybrid model based on cognitive diagnosis theory（H-CD）. According to the actual situation of the learner, the fixed question bank and the question bank exist simultaneously, and has a dynamic selection strategy, which is based on the learner's different knowledge structure. The test database dynamically updates the student model, providing an effective guarantee for the adaptive system dynamic recommendation.

II. CD-CAT

A. Commonly used cognitive diagnostic models

By 2007, there are more than 60 kinds of cognitive diagnostic models[5]. More representative of the cognitive diagnosis model are: Linear Logistic Test Model (LLTM)[6], Rule Space Model (RSM)[7], Deterministic Input, Noisy-And gate Model (DINA)[8], Noisy Input, Deterministic-And gate Model (NIDA)[9], Fusion Model (FM)[10], Attribute Hierarchy Model (AHM)[11]。

Among them, the RSM model expresses the respondent's response to the project as an attribute mastering model, which is associated with a cognitive skill. One of the assumptions of the model is that one can use a certain cognitive attribute to describe the test item, or use a set of attribute control patterns to represent the knowledge structure of the testee. This kind of attribute control model is usually not easily observed. And this unobservable cognitive property can be characterized by some observable item response pattern.

The AHM model is an important variant of Tatsuoka's regular space model, proposed by Leighton et al. in 2004. The difference between AHM and RSM lies in the fact that AHM emphasizes the need to first determine the hierarchical relationship between cognitive attributes and to perform other tasks on the premise of correctly identifying attribute-level relationships. Therefore, AHM is more logical. And the hierarchical compatibility of the attributes can be verified by the Hierarchical Compatibility Index (HCI).

The DINA model adds the error and guess parameters of the item itself to the item-attribute parameter. The error parameter refers to the probability that the subject has mastered the attributes of the item and answered the question incorrectly. The guessing parameter refers to the probability that the subject did not fully grasp the attributes of the item but answered the question.

B. CD-CAT

Computer Adaptive Testing (CAT) refers to the use of a computer as a medium, in accordance with the different conditions of learners, from the system to select the most suitable question for the learner to answer questions, test the candidates, and automatically test the learner test results of a test form. General CAT consists of several parts: Item Bank, Starting point, Parameters Estimate and Items Selection Method.

The difference between cognitive diagnosis theory and traditional CAT is that it does not provide a single score as an overall assessment of the subject, but rather pays more attention to the subject's individual condition - its knowledge structure. The scope of assessment of cognitive diagnosis generally includes the subject’s individual knowledge structure, processing process or cognitive process, which can all be referred to as attributes. The cognitive results are generally presented in the form of reports. Mislevy R.J believes that the theory of cognitive diagnosis is a new generation of test theory following the standard test theory, which is a combination of cognitive psychology and modern measurement. It is the future direction of adaptive testing. Fig. 2 shows the unified model of CD-CAT.

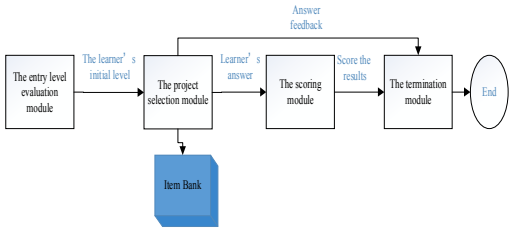

Fig.2. A unified model of CD-CAT

Entry level evaluation module: In general, CD-CAT has two methods for selecting topics, one is not considering the attributes of the questions, and all subjects' initial questions are randomly selected from the question bank. The other is to consider item attribute vector factors, for example, to randomly generate a subject with some initial knowledge state, and then use the title strategy to select the first item [12] based on this randomly generated knowledge state.

Project selection module: The project selection module is the core component of CAT and reflects the personalization and intelligence of the test. Among the CD-CAT methods, Kullback-Leibler (KL) and Shannon Entropy (SHE) have relatively good experimental results and high test accuracy. Therefore, this article mainly implements these two selection strategies for researchers to use [13].

Scoring module: Scoring module is an important part of CD-CAT system. Section 2.1 of this article is a common method introduction.

Terminate evaluation module: CD-CAT termination rules also have fixed length and variable length two. In the fixed length of CAT, each subject received the same number of items to test, the system is easy to achieve, test length can be flexible control, from the concept point of view is also more traditional, easy for people to accept.

III. CONSTRUCTION OF CAT HYBRID MODEL BASED ON COGNITIVE DIAGNOSIS

A. CAT hybrid item bank design based on cognitive diagnosis

The item bank serves as the basis for the selection strategy and plays an important role in the adaptive test results.

Hybrid item bank will include fixed question bank and sub-question bank. Among them, the fixed question bank is based on basic information, and the classification question bank is distinguished by different categories.

In addition to the content of traditional options and the correct and incorrect scores, each question will determine the learner's ability in different cognitive states and different knowledge processing. Each item responds to at least one or all of its abilities. The entire class of subjects must achieve full capacity coverage and must not be omitted.

The test is generally conducted by selecting the title of the fixed question bank and the order of the questions after the ranking question bank.

B. CAT hybrid selection strategy (H-CD) model based on cognitive diagnosis

Whether it is a common simple model, such as RSM model, AHM model or a more complex model, such as a unified model, there are more or less limitations. The simple model has fewer parameters and the description of psychological phenomena is not enough, so it is difficult to expand the design. Complex models are faced with the problem that the parameters cannot be estimated and the model is difficult to implement and apply.

In order to solve the above problems, this paper adopts an adaptive hybrid model with relatively complete parameters. The adaptive hybrid model not only has a complete parameter system, but also has parameters that can be estimated and the model is simple to implement. Through the probability formula, the model establishes the relationship between items and attributes. Through the probability formula, we can get the project's expected response pattern, and then carry out cognitive diagnosis.The comprehensive model is shown in Equation (1), Equation (2), Equation (3).

$$P(\mathrm{x}_{\mathrm{ij}}=1\mid\alpha_j)=(1-s_i)^{\alpha_{\mathrm{jk}}\mathrm{q}_{\mathrm{ik}}}\pi_i^*\prod_{k=1}^{k}r_{ik}^{*(1-\alpha_{jk})\times q_{ik}}P_{\mathrm{c}_i}(\theta_j) \quad (1)$$

$P(\mathrm{x}_{\mathrm{ij}}=1\mid\alpha_j)$ —Indicates the probability of correcting item i in the case that the subject's attribute mastering mode is α_j;

α_j —Testee j's attribute;

s_i —Test subject j's probability of error on item i ;

$$\pi_i^*=\prod_{k=1}^{K}P(Y_{ijk}=1\mid\alpha_{jk}=1)^{q_{ik}} \quad (2)$$

π_i^* is the project parameter. It refers to the probability that the subject who has mastered all attributes correctly responds to item I and can also be seen as the difficulty parameter based on the Q matrix;

$$\mathrm{r}_{ik}^*=\frac{P(Y_{ijk}=1\mid\alpha_{jk}=0)}{P(Y_{ijk}=1\mid\alpha_{jk}=1)} \quad (3)$$

It means that the subject correctly reflects the probability of item i without mastering all attributes.

c_i —It is an index to investigate the completeness of the Q matrix.

$$P_{\mathrm{c}_i}(\theta)=\{1+\exp[-1.7(\theta_j+c_j)]\}^{-1} \quad (4)$$

It indicates the probability that the subject correctly used attributes not included in the Q matrix。

From this, it can be seen that the comprehensive model uses a series of parameters to characterize the completeness of the Q matrix, the probability of students' mistakes in answering, and the residual ability, which is a very complete model. The model completes the estimation of the student's knowledge state, describes the relationship between the project and each attribute, and its parameters can be identified.

IV. STUDENT MODEL CONSTRUCTION BASED ON CAT HYBRID MODEL

A. Student Model Introduction

The student model is a data structure used to represent the learner's current knowledge state in the adaptive learning system, reflecting the student's individual characteristics, knowledge state, and cognitive ability. The accuracy of the information in the student model directly affects the quality of the system recommendation service. In order to improve the portability of student data, data sharing between different systems is facilitated, while ensuring the privacy, security and integrity of student data.

Fig.3 shows the workflow diagram of the student model. The test section is the main discussion part of this article and plays a very important role in the overall curriculum vitae.

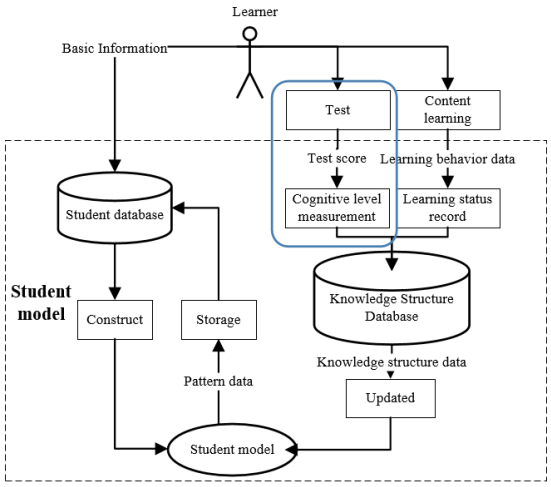

Fig.3. The workflow diagram of the student model.

The basic student model construction process is as follows:

Step1: After the first registration, the basic information is collected. On the one hand, it is used to build a static model for students.

Step2: Students can select content learning and test after login to the system. Students' behavior data is collected during the learning process to record the current learning progress and update the knowledge status data.

Step3: When students finish the current content, they need to answer the corresponding test questions.

Step4: If students have already learned something, they can go directly to the test session and avoid wasting time.

Step5: After the students complete the learning or testing session, they dynamically update the model based on the learning behavior and cognitive level data.

Step6: Finally, the updated model data is stored in the student database to facilitate subsequent system calls.

B. The student model based on H-CD

The adaptive test section can be divided into two scenarios: Fig. 4 shows the data items and storage structure of the student model.

1) Learners register for the first time: Open the fixed question bank immediately when registering for the first time. The fixed question bank will contain basic learner information such as age and gender. When the fixed question is answered, start the question bank. Take the age grading as an example. For example, if the learner selects the option “0-5 years old”, the title is selected from the corresponding grading question bank. If the answerer’s 80% title error or 80% is correct, automatic downgrade or upgrade continues to select. Initial test scores will be saved at all times.

2) Learner's learning course test: According to the initial test scores, the entity in the question bank is cold-started, and the question bank and learner status information are dynamically updated according to the established CAT hybrid choice questioning strategy model based on cognitive detection.

In addition to affecting the function of the adaptive engine recommendation part of the adaptive learning system, the test results will also generate radar maps based on learner competence.

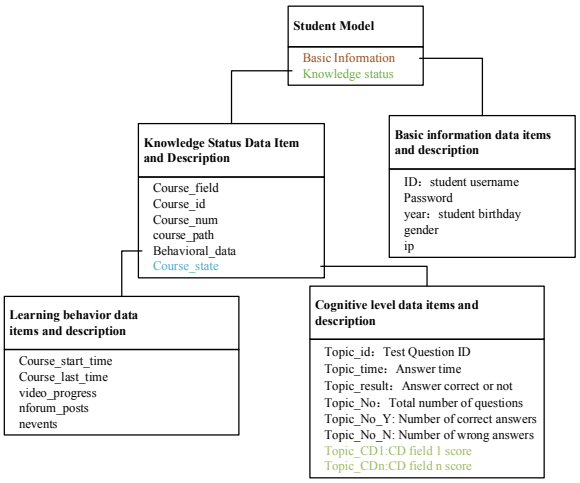

Fig.4. The data items and storage structure of the student model

V. EXPERIMENTAL RESULTS

In order to verify the hybrid (H-CD) model proposed in this paper, we conducted an experimental analysis using the developed CD-CAT system.

Based on science experiment classes for children aged 5-12 years and elementary schools, the curriculum consists of 256 sessions. According to 8 different age groups, each age group is divided into 16 sections for introduction class A and advanced class B. Each class corresponds to 20 question bank test questions, a total of 5120 questions. The topics are all single-choice questions. According to the analysis of the curriculum, according to the requirements of cognitive diagnostic theory, each question corresponds to one of the six cognitive abilities of "remember," "understand," "analysis," "apply," "evaluate," and "create." The 20 questions corresponding to each lesson cover all six cognitive abilities. All the tested items are strictly inspected and designed by the cooperative school teachers. Therefore, the quality of the project, the distribution of each ability level, and the determination of attributes and hierarchical structure all meet the needs of the experiment.

TABLE I. COMPARISON OF DIFFERENT MODELS

Model	CA	MA	TS	TO
RSM	0.8265	0.6889	301.77	0.0852
AHM	0.8916	0.7143	299.66	0.0683
LLTM	0.8750	0.7099	302.78	0.4448
DINA	0.9125	0.7228	300.69	0.0426
H-CD	0.9366	0.9011	300.07	0.0342

In order to reflect the advantages and disadvantages of the integrated model, we have implemented models such as RSM, AHM, LLTM, DINA, and hybrid model H-CD, respectively, through the system. Then through 320 students, 16 simulation experiments were performed according to different cognitive diagnostic models. The four indicators of classification accuracy (CA), marginal accuracy (MA), test statistic (TS), and test overlapping (TO) were analyzed. . Data analysis is shown in Table 1 and Fig. 5, Fig. 6, Fig. 7, and Fig. 8.

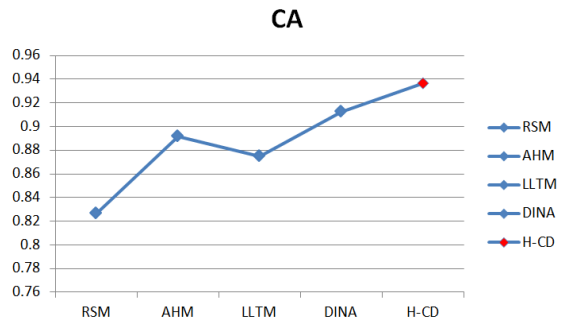

Fig.5 CA of different models

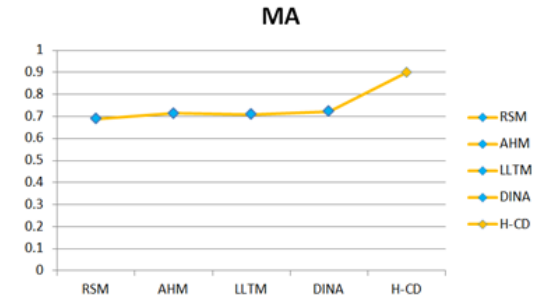

Fig.6. MA of different models

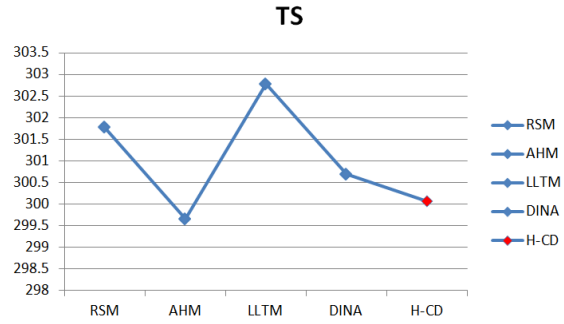

Fig.7. TS of different models

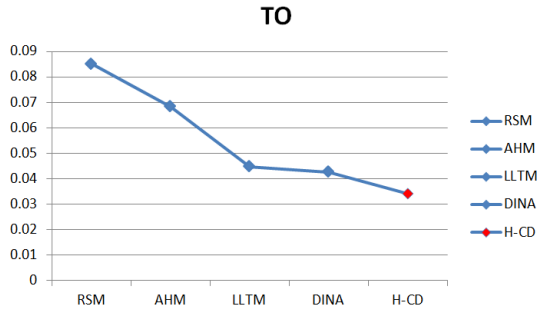

Fig.8. TO of different models

Compared with other models, H-CD has good performance in terms of CA, MA, TS, TO: In CA, H-CD, DINA model and AHM model are in similar levels; in TO, IM model and NIDA model The overlap rates are similar, but all are at a low level. The verification shows that H-CD is a complete and reliable model for accurate cognitive diagnosis.

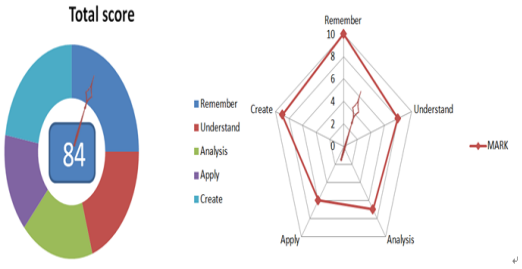

Figure 9 ： A partial test report based on CD

At the same time, each learner will get a test report after the test is over. Fig.9 shows a partial test report based on cognitive diagnostics for different capabilities.

VI. CONCLUSIONS

The computerized adaptive system based on cognitive diagnosis is to analyze the knowledge of students' examinations, so as to better remedy the students' insufficiency and provide scientific guidance for teaching students in accordance with their aptitude. . Combining cognitive science, psychometrics, teaching research, artificial intelligence and other theories, CAT system based on cognitive diagnosis has become a new research direction. In this paper, the specific implementation of cognitive computerized adaptive system and the establishment of unified model are studied. A hybrid question bank construction and hybrid topic selection strategy are proposed. After experiment, H-CD has good comprehensive characteristics. This paper applies the CD-CAT model to the learning model of the adaptive learning system to make the student model more dynamic and complete.

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*

[2] Brusilovsky P. Methods and techniques of adaptive hypermedia[M]//Adaptive hypertext and hypermedia. Springer Netherlands, 1998: 1-43.

[3] Brusilovsky P, Eklund J, Schwarz E. Web-based education for all: a tool for developmentadaptive courseware[J]. Computer Networks and ISDN Systems, 1998, 30(1-7): 291-300.

[4] Brusilovsky P, Schwarz E, Weber G. ELM-ART: An intelligent tutoring system on WorldWide Web[C]//International Conference on Intelligent Tutoring Systems. Springer BerlinHeidelberg, 1996: 261-269.

[5] DeBra. AHA![EB/OL].http://aha.win.tue.nl/,2007.

[6] Fu, J.,& Li, Y. (2007). Cognitive Diagnostic Psychometric Models: An Integrative Review". Paper presented at the the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

[7] Fischer G H. The linear logistic test model as an instrument in educational research ☆ [J]. Acta Psychologica, 1973, 37(6):359-374.

[8] Tatsuoka K K. Rule Space: An Approach for Dealing with Misconceptions Based on Item Response Theory[J]. Journal of Educational Measurement, 1983, 20(4):345-354.

[9] Torre J D L. DINA Model and Parameter Estimation: A Didactic[J]. Journal of Educational and Behavioral Statistics, 2009, 34(1):115-130.

[10] Maris E. Estimating multiple classification latent class models[J]. Psychometrika, 1999, 64(2):187-212.

[11] Hartz M C. A Bayesian Framework for the Unified Model for Assessing Cognitive Abilities: Blending Theory With Practicality[J]. American Journal of Gastroenterology, 2002, 95(4):906–909.

[12] Leighton J P, Gierl M J, Hunka S M. The Attribute Hierarchy Method for Cognitive Assessment: A Variation on Tatsuoka's Rule-Space Approach[J]. Journal of Educational Measurement, 2004, 41(3):205–237.

[13] Lomber S G,Payne B R,CornwellP,Long K D. Perceptual and cognitive visual functions of parietal and temporal cortices in the cat.[J]. Cerebral cortex (New York, N.Y. : 1991),1996,65:.

[14] GongjunXu,ChunWang,Zhuoran Shang. On initial item selection in cognitive diagnostic computerized adaptive testing[J]. Br J Math Stat Psychol,2016,693:.