Research article

Optimizing an English text reading recommendation model by integrating collaborative filtering algorithm and FastText classification method

Ke Yan [*]

Department of Public Instruction, Nanyang Medical College, Nanyang, 473000, Henan, China

ABSTRACT

To comprehend the genuine reading habits and preferences of diverse user cohorts and furnish tailored reading recommendations, this study introduces an English text reading recommendation model designed specifically for long-tail users. This model integrates collaborative filtering algorithms with the FastText classification method. Initially, the integrated collaborative filtering algorithm is explicated, followed by the calculation of the user's interest distribution across various types of English texts, achieved through an enhanced Ebbinghaus forgetting curve and analysis of user reading behaviors. Subsequently, an intelligent English text reading recommendation is generated by amalgamating collaborative filtering algorithms with association rule-based recommendation algorithms. Through optimization of the recommendation generation process, the model's recommendation accuracy is enhanced, thereby augmenting the performance and user satisfaction of the recommendation system. Finally, a comparative analysis is conducted with respect to the Top-N algorithm model, matrix factorization-based algorithm model, and FastText classification model, illustrating the superior recommendation accuracy and F-Measure value of the proposed model. The study findings indicate that when the recommendation list contains 10, 30, 50, and 70 texts, the recommendation accuracy of the proposed algorithm model is 0.75, 0.79, 0.8, and 0.74, respectively, outperforming other algorithms. Furthermore, as the number of texts increases, the F-Measure of all four models gradually improves, with the final F-Measure of the proposed model reaching 0.81. Notably, the F-Measure of the English text reading recommendation model proposed in this study significantly surpasses that of the other three recommendation methods. Demonstrating commendable performance in recall rate, root mean square error, normalized cumulative gain, precision, and accuracy, the model adeptly reflects user reading interests, thereby enhancing the accuracy of text recommendations and the overall system performance. The study findings offer crucial insights and guidance for enhancing the accuracy and overall efficacy of English text recommendation systems.

1. Introduction

In today's internet era, personalized recommendation systems have become crucial driving factors for user engagement in English

text reading. These systems leverage advanced algorithms and data analytics techniques to provide customized reading suggestions based on users' interests, preferences, and behavioral patterns. However, despite significant progress in enhancing user experience and promoting reading activities, these systems still face a range of challenges, including the problem of long-tail users and other aspects [1,2]. Long-tail users refer to those user groups with relatively uncommon or more diverse interests or preferences. The interest areas of these users may be specific, while traditional recommendation systems often tend to provide popular or common content for mainstream users. As a result, long-tail users are often overlooked, leading to them not receiving recommendations tailored to their interests. Addressing the issue of long-tail users requires personalized recommendation systems to more finely analyze user behavior and interests, thereby providing recommendations that better meet the needs of this user segment [3,4]. Recommendation systems typically generate recommendations based on users' past behavior and preferences, but these results often lack interpretability. Users want to understand why they are seeing certain recommendations, in order to comprehend the operation principles of the recommendation system and increase trust in the recommended results. Therefore, improving the interpretability of recommendation systems becomes crucial, which means the system needs to be able to clearly explain the recommendation logic and basis behind the recommended results [5,6]. Personalized recommendation systems need to handle a large amount of textual data in order to accurately understand users' reading preferences and generate corresponding recommendations. However, processing textual data may face challenges such as insufficient accuracy and low efficiency. These issues may stem from the complexity of textual data and limitations in current text processing techniques. Therefore, enhancing the text data processing capability of personalized recommendation systems is an important task that requires leveraging advanced text processing techniques and algorithms to address [7]. Collaborative filtering and association rules are two important algorithmic techniques in personalized recommendation systems. Collaborative filtering algorithms make recommendations based on the similarity between users or items and can be divided into user-based collaborative filtering and item-based collaborative filtering. These algorithms have been widely researched and applied, achieving success in many recommendation systems. In recent years, with the increase in data volume and improvement in computational power, collaborative filtering-based personalized recommendation systems have been continuously optimized and improved, leading to the emergence of many new algorithms and technologies such as latent semantic model-based collaborative filtering, social network-based collaborative filtering, etc. [8]. On the other hand, association rules are a technique for discovering relationships between items in a dataset by analyzing the co-occurrence of items in the dataset. In personalized recommendation systems, association rules can be used to discover associations between users' purchase or browsing behaviors, thereby providing personalized recommendations. In recent years, with the development of data mining and machine learning technologies, research on association rules in personalized recommendation systems has been deepening, leading to the emergence of many new algorithms and methods such as association rules based on sequence pattern mining, association rules based on user behavior sequences, etc. [9]. However, currently, these two individual approaches have certain limitations in addressing the problem of long-tail users, and more advanced methods are needed to improve the performance of personalized recommendation systems and better meet users' personalized needs [10].

The objective of this study is to augment the precision and overall efficacy of the English text reading recommendation system. This objective is attained through the integration of collaborative filtering algorithms and FastText classification methods, adoption of recommendation algorithms founded on association rules to capture robust rules, and refinement of the recommendation generation process to enhance recommendation accuracy and user contentment. Furthermore, the proposed approach undergoes comparison and analysis against alternative algorithmic models to substantiate its superiority. The study commences with a comprehensive elucidation and examination of the integrated collaborative filtering algorithm. Subsequently, it employs a recommendation algorithm based on association rules to capture robust rules for text reading recommendations, synergizing with collaborative filtering algorithms to yield personalized recommendation outcomes. Ultimately, to affirm the superiority of the proposed algorithmic model, a comparative analysis is conducted vis-à-vis the Top-N algorithm model, matrix factorization-based algorithm model, and FastText classification model in terms of recommendation accuracy and F-measure values. The novelty of this study resides in the fusion of collaborative filtering algorithms and FastText classification methods. Through enhancements to the forgetting curve and the incorporation of recommendations grounded on association rules, personalized recommendations tailored for long-tail users are realized. This advancement bolsters the accuracy of the personalized recommendation model, endeavoring to furnish English text reading recommendations that more closely resonate with users' interests and preferences. The significance of this study lies in the optimization of personalized recommendation systems, particularly within the domain of English text reading. By amalgamating collaborative filtering algorithms, recommendation algorithms based on association rules, and FastText classification methods, the study tackles challenges inherent in conventional recommendation systems, such as suboptimal recommendation accuracy and diminished satisfaction among long-tail users. In doing so, this study yields a notable contribution to the realm of recommendation systems, proffering more precise and gratifying English text reading recommendations and propelling the advancement of personalized recommendation technology.

The contribution of this study encompasses both theoretical and practical dimensions. Theoretically, it presents a model for tailored English text reading recommendations aimed at long-tail users, integrating collaborative filtering algorithms with the FastText classification method. This integration, coupled with an enhanced Ebbinghaus forgetting curve and association rule-based recommendation algorithms, significantly advances the field of personalized recommendation. The introduction of this integrated approach addresses gaps in existing research and establishes a new theoretical framework for the future evolution of recommendation systems. From a managerial standpoint, the English text reading recommendation model proposed herein not only enhances theoretical comprehension but also holds substantial managerial implications. The model adeptly discerns the reading preferences of long-tail users, thereby offering more precise recommendation services for reading service providers such as libraries and online bookstores. Operationalizing this model in practical management applications can elevate the accuracy and user satisfaction of reading recommendation systems, consequently fostering user loyalty and utilization rates. Furthermore, from a policy perspective, this study provides valuable insights for policy formulation. Given the prevalence of digital reading and online reading services, governments and

relevant institutions are tasked with formulating policies to regulate and foster the development of this domain. The English text reading recommendation model proposed in this study can serve as a guiding framework for governmental bodies, aiding in the formulation of more intelligent and personalized reading recommendation policies. This, in turn, can facilitate the wholesome advancement of the digital reading industry, amplifying public interest and efficacy in reading endeavors.

2. Literature review

With the rapid advancement of the internet and the proliferation of big data, personalized recommendation systems have emerged as pivotal tools for enriching user experiences and catering to individual needs. Within the realm of recommendation systems, text classification and collaborative filtering algorithms stand as prominent methodologies, furnishing robust frameworks for the development of recommendation models. By delving into both user behavioral data and textual content, these foundational technologies facilitate a more nuanced understanding of user interests and preferences, thereby facilitating the delivery of personalized recommendation services. Wang et al. (2023) underscored the indispensable role of recommendation systems in information retrieval and filtering. By actively mining and analyzing users' historical behavioral data to discern precise user preferences, such systems find application in project decision-making and resource optimization within the cultural and creative industries, exhibiting commendable levels of intelligence, accuracy, and personalization [11]. Yannam et al. (2023) proposed a rating prediction approach for metadata that amalgamates multi-layer perceptrons with generalized matrix factorization, refining recommendation systems through neural collaborative filtering techniques [12]. Vuong Nguyen et al. (2023) suggested that the integration of collaborative filtering methods with content analysis based on word embeddings can mitigate issues of data sparsity and user rating cold starts inherent in conventional recommendation systems [13]. Venkatesan (2023) posited collaborative filtering as the predominant method in contemporary recommendation systems, with matrix factorization commonly employed to address associated challenges [14]. Karabila et al. (2023) discovered that merging sentiment analysis of textual data with collaborative filtering techniques utilizing ensemble learning methodologies can furnish users with more precise and personalized suggestions. This fusion approach notably enhances the accuracy of collaborative filtering algorithms, culminating in refined recommendation outcomes [15]. Alatrash and Priyadarshini (2023) combined alternating least squares collaborative filtering with fine-grained sentiment analysis, proposing three fine-grained sentiment analysis models grounded in attention mechanisms and bidirectional Long Short-Term Memory. Their findings yielded an improved accuracy of 93.39 % [16]. Mudavath and Negi (2024) devised a luggage item recommendation system employing FastText word embeddings and association rule mining to deliver personalized suggestions to users. This result demonstrated the efficacy of the recommendation model based on FastText and ARM across metrics including coverage, support, confidence, lift, leverage, and conviction [17]. Pilato et al. (2023) indicated that incorporating user historical behavior and text information into the FastText model enhances the system's ability to capture user interests and preferences accurately, thereby yielding more targeted recommendation results. The FastText model exhibits exceptional proficiency in handling large-scale text data and is well-suited for personalized recommendation scenarios, thereby offering robust support for enhancing user experiences and recommendation system efficacy [18].

The preceding literature review underscores the prevailing focus in current recommendation systems research on algorithmic integration, particularly the amalgamation of text analysis with collaborative filtering techniques aimed at refining the accuracy and personalization of recommendation systems. Concurrently, endeavors are directed towards mitigating challenges such as data sparsity and user rating cold start predicaments by leveraging users' historical behaviors and textual information to heighten the precision of personalized recommendations, thereby augmenting user experiences. Regarding practical applications, researchers emphasize the imperative of evaluating recommendation system performance in real-world contexts and employ diverse evaluation metrics to substantiate their efficacy. Nevertheless, notwithstanding some strides made in experimental domains, persistent challenges persist, including the dependability of recommendation systems, scalability concerns associated with recommendation lists, and the organization of user preference attributes. Prior studies on English text recommendation remain circumscribed, potentially attributable to the intricacies and variegated nature of English textual data, which compound the challenges faced by recommendation systems in comprehending and processing textual content. English text embodies diverse linguistic styles, semantic nuances, and grammatical intricacies, necessitating recommendation systems to wield robust natural language processing capabilities to adeptly decipher and analyze textual content. This study endeavors to redress prospective issues inherent in recommendation systems based on antecedent scholarship. By introducing innovative methodologies and comprehensive strategies, the aspiration is to contribute towards the optimization of English text reading recommendation systems.

3. Research theory and English Text Reading recommendation model

3.1. Text classification

Text classification constitutes a supervised learning endeavor geared towards automatically categorizing textual data. Within the ambit of text classification, textual data undergoes preprocessing stages encompassing tasks such as tokenization, stop word elimination, stemming, or lemmatization, thereby rendering it amenable to classification. Subsequently, feature extraction methodologies are deployed to translate the textual content into numerical or vectorized representations. A classifier model is then trained accordingly. Following the completion of training, the classifier model stands poised to categorize novel, unlabeled textual inputs into their respective classes [19,20]. Fig. 1 (a, b) presents the division of features and models used in text classification.

3.2. Collaborative filtering algorithm

Collaborative filtering stands as a ubiquitous methodology within recommendation systems, offering tailored recommendations to users through the analysis of user behavior data and item correlations. Its widespread utilization spans across diverse domains including e-commerce, cinema, music, and news, furnishing users with personalized suggestions attuned to their inclinations and predilections [21]. Within collaborative filtering, two primary variants exist: user-based collaborative filtering, predicated on user similarity, and item-based collaborative filtering, rooted in item associations [22]. Fig. 2 (a, b) shows the basic principles of user-based collaborative filtering and item-based collaborative filtering.

In Fig. 2 (a, b), user-based collaborative filtering entails the identification of users sharing akin interests with the focal user, subsequently recommending items endorsed by these akin users to the target user. Conversely, item-based collaborative filtering suggests alternative items associated with a specified item predicated on user ratings or engagements.

Within the user-based collaborative filtering algorithm, user similarities are quantified through metrics such as cosine similarity. Let x represent the target user and z denote the item. The anticipated rating within the user-based collaborative filtering algorithm is expressed as Equation (1):

$$\acute{R}_{x,z} = \overline{R}_x + \frac{\sum\limits_{y \in N(x,z)} \sin(x,y)\left(R_{y,z} - \overline{R}_y\right)}{\sum\limits_{y \in N(x,z)} |\sin(x,y)|} \tag{1}$$

Equation (1) represents the user-based collaborative filtering algorithm. $\acute{R}_{x,z}$ denotes the pre0dicted rating of user x for item z. $\overline{R}_x$ is the average rating of the user x. $N(x,z)$ is the set of items z that have been interacted with by user x. $\sin(x,y)$ represents the similarity between user x and user y. $R_{y,z}$ represents the actual rating of user y for item z. $\overline{R}_y$ is the average rating of the user y.

In the context of the item-based collaborative filtering algorithm, the quantification of the affinity between items may be conducted through methodologies like cosine similarity. Herein, let the user of interest be denoted as x and the item under consideration as z. The anticipated rating within the framework of the item-based collaborative filtering algorithm is expressible through Equation (2).

$$\acute{R}_{x,z} = \overline{R}_z + \frac{\sum\limits_{m \in N(z,x)} \sin(z,m)\left(R_{x,m} - \overline{R}_m\right)}{\sum\limits_{y \in N(m,x)} |\sin(z,m)|} \tag{2}$$

Equation (2) represents the item-based collaborative filtering algorithm, where $\acute{R}_{x,z}$ denotes the predicted rating of user x for item z. $\overline{R}_z$ is the average rating of an item z. $N(z,x)$ is the set of items z that have been interacted with by user x. $\sin(z,m)$ represents the similarity between item z and item m. $R_{x,m}$ represents the actual rating of user x for item m, and $\overline{R}_m$ is the average rating of an item m.

The collaborative filtering algorithm process is illustrated in Fig. 3.

Characterization	Concrete meaning		Classification Models	Model name
Unique hot code	Solo thermal encoding uses N-bit status registers to encode N states, each with its own individual register bit and only one of which is valid at any given time.		The shallow learning model	Naive Bayes model
				Support vector machine
Word vector	Word vectoring is a technique for mapping words in natural language to a real number vector space.			K-nearest neighbor
				Decision tree
Bag-of-words model	Bag-of-words modeling is a method of text vectorization that converts plain text into a vector representation.			Random forest
			Deep learning models	Rule embedded neural networks
				Multi-layer perception
N-element model	The N-element grammar model utilizes the previous N-1 words to predict the next word.			Recurrent neural networks
				Convolutional neural networks
Word frequency-inverse document frequency	Word Frequency - Inverse Document Frequency can quantitatively represent strings of words, phrases, etc.			Attention mechanisms
				Transformer
				Graph convolutional neural networks

(a)   (b)

Fig. 1. Features and Model Categories of Text Classification: (a) feature representation for text categorization; (b) model classification for text categorization

Within Fig. 1 (a, b), prominent features employed in text classification consist of the bag-of-words model, word embeddings, and one-hot encoding. Meanwhile, the suite of model algorithms applied in text classification encompasses naive Bayes, support vector machines, and deep learning models.
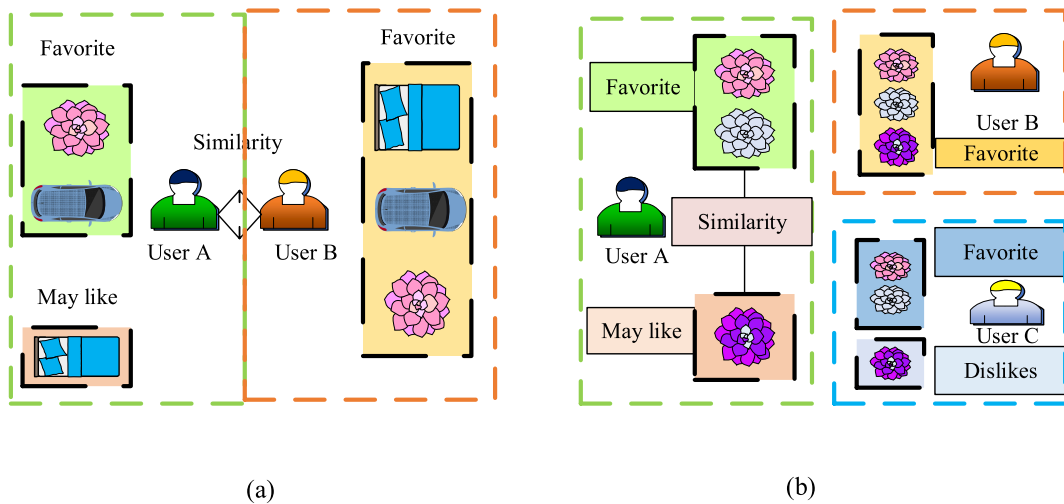
(a)   (b)

Fig. 2. Principles of User-Based Collaborative Filtering and Item-Based Collaborative Filtering: (a) Principles of user-based collaborative filtering algorithms; (b) Project-based collaborative filtering algorithm principle.

In Fig. 3, the collaborative filtering algorithm undertakes an analytical process involving the examination of user behavior data or item affinities to ascertain the most akin neighbors to the target user or item. Leveraging the behaviors or ratings of these neighbors, personalized recommendations or prognostications are formulated, thereby facilitating personalized recommendations.
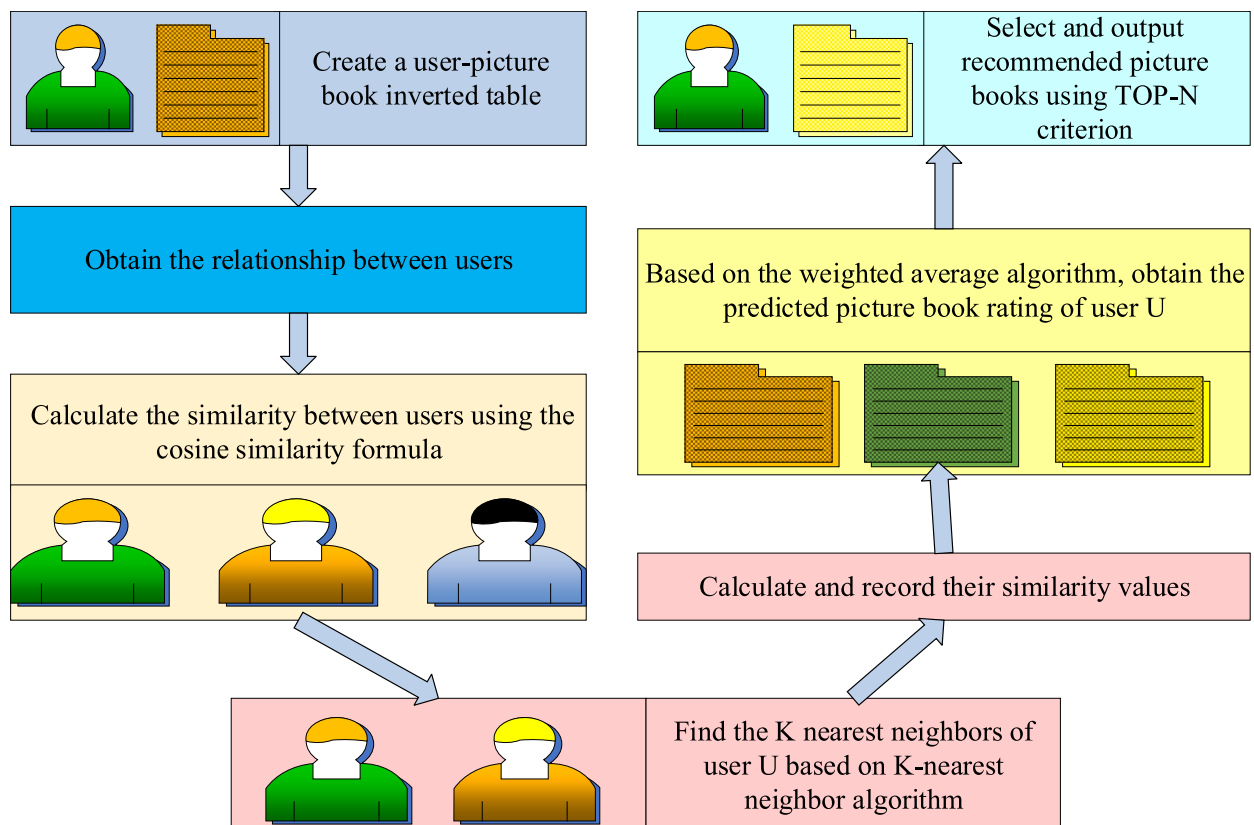

Fig. 3. Collaborative filtering algorithm process.

3.3. FastText classification

FastText classification stands as an adept text classification algorithm that amalgamates the attributes of the bag-of-words model and word embeddings. Renowned for its exemplary performance across various domains including text classification, sentiment analysis, and spam filtering, FastText efficaciously processes textual data. By treating text as a conglomerate of words and converting it into a fixed-length vector representation, FastText dispenses with word order, thereby expediting processing. Leveraging word embeddings to delineate words, FastText encapsulates semantic nuances and contextual relationships inherent within the text [23]. Structurally akin to the continuous bag-of-words model (CBOW), the architecture of the FastText classification model aligns with CBOW's framework.

Let α represent text sample, then the probability distribution of each category F is calculated as shown in Equation (3):

$$Q(F|\alpha) = \frac{e^{K_F L_\alpha}}{\sum\limits_{\acute{F}} e^{K_F L_\alpha}} \tag{3}$$

Equation (3) represents that $Q(F|\alpha)$ is the probability of category F given the text sample α. K_F represents the weight vector of category F, and L_α represents the vector representation of the text sample α.

CBOW is a widely used model for learning word vectors in natural language processing [24]. The structures of CBOW and FastText models are shown in Fig. 4 (a, b).

3.4. Association rule recommendation algorithm

The association rule recommendation algorithm represents a technique aimed at unveiling the relationships between items by extracting frequent itemsets from past user behavior data [25]. Employing support and confidence metrics, it evaluates the frequency and robustness of associations among item combinations, facilitating the identification of implicit linkages between items devoid of reliance on user-specific data [26]. Fig. 5 (a, b) illustrates the configuration and operation of the association rule recommendation algorithm.

Fig. 5 (a, b) illustrates the structural framework of the association rule recommendation algorithm, which hinges upon the mining of frequent itemsets and the generation of association rules. The algorithm's procedural workflow encompasses the identification of frequent item sets, the derivation of association rules to unveil inter-item relationships, and their subsequent application towards
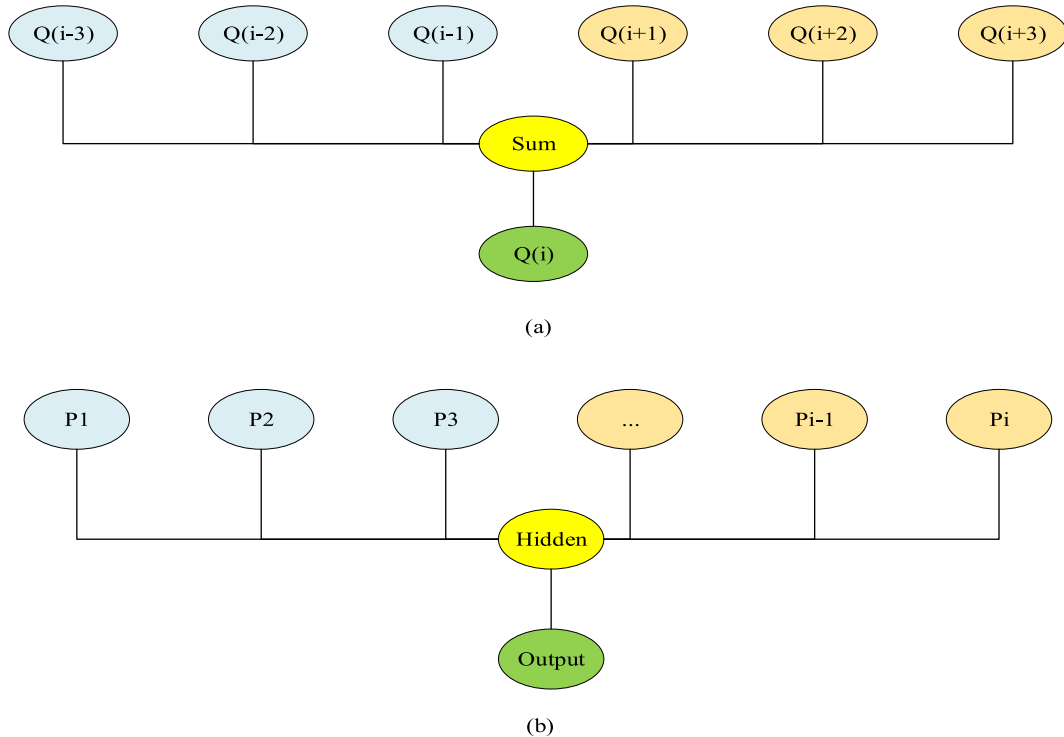
(a)

(b)

Fig. 4. CBO Structure and FastText Model Structure: (a) CBOW structure; (b) FastText model structure

In Fig. 4 (a, b), the CBOW model forecasts the target word through the averaging of contextual word vectors. Serving as an extension of CBOW, the FastText model delves deeper by dissecting words into character-level components and constructing vector representations at the subword level. This augmentation enables enhanced performance in text classification and the computation of word similarities.
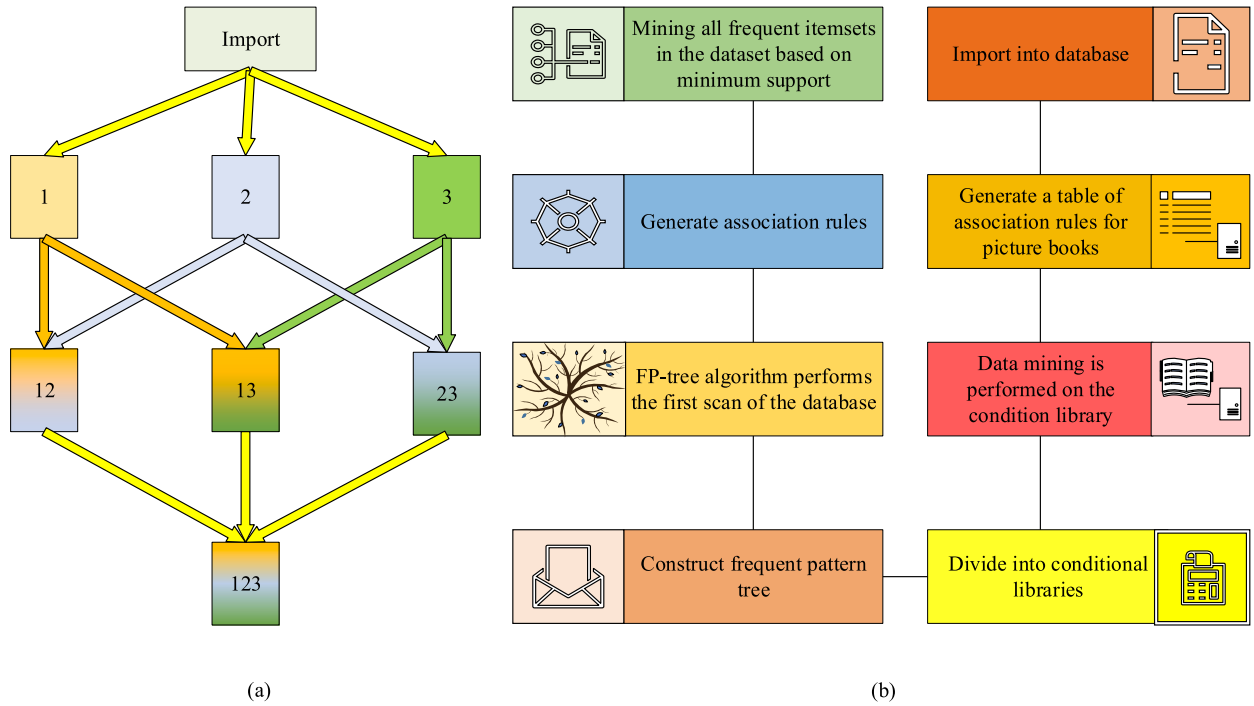
Fig. 5. Structure and Process of the Association Rule Recommendation Algorithm: (a) Association rule recommendation algorithm structure; (b) Flow of association rule recommendation algorithm.

personalized recommendations.

3.5. Ebbinghaus forgetting curve

The Ebbinghaus forgetting curve, introduced by psychologist Hermann Ebbinghaus in 1885, elucidates the phenomenon wherein memory retention exhibits a gradual decline subsequent to the process of learning [27].

H represent the initial memory strength, and *T* denote the time elapsed since the initial learning. The forgetting rate *S* can be calculated as shown in Equation (4):

$$S = H \times e^{-\frac{T}{\beta}} \tag{4}$$

Equation (4) represents that e is a constant term, and β is the time constant of the forgetting curve, which controls the forgetting rate. As time *T* increases, *S* gradually increases, indicating that memory weakens over time. The slope of the forgetting curve is determined by the constant β, where a smaller β leads to a faster forgetting rate and vice versa. The Ebbinghaus forgetting curve, obtained based on the forgetting rate calculation formula, is presented in Table 1.

Equation (5) delineates the estimation of users' memory retention rate concerning the texts they have perused, leveraging the principles elucidated by the Ebbinghaus forgetting curve. The computation of the memory retention rate is explicated through Equation (5):

$$O = e^{-\frac{T}{\beta}} \tag{5}$$

Table 1
Ebbinghaus forgetting curve.

Time interval	Memory test
Just finished memory	100 %
After 20 min	58.20 %
1 h later	44.20 %
8–9 h later	35.80 %
One day after	33.70 %
Two days later	27.80 %
Six days later	25.40 %
One month later	21.10 %

Equation (6) represents that the user's interest in different texts can be calculated based on the user's behavior and memory retention rate. Let $O_{\theta,i}$ denote the memory retention rate for text i at the θ-th reading, $O_{\theta,\delta}$ represent the memory retention rate for all δ texts at the θ-th reading, and $b_{\delta,\theta}$ be the user's interest in all δ texts at the θ-th reading. The calculation of interest proportion is shown in Equation (6):

$$B_i = \frac{O_{\theta,i}}{O_{\theta,\delta} b_{\delta,\theta}} \tag{6}$$

3.6. The English text reading recommendation model design

Personalized recommendation refers to an academic service within recommendation systems aimed at tailoring the most pertinent content, products, or services to individual users based on their specific interests, preferences, historical behaviors, and other pertinent information. These systems utilize collaborative filtering, matrix factorization, and deep learning techniques to analyze and model individual user characteristics, thereby facilitating accurate recommendations and enhancing user experience and satisfaction [28].

The Top-N algorithm, comparatively, is a relatively straightforward recommendation approach that typically relies on popular trends or basic rules to directly suggest a selection of the most popular or latest items to users. Nonetheless, this method lacks personalized considerations and may not effectively cater to the requirements of long-tail users [29].

The matrix factorization-based recommendation algorithm, on the other hand, is a classical collaborative filtering method that predicts users' interests in unread items by decomposing the rating matrix representing interactions between users and items. While this approach partially incorporates personalized factors, it also faces challenges such as data sparsity and cold start problems, especially in the case of long-tail users, which can potentially introduce recommendation biases [30].

The Top-N algorithm lacks a fixed formula but typically employs a rating prediction equation to calculate item scores. Conversely, the matrix factorization-based recommendation algorithm commonly employs SVD for computation. Denoting the user-item rating matrix as U, it can be decomposed into three new matrices, with their relationship depicted as shown in Equation (7):

$$U = W \times V \times P \tag{7}$$

In Equation (7), W represents the user matrix. V represents the singular value matrix. P represents the transposition of the item matrix.

To address dimensionality reduction in the rating matrix, with a consideration of n users and k items, the computation of user-item ratings is delineated in Equation (8):

$$\acute{R}_{n,k} = \sum_{d=1}^{D} W_{n,d} P_{d,k} \tag{8}$$

Equation (8) represents the calculation of user-item ratings after dimensionality reduction. $W_{n,d}$ denotes the d-th feature vector of user n, $P_{d,k}$ represents the d-th feature vector of item k, and D is the dimensionality of the feature vectors.

The steps of the Top-N algorithm and the structure of the matrix factorization algorithm model are shown in Fig. 6 (a, b).
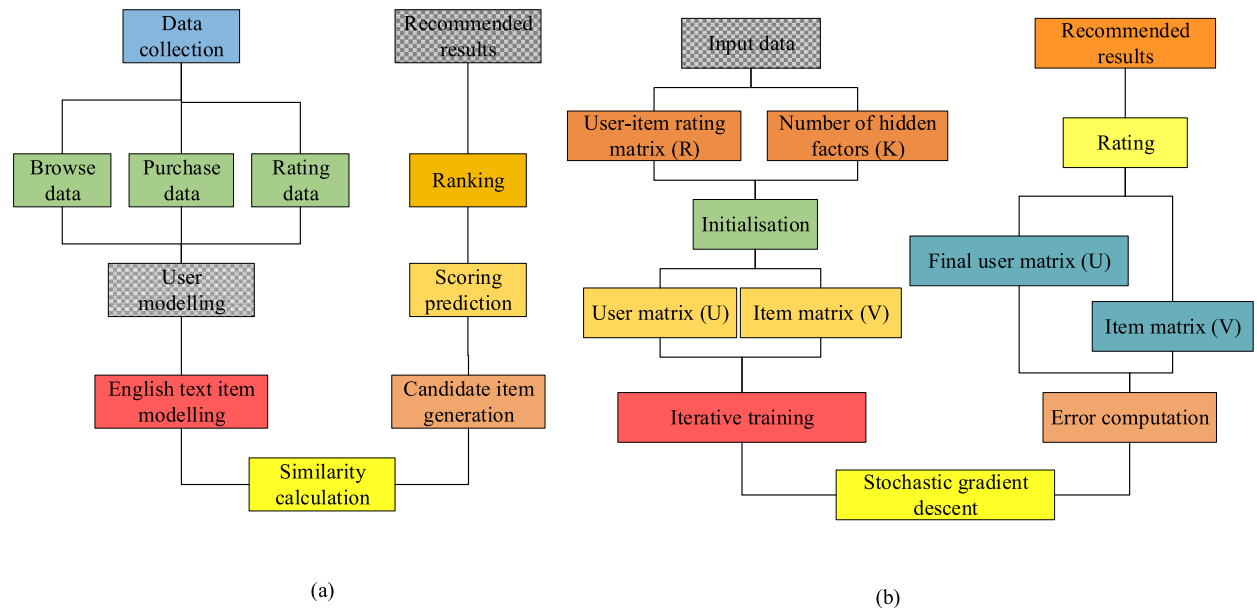

Fig. 6. Top-N Algorithm Steps and Matrix Factorization Algorithm Model Structure: (a) Top-N algorithm steps; (b) Matrix decomposition algorithm model structure.

In Fig. 6 (a, b), the Top-N algorithm proceeds by leveraging users' historical behavior data to forecast their preferences for unrated items. The structure of the matrix factorization algorithm model entails decomposing the rating matrix into low-dimensional user and item feature vector matrices. This facilitates the prediction of user ratings for unrated items, thereby achieving personalized recommendations within the recommendation system.

English text reading refers to the activity of enhancing English reading and comprehension abilities through the perusal of English books, articles, news, blogs, and other textual materials. It plays a pivotal role in language acquisition, contributing to the expansion of vocabulary, improvement of reading comprehension skills, and deepening of understanding in English grammar and expression. The process of uploading and classifying English text reading is shown in Fig. 7 (a, b).

In Fig. 7 (a, b), the process of uploading and classifying English text reading primarily encompasses the submission of text data to the system, subsequent preprocessing, and feature extraction within the system, followed by text categorization into various classes or topics. This process facilitates the provision of intelligent English text reading recommendation and classification services.

This study employs a model that amalgamates collaborative filtering algorithms and FastText classification methods, integrating refined forgetting curves and association rule recommendations. Such an approach is poised to enhance the accuracy and personalization of recommendations, thereby better catering to the needs of long-tail users. The framework of the English text reading recommendation model is depicted in Fig. 8.

In Fig. 8, this study conducts an analysis at two tiers: user historical behavior and current text data. Initially, text preprocessing is executed to generate vectors, succeeding which dense text vectors are created and a rapid text model is constructed. Subsequently, English text data undergoes processing to yield text classification outcomes. In the subsequent phase, contextual cues are extracted from user behavior data, employing the Ebbinghaus forgetting curve to compute memory retention rates and interest level percentages. Historical user behavior data is amassed, and the ultimate interest percentage is determined. In the final stage, the produced text classification results are amalgamated with the calculated final interest percentage to ascertain the optimal recommendation coefficients. This intelligence is utilized for text filtration, ultimately culminating in the generation of the recommended list.

3.7. Experimental data design

The English text reading recommendation model, based on collaborative filtering algorithms and FastText classification as explored in this study, leverages the Book-Crossing dataset. The Book-Crossing dataset, publicly accessible, serves as a prominent resource for investigating and assessing recommendation systems and collaborative filtering algorithms. It encompasses anonymous user ratings and critiques of books, alongside detailed book metadata like the International Standard Book Number (ISBN), title, and author. User ratings span a scale from 1 to 10, denoting the least to the highest liking, respectively, with some users supplementing textual reviews. This dataset holds paramount importance in the refinement and validation of recommendation algorithms, fortifying the efficacy of personalized recommendations. The dataset's openness fosters resource-sharing within academia and industry, thereby nurturing advancements in recommendation systems development.
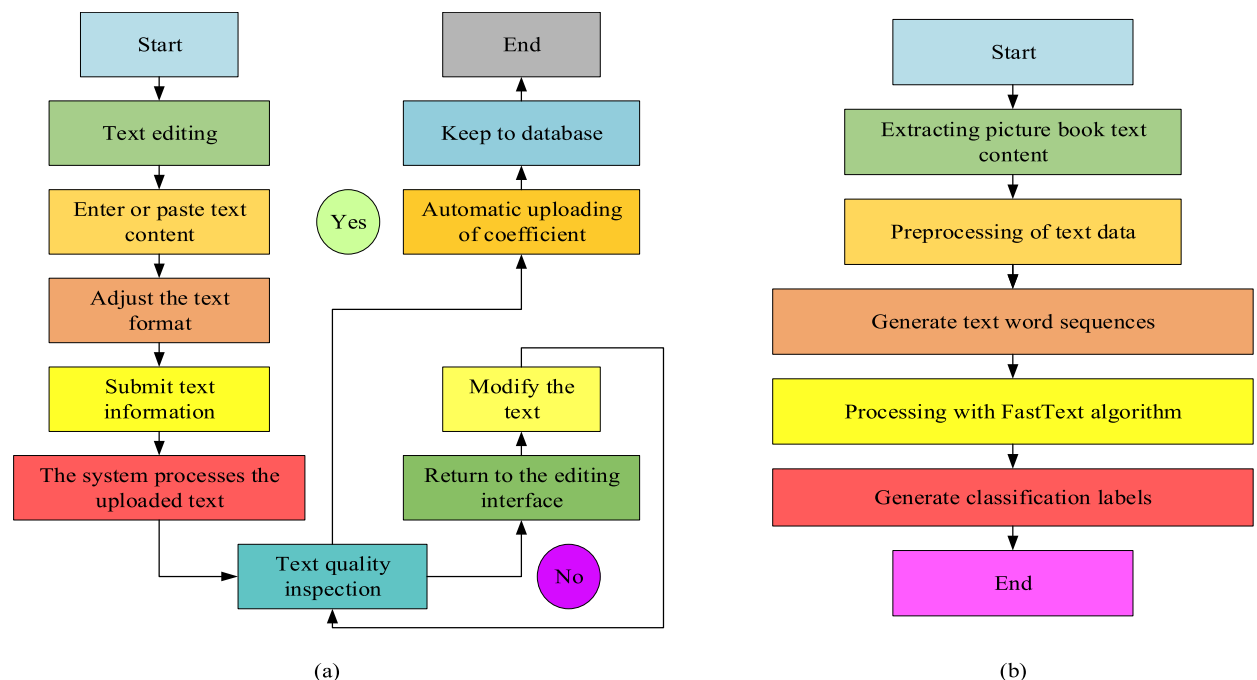

Fig. 7. English Text Reading Upload and Classification Process:(a) English text uploading process; (b) English text categorization process.
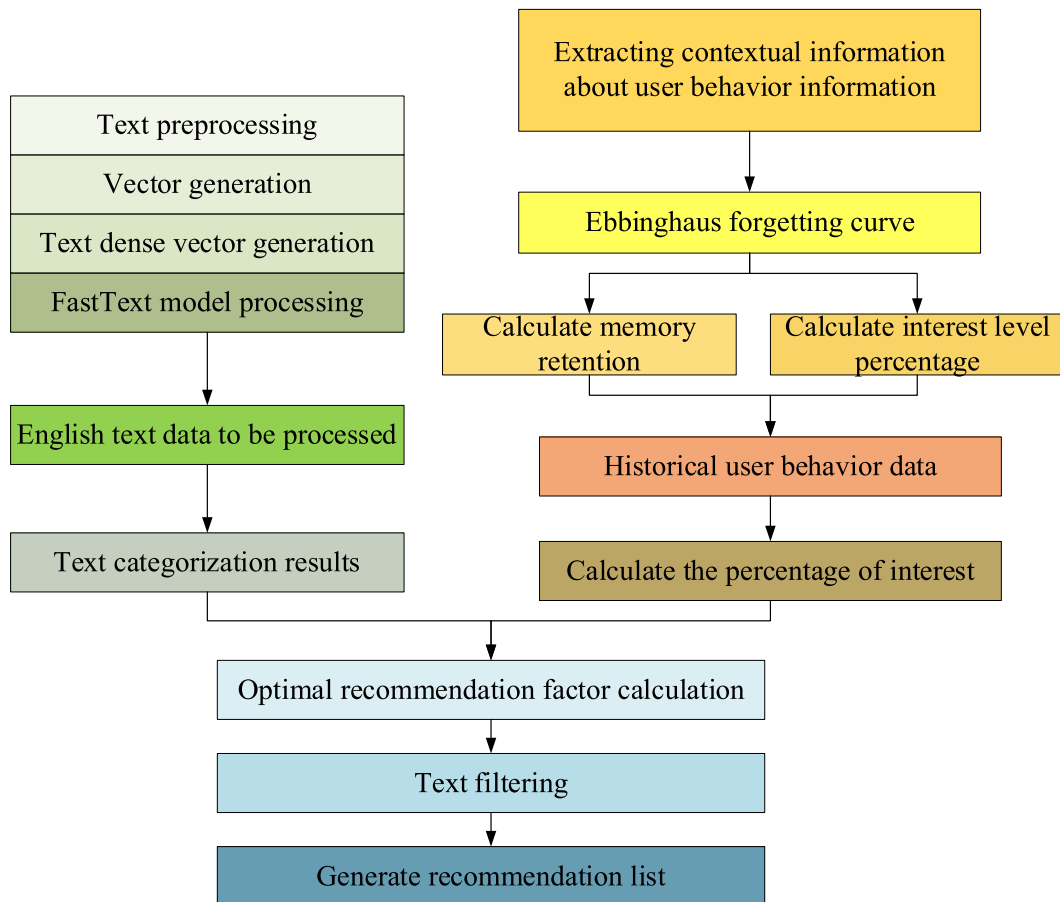
Fig. 8. English text reading recommendation model structure.

Several factors underpin the selection of the Book-Crossing dataset for experimentation. Primarily, its richness and diversity, encapsulating a vast array of user ratings and reviews across diverse book genres and domains, facilitate a comprehensive consideration of users' multifaceted interests and preferences, thereby enriching the accuracy and breadth of recommendations. Secondly, its public availability ensures unrestricted accessibility for research and development endeavors, propelling progress in the recommendation systems domain. Moreover, the dataset's real-world user ratings and reviews render it highly practical for evaluating model performance, offering insights into its efficacy and utility in real-world contexts. This augments comprehension of the model's effectiveness, furnishing robust support and guidance for its practical deployment. Lastly, as a widely embraced benchmark dataset in both academic and industrial circles, its adoption enhances the model's result comparability and generalizability, thus bolstering research reproducibility and findings dissemination. Therefore, the selection of the Book-Crossing dataset is predicated upon its richness, accessibility, practical utility, and widespread adoption, rendering it an optimal substrate for the study and appraisal of recommendation systems.

The rationale behind selecting the FastText model in this study encompasses several facets. Firstly, FastText stands out as a rapid and efficient text classification model, particularly adept at handling extensive text datasets. Its notable performance in text classification tasks facilitates swift training and attains commendable accuracy. Secondly, the FastText model exhibits proficiency in managing text data across multiple languages, showcasing robust cross-lingual generalization capabilities. This attribute renders it well-suited for addressing English text recommendation tasks, particularly in scenarios spanning international interfaces and multilingual contexts. Lastly, the FastText model is characterized by a straightforward and intuitive interface, along with a comprehensible training process, facilitating swift adoption and experimentation by researchers. Its user-friendly nature positions it as one of the widely embraced tools for text classification in practical projects. In contrast to contemporary methodologies, the FastText model boasts distinctive traits and advantages. Notably, its hallmark attributes include rapid training speed and efficient performance. When juxtaposed with cutting-edge deep learning models like BERT and Transformer, the FastText model typically incurs lower computational costs and exhibits accelerated training speeds, particularly excelling in handling extensive text datasets. Additionally, the FastText model demonstrates commendable performance even with limited data volumes, rendering it more adaptable to resource-constrained environments. Conversely, some contemporary deep learning models may necessitate substantial training data and computational resources to achieve optimal outcomes. Moreover, the FastText model adeptly manages text data across diverse

languages, encompassing both abundant and unconventional linguistic contexts. This versatility renders it highly suitable for cross-lingual tasks and multilingual settings, whereas certain contemporary models may require additional adaptations and enhancements to accommodate various languages. Furthermore, the FastText model's simplistic interface and training process contribute to its ease of use and comprehension. In contrast, some contemporary deep learning models may entail more intricate setup and tuning procedures. Detailed parameter configurations for this study are elucidated in Table 2.

The primary objective of this study is to significantly enhance user satisfaction and engagement with English text reading materials, particularly focusing on long-tail user groups, through the implementation of personalized recommendations, improved accuracy and precision, heightened user engagement, and augmented user trust. Personalized recommendations are instrumental in catering to the diverse needs of long-tail user groups, thereby fostering greater interest and engagement with the recommended materials. Furthermore, optimized recommendation algorithms have the potential to refine the accuracy and precision of the recommendation system, thereby instilling greater user trust and subsequently elevating their satisfaction and engagement levels. These initiatives aim to encourage users to actively participate in English text reading endeavors and to enthusiastically explore and appreciate the recommended reading materials.

The preprocessing of English text encompasses several sequential steps. Initially, non-textual components such as spaces, symbols, and paragraph separators are eliminated using regular expressions. Subsequently, spell checking and correction procedures are conducted utilizing the pyenchant library to rectify any potential spelling inaccuracies. Stemming and lemmatization techniques, facilitated by the Natural Language Toolkit (NLTK) library, are then applied to derive the original forms of words. Consequently, text normalization procedures convert the text to lowercase to mitigate case discrepancies and enhance statistical accuracy. Filtering out stop words, accomplished through the stopwords library, effectively removes irrelevant terms such as "a" and "to" from the text. Finally, feature processing operations are carried out, with vectorization and the Hash Trick being common methodologies. Among these techniques, vectorization emerges as the predominant approach, with this study opting for sparse vectors to represent word vectors, thereby constructing a word vector space to encapsulate text features.

4. Analysis of results of the English Text Reading recommendation model based on collaborative filtering algorithms and FastText classification

4.1. Analysis of the accuracy and F-measure values of the English Text Reading Recommendation Models

Fig. 9 (a, b) shows the comparison results obtained after analyzing the accuracy and F-measurement values of the English text reading recommendation model.

Fig. 9 (a, b) presents a comparative analysis of the accuracy of various English text reading recommendation models. Notably, when the corpus sizes are 10, 30, 50, and 70, the proposed model yields recommendation accuracies of 0.75, 0.79, 0.8, and 0.74, respectively. In contrast, the Top-N algorithm model achieves accuracies of 0.32, 0.54, 0.57, and 0.49, while the matrix factorization model attains accuracies of 0.39, 0.59, 0.56, and 0.65. Moreover, the unimproved FastText classification model exhibits a slightly superior performance compared to the matrix factorization model but significantly lower than the proposed model. Evidently, the proposed model excels in accuracy, offering more precise recommendations compared to its counterparts. It distinctly outperforms alternative algorithms, particularly benefiting long-tail users with personalized recommendations. Furthermore, the F-Measure values for the English text reading recommendation models underscore the superiority of the proposed model. For instance, with corpus sizes of 10, 30, 50, and 70, the proposed model achieves F-Measure results of 0.64, 0.75, 0.79, and 0.81, respectively, surpassing the matrix factorization model's corresponding values of 0.39, 0.47, 0.59, and 0.68. Similarly, the Top-N algorithm model registers F-Measure values of 0.49, 0.58, 0.65, and 0.7, whereas the unimproved FastText classification model exhibits F-Measure values higher than the matrix factorization model but notably lower than the proposed model. Overall, the proposed model exhibits a substantial advantage in terms of F-Measure values, indicating superior recommendation performance tailored to varying corpus sizes of English texts.

Additionally, Fig. 10 presents a detailed comparative analysis of diverse recommendation models for English text reading.

In Fig. 10, the comparative analysis of various recommendation models for English text reading reveals compelling insights into

Table 2
Specific parameter settings.

| Parameters	Value
Optimiser	Adagrad
Learning Rate	0.1
Learning rate update speed	100
Word vector dimension	100
N value in the TOP-N recommendation model	{5, 10, 15, 20}
Number of latent factors in matrix decomposition recommendation model	50
Number of threads in FastText classification model	12
Value of n-gram in FastText classification model	1
Maximum number of iterations of the algorithm	500
Ranking of the ranking function	300
lambda value of the anonymous function	0.1
Matrix parameters of user interest	1
Matrix parameters of user disinterest	0
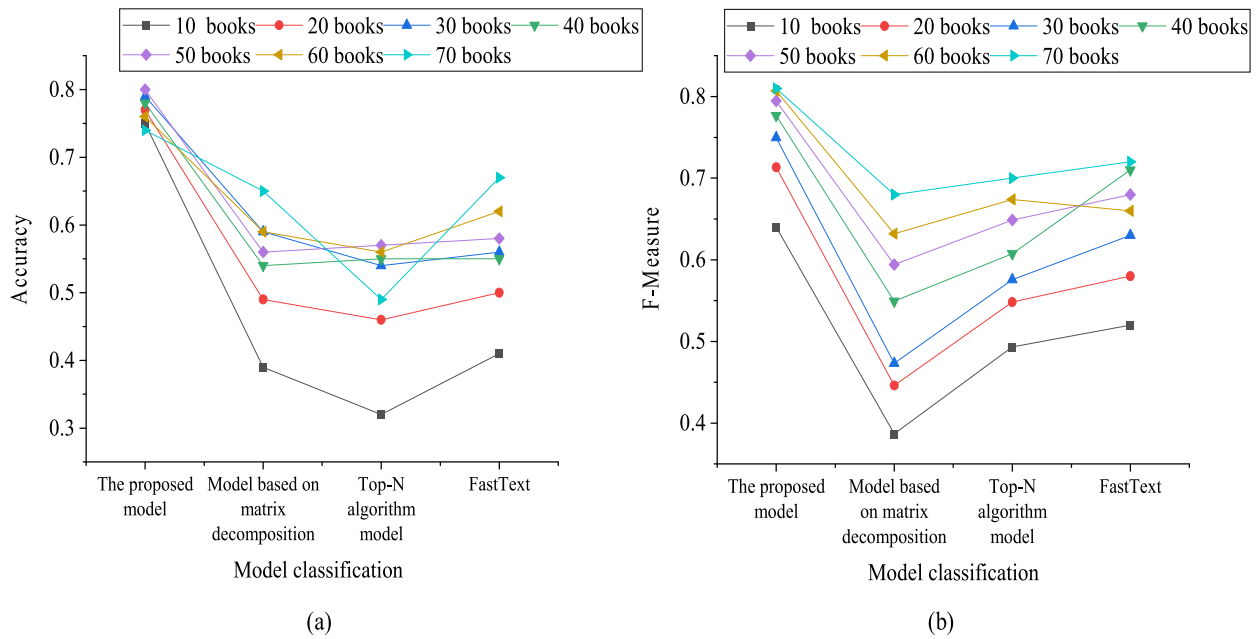
Fig. 9. Analysis of Accuracy and F-Measure Values of English Text Reading Recommendation Models:(a) Analysis of accuracy comparison results; (b) F-Measure value comparison result analysis.
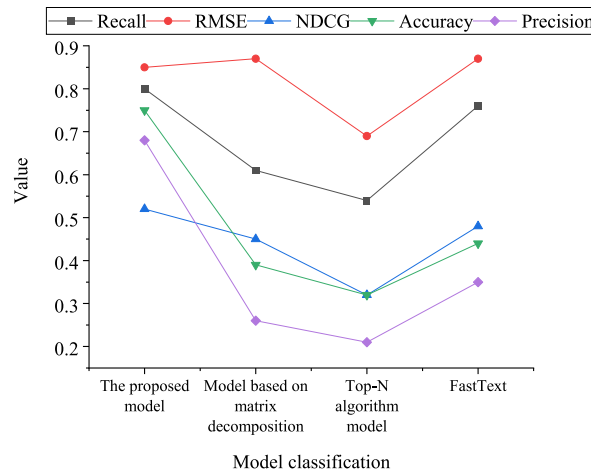

Fig. 10. Comparative analysis of different recommendation models for English text reading.

their performance. The proposed model in this study stands out for its exceptional performance across multiple evaluation metrics. Notably, the recall rate of 0.8 indicates the model's adeptness at capturing 80 % of user preferences accurately. With a Root Mean Squared Error (RMSE) of 0.85, the model demonstrates minimal deviation between predicted and actual values, underscoring its predictive accuracy. The Normalized Discounted Cumulative Gain (NDCG) attains a commendable score of 0.52, suggesting the model's consideration of item relevance in generating recommendation lists, thereby augmenting recommendation quality. Moreover, the model achieves an accuracy of 0.75, indicating its proficiency in predicting user interest accurately, while boasting a precision of 0.68, denoting a high hit rate among recommended items. Conversely, the matrix factorization-based model exhibits subpar performance across metrics such as recall rate, RMSE, NDCG, and accuracy. Similarly, the Top-N algorithm model trails behind in various indicators. Although the FastText model demonstrates respectable performance in recall rate and NDCG, its accuracy and precision fall slightly short compared to the proposed model in this study. Therefore, the model advocated in this study surpasses other comparative models in terms of accuracy and the quality of personalized recommendations. These findings affirm the efficacy of the proposed model in delivering personalized English text reading recommendations tailored for long-tail users.

4.2. Discussion

Long-tail users denote a minority subset within a user cohort with diverse and often niche preferences, posing a challenge for conventional recommendation systems to adequately address. In the realm of English text reading, discerning the reading inclinations of various user segments, particularly long-tail users, is pivotal for delivering tailored recommendations. This study devises an English text reading recommendation framework customized for long-tail users, amalgamating collaborative filtering algorithms with FastText classification techniques. Initially, it elucidates the fusion of collaborative filtering algorithms while leveraging an enhanced Ebbinghaus forgetting curve to compute user interests and forgetting rates across diverse English text categories. Subsequently, it employs an association rule-based recommendation algorithm in tandem with collaborative filtering to furnish insightful recommendations, thereby augmenting recommendation precision and user contentment. Ultimately, the model undergoes comparative analysis against Top-N algorithms and matrix factorization-based models to assess its efficacy and limitations. In contrast to prior research endeavors, this study achieves heightened recommendation accuracy and personalization by tailoring a recommendation framework explicitly for long-tail user cohorts and integrating collaborative filtering algorithms with FastText classification methodologies. Empirical findings underscore the notable superiority of the proposed algorithm over traditional Top-N and matrix factorization-based models in refining the recommendation generation process, as evidenced by enhanced recommendation accuracy and F-Measure metrics. This substantiates the efficacy of the proposed algorithm in elevating recommendation performance and catering to the personalized requisites of long-tail users, thereby ameliorating recommendation precision and user satisfaction within the recommendation ecosystem. Moreover, the study advocates for the application of intelligent recommendation technologies in educational and reading domains, envisaging the provision of superior-quality personalized services to users while furnishing valuable insights for recommendation system development and refinement. Nevertheless, in comparison to antecedent research endeavors, this study falls short in fully exploring the model's complexity and computational overhead. The integration of multiple algorithms may escalate model intricacy and computational demands, potentially compromising practicality and scalability. Oversight of comprehensive discussions on these facets may obscure the model's operational viability in real-world applications.

The primary findings of this study revolve around the development of a bespoke English text reading recommendation model tailored to long-tail users, achieved through the integration of collaborative filtering algorithms and FastText classification methodologies. Through the meticulous amalgamation of diverse algorithms and optimization techniques, this investigation effectively surmounted the inherent constraints of collaborative filtering algorithms and FastText classification methods in personalized recommendations, consequently elevating recommendation system accuracy and user contentment. The principal challenge lay in attaining heightened precision in personalized recommendations for long-tail users. This study adeptly tackles this challenge by augmenting the Ebbinghaus forgetting curve, introducing rule-based recommendation algorithms, and accounting for user reading behavior. Comparative to conventional Top-N algorithmic models and matrix factorization-based counterparts, the proposed model showcases exceptional performance in recommendation accuracy and F-Measure metrics, particularly within long-tail user contexts. The outcomes of this study underscore methodological innovation and substantial advancement in addressing pragmatic challenges within recommendation systems. By pioneering the integration of diverse recommendation algorithms and crafting a personalized recommendation framework tailored to long-tail user cohorts, this study bridges extant research lacunae and furnishes viable remedies for enhancing recommendation system efficacy and user satisfaction. The novelty and scholarly contribution of this investigation lie in its seminal strides toward addressing the inadequacies in meeting long-tail user needs and ameliorating recommendation accuracy within traditional recommendation paradigms, thereby offering robust support and direction for the evolution of personalized recommendation technologies. While this study demonstrates certain advantages in terms of recall rate, root mean square error, NDCG, accuracy, and precision, it nonetheless harbors the following limitations. Firstly, the algorithmic integration process in the study may not fully account for the specific needs and disparities among particular scenarios or user groups, resulting in less precise recommendations for certain user cohorts. For instance, factors such as age groups, interests, and cultural backgrounds might influence users' receptivity to recommended content. Hence, future research should delve deeper into understanding the diverse needs and preferences of various user groups to achieve more precise personalized recommendations. Secondly, the study's utilization of a limited sample size may lead to less accurate and effective calculations of metrics such as interest and forgetting rates. Smaller samples are prone to introducing biases, thus constraining a comprehensive understanding of user behavior. Consequently, future research should expand sample sizes and employ more data to ensure the reliability and robustness of results. Moreover, while the study mentions personalized recommendations for long-tail users, practical understanding of this demographic's needs and preferences remains inadequate. Long-tail users typically exhibit more specialized and diversified interests, necessitating finer recommendation systems to cater to their requirements. Future research should prioritize enhancing comprehension of long-tail users to improve the model's personalized recommendation capabilities towards them. Additionally, although the study mentions the calculation of interest and forgetting rates, there may exist certain limitations in the calculation methods. Future research could explore improved methods to calculate forgetting rates more accurately, such as integrating psychological principles or using more experimental data to validate the model's effectiveness. The study mentions deep learning-based recommendation models but does not delve into them extensively. Future research could consider integrating more advanced recommendation algorithms and technologies, such as deep learning and natural language processing, to further enhance recommendation system performance and accuracy. In terms of study methodology, collaborative filtering algorithms encounter certain challenges in data sparsity and cold start issues. In long-tail user scenarios, traditional collaborative filtering algorithms may fail to provide accurate recommendations due to data sparsity. Additionally, when new users or items join the system, collaborative filtering algorithms face cold start problems, i.e., they cannot provide effective personalized recommendations. Hence, more strategies need to be incorporated into the algorithms to address data sparsity and cold start issues. FastText classification methods are primarily used for text classification tasks, which may have certain limitations in

English text reading recommendations. FastText may not effectively capture the semantic information of English texts, thereby affecting recommendation accuracy. Therefore, future research could combine other algorithms or feature engineering methods to improve recommendation accuracy. In result analysis, for comparing differences between different models, the study employs metrics such as F-Measure. In future research, p-values and confidence intervals could be incorporated to enhance model analysis and more objectively evaluate model effectiveness and generalizability, thereby enhancing research credibility and reliability.

5. Conclusion

This study proposes a personalized English text reading recommendation model specifically tailored for long-tail users, devised through the integration of collaborative filtering algorithms and FastText classification methods. The methodology entails a meticulous examination of collaborative filtering algorithms, leveraging an enhanced Ebbinghaus forgetting curve in conjunction with user reading behavior to compute the percentage of user interest across various types of English texts. Subsequently, the model employs a rule-based recommendation algorithm to establish robust text reading recommendation rules. By amalgamating collaborative filtering algorithms, the model generates recommendations tailored to the distinct characteristics of long-tail user groups, thereby achieving intelligent English text reading recommendations. Through the optimization of the recommendation generation process, the model enhances recommendation accuracy, consequently enhancing the performance and user satisfaction of the English text reading recommendation system. To underscore the advantages of this algorithmic model, the study compares and scrutinizes the performance of Top-N algorithm models, matrix factorization-based algorithm models, and basic FastText classification models in terms of recommendation accuracy and F-Measure. The study findings demonstrate the exceptional performance of the proposed model in recommendation accuracy and F-Measure across varying text quantities. Specifically, when the text quantity is 50, the model attains the highest recommendation accuracy of 0.8, significantly surpassing the Top-N algorithm model's 0.57. Moreover, with a text quantity of 70, both the matrix factorization-based recommendation algorithm model and the basic FastText classification model also achieve the highest recommendation accuracy of 0.65 and 0.67, respectively. As the text quantity increases, the F-Measure of all models gradually improves, with the proposed model reaching the highest final F-Measure of 0.81, surpassing the performance of the other three models. These results underscore significant advantages of the proposed model in providing personalized English text reading recommendations for long-tail users, offering reliable reference and guidance for enhancing the performance of recommendation systems. The practical significance of this study lies in enhancing the accuracy and user satisfaction of the English text reading recommendation system through the development of a personalized recommendation model for long-tail users. This not only introduces novel ideas and methodologies for advancing recommendation system technology but also contributes positively to meeting diverse reading needs, improving recommendation system efficiency, and fostering the development of related fields. Additionally, one of the key findings of this study is that by integrating multi-layer perceptrons and generic matrix factorization with metadata, the accuracy of rating prediction is enhanced, leading to improved precision and user satisfaction of recommendations. In comparison to other methods outlined in the literature review, this approach of combining metadata may offer significant advantages in recommendation accuracy, particularly in addressing data sparsity and cold-start problems. Furthermore, this study identifies certain drawbacks of recommendation systems, such as data sparsity and user rating cold-start problems. However, through the introduction of novel methods such as sentiment analysis of text data and ensemble learning methods, these challenges can be mitigated, thus enhancing the performance and accuracy of recommendation systems. Compared to other methods discussed in the literature review, this comprehensive utilization of diverse technologies may more effectively tackle the challenges encountered by recommendation systems.

Data availability

All data generated or analysed during this research are included in this published article (and its Supplementary Information files).

CRediT authorship contribution statement

Ke Yan: Writing – review & editing, Writing – original draft, Validation, Software, Resources, Methodology, Formal analysis, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e30413.

References

[1] X. Luo, Efficient English text classification using selected machine learning techniques, Alex. Eng. J. 60 (3) (2021) 3401–3409.
[2] K.D. Thompson, I.M. Umansky, W.J. Rew, Improving understanding of English learner education through an expanded analytic framework, Educ. Pol. 37 (5) (2023) 1315–1348.
[3] C. Sardianos, I. Varlamis, C. Chronis, G. Dimitrakopoulos, A. Alsalemi, Y. Himeur, et al., The emergence of explainability of intelligent systems: delivering explainable and personalized recommendations for energy efficiency, Int. J. Intell. Syst. 36 (2) (2021) 656–680.
[4] F. Gurcan, N.E. Cagiltay, Research trends on distance learning: a text mining-based literature review from 2008 to 2018, Interact. Learn. Environ. 31 (2) (2023) 1007–1028.
[5] W. Cherif, A. Madani, M. Kissi, Text categorization based on a new classification by thresholds, Prog. Artif. Intell. 10 (4) (2021) 433–447.
[6] F. El-Alami, S.O. El Alaoui, N.E. Nahnahi, Contextual semantic embeddings based on fine-tuned AraBERT model for Arabic text multi-class categorization, J. King Saud Univ. Comput. Inf. Sci. 34 (10) (2022) 8422–8428.
[7] P. Nitu, J. Coelho, P. Madiraju, Improvising personalized travel recommendation system with recency effects, Big Data Mining Anal. 4 (3) (2021) 139–154.
[8] K.K. Yadav, H.K. Soni, G. Yadav, M. Sharma, Collaborative filtering based hybrid recommendation system using neural network and matrix factorization techniques, Int. J. Intell. Syst. Appl. Eng. 12 (8s) (2024) 695–701.
[9] M. Al-Hassan, B. Abu-Salih, E. Alshdaifat, A. Aloqaily, A. Rodan, An improved fusion-based semantic similarity measure for effective collaborative filtering recommendations, Int. J. Comput. Intell. Syst. 17 (1) (2024) 1–18.
[10] G. Behera, N. Nain, R.K. Soni, Integrating user-side information into matrix factorization to address data sparsity of collaborative filtering, Multimed. Syst. 30 (2) (2024) 64.
[11] Z. Wang, Y. Deng, S. Zhou, Z. Wu, Achieving sustainable development goal 9: a study of enterprise resource optimization based on artificial intelligence algorithms, Resour. Pol. 80 (2023) 103212.
[12] V.R. Yannam, J. Kumar, K.S. Babu, B. Sahoo, Improving group recommendation using deep collaborative filtering approach, Int. J. Inf. Technol. 15 (3) (2023) 1489–1497.
[13] L. Vuong Nguyen, T.H. Nguyen, J.J. Jung, D. Camacho, Extending collaborative filtering recommendation using word embedding: a hybrid approach, Concurrency Comput. Pract. Ex. 35 (16) (2023) e6232.
[14] S. Venkatesan, A recommender system based on matrix factorization techniques using collaborative filtering algorithm, NeuroQuantology 21 (5) (2023) 864.
[15] I. Karabila, N. Darraz, A. El-Ansari, N. Alami, M. El Mallahi, Enhancing collaborative filtering-based recommender system using sentiment analysis, Future Internet 15 (7) (2023) 235.
[16] R. Alatrash, R. Priyadarshini, Fine-grained sentiment-enhanced collaborative filtering-based hybrid recommender system, J. Web Eng. 22 (7) (2023) 983–1035.
[17] R. Mudavath, A. Negi, A multi-tiered solution for personalized baggage item recommendations using FastText and association rule mining: a novel approach 1 (1) (2024) 28.
[18] G. Pilato, F. Persia, M. Ge, T. Chondrogiannis, D. D'Auria, A modular social sensing system for personalized orienteering in the COVID-19 era, ACM Trans. Manag. Inf. Syst. 14 (4) (2023) 1–26.
[19] A.V. Turukmane, N. Alhebaishi, A.M. Alshareef, O.M. Mirza, A. Bhardwaj, B. Singh, Multispectral image analysis for monitoring by IoT based wireless communication using secure locations protocol and classification by deep learning techniques, Optik 271 (2022) 170122.
[20] M. Bayer, M.A. Kaufhold, C. Reuter, A survey on data augmentation for text classification, ACM Comput. Surv. 55 (7) (2022) 1–39.
[21] L.V. Nguyen, Q.T. Vo, T.H. Nguyen, Adaptive KNN-based extended collaborative filtering recommendation services, Big Data and Cognitive Computing 7 (2) (2023) 106.
[22] M. Tr, V. Vinoth Kumar, S.J. Lim, UsCoTc: improved Collaborative Filtering (CFL) recommendation methodology using user confidence, time context with impact factors for performance enhancement, PLoS One 18 (3) (2023) e0282904.
[23] M. Umer, Z. Imtiaz, M. Ahmad, M. Nappi, C. Medaglia, G.S. Choi, et al., Impact of convolutional neural network and FastText embedding on text classification, Multimed. Tool. Appl. 82 (4) (2023) 5569–5585.
[24] D. Qiu, H. Jiang, S. Chen, Fuzzy information retrieval based on continuous bag-of-words model, Symmetry 12 (2) (2020) 225.
[25] O. Dogan, A recommendation system in e-commerce with profit-support fuzzy association rule mining (p-farm), J. Theor. Appl. Electron. Commer. Res. 18 (2) (2023) 831–847.
[26] X. Xia, Learning behavior mining and decision recommendation based on association rules in interactive learning environment, Interact. Learn. Environ. 31 (2) (2023) 593–608.
[27] J.M.J. Murre, A.G. Chessa, Why Ebbinghaus' savings method from 1885 is a very 'pure'measure of memory performance, Psychon. Bull. Rev. 30 (1) (2023) 303–307.
[28] A. Marchand, P. Marx, Automated product recommendations with preference-based explanations, J. Retailing 96 (3) (2020) 328–343.
[29] W.X. Zhao, Z. Lin, Z. Feng, P. Wang, J.R. Wen, A revisiting study of appropriate offline evaluation for top-N recommendation algorithms, ACM Trans. Inf. Syst. 41 (2) (2022) 1–41.
[30] J. Ha, MDMF: predicting miRNA–disease association based on matrix factorization with disease similarity constraint, J. Personalized Med. 12 (6) (2022) 885.