International Journal for Housing Science and Its Applications
Publish August 10, 2025. Volume 46, Issue 4 Pages 4215-4228

https://doi.org/10.70517/ijhsa464273

Research on Computational Methods for Optimizing English Vocabulary Learning Paths in Higher Vocational Colleges under Big Data Environment

Tingting Liu[1,*]
[1] Basic Teaching Department, Henan Polytechnic, Zhengzhou, Henan, 450046, China
Corresponding authors: (e-mail: 22036@hnzj.edu.cn).

Abstract With the purpose of meeting the personalized needs of higher vocational students and optimizing the English vocabulary learning path, a personalized recommendation learning system for English vocabulary is designed based on big data. The improved K-means algorithm is used for clustering analysis of learner features, and the English vocabulary hybrid recommendation model is constructed by combining the user-based collaborative filtering recommendation algorithm and the item-based collaborative filtering recommendation algorithm. Then the questionnaire is used to evaluate the English vocabulary personalized recommendation learning system. The sample learners are categorized into six types of different English learning levels, with intermediate level students accounting for the largest proportion of 31.4%. The hybrid recommendation model in this paper has a certain improvement compared with the single collaborative filtering recommendation algorithm, with MAE values of 0.606 and 0.514 for different data, showing better music vocabulary recommendation. The English vocabulary personalized recommendation learning system is affirmed by most learners, more than 80% of the learners are satisfied with the reasonableness of its English vocabulary recommendation, and more than 60% of the learners think that it can promote the interest and motivation of learning English vocabulary. The development of English vocabulary personalized recommendation learning system is of practical significance to further promote the improvement of English vocabulary learning.

Index Terms collaborative filtering, K-means, hybrid recommendation model, learner characteristics, English vocabulary learning

I. Introduction

With the acceleration of globalization, English, as an international common language, has become an important teaching content in the field of education [1]. English vocabulary teaching is an important part of language learning and plays a fundamental role in improving students' language ability [2], [3]. Traditional English vocabulary teaching is usually teacher-driven, focusing on vocabulary memorization and mechanical training [4]. Teaching content mostly relies on textbooks and vocabulary lists, and teachers adopt ways of explaining the meaning of words, pronunciation and using example sentences to help students accumulate vocabulary [5]. This teaching method is relatively single, and students mainly engage in rote memorization, reciting the spelling and meaning of vocabulary words, lacking the understanding and contextual experience of the actual use of vocabulary words [6]-[8]. At the same time, traditional teaching tends to emphasize the transmission and memorization of written knowledge, neglecting the practical application of vocabulary and the communicative function of language [9], [10]. The evaluation of vocabulary learning is mostly based on exams, which increases students' psychological pressure in the learning process and may lead to students' lack of interest in vocabulary learning [11]. The use of transmission-based vocabulary teaching methods for a long time cannot really stimulate students' motivation to learn, and students' mastery of vocabulary is often unsatisfactory [12], [13]. In this context, how to effectively integrate the concept of wisdom education into English vocabulary teaching to improve students' learning efficiency and interest has become an important research topic for current English teachers [14], [15].

In this paper, we propose an English vocabulary recommendation system based on personalized recommendation, and introduce the system architecture and functional flow. For the learner data in the system, the K-means algorithm based on density and weight improvement is used for clustering analysis to divide students into groups with different English learning levels. In view of the shortcomings of the current implicit scoring model, time-awareness is introduced into the implicit scoring algorithm, and then the user-based collaborative filtering recommendation algorithm and item-based collaborative filtering recommendation algorithm are integrated to

design a hybrid recommendation model for English vocabulary. Subsequently, the Alg and Geo datasets are used as samples to analyze the MAE values of different similarity metrics with the increase of the number of nearest neighbors, and to test the effectiveness of the selected similarity calculation methods. The hybrid recommendation algorithm is also compared and analyzed with individual recommendation algorithms to explore the effectiveness of this paper's method for English vocabulary recommendation. Finally, a questionnaire survey is conducted on the students of a higher vocational college to obtain the application effect of the English vocabulary personalized recommendation learning system from three aspects: system satisfaction, learning attitude and learning effect.

II. Optimization of English Vocabulary Learning Paths Based on Big Data

In recent years, education academics have been exploring in the fields of learning analysis, learning strategy exploration, teaching evaluation reform, personalized education and education management methods, etc. With the help of big data thinking, the education mode is gradually transformed into digital personalized adaptive learning. Based on big data technology, this paper explores the optimization path of English vocabulary learning in higher vocational colleges and proposes an English vocabulary learning system based on personalized recommendation, which forms learner characteristics by mining and acquiring basic information and behavioral and interactive data of the users, and realizes the intelligent recommendation of vocabulary's source, difficulty, time, frequency, quantity, and presentation form.

II. A. System architecture

The system architecture of English vocabulary learning platform based on personalized recommendation is shown in Figure 1, which is divided into three layers: user layer, business layer and data layer. The data layer provides data storage services and assumes the responsibility of ensuring reliable and safe data. According to the actual needs of the system, the data layer stores the user information database, user log behavior database, vocabulary database, corpus database, comment data published by users, test data and so on. The business layer implements the core business logic of the recommendation system, including similar word mining, similar user mining, recommending words to learners using collaborative filtering algorithms, locating users' learning styles through clustering algorithms, and adjusting the pushing method. The user layer is responsible for the interaction between learners and the system, the server responds to user requests and displays the content results, such as word learning, information registration, liking and commenting, corpus uploading, completing tests and other functions. All user behavior data generated by this layer will be recorded in the log database.

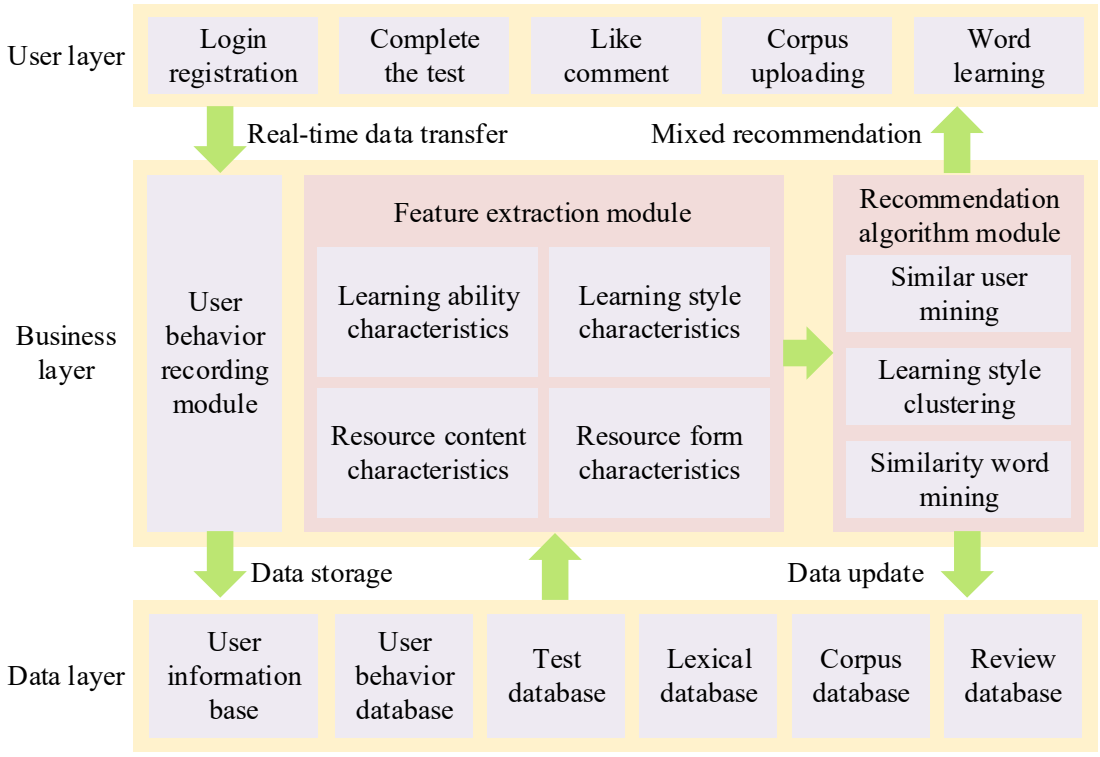


Figure 1: The architecture of the English lexical translation platform

II. B.Functional Flow Design

II. B. 1) Word learning

The system will recommend adapted test questions for users with different needs based on the information registered during the user's pre-registration, such as recommending past exams or high-quality prediction questions for users who choose the Grade 4 or 6 exam as their learning goal. Based on the learning data, the system can find similar users and words with high relevance for the user, forming a recommended word list for the user to learn. When learning a new word, users can choose whether to bookmark it as a new word, comment on it, like it or skip to the next word. As the cycle continues, the user's vocabulary book grows, and each time the software is reopened to begin learning, the system will recommend a new batch of words based on an updated version of the vocabulary base.

II. B. 2) Updating of the corpus

In the process of language learning, students from different professional backgrounds have individualized needs for English vocabulary. The system adopts a user-customizable corpus for students with different professional backgrounds, allowing users to upload different forms of corpus materials including: speech videos, interview recordings, news reports, professional papers, etc., to supplement the existing professional corpus. Through the co-occurring word matrix and trained word vectors, co-occurring words and similar words can be mined to prepare for the subsequent personalized word recommendation. The administrator can review the corpus materials uploaded by users and continuously extract the latest materials to update the existing corpus.

II. B. 3) Phase testing

The system develops the function of phase testing, users can choose to retest at any time when they feel their vocabulary has been greatly improved, or the last test can not accurately reflect their true level, the system constantly updates the latest recommended vocabulary list according to the user's test results, to ensure that the user memorizes words efficiently.

II. B. 4) Comments and Likes

The system opens up the function of users' comments on words and the permission of liking the wonderful high-quality comments, which makes the group intelligence of users show and increases the form of interaction when users memorize words. The user's word commenting and liking functions satisfy the users' emotional demands when using the software for learning, and increase the users' sense of communication and interaction when using the software to memorize words.

III. Learner profiling

Cluster analysis of learners in the English vocabulary learning system based on personalized recommendation to provide a more scientific basis for English vocabulary recommendation.

III. A. Clustering methods

III. A. 1) K-means algorithm

K-means clustering algorithm is a widely used and division-based clustering algorithm. The traditional K-means clustering algorithm uses the sum of squared errors as a criterion function for judging the clustering effect.K-means clustering algorithm divides the set containing R samples into K disjoint clusters, where the samples belonging to the same cluster have a higher degree of similarity, and the samples between the clusters of different classes have a lower degree of similarity.The basic idea of the K-means clustering algorithm is as follows: firstly, K samples are randomly selected as the initial clustering centers in a set containing R samples. The basic idea of K-means clustering algorithm is: first in a set containing R samples, randomly select K samples as the initial clustering centers, according to the Euclidean distance from each sample to the K centers, the samples are assigned to the most similar clustering centers, so as to obtain the K clusters that do not intersect each other. The new centers of the K clusters are recalculated, and then the R samples are assigned to the most similar class clusters according to the Euclidean distance principle again. Keep repeating this process iteratively until the K class clusters centers do not change anymore, thus obtaining K stable class clusters that do not intersect each other of the original set.

III. A. 2) Improved K-means

The traditional K-means clustering algorithm has randomness in the determination of the initial clustering centers, and also may select isolated or noisy points, which may lead to the clustering results are inconsistent with the real distribution of the sample dataset, and do not get the correct clustering results. In this paper, based on the principle of the highest degree of closeness (i.e., the lowest variance) of the sample space, and then using the average

distance after the introduction of weights as the radius, K initial clustering centers located in different regions are selected, and a K-means clustering algorithm combining the sample density and the improvement of weights is proposed. Optimizing the selection of initial clustering centers based on the principle of minimum variance avoids its randomness, while the initial clustering centers are selected from clusters that are far away from each other thus avoiding that they are located in the same cluster, and the weights are introduced in order to reduce the influence of outliers on the clustering results.

Assume that the given data set to be clustered $M=\{x_1,x_2,x_3,\cdots,x_n\}$ and each sample point is m dimensional, denoted as $x_i=\{x_{i1},x_{i2},x_{i3},\cdots,x_{im}\}(i=1,2,\cdots,n)$.

The weights of the different dimensional data of the sample points for calculating the distance are calculated as:

$$w_{id}=\frac{x_{id}}{\frac{1}{n}\sum_{i=1}^{n}x_{id}} \tag{1}$$

where x_{id} denotes the value of the d rd component of the i nd sample and $\frac{1}{n}\sum_{i=1}^{n}x_{id}$ denotes the mean value of the d th component in the sample data set.

The Euclidean distance between the sample points x_i,x_j is:

$$d(x_i,x_j)=\sqrt{(x_i-x_j)^T(x_i-x_j)} \tag{2}$$

The average distance from sample point x_i to all sample points is:

$$m_i=\frac{1}{n}w_{id}\sum_{j=1}^{n}d(x_i,x_j) \tag{3}$$

The variance of sample point x_i is:

$$var_i=\frac{1}{n-1}\sum(d(x_i,x_j)-m_i)^2 \tag{4}$$

The average Euclidean distance of the sample data points is:

$$cmean=\frac{1}{n(n-1)}w_{id}\sum_{i=1}^{n}\sum_{j=1}^{i}d(x_i,x_j) \tag{5}$$

The clustering error squared is:

$$E=\sum_{j=1}^{K}\sum_{x_l\in C_j}\left|x_l-C_j\right|^2 \tag{6}$$

The density of the sample points is measured by the variance, and according to the maximum density, the average of the distances weighted in the sample space is used as the radius, and K sample point located in a different area with the smallest sum of squared errors is selected as the initial clustering center. In this paper, we consider selecting the initial clustering center based on the closeness and weight of the sample space. If we want to achieve the effect of convergence of clusters, we first calculate the variance of the sample data points through equation (4), find the sample point with the highest density as the initial clustering center, and take the average of the weighted distance of the sample data points as the radius, and the sample data points located in this region constitute set M_1. The sample point with the smallest variance in $M-M_1$ is selected as the radius of the average of the assigned weight distances of all the remaining sample data points that are in the region constituting set M_2. The above steps are repeated until the dataset to be clustered M is divided into K sets located in different regions. Then the mean value of each set $M_1,M_2,...,M_K$ is used as the new clustering center of that set. Finally the sum of squares of the errors of the clusters is then calculated from equation (6).

III. A. 3) Basic steps

1) Determine the initial clustering center:

(1) Calculate the variance of each sample point from Eqs. (1)-(4), and find the sample point x_i^1 with the smallest variance as the initial clustering center. According to Eq. (5), the average Euclidean distance of the sample data points is obtained as the radius, so the first clustering set is obtained:

$$M_1, M_1 = \langle d(x_j, x_i^1) < cmean, j = 1, 2, \cdots, n \rangle \tag{7}$$

(2) Let $M = M - M_1$, find the sample point x_i^2 with the smallest variance in the new data set as the center of clustering, and use the average distance of the sample points as the radius to get the second set of clusters M_2:

$$M_2 = \left\{ d(x_j, x_i^2) < cmean, j = 1, 2, \cdots, n \right\} \tag{8}$$

(3) Repeat the above steps until K mutually disjoint set is found.

2) Construct the initial division:

(1) Obtain the Euclidean distance from each sample point to the K selected initial clustering centers according to equation (2), and then divide the sample points into the nearest classes to constitute the initial division.

(2) Calculate the mean value of each class of the initial division and use it as the new center of the class.

(3) Calculate the error sum of squares of the clustering results from equation (6).

3) Iterative update clustering:

(1) Obtain the new clustering center based on the last clustering, calculate the Euclidean distance from the sample data points to the new center from equation (2) and assign it to the nearest class.

(2) Calculate the mean value of each class and use it as the new center of the class.

(3) Calculate the sum of error squares of the clustering result from equation (6).

(4) Compare this result with the sum of error squares of the last clustering, and if $E' - E \prec 10^{-10}$, the center of clustering is satisfied that the center of clustering is no longer changed, the iteration is terminated, and the clustering result is output. Otherwise, continue the above steps until the clustering result converges.

III. B. Cluster analysis and results

In this paper, the English proficiency test data of 500 students were randomly selected from an online learning platform for cluster analysis. The proficiency test is divided into four modules: listening, speaking, reading and writing, and the score of each module is calculated in percentage. The cluster analysis was started by importing the English proficiency test data.

III. B. 1) Input variables

Using the improved K-means algorithm, drag the variables that need to be clustered into the “Variables” box. In this paper, we only need to cluster the students' listening, speaking, reading and writing scores, so we put the Listening scores, Speaking scores, Reading scores and Writing scores into the “Variables” box, In this paper, we only need to cluster the students' listening, speaking, reading and writing scores.

III. B. 2) Determining the number of clusters

Since the data are unlabeled, the selection of the number of clusters k is the most critical part of the K-means clustering process and also the most difficult. In order to assess whether the k value is suitable for the case of this paper, the contour coefficient (SC) and the variance ratio criterion (CH) are chosen to evaluate the performance of the subsequently generated model. Figure 2 shows the variation of the values of SC and CH with the number of clusters k. The peak value of SC is 0.287 when k = 5 and the peak value of CH is 5834 when k = 6. However, since the CH score decreases by 4.68% at k = 5 and SC only decreases by 3.48% at k = 6, it is not possible to determine the performance of the model when the clustering is best when k = 6.

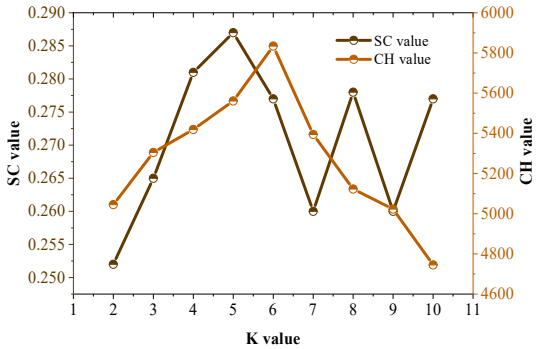


Figure 2: The value of SC and CH varies with the number of clustering k

III. B. 3) Determining the clustering method

There are two main methods of K-means clustering: iteration and classification and classification only. Iteration and classification method refers to first determine the initial category center point, and then according to K-means clustering algorithm for iterative classification. Only classification method refers to only according to the initial category center point classification, no longer do iterative operations. In this paper, the former method is used.

III. B. 4) Clustering results output

After setting the above three steps, click “OK” to get the results of K-means clustering, the number of iterations is 5. At the same time, the final clustering center results are shown in Figure 3. Here, the number of clusters corresponds to the English learning levels according to the clustering variables, i.e., 4 stands for introductory level, 3 stands for elementary level, 5 stands for intermediate level, 2 stands for intermediate-advanced level, 6 stands for advanced level, and 1 stands for proficient level. The number of students in each cluster can also be obtained by calculation, and the number of students in each cluster is shown in Figure 4. The percentage of students whose ELL level is intermediate is the largest, 31.4%, followed by beginner and introductory levels, which account for more than 18%, and there are fewer students whose ELL level is advanced and proficient, which account for less than 10% of the total number of students.

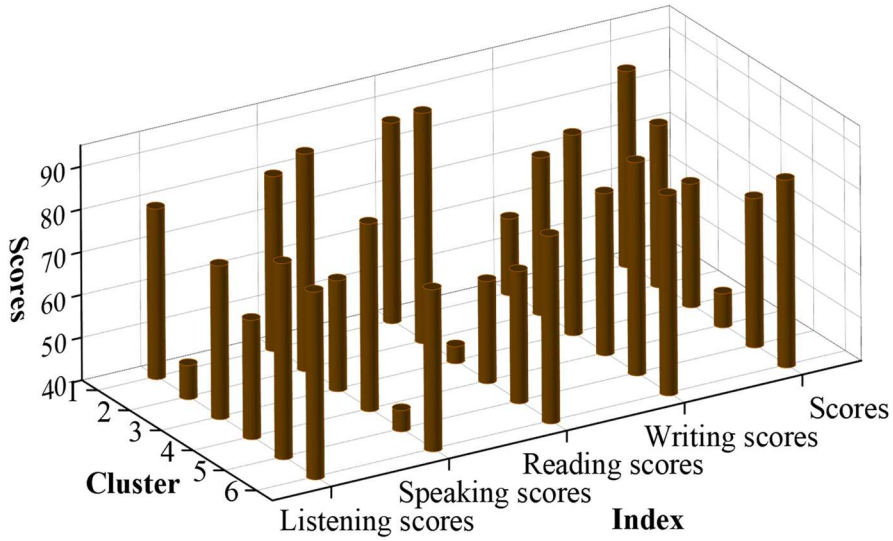


Figure 3: Final cluster center

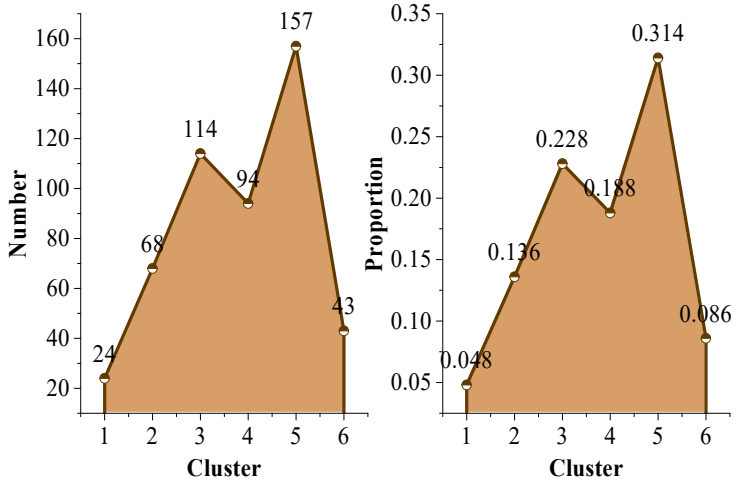


Figure 4: Student number of each cluster

IV. English Vocabulary Recommendation Model

Based on the English vocabulary recommendation module in the personalized recommendation system for English vocabulary learning, this chapter improves the implicit scoring model and proposes a hybrid recommendation algorithm based on items and based on user collaborative filtering algorithm.

IV. A. Time-aware implicit scoring models

IV. A. 1) Implicit scoring formula

The user's interest in the project is captured with the help of an effective implicit rating model to construct a more accurate user interest model.

First, the total duration of user visits to the project is calculated:

$$Sum\_time_{u,i} = \sum_{\varepsilon=1}^{M} Behave\_time_{u,i} \tag{9}$$

Then, the user's completeness of the project is calculated:

$$Item\_completion_{u,i} = \frac{Sum\_time_{u,i}}{\sum_{\varepsilon=1}^{M} Item\_time_i} \tag{10}$$

Finally, the implicit score calculation formula is obtained as:

$$R_{u,i} = \beta \cdot Item\_completion_{u,i} \tag{11}$$

where β is a rating cap score. M denotes the total number of visits to the item i by the user u, $Behave\_time_{u,i}$ denotes the single time duration of the user u visiting the item i, $Sum\_time_{u,i}$ denotes the cumulative time duration of the user u visiting the item M times i, $Item\_time_i$ denotes the time duration of the item i itself, $Item\_completion_{u,i}$ is the completeness of the user u visiting the item i, and $R_{u,i}$ is the value of the user's u rating of the item i.

IV. A. 2) Improvement of implicit scoring formulas

Most of the implicit rating models for obtaining users' ratings of items by mining their implicit feedback behaviors do not consider the factor of time context. For this reason, this paper adds the time factor to the implicit rating formula in the previous section. Define a number of time windows, denoted as $TW = \{TW_1, TW_2, ..., TW_n\}$, that match the user's behavioral cycle.

The time-aware implicit scoring model is:

$$Ratings\_Interest_u = \begin{Bmatrix} TW_1 : ((I_1, R_1), (I_2, R_2), ..., (I_m, R_m)) \\ TW_2 : ((I_1, R_1), (I_2, R_2), ..., (I_m, R_m)) \\ ... \\ TW_i : ((I_1, R_1), (I_2, R_2), ..., (I_m, R_m)) \\ ... \\ TW_n : ((I_1, R_1), (I_2, R_2), ..(I_m, R_m)) \end{Bmatrix} \tag{12}$$

where (I_m, R_m) denotes the items visited by the user and their corresponding ratings.

Finally, the user dynamic rating model is obtained as:

$$UDRM = y_1 TW_1 + y_2 TW_2 + ... + y_i TW_i + ... + y_n TW_n \tag{13}$$

where the value of $y_i (1 \le i \le n)$ is based on the dynamic activation of a specific time window, 1 when activated and 0 when not activated.

IV. B. User-based collaborative filtering recommendation algorithm

(1) Setting of time window

By analyzing the characteristics of user behavior, set up a time window model that meets the application scenario: $TW = \{TW_1, TW_2, ..., TW_n\}$.

(2) User modeling

Step1: Calculate the user's weight preference for the item type (i.e., label) under the same time window TW_n, $w_{u,t}, w_{u,t}$ represents the weight preference of user u for label t, and saves it as the user's dynamic interest model UDIM.

Step2: Based on the time-aware implicit rating model, compute to obtain the user's rating $r_{u,i}$ of the item under the same time window TW_n, where $r_{u,i}$ denotes the user's u rating of the item i, and save it as the user rating model UDRM.

(3) Establish the user-item rating matrix

Based on the user dynamic rating model UDRM, a "user-project" rating matrix $R_{u,i}$ is built corresponding to time window TW_n, and $R_{u,i}$ represents the user-project rating matrix.

(4) Calculate the similarity, and find the K nearest neighbors of target user u.

The cosine similarity formula (14) is chosen to calculate the user-to-user similarity $sim(u,v)$ in the user-item scoring matrix $R_{u,i}$, and the top K set of nearest neighbor users $S(u,K)$ are selected based on the magnitude of the similarity:

$$sim(u,v)=\frac{\sum_{i} r_{u,i} r_{v,i}}{\sum_{i} r_{u,i}^2 \sum_{i} r_{v,i}^2} \tag{14}$$

(5) User preference prediction

The user preference prediction formula is calculated as follows:

$$p(u,i)=\frac{\sum_{v\in N(i)} sim(u,v) r_{v,i}}{\sum_{v\in N(i)} sim(u,v)} \tag{15}$$

where, $S(v,K)$ represents the set of K users with high similarity to the target user u, $N(i)$ is the set of users who have behaved towards the item i, $sim(u,v)$ represents the similarity between the user u and the user v, $r_{v,i}$ represents the ratings of the target user u's neighboring user v on the item i, and $p(u,i)$ represents the predicted preference value of the target user u on the to-be-recommended item i.

(6) Generation of preliminary recommended item set

According to the size of the calculated preference value of target user u to be recommended item i, the first M items are taken as the preliminary recommended item set P.

(7) Generation of item-label matrix for preliminary recommended item set

Use the vector model to obtain the item-label matrix $V_{i,t}=\{tag_1, tag_2, ..., tag_1\}$, which represents the influence value of label t on item i, and the value of tag_t is 1 if the item contains the label, and 0 otherwise.

(8) Generation of final recommended itemset T

The user-label interest degree in the user dynamic interest model UDIM is transformed into the user-label matrix for the corresponding time window, and Equation (16) is used to calculate the interest preference value of the target user u for the preliminary recommended item set P, and then sort the preference values, and ultimately select the N items with higher predicted preference values as the final recommended item set T and recommend them to the target user u.

The formula for calculating the user's preference value for the preliminary recommended item set is as follows:

$$p'(u,i)=\sum_{i\in P} w_{u,t} p(u,i) V_{i,t} \tag{16}$$

where $w_{u,t}$ denotes the weighted preference of user u for label t, $p(u,i)$ denotes the predicted preference value of target user u for the to-be-recommended item i, and $V_{i,j}$ denotes the weighted value of label t for item i.

IV. C. Item-based collaborative filtering recommendation algorithm

(1) Setting of time window

By analyzing the characteristics of user behavior, set up a time window model that meets the application scenario: $TW=\{TW_1, TW_2, ..., TW_n\}$.

(2) User modeling

Step1: The weight preference of the user for the item type (i.e., the label) under the same time window TW_n is calculated, and $w_{u,t}, w_{u,t}$ represents the weight preference of user u for label t, and is saved as the UDIM of the user's dynamic interest model.

Step2: Based on the time-aware implicit rating model, the user's rating $r_{u,i}$ of the item under the same time window TW_n is computed, where $r_{u,i}$ denotes the user u rating of the item i, and saved as the user dynamic rating model UDRM.

(3) According to the user dynamic rating model UDRM, establish the "item-user" rating matrix $R_{i,u}$ for the corresponding time window TW_n, where $R_{i,u}$ represents the item-user rating matrix.

(4) Calculation of similarity degree

The cosine similarity formula (17) is chosen to calculate the similarity $sim(i,j)$ between item i and item j:

$$sim(i,j)=\frac{\sum_{u} r_{i,u}r_{j,u}}{\sum_{u} r_{i,u}^2 \sum_{i} r_{i,u}^2} \tag{17}$$

(5) Preliminary calculation of user preference prediction

The user preference prediction formula is calculated as follows:

$$p(u,i)=\frac{\sum_{j\in I(u)} sim(i,j)r_{j,u}}{\sum_{j\in I(u)} sim(i,j)} \tag{18}$$

where $I(u)$ is the set of items visited to user u, $sim(i,j)$ represents the similarity between item i and item j, $r_{j,u}$ represents the rating of item j by target user u, and $p(u,i)$ denotes the predicted preference value of item i to be recommended by target user u.

(6) Generation of preliminary recommended item set

According to the size of the calculated preference value of the target user u for the item i to be recommended, the first 14 items are sorted in descending order, and the first M items are used as the preliminary recommended item set P.

(7) Generation of item-label matrix for preliminary recommended item set

Use the vector model to obtain the item-label matrix $V_{i,t}=\{tag_1,tag_2,...,tag_t\}$, which represents the value of the influence of the label t on the item i, and the value of tag_t is 1 if the item has the label, and 0 otherwise.

(8) Generation of the final recommended item set T (use of cascade method)

The user-label interest degree in the user dynamic interest model UDM is transformed into the user-label matrix of the corresponding time window, and the interest preference value of the target user v for the preliminary recommended item set P is calculated using Equation (19), and then the preference values are sorted, and finally the N items with higher predicted preference values are selected as the final recommended item set T and recommended to the target user V.

The formula for calculating the user's preference value for the preliminary recommended item set is as follows:

$$p'(u,i)=\sum_{i\in P} w_{u,t}p(u,i)V_{i,t} \tag{19}$$

where $w_{u,t}$ denotes the weighted preference of user u for label t, $p(u,i)$ denotes the predicted preference value of target user u for the to-be-recommended item i, and $V_{i,t}$ denotes the weighted value of label t for item i.

IV. D. Hybrid recommendation algorithms

Figure 5 shows the flow of the hybrid recommendation algorithm proposed in this paper, which is described as follows:

(1) Use temporal context pre-filtering technology to establish user micro-documents.

(2) Judge the number of user behaviors of the user micro-document, if the number of user behaviors is less than the specified threshold, then execute (3), otherwise execute (4).

(3) Establish a sparse user document, invoke the project popularity recommendation algorithm, and directly generate the current most popular top-N projects for the target user.

(4) Create dense user profile, construct user interest model, and get user-label matrix. At the same time, calculate the total number of users and the total number of projects at this time, determine whether the total number of users is greater than the total number of projects, if so, execute (5), otherwise execute (6).

(5) Call the item-based collaborative filtering algorithm to generate the top-M pre-recommendation list.

(6) Call the user-based collaborative filtering algorithm to generate the top-M pre-recommendation list.

(7) Build item-label matrix for the obtained top-M items.

(8) Calculate user-item preferences and output top-N recommendation list.

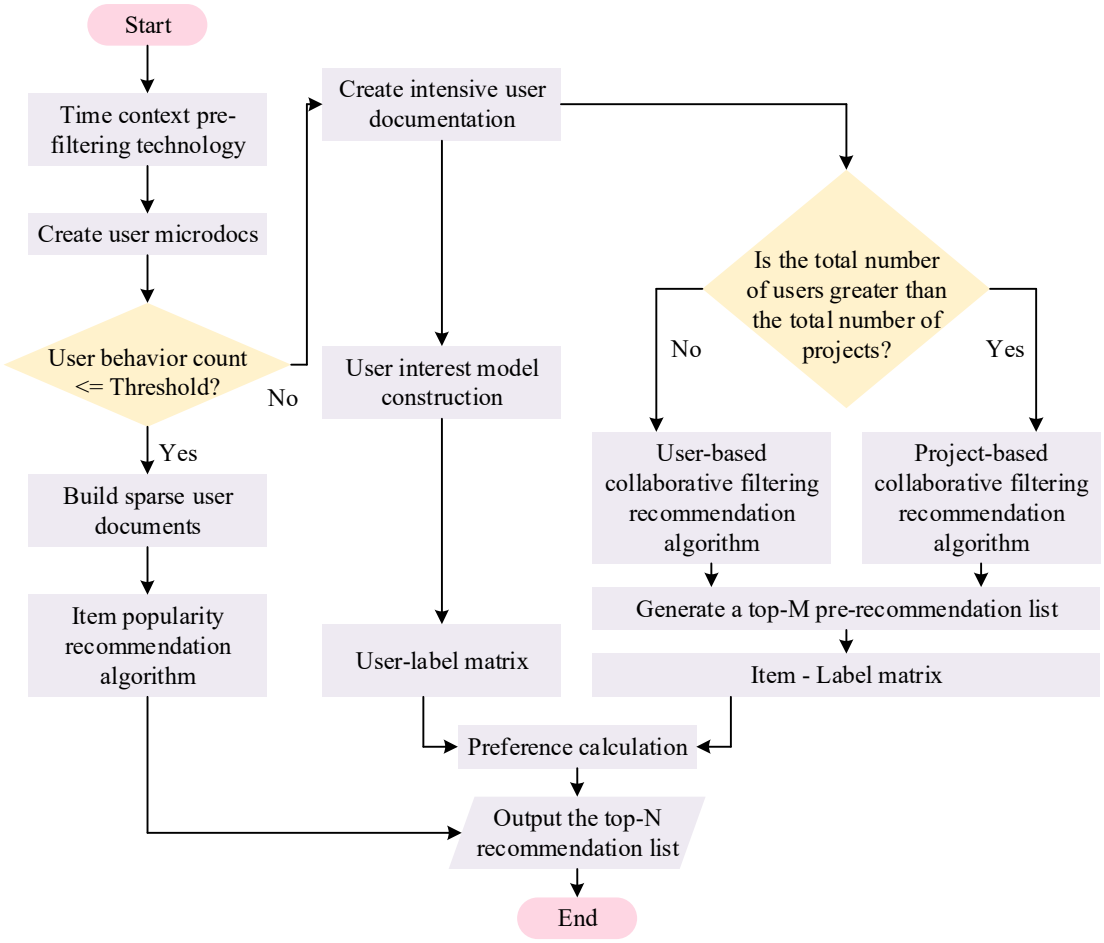


Figure 5: The process of mixing the recommended algorithm

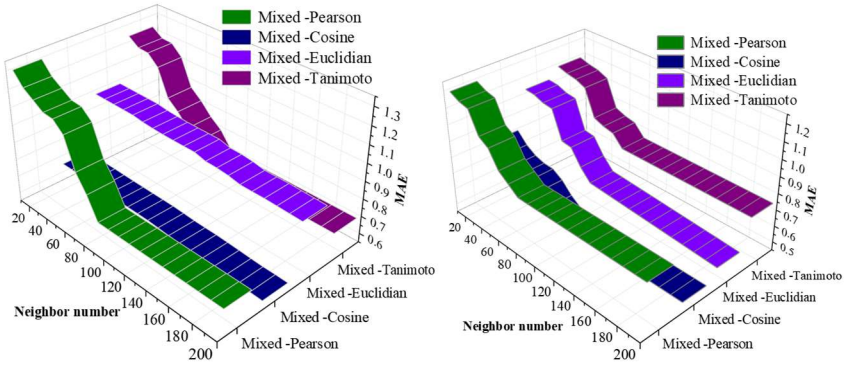
IV. E. Experiments and Analysis of Results

IV. E. 1) Data sets

The experiments in this paper use two real datasets from the external environment of E-Learning, the Algebra 2005-2006 (Alg) dataset and the Geometry 2006-2007 (Geo) dataset, extracted from the Cognitive Tutoring System and published by the Pittsburgh Science Learning Center. Both datasets contain implicit information about learners' interactions with the tutoring system and learning resources. In order to evaluate the performance of the algorithm, the datasets need to be divided into two parts: the training set (80%) and the test set (20%). The main purpose of the experiment is to test the prediction accuracy of the newly proposed method and MAE is selected for evaluation.

IV. E. 2) Tests of Similarity Measures

Mixed recommendation algorithms (Mixed) with different similarity metrics are considered to compare the performance: Mixed -Pearson, Mixed -Cosine, Mixed -Euclidian, Mixed -Tanimoto.Comparison of the MAE values of Mixed algorithms with different similarity metrics is shown in Fig. 6, and (a) and (b) represent the Alg and Geo datasets for the experimental results. In order to determine the optimal value of the number of nearest neighbors K for the KNN method on the Alg and Geo datasets, the experiments were selected to control the range between 10 and 200. By increasing the value of K, it was found that the Mixed algorithm using the Cosine similarity measure obtained better prediction accuracy with stabilized MAE values of 0.603 and 0.517 on the Alg and Geo datasets, demonstrating the superiority of the cosine similarity computation method chosen in this paper.



(a)Alg dataset (b)Geo dataset

Figure 6: Comparison of the MAE values of mixed algorithms of different similarity metrics

IV. E. 3) Comparative experimental results

The accuracy of the algorithm proposed in this paper is further verified in the following experiments, using the K value of 150, selecting the user-based collaborative filtering recommendation algorithm (UCF) and item-based collaborative filtering recommendation algorithm (ICF) as the comparison method, and comparing the MAE values under different incremental data to analyze the performance of the different algorithms for the recommendation of English vocabulary, and the results of the comparison experiments of the different recommendation methods are shown in Fig. 7 shows. The Mixed algorithm performs better in any case in the Alg and Geo datasets, with lower MAE values than the other two collaborative filtering recommendation algorithms, remaining at 0.606 and 0.514, reflecting its more accurate English vocabulary recommendation performance.

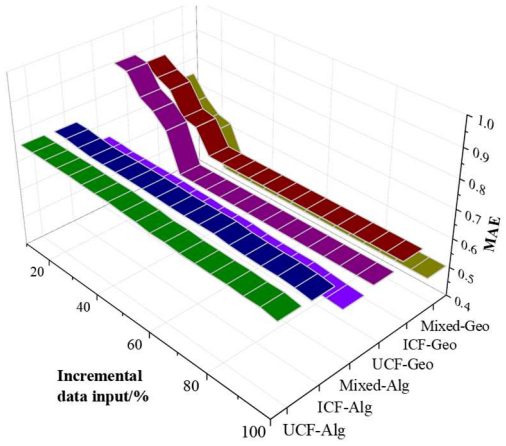


Figure 7: Comparison results of different recommended methods

V. Systematic evaluation and analysis

V. A. Survey design

The main purpose of this study is to explore the optimization path of English vocabulary learning in higher vocational colleges and universities, and to recommend personalized English vocabulary learning resources for learners. After the design of English vocabulary learning recommendation platform for learners' functional requirements is completed and implemented, the application effect of the system is to be tested, and the problems appearing in the platform are modified according to the results of the test, so as to further improve the system. In order to verify whether the English vocabulary resource recommendation platform improves learners' vocabulary learning efficiency and recommends appropriate vocabulary resources according to learners' needs to a certain extent, the application effect of the platform is verified through the questionnaire survey method.

The questionnaire survey targets were selected from non-English majors in a higher vocational college, and the feedback information of learners after using the platform was collected. The questionnaire on the use effect of the English vocabulary recommendation system was designed to evaluate the system's evaluation indexes from the

dimensions of satisfaction, learning attitude, and learning results of the English vocabulary learning recommendation platform. A total of 125 questionnaires were distributed and 108 were effectively collected.

V. B. Findings

V. B. 1) System satisfaction analysis

The satisfaction survey of the system was analyzed in terms of whether the learners supported the behavioral data being accessed and used X1, whether the system resources were comprehensively designed X2, and whether the recommended vocabulary met the individualized needs X3. The results of the system satisfaction survey are shown in Figure 8, with A~E denoting very much, quite much, generally, not quite much and not at all.64.14% of the learners strongly support the use of learner behavioral data, and 65.04% of the learners think that the system's resources for English learning are very comprehensive, while 56.68% of the learners and 26.97% think to varying degrees that the English vocabulary recommendation meets the needs, indicating that most learners think that the vocabulary recommendation function of the system can basically meet the learners' personalized vocabulary needs, which shows that the recommendation function of the system has a good recommendation effect. On the whole, most learners recognize the satisfaction of the system to varying degrees, and the analysis of learners' satisfaction shows that the English vocabulary learning recommendation system has a certain degree of effectiveness.

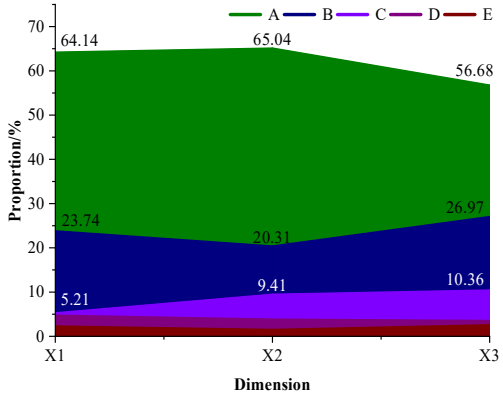


Figure 8: Results of systematic satisfaction survey

V. B. 2) Analysis of learning attitudes

The system was analyzed in terms of whether it is a powerful tool to assist vocabulary learning X4, whether the system can increase learners' interest in learning vocabulary X5, and whether the system improves learners' motivation to learn vocabulary X6. The results of the survey on learning attitudes are shown in Figure 9. 56.91% and 25.31% of the learners think that the system is a better tool to assist in learning English vocabulary to varying degrees, and 63.09% and 62.45% strongly agree that the system can increase learners' interest and motivation in learning English vocabulary. In general, the system has good positive effects on learning English vocabulary.

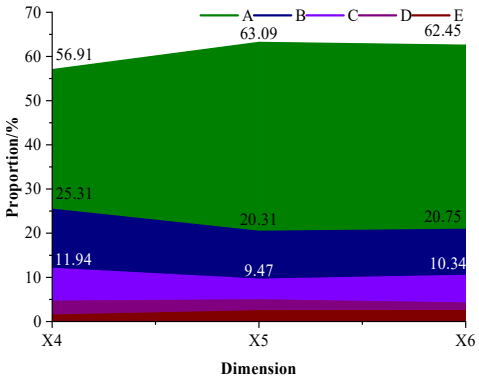


Figure 9: The results of the study attitude

V. B. 3) Analysis of learning outcomes

The results of the learning effect survey are shown in Figure 10. 54.61% and 27.09% of the learners think that the system can help learners build the overall structure between vocabularies X7, whether the system can improve the learners' efficiency in learning English vocabulary X8, and the overall learning effect improvement X9. 54.61% and 27.09% of the learners think that the system can help the learners build the overall structure between vocabularies in varying degrees and is useful for memorizing 58.26% of the learners agree that the system can improve the efficiency of learning English vocabulary, and 56.59% and 26.51% of the learners think that the system can improve the overall learning effect to different degrees. Overall learners recognized that the system is helpful in memorizing vocabulary.

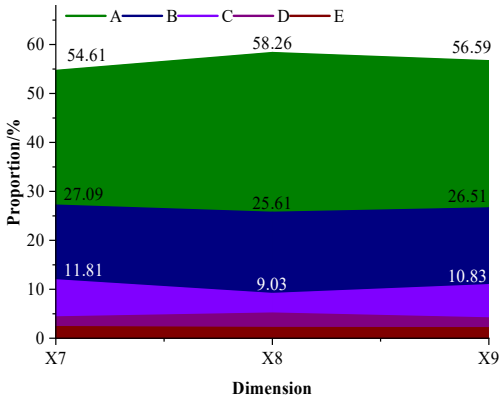


Figure 10: Results of the study effect

VI. Conclusion

In order to optimize the English vocabulary learning path in higher vocational colleges and universities, this paper proposes a personalized recommendation English vocabulary learning platform based on big data technology, analyzes the learner characteristics in it by using clustering methods, and designs a hybrid recommendation model for English vocabulary that integrates user-based and item-based collaborative filtering recommendation algorithms. Finally, the English vocabulary personalized recommendation learning system is evaluated through a questionnaire survey.

(1) The English learning levels of the students are classified into six categories through cluster analysis: introductory, beginner, intermediate, intermediate-advanced, advanced, and proficient, in which intermediate students account for 31.4%, the largest number, followed by beginner and introductory levels.

(2) The superiority of the proposed similarity measure and English vocabulary recommendation model is verified. The hybrid recommendation model using the cosine similarity calculation method has lower MAE values than other approaches in different datasets, with MAE values of 0.603 and 0.517. Meanwhile, the hybrid recommendation algorithms have lower MAE values than the individual adoption of the user-based collaborative filtering recommendation algorithms and item-based collaborative filtering recommendation algorithms, with MAE values of 0.606 and 0.514, respectively.

(3) The designed English vocabulary personalized recommendation learning system achieves good evaluations in the three dimensions of satisfaction, learning attitude and learning effect. 83.65% of the learners say that the English vocabulary recommendation meets the needs, more than 60% of the learners think that the system can greatly increase the interest and motivation in learning English vocabulary, and 58.26% of the learners are very much in favor of the system's effectiveness in improving the learning of English vocabulary. 58.26% of the learners recognize the effect of the system on improving the efficiency of English vocabulary learning.

The English vocabulary personalized recommendation system designed in this paper can provide learners with personalized recommendation services and improve the quality of their English learning. In the subsequent research, the learner model and recommendation model can be further improved to provide students with more humanized online learning services.

References

[1] Liu, J., & Zhang, J. (2018). The effects of extensive reading on English vocabulary learning: A meta-analysis. English language teaching, 11(6), 1-15.

[2] Liao, L. (2023). Artificial intelligence-based English vocabulary test research using log analysis with virtual reality assistance. Computer-Aided Design and Applications, 20(S9), 23-39.

[3] Li, J. (2022). Adaptive learning model of english vocabulary based on blockchain and deep learning. Mobile Information Systems, 2022(1), 4554190.
[4] Boontam, P. (2022). The Effect of Teaching English Synonyms through Data-Driven Learning (DDL) on Thai EFL Students' Vocabulary Learning. Shanlax International Journal of Education, 10(2), 80-91.
[5] Ehara, Y. (2018, May). Building an English vocabulary knowledge dataset of Japanese English-as-a-second-language learners using crowdsourcing. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).
[6] Tlili, A., Hattab, S., Essalmi, F., Chen, N. S., Huang, R., Chang, M., & Solans, D. B. (2021). A smart collaborative educational game with learning analytics to support English vocabulary teaching. IJIMAI, 6(6), 215-224.
[7] Cui, J. (2020). Application of deep learning and target visual detection in English vocabulary online teaching. Journal of Intelligent & Fuzzy Systems, 39(4), 5535-5545.
[8] Cui, Y. (2024). Application of Artificial Intelligence Technology in College English Vocabulary Teaching. Development, 6(7), 196-200.
[9] Liu, J. (2024). Enhancing English language education through big data analytics and generative AI. Journal of Web Engineering, 23(2), 227-249.
[10] Zhu, J., Zhu, C., & Tsai, S. B. (2021). RETRACTED: Construction and Analysis of Intelligent English Teaching Model Assisted by Personalized Virtual Corpus by Big Data Analysis. Mathematical Problems in Engineering, 2021(1), 5374832.
[11] Barabadi, E., & Khajavi, Y. (2017). The effect of data-driven approach to teaching vocabulary on Iranian students' learning of English vocabulary. Cogent Education, 4(1), 1283876.
[12] Zhu, Q. (2023). Empowering language learning through IoT and big data: An innovative English translation approach. Soft Computing, 27(17), 12725-12740.
[13] Peng, X. (2022). Holistic language teaching method in college English vocabulary teaching under big data and multimedia environment. Scientific Programming, 2022(1), 4250202.
[14] Fang, H., & Caili, L. (2021, September). The Application of Big Data Analysis in College English Classroom Vocabulary Memory and Learning Efficiency Teaching. In 2021 4th International Conference on Information Systems and Computer Aided Education (pp. 370-375).
[15] Soruç, A., & Tekin, B. (2017). Vocabulary learning through data-driven learning in an English as a second language setting. Educational Sciences: Theory & Practice, 17(6).