Research Article

Research on Recommendation of Personalized Exercises in English Learning Based on Data Mining

Lina Zhou and Chunxia Wang
Baotou Medical College, Baotou 014040, China

Correspondence should be addressed to Chunxia Wang; 102002066@btmc.edu.cn

Received 16 October 2021; Revised 14 November 2021; Accepted 19 November 2021; Published 21 December 2021

Academic Editor: Hai Dong

Copyright © 2021 Lina Zhou and Chunxia Wang. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Aiming at the problems of traditional method of exercise recommendation precision, recall rate, long recommendation time, and poor recommendation comprehensiveness, this study proposes a personalized exercise recommendation method for English learning based on data mining. Firstly, a personalized recommendation model is designed, based on the model to preprocess the data in the Web access log, and cleaning the noise data to avoid its impact on the accuracy of the recommendation results is focused; secondly, the DINA model to diagnose the degree of mastery of students’ knowledge points is used and the students’ browsing patterns through fuzzy similar relationships are clustered; and finally, according to the clustering results, the similarity between students and the similarity between exercises are measured, and the collaborative filtering recommendation of personalized exercises for English learning is realized. The experimental results show that the exercise recommendation precision and recall rate of this method are higher, the recommendation time is shorter, and the recommendation results are comprehensive.

1. Introduction

Educational data mining is an important research direction of personalized teaching assistance. In recent years, network video teaching, especially the rise of online classroom at home and abroad, has rapidly accumulated a large number of pure electronic educational data, which provide rich materials for educational data mining research [1]. Exercise training is an important part of education, and its personalized recommendation is of great significance. According to the personalized situation of different students, targeted exercise recommendation can effectively improve the teaching quality [2]. However, personalized exercise recommendation still faces great challenges. First of all, how to accurately obtain students’ knowledge mastery and knowledge points that have not been mastered according to students’ history learning behavior, so as to accurately model students, still has a huge room for improvement [3]. Secondly, how to make a reasonable electronic modeling of the knowledge points that students need to master, and make a reasonable personalized exercise recommendation according to the students’ cognitive level model, so that students can more quickly and accurately check deficiencies and fill gaps, which is also one of the key concerns of researchers [4].

In view of the above problems, relevant scholars have conducted in-depth research and achieved some results. Among them, reference [5] proposed a personalized test question recommendation method based on deep self-encoder and secondary collaborative filtering; firstly, considering the students’ cognition of knowledge points, the secondary collaborative filtering test question recommendation based on knowledge points is carried out, and then, the project response theory and depth self-encoder are applied to predict the students’ scores and comprehensive scores related to recommended knowledge points in the recommended test questions. Finally, the prediction results are jointly judged and the difficulty of the final personalized recommended test questions is controlled, generating a final list of recommended questions. Through comparative experiments, this method can realize the personalized recommendation of test questions. Reference [6] proposes a
literature personalized recommendation model based on three dimensions, which identifies users’ points of interest through the collaboration of three dimensions: expert weight dimension, user dimension, and context perception dimension. The recommendation model uses the analytic hierarchy process and entropy weight method to quantify expert opinions and uses the potential Dirichlet distribution and KL divergence to calculate quantitative user similarity. Through the user’s social annotation behavior, search behavior, and browsing behavior, the user’s emotional tendency is obtained, and the time factor is introduced to quantify the user’s emotion. Finally, the “maximum frequency value” is introduced to determine the recommendation index of each dimension, and the literature comprehensive recommendation index is obtained by weighted calculation. Taking the university library as the experimental platform, the method is verified. The experimental results show that the method has good recommendation performance. Reference [7] proposes a personalized exercise recommendation method combining in-depth knowledge tracking model and collaborative filtering method. This method first models students’ knowledge with in-depth knowledge tracking model, then combines collaborative filtering method to calculate students’ correct probability of exercises, and recommends exercises within a certain difficulty range to students according to this probability. This method refers to the personal knowledge level and the nearest neighbor information of students in similar situations at the same time, has better model accuracy, and can recommend suitable content according to the difficulty range. Finally, the effectiveness of this method is verified by experiments.

Although the above traditional methods realize the recommendation function, they still have some room for improvement in precision, recall, recommendation time, and comprehensiveness. Therefore, this study proposes a personalized exercise recommendation method for English learning based on data mining. The main research contents of this method are as follows:

1. A personalized recommendation model is designed to preprocess the data;
2. Cleaning the data to reduce the influence of noise data on the recommendation result is focused;
3. The Dina model is used to diagnose students’ mastery of knowledge points, and the fuzzy similarity relationship is used to cluster students’ browsing patterns;
4. The similarity measurement is carried out, and the collaborative filtering recommendation of personalized exercises in English learning is realized according to the measurement results; and
5. Through comparative experimental verification, the advantages of this method are obtained and its application value is verified.

Our contribution is threefold:

1. Aiming at the problems of traditional method of exercise recommendation precision, recall rate, long recommendation time, and poor recommendation comprehensiveness, this study proposes a personalized exercise recommendation method for English learning based on data mining.

2. We design a personalized recommendation model, based on the model to preprocess the data in the Web access log, and focus on cleaning the noise data to avoid its impact on the accuracy of the recommendation results.

3. The experimental results show that the exercise recommendation precision and recall rate of this method are higher, the recommendation time is shorter, and the recommendation results are comprehensive.

The remainder of this study is organized as follows: Section 2 introduces the personalized recommendation model design; Section 3 discusses the recommendation method of personalized exercises for English learning based on data mining; Section 4 discusses simulation experiment and analysis; and Section 5 presents the conclusions of the study.

2. Personalized Recommendation Model Design

The current network teaching platform system structure generally consists of three parts, namely education resource library, learning platform, and users. The educational resource library is a media server that stores various types of educational resources; users are students; the learning platform is a Web server that displays teaching resources to users through the Web, and users can freely choose learning resources. The personalized network teaching platform improves the learning platform in the original teaching platform and introduces a personalized service module here, so that the network teaching platform can recommend exercise resources for students in a targeted manner according to their personality characteristics. The schematic diagram of the personalized recommendation model is shown in Figure 1.

The function of the personalized recommendation model is to track changes in students’ interest in the network teaching system, use the knowledge obtained from data mining to dynamically recommend teaching resources to students from teaching resources, and provide customized teaching resources for users according to the resource interest characteristics described by users. The model includes 6 main modules, and they are preprocessing module, association rule mining module, personalized customization module, user feature extraction module, personalized computing module, and personalized extraction module. There are also 4 resource libraries in the model [8], namely Web access transaction library, teaching resource library, frequent item set library, and user feature library. The following is a detailed analysis of the various modules of the recommended model.

2.1. Preprocessing Module. The main function of the preprocessing module is to clean and filter the server-side access logs, with the purpose of obtaining transaction data that
meet the requirements of association rule mining. The preprocessing work is mainly divided into four parts: data cleaning, user authentication, session authentication, and sequence identification.

1. Data cleaning: the spam in the server’s original log is filtered out, such as invalid access records embedded in HTML files. These access records are meaningless entries for mining. What is really needed in the recommendation is to contain teaching content HTML access record [9];

2. User authentication: the task of user authentication is to obtain a collection of all paths accessed by the same client;

3. Session authentication: the access sets of the identified users are distinguished, and the different sessions of the proposed user are found out; and

4. Sequence identification: the purpose of sequence identification is to find the user’s meaningful access subsequences. The result of session authentication is a collection of user access sequences, and sequence identification is to find a number of meaningful subsequences in this collection. The multiple relatively independent access subsequences for the user in a session are formed.

2.2. Association Rule Mining Module. The function of the association rule mining module is to find out all the Web access rules based on the Web access transaction database. The work done by this module has nothing to do with the business domain. The method used is the classic Apriori algorithm. The output of this module is Web access frequent item sets, and these frequent item sets are knowledge reflecting the learning trajectory of students [10].

2.3. Personalized Customization Module. The personalized customization module provides students with customized page content based on the user’s selected teaching resource classification, course options, discussion topics, and other interest characteristics. For example, the user will select the content of three courses in the online teaching platform and fill in other field interest options, so that every time a student logs in, the system can provide the student with the teaching content of these three courses and exercise data in other fields. When new teaching resources in these fields are generated, the system will promptly renew.

2.4. User Feature Extraction Module. The purpose of user feature extraction is to extract access interest information from each session of the user. The user interest feature is described through the student submodel, which defines the user’s basic information and interest feature information. After each user logs in, the system starts to record the user’s access tracks and filter out nonteaching contents (such as navigation and reference pages) from these access tracks and stops this recording until the end of the user’s access session. The result of this module is the students’ access interest characteristics. These characteristics describe the sequence sets of teaching resources accessed by the students in the last session. These access sequence sets are written into the user feature library at the end of the students’ current session.

2.5. Personalized Computing Module. The personalized computing module inputs the visit sequence of the user’s most recent session into the recommendation algorithm, finds out the strong rules for Web visits by matching the corresponding frequent item sets, and stores the consequence of this rule in the user feature database as the recommended set. It recommends users the next time they log in. In this way, at the beginning of each new conversation, the students will get the system-tailored recommended content, and the function of recommending relevant content to the students is completed through the personalized recommendation algorithm.
2.6. Personalized Extraction Module. The personalized extraction module extracts personalized content from the user feature library every time a user logs in. These personalized contents are obtained by the personalized computing module using the interest features of the user when they log in last time.

3. Recommendation Method of Personalized Exercises for English Learning Based on Data Mining

3.1. Data Preprocessing. The main source of data in the Web access log is large amounts of data that are recorded in the Web access log of user visits stored directly from these data, because there is a large number of noise data in the Web access log. The existence of noisy data will interfere with the accuracy of data mining. Therefore, before using data mining algorithms to analyze Web access logs, it must be preprocessed to obtain user access sequence information that meets the requirements of data mining.

Web log preprocessing is roughly divided into five steps: data cleaning, user authentication, session authentication, path integration, and sequence identification. Since most customer information is shielded under the proxy server, it is necessary to distinguish different customer records under the same proxy server. Each user may have several visits, and it is necessary to distinguish different sessions of the same user. Customers may go back and forth, and refresh the page while browsing. The user access sequence obtained from the log file needs to be path integrated to get a complete user access path. The session generated by the user during one visit contains not only one session, but also may have several relatively independent visit sequences. Therefore, sequence identification is required. The process of data preprocessing is shown in Figure 2.

In data preprocessing, a large amount of data need to be processed. In this process, more interference data will be generated. Therefore, data cleaning is a very important part of it. Data cleaning refers to the removal of redundant records in the Web access log. Each webpage on the Web server is specified through a separate link. When a user sends an access request for a page, the graphics, scripts, images, and other resources contained in the page will be automatically downloaded and written into the access log, and these contents are noise data for data mining. The data cleaning process is shown in Figure 3.

3.2. Diagnosis of Students’ Knowledge of Mastery. To obtain the student’s learning status (student personality status), on the basis of data preprocessing and data cleaning, the DINA model [11] is used to diagnose the degree of mastery of students’ knowledge points.

Assume each student \( P \) as a knowledge point mastery degree vector \( K = [k_1, k_2, \ldots, k_n] \), where each dimension corresponds to a knowledge point, \( k_n = 1 \) represents that student \( P \) has knowledge point \( n \), and \( k_n = 0 \) represents that student \( P \) has not mastered knowledge point \( n \).

Given student \( P \)'s knowledge point master vector \( K \), for student \( P \) unanswered test question \( A_i \), the potential answer of student \( P \) to test question \( A_i \) can be obtained according to the following formula:

\[
f_{ik} = \prod_{i=1}^{n} \xi_{ik}^{a_i} \times A_i, \tag{1}
\]

where \( \xi_{ik} \) represents the student’s test score matrix; \( f_{ik} = 0 \) represents that student \( P \) cannot answer test question \( A_i \) correctly; and \( f_{ik} = 1 \) represents that student \( P \) can answer test question \( A_i \) correctly.

In addition, the DINA model also introduces test question parameters (error rate) and guessing rate to model students’ answering conditions in the real state. Specifically, student \( P \)'s response to test question \( A_i \) is expressed by the following formula:

\[
P(A_i) = E_k(1 - M_{il} \times M_{lj}), \tag{2}
\]

where \( E_k \) represents all the knowledge points examined in the test question set; \( M_{il} \) represents the student’s record of doing the questions; and \( M_{lj} \) represents the relationship between the test questions and the knowledge points.

Because the relationship between knowledge points is considered to be "connected" in the DINA model, the error rate \( D_z \) is defined as the probability that a student who has mastered all the skills required for the test question will still be unable to answer the test question \( A_i \) correctly; the guess rate \( D_z \) is defined as the probability that a student who has not mastered all the skills required to answer the question \( A_i \) correctly.

The DINA model uses the EM algorithm [12] to maximize the edge likelihood of formula (2), thereby obtaining the parameter estimates of \( D_z \) and \( D_z \). The knowledge point mastery vector \( K \) of student \( P \) can be determined by maximizing the posterior probability of the student’s test score, so as to obtain the student’s dichotomous knowledge point mastery vector. The formula for calculating the posterior probability of student test score is as follows:

\[
\hat{R} = \sum_{i,j=1}^{n} \left( a_i - a_j \right) \left( \varphi(x_i), \varphi(x_j) \right),
\]

\[
= \sum_{i,j=1}^{n} \left( a_i - a_j \right)^2 \tag{3}
\]

\[
= \arg \max K \left(P \mid K, D_z, D_z \right),
\]

where \( a_i \) represents the potential feature vector of students; \( a_j \) represents the potential feature vector of test questions; \( \varphi(x_i) \) represents the potential factors of students; and \( \varphi(x_j) \) represents the potential factors of test questions.

After obtaining the mastery of students’ knowledge points, students’ browsing patterns can be clustered and used for exercise recommendation in combination with the mastery of students’ knowledge points and the examination of knowledge points of the test questions to be recommended.

3.3. Clustering Algorithm of Student Browsing Patterns Based on Fuzzy Similarity Relations. In the online learning mode, the process of student learning is the process of activities in
the distance education website. Each activity of the student is a click operation on a page object on the learning website, and these click operations are completely recorded in the log file middle. Through data mining of the log files left by students visiting the learning website, we can find the hidden patterns, reveal students’ preference for access paths, find the trends and laws of students’ access paths, and help understand students’ learning behavior, so as to improve the structure of the site and provide personalized services for students.

According to the actual situation, in relation to a criterion or a certain method, a number in the interval [0, 1] is assigned to each element in the universe of $G$, which is called the similarity coefficient, and its size indicates that the two elements are each other, degree of closeness or similarity [13].

Let $S_{ij}$ denote the similarity coefficient between elements $u_i$ and $u_j$, where

$$
S_{ij} = \begin{cases} 1, & S_{ij} \geq \mu_i, \\
0, & \text{other.}
\end{cases}
$$

Then, $V_{ij}$ is called the cutoff matrix of $\mu$. Obviously, it is a Boolean matrix. The element in the $V$ matrix is a symmetrical binary variable, which describes the similarity between objects.

Since $V$ defined in this way is an equivalent relationship on set $Y$, then $V$ can uniquely determine a division of set $Y$, and it can also classify set $Y$ according to its $\mu$-section relationship. Different section relationships can get a different classification.

Clustering can divide the pattern set into several classes, but there may not be a clear boundary between classes, which means that there is overlap between classes. A model may belong to multiple classes with different membership

![Figure 2: Data preprocessing flow chart.](image)

![Figure 3: Data cleaning flow chart.](image)
degrees. In addition, the log file on the web server contains the access sequence of specific students accessing the web page, which is not a real value vector, and the length of different sequences is different, so it is transformed into a real value vector of equal length by using the method of fuzzy mathematics. Forms are easy to compare the degree of similarity between patterns, so as to cluster student behavior affairs.

Student behavior affairs are a data collection of multiple student browsing behaviors. Assuming there are M students, there is transaction set \( H \), which contains \( m \) different student affairs, and its representation is as follows:

\[
H = \{h_1, h_2, \ldots, h_m\},
\]

(7)

where \( h_i \) represents the sequence of the \( i \)th student.

Assume \( \delta \) is a complete sequence of pages clicked by different students, and its representation is as follows:

\[
\delta = [\delta_1, \delta_2, \ldots, \delta_n].
\]

(8)

Each \( h_i \in H \) is a non-empty subset of \( \delta \).

To better measure the similarity between any two objects, each student’s behavioral affair is first converted into the form of a real-valued vector with equal length.

For any student behavior transaction \( h \in H \), it can be expressed as a real-valued vector, and its form can be expressed as follows:

\[
h = [f_1, f_2, \ldots, f_r].
\]

(9)

Of which

\[
f_r = \begin{cases} 
1, & f_r \in h, \\
0, & \text{other.}
\end{cases}
\]

(10)

In this way, each student transaction is represented as a vector with equal length, and each element in the vector is 1 or 0, so as to realize student behavior transaction clustering.

### 3.4. Similarity Measure

Based on the clustering results of the student browsing mode, the exercises to be recommended are referred to by matrix \( J \). The rows in the matrix represent the students, and the columns represent the recommended exercises. The matrix scoring value in a certain range of values reflects the differences in the recommended exercises by the students. The degree of preference, if there is only 1 or 0 in the matrix data, means that the user has only two choices: like or dislike. The following matrix is used to describe students’ interest and preference for exercises:

\[
J_p = \begin{bmatrix} 
\hat{j}_{1,1} & \hat{j}_{1,2} & \cdots & \hat{j}_{1,p} \\
\hat{j}_{2,1} & \hat{j}_{2,2} & \cdots & \hat{j}_{2,p} \\
\vdots & \vdots & \ddots & \vdots \\
\hat{j}_{p,1} & \hat{j}_{p,2} & \cdots & \hat{j}_{p,n}
\end{bmatrix}.
\]

(11)

For the exercise data unknown to students, if want to accurately predict students’ interests and preferences, cosine similarity is used to measure the similarity between neighborhood exercises [14].

The exercise data as the word frequency vector and the student’s score as \( N \) vector are set. The angle cosine between the two word frequency vectors is used to describe the degree of similarity between the exercises. The angle cosine between the two exercise vectors can reflect the difference between the exercises, and degree of similarity. If the student does not give a score for the problem data, the score value is 0.

Given that two students are \( P_1 \) and \( P_2 \), and their scores on the exercises are \( \omega_1 \) and \( \omega_2 \), the cosine similarity \( \text{sim}(P_1, P_2) \) between the exercises is calculated as follows:

\[
\text{sim}(P_1, P_2) = \frac{P_1 \cdot P_2}{\|P_1\| \times \|P_2\|}.
\]

where \( \|P_1\|^2 \) and \( \|P_2\|^2 \) represent the scores of students \( P_1 \) and \( P_2 \) on the exercises.

Considering that there are certain differences in students’ scoring standards, the cosine similarity calculation method is optimized, and the average score of the exercises is removed to reduce the degree of scoring difference.

Based on the corresponding problem sets \( P_1 \) and \( P_2 \) scored by students \( I_u \) and \( I_v \), and the intersection \( I_c \) of the exercises scored by the two students, the following formula for calculating the improved cosine similarity \( \text{sim}(I_u, I_c) \) is constructed:

\[
\text{sim}(I_u, I_v) = \sum_{u,v=1}^{n} (R_u - R_v) \times (R_u - R_v),
\]

(13)

where \( R_u \) and \( R_v \) represent the number of exercises in the problem sets \( I_u \) and \( I_v \), respectively; \( R_u \) and \( R_v \) represent the average scores of students \( P_1 \) and \( P_2 \), respectively.

To obtain the similarity more accurately with respect to the problem \( I_c \) that was scored by the two students, the following similarity expression can be used:

\[
\text{sim}(I_c) = \sum_{c=1}^{n} (R_c - R_c)^2,
\]

(14)

where \( R_c \) and \( R_c \) both represent the student’s score for problem \( I_c \).

### 3.5. Collaborative Filtering Recommendation of Personalized Exercises for English Learning

According to the measured student similarity and the simultaneous exercise similarity relationship, combined with the similarity relationship between the fuzzy attribute characteristics of the exercises, the comprehensive similarity between the exercises is solved by weighted fusion. The calculation formula is shown in the following equation:

\[
\text{sim}(I_d) = \omega_1 \text{sim}(I_u, I_c) + \omega_2 \text{sim}(I_c),
\]

(15)

where \( \omega_1 \) represents the cooperative similarity value of exercises under the scoring matrix; \( \omega_2 \) represents the similarity value of exercises based on fuzzy attribute characteristics. \( \omega_1 + \omega_2 = 1 \), according to the sparsity of scoring data, and \( \omega_1 \) and \( \omega_2 \) can be clarified.
Assuming that set \(S_u\) is the neighborhood set of exercise \(I_u\), and \(S_v\) is the neighborhood set of exercise \(I_v\), through the similarity between the exercises and the sets \(S_u\) and \(S_v\), combined with the score results of the neighborhood user set, the estimated score \(\Phi_{u,v}\) of exercises \(I_u\) and \(I_v\) is solved, and the calculation formula is as follows:

\[
\Phi_{u,v} = (\overline{R}_u + \overline{R}_v) \times \frac{\sum_{i=1}^{N} \text{com}_\text{sim}(I_u, I_v)}{\sum_{i=1}^{N} \text{com}_\text{sim}(I_u, I_v)} \theta^2
\]

(16)

where the comprehensive similarity between the two exercises is \(\text{com}_\text{sim}(I_u, I_v)\), based on the average scores \(\overline{R}_u\) and \(\overline{R}_v\) of exercises \(I_u\) and \(I_v\), to complete the collaborative filtering recommendation of personalized exercises for English learning [15].

4. Simulation Experiment

A simulation experiment is designed to verify the effectiveness of the personalized exercise recommendation method for English learning based on data mining. Taking the personalized test question recommendation method based on deep self-encoder and secondary collaborative filtering and the literature personalized recommendation model based on the three-tier dimension as the comparative method, this study makes a comparative analysis with the method in this study and draws the corresponding conclusions.

4.1. Data Set Description. To verify the feasibility and accuracy of the recommendation method of personalized English learning exercises based on data mining, this article uses an exercise data set, which is a university "C language programming" course exercises and the course examination answer records in the past 5 years.

Exercise data set is composed of the following data: (1) the exercise database of C language programming course, which marks knowledge points according to expert knowledge, contains 1653 exercises including multiple-choice questions, blank filling questions, judgment questions, program questions, and programming questions, involving 237 knowledge points, and each question is marked by 1~6 knowledge points; (2) the examination answer records of the course in recent 5 years contain 1069 answer data. In this study, the answer data are normalized and preprocessed, the answer scores of objective questions and subjective questions are mapped to [0, 1], also known as score ratio, and 1069 answer records are randomly divided into 3 copies. 10 experiments are carried out by means of cross-validation, and the average value is taken as the experimental result. The exercise data set is described in Table 1.

The experiment uses Windows 7 64 bit operating system, the CPU is Intel® Core™ i7-4700MQ, the memory size is 16 GB, the hard disk size is 2 TB, and the experiment language is MATLAB R2012a version.

4.2. Analysis of Experimental Results

4.2.1. Precision and Recall%. In recommendation research, precision and recall are two commonly used evaluation indicators, which are different from common e-commerce recommendation methods. Exercise recommendation focuses on digging out students' knowledge of knowledge points. The table is the evaluation index to compare different methods.

\[
\text{precision} = \frac{\theta_a}{\theta_a + \theta_b},
\]

(17)

\[
\text{recall} = \frac{\theta_a}{\theta_a + \theta_c},
\]

where \(\theta_a\) represents the number of correct recommendations; \(\theta_b\) represents the number of actual recommendations; and \(\theta_c\) represents the number of incorrect recommendations.

Exercises to target students are recommended, and the recommended precision and recall rates to obtain comparative experimental results are used. The experimental results are shown in Table 2.

It can be seen from the results in Table 2 that in terms of exercise recommendation precision and recall, this method is superior to the personalized test recommendation method based on deep self-encoder and secondary collaborative filtering and the literature personalized recommendation model based on the three-tier dimension. The highest precision and recall rates reach 98.1% and 90.2%, respectively. This method uses the Dina model to diagnose the mastery degree of students' knowledge points and has an in-depth understanding of students' learning situation. It can not only reflect students' learning situation, but also improve the quality of exercise recommendation. It verifies that this method has high pertinence of knowledge points and the accuracy of exercise recommendation.

4.2.2. Recommended Time/S. Next, a comparative experiment is conducted on the recommended time of individual exercises for English learning under the three methods. 1400 pieces of data information are randomly selected from the exercise data set, and the recommended time is used as the evaluation standard. The comparison result is shown in Figure 4.

It can be seen from Figure 4 that the exercise recommendation time of the method in this study is significantly lower than that of the personalized test question recommendation method based on deep self-encoder and secondary collaborative filtering and the literature personalized recommendation model based on the three-tier dimension. Although the recommendation time of the three methods shows a gradual increasing trend with the increase in data information, however, the increasing trend of this method is significantly lower than that of traditional methods. The maximum recommended time of this method is only 6.5 s, while the maximum recommended time of the two traditional methods is 12.0 s and 15.0 s, respectively. It can be seen that this method speeds up the recommendation speed of exercises.

4.2.3. Recommended Comprehensiveness. To further test the application value of the recommendation method, the comprehensiveness of exercise recommendation as the
evaluation index is taken, and different methods are compared. The results are shown in Figure 5. Among them, the comprehensiveness of recommendation is expressed by numerical value, specifically 0.1–1.0. The larger the value is, the more comprehensive the recommendation result is.

By analyzing the data in Figure 5, it can be seen that when the data information is 100, the recommended comprehensive coefficient of the method in this study is 0.9, and the recommended comprehensive coefficients of the personalized test question recommendation method based on deep self-encoder and secondary collaborative filtering and the literature personalized recommendation model based on the three-tier dimension are 0.72 and 0.75, respectively; when the data information is 2000 pieces, the recommendation overall coefficient of this method is 0.7, the recommendation overall coefficients of personalized test question recommendation method based on deep self-encoder and secondary collaborative filtering and literature personalized recommendation model based on the three-tier dimension are 0.38 and 0.5, respectively. The comparison shows that the recommendation of this method is more comprehensive, which shows that this method can recommend more comprehensive exercise resources for students and help students practice better.

5. Conclusion

To solve the problems of low precision, recall, long recommendation time, and poor comprehensiveness of exercise recommendation in traditional methods, this study proposes a personalized exercise recommendation method for English learning based on data mining. The main innovations of this method are as follows:

1. Design a personalized recommendation model, preprocess the data in the Web access log, clean the

---

Table 1: Details of exercise data set.

<table>
<thead>
<tr>
<th>Data set number</th>
<th>Number of students/person</th>
<th>Number of questions/piece</th>
<th>Number of exercises submitted/time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3528</td>
<td>204</td>
<td>254698</td>
</tr>
<tr>
<td>2</td>
<td>2706</td>
<td>1127</td>
<td>321486</td>
</tr>
<tr>
<td>3</td>
<td>1983</td>
<td>580</td>
<td>367190</td>
</tr>
</tbody>
</table>

Table 2: Comparison of precision and recall.

<table>
<thead>
<tr>
<th>Data set</th>
<th>Method of this article</th>
<th>Personalized test question recommendation method based on deep self-encoder and secondary collaborative filtering</th>
<th>Literature personalized recommendation model based on three dimensions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>Precision</td>
</tr>
<tr>
<td>1</td>
<td>97.5</td>
<td>89.7</td>
<td>90.2</td>
</tr>
<tr>
<td>2</td>
<td>98.1</td>
<td>86.3</td>
<td>89.6</td>
</tr>
<tr>
<td>3</td>
<td>97.6</td>
<td>90.2</td>
<td>87.4</td>
</tr>
</tbody>
</table>

Figure 4: Comparison of recommended time for exercises.

Figure 5: Comprehensiveness of exercise recommendations.
data to avoid the impact of noise data on the recommendation results, and improve the accuracy of the recommendation results;

(2) diagnose the degree of mastery of students’ knowledge points by the Dina model, and cluster the browsing patterns of students by fuzzy similarity relationship; and

(3) Measure the similarity, including the similarity between students and exercises, and finally, realize the collaborative filtering recommendation of personalized exercises in English learning.

The experimental results show that the highest accuracy and recall rates of exercise recommendation in this method are 98.1% and 90.2%, respectively, and the highest recommendation time is only 6.5 s, and the comprehensiveness coefficient of exercise recommendation is high, indicating that the recommendation effect of this method is good. In the future work, we can further optimize the information storage mode, query mode, and resource management, to conduct accurate query in the case of complex data.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study was supported by the Project of the Federation of Social Sciences of Inner Mongolia Autonomous Region: A Study on the Ideological and Political Teaching Model of College English Course based on “online and offline” Hybrid Teaching (No. 20WY15) and 2022 University Scientific Research Project of Inner Mongolia Autonomous Region: Student Behavior Analysis of English online education from the perspective of big data (No. NJSY22064).

References


