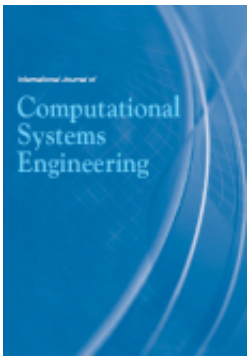
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Computational Systems Engineering

ISSN online: 2046-3405 - ISSN print: 2046-3391
https://www.inderscience.com/ijcsyse

A data mining-based approach to integrating multimedia English teaching resources

Ran Li

DOI: 10.1504/IJCSYSE.2022.10053047

Article History:

Received:	19 September 2022
Accepted:	11 November 2022
Published online:	19 March 2024

A data mining-based approach to integrating multimedia English teaching resources

Ran Li

Department of Foreign English,
Xingtai University,
Xingtai, 054001, China
Email: ranlischolar@yandex.com

Abstract: With the rapid development and popularisation of information technology, great changes have taken place in the field of education. Multimedia technology has gradually become an important means of English teaching. However, the integration of multimedia English teaching resources needs to be improved. Therefore, the D-K-means algorithm is formed by adding the splitting and aggregation operations to the clustering process of K-means algorithm to cluster the teaching resource data, and the apriori algorithm is used to find the valuable association rules in the data. Finally, the simulation experiment is carried out. The results show that the objective function value of the final solution of D-K-means algorithm is 108.64, which has stronger search ability. After combining with apriori algorithm, the accuracy of data mining can reach about 96%. In practical application, the teaching resources after the integration of this method have significantly improved students' English abilities, which show that this method can effectively tap and integrate English teaching resources, and provides a realisable path to improve the quality of English teaching.

Keywords: data mining; K-means; English language teaching resources; association rules; apriori.

Reference to this paper should be made as follows: Li, R. (2024) 'A data mining-based approach to integrating multimedia English teaching resources', *Int. J. Computational Systems Engineering*, Vol. 8, Nos. 1/2, pp.1–9.

Biographical notes: Ran Li obtained her Master's in English Language and Literature (2009) from Hebei Normal University, China. Presently, she is working as an English teacher in Xing Tai University. She is responsible for teaching American literature in foreign language department. She is also serving and served as a reviewer for national, international conferences and journals. She has published articles in more than 20 Chinese reputed peer reviewed journals and conferences proceedings. Her areas of interest include American literature, British literature, etc.

1 Introduction

With the advent of the information age, all walks of life have begun to undergo 'digital reform'. In the field of education, multimedia teaching is becoming more and more widely used and popular. Multimedia teaching tools play an important role in the teaching of listening, writing, reading, translating and speaking in English, which is mainly supported by information technology and provides students with a three-dimensional teaching experience through a combination of animation, text, sound and images (Guan et al, 2018). Khan and Ghosh (2018) explored the relationship between and student performance, and the results showed that multimedia teaching had a positive impact on students' classroom and learning performance. Therefore, education should actively promote the reform of traditional teaching methods, make full use of the advantages of network technology, and achieve information-based teaching. However, at present, there is a lack of efficient and accurate integration methods for the vast amount of English teaching resources, resulting in the advantages of multimedia teaching not being fully reflected in students and classrooms. Data mining, as a hot issue in the field of database and artificial intelligence nowadays, can reveal potential value information from a large amount of data as well as provide technical support for the integration of resources. Liu et al. (2017) aiming at the problem that a large number of teaching resource data are difficult to sort out, used data mining algorithms to explore the internal relationship of teaching resource data, and the results showed that it improved the utilisation of teaching resources. However, few research has been conducted on the use and improvement of specific algorithms for data mining, resulting in resource integration still facing problems such as low accuracy rates. Therefore, the study was conducted to improve the efficiency and quality of resource integration by optimising the K-means algorithm in data mining and combining it with the apriori algorithm after clustering.

2 Related work

In recent years, multimedia English teaching has been widely used in the classroom. The research on the integration of English teaching resources has also been paid attention to by professionals at home and abroad, with many research achievements been made on this basis. Among them, the improvement and application of K-means algorithm has a strong reference significance for the integration of teaching resources. Liu and Tsai (2021) constructed a multiword sense word vector model based on sparse soft clustering and non-negative matrix decomposition for the recognition problem of English fuzzy text, and extracted the multi-sense words by non-negative matrix decomposition. Mixed semantics were extracted and sparse soft clustering was used to partition multiple word senses into related features, which proved to be operationally effective. Bai et al. (2017) proposed a fast clustering technique to reduce the computational cost of the K-means algorithm under large-scale datasets, to enhance the scalability of the discovery density peak clustering algorithm, and to apply accelerated algorithm to compute distances containing less, and the results showed that the algorithm can effectively describe clusters of arbitrary shapes. Xiao-Yu et al. (2017) proposed a strategy to optimise the parallel K-means algorithm using Hadoop in order to solve the problem of scarce resources and high computational complexity, and introduced the canopy algorithm to initialise the clustering centres, merging clusters between canopy layers to improve the efficiency of text clustering. Jing and Wang (2017) addressed the problem of personalisation and randomness of user annotation, used maximum-minimum similarity to construct initial centres, firstly improved the traditional K-means algorithm, and proposed a new label clustering algorithm, the results of which run on MATLAB proved to improve the accuracy of clustering. Zhang et al. (2018) addressed the problem of determining the best initial seed and the most suitable number of clusters k, and improved the K-means algorithm by using depression. The inter-class distance, the average sample distance within a class and the density of the sample dataset were first calculated, and then the product of the inverse of the average inter-sample distance, the sample density and the inter-cluster distance was used as the weight product, and finally tested on the dataset in the University of California Irvine (UCI) learning repository, and the results showed that it had better clustering results (Zhang et al., 2018). Joshi et al. (2019) took into account the problem of constrained teaching resource scheduling, proposed a teaching-based optimisation algorithm that uses exams and self-study as additional features to enhance development and exploration, and used genetic algorithms for a comparative analysis of this work, with simulation results showing that it improved the efficiency of teaching resource scheduling. It can be seen that the improvement of K-means algorithm can achieve better clustering effect. The focus of the study is the integration of English teaching resources, so it can provide methodological and theoretical reference for the clustering analysis of English teaching resources.

At the same time, the current research on random forest algorithm and data mining algorithm provides a further theoretical and methodological reference for the integration of teaching resources. For determining the management between diseases and metabolites, Tie et al. (2022) combined the random forest algorithm with depth walking, and developed a new metabolite disease association prediction algorithm based on this. The results show that it has good performance in prediction. Liu et al. (2022) designed a classification framework based on random forests, which quantifies the importance of battery manufacturing features and their impact on electrode performance classification through the change of Gini coefficient, that is, predictive correlation degree. The results prove the effectiveness of this method. Lei et al. (2017) proposed an algorithm for mining seemingly unrelated useful knowledge in data, a fuzzy association rule mining algorithm that introduces fuzzy set theory into association rule mining was proposed to mine hidden knowledge by extending the affiliation of factors in the set, and simulation results showed that it can effectively suppress the loss of data at the edge of association rules. Qi et al. (2021) addressed the problem of low efficiency of teaching resource management and proposed to design intelligent management of courses using Web technology. It first collects user input data, then compiles and transfers it to the storage layer through an encoder, and then applies an information visualisation method using large-scale hierarchy to present browser data, which has been shown to have high operational speed. Chakraborty et al. (2019) in order to solve the problem of large clustering results due to initial guessing of clustering centres The K-means algorithm was combined with genetic algorithm and volume measure algorithm to predict the optimal value of the initial clustering centre, and the results showed that the algorithm improved the prediction efficiency. Guney et al. (2020) proposed a data mining method that combines clustering and association rules in order to analyse the relevance of the described resources and association rule analysis of the generated resource content, and the results validated the effectiveness of the method.

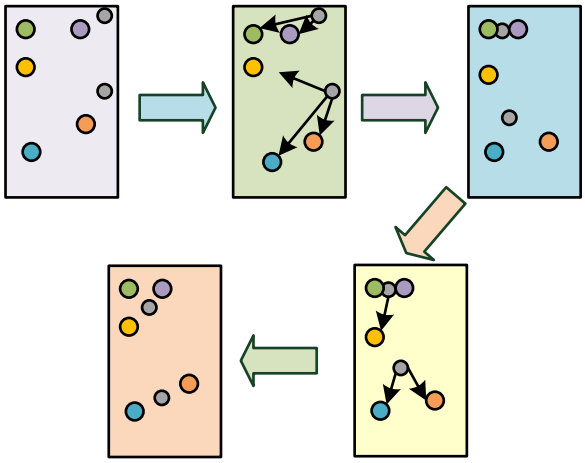
The integration of teaching resources is the focus of current research. Through the improvement of K-means algorithm, the clustering analysis of resources can be better realised. At the same time, association rules can explore the internal relationship between data resources, but there is little research on the combination of the two. Therefore, the combination of K-means algorithm and association rules is studied to better integrate multimedia English teaching resources.

3 Integration of multimedia English teaching resources based on data mining

3.1 Data mining based on improved K-means algorithm

Data mining is the process of extracting knowledge and information from incomplete, fuzzy, random and large amounts of actual data, which is not known beforehand, but which is implicitly useful. While information and data are also a form of knowledge, data mining, as a cross-discipline, can transform the application of data from low-level simple queries to extracting knowledge from data to support decision making (Gong et al., 2018). Data mining has five categories of functions: association analysis, clustering, concept description, automatic prediction of trends and behaviours, and deviation detection, which can discover meaningful implicit knowledge. K-means algorithm, as a commonly used divisional clustering algorithm, is easy to operate, simple in principle, and can maintain scalability and high execution efficiency when dealing with large datasets, and is widely used in data integration. However, with the development of data transmission and storage technology, the data stored today, including teaching resources, continues to explode, especially in the face of a certain mining target, irrelevant data is too numerous, bringing great difficulties to data collection and collation. The K-means algorithm is heavily influenced by the initial centre of mass, and the number of clusters needs to be predetermined before clustering, which leads to an increase in the amount of computing tasks and gradually becomes less effective when dealing with the current huge large datasets. Therefore, a splitting operation and an aggregation operation are added to the traditional K-means algorithm to form the D-K-means algorithm. This can increase the representativeness of the clustering results, avoid the problem of insignificant clustering effect, and thus promote the efficiency of the algorithm operation.

Figure 1 Sketch map of K-means algorithm (see online version for colours)



The K-means algorithm is an iterative solution-based cluster analysis algorithm, which first divides the selected data into K groups, and among these, K objects are randomly selected to form the initial cluster centres. On the basis of calculating the spacing between the clustering centres and the objects, the objects are assigned in close proximity (Wang et al., 2017). The assigned objects and clustering centres act as clusters, and each assignment of samples causes the existing objects in the clusters to be recalculated until the termination condition is satisfied. Typically, the termination condition is no change in the cluster centres or a local minimum of the sum of squared errors. The K-means algorithm divides the data samples based on similarity, and its Euclidean distance formula is shown in equation (1).

$$d\left(x_i, x_j\right) = \sqrt{\sum_{k-1}^{n}\left(x_{ik} - x_{jk}\right)^2} \tag{1}$$

In equation (1), x_{ik} and x_{jk} represent the samples located in the k cluster, while x_i and x_j both represent the samples located in the dataset. The mean squared deviation is the objective criterion function of the K-means algorithm, as shown in equation (2).

$$E = \sum_{j=1}^{c}\sum_{k=1}^{n_j}\left\|x_k - m_j\right\|^2 \tag{2}$$

In equation (2), E represents the sum of mean squared differences based on the data elements of the sample, x_k represents the data elements contained in the selected sample, and m_j represents the cluster centre of the j cluster. The clustering enables a set of optimal divisions to be obtained that remain compact within clusters and maximally independent between clusters, while the cluster analysis is judged on the basis of the sum of squared errors, as shown in equation (3).

$$\begin{cases} c_i = \dfrac{1}{N_i}\displaystyle\sum_{\substack{j=1 \\ x_j \in C_i}}^{n} x_j \\ J = \displaystyle\sum_{i=1}^{k}\sum_{\substack{j=1 \\ x_j \in C_i}}^{n} dis\left(x_j, c_i\right)^2 \end{cases} \tag{3}$$

In equation (3), $dis(x_j, c_i)$ represents the Euclidean distance between x_j and c_i, x_j represents the data samples in class C_i, N_i represents the number of data objects in cluster i, and c_i represents the mean value of the data objects. After the first clustering, the mean is obtained from all the samples within the cluster centre and the vector of cluster centres within all the classes is corrected, using the formula shown in equation (4).

$$z_j = \frac{\sum_{x \in s_j} x}{N_j}, \qquad j = 1, 2, 3, \ldots, k \tag{4}$$

In equation (4), N_j represents the number of samples in each group, S_j represents the clustering groups and z_j represents the cluster centres. In order to accurately determine the degree of similarity between samples within a class, the required measure of the sample is calculated. The average

distance between the cluster centres and the samples is shown in equation (5).

$$D_j = \frac{\sum_{x \in s} \left\| x - z_j \right\|}{N_j}, \quad j = 1, 2, 3, \ldots, k \tag{5}$$

In equation (5), D_j represents the average distance of each cluster centre from the selected sample. Therefore, the total mean distance is derived as shown in equation (6).

$$\bar{D} = \frac{1}{N} \sum_{i=1}^{k} \sum_{x \in s_j} \left\| x - z_j \right\| \tag{6}$$

In equation (6), $\bar{D}$ denotes the total mean distance. The results of the previous clustering are split or merged. The splitting process can yield more clustering centres, while the samples of the two classes tend to be too close together and therefore need to be merged as well. The splitting process starts by calculating the standard deviation of the data samples to the centroids, as shown in equation (7).

$$\sigma_j = \sqrt{\frac{\sum_{x \in s_j} \left(x - z_j \right)^2}{N_j}} \tag{7}$$

In equation (7), z_j is the cluster centre, N_j represents the number of samples and σ represents the standard deviation. The standard deviations of all groups are then compared together to obtain the maximum value. If the maximum value obtained is greater than the maximum standard deviation of the samples in the class, and the calculated mean distance is greater than the total mean distance, and the number of samples in the class exceeds the specified maximum, or the number of clusters is less than half of the specified number, then it is split into two groups, with the cluster centres shown in equation (8).

$$\begin{cases} z_j^- = \rho \sigma_{\max} + z_j \\ z_j^+ = z_j - \rho \sigma_{\max}, 0 < \rho < 1 \end{cases} \tag{8}$$

In equation (8), $\sigma_{\max}$ represents the maximum of all standard deviations, and z_j^+ and z_j^- are the clustering centres of the grouping. The merging operation is to compare the cluster centres between the two groups, as shown in equation (9).

$$D_{ij} = \left\| z_i - z_j \right\|, \quad i = 1, 2, 3, \ldots, k-1; \; j = i+1, \ldots, k \tag{9}$$

In equation (9), z_i and z_j are the cluster centres. The minimum value is found from all the average distances, and if it is less than the minimum value of the distance calculated between the cluster centres, it is merged and the new cluster centres are calculated as in equation (10).

$$z_i^* = \frac{N_j z_j + N_i z_i}{N_j + N_i} \tag{10}$$

In equation (10), z_i^* denotes the cluster centres of the new clusters after grouping. The cluster centre vector is weighted according to the number of samples in its class. When a pair is merged, the number of clusters is subtracted by 1. When the clustering is complete, the computation is stopped according to the stop iteration condition, and the groupings and corresponding centroid vectors are output when the computation is stopped.

3.2 Integration of English teaching resources based on apriori algorithm

There are some important rules between multimedia English teaching data. By analysing these rules, we can get the hidden relationship between teaching resource data. To further analyse the clustering results obtained from the improved K-means algorithm, association rules are introduced to mine valuable interrelationships. Association rules reflect the association and interdependence of things with other things and are an important technique for data mining. Association rules find dependencies between multiple sub-domains that satisfy both given confidence and support by describing the intrinsic relationships between data items in the selected database (Debbagh and Jones, 2017). Association rule mining is to clarify the goal of mining and prepare the data to ensure the quality of the mining goal. Apriori algorithm is one of the classical algorithms for mining relationships between data by finding association rules that meet the minimum confidence and minimum support in a given dataset, thereby filtering a large number of useless rules while ensuring that the required rules are not skipped. Apriori algorithms are easy to grasp, simple in structure and derivation, and in most cases can increase the efficiency of the algorithm and guarantee the accuracy of the results obtained. At the same time, the apriori algorithm significantly reduces the size and number of candidates to be detected, and its method of obtaining the frequency set is based on the relationship between the candidate set and the frequency set. On the basis of pruning the candidate item set, the smaller size and scale of the marquis set is obtained, which ensures the efficiency and accuracy of the algorithm results (Al-Yaseen et al., 2017). The apriori algorithm has two main steps, i.e., firstly, the full set of frequent items is obtained, and in this way the association rules are obtained. Its operational framework is shown in Figure 2.

The apriori algorithm first mines the frequent itemsets and the entire set of items with support greater than the minimum support threshold is subordinated to the frequent itemset sum. Therefore, it must ensure that the predetermined threshold is lower than the frequency of the frequent itemset. The required strong association rules are then generated, with the frequent itemsets as the basis and the support and minimum confidence as the measure required to reach the predetermined threshold. The desired rules are then generated, including the items of the set, and the concept of a medium rule is used as the base, with only one item on the right-hand side of each rule. The minimum confidence is then set as the judgement criterion and only rules that exceed the predetermined threshold range need to be retained. The apriori algorithm processes the collected data, ranks each item in the data and sets the minimum

support, and then finds all the individual items in it that do not duplicate by searching the entire data and placing these individual items into a set. The support of the individual items is then calculated and the formula is shown in equation (11).

$$support(x) = \frac{number(x \subseteq T)}{|D|} \tag{11}$$

In equation (11), *support* represents the support, D is the transaction database, and T represents the transaction, which is also a subset of the set of items. If the resulting support does not match the minimum support, the corresponding items are removed and a new set is constructed as the marquee set. According to the support formula, the items in the item set are judged in relation to the minimum support, and the items that do not match the condition are discarded to obtain the frequent item set, which is then merged to obtain the new set of options, as shown in equation (12).

$$C_k = L_{(k-1)n} \cup L_{(k-1)m} \tag{12}$$

In equation (12), C_k represents the new candidate set, k represents the k candidate set, L represents the frequent item set, n and m are the n and m sets respectively, and m is different from n. The new candidates are obtained by taking the concatenated set of the two sets through two-by-two matching, and the new candidates are traversed by the formula shown in equation (13).

$$C_j = C_k(j) \tag{13}$$

In equation (13), C_j represents a subset of the new marquee set and j represents the sequence of subsets in C_j. The size of the number of subsets is then calculated as shown in equation (14).

$$count := \begin{cases} count, & C_j \not\subset I_i \\ count + 1, & C_j \subset I_i \end{cases} \quad (j = 0, 1, 2, \ldots)\ (i = 1, 2, 3, \ldots, N) \tag{14}$$

In equation (14), N represents the sequence, i represents the first set of items, and C_j represents a subset of C_k. The initial value of is 0 and the maximum value of is jk. When an item set of $C_j \subset I_i$ is present, *count* is added to the original by 1, otherwise *count* is output depending on whether each item of C_j is included in I_i. Set the minimum confidence level, which is the probability that one item set will occur while another item set occurs, as shown in equation (15).

$$con(x \Rightarrow y) = \frac{support(x \Rightarrow y)}{support(x)} = \frac{P(xy)}{P(x)} = P(y|x) \tag{15}$$

In equation (15), x, y represents the set of items and x is not equal to y, represents the probability that the item set contains the item set *con* $(x \Rightarrow y)$ xy *support*$(x \Rightarrow y)$ represents the probability that the, dataset contains both x and y, which is $P(xy)$. $P(x)$ represents the probability that the item set x is contained in the database and represents the probability that the item set $P(y|x)$, y is contained in the item set x. The confidence level of the frequent itemset is then obtained by using the confidence formula and comparing it with the minimum confidence level to find the set of frequent items that meet the requirements. The flow of frequent itemset mining is shown in Figure 3.

Figure 2 Association rule operation frame diagram (see online version for colours)

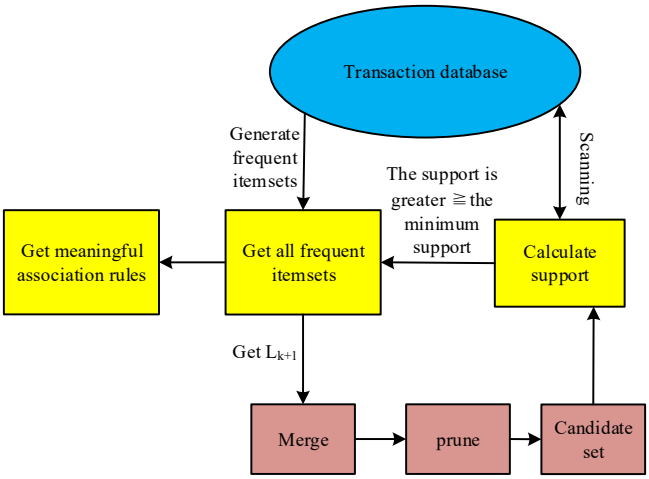
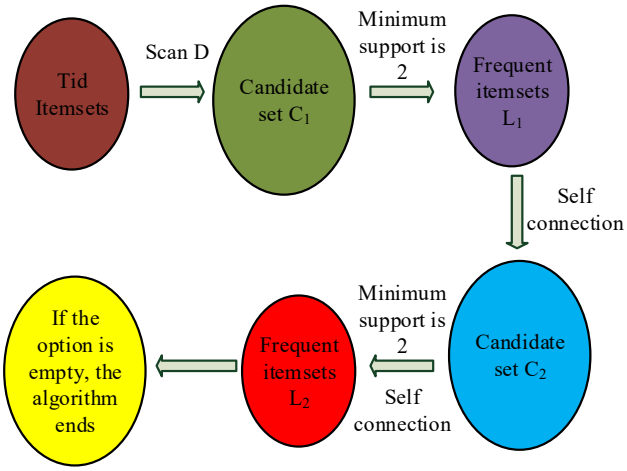


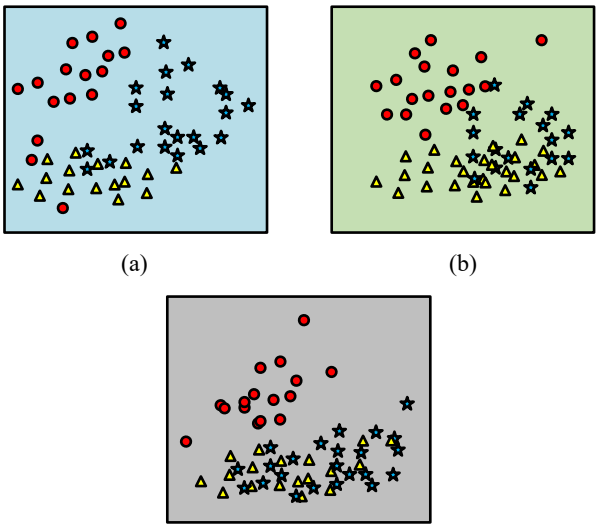
Figure 3 Flow chart of frequent itemset mining (see online version for colours)



4 Analysis of application effects

Using the improved K-means algorithm for clustering analysis of multimedia English teaching resources, the D-K-means algorithm first needs to be tested for its performance to test its clustering effect. The traditional K-means algorithm and SAKM algorithm were selected for performance testing and compared with the D-K-means algorithm proposed in the study. The clustering effects of the three in the same dataset are shown in Figure 4.

Figure 4 Clustering rendering of three algorithms, (a) K-means clustering results (b) SAKM clustering results (c) D-K-means clustering results (see online version for colours)



As can be seen from Figure 4(a), the clustering results of the traditional K-means algorithm have fewer points that intersect in different categories, basically none. In contrast, in Figure 4(b) and Figure 4(c), the point sets of both the D-K-means algorithm and the SAKM algorithm intersect, but the point sets in different categories obtained by the D-K-means algorithm intersect more than those of the SAKM algorithm, indicating that the improved algorithm has a stronger global search capability and is better able to avoid falling into the situation of local minima. The results of the SAKM algorithm and the D-K-means are shown in Figure 5.

As can be seen from Figure 5, since the magnitude of the control parameter values is gradually decreasing, then the point located in the bottom left corner of Figure 5 is the end of the change value, while the point in the bottom right corner is the starting point. the initial solution objective function obtained by the SAKM algorithm is larger than that of the D-K-means algorithm, while the objective function values of the two final solutions are 79.80 and 108.64 respectively, with the D-K-means algorithm being much larger than the SAKM algorithm is much larger than the SAKM algorithm, indicating that the improved algorithm is better at optimising the objective function. The three algorithms were then run 30 times on the same dataset and the mean, minimum and maximum values of the objective function, the average CPU time and the error with the actual centre were compared, as shown in Table 1.

Table 1 Summary of operation results of the three algorithms

Algorithm	*Average value*	*Maximum value*	*Minimum value*	*Average CPU time*	*Error from actual centre*
SAKM	126.81	182.53	101.36	13.21	8.79
K-means	176.47	205.94	136.24	0.06	16.88
D-K-means	72.12	86.25	75.44	1.26	2.36

As can be seen from Table 1, the traditional K-means algorithm requires an average running time of 0.06 s, the least running time, but is more sensitive to the initial values due to its larger range of performance functions. The SAKM algorithm has a reduced actual centre error with respect to the cluster centres, but requires a significantly higher time of 13.21 s compared to the K-means algorithm. The D-K-means algorithm, on the other hand, required 1.26 s, which was less different from the traditional K-means algorithm, but it yielded the best results with an error of only 2.36 from the actual centre. Based on the improved K-means algorithm, the apriori algorithm with association rules was added to classify and extract the teaching resource data. The data mining results of the D-K-means algorithm, the apriori algorithm and the algorithm with the combination of both are shown in Figure 6.

Figure 5 Result diagram of clustered objective function values and control parameters, (a) K-means (b) D-K-means (see online version for colours)

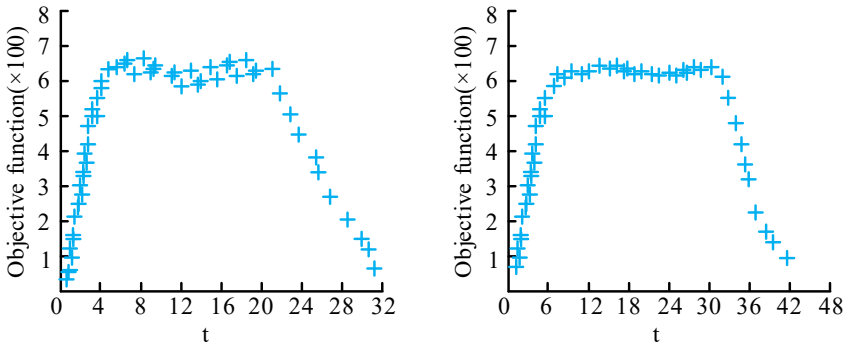


Figure 6 Comparison chart of accuracy of three algorithms (see online version for colours)

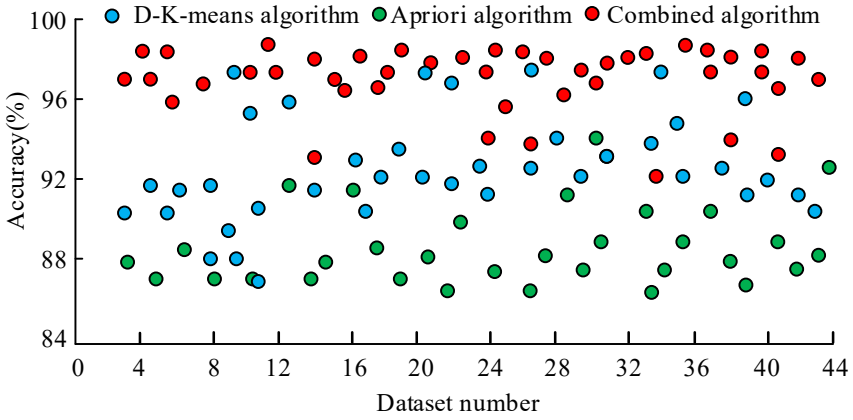
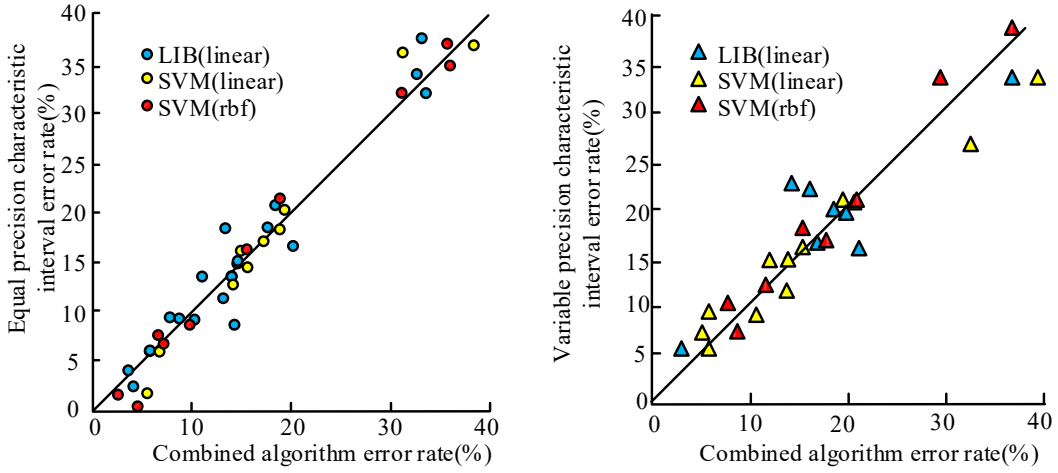


Figure 7 Classification error rate of different equivalent intervals, (a) equal precision feature interval under different classifiers (b) variable precision feature interval (see online version for colours)



From Figure 6, we can see that the apriori algorithm is able to maintain an accuracy of around 88% in clustering, and the accuracy fluctuates greatly. The D-K-means algorithm was able to maintain an accuracy of around 92% with slight fluctuations. The algorithm is then validated by example. The combined algorithm was applied to the clustering of multimedia teaching resources, i.e. the clustering analysis of English listening, reading, speaking, writing and translation teaching resources, and the experimental results are shown in Figure 7.

Figure 7 shows the classification error rate of this algorithm for multimedia ELT resources in different equal-value intervals. Figure 7(a) shows the results in the equal precision eigenvalue interval, while Figure 7(b) shows the results in the variable precision eigenvalue interval, where the numerical points represent different datasets and the distance between the numerical points and the reference line is the difference in error rate. Therefore, it can be seen from Figure 7 that the classification error rate of the D-K-means algorithm combined with the association rule is smaller, with an error rate of less than 1%, and the clustering effect is good and efficient, both in the equal-precision eigenvalue interval and in the variable-precision eigenvalue interval. Therefore, the algorithm is applied to the actual multimedia English teaching. A college English major is selected as the experimental research object, and the integrated multimedia teaching resources are used for teaching. Then, the proposed algorithm is used to explore the correlation between multimedia English teaching resources and teaching effects. The data results are shown in Table 2.

Table 2 Statistical results of the correlation between English teaching quality and multimedia English teaching resources

Category	*Listening teaching resources*	*Translation teaching resources*	*Writing teaching resources*	*Oral teaching resources*	*English reading teaching resources*
English communicative competence	0.12	0.20	0.22	0.28	0.26
Comprehensive quality of students	0.21	0.24	0.18	0.24	0.19
Classroom performance	0.23	0.20	0.20	0.21	0.18
Basic English ability	0.16	0.14	0.15	0.20	0.12

igure 8 Student achievement results before and after the experiment, (a) statistics of English classroom scores of the top 50% students (b) the scores of the rest 50% student in English class (see online version for colours)

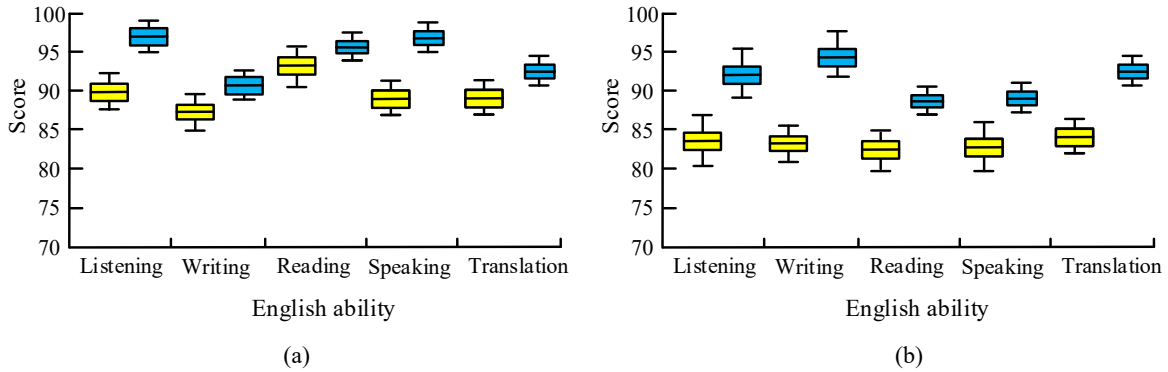


Table 2 shows that all five multimedia teaching resources – listening, writing, reading, speaking and translation – have an impact on students' professional competence, classroom performance, comprehension and English communicative competence. At the same time, the students' examination results before and after the experiment were counted to further test the impact of the integrated multimedia English teaching resources on the actual teaching effect, as shown in Figure 8.

As can be seen from Figure 8, students' scores in listening, translation, speaking, reading and writing all improved after the integration of multimedia ELT resources. The top 50% of students improved significantly, with listening and speaking improving the most, by 6% and 8% respectively. Students in the bottom 50% also improved more, with the largest gains in listening and writing at 7% and 10% respectively. The combination of the two algorithms has resulted in significant improvements in student performance and improved quality of teaching and learning.

5 Conclusions

Multimedia technology is an indispensable auxiliary tool in current English teaching. Aiming at the integration of multimedia English teaching resources, this paper proposes a data mining method based on improved K-means algorithm, which is combined with apriori algorithm for resource integration. The experimental results show that compared with SAKM algorithm and traditional K-means algorithm, D-K-means algorithm can better optimise the objective function in the same dataset. In the data mining results, the accuracy of apriori algorithm is about 88%, but the fluctuation is large. The D-K-means algorithm is about 92%. The combined accuracy of the two algorithms can reach 96% or more, and the floating is small. It has strong stability and high accuracy. In the cluster analysis of teaching resources, the error rate of the combined algorithm is less than 1%, and the clustering effect and efficiency are good. After using this method to integrate resources, students at all stages have improved in the five core English abilities of writing, speaking, reading, listening and translation. It shows that this method can effectively and accurately mine the correlation and laws between multimedia English teaching data, and effectively classify the teaching data, thus realising the further integration of English teaching resources. However, the research has not improved the apriori algorithm, so it needs to be further explored to get a better integration effect of teaching resources.

References

Al-Yaseen, W.L., Othman, Z.A. and Nazri, M. (2017) 'Multi-level hybrid support vector machine and extreme learning machine based on modified K-means for intrusion detection system', *Expert Systems with Applications*, Vol. 67, pp.296–303.

Bai, L., Cheng, X., Liang, J., Shen, H. and Guo, Y. (2017) 'Fast density clustering strategies based on the k-means algorithm', *Pattern Recognition*, Vol. 71, pp.375–386.

Chakraborty, S., Raj, S. and Garg, S. (2019) 'Selection of 'K' in K-means clustering using GA and VMA', *International Journal of Data Science*, Vol. 4, No. 1, pp.63–78.

Debbagh, M. and Jones, W.M. (2017) 'Examining English language teachers' TPACK in oral communication skills teaching', *Journal of Educational Multimedia & Hypermedia*, Vol. 27, No. 1, pp.43–62.

Gong, X., Wang, Z. and Wang, L. (2018) 'Research on financial early warning model for papermaking enterprise based on particle swarm K-means algorithm', *Paper Asia*, Vol. 34, No. 6, pp.41–45.

Guan, N., Song, J. and Li, D. (2018) 'On the advantages of computer multimedia-aided English teaching', *Procedia Computer Science*, No. 131, pp.727–732.

Guney, S., Peker, S. and Turhan, C. (2020) 'A combined approach for customer profiling in video on demand services using clustering and association rule mining', *IEEE Access*, Vol. 8, No. 99, pp.6–10.

Jing, Y. and Wang, J. (2017) 'Tag clustering algorithm LMMSK: improved K-means algorithm based on latent semantic analysis', *Journal of Systems Engineering and Electronics*, Vol. 28, No. 2, pp.374–384.

Joshi, D., Mittal, M.L., Sharma, M.K. and Kumar, M. (2019) 'An effective teaching-learning-based optimization algorithm for the multi-skill resource-constrained project scheduling problem', *Journal of Modelling in Management*, Vol. 14, No. 4, pp.1064–1087.

Khan, A. and Ghosh, S.K. (2018) 'Data mining based analysis to explore the effect of teaching on student performance', *Education & Information Technologies*, Vol. 23, No. 4, pp.1–21.

Lei, Y.R., Lei, L. and Liu, L.Q. (2017) 'Application of fuzzy association rules in the analysis on higher vocational college students' performance', *Journal of Computer*, Vol. 28, No. 1, pp.1–12.

Liu, K., Hu, X. and Zhou, H. (2022) 'Feature analyses and modeling of lithium-ion battery manufacturing based on random forest classification', *IEEE/ASME Transactions on Mechatronics*, Vol. 26, No. 6, pp.2944–2955.

Liu, L. and Tsai, S.B. (2021) 'Intelligent recognition and teaching of English fuzzy texts based on fuzzy computing and big data', *Wireless Communications and Mobile Computing*, Vol. 2021, No. 1, pp.1–10.

Liu, L.D., Zhou, R.Q., Yi-Lin, W.U. (2017) 'Optimal allocation strategy for teaching resources based on classification data mining', *Modern Computer*, No. 2, pp.5–9.

Qi, S., Li, S. and Zhang, J. (2021) 'Designing a teaching assistant system for physical education using web technology', *Mobile Information Systems*, Vol. 2021, No. 6, pp.1–11.

Tie, J., Lei, X. and Pan, Y. (2022) 'Metabolite-disease association prediction algorithm combining deepwalk and random forest', *Tsinghua Science and Technology*, Vol. 27, No. 1, pp.58–67.

Wang, L., Lu, Z.M., Ma, L.H. and Feng, Y.P. (2017) 'VQ codebook design using modified K-means algorithm with feature classification and grouping based initialization', *Multimedia Tools and Applications*, Vol. 77, No. 10, pp.1–16.

Xiao-Yu, L.I., Li-Ying, Y.U., Lei, H. and Tang, X. (2017) 'The parallel implementation and application of an improved k-means algorithm', *Journal of University of Electronic Science and Technology of China*, Vol. 46, No. 1, pp.61–68.

Zhang, G., Zhang, C. and Zhang, H. (2018) 'Improved k-means algorithm based on density canopy', *Knowledge-Based Systems*, pp.289–297.