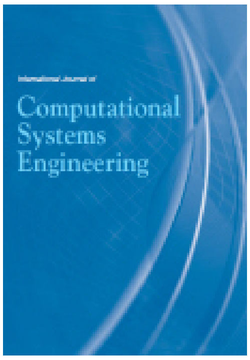
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Computational Systems Engineering

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

Analysis of word vector combined with group intelligence perception based on STEM concept for ELT word recommendation strategy

Ke Qin, Yang Zhou, Daniel Kvitrud

DOI: 10.1504/IJCSYSE.2024.10056267

Article History:

Received:	09 August 2022
Last revised:	01 November 2022
Accepted:	27 December 2022
Published online:	19 March 2024

Analysis of word vector combined with group intelligence perception based on STEM concept for ELT word recommendation strategy

Ke Qin*

School of Foreign Languages,
Guizhou University of Finance and Economics,
Guiyang, 550000, China
Email: keqingufe@163.com
*Corresponding author

Yang Zhou

Guizhou University of Finance and Economics,
Guiyang, 550000, China
Email: zy19851126@126.com

Daniel Kvitrud

Athletic Department,
Saint John's University,
Collegeville, 56321, USA
Email: dankvitrud@gmail.com

Abstract: With the rapid development of China's education concept, the mode of English learning has produced great changes, and the learning methods of ELT words have gradually developed towards intelligence. For the problem of ELT word recommendation strategy, an optimised CWSAR algorithm model based on STEM concept, combining collaborative filtering algorithm, word vector semantic perception, context perception and crowdsensing is proposed. The optimal parameter experiments and comparison experiments are conducted for this algorithm model. The experimental results show that CWSAR algorithm is better than the two CF algorithms in the experiment of ordinary data sets. The CWSAR algorithm can better complete the work of English teaching word recommendation and provide effective help for English teaching.

Keywords: word recommendation; collaborative filtering; word vector; group wisdom perception.

Reference to this paper should be made as follows: Qin, K., Zhou, Y. and Kvitrud, D. (2024) 'Analysis of word vector combined with group intelligence perception based on STEM concept for ELT word recommendation strategy', *Int. J. Computational Systems Engineering*, Vol. 8, Nos. 1/2, pp.30–39.

Biographical notes: Ke Qin received her Bachelor of Arts from Chongqing Jiaotong University, Chongqing, China in 2008. She received her Master's in ESP from Chongqing University, Chongqing, China in 2011. Presently, she is working as a Lecturer in the School of Foreign Language, Guizhou University of Finance and Economics, China. Her areas of interest are English teaching, corpus linguistics and word embedding.

Yang Zhou obtained his BSc from Guizhou University, Guiyang, China in 2008. He received his Master's in HYPERLINK 'javascript'; Food Science from Guizhou University, Guiyang, China in 2012. Presently, he is working in the Library, Guizhou University of Finance and Economics, China. His area of interest is library information technology.

Daniel Kvitrud is a Physical Education Teacher and Strength and Conditioning Coach. Working across grade levels, he delivers instruction on creating and maintaining healthy lifestyles and physical activity. As the MSHSL Varsity Sports Official in Minnesota State High School League, he directed thousands of Minnesota high school athletes in competitions with knowledge of application of rules in specific sports. He also provides crisis mediation and solution during intense periods of contests.

1 Introduction

As educational reform continue to advance, the value of the STEM education concept has begun to emerge, and it has had a positive impact on the development of education in various subjects (Ortiz-Revilla et al., 2022). As a basic language subject, English, in the process of combining with the STEM concept, not only develops itself in the direction of diversification and multidisciplinary integration, but also acts as a language carrier for other subjects (Ozkaya et al., 2022). Therefore, ELT research is particularly important. Words are the basis of English teaching, thus the research on ELT word learning has become a key topic in related industries (Chang et al., 2021). With the popularity of smart phones, the method of ELT word learning tends to use these tools. Improving the vocabulary recommendation ability of ELT word software is an important means to improve the efficiency of English learning (Wang, 2021). At present, all kinds of ELT software lack research on teaching word recommendation strategies and comprehensive analysis of multiple dimensional data related to words themselves and users' learning habits (Zhang et al., 2021). Therefore, this study explores the research on ELT word recommendation strategies. The difficulty of the study is that the current methods of recommending educational resources are difficult to be compatible with the more advanced stem educational concepts, so that the existing methods of recommending English teaching words cannot match the current educational needs. Therefore, this research work has great application significance.

2 Related work

Yan et al. (2022) found that data sparsity reduces the accuracy and diversity of collaborative filtering recommendation algorithms, and proposed a coverage coarse-grained computational model for collaborative filtering recommendation algorithm optimisation. MovieLens dataset was used to test the model against six classical algorithms, and the results showed that the optimised algorithm not only enhanced robustness with the same time complexity, but also had significantly higher recommendation diversity and accuracy. Joorabloo et al. (2020) and his experimental partners proposed a model that takes into account the future of user-item similarity trends and reordering their neighbourhood sets, which can be applied to a wide range of CF methods based on similarity between users and items. The method was applied to a set of CF algorithms on two benchmark datasets and it significantly improved the performance of the original CF recommender. Zhang et al. (2020) summarised the key techniques of collaborative filtering recommendation algorithms and its development difficulties. He analysed the problems of different techniques, looked forward to the application prospects of collaborative filtering technology, and concluded that collaborative filtering algorithms can adapt to most recommendation requirements. Liu et al. (2021) proposed a hybrid news recommendation algorithm combining content-based recommendation algorithm and collaborative filtering. TF-IDF method and K-means clustering techniques were used to extract and process features of news content, while singular value decomposition technique was applied to solve the matrix sparsity problem in traditional collaborative filtering algorithms. Wu et al. (2020) believed that crisp scores could not effectively measure the uncertainty of user preferences, and designed a collaborative filtering recommendation algorithm based on interval-valued fuzzy numbers. The experiment proved that compared with other fuzzy and traditional algorithms this algorithm can effectively improve the prediction accuracy and rank accuracy with less time overhead.

Lu et al. (2019) and other researchers proposed a new user-based collaborative filtering recommendation method combined with privacy concern strength, and experiments showed that the method had more advantages than other algorithms to reduce the impact of users' privacy concerns on mobile personalised services. Yuan et al. (2021) and his team addressed the problem that a single model of traditional recommendation systems cannot accurately capture user preferences. They proposed a hybrid movie recommendation system and optimisation method based on weighted classification and user collaborative filtering algorithms. Experiments based on the Douban movie dataset verified that the algorithm solved the shortcomings of the single algorithm model to a certain extent and improved the recommendation effect. Chen et al. (2021) proposed a CF recommendation system (KDPCF) based on k-means clustering and users' information to solve the existing differential private CF recommendation system. KDPCF first classifies the dataset by k-means clustering and appropriately adjusts the size of the target category to which the target user belongs in order to use only the users in the appropriately sized target category for recommendation. The experimental results showed that KDPCF has significant performance improvement over existing systems. Cai et al. (2022) proposed friend-aware graph collaborative filtering (FG-CF), a framework that merges social information into the user POI graph. Moreover, they proposed a new message aggregation function to update user and POI embeddings. Extensive experiments on two large-scale LBSN datasets proved that the model outperforms several advanced approaches. Tang et al. (2021) proposed a dynamically evolving multigraph collaborative filtering (DMGCF) model to mine and reuse edge information constructing user and item graphs based on user-item bipartite graphs and embeddings to exploit inter-user and inter-item relationships. It is a new dynamic evolutionary mechanism that collaboratively updates and enhances embeddings and graphs during the learning process to produce better embeddings, user-item relationships, and rating scores. A series of experiments were conducted on real datasets, and the experimental results demonstrated the effectiveness of the method. Li and Zhou (2020) and his team used online digital movie recommendation as an example, and used a combination of theoretical analysis and experiments to illustrate the collaborative filtering recommendation algorithm to achieve the principle and process of personalised recommendation.

Through the in-depth research on the application of recommendation algorithms by many domestic and foreign researchers, it can be seen that many scholars have selected collaborative filtering algorithms when recommending information resources. Collaborative filtering algorithms can also be used in the selection of English teaching words recommendation strategies. Therefore, this research is based on collaborative filtering algorithm, fully considering the particularity of English learning word recommendation, and specifically combining word vector and swarm intelligence perception to match and improve the content of the recommendation algorithm. It is expected to play a certain role in promoting the research of English teaching word recommendation methods.

3 A study of word vector combined with group intelligence perception for ELT word recommendation strategy based on STEM concept

3.1 Research on context-aware ELT word recommendation algorithm

With the development of technology, the methods and concepts of subject education are gradually changing, and the learning of English words has also changed from the once rote memorisation mode to the intelligent associative learning mode nowadays, which also puts forward new requirements for the recommendation strategies of ELT words (Kang, 2020). The core idea of contextual word sense awareness recommendation (CWSAR) algorithm is to introduce users' past usage habits, sense users' contextual information, and use this information to recommend ELT words of higher interest to users.

Based on the idea of collaborative filtering algorithm, the similarity of users is used as the core condition for recommendation, and the similarity calculation of users consists of two modules, which are context-aware module and interest-aware module. Combining the two modules, the similarity of users is calculated and used as the basis to match the similar neighbours. Finally, the user's recommendation list is generated based on the calculated recommendation list of all neighbouring users, and the final recommendation result is obtained. The interest perception module relies on a combination of the user's word browsing records and semantic perception, where semantic perception is the key unit. The matching of English words needs to be done using semantics to accurately match the user's interest. The traditional matching mechanism often uses the similarity of letters for matching, which leads to very poor correlation between words. Therefore, the matching of English words should be combined with semantics and constructed with the help of word vector methods to accurately match words of higher interest based on users' English teaching word browsing records. The Euclidean distance method between feature vectors is used to define the interest relevance sim_{ia} of users, which is calculated as shown in equation (1).

$$sim_{ia}(a, b) = 1 - \sqrt[2]{\frac{\sum_{i=1}^{k} (f_a)_i^2 - (f_b)_i^2}{k}} \tag{1}$$

In equation (1), a, b denotes user a and user b respectively. $f_a \in R^k$ denotes the feature vector of user a, $f_b \in R^k$ denotes the feature vector of user b, and k denotes the number of ELT words in users' browsing records. When calculating users' interest similarity, the similarity value of each component is calculated, and the loop is traversed, and then all words with high similarity are updated. Two problems arise in the finding process, which are too few ELT word browsing records of the target users, and too few ELT word browsing records of the matched users. If the matched user has too few ELT word browsing records, the match is dropped and a new similar user is found. If the target user has too few ELT word browsing records, the calculation will ignore the interest similarity and omit the search step, and the user context-awareness module will prevail. The context-aware module will get some unrestricted information of users, such as age, gender and location area, with the purpose of determining their education level, learning stage and other information for context matching and getting more suitable recommendations. According to the current education situation in China, the priority is to match age similarity, and learners in the same age stage have mostly the same learning goals. The age similarity sim_{ag} is calculated as shown in equation (2).

$$sim_{ag} = 1 - \frac{|A_a - A_b|}{\min(A_a, A_b)} \tag{2}$$

In equation (2), A_a, A_b represents the age of user a and b, respectively. The minimum value is used as the denominator in equation (2) because English learning exhibits a relatively obvious incremental pattern, often progressing slowly and acquiring less knowledge in the early stages of learning. As the age increases, the amount of learning increases and a huge leap is made across the plateau period. Therefore, the minimum value is used as the denominator to reflect the maximum effect of age. Considering that the effect of value $|A_a - A_b|$ is minimal when both users are older, the influence is corrected by adding the age window W_{ag}, where users with the same window are at the same learning stage and users without the window are not at the same learning stage, at which point the age similarity is calculated as shown in equation (3).

$$sim_{Wa} = 1 - \frac{\min(|A_a - A_b|, W_{ag})}{W_{ag}} \tag{3}$$

Regional differences also have a large impact on English learning, as educational resources and educational attainment vary from region to region. Therefore, the regional similarity sim_{re} is introduced to calculate it, which is shown in equation (4).

$$sim_{re} = 1 - \frac{\min\left(dis(a,b), W_{dis}\right)}{W_{dis}} \tag{4}$$

In equation (4), W_{dis} indicates the region window, and the users in the same region window are considered in the same learning stage. $dis(a, b)$ indicates the distance of the region location difference between user a and b. The region similarity is actually calculated by constructing a circle of radius W_{dis} with the region where the user a is located as the centre, and the region similarity is the distance from the centre of the circle. When W_{dis} is the maximum region location difference distance, the region similarity is calculated for all users, and vice versa for only some users. There are certain problems and discrepancies with the facts, whether using age or region alone as the key information of the context. Therefore, to calculate the user's contextual similarity sim_{cs} needs to be combined the age similarity sim_{ag} and region similarity sim_{re}, which is calculated as shown in equation (5).

$$sim_{cs} = \alpha \cdot sim_{ag} + \beta \cdot sim_{re} \tag{5}$$

In equation (5), α represents the weight of user's age similarity sim_{ag}, β represents the weight of user's region similarity sim_{re}, and the sum of α and β is 1. By adjusting the values of α and β, the weight of age and region influence in user's contextual similarity can be adjusted. The user similarity sim_{us} is calculated based on the combination of the user's interest relevance sim_{ia} and the user's contextual similarity sim_{cs}, which is shown in equation (6).

$$sim_{us} = \lambda \cdot sim_{ia} + \gamma \cdot sim_{cs} \tag{6}$$

In equation (6), λ represents the weight of interest relevance sim_{ia} and γ represents the weight of context similarity sim_{cs}, the sum of them should also equal to 1. Similarly, adjusting the proportion of the two weights can adjust the degree of influence of the corresponding parameters. After calculating the user similarity, the corresponding neighbours are matched according to the similarity. In this case, a user similarity threshold ρ is introduced, and when the user similarity exceeds the target threshold, the corresponding user is added to the neighbour list. After matching the neighbours, the ELT words that match the suitability are selected and added to the recommendation list based on the browsing records of each neighbour. The selection conditions of recommended ELT words include the rating mechanism of ELT words as the ranking basis and screening mechanism for word recommendation to users, and the related calculation method is shown in equation (7).

$$sim_w(y) = \max\left\{s \,\middle|\, s = sim_{ss}(x, y), x \in W_d\right\} \tag{7}$$

In equation (7), $sim_w(y)$ denotes the semantic similarity of the recommended words, W_d denotes the browsing records of the user d. y denotes the recommended ELT words corresponding to the target neighbour c, and $sim_{ss}(x, y)$ denotes the evaluation function of semantic similarity. Based on these, a common evaluation system can be derived as shown in equation (8).

$$score_1 = sim_w\left(w_i\right) \tag{8}$$

The evaluation system shown in equation (8) directly adopts semantic similarity as the final result, and there will be the problem that words with low similarity but high repetition rate keep failing to appear in the recommended list, and the list is always occupied by words with higher similarity. To solve this problem, the number of occurrences of words is introduced into the evaluation system to constitute a new evaluation system, as shown in equation (9).

$$score_2 = D_i \cdot sim_w\left(w_i\right) \tag{9}$$

In equation (9), D_i indicates the number of times the teaching vocabulary w_i appears in the recommendation list of neighbouring users. The evaluation system with equation (9) can make more reasonable recommendations. The combination of context aware content into the recommendation algorithm will help users to fit more closely with social groups and achieve social expectations.

3.2 Word vectors combined with collaborative filtering and group wise perception

The processing of ELT words needs to be performed using machine language. Traditional vocabulary processing is often divided by alphabetic characters without considering the word meanings of ELT words. To solve this problem, a word vector module is introduced to process ELT words. Firstly, latent semantic analysis (LSA) is used to process the ELT words, and the words are added to the vector space of the word mapping together with their corresponding paraphrase information documents. Combined with latent Dirichlet allocation (LDA), the document-to-topic and topic-to-word probability distributions of probabilistic LSA are first validated before the distribution, changing the process into a Bayesian form, as shown in Figure 2.

As shown in Figure 2, in the ELT word vector construction process, a random sample representing the topic distribution θ of the specified document is made according to the LDA's $Dir(\alpha)$, then a topic H is selected according to θ, followed by a random sample word distribution δ from another $Dir(\beta)$, from which the word W is selected by combining the two. Since the processing of the word vector model takes up a lot of space, which leads to slow or obstructed operation of the algorithm model. To address this problem, neural network language model (NNLM) is introduced, in which each vocabulary corresponds to a separate feature vector. A model is simulated that can obtain the combination probability from the input word sequence, which is to be in a continuous smoothing state. In this way, the weights of the word vectors and the parameters that appear in the model are continuously learned for the vocabulary to reach optimisation, and the structure of the NNLM is shown in Figure 3.

Figure 1 CWSAR algorithm framework (see online version for colours)

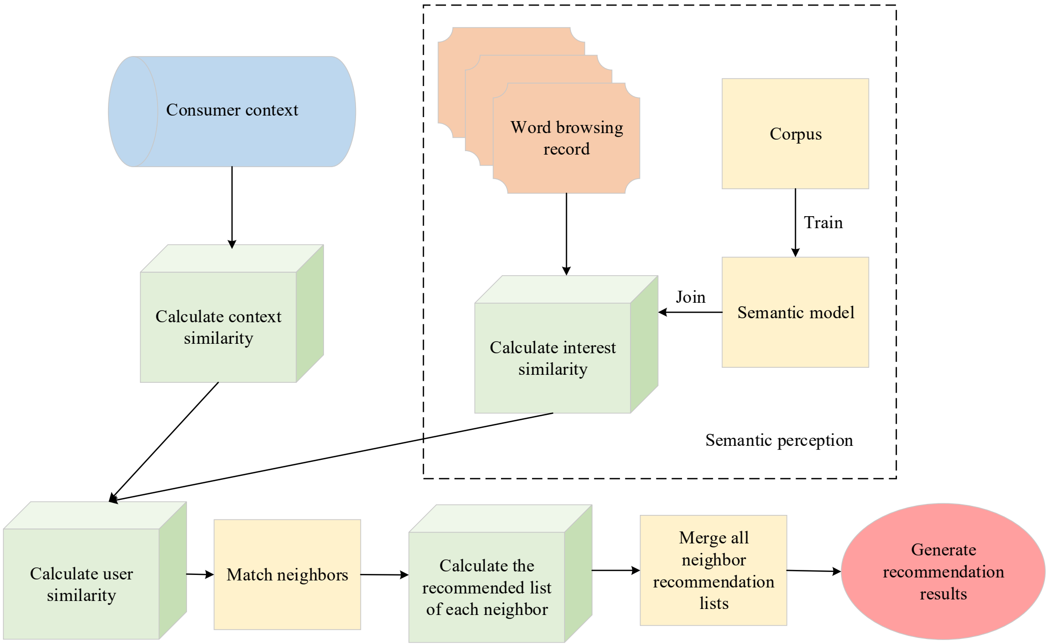


Figure 2 Schematic diagram of LDA model (see online version for colours)

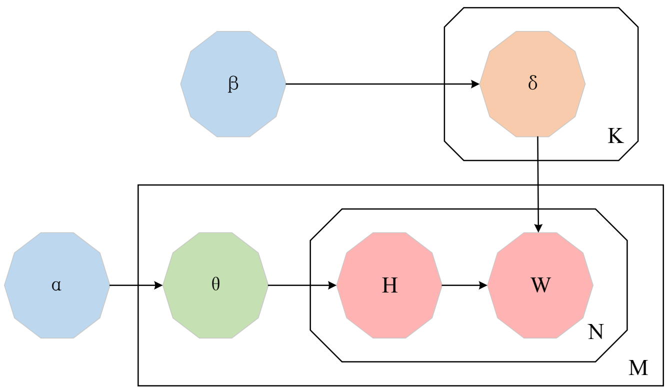


Figure 3 NNLM structure diagram (see online version for colours)

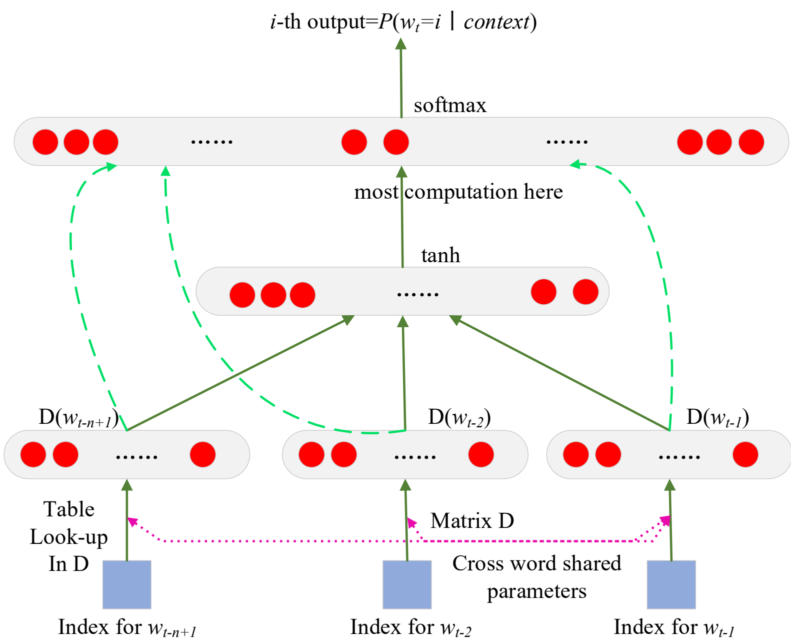


Figure 4 Schematic diagram of CBOW and skip-gram (see online version for colours)

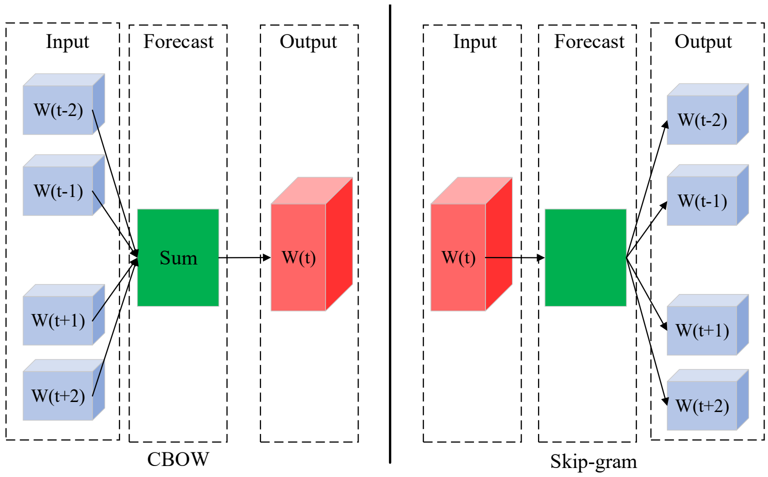


Table 1 Crowdsensing data list

Element	*Explain*	*Data*	*Example*
Character	Users participating in the behaviour	User unique ID	XXXXXXJFAHTWID856DXXX
Time	When the behaviour occurred	The system automatically obtains the time	2022-03-01 18:12:10
Place	Place where the behaviour occurred	User IP and location information	192.168.133.25
Way	How behaviour occurs	Use environment	Iphone13, ios15.0.2
Behaviour	Content of behaviour	Behaviour attribute	Goods_name: books

The NNLM model contains a large number of parameters to learn, and it needs to be simplified by adapting the word vector neural network using continuous bag-of-words (CBOW) model and skip-gram (Bangyal et al., 2021). CBOW acts on context-awareness to predict current ELT words, and skip-gram is the inverse model of CBOW that acts on current ELT words to predict the context, as shown in Figure 4.

The combination of CBOW and skip-gram can effectively optimise the NNLM model, speed up the word vector construction, and improve its accuracy. The construction of word vector helps to strengthen the machine's understanding of English words and avoid the occurrence of mismatched recommended words. After the word vector module is built, it needs to be combined with collaborative filtering and group wisdom perception. Collaborative filtering is different from content recommendation, and is based on the information of items (Qasim et al., 2022). By default similar users have similar rating system for items, so to accurately predict the rating s_a of the user a for the item k, the neighbouring users j who are similar to the target user have to be found. The average of their ratings for the item s_v, k is our predicted rating s_a, and the predicted rating is calculated as shown in equation (10).

$$s_a = \frac{1}{|N_i(a)|} \sum_{v \in N_i(a)} s_v \qquad (10)$$

Because there are multiple neighbours of the target user and the influence varies among neighbours, it is necessary to add weight values to the neighbours' scores η to obtain a more accurate weighted average as the predicted score, which is calculated as shown in equation (11).

$$s_a{}' = \frac{\sum_{v \in N_i(a)} \eta s_v}{\sum_{v \in N_i(a)} |\eta|} \qquad (11)$$

Another situation occurs when referring to the ratings of neighbouring users: different ratings are generated for the same item. It is then necessary to set aside the ideal conditions and normalise the user's ratings to derive the user's true rating R_{sc}, at which point the user's predicted rating is calculated as shown in equation (12).

$$s_a'' = h^{-1}\left(\frac{\sum_{v \in N_i(a)} \eta R_{sc}}{\sum_{v \in N_i(a)} |\eta|}\right) \qquad (12)$$

By adding the potential factors to the user rating system, adding the vector o_i to represent the potential factors possessed by the item i, and adding the vector p_a to represent the level of interest of the user a in the corresponding factor, the user rating of the item s_i is calculated as shown in equation (13).

$$s_i = \mu + b_a + b_i + o_i{}^T q_a \qquad (13)$$

In equation (13), μ represents the average rating. b_a and b_i represent the respective influence of users and items on the rating, and $o_i{}^T q_a$ represents the interaction between users and items. The user rating system is used to establish a comprehensive factor evaluation of the items, and the item ratings are obtained as shown in equation (14).

$$s_i' = \mu + b_a + b_i + o_i^T \left(q_a + |R(a)|^{\frac{1}{2}} \sum_{j \in R(a)} y_j \right) \quad (14)$$

In equation (14), $R(a)$ denotes all items rated by the user a, and the items are analysed by combining the composite factors with the time function to obtain the final rating system. As shown in equation (15), where t_{ai} denotes the time influence factor.

$$\begin{aligned} ts_i' &= \mu + b_a(t_{ai}) + b_i(t_{ai}) \\ &+ o_i^T \left(q_a(t_{ai}) + |R(a)|^{\frac{1}{2}} \sum_{j \in R(a)} y_j \right) \end{aligned} \quad (15)$$

Crowdsensing is a new interaction model that relies on mobile devices to collect data and utilise it (Yang et al., 2020). Crowdsensing can help to accelerate the understanding of users and their neighbours and provide recommendations more accurately. Crowdsensing mainly records some non-sensitive data of users with the aim of understanding their behavioural habits, and the specific data list is shown in Table 1.

Combining these attributes of users can speed up the analysis of users' habits and hobbies, so as to determine attributes such as users' interests and learning stages, and improve the accuracy of recommended English teaching words (Haider Bangyal et al., 2022). At the same time, the crowdsensing analysis helps the algorithm to continuously learn more user characteristics and accelerate the recommendation rate.

4 Experiments and analysis of the CWSAR algorithm model

The experiments were conducted in the same experimental environment, and the comparative analysis of the algorithmic models was carried out in Python under Ubuntu operating system. The Frappe dataset and a self-built ETW dataset of ELT words were used for the experiments. The ratio of training set, test set and validation set in the experiment is 7:2:1. CWSAR is compared with UCF, ICF, context-aware semantic aware recommendation (SAR) and semantic-aware context recommendation (CR) algorithm models on both datasets. In the comparison, the interest similarity weight λ, age similarity weight α, and user similarity threshold ρ present in each algorithm model are taken as 0.5. The Precision@N and Recall@N values of the five algorithms are compared using the Top-N recommendation task, and the Precision@N and Recall@N values of each model under the two datasets are recorded ten times, and finally the average value is taken for recording. The performance of each algorithm model on the Frappe dataset is shown in Table 2.

As can be seen from Table 2, CWSAR performs the best in the Frappe dataset experiments, even though this dataset is not an English word dataset, indicating that the optimisation effect of combining crowdsensing and context perception in CWSAR is obvious. The difference between ICF, UCF, CR and SAR is not significant, due to the fact that each algorithm weight is adjusted to 0.5 without special bias and the content of this dataset is more average. The poor performance of the SAR algorithm model is due to the fact that the semantic perception module does not have room to play under this data. The performance of each algorithmic model on the ELT word ETW dataset continues to be compared, as shown in Table 3.

Table 2 Experimental results on dataset Frappe

Algorithm model	*Precision@5*	*Precision@10*	*Recall@5*	*Recall@10*
CWSAR	0.842	0.673	0.263	0.356
UCF	0.718	0.632	0.231	0.273
ICF	0.705	0.618	0.234	0.258
CR	0.715	0.628	0.233	0.264
SAR	0.681	0.606	0.225	0.248

Table 3 Experimental results on dataset ETW

Algorithm model	*Precision@5*	*Precision@10*	*Recall@5*	*Recall@10*
CWSAR	0.875	0.718	0.286	0.374
UCF	0.523	0.416	0.159	0.184
ICF	0.515	0.409	0.151	0.176
CR	0.613	0.541	0.183	0.211
SAR	0.776	0.675	0.252	0.282

Comparing Table 2 with Table 3, it can be seen that CWSAR performs the best in the experiment on the ETW dataset and outperforms itself on the Frappe dataset. This indicates that the word vector semantic perception module plays a role in English vocabulary recommendation, and similarly the larger improvement of the SAR algorithm also indicates that the semantic module plays a role in English word recommendation. The extremely poor performance of UCF and ICF in ELT word recommendation indicates that accurate recommendation cannot be achieved without semantic understanding of English words, and the deteriorating performance of the CR algorithm model also reflects this. Combining the experimental results of the two data sets, it can be concluded that CWSAR is better than the two CF algorithms in the experiment of ordinary data sets. Its Precision@N value increases by 13.4%, and Recall@N increases by 24.3% in value; CWSAR is better than two CF algorithms in ETW dataset experiment of English teaching words. Its Precision@N increases by 71.0% and Recall@N value increases by 40.2%. Keeping the age similarity weight α and the user similarity threshold ρ unchanged, the performance of the CWSAR algorithm model on the ETW dataset is compared for different values of the interest similarity weight λ, and the experimental results are shown in Figure 5.

As can be seen from Figure 5, keeping other parameters constant, Precision@N value, Recall@N value and F1@N value all increase with the increase of interest similarity weight λ until it reaches the maximum value at $\lambda = 0.8$, and then starts to decrease at a faster rate. The growth rate is slower when λ is in the interval of 0.05~0.25 and 0.5~0.8, and increases significantly when it is in the interval of 0.3~0.5. The experimental results indicate that interest similarity has a greater impact on the recommendation accuracy of the algorithm compared to contextual association similarity, and the accuracy of the algorithm decreases significantly when the interest similarity weight is lower than 0.3 or higher than 0.9, and reaches the optimal effect when the interest similarity weight is 0.8. Keeping the interest similarity weight λ and user similarity threshold ρ unchanged, the performance of the CWSAR algorithm model on the ETW dataset is compared under different values of age similarity weight α, and the experimental results are shown in Figure 6.

As can be seen from Figure 6, keeping other parameters constant, the Precision@N, Recall@N and F1@N values increase with the increase of age similarity weight α until the maximum value is reached at $\alpha = 0.7$, and then start to decrease, and the accuracy of the algorithm starts to drop sharply after the value of α is greater than 0.8. The growth rate is slow when α is in the interval of 0.05~0.2 and 0.5~0.7, and the increase is extremely obvious when it is in the interval of 0.2~0.3. The experimental results show that age similarity and region similarity have almost equal impact on the recommendation accuracy of the algorithm, and the algorithm is at a lower level when the age similarity weight is lower than 0.4 or higher than 0.8, and reaches the optimal effect when the age similarity weight is 0.7. Keeping the age similarity weight α and interest similarity weight λ unchanged, the performance of the CWSAR algorithm model on the ETW dataset is compared for different values of the user similarity threshold ρ, and the experimental results are shown in Figure 7.

As can be seen from Figure 7, Precision@N values and Recall@N values do not change significantly, and F1@N values show a slow growth trend until reaching the maximum value at $\rho = 0.6$. Then start to decline slowly. When ρ is in the interval of 0.75~1, the algorithm accuracy decreases significantly, and when $\rho = 1$, all values go to 0. Because the similarity threshold ρ is a screening coefficient, when it is small, it is equivalent to no screening effect, and the algorithm accuracy stays at a more average level. When $\rho = 0.6$, the screening effect is the most obvious and the algorithm accuracy reaches the highest. When $\rho > 0.75$, the screening threshold is too high, resulting in the use of fewer selected neighbours and the algorithm accuracy starts to drop sharply.

5 Conclusions

This study addresses the problem of recommendation strategies for ELT words, and proposes the CWSAR algorithm model based on the STEM concept, combining collaborative filtering algorithm, word vector semantic understanding module, crowsensing and context perception. The results show that the CWSAR algorithm model achieves optimal recommendation accuracy when the interest similarity weight λ is taken as 0.8, the age similarity weight α is taken as 0.7, and the user similarity threshold ρ is taken as 0.6. CWSAR algorithm is better than the two CF algorithms in the experiment of ordinary data sets. Its Precision@N Value increases by 13.4%, and Recall@N increases by 24.3% in value; In the ETW data set experiment of English teaching words, two CF algorithms are compared in Precision@N increased by 71.0% in value Recall@N value increased by 40.2%. There are also some deficiencies in this study, such as the insufficient number of experimental data sets, and the self built English teaching word data sets are not complex and complete enough. It is expected that more data sets can be used for experiments in the future research to improve the accuracy of the experimental results and help improve the research of English teaching word recommendation methods. At the same time, in the future, more English teaching words will be adapted to enhance the practical value of the research and help students at different learning levels to improve their English learning ability.

Figure 5 Experimental results of different interest similarity weights (see online version for colours)

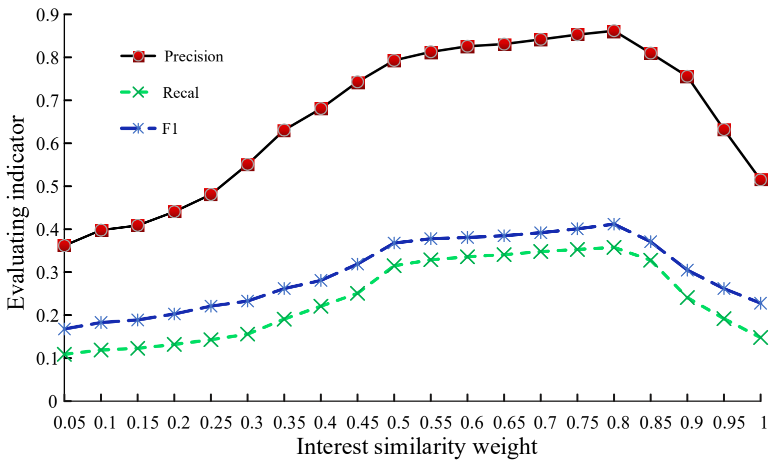


Figure 6 Experimental results of different age similarity weights (see online version for colours)

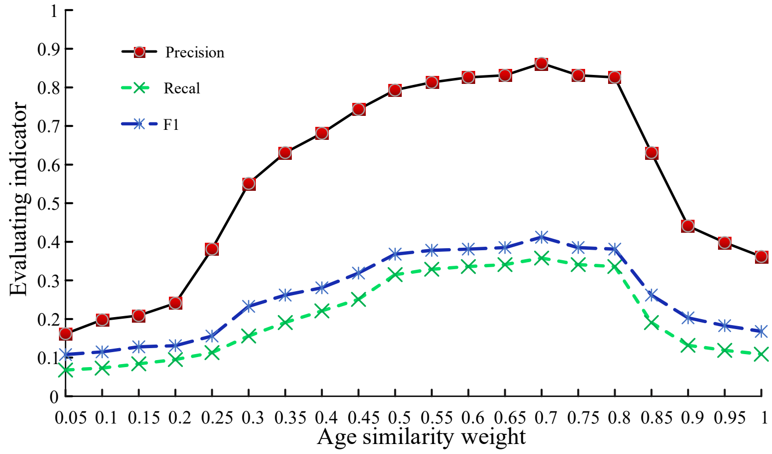
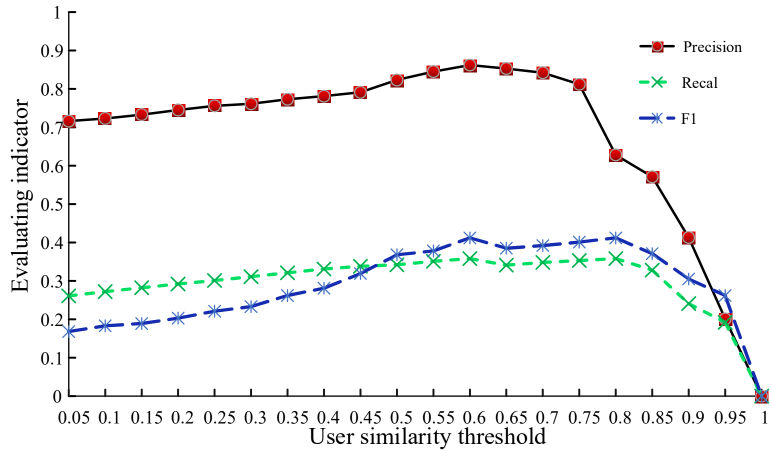


Figure 7 Experimental results of similarity threshold of different users (see online version for colours)



References

Bangyal, W.H., Qasim, R., Ahmad, Z. et al. (2021) 'Detection of fake news text classification on COVID-19 using deep learning approaches', *Computational and Mathematical Methods in Medicine*, Vol. 2021, No. 12, pp.1–14.

Cai, Z., Yuan, G., Qiao, S. et al. (2022) 'FG-CF: friends-aware graph collaborative filtering for POI recommendation', *Neurocomputing*, Vol. 488, No. 1, pp.107–119.

Chang, Y., Li, B. and Lu, J. (2021) 'English vocabulary teaching from a cognitive perspective', *Open Access Library Journal*, Vol. 8, No. 9, pp.1–8.

Chen, Z., Wang, Y., Zhang, S. et al. (2021) 'Differentially private user-based collaborative filtering recommendation based on k-means clustering', *Expert Systems with Applications*, Vol. 168, No. 8, pp.114366–114366.

Haider Bangyal, W., Hameed, A., Ahmad, J. et al. (2022) 'New modified controlled bat algorithm for numerical optimization problem', *Computers, Materials & Continua*, Vol. 70, No. 2, pp.2241–2259.

Joorabloo, N., Jalili, M. and Ren, Y. (2020) 'Improved collaborative filtering recommendation through similarity prediction', *IEEE Access*, Vol. 8, No. 1, pp.202122–202132.

Kang, P. (2020) 'A study of college english autonomous learning model based on computer and network', *IOP Conference Series Materials Science and Engineering*, Vol. 750, No. 1, pp.012031–012031.

Li, S. and Zhou, D. (2020) 'Research on application of collaborative filtering algorithm in digital movie recommendation', *Journal of Physics: Conference Series*, Vol. 1651, No. 1, pp.012091–012091.

Liu, J., Song, J., Li, C. et al. (2021) 'A hybrid news recommendation algorithm based on k-means clustering and collaborative filtering', *Journal of Physics: Conference Series*, Vol. 1881, No. 3, pp.032050–032050.

Lu, Q., Guo, F. and Zhang, R. (2019) 'User-based collaborative filtering recommendation method combining with privacy concerns intensity in mobile commerce', *International Journal of Wireless and Mobile Computing*, Vol. 17, No. 1, pp.63–70.

Ortiz-Revilla, J., Greca, I.M. and Arriassecq, I. (2022) 'A theoretical framework for integrated stem education', *Science & Education*, Vol. 31, No. 2, pp.383–404.

Ozkaya, A., Bulut, S. and Sahin, G. (2022) 'Investigation of the impact of stem education on the creative design skills of prospective teachers', *Asian Journal of Contemporary Education*, Vol. 6, No. 1, pp.1–15.

Qasim, R., Bangyal, W.H., Alqarni, M.A. et al. (2022) 'A fine-tuned BERT-based transfer learning approach for text classification', *Journal of Healthcare Engineering*, Vol. 2022, pp.1–17.

Tang, H., Zhao, G., Bu, X. et al. (2021) 'Dynamic evolution of multi-graph based collaborative filtering for recommendation systems', *Knowledge-Based Systems*, Vol. 228, No. 1, pp.107251–107251.

Wang, Q. (2021) 'Application of recommendation algorithm and big data technology in computer English corpus database', *Journal of Physics: Conference Series*, Vol. 2083, No. 3, pp.032092–032092.

Wu, Y., Zhao, Y. and Wei, S. (2020) 'Collaborative filtering recommendation algorithm based on interval-valued fuzzy numbers', *Applied Intelligence*, Vol. 50, No. 3, pp.2663–2675.

Yan, H.C., Wang, Z.R., Niu, J.Y. et al. (2022) 'Application of covering rough granular computing model in collaborative filtering recommendation algorithm optimization', *Advanced Engineering Informatics*, Vol. 51, No. 1, pp.101485–101485.

Yang, Q., Liang, Y. and Hu, X. (2020) 'Research on assignment algorithm of personalized intelligent recommended computing based on group intelligence perception', *Journal of Physics: Conference Series*, Vol. 1648, No. 3, pp.032008–032008.

Yuan, Z., Lee, J.H. and Zhang, S. (2021) 'Optimization of the hybrid movie recommendation system based on weighted classification and user collaborative filtering algorithm', *Complexity*, Vol. 2021, No. 6, pp.1–13.

Zhang W, Wang B, Zhou L (2021) 'Analysis of text feature of English corpus with dynamic adaptive recommendation algorithm fused with multiple data source English language', *Microprocessors and Microsystem*s, Vol. 6, pp.104075–104075.

Zhang, Z., Peng, T. and Shen, K. (2020) 'Overview of collaborative filtering recommendation algorithms', *IOP Conference Series: Earth and Environmental Science*, Vol. 440, No. 2, pp.022063–022063.