Journal of Data Processing

DLINE JOURNALS

Print ISSN: 2278 – 6481
Online ISSN: 2278 – 649X

JDP 2025: 15 (4)
https://doi.org/10.6025/jdp/2025/15/4/175-182

Modeling Clustering Approaches to Recommender Systems in Language Datasets

Chen Wang
School of Physical Education
Xi'an Fanyi University, Xi'an, Shaanxi
710105. China
13892899088@126.com

ABSTRACT

This paper proposes a novel approach to improving personalized recommendations in English resource libraries. Addressing challenges such as information overload, lack of personalization, cold start problems, and algorithmic complexity, the study leverages the K-means clustering algorithman unsupervised machine learning technique to group users and resources based on similarity. By transforming multidimensional resource attributes (e.g., topic, proficiency level, target audience) into onedimensional data via dimensionality reduction, the system improves storage, search, and recommendation efficiency. The model integrates user English proficiency data, processes it via dynamic multimodal modeling and principal component analysis, and clusters users for tailored suggestions. Experimental results demonstrate that the K-means based system outperforms traditional collaborative filtering in recommendation accuracy, although recall rates vary with list size. User activity data reveal consistent personalized recommendation usage (~40% daily peak) and search spikes at 15:00 and 21:00, likely reflecting student and working user learning patterns. The study concludes that the proposed method effectively enhances resource management and user satisfaction, offering a scalable and efficient solution for digital English libraries. Future work includes refining algorithms to improve accuracy and user experience further.

Keywords: Personalized Recommendation, English Resource Library, K-means Algorithm, Clustering Analysis, Dimensionality Reduction, Information Retrieval

Received: July 2025, Revised 28 August 2025, Accepted 14 September 2025

1. Introduction

With the swift advancement of information technology and the rise of digital libraries, we find ourselves in an age characterized by an abundance of information. The wealth of online resources simplifies information retrieval. However, it also introduces a new challenge for users: how to sift through this extensive content to find resources that meet their specific needs. Confronted with information overload, traditional search techniques have proven insufficient to meet user requirements. Consequently, personalized recommendation systems have increasingly become a prominent area of study in the realm of information retrieval.

Personalized recommendation systems evaluate users' past behaviors, preferences, and interests to deliver tailored content suggestions, significantly enhancing the efficiency and quality of information retrieval. Among the various approaches to personalized recommendations, the K-means algorithm has gained notable attention due to its straightforwardness and effectiveness. The K-means algorithm is a method of clustering analysis that organizes data points into K clusters, ensuring that data points within the same cluster are more alike, while those in different clusters are more distinct. This clustering based approach to personalized recommendations provides users with tailored resource suggestions, thereby increasing their satisfaction with the information provided.

This paper intends to investigate the use of the personalized recommendation method based on the K-means algorithm in English resource library information guidance. We will conduct a thorough review of the most recent developments in personalized recommendation systems on both national and international scales and offer a comprehensive introduction to the technical foundations and application mechanisms of the K-means algorithm. Through the design of well organized experiments and subsequent data analysis, we will validate the effectiveness and practicality of the proposed method. It is anticipated that this research will serve as a valuable reference for the ongoing advancement of personalized recommendations in English resource libraries, providing users with more innovative and more tailored information recomme ndation services.

2. English Resource Library and its Challenges

The inception of English resource libraries stemmed from the necessity to organize and oversee a vast volume of information. With the evolution of the internet and digital technology, countless English language resources have surfaced online. To better satisfy the requirements of students, researchers, educators, and enthusiasts involved in English learning and research, English resource libraries have been established. These libraries compile a significant number of digital and physical resources, offering users a convenient platform to easily access and utilize a variety of information related to the English language. The applications of English resource libraries are varied. Students and language learners can discover a wealth of educational materials, such as textbooks, language exercises, interactive tutorials, and audiovisual resources, to enhance their language skills. Researchers and academics can access scholarly articles, research documents, journals, and reference materials to support comprehensive research and stay up to date on the latest advancements in their fields. Further more, English resource libraries play a crucial role in fostering cultural exchange and understanding. By providing literary works, art collections, historical documents, and multimedia content, these libraries promote exploration and appreciation of English culture. They also act as conduits for intercultural communication and gratitude.

With the emergence of the contemporary digital age, English resource libraries have transitioned from conventional physical collections to online platforms, offering increased accessibility and ease of use [8]. Users can access these libraries at any time and from any location via digital interfaces such as web portals and mobile apps, facilitating global information exchange. The inception and utilization of English resource libraries have significantly transformed the manner in which we access and leverage English language materials. They have become essential tools for learners, researchers, and enthusiasts, propelling global English education and research, while cultivating an international network of English language aficionados and students. As technology continues to advance, English resource libraries will further evolve, adapting to users' changing needs and providing more extensive, tailored content to meet a variety of user requirements.

Nevertheless, several obstacles confront the recommendations from English resource libraries:
(1) Information overload: With the rise of the internet age, English resource libraries have amassed a tremendous amount of information and content, resulting in information overload for users. It becomes difficult for users to pinpoint information that suits their individual needs amidst such extensive resources, which can leave them feeling inundated.

(2) Insufficient personalized recommendations: Conventional resource library recommendations often depend on basic popularity rankings or generic suggestions, failing to address the unique needs of diverse users. Given that users have varying learning objectives, interests, and professional backgrounds, personalized recommendations based on their specific traits are necessary to deliver more relevant content.

(3) Cold-start issue: For newcomers or newly introduced resources, traditional recommendation algorithms often find it challenging to provide accurate recommendations due to a lack of adequate user behavior data or resource correlation information, leading to the cold start dilemma. New users may encounter inaccurate or limited suggestions.

(4) Limited recommendation transparency: For users, the mechanisms behind recommendation algorithms are frequently unclear, lacking explanations for the suggested outcomes. Users may not understand the reasons behind specific recommendations, which can diminish their confidence in the recommendation system.

(5) Complexity of algorithms: Personalized recommendation algorithms often require extensive computations and processing, especially when dealing with a large number of users and resources, resulting in heightened algorithm complexity that can affect the real time performance and efficiency of the recommendation system. Tackling these challenges necessitates the adoption of innovative recommendation algorithms and technologies, such as machine learning based personalized suggestions, collaborative filtering, deep learning, along with more detailed user profiles and resource tagging to offer more precise, personalized, and explanatory recommendation insights.

3. K-means Algorithm and Model Establishment

3.1 K-means Algorithm

The K-means algorithm is a frequently utilized clustering technique that seeks to partition the objects in a dataset into K groups, where data points within the same group exhibit greater similarity, while those in different groups show dissimilarity. It operates as an unsupervised learning method, meaning it does not require predefined categories but instead autonomously performs clustering based on the data's characteristics [9]. The fundamental operation of K-means starts with the random selection of K initial cluster centers, which can be picked as K data points at random from the dataset or identified through alternative methods. Each data point in the dataset is subsequently allocated to the cluster whose center is nearest to it. The mean of the points within each cluster is computed, and this mean becomes the new center for that cluster [10]. This procedure is repeated until the cluster centers change only minimally, or a set number of iterations is achieved. The objective of the K-means algorithm is to minimize the total squared errors within the clusters, which signifies the total distances between data points and their respective cluster centers. This optimization is accomplished through iterative updates until a convergence condition is satisfied. Given K-means's sensitivity to the number of iterations and the selection of initial cluster centers, it is a common practice to execute the algorithm multiple times and choose the optimal outcome based on its effectiveness.

The K-means algorithm is widely used for clustering tasks, including market research, image segmentation, social network analysis, and more. While K-means is straightforward and user friendly, it is prone to being affected by the choice of initial cluster centers and may become stuck in local optima. Therefore, depending on the particular challenge, various optimization techniques or enhanced versions of the K-means algorithm might be utilized to improve clustering results.

3.2 Model Establishment for English Resource Library

The English resource database employed in this study utilises a CS hybrid framework, incorporating modules for data storage, analysis, and output within the system, which can manage relatively intricate, large amounts of user data. Likewise, the architecture of the proprietary data system is built on a CS hybrid model, facilitating seamless integration between the two systems for collaborative data collection and everyday user operations. A shared benefit of both systems is that the system administrator can directly interact with the server without intermediaries, issuing commands and retrieving necessary information. The design of the user information management platform, along with the proprietary data system, is illustrated in Figure 1.

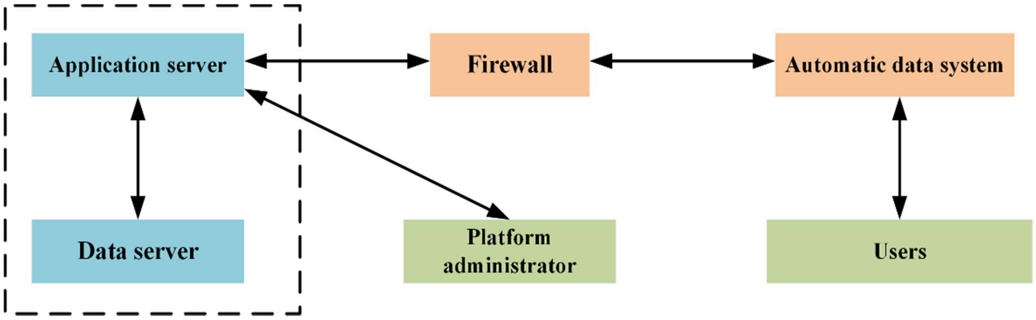

Figure 1. Structure Diagram of the User Information Management Platform and Self owned Data System in the English Resource Library

As shown in the above structure diagram (See Figure 1) of the user information management platform and the self owned data system, the application server and data server can autonomously exchange data. The application data server can be used directly by the resource library administrator or accessed by users through their own data system after passing through the firewall.

Using the professional English resource library established based on the K-means algorithm, users' proficiency levels in English are captured, and data mining is conducted using the model's built in functionality. In the construction of the model, considering the different data categories in the English resource library, the data positioning state estimation vector is given by:

$$\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_n) \neq 0 \qquad (1)$$

The diversified user English proficiency related data is transmitted to the application server, and during the data scheduling process, the characteristic vectors are denoted as:

$$y^{(k)} = \left[y_1^{(k)}, y_2^{(k)}, \cdots, y_{N_{k-1}}^{(k)} \right]^T \qquad (2)$$

$$z^{(k)} = \left[z_1^{(k)}, z_2^{(k)}, \cdots, z_{N_k}^{(k)} \right]^T \qquad (3)$$

Where and represent the system's linear horizontal and vertical inputs, respectively.

After processing, a dynamic multimodal model of the user's English proficiency index is obtained, which is based on linearity or approximate linearity. Using principal component analysis, the index information is dynamically and fractally stored in a tree like structure, establishing a linear dynamic system that captures various factors influencing English proficiency in the model. The fitting of diversified data in the model can be represented by formula (4):

$$R_\beta Y = U\left\{ E \in U / R \middle| c(E, Y) \leq 1 - \beta \right\} \qquad (4)$$

Finally, through the above dynamic fractal design, this experiment achieved the model establishment of user English proficiency data based on the K-means algorithm, and after multiple iterations, personalized recommendations can be made based on this data.

4. Effectiveness and Application of the Personalized English Resource Recommen dation System

To verify the feasibility of the K-means algorithm used in this experiment, the accuracy and recall of the system model were evaluated across different search and recommendation lists. The results obtained are plotted in Figure 2.

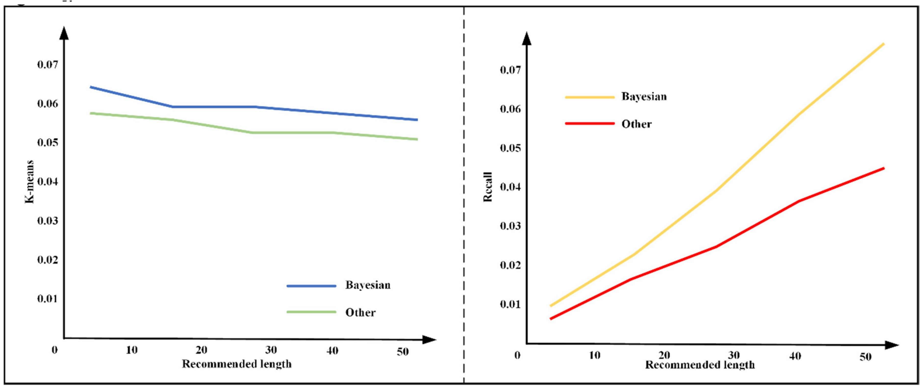

Figure 2. Comparison of Accuracy and Recall Rate

The left portion of Figure 2 illustrates the accuracy comparison. From the graph, it is evident that after implementing the K-means algorithm to filter and select features from English resources and to provide recommendations, the accuracy of those recommendations is markedly superior to that of standard collaborative filtering recommendations. The right portion of Figure 2 displays the recall rates achieved by two distinct algorithms. It is apparent that as the English resource search and recommendation lists expand, the recall rates for both algorithms progressively rise. When the recommendation list reaches 45, an acceptable recall rate is attained. At this juncture, the recall rate for English resource data search and recommendation utilizing the K-means algorithm model stands at 4.62%, while the recall rate for the non standard collaborative filtering music data search and recommendation is approximately 6.29%. This comparison suggests that the recommendation performance following filtering with the K-means algorithm model surpasses that of the non standard collaborative filtering recommendation system.

The English resource library is rich and varied, with each resource describable across three dimensions: content topic, knowledge level, and intended users. However, as the volume of resources increases and data scales expand, managing extensive multidimensional data becomes increasingly complex. To enhance resource library management and improve search and recommendation efficiency, this study employs a data dimensionality reduction approach that transforms multifaceted data into a one dimensional format. In the dimensionality reduction process, an appropriate algorithm is utilized to preserve the most critical information from the data and project it into a lower dimensional space. By reducing the dimensions, resource information can be stored and managed more effectively in one dimensional data, minimizing data storage requirements and query time complexity.

The one dimensional data not only enhances the efficiency of resource library management but also accelerates the calculation of search results and personalized recommendations. This improvement arises because one-dimensional data eliminates redundant features and accurately captures the essential characteristics and attributes of resources. Consequently, the system can more precisely align with users' needs during searches and recommendations, significantly boosting the accuracy and user satisfaction of the suggestions provided. Furthermore, the one dimensional data resulting from dimensionality reduction also helps expedite the calculation of recommendation algorithms, reducing complexity and resource consumption. This enables the recommendation system to effectively handle the increasing data scale and growing user numbers, maintaining optimal response speed and performance. Based on a phase of data collection, the outcomes of user search and personalized recommendation data derived from this research are presented in Figure 3.

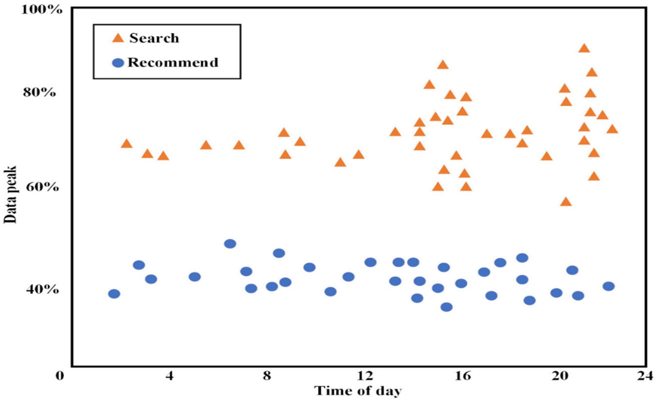

Figure 3. User Search and Personalized Recommendation Data

Figure 3 illustrates the daily average peak data for searches and tailored recommendations over a specific timeframe, revealing that the peak recommendation data for users hovers around 40% consistently throughout the day. This suggests that the English resource library system utilized in this study is stable and capable of computing and delivering personalized recommendation results while utilizing relatively minimal resources, thereby fulfilling users' requirements. Conversely, the search data peaks occur at 15:00 and 21:00, leading this research to deduce that these intervals correspond to when student users have the highest demand for classes and self study, as well as when working users seek opportunities for self learning and personal growth, resulting in these search spikes.

5. Conclusion

This paper introduces a personalized recommendation approach leveraging the K-means algorithm, specifically applied to English resource library information suggestions. The objective is to attain efficient management of the resource library while providing more precise personalized recommendations via data dimensionality reduction and clustering analysis. Initially, this article surveys the latest advancements in personalized recommendation systems and evaluates the benefits and drawbacks of various recommendation techniques and technologies. Among the numerous methods, the K-means algorithm has garnered significant attention for its simplicity and efficacy. As a clustering technique, it can categorize users and resources into distinct groups for the purpose of personalized recommendations.

Subsequently, this article elaborates on the technical foundations and operational mechanism of the K-means algorithm. By reducing the data dimensionality, the three dimensional characteristics of the resource library are transformed into a one dimensional data format, minimizing data redundancy and enhancing the efficiency of resource library management. The K-means algorithm is then employed for clustering analysis to group users and resources, resulting in the creation of personalized user profiles and resource clusters.

In the section dedicated to experimental design and analysis, we assess the efficacy of the proposed solution. By contrasting it with conventional recommendation methods, we evaluate the accuracy and performance of the recommendations. The findings indicate that the personalized recommendation solution founded on the K-means algorithm excels in recommendation accuracy and user satisfaction, effectively addressing the challenges encountered in English resource library information recommendations.

In conclusion, the personalized recommendation approach based on the K-means algorithm presented in this paper offers an efficient and precise method for English resource library information recommendations. Through the processes of data dimensionality reduction and clustering analysis, optimized management of the resource library and a quicker, more user friendly personalized recommendation experience are achieved. This solution significantly contributes to the ongoing development of the English resource library and enhances the user experience, offering valuable insights for advancing and applying information recommendation technology. While the personalized recommendation solution utilizing the K-means algorithm proves effective for English resource library information recommendations, further research and refinements are essential. Future endeavors may focus on exploring and enhancing recommendation algorithms to boost accuracy and user experience.

References

[1] Celebi, M., Emre, H. A., Kingravi, P. A., Vela. (2013). A Comparative Study of Efficient Initialization Methods for the K-Means Clustering Algorithm. *Expert Systems with Applications*, 40 (1), 200-210.

[2] Haizhou, Wang., Song, M. Ckmeans. (2011). 1d.dp:Optimal k-means Clustering in One Dimension by Dynamic Programming. *The R Journal,* 3 (2), 29-33.

[3] Mcgee, Iain. (2012). Collocation Dictionaries as Inductive Learning Resources in Data Driven Learning An Analysis and Evaluation. *International Journal of Lexicography,* 25 (3), 319-361.

[4] Huizhong, Shen, Yuan, Y., Ewing, R. (2015). English learning websites and digital resources from the perspective of Chinese university EFL practitioners. *Recall,* 27 (02), 1-21.

[5] Paul, Light., Crook, C., White, S. (2000). Learning sites: networked resources and the learning community. *Journal of Computer Assisted Learning,* 6 (1), 187-194.

[6] Biesenbach-Lucas, S. (2007). Students Writing Emails To Faculty: An Examination of E-Politeness Among Native And Non Native Speakers of English. *Language Learning Technology*, 11 (2), 59-81.

[7] Lago, A. M. F. (2016). Tourism websites in English as a source for the autonomous learning of specialized terminology. *Ibérica,* (31), 109-126.

[8] Steven, Corsello, M., et al. (2017). The Drug Repurposing Hub: a next-generation drug library and information resource. *Nature Medicine,* 23 (4), 405-408.

[9] Hyejoo, Lee., et al. (Lee). Paternal age related schizophrenia (PARS): Latent subgroups detected by k-means clustering analysis. *Schizophrenia Research*, 128 (1-3), 143-149.

[10] Cuong Manh, Do., Javidi, B. (2010). 3D Integral Imaging Reconstruction of Occluded Objects Using Independent Component Analysis based K-Means Clustering. *Journal of Display Technology,* 6 (7), 257-262.