Research on the Construction of a Personalized Learning Model Based on Data Mining Technology

Wenjuan Shao, Xiaoxiao Gu, Kun Liu*

College of Applied Science and Technology, Beijing Union University, Beijing, China

**Corresponding Author*

Abstract: Aiming at the limitation that the existing personalized learning system relies on a single model and static path planning, an integrated learning model based on data mining technology is proposed. The model constructs dynamic learner files and subject knowledge maps, designs a hybrid recommendation engine integrating content-based filtering, collaborative filtering and association rules, and innovatively combines knowledge maps with sequential pattern mining to generate personalized learning paths. The experimental results on OULAD data set show that the model is superior to the benchmark model in recommendation accuracy (0.45) and NDCG (0.59). In addition, it achieves a significantly higher completion rate of learning path (85%) and score improvement rate (23%), which verifies its effectiveness. This study provides an innovative solution to overcome the limitations of the current personalized learning system and has important reference value for promoting the intelligent development of personalized education services.

Keywords: Personalized Learning; Data Mining; Recommender Systems; Knowledge Graph

1. Introduction

With the wide adoption of online education, the overload of learning resources and the monotony of traditional teaching mode make "how to provide personalized learning support for students" a core challenge in the field of educational technology. Data mining technology can find learning patterns and individual differences from a large number of educational data sets, which provides an important way to build an intelligent adaptive learning model.

Currently, scholars both domestically and internationally have conducted extensive research on personalized learning recommendations. Zhang(2023) constructed a weight model to mine learning needs by analyzing users' learning behavior and dynamic characteristics, and based on vector similarity matching for resource recommendation. His system effectively improved student performance in practice[1]. Wang(2022) systematically designed a personalized recommendation system framework that covers learner classification, resource classification, and hybrid recommendation algorithms from the perspective of educational big data, emphasizing the importance of considering learners from multiple dimensions[2]. Wang(2023) focuses on using collaborative filtering algorithms to analyze student behavior and proposes an improved null filling method to address data sparsity issues, in order to enhance the accuracy of recommendations[3]. However, existing research still has significant limitations: most models rely on a single recommendation technique, lack sufficient characterization of learners' dynamic changes, and lack an end-to-end solution that integrates multiple data mining techniques, resulting in limited overall effectiveness of personalized services.

In view of the above shortcomings, this paper aims to build a complete personalized learning model. The core innovations of this study are as follows: firstly, a multi-level model framework integrating learner analysis, knowledge state diagnosis and path planning is proposed, and the closed-loop integration of services is realized; Secondly, the semantic association of knowledge map is innovatively combined with the temporal pattern of sequential pattern mining in technology to generate a more scientific learning path; Thirdly, a hybrid recommendation strategy is designed, which uses content, collaborative filtering and association rules to improve the recommendation accuracy and alleviate the cold start problem. Through the above design, this study aims to provide a more comprehensive and effective solution to the challenges faced by the existing personalized learning system.

2. Overview of Relevant Technologies

2.1 Cluster Analysis

Cluster analysis is an unsupervised learning technique used to partition samples in a dataset into intrinsic, similarity based groups[4]. In the proposed model, the K-Means algorithm is mainly used to automatically cluster learners based on their behavioral data (such as login frequency, homework completion rate) and cognitive data (such as test scores). This process identifies different types of learners (such as high achievers, diligent learners, struggling learners), providing a foundation for implementing differentiated teaching strategies at the group level.

2.2 Collaborative Filtering Recommendation

Collaborative filtering is a core algorithm for recommender systems[5], operating on the fundamental premise that "similar users tend to prefer similar items." It analyzes the historical user-item interaction matrix (e.g., ratings, clicks) to discover user interest similarities (user-based) or item content similarities (item-based), subsequently predicting and recommending new items that a target user might find interesting. This model incorporates collaborative filtering as a key component of its hybrid recommendation approach to uncover potential learning resource preferences.

2.3 Association Rule and Sequential Pattern Mining

Association rule mining aims to discover interesting relationships between itemsets in a dataset. For instance, the Apriori algorithm can mine rules such as "learners who studied knowledge point A often also study knowledge point B"[6]. Sequential pattern mining extends this by focusing on the temporal order of data, where algorithms like PrefixSpan can identify frequent learning sequences, such as "completing a conceptual introduction before attempting basic exercises." Together, these two techniques provide data-driven support for the model's associated resource recommendations and the generation of scientifically-sequenced learning paths.

2.4 Knowledge Graph

A knowledge graph is a graph-structured method of knowledge representation that describes real-world concepts and their interconnections in a structured format using (entity, relation, entity) triples[7]. In this model, we construct a subject-specific knowledge graph, representing knowledge points as nodes and the logical relationships between them (e.g., prerequisite, successor, belongs-to) as edges. This graph provides a structured knowledge framework for the entire model, serving as the semantic foundation for accurate knowledge state diagnosis and logically-sound learning path planning.

3. Framework

The personalized learning model proposed in this paper aims to provide end-to-end personalized services for learners, spanning from diagnosis to planning, through a systematic data mining pipeline.

3.1 Overall Framework Design

The model adopts hierarchical design concept and is divided into four logical layers from bottom to top: data layer, feature and configuration file layer, core algorithm layer and application service layer, as shown in Figure 1. This architecture ensures a smooth and systematic transformation from raw data to personalized services.

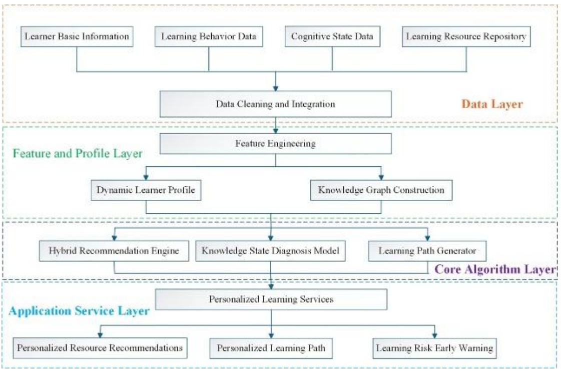
Learner Basic Information
Learning Behavior Data
Cognitive State Data
Learning Resource Repository
Data Cleaning and Integration
Data Layer
Feature Engineering
Feature and Profile Layer
Dynamic Learner Profile
Knowledge Graph Construction
Hybrid Recommendation Engine
Knowledge State Diagnosis Model
Learning Path Generator
Core Algorithm Layer
Personalized Learning Services
Application Service Layer
Personalized Resource Recommendations
Personalized Learning Path
Learning Risk Early Warning

Figure 1. Overall Framework of the Model

(1) Data Layer： Serving as the foundation, the Data Layer is responsible for collecting and preprocessing multi-source, heterogeneous educational data. This includes learners' basic information, behavioral logs (e.g., video watching, exercise attempts), and structured learning resource data. It provides high-quality data input for the upper layers.

(2) Feature & Profile Layer： This layer acts as the connecting link. On one hand, it constructs effective features from the preprocessed data and

utilizes techniques like clustering to form multi-dimensional, dynamic learner profiles. On the other hand, it constructs a subject knowledge graph, providing a structured semantic representation of the knowledge system.

(3) Core Algorithm Layer：This layer functions as the model's brain, integrating three core modules: a knowledge state diagnosis model based on machine learning methods, a hybrid recommendation engine incorporating multiple strategies, and a learning path generator that combines knowledge graphs with sequential patterns. These modules collaborate to accomplish intelligent decision-making.

(4|) Application Service Layer：As the output interface of the model, this layer translates the decision results from the Core Algorithm Layer into concrete, perceivable personalized services, which are ultimately delivered to learners and instructors.

3.2 Core Modules

3.2.1 Learner profile and knowledge graph module

This module aims to describe learners' attributes and knowledge structure comprehensively. Learner profile is not a static tag set, but a dynamic and newer combination[8]. Its dimensions include: cognitive level, quantified by historical performance and knowledge state diagnosis results; Learning style, determined by questionnaire survey and behavioral data analysis (for example, preference for video or text); Interest preference, based on implicit feedback calculation, such as resource clicks, bookmarks and stay time; Learning state includes real-time monitoring of participation, effort and emotional state (for example, inferring from interaction frequency).

Concurrently, the Knowledge Graph module clearly defines the logical relationships between knowledge points—such as prerequisite, successor, part-of, and whole—using (entity, relation, entity) triples. This provides an indispensable semantic map and set of constraints for learning path planning.

3.2.2 Knowledge state diagnosis and hybrid recommendation engine

The Knowledge State Diagnosis module maps a learner's exercise records onto the knowledge graph and utilizes models like Random Forest or Bayesian Knowledge Tracing to dynamically infer the learner's mastery probability for each knowledge point, forming a precise cognitive ability profile[9].

The Hybrid Recommendation Engine synthesizes three recommendation strategies to overcome the limitations of any single algorithm: Content-based Filtering recommends resources similar in content to those the learner has previously liked; Collaborative Filtering identifies "neighbors" with interests similar to the target learner and recommends resources those neighbors have engaged with but the target has not; Association Rule-based Filtering employs algorithms like Apriori[10] to discover strong associations between knowledge points, facilitating complementary recommendations.

Finally, the recommendation results from these three strategies are weighted and fused to generate a final list of recommended resources that is both accurate and diverse.

3.2.3 Learning path generation and service output

Learning path generation is a key innovation of the model. It is not a simple content sequence but is generated by using the knowledge graph as a static skeleton to ensure the path adheres to subject logic, while simultaneously using frequent patterns mined through sequential pattern mining as a dynamic reference to incorporate the experiences of successful learners[11]. On this basis, graph search algorithms generate an optimal learning sequence from the learner's current knowledge state to the target state.

All the results of intelligent computing are gathered in the application service layer, which is manifested as a specific service: personalized learning path planning, which draws a clear learning roadmap customized according to learners' cognitive patterns; Personalized resource recommendation, pushing the most suitable resource at each node of the path-calculated by the hybrid engine; And learning risk early warning, which actively identifies learners with learning difficulties or risk of dropping out of school according to knowledge status and behavior data, and then triggers an intervention mechanism.

4. Experiment and Analysis

4.1 Experimental Environment and Dataset

Experimental environment: The experiment was conducted on a hardware platform equipped with Intel Core i7-12700H processor and 32GB RAM.

The software environment is Windows 11 operating system, Python 3.9 is used as programming language, and mainstream machine learning and data processing libraries, such as Scikit-learn, Pandas and Numpy, are used to build the model.

Dataset: The experiments employed the publicly available and widely-cited educational dataset "OULAD" (Open University Learning Analytics Dataset). This dataset originates from a real online course at the Open University in the UK, containing virtual learning behavior records for over 30,000 students, with a data time span of one year. OULAD is recognized as a benchmark dataset in the educational data mining field due to its large scale, rich dimensionality, and authenticity. We extracted data highly relevant to this study, specifically including:

(1) Student Information: Students' final scores and associated course modules.

(2) Learning Behavior Data: Clickstream data (number of accesses and access duration for different learning resources), assignment submission records, and scores.

(3) Course Structure Data: Course chapters, learning activities, and their explicit prerequisite and successor relationships, which provided a perfect structural foundation for building the knowledge graph.

4.2 Experimental Design and Evaluation Metrics

To validate the effectiveness of the proposed integrated model, we designed the following comparative experiments:

(1) Baseline Model 1 (CF-Only): Utilized only a collaborative filtering-based recommendation algorithm, representing a common but functionally singular personalized learning approach.

(2) Baseline Model 2 (KGraph-Static): Employed the knowledge graph for static learning path planning, where paths were generated solely based on the logical relationships between knowledge points, without considering historical successful sequential patterns.

(3) Proposed Model (IM-PL): Our complete proposed model, incorporating dynamic learner profiling, hybrid recommendation, and a dynamic path generator that integrates the knowledge graph with sequential pattern mining.

For a comprehensive evaluation, we adopted the following three categories of metrics:

(1) Recommendation Effectiveness: Used Precision and Normalized Discounted Cumulative Gain (NDCG) to measure the accuracy and ranking quality of the recommended resource lists.

(2) Path Effectiveness: Used the Path Completion Rate and the Average Score Improvement Rate as key business metrics to directly measure the practicality and effectiveness of the learning paths.

(3) Prediction Effectiveness: Used the Mean Absolute Error (MAE) to evaluate the accuracy of the knowledge state diagnosis model in predicting student scores.

4.3 Experimental Results and Analysis

4.3.1 Comparative analysis of recommendation effectiveness

The performance comparison of the three models on the resource recommendation task is shown in Table 1 below:

Table 1. Comparison of Resource Recommendation Effectiveness across Models

Model	Precision	NDCG@10
CF-Only	0.31	0.42
KGraph-Static	0.35	0.48
IM-PL (Ours)	0.45	0.59

The results indicate that the model proposed in this article is significantly better than the two benchmark models in both indicators. This fully demonstrates the effectiveness of the hybrid recommendation strategy, which integrates content, collaboration, and association rules to provide richer and more accurate recommendations than single collaborative filtering (CF Only). At the same time, its performance is also better than static recommendation (KGraph Static) that relies solely on knowledge structure, indicating that dynamic learner portraits provide more personalized basis for recommendation engines.

4.3.2 Analysis of path effectiveness and final scores

To examine whether the generated learning paths genuinely enhance learning outcomes, we recorded the performance of learners guided by different models, as shown in Table 2:

Table 2. Learner Performance Records

Model	Path Completion Rate	Average Score Improvement Rate
KGraph-Static	68%	12%
IM-PL (Ours)	85%	23%

It can be seen that this model has achieved overwhelming advantages in both path completion rate and performance improvement rate. This result strongly validates the feasibility and superiority of combining knowledge graphs with sequential pattern mining, which is the core innovation of this study. Although KGraph Static is logical, it may be tedious or inefficient; And the learning path generated by our model is not only logically smooth, but also incorporates the experience patterns of successful learners in history, which can better stimulate and maintain learners' learning motivation, ultimately leading to better learning outcomes.

4.3.3 Analysis of knowledge state diagnosis performance

In the task of predicting students' final grades, the MAE value of the diagnosis module integrated in this model is 8.5, which shows good prediction accuracy. This means that the model can identify students with academic risks earlier and relatively accurately, and provide reliable data support for subsequent early warning and intervention.

4.4 Discussion of Results

The experimental results show that the integrated personalized learning model proposed in this paper shows significant advantages in multiple evaluation dimensions. First of all, in resource recommendation, the hybrid strategy effectively improves the recommendation accuracy and user satisfaction. Secondly, and most importantly, in the path planning that directly affects the learning results, the dynamic path generation method integrating knowledge graph logic and sequential pattern empiricism has been proved to be feasible and efficient, and successfully solved the problem of poor adaptability of static paths. In a word, the experiment fully verified the innovation of this study: an integrated, end-to-end model framework and key technology integration strategy can systematically improve the quality and effect of personalized learning services.

5. Conclusion

This study systematically constructs a personalized learning model based on data mining technology to meet the personalized needs of online learning. The core contribution of this model is to propose an integration framework, which organically integrates dynamic learner modeling, knowledge state diagnosis, hybrid recommendation engine, personalized learning path generation and other modules. This model provides structural semantic support for knowledge system by introducing knowledge graph, and captures the temporal pattern of historical successful paths by combining sequential pattern mining, thus realizing the unity of logical rigor and empirical validity of path generation. The experimental part is based on the real OULAD public data set. The results show that compared with the traditional single recommendation or static path planning method, the model proposed in this paper has achieved significant advantages in key indicators such as recommendation accuracy, path completion rate and student performance improvement rate, which effectively verifies the rationality of the model framework and the feasibility of core innovations.

Although this study has achieved some positive results, there are still several limitations that need to be improved. Firstly, the initial performance of the model depends to some extent on the size and quality of the data. When facing new users with sparse data or insufficient annotation information, the speed and accuracy of personalized service initiation may be affected. Secondly, the current model mainly relies on learners' cognitive and behavioral data for decision-making, and the consideration of non cognitive factors such as learning emotions and motivations has not been fully incorporated into the model, which to some extent limits its ability to perceive the overall picture of learners. In addition, the dynamic adjustment mechanism of the model is currently mainly based on phased learning data, and there is still potential for optimizing its response mechanism in achieving real-time and high-frequency adaptive feedback.

Looking forward to the future, this study can be deepened and expanded from the following directions on the existing basis. Technically, we can explore the introduction of reinforcement learning and other mechanisms to enhance the online decision-making and real-time adjustment ability of the model in complex dynamic environment. In the dimension, we can try to integrate multimodal data (such as voice, video, etc.).) to capture the emotion and cognitive state in the learning process, so as to build a more comprehensive learner model with humanistic care. In application, the framework and strategy of this model can be extended to a wider range of lifelong learning scenarios, such as vocational skills training and internal training of enterprises, to verify its universality and potential for large-scale application, and to make more contributions to the deepening development of personalized education.

Acknowledgments
This work was supported by the project of Beijing Digital Education Research Project (BDEC2023619047).

References

[1] Zhang, J (2023). Design of a Personalized Recommendation System for Online Learning Platforms Based on Data Mining. Software, 44(12), 44-46.
[2] Wang, Y. J. (2023). Research on Personalized Learning Models Based on Data Mining Technology. Journal of Anhui Open University, (03), 92-96.
[3] Wang, R. X. (2022). Research on the Design of Personalized Recommendation Systems for Online Learning Platforms. Computer Knowledge and Technology, 18(32), 41-43.
[4] Esteban, A., Zafra, A., Ventura, S. (2020). A clustering-based approach to support educational decision-making with multiple criteria. Computers & Education, 152, 103880.
[5] Ma, H. W., Zhang, G. W., Li, P. (2009). A survey of collaborative filtering recommendation algorithms. Mini-Micro Systems, 30(7), 1282–1288.
[6] Liu, R., Zhao, W. (2023). A Survey of Knowledge Tracing on Structured Knowledge Graph. ACM Transactions on Intelligent Systems and Technology, 14(4), 1-27.
[7] Liu, Q., Li, Y., Duan, H., et al. (2016). A survey on knowledge graph construction techniques. Journal of Computer Research and Development, 53(3), 582–600.
[8] Liu, A., Gao, W. (2024). Research on incremental learner profile construction based on reconfigurable knowledge graph. Office Informatization, 29(14), 38-41.
[9] Liang, Q., Wang, B, Zhang, Q. (2024). Research on teaching cognitive diagnosis model based on SOM neural network. Modern Educational Technology, 34(9), 59–70.
[10]Wang, C. (2023). An improved constrained multidimensional Apriori algorithm. Application of Electronic Technique, 49(10), 100-105.
[11]Xia, S. (2025). Research on multimodal representation and analysis path for human-computer collaborative learning ability. e-Education Research, 46(11), 54-61.