ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 1 - 2026

K-means clustering improves the efficiency of English vocabulary learning

Yannan Li[1] *and Jingwen Liu*[1,*]

[1] School of Humanities and Education, Jinan Preschool Education College, Jinan 250000, Shandong, China

SUMMARY: *In view of the obvious differences in learning rhythms, the homogeneity of review arrangements and the instability of learning output per unit time in English vocabulary learning, this paper proposes a vocabulary learning efficiency improvement method based on K-means clustering. Based on the behavioral data collected by the online learning platform, such as login frequency, learning duration, spelling accuracy, context question completion rate, review interval and forgetting backoff times, this paper constructed learners 'multi-dimensional feature vectors, and completed learners 'grouping after standardized processing. On this basis, it extracted vocabulary mastery preferences and generated vocabulary learning paths for individual differences. The experimental results show that the accuracy of vocabulary test in the experimental group reaches 83.6%, which is 8.8 percentage points higher than that in the control group. The number of words mastered per unit time was increased from 18.9 /h to 24.7 /h, the 7-day retention rate was increased by 10.7 percentage points, and the average review response time was shortened to 19.8 minutes. The results show that K-means clustering can effectively identify the structural differences of learners in learning engagement, memory retention and task response, and improve the efficiency of English vocabulary learning through vocabulary task scheduling and review node optimization.*

KEYWORDS: *K-means clustering; English vocabulary learning; Learning behavior analysis; Personalized intervention*

1 Introduction

With the continuous expansion of digital learning environment, English vocabulary learning has gradually shifted from traditional paper memory to the collaborative promotion mode of platform, data and intelligence. Learners' click, stay, spelling, review, test and error correction behaviors in mobile terminals, online course platforms and vocabulary training systems will continuously generate computable process data, which provides a new technical basis for identifying learning differences, optimizing teaching intervention and improving vocabulary learning efficiency [1, 2]. However, from the perspective of practical application, the current English vocabulary learning system still faces two prominent quantitative problems. First, the learning behavior data are obviously discrete and heterogeneous, and different learners have large differences in learning frequency, review cycle and mastery speed, which makes it difficult for the unified push strategy to stably adapt to diverse needs. Second, most platforms still rely on experience threshold or average progress to organize learning content, which is slow to respond to changes in individual memory state, and prone to the phenomenon of

*675781775@qq.com
https://doi.org/10.65102/is2026193 Published: 30 April, 2026

"repeated investment of mastered words" and "intervention lag of weak words", thus reducing the learning benefit per unit time [3-5]. Therefore, how to use data mining methods to extract stable patterns from complex learning behaviors and form differentiated vocabulary learning paths has become a key issue in the research of intelligent English vocabulary teaching.

There are many achievements on data analysis and adaptive support in foreign language learning. Warschauer et al. pointed out that educational data mining is promoting language learning research from result description to process modeling, and behavior trajectories, task records and interaction frequency in the learning platform can be transformed into important features reflecting learning states [6]. Bravo-Agapito et al. believe that data mining technology has strong applicability in foreign language learning scenarios, especially suitable for dealing with problems such as learner difference identification, resource matching and learning risk early warning, but the effect of the model highly depends on the way of feature organization and data quality [7]. Xie et al. pointed out in their review of adaptive learning research that in order to truly improve the learning effect, the personalized learning system not only needs to identify learners' differences, but also needs to transform the difference recognition results into an executable learning content scheduling mechanism [8]. Xu built an adaptive English vocabulary learning system based on machine learning, and verified the value of behavioral data for the generation of learning support strategies [9]. Wilschut et al. introduced the model-driven mechanism into the vocabulary learning process and combined with automatic speech recognition to improve the timeliness of learning feedback [10]. These studies show that the improvement of English vocabulary learning efficiency is not only a teaching method problem, but also a computer application problem oriented to multidimensional data modeling and algorithm optimization.

However, there is still room for further breakthroughs in the existing results. Some studies pay more attention to the prediction of learning results or risk identification, and the description of the internal difference structure of learners is still rough, so it is difficult to explain why the same training task can produce significantly different efficiency performance among different learners [11, 12]. Although some studies emphasize personalized learning, they are mostly based on preset rules, questionnaire classification or single performance index, and lack of joint modeling of review rhythm, error distribution, context recognition ability and forgetting backoff behavior in the learning process, resulting in insufficient dynamic adaptation ability of recommendation results [13-15]. Vanbecelaere et al. 's research shows that "whether adaptive or not" in the digital learning environment is not the only determining factor, but more importantly, whether the system can accurately identify the actual state of learners, and adjust the task difficulty and presentation order accordingly [16]. In this sense, if learners are not effectively divided into groups, the subsequent intervention design is easy to stay at the surface level of personalization, and it is difficult to form a real fine-grained support for behavior rules.

Clustering analysis provides a more feasible technical path to solve the above problems. Compared with supervised learning, the clustering method does not rely on pre-labeled category labels, and is more suitable for dealing with the intertwined scenarios of weak labels, sparse records and individual differences that are common in English vocabulary learning [17, 18]. Among them, K-means algorithm has the characteristics of clear structure, high computational efficiency, and easy deployment and connection with the learning platform. According to the distance relationship between samples, learners can be divided into several groups with similar behavior patterns, and then the common characteristics of different groups in vocabulary memory, review preference and task response can be further extracted [19, 20]. For English vocabulary learning, this method is not simply "grouping students", but uses the computer to vectorize the multidimensional learning behavior, iterative clustering and center update, identify stable learning types from discrete data, and provide data basis for subsequent

vocabulary push, exercise frequency adjustment and path generation.

Based on this, this paper focuses on the theme of K-means clustering to improve the efficiency of English vocabulary learning, and tries to construct a learner clustering and intervention optimization model for English vocabulary learning process. Starting from the multi-dimensional behavior data collected by the learning platform, this paper encodes and standardizes learners 'vocabulary learning behavior features, uses the K-means algorithm to complete learners 'grouping, and extracts vocabulary mastery preference features on this basis to generate personalized learning paths for different groups. This paper argues that the improvement of English vocabulary learning efficiency depends not only on the increase of the number of exercises, but also on the identification of learners 'task needs, training rhythm and reinforcement time points by computer algorithms, so as to realize the collaborative optimization of learning resource allocation and cognitive load control.

2 Design of English vocabulary learning efficiency improvement method based on K-means clustering

2.1 Classification and feature analysis of English vocabulary learning behavior data

In the research of improving the efficiency of English vocabulary learning based on K-means clustering, the classification and feature analysis of learning behavior data are not an auxiliary link, but the computational basis of subsequent clustering modeling, preference extraction and path generation. English vocabulary learning is ostensibly reflected in the teaching activities such as memorization, spelling, discrimination and review. In fact, it corresponds to a group of continuously changing learning behavior sequences, including the time of learners entering the system, the duration of task stay, the completion of task, the word category in the error set, the stability of review interval and the difference of responses to different types of questions. All these information can be collected by the learning platform in real time and converted into computable data. If the total score or single test results are still used as the basis for learning state judgment, it is difficult to identify the differences in vocabulary mastery speed, forgetting rhythm and practice preference of different learners, and the intervention strategies generated by the system are easy to stay at the average level, which is difficult to truly improve the learning benefits per unit time [1, 21]. Therefore, before K-means clustering modeling, it is necessary to first structurally classify English vocabulary learning behavior data and analyze the learning implications reflected by different types of data.

From the perspective of the data sources of the learning platform, this paper divides the English vocabulary learning behavior data into three categories. The first is basic access data, which mainly records the operation trajectory of learners after entering the system, such as login frequency, single learning time, learning time distribution, task startup times and page stay time. This kind of data can reflect the degree of learning engagement and usage habits, and is an important basis for judging learning activity and learning rhythm. The second is task performance data, including spelling accuracy, word sense discrimination score, example sentence matching completion rate, dictation pass rate, vocabulary repetition success rate, and error correction times. This kind of data is directly related to the quality of vocabulary mastery, and can be used to describe learners 'memorization effect and output ability. The third is the review adjustment data, which mainly covers the review interval, the number of forgotten back-off, the proportion of repeated learning, the frequency of wrong questions and the recovery time after task interruption. Compared with the first two types of data, this part can better reflect the characteristics of retention ability and self-regulation in the learning process, and is particularly

critical for identifying the two types of learners who "learn quickly but forget quickly" and "progress slowly but maintain stability". The classification results of the English vocabulary learning behavior data are shown in Table 1.

Table 1: Classification structure of English vocabulary learning behavior data

Data Category	Specific Indicators	Main Reflected Content
Basic Access	Login Frequency, Learning Duration, Page Dwell Time, Task Launch Count, Learning Time Distribution	Level of Learning Engagement, Platform Usage Habits, Learning Rhythm
Task Performance	Spelling Accuracy, Word Meaning Discrimination Score, Example Sentence Matching Completion Rate, Dictation Pass Rate, Number of Error Corrections	Quality of Vocabulary Acquisition, Semantic Understanding Ability, Output Performance
Review Regulation	Review Interval, Number of Forgetting Rollbacks, Repetition Learning Ratio, Error Re-practice Frequency, Recovery Duration	Memory Retention, Forgetting Characteristics, Self-Regulation Ability

This classification has direct significance for subsequent K-means clustering. On the one hand, different categories of data have large differences in dimension, distribution form and fluctuation amplitude. If the original behavior log is directly input into the clustering model without classification and feature analysis, it is easy to cause abnormal traction of high-frequency variables on the clustering center, so that the real explanatory learning differences are covered up. On the other hand, the classified data is more convenient for feature screening and standardization, and the originally scattered log records can be converted into a unified learner feature vector, so as to improve the stability and interpretability of clustering results [7, 8]. For example, it is not sufficient to judge learning motivation based on login frequency alone, because high-frequency login does not necessarily mean effective learning. However, if the login frequency was jointly analyzed with the length of stay, the correct rate and the review interval, the two types of learning behavior patterns of "high activity and low effectiveness" and "medium activity and high efficiency" could be more accurately distinguished.

At the data structure level, the learner behavior information after classification and collation can be further expressed as a "learner-feature" matrix. Each row in the matrix represents a learner, each column corresponds to a behavior feature, and the numerical value represents the statistical result or normalized performance of the learner in the corresponding dimension. Different from the "user-item" rating matrix in the traditional recommendation system, this matrix is not centered on the resource preference, but on the behavior performance in the process of English vocabulary learning, so it is more suitable for learners 'state clustering and behavior pattern recognition. Table 2 presents the basic form of the learner behavior characteristic matrix.

Table 2: Schematic representation of learner behavior characteristic matrix

Learner	Login Frequency	Average Learning Duration / min	Spelling Accuracy / %	Review Interval / d	Number of Forgetting Rollbacks	Contextual Question Completion Rate / %
Learner 1	18	24.6	82.4	2.1	3	79.8
Learner 2	11	17.3	68.7	4.8	8	63.5
Learner 3	26	21.9	90.3	1.9	2	88.1
…	…	…	…	…	…	…
Learner n	14	19.5	74.6	3.7	5	71.2

As can be seen from Table 2, there are obvious individual differences and unbalanced characteristics in the English vocabulary learning data. Some learners logged in frequently and had a long learning time, but the improvement of accuracy rate was limited, indicating that their learning input had not been effectively transformed into mastery quality. Some learners' overall learning time was not outstanding, but showed higher stability in spelling, context judgment and review rhythm, indicating that their learning strategies were more effective. Similar differences are usually difficult to capture in time if they are only judged by classroom observation or final test. With the help of learning platform records and computer feature extraction, these implicit differences can be made explicit into analysiable numerical structures. Before entering the K-means model, the data in this category still needs to be processed, including imputation of missing values, correction of outliers, coding of categorical features, and standardization of continuous variables. The reason is that K-means essentially relies on the distance calculation between the sample and the cluster center. If the feature dimension is inconsistent, the variable with a large range of values such as "learning time" will have too strong influence on the clustering result, and the discriminant value of the low-scale variable such as "the number of forgetting backoffs" may be weakened. The unified and standardized processing of multi-source behavior data can make the features of different dimensions participate in the clustering in the same computational space, so as to reflect the similarities and differences between English vocabulary learners more truly.

2.2 Construction of learner clustering model based on K-means algorithm

After completing the classification, feature extraction and standardization of English vocabulary learning behavior data, the learning efficiency improvement model enters the critical stage of cluster modeling. For English vocabulary learning, different learners have obvious differences in the dimensions of login frequency, single learning time, spelling accuracy, review interval, forgetting back-off times and context task completion. These differences do not appear in isolation, but often act on the vocabulary acquisition effect in a combined manner. If the unified threshold is still used to divide the learning level, it is easy to classify the learners with significant differences into the same type, which weakens the pertinence of personalized intervention. Based on this, this paper introduces the K-means algorithm to cluster and analyze the learners 'multi-dimensional behavior characteristics, so as to form a cluster structure that can be used for subsequent preference recognition and path generation.

K-means is a typical unsupervised clustering method based on distance measurement. Its basic idea is to divide the sample set into K clusters under the condition of a given cluster number K, so that the similarity of samples in the same cluster is as high as possible, and the difference between different clusters is as obvious as possible. Let the set of learner samples

after preprocessing be:

$$X = \{x_1, x_2, \cdots, x_n\} \tag{1}$$

Here, n denotes the number of learners and $x_i \in \mathbb{R}^d$ denotes the d-dimensional feature vector of the ith learner. The feature vector consists of login frequency, average learning time, spelling accuracy, word sense discrimination accuracy, review interval, forgetting backoff times and context question completion rate. The objective of K-means is to minimize the sum of squared errors from samples to the cluster centers to which they belong, and its objective function can be expressed as:

$$J = \sum_{j=1}^{K} \sum_{x_i \in C_j} \| x_i - \mu_j \|^2 \tag{2}$$

Here, C_j denotes the JTH cluster, μ_j denotes the center vector of this cluster, and $\|x_i - \mu_j\|^2$ denotes the square Euclidean distance between the sample and the cluster center. The meaning of equation (2) is that by constantly adjusting the sample attribution and cluster center position, the dispersion degree within the cluster decreases, so as to obtain a more stable learner behavior distribution. In the specific calculation process, K sample points are randomly selected from all the samples as the initial cluster centers, which are denoted as:

$$\mu_1^{(0)}, \mu_2^{(0)}, \cdots, \mu_K^{(0)} \tag{3}$$

Then, the distance between each learner sample xi and each cluster center is calculated, and the distance calculation formula is as follows:

$$d_{ij} = \|x_i - \mu_j\| = \sqrt{\sum_{t=1}^{d} (x_{it} - \mu_{jt})^2} \tag{4}$$

Here, d_{ij} represents the Euclidean distance between the ith learner and the JTH cluster center, x_{it} represents the value of sample x_i on the TTH dimensional feature, and μ_{jt} represents the value of the JTH center on the corresponding dimension. According to the minimum distance principle, the sample can be assigned to the nearest cluster, i.e:

$$x_i \in C_j \quad \text{if} \quad d_{ij} = \min_{1 \le r \le K} d_{ir} \tag{5}$$

After completing one sample assignment, each cluster center needs to be recalculated based on the newly formed clusters. The center update formula for the JTH cluster is given by:

$$\mu_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i \tag{6}$$

Here, $|C_j|$ represents the number of samples within the JTH cluster. Equation (5) is essentially the average of all learners' feature vectors in a cluster, which reflects the average state of the learners in the multi-dimensional learning behavior. The updated centers are then used in the next round of distance calculation and sample redistribution until the center changes

in the two adjacent rounds are small enough or the objective function J converges. In this paper, the following stopping condition is used:

$$\sum_{j=1}^{K} \| \mu_j^{(t+1)} - \mu_j^{(t)} \| < \varepsilon \tag{7}$$

Here, ε is the preset convergence threshold and t represents the number of iterations. When the overall change of the cluster center is lower than the threshold, it can be considered that the learner clustering result has become stable.

K-means is sensitive to the scale of the input features. Without normalization, a variable with a large range of values, such as average study time, can overweight the distance calculation, suppressing features that are also critical to learning efficiency, such as review interval or number of forgotten backoffs. To this end, this paper uses the Z-score normalization method to normalize the features of each dimension before clustering, and its expression is:

$$z_{it} = \frac{x_{it} - \bar{x}_t}{s_t} \tag{8}$$

where $\bar{x}_t$ and s_t represent the mean and standard deviation of the t-th feature, respectively. After standardization, different indicators are mapped into similar computational Spaces, and the distance measure can more truly reflect the behavioral similarity between learners.

In the determination of cluster number K, this paper combines the sum of squared error and contour coefficient to make a comprehensive judgment. The sum of squared errors decreased gradually with the increase of K, but when the decrease slowed down obviously, the significance of continuing to increase the number of clusters was limited. The silhouette coefficient can further measure the intra-cluster closeness and inter-cluster separation. By comparing the clustering performance under different K values, the clustering scheme with both interpretability and stability was finally selected. After grouping, learners can be divided into several categories, such as high-frequency steady state type, input fluctuation type, inefficient repetition type and potential improvement type. Different categories are not directly equivalent to the level of learning performance, but reflect the differences in their behavioral structure in the process of English vocabulary learning.

2.3 Vocabulary mastery preference extraction method based on multi-dimensional learning features

On the basis of learners 'grouping results, this paper constructs a vocabulary acquisition preference extraction method integrating multi-dimensional learning features, which maps learning behavior features, task performance features and memory retention features into the same computational space, so as to describe learners 'acquisition preferences for different vocabulary types and learning styles. The "preference" in English vocabulary learning is not simply an interest choice, but a stable performance tendency formed by learners in the long-term training process. It is reflected not only in the differences in response to spelling, word sense discrimination, context matching and review and review tasks, but also in the mastery efficiency of different lexical objects such as high-frequency words, abstract words, phrase collocations and polysemy. In order to avoid misjudgment caused by a single index, the multi-dimensional features of learners are represented as vectors in this paper:

$$u_i = [f_{i1}, f_{i2}, \ldots, f_{it}] \tag{9}$$

u_i represents the ith learner's preference input vector, and f_{it} represents the t-dimensional learning feature, including learning frequency, average learning time, spelling accuracy, review interval, forgetting backoff times, context question completion rate, and the proportion of wrong questions to practice again. At the same time, the lexical objects are represented as feature vectors:

$$w_j = [g_{j1}, g_{j2}, \dots, g_{jk}] \tag{10}$$

Here, w_j represents the attribute vector of the JTH word or task unit, and g_{jk} reflects its characteristics such as difficulty level, word frequency level, semantic abstraction degree, collocation complexity and context dependence degree. Through this bidirectional representation, learner states and lexical attributes can be entered into a unified modeling framework.

Considering that different features do not contribute equally to vocabulary mastery preferences, this paper uses weighted mapping to extract learners' latent preference representations on specific vocabulary dimensions. The implicit preference vector on the learner side is defined as:

$$h_i = \sigma(W_u u_i + b_u) \tag{11}$$

The attribute embedding on the lexical side is denoted as:

$$r_j = \sigma(W_w w_j + b_w) \tag{12}$$

where W_u and W_w represent learner feature transformation matrix and vocabulary feature transformation matrix respectively, b_u and b_w are bias terms, and $\sigma(\cdot)$ is a nonlinear activation function. After mapping, learning behavior features and lexical attribute features with different original dimensions and dimensions are compressed into the same latent space, which is convenient for subsequent similarity calculation and preference strength analysis.

Static mapping alone is still not sufficient to reflect true preference changes during learning. English vocabulary learning is obviously sequential, and the performance of the same learner in mastering vocabulary types in different learning stages is not constant. For example, in the initial stage, learners rely more on high-frequency basic words and spelling repeat training. As learning progresses, their attention will gradually turn to context discrimination and collocation application. Therefore, this paper further introduces the time decay mechanism to assign different weights to the learning records in different periods. The time weight of the t-th learning action is defined as:

$$\alpha_t = \frac{e^{-\lambda(T-t)}}{\sum_{s=1}^{T} e^{-\lambda(T-s)}} \tag{13}$$

Here, T is the total number of learning within the current statistical window and λ is the time decay coefficient. This equation indicates that the learning behavior closer to the current moment will contribute more to the preference judgment, while the influence of earlier recordings will be moderately weakened. After processing in this way, the model can more sensitively identify the change of the learner's recent mastery state.

On this basis, the dynamic preference representation of learners can be obtained by aggregating the behavioral feature sequence of learners in a time window:

$$p_i = \sum_{t=1}^{T} \alpha_t \, h_i^{(t)} \tag{14}$$

where $h_i^{(t)}$ represents the implicit feature representation of the learner after the TTH learning. The vector is not a simple average, but a comprehensive consideration of recent performance, historical behavior and time decay, which is more suitable to describe the real tendency in the process of vocabulary acquisition. In order to measure the matching degree between learners and a certain type of lexical objects, this paper uses the normalized correlation scoring function to construct the preference strength value:

$$s_{ij} = \frac{p_i^{\top} r_j}{\|p_i\| \, \|r_j\|}. \tag{15}$$

where, s_{ij} represents the preference score of the ith learner for the JTH lexical task, and the larger the value is, the more matching the lexical object of this type with the learner's current mastery characteristics. If a learner's performance is stable under the conditions of context judgment, example sentence recognition and medium review interval, its s_{ij} tends to be high on the contextual vocabulary task. On the contrary, if the spelling accuracy is low and the forgetting back-off is frequent, the system will recognize that the spelling is still more dependent on the mechanical repetition type of training. At the same time, this preference extraction is not used to simply recommend "like content", but to improve learning efficiency. The preference vector constructed in this paper essentially reflects learners' effective mastery tendency under different vocabulary types and task forms. It not only retains the group structure information provided by clustering, but also further mines the fine-grained differences within the group. Through this process, the system can advance from "knowing which class the learner belongs to" to "knowing what the learner is currently more suitable for learning and how to learn more effectively", thus providing a computable and interpretable basis for the generation of personalized learning paths in the next section.

2.4 Vocabulary learning path generation method for personalized intervention

After learner grouping and vocabulary acquisition preference extraction, the focus of the model is no longer on the recognition of learning states, but on the further transformation of the above calculation results into executable learning arrangements. For English vocabulary learning, what really affects the efficiency is not "how much is learned", but the order of vocabulary presentation, the configuration of task difficulty, the control of review time point and whether the training form is reasonable. If the system still adopts the unified vocabulary, fixed frequency and homogeneous exercise strategy, it is difficult to form substantial efficiency improvement even if the learner clustering results and preference vectors have been obtained. Therefore, on the basis of K-means clustering and multi-dimensional preference features, this paper constructs a vocabulary learning path generation method for personalized intervention, which integrates learner status, vocabulary attributes and task scheduling mechanism into the path optimization framework.

In this paper, vocabulary learning path is defined as an ordered sequence of vocabulary task set, presentation order, exercise mode and review node which is dynamically determined by the system according to learners' current characteristic state in a given learning cycle. Let the

dynamic preference vector of learner i at the current moment be p_i and the set of candidate lexical tasks be denoted as:

$$\mathcal{V}_i = \{v_1, v_2, \dots, v_m\} \tag{16}$$

Here, m represents the number of lexical tasks to be scheduled. Each task vj simultaneously contains attributes such as lexical difficulty, word frequency rank, semantic abstraction, question type form, and expected learning duration. In order to measure whether a certain lexical task is suitable for the current learner, this paper defines the lexical task matching score as

$$R_{ij} = \eta_1 s_{ij} + \eta_2 d_j + \eta_3 c_j - \eta_4 l_j \tag{17}$$

Here, R_{ij} represents the comprehensive adaptation score of learner i to task v_j; s_{ij} is the preference strength value obtained in the previous section; d_j indicates the degree of matching between the task difficulty and the learner's current ability interval. c_j indicates how well the task covers the current weak lexical category; l_j is the task load coefficient, which reflects the learning time cost and cognitive pressure. $\eta_1, \eta_2, \eta_3, \eta_4$ are the weight parameters. This formula shows that path generation is not simply selecting "the most favorite content", but striking a balance between interest bias, ability adaptation, weak reinforcement and load control.

In the candidate task screening stage, the system does not directly push all vocabulary items to learners, but sorts the tasks according to their comprehensive scores, and selects the top K tasks with higher scores to form a candidate set of personalized paths:

$$\mathcal{P}_i^{(0)} = \mathrm{Top} - \mathrm{K}(R_{i1}, R_{i2}, \dots, R_{im}). \tag{18}$$

Here $\mathcal{P}_i^{(0)}$ can be viewed as the set of initial learning paths. This method has two functions. Firstly, it can avoid the selection redundancy and system burden caused by too many tasks. Second, we can focus our push content on vocabulary items that are most likely to bring efficiency gains at the moment, rather than broadening learning coverage in general.

However, the candidate set obtained only by static ranking is still not enough to form a complete learning path. English vocabulary learning has obvious stages and progressiveness. Some words are suitable for pre-memorization tasks, while others are more suitable for consolidation and transfer. Based on this characteristic, this paper further introduces the path order optimization mechanism. Assuming that the transfer cost from the a-th task to the b-th task in the candidate path is $\phi(v_a, v_b)$, the total cost of a learning path Π_i with length K can be expressed as ϕ(va,vb), which comprehensively considers the difficulty jump, the frequency of question type switching and the risk of memory interference:

$$Q(\Pi_i) = \sum_{t=1}^{K} \omega_t R_{i\pi_t} - \sum_{t=1}^{K-1} \phi\left(v_{\pi_t}, v_{\pi_{t+1}}\right) \tag{19}$$

Here, π_t represents the task number corresponding to the t-th position in the path, and ω_t is the position weight coefficient. The objective of this formula is to maximize the overall benefit of the path, that is, to ensure that a single task itself has a high fitness, and control the jump between adjacent tasks is too large or repeat too dense. In this way, the generated paths are no longer simple stacks of high-scoring items, but have clearer structural continuity.

Considering that the effect of vocabulary retention is closely related to the review

arrangement, this paper adds a review time adjustment function to the path generation. For a lexical task vj that has entered the learning path, its next review trigger moment is defined as:

$$T_j^{review} = T_j^{learn} + \rho \cdot (1 + \kappa_j) \tag{20}$$

Here, T_j^{learn} is the first learning time, ρ is the base review interval, and κ_j is the individual moderator calculated based on error rate, reaction time, and the number of forgetting backoffs. If the error rate of a certain word is high in the recent training, κ_j is small and review will be triggered in advance. If the learner has shown a relatively stable grasp, the review node moves back moderately. Through this mechanism, path generation realizes the linkage of "task sequencing" and "review scheduling", and the review process is no longer regarded as an external additional operation. According to the above design, the generation process of personalized vocabulary learning path constructed in this paper is shown in Figure 1.

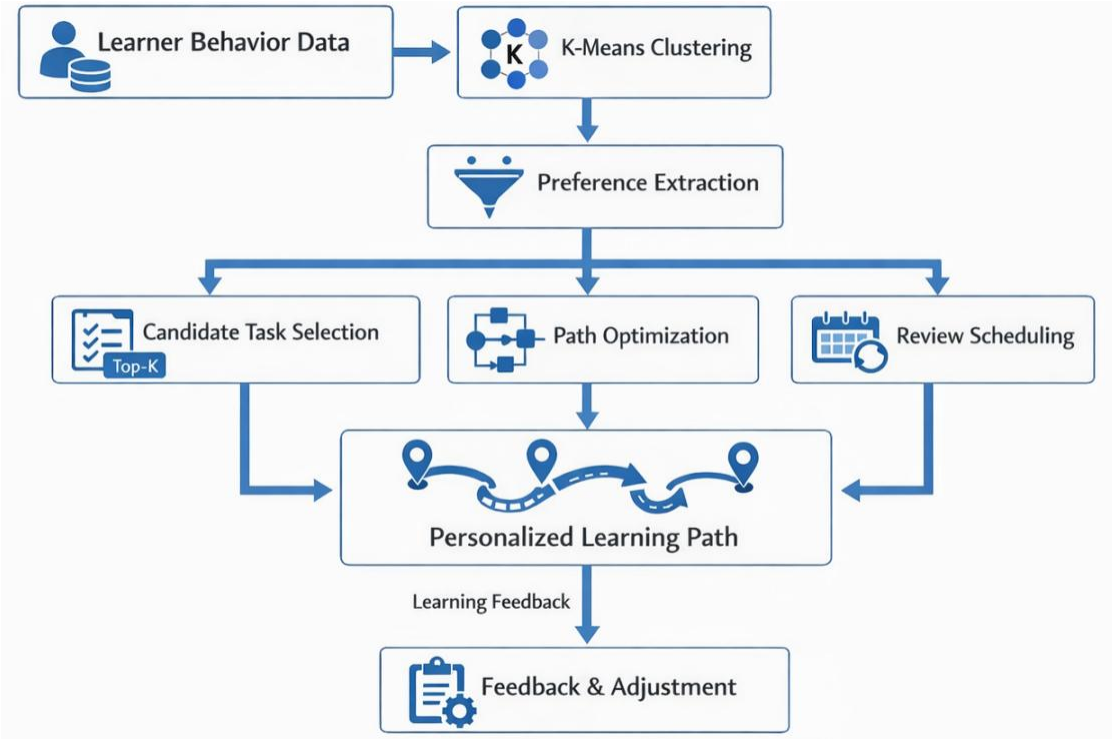

Figure 1: Flowchart of vocabulary learning path generation for personalized intervention

3 Experimental Analysis

3.1 Experimental platform and data source Settings

In order to test the actual effect of K-means clustering method in improving the efficiency of English vocabulary learning, this paper deployed the constructed model on an online English vocabulary learning platform and carried out experiments combined with real learning logs. The platform is implemented by a four-layer structure of "front-end interaction-business processing-algorithm service-data storage". The front-end is responsible for vocabulary learning task presentation, test result feedback and learning progress visualization, and the business layer is responsible for user authentication, task scheduling, learning record management and result return. The service layer of the algorithm completes the feature extraction of learning behavior, K-means clustering analysis, vocabulary mastery preference

calculation and personalized path generation. The data layer uniformly stores user information, vocabulary attributes, answer records, review logs and model output results. JSON format was used for data exchange between each module of the platform, which could stably support the closed-loop operation of learning behavior collection, feature transmission and algorithm call. Its general framework is shown in Figure 2.

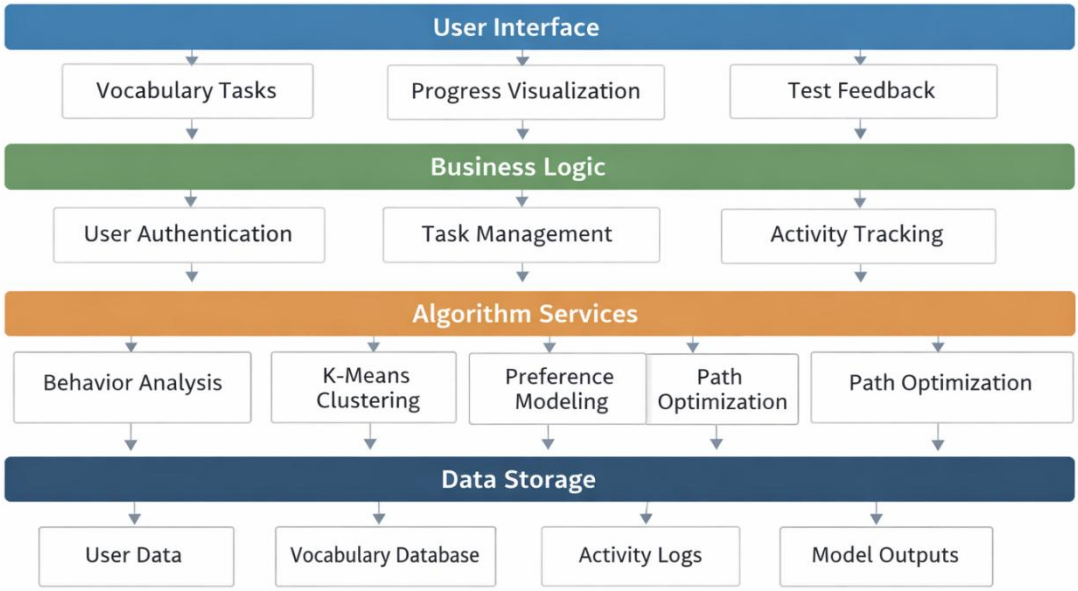

Figure 2: Structure of the experimental platform for improving vocabulary learning efficiency

From the perspective of platform implementation mechanism, K-means clustering is not regarded as a static statistical tool in isolation, but is embedded into the continuous analysis link of English vocabulary learning process data. After learners complete the operations such as login, word selection learning, spelling test, word sense discrimination, context judgment, error practice and periodic review in the platform, the system automatically records the information such as timestamp, task type, accuracy, reaction time, dwell time, review interval and forgetting back-off. Then, the algorithm service layer cleaned, aggregated and standardized the original log, transformed the discrete behavior into a structured feature vector, and then entered the K-means model to complete the learner clustering. The clustering results are not presented to the user directly, but are linked with the subsequent preference extraction and path generation modules to adjust the word push order, task difficulty and review nodes. This deployment mode ensures that the experiment does not stay at the level of offline data analysis, but can reflect the dynamic intervention ability of the computer system in the teaching scene.

The experimental data comes from the learning records of an English vocabulary learning platform in a university for 8 consecutive weeks. The original data contains a total of 30214 behavior logs involving 211 learners. In order to ensure the validity of the experiment, this paper preprocessed the data in three steps. Firstly, it deleted the records with duplicate submission, abnormal interruption and severe field missing. Second, the extreme value which obviously deviates from the normal range is corrected. Thirdly, multiple records are aggregated by learners to construct a feature matrix that can be used for clustering analysis. After processing, there are 186 valid samples and 26857 valid learning records. The samples cover a variety of task scenarios such as basic word recognition, spelling training, word sense discrimination, context matching and review consolidation, which can reflect the process characteristics of English vocabulary learning completely.

In order to enhance the interpretability of the experiment, the vocabulary tasks selected by

the platform are not a single vocabulary list, but are organized according to word frequency level, difficulty level, semantic abstraction degree and application scenario. The vocabulary resources include not only high-frequency basic words, but also middle and high order academic words, common phrase collocations and easily confused polysemy. The task format covers five categories: spelling input, word sense selection, example sentence matching, dictation recognition and interval review. The learner side data included variables such as grade, learning period preference, weekly login frequency, average learning time, spelling accuracy, context question completion rate, review interval, and forgetting back-off times. The main parameters of the experimental platform and data sources are shown in Table 3.

Table 3: Main parameters of experimental platform and data sources

Parameter Category	Specific Content
Platform Architecture	Front-end Interaction Layer, Business Processing Layer, Algorithm Service Layer, Data Storage Layer
Development and Operating Environment	Python 3.11, Flask, MySQL 8.0, Vue 3
Data Exchange Format	JSON
Number of Raw Logs	30,214
Valid Learning Records	26,857
Number of Learner Samples	186
Data Collection Period	Continuous 8 Weeks
Vocabulary Task Types	Spelling Training, Word Meaning Discrimination, Example Sentence Matching, Dictation Recognition, Review Consolidation
Learner Characteristics	Login Frequency, Learning Duration, Accuracy Rate, Response Time, Review Interval, Number of Forgetting Rollbacks
Vocabulary Resource Types	High-Frequency Basic Words, Academic Words, Phrase Collocations, Polysemous Words
Clustering Input Dimensions	8-Dimensional Core Behavioral Features
Model Output	Learner Clustering Results, Preference Features, Personalized Learning Paths

In the model running setting, this paper uses the standardized 8-dimensional behavioral features as the input of K-means, and the number of clusters is determined by the elbow method and the contour coefficient. During the experiment, the platform updates the feature cache once every round of learning task is completed, and the cluster to which the learner belongs is recalcuated in a fixed time window to reduce the interference of short-term fluctuations on the clustering results. At the same time, the system retains the pre-test and post-test data for subsequent evaluation of whether the model really improves the vocabulary learning efficiency. On the one hand, this experimental setup ensures the authenticity and continuity of data sources, and on the other hand, it also makes the "behavior collection, feature construction, cluster analysis and path intervention" form a complete and verifiable chain, which lays a foundation for the effect test and index analysis in the next section.

3.2 Test and analysis of the improvement effect of English vocabulary learning efficiency

In order to verify the improvement effect of the method proposed in this paper on the efficiency

of English vocabulary learning, further comparative tests are carried out on the basis of the above experimental platform and data environment. The test subjects were 186 students who participated in the platform learning. According to the initial vocabulary level and learning activity, 93 students in the experimental group and 93 in the control group were formed. The experimental group adopted the intervention process of "learning behavior feature extraction - K-means clustering - preference identification - personalized path push", while the control group used the original unified vocabulary push and fixed review mechanism of the platform. The whole test period is 8 weeks, and the system outputs the phased evaluation results at 2, 4, 6, and 8 weeks respectively. In order to avoid the influence of single test fluctuation on judgment, this paper defined learning efficiency as the effective mastery increment in unit learning time, and comprehensively analyzed the vocabulary test accuracy, the number of words mastered per unit time, the 7-day retention rate, the repetition error rate and the average review response time.

Five-fold cross validation was used to evaluate the stability of the model. In each round of validation, the training set, validation set and test set are divided in a 6:2:2 ratio to reduce the accidental error caused by sample division. The test results show that the personalized intervention based on K-means clustering is superior to the traditional unified push method in multiple dimensions. Table 4 shows that the correct rate of vocabulary test in the control group is 74.8%, and that in the experimental group is 83.6%, an increase of 8.8 percentage points. The number of words mastered per unit time increased from 18.9 /h to 24.7 /h, with an increase of 30.7%. This shows that the learning path generated after clustering can more effectively improve the conversion efficiency between learning input and learning output.

Table 4: Comparison of learning efficiency improvement effects

Indicator	Control Group	Experimental Group	Improvement
Vocabulary Test Accuracy / %	74.8	83.6	+8.8
Number of Words Mastered per Unit Time / words·h^{-1}	18.9	24.7	+5.8
7-Day Retention Rate / %	68.5	79.2	+10.7
Repeated Error Rate / %	21.6	12.4	-9.2
Average Review Response Time / min	31.4	19.8	-11.6

From the perspective of memory retention and review effect, the retention rate of the experimental group was 79.2%, which was significantly higher than that of the control group (68.5%), and the increase was 10.7 percentage points. The repeated error rate decreases from 21.6% to 12.4%, a decrease of 9.2 percentage points. At the same time, the average review response time of the experimental group was shortened to 19.8 minutes, while that of the control group was 31.4 minutes, a reduction of 11.6 minutes. The above results show that the proposed method can not only improve the short-term recognition accuracy, but also identify weak words faster and trigger review intervention in advance, thereby improving the quality of vocabulary consolidation.

As shown in Figure 3, in the second week, the vocabulary accuracy of the experimental group was 71.5%, higher than that of the control group (69.2%). By week 4, the gap had widened to 5.4 percentage points. At week 8, the experimental group increased to 83.6%, and the control group was 74.8%, and the difference between the two groups increased to 8.8 percentage points. It can be seen that the experimental group always maintains a steeper growth slope in the test cycle, indicating that the learning path based on K-means clustering does not only produce short-term promotion in the early stage, but can continuously modify the task

sequence and review rhythm with the continuous accumulation of learning logs, so that the system intervention keeps a high match with the actual state of learners.

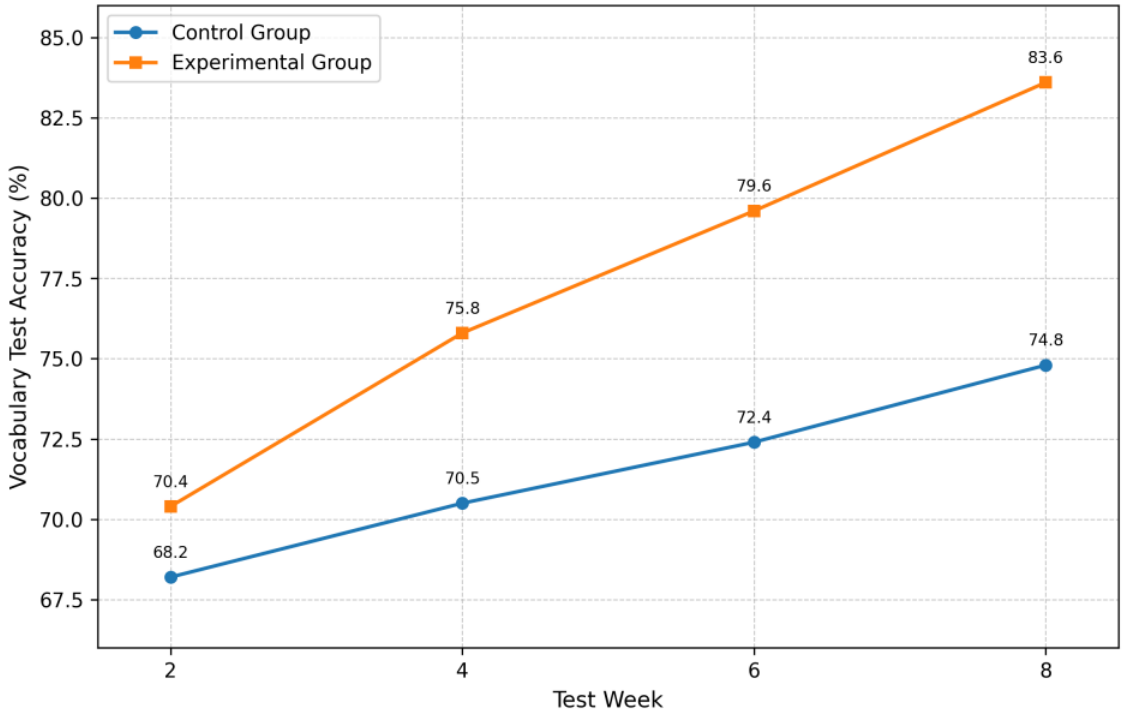

Figure 3: Trends of learning efficiency in different testing weeks

On the whole, the proposed method shows good stability and practicability in the English vocabulary learning efficiency improvement test. Its advantage does not come from the reinforcement of a single question type, but lies in the continuous collection, clustering recognition and dynamic feedback of learning behavior by computer, so as to transform the originally scattered learning records into executable personalized intervention basis, and then realize the fine regulation of vocabulary learning process.

3.3 Analysis of key performance indicators and validation of model validity

After completing the learning efficiency improvement effect test, this paper further verifies the effectiveness of the model from three aspects of clustering quality, operating efficiency and application stability. For the English vocabulary learning platform, K-means clustering is not an independent offline statistical step. It is directly involved in learner clustering, preference extraction and path scheduling. Therefore, its performance not only relates to whether the clustering results are clear, but also affects whether the subsequent personalized intervention can be stably implemented. Therefore, this paper selects the contour coefficient, the number of clustering iterations, the average response time, the accuracy of path matching and the improvement of 7 d retention rate as the key indicators to comprehensively analyze the performance of the model.

Experimental results show that the proposed method has good stability in both clustering quality and system application. The silhouette coefficient reached 0.71, indicating that the learners had high intra-cluster similarity and obvious inter-cluster differences, and the clustering results had good structural interpretation ability. The average number of iterations is 9, which indicates that the model can converge quickly under the current sample size, and there

is no obvious center oscillation phenomenon. At the same time, the average response time of the system is controlled at 214 ms, which can meet the basic requirements of real-time intervention in online learning platform. More noteworthy is that the accuracy of path matching reaches 86.9%, which indicates that the vocabulary learning path generated by clustering results and preference features has a high consistency with the actual completion of learners. The 7-day retention rate increased by 10.7 percentage points, which also verified the effectiveness of the model intervention from the level of memory retention. The relevant results are shown in Table 5.

Table 5: Analysis results of key performance indicators

Indicator	Test Result	Description
Silhouette Coefficient	0.71	The clustering structure is relatively clear.
Average Number of Iterations / times	9	The convergence speed is relatively stable.
Average Response Time / ms	214	It meets the real-time calling requirements of the platform.
Path Matching Accuracy / %	86.9	The consistency between path generation and learning performance is relatively high.
Improvement in 7-Day Retention Rate / percentage points	10.7	It has a significant promoting effect on memory consolidation.

As can be seen from Table 5, the proposed model does not significantly increase the system burden due to the introduction of multidimensional behavior characteristics and path generation mechanism, but instead achieves relatively stable learner identification and task scheduling in a short response time. The silhouette coefficient and the path matching accuracy together show that K-means clustering can not only complete the behavior clustering, but also provide an interpretable structural basis for subsequent personalized intervention. The increase in retention rate further indicates that the advantage of this method is not only reflected in short-term test scores, but also reflected in the continuous improvement of vocabulary memory. Comprehensive analysis shows that the model constructed in this paper achieves a better balance between computational efficiency, clustering effectiveness and teaching application value, and can provide reliable technical support for improving the efficiency of English vocabulary learning.

4 Discussion and comparative analysis

From the experimental results, the English vocabulary learning efficiency improvement method based on K-means clustering proposed in this paper shows more stable application advantages than the traditional unified push mode. The correct rate of vocabulary test in the experimental group reached 83.6%, which was 8.8 percentage points higher than that in the control group. The number of words mastered per unit time increased from 18.9 /h to 24.7 /h, the 7-day retention rate increased by 10.7 percentage points, and the repeat error rate decreased by 9.2 percentage points. This shows that the effectiveness of this method does not come from simply increasing the amount of practice, but lies in the continuous collection, clustering recognition and path reconstruction of learning behavior by computer, so that the presentation order of vocabulary tasks, review rhythm and learners 'state are matched to a higher degree. Compared

with the conventional model that relies on fixed vocabulary and empirical rules, K-means clustering can identify the structural differences in learners 'engagement frequency, forgetting speed and task response faster, so as to improve the learning output per unit time.

Further analysis shows that the advantages of the proposed method are mainly reflected in two aspects. On the one hand, the grouping of learners makes the system get rid of the extensive way of "organizing teaching according to the average level". Although different learners are in the same course environment, their vocabulary mastery process is not synchronized. If they continue to use the same training sequence and review cycle, some learners will often be repeatedly engaged, and some learners will lag behind the intervention. In this paper, the learning behavior vector is mapped into a relatively clear cluster structure by K-means, and the differential intervention is completed by combining the preference characteristics and the path scheduling strategy. Therefore, good results are achieved in the improvement of accuracy rate and retention rate. On the other hand, the method also maintains good deployability at the system response level, with an average response time of 214 ms and an average number of iterations of 9, indicating that the computational cost is still within the acceptable range of the online learning platform, and the real-time performance of the system is not significantly reduced by the introduction of clustering and path optimization.

However, the model in this paper still has some limitations. K-means essentially depends on the distance relationship between samples, and is sensitive to the initial clustering center and feature scale. When the learner's behavior changes in a short period of time, the clustering results may be phased offset, which affects the stability of subsequent path push. In addition, the modeling is mainly based on structured learning logs, and the utilization of unstructured data such as speech following performance, open paraphrase answering and example sentence generation is still insufficient, which means that the system still has room for improvement in the recognition ability of deep vocabulary mastery states. The subsequent research can consider the introduction of time attenuation mechanism and incremental update strategy to enhance the real-time response ability of the model to the change of learning state. At the same time, natural language processing methods can be combined to process text and speech feedback data to further improve the fineness of personalized intervention.

5 Conclusion

With the continuous development of online learning platforms, learning analysis technologies and educational data mining methods, English vocabulary learning is gradually shifting from experience-driven to data-driven. Focusing on the topic of K-means clustering to improve the efficiency of English vocabulary learning, this paper constructs an overall method framework consisting of learning behavior characteristics analysis, learners 'grouping, vocabulary acquisition preference extraction and personalized learning path generation. The results show that K-means clustering can effectively identify the differences of learners in learning frequency, correct rate, review rhythm and forgetting characteristics, and further convert these differences into executable path intervention basis, so as to improve the efficiency of content matching and review organization in the vocabulary learning process. The experimental results show that the proposed method has good application effect. Compared with the traditional unified push mode, the accuracy of vocabulary test in the experimental group reached 83.6%, which was 8.8 percentage points higher. The number of words mastered per unit time increased from 18.9 /h to 24.7 /h, and the 7-day retention rate increased by 10.7 percentage points, indicating that this method not only improves the short-term learning effect, but also enhances the stability of vocabulary memory. At the same time, the average response time of the system is 214 ms, which shows that the model has certain real-time deployment ability in the online learning platform.

In general, the introduction of K-means clustering algorithm into English vocabulary learning scenarios can provide reliable computational support for learning behavior recognition, learning path adjustment and teaching intervention optimization. This paper provides a reference method for the intelligent improvement of English vocabulary learning platform, and also provides an experimental basis for the subsequent research on more elaborate adaptive vocabulary learning.

References

[1] Chen C H, Yang S J H, Weng J X, et al. Predicting at-risk university students based on their e-book reading behaviours by using machine learning classifiers[J]. Australasian Journal of Educational Technology, 2021, 37(4): 130-144.

[2] Le Quy T, Friege G, Ntoutsi E. A review of clustering models in educational data science toward fairness-aware learning[J]. Educational data science: Essentials, approaches, and tendencies: Proactive education based on empirical big data evidence, 2023: 43-94.

[3] Papadogiannis I, Wallace M, Karountzou G. Educational data mining: A foundational overview[J]. Encyclopedia, 2024, 4(4): 1644-1664.

[4] Schmitt N. Instructed second language vocabulary learning[J]. Language teaching research, 2008, 12(3): 329-363.

[5] Webb S. The effects of repetition on vocabulary knowledge[J]. Applied linguistics, 2007, 28(1): 46-65.

[6] Warschauer M, Yim S, Lee H, et al. Recent contributions of data mining to language learning research[J]. Annual Review of Applied Linguistics, 2019, 39: 93-112.

[7] Bravo-Agapito J, Bonilla C F, Seoane I. Data mining in foreign language learning[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2020, 10(1): e1287.

[8] Xie H, Chu H C, Hwang G J, et al. Trends and development in technology-enhanced adaptive/personalized learning: A systematic review of journal publications from 2007 to 2017[J]. Computers & Education, 2019, 140: 103599.

[9] Xu Y. An adaptive learning system for english vocabulary using machine learning[J]. Mobile Information Systems, 2022, 2022(1): 3501494.

[10] [Wilschut T, Sense F, van Rijn H. Speaking to remember: Model-based adaptive vocabulary learning using automatic speech recognition[J]. Computer Speech & Language, 2024, 84: 101578.

[11] Kojic-Sabo I, Lightbown P M. Students' approaches to vocabulary learning and their relationship to success[J]. The Modern Language Journal, 1999, 83(2): 176-192.

[12] Mizumoto A, Takeuchi O. Examining the effectiveness of explicit instruction of vocabulary learning strategies with Japanese EFL university students[J]. Language Teaching Research, 2009, 13(4): 425-449.

[13] Vanbecelaere S, Van den Berghe K, Cornillie F, et al. The effectiveness of adaptive versus non-adaptive learning with digital educational games[J]. Journal of Computer Assisted Learning, 2020, 36(4): 502-513.

[14] Alamer A, Sonbul S, El-Dakhs D A S. Revisiting the validity of the vocabulary learning strategies questionnaire using the confirmatory composite analysis (CCA): Setting new directions for the field[J]. International Journal of Applied Linguistics, 2025, 35(1): 193-217.

[15] Fan T, Zhang S. Sparse Information Filtering for English Language Repositories Using Multilevel Interactive Attention Mechanism[J]. Informatica, 2025, 49(33).

[16] Alaff A, Uluyol Ç. Integrating Equation-Based Labeling and Classification for Adaptive Turkish Vocabulary Acquisition[J]. Informatica, 2025, 49(27).

[17] Wang B. A hybrid fuzzy logic and deep learning model for Corpus-Based German Language learning with NLP[J]. Informatica, 2025, 49(21).

[18] Jaikrishnan S, Ismail H H. A review on vocabulary learning strategies used in learning English as a second language[J]. International Journal of Academic Research in Business and Social Sciences, 2021, 11(9): 297-309.

[19] Ghalebi R, Sadighi F, Bagheri M S. Vocabulary learning strategies: A comparative study of EFL learners[J]. Cogent Psychology, 2020, 7(1): 1824306.

[20] Fan N. Strategy use in second language vocabulary learning and its relationships with the breadth and depth of vocabulary knowledge: A structural equation modeling study[J]. Frontiers in psychology, 2020, 11: 752.

[21] Wang H. An Empirical Study of Data Mining Technology in English Learning Outcome Prediction[J]. International Journal of e-Collaboration (IJeC), 2024, 20(1): 1-14.

[22] Li M. Research on the construction of English vocabulary learning recommendation system based on multi-objective crow search algorithm[J]. Systems and Soft Computing, 2025, 7: 200304.