frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 17 October 2024
DOI 10.3389/fphy.2024.1491365

SABTR: semantic analysis-based tourism recommendation

Jiao Li[1,2], Huajian Xue[3,4]*, Qigui Tang[1], Hailiang Wang[5] and Tieliang Gao[1,2]*

[1]Key Laboratory of Data Analysis and Financial Risk Prediction, Xinxiang University, Xinxiang, China, [2]Business School, Xinxiang University, Xinxiang, China, [3]College of Mathematics and Computer Science, Tongling University, Tongling, China, [4]Anhui Engineering Research Center Of Intelligent Manufacturing of Copper-based Materials, Tongling University, Tongling, China, [5]College of Electronic Information and Optical Engineering, Nankai University, Tianjin, China

Check for updates

OPEN ACCESS

EDITED BY
Xuzhen Zhu,
Beijing University of Posts and Telecommunications (BUPT), China

REVIEWED BY
Shimin Cai,
University of Electronic Science and Technology of China, China
Wei Wang,
Chongqing Medical University, China
Run-Ran Liu,
Hangzhou Normal University, China

*CORRESPONDENCE
Huajian Xue,
✉ xuehj@tlu.edu.cn
Tieliang Gao,
✉ chrisgtl2012@163.com

RECEIVED 04 September 2024
ACCEPTED 02 October 2024
PUBLISHED 17 October 2024

CITATION
Li J, Xue H, Tang Q, Wang H and Gao T (2024)
SABTR: semantic analysis-based tourism recommendation.
Front. Phys. 12:1491365.
doi: 10.3389/fphy.2024.1491365

Online tourism spot recommendations, as a key component of tourism services, aim to present travel options that align with users' personal preferences. However, current recommendation systems often underperform due to the sparsity of tourism data and the wide variance in user preferences. To address this challenge, we propose a Semantic Analysis-Based Tourism Recommendation framework, abbreviated as SABTR (Semantic Analysis-Based Tourism Recommendation). The framework comprises two stages: Firstly, Latent Dirichlet Allocation (LDA) models are utilized to deeply mine data between users and attractions, constructing two core matrices: the user similarity matrix and the attraction similarity matrix. Secondly, based on the user similarity matrix, similarity calculation methods are applied to predict ratings for tourism spots that users have not yet evaluated. Simultaneously, within the attraction similarity matrix, probability distributions for each attraction across various thematic interests are calculated. When the system identifies a user's interest in specific types of attractions, SABTR can select a series of related attractions from associated interest tags. Then, these candidate attractions are ranked according to both known and predicted user ratings, ultimately forming personalized attraction packages recommended to users. Extensive experiments have demonstrated that compared to existing tourism recommendation solutions, our method significantly improves the quality of attraction recommendations and enhances user satisfaction.

KEYWORDS

LDA, tourism recommendation, semantic analysis, similarity of users, rating prediction

1 Introduction

With the continuous development of tourism resources and the rapid advancement of information technology, a large amount of tourism resource information can be easily accessed by users through websites and travel applications. However, confronted with a vast array of options for tourist attractions, users often feel confused and hesitant when making choices. To improve the experience of tourists when selecting travel services, various tourism recommendation solutions have been introduced by both industry and academia. These aim to provide better travel experiences and intelligent services for tourists.

Hsieh et al. proposed a Bi-LSTM model in [1] to train on user travel time series data, predicting the migration of users' interest in tourist attractions through adaptively learned parameters. Ma et al. in [2] leveraged differential game theory and Bellman's continuous dynamic programming theory to generate more personalized low-carbon travel plans for tourists, enhancing their environmental awareness and stimulating low-carbon, efficient, and sustainable development within the tourism supply chain. Yao et al. proposed a new Neural Network-enhanced Hidden Markov Structure Time Series Model in [3]. The model uses a neural network for trends and a hidden Markov model with four parts for seasonality: cyclical patterns, unexpected events, event intensity, and random errors. It was tested on US tourism data from 12 countries to suggest travel packages.

These tourism recommendation schemes have promoted the development of the tourism industry and enhanced the overall level of tourism public services. However, these methods do not delve deeply into the characteristics of tourism recommendations. Firstly, user-tourism data is quite sparse, making it difficult for traditional similarity algorithms to uncover the diverse interest distributions of users. Secondly, user ratings for tourism items are also sparse, making it hard to determine users' preferences for tourism items. Lastly, the temporal context of tourists choosing attractions is also a factor that needs to be considered in tourism recommendations. In response to these characteristics of tourism recommendations, we designed a probabilistic semantic analysis-based tourism recommendation algorithm called SABTR. This algorithm can extract user interests from sparse datasets, predict missing ratings for tourism items by tourists, and finally generate a list of tourism items that match tourists' preferences based on their behavior. The proposed SABTR approach can integrate tourists' hidden preferences (such as clicks and favorites) with their direct preferences (such as ratings and likes), ensuring the accuracy of tourists' interests while also ensuring the diversity of user interests. The specific work is as follows:

- We use a semantic analysis model to obtain the distribution of tourist interests by training history records of tourists. Based on this distribution, we design a user similarity algorithm. By aggregating ratings between similar users, we can infer missing tourism item evaluations for users.
- We design an online tourism recommendation scheme. When a tourist clicks on an interesting tourism item, we analyze the interest distribution associated with the item and its ratings to recommend high-rated tourism items that align with the tourist's interests.

The proposed scheme has been extensively tested on experimental datasets, and compared to other baseline algorithms, our method shows better accuracy and recall in tourism recommendations. Diverse interest-based attraction recommendations also provide a better service experience for users. The remaining sections of this paper are organized as follows: Initially, we will provide a review of the existing literature, clearly delineating the differences between the methodologies proposed in this study and those currently employed. Subsequently, we will present a detailed description of the framework of the proposed scheme, explaining how it effectively extracts user interests and predicts missing ratings for tourism projects. Furthermore, we will evaluate the effectiveness and efficiency of the proposed scheme through a series of experiments, summarizing the advantages and shortcomings of the algorithm in the conclusion section, and providing an outlook on future research work.

2 Related works

In this section, we review previous research achievements in tourism recommendation.

Yang et al. conducted an online survey collecting data from 496 users in the Ctrip dataset [4] and performed extensive experiments using Partial Least Squares Structural Equation Modeling (PLS-SEM) on the data. They concluded that perceived personalization, the visual appearance of tours, and the quality of provided travel information can meet users' personalized needs. Gasmi et al. [5] consider itinerary planning and travel recommendations as crucial tasks in tourism personalization. Since tourists are typically unfamiliar with points of interest (POIs) in new cities, They must choose and arrange points of interest (POIs) that suit their preferences, considering factors like starting point and travel time. Researchers suggest using Multi-Objective Evolutionary Algorithms (MOEAs) to find recommendations that balance two goals. Their experimental results on a dataset from Flicker demonstrate the efficiency of the proposed algorithm in generating personalized itinerary recommendation rules, which can help tourists plan their trips in unfamiliar towns. Ding et al., in reference [6], considered the travel itinerary planning problem under a total time constraint and uncertain travel times. This problem requires making a two-stage decision: first, selecting tourist attractions from a set of candidates to maximize the popularity utility for the tourist; second, planning the visiting sequence of these attractions under random travel times to maximize the activity utility for the tourist. Therefore, the paper constructs a two-stage stochastic optimization model with chance constraints for recommending tourist attractions. Compared to the benchmark model, the proposed model improves the recommendation accuracy by nearly 40%. Chen et al., in reference [7], suggest a model called Dynamic Trust Network-based Fuzzy Group Recommendation (DTN-FGR). It turns user ratings into Fuzzy Preference Relations to handle varying evaluation standards. It also uses a PageRank method to calculate user trust scores. This DTN-FGR model shows the best consistency compared to other group recommendation models. Liu et al. [8] suggest using historical check-ins from Location-Based Social Networks (LBSNs) to understand user preferences and boost tourism. A new privacy-focused POI recommendation model is proposed, combining a simplified Graph Convolutional Neural Network (GCN) with user privacy settings. This model offers efficient POI suggestions while safeguarding user privacy. Chen et al. [9] Show through research that metaverse tourism differs from physical travel. Experts say that tailor-made travel choices, socializing, immersive experiences, and getting visitor feedback can significantly improve the travel experience. Ding et al., in reference [10], sought to understand what motivates customers to leave positive or negative feedback. Analyzing over 10,000 Airbnb reviews, researchers used a structural topic model to uncover hidden themes linked to recommendation intentions. They found that positive feedback is mainly driven by the enjoyment of the experience, whereas negative feedback is linked to practical concerns and utilitarian value. Gamidullaeva et al., in reference [11], highlight the importance of combining diverse approaches to create a universal system for recommending travel information when customizing itineraries. The research goal is to introduce a concept for a system that can suggest personalized travel routes. This concept includes processes for gathering and preparing data to create tourism offerings, techniques for tailoring these offerings to individual preferences, and the key steps to put these techniques into action. Chen et al., in reference [12], suggest a framework called GRM-RTrip, which uses graph networks to understand Points of Interest (POIs) from different angles, calculating the chances of moving from one POI to another. This information is then used to predict what users might like. The system treats trip planning like a game, using smart learning to create trips that give the best experience. Tests show it does better than other ways of suggesting trips. Nilashi, in reference [13], claims that Multi-Criteria Collaborative Filtering (MCCF), which considers various product features, offers more dependable and efficient recommendations on shopping sites. The study introduces a new recommendation agent using MCCF to enhance travel site recommendation systems. Extensive testing proves this method can accurately suggest relevant travel options to users, even with limited data.

Majid et al., in reference [14], examined sustainability and tourist involvement as key factors for sustainable development in tourism. They identified 23 AI innovations that could shape future research in this area. The study points out a current shortfall in AI solutions that fully address sustainability and tourist interaction. It also highlights blockchain's potential to revolutionize tourism and hospitality due to its transparency and efficiency. Jain et al. [15] analyzed 56 papers from 2012 to 2022 to uncover gaps in how technology is viewed in tourism. The research summarized key issues and proposed future study paths using a TCM framework. It also positioned tourism as a prime candidate for sustainable virtual investments in the metaverse. Kou et al., in reference [16], explore the application of the Balanced Scorecard for evaluating sustainable investment options in sectors like metaverse tourism. They suggest a hybrid approach combining quantum, spherical, and fuzzy decision-making to prioritize sustainable investment opportunities in the metaverse's tourism sector. Zheng et al. [17] concentrate on disabled tourists who are otherwise capable of traveling, recognizing that tourism could open up new patient-centered care options. The study discusses the challenges of conducting empirical research with tourists who have mental health issues. The paper recommends strategies like setting clear participant criteria, using randomized controlled trials, and adopting comprehensive health research methods. The research could guide tourism management and marketing efforts aimed at these groups. Abbasi-Moud et al. [18] suggested a tourism recommendation system based on user preferences. It starts by gathering user reviews from travel social networks to identify their likes. The reviews are then cleaned up, grouped by topic, and analyzed for sentiment to understand what tourists want. For each point of interest (POI), features are extracted from all the reviews about it. The system then suggests POIs that best match a user's preferences by comparing them semantically. This approach aims to improve on the inaccuracy of standard travel route recommendation algorithms. Esmaeili et al. [19] suggested a social commerce-based hybrid recommendation system to tailor tourist attraction lists to individual tourists, considering their preferences, trust, reputation, social ties, and communities. The method, which factors in multiple elements, was found to be superior to standard collaborative filtering, content-based, and hybrid recommendation techniques in experiments. Cheng et al., in reference [20], suggested an algorithm for recommending travel routes that considers users' interests and the distances between places. It begins by examining users' past travel patterns. The algorithm then determines users' preferences for certain themes and distances based on how long they spend at each attraction. It calculates the best route considering time limits, starting and ending locations. Tests using data from Flickr indicate that this algorithm is more accurate and has better recall than existing methods.

Previous methods identified similar tourism resources by calculating similarity, which could potentially lead to the echo chamber effect, limiting users' ability to discover potential points of interest. In contrast, our proposed algorithm, SABTR (Semantic Analysis Based on User's Behavioral Traces), aims to identify users' interests through semantic analysis. By analyzing users' behaviors such as clicks, favorites, and ratings on tourism resources, the algorithm determines users' preferences for specific types of tourism resources, rather than simply finding similar resources. This approach significantly broadens the scope of users' interests and enhances the accuracy of the tourism recommendation system by employing a rating-based sorting mechanism within similar interest resources, thereby better meeting users' personalized needs.

3 The proposed SABTR method

3.1 The overview of the SABTR

As shown in Figure 1, the proposed SABTR framework includes both an offline training component and an online analysis component. In the offline training phase, records of tourists' visits to tourist attractions are input into the SABTR framework for matrix factorization. The semantic analysis algorithm LDA (Latent Dirichlet Allocation) within SABTR can decompose the tourist-tourist attraction data into two matrices using the Gibbs sampling algorithm [21]: the tourist-interest topic matrix and the interest topic-tourist attraction matrix. In the tourist-interest topic matrix, the distribution of a tourist's interests is considered as the feature vector of the tourist, and then users with similar interests are clustered based on the similarity of these feature vectors. In the interest topic-tourist attraction matrix, for each tourist attraction, the topic distribution is counted and ranked according to the topic probability values.

For the online recommendation phase, when a user clicks on an interesting tourist attraction or rates one, based on the topic distribution of this attraction, several items from each topic are selected and added to the candidate recommendation list according to their topic probabilities. Then, the candidate items are sorted by their predicted ratings, and a suitable recommendation list is generated and sent to the user. The number of interest topics, the length of the recommendation list, and how many tourist attractions are returned for each interest topic will be determined through extensive experimentation in the experimental section.

3.2 Semantic analysis in the training phase

In real-life scenarios, tourists often select travel destinations based on their personal travel preferences. From the perspective of

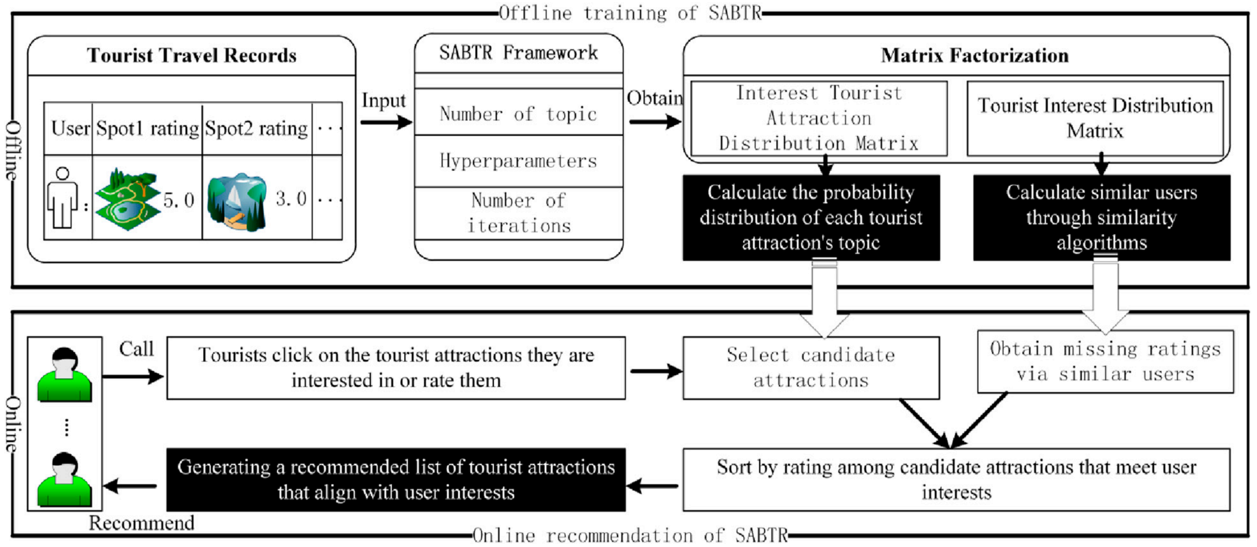
Offline training of SABTR
Offline
Tourist Travel Records
User
Spot1 rating
Spot2 rating
5.0
3.0
Input
SABTR Framework
Number of topic
Hyperparameters
Number of iterations
Obtain
Matrix Factorization
Interest Tourist Attraction Distribution Matrix
Tourist Interest Distribution Matrix
Calculate the probability distribution of each tourist attraction's topic
Calculate similar users through similarity algorithms
Online
Call
Tourists click on the tourist attractions they are interested in or rate them
Select candidate attractions
Obtain missing ratings via similar users
Generating a recommended list of tourist attractions that align with user interests
Sort by rating among candidate attractions that meet user interests
Recommend
Online recommendation of SABTR

FIGURE 1
The Overview of SABTR Scheme. The SABTR scheme includes two stages: offline training and online recommendation. During the offline training phase of the model, a probability vector reflecting tourists' interest preferences was constructed by applying semantic analysis algorithms, and the distribution probability of attractions corresponding to each theme was analyzed. Based on these two vector matrices, the model is capable of predicting the missing ratings for attractions and determining the thematic affiliation of attractions. In the online recommendation phase, the model utilizes the thematic probability distribution of attractions and the predicted missing ratings to generate a list of recommended attractions with rating information for tourists.

probabilistic topic models, the process of tourists choosing attractions can be broken down into two steps: first, tourists pick out themes from a variety of travel topics that interest them; then, they select specific attractions to visit within those themes. The goal of a travel recommendation system is to analyze tourists' historical data, uncover their latent interests, and recommend attractions that align with their preferences.

Since tourists' interests are a latent variable, in our proposed Semantic Analysis-Based Tourist Recommendation system (SABTR), we employ the LDA (Latent Dirichlet Allocation) model to construct tourists' interest themes. LDA is a soft-clustering model that allows data points to be assigned to multiple categories with different probabilities, which means that attractions belonging to the same category share similar latent semantic features. Consequently, attractions with similar semantic features can be recommended to tourists who are interested in these features. Let the set of tourists be denoted as U, the set of themes as Z. In the context of travel recommendations, the themes associated with attractions can be considered as the interests of the users. Let the set of attractions be denoted as S. Let a user's attraction record be represented as a vector $\vec{s}$, and the thematic affiliation of each attraction as a vector $\vec{z}$. Then, the user semantic analysis in the proposed scheme is how to derive the thematic interests $\vec{z}$ of attractions based on the user's attraction record $\vec{s}$, i.e., solving for $p(\vec{z} \mid \vec{s})$. In the scheme, $(\vec{s}, \vec{z})$ is considered as random variables, and the distribution of the variables is shown in the following formula:

$$p\,(\vec{s}, \vec{z} \mid \alpha, \beta) = \prod_{k=1}^{K} \frac{\Delta(\vec{n}_k + \beta)}{\Delta(\beta)} . \prod_{m=1}^{M} \frac{\Delta(\vec{n}_m + \alpha)}{\Delta(\alpha)},\ \vec{n}_m = \left\{\vec{n}_m^{(k)}\right\}_{k=1}^{K} \tag{1}$$

where $\vec{\mathrm{n}}_m$ refers to the m-th tourist's topic distribution, and $\vec{\mathrm{n}}_k$ refers to the distribution of attractions for the k-th topic. $\vec{\mathrm{n}}_m^{(\mathrm{k})}$ represents the number of attractions in the k-th topic of the m-th tourist, and α and β are the hyperparameters of the Dirichlet distribution, while $\Delta(\alpha)$ and $\Delta(\beta)$ are the regularization factors in the Dirichlet distribution.

In the proposed SABTR algorithm, in order to cluster tourists and attractions, it is necessary to solve for $p(\mathrm{z_k}|\mathrm{u_m})$ and $p(\mathrm{s_t}|\mathrm{z_k})$ within the aforementioned probability distribution. $p(\mathrm{z_k}|\mathrm{u_m})$ refers to the probability of the m-th tourist's the k-th topic, which can be represented by θ_{mk}, and $p(\mathrm{s_t}|\mathrm{z_k})$ refers to the probability of the t-th attraction belonging to topic k, which can be represented by ϕ_{kt}. Since the topic interests are latent variables, it is difficult to directly estimate parameters $p(\mathrm{z_k}|\mathrm{u_m})$ and $p(\mathrm{s_t}|\mathrm{z_k})$ using maximum likelihood estimation. Therefore, this paper employs the Gibbs sampling algorithm to estimate these parameters.

In the initial step, each attraction is assigned a random topic, then during the sampling process, the topic transition probability of the target attraction is obtained using the interest distribution of other attractions (excluding the target attraction). Assuming an observed variable for an attraction $s_i = t$, where i = (m, n) is a subscript indicating the travel record of the n-th attraction for tourist $\mathrm{u_m}$. Using Bayes' theorem, we can obtain the sampling expression for the interest of attractions (i.e., the conditional probability of the attractions), as shown in the following formula:

$$\begin{aligned} p\left(\mathrm{z_i}\middle|\vec{\mathrm{z}}_{\neg\mathrm{i}}, \vec{\mathrm{s}}\right) &= \frac{p(\vec{\mathrm{s}}, \vec{\mathrm{z}})}{p(\vec{\mathrm{s}}, \vec{\mathrm{z}}_{\neg\mathrm{i}})} = \frac{p\left(\vec{\mathrm{s}}\middle|\vec{\mathrm{z}}\right)}{p(\vec{s}_{\neg\mathrm{i}}|\vec{\mathrm{z}}_{\neg\mathrm{i}})} . \frac{p(\vec{\mathrm{z}})}{p(\vec{\mathrm{z}}_{\neg\mathrm{i}})} \propto \frac{\Delta(\vec{\mathrm{n}}_z + \beta)}{\Delta(\vec{\mathrm{n}}_{z,\neg\mathrm{i}} + \beta)} . \frac{\Delta(\vec{\mathrm{n}}_m + \alpha)}{\Delta(\vec{\mathrm{n}}_{m,\neg\mathrm{i}} + \alpha)} \\ &= \frac{n_{k,\neg i}^{(\mathrm{t})} + \beta}{\sum_{t=1}^{V}\left(n_{k,\neg i}^{(\mathrm{t})} + \beta\right)} . \frac{n_{\mathrm{m},\neg i}^{(\mathrm{k})} + \alpha}{\left[\sum_{k=1}^{K}\left(n_{\mathrm{m},\neg i}^{(\mathrm{k})} + \alpha\right)\right] - 1} \end{aligned} \tag{2}$$

where $\vec{\mathrm{z}}_{\neg \mathrm{i}}$ represents the current topic setting of all attractions except for attraction s_i, $n_{k,\neg i}^{(\mathrm{t})}$ indicates the number of other attractions (excluding attraction s_i) that have been assigned interest k, and $n_{\mathrm{m},\neg i}^{(\mathrm{k})}$ indicates the number of times attractions other than s_i, which belong to interest k, have been selected by tourists.

The conditional probability of an attraction's interest can be obtained from Equation 2, where in each iteration, every attraction is assigned a new interest through a roulette wheel algorithm. After the model converges, each attraction in every tourist's historical record will be assigned a theme. $\vec{\theta}_m$ refers to the topic distribution of the m-th tourist, and $\vec{\phi}_k$ refers to the attraction distribution of the k-th topic. The distributions of $\vec{\theta}_m$ and $\vec{\phi}_k$ follow the Multinomial distribution, and the prior of these two distribution belong to the Dirichlet distribution. By leveraging the conjugate property of the Dirichlet-Multinomial, it can be deduced that the posterior distributions of $\vec{\theta}_m$ and $\vec{\phi}_k$ follow the Dirichlet distribution. We can obtain the interest distribution of tourists θ_{mk} and the attraction distribution of interests ϕ_{kt} via the expectations of $Dir\left(\vec{\theta}_m | \vec{\mathrm{n}}_{m,\neg \mathrm{i}} + \alpha\right)$ and $Dir\left(\vec{\theta}_m | \vec{\mathrm{n}}_{m,\neg \mathrm{i}} + \alpha\right)$. The expressions are as follows:

$$\begin{aligned} \theta_{mk} &= p\left(\mathrm{z_k}|\mathrm{u_m}\right) = \frac{n_{\mathrm{m},\neg i}^{(\mathrm{k})} + \alpha_k}{\sum_{k=1}^{K}\left(n_{\mathrm{m},\neg i}^{(\mathrm{k})} + \alpha_k\right)} \\ \phi_{kt} &= p\left(\mathrm{s_t}|\mathrm{z_k}\right) = \frac{n_{k,\neg i}^{(\mathrm{t})} + \beta_t}{\sum_{t=1}^{V}\left(n_{k,\neg i}^{(\mathrm{t})} + \beta_t\right)} \end{aligned} \tag{3}$$

3.3 The creation of the tourism recommendation list

When parameters $p\left(\mathrm{z_k}|\mathrm{u_m}\right)$ and $p\left(\mathrm{s_t}|\mathrm{z_k}\right)$ are obtained, candidate attraction selection and similar user selection can be performed. Based on the obtained $p\left(\mathrm{z_k}|\mathrm{u_m}\right)$ the tourist's interest characteristics are transformed into an interest distribution vector. Let the interest distribution vector for the m-th user be denoted as $\bar{u}_m$, then the interest distribution vector for $\bar{u}_m$ is shown in the following formula:

$$\bar{u}_m = \left[p\left(\mathrm{z_1}|\mathrm{u_m}\right), p\left(\mathrm{z_2}|\mathrm{u_m}\right), \ldots p\left(\mathrm{z_k}|\mathrm{u_m}\right)\right] \tag{4}$$

Based on the cosine similarity formula for vectors, the similarity between users can be obtained, as shown in the following formula:

$$Sim\_val\left(u_o, u_i\right) = \mathrm{Cos}\left(\bar{u}_0, \bar{u}_i\right) = \frac{\bar{u}_0 \cdot \bar{u}_i}{\|\bar{u}_0\| \times \|\bar{u}_i\|} \tag{5}$$

where $Sim\_val\left(u_o, u_i\right)$ represents the numerical similarity between the target tourist u_o and other tourists u_i. $\|\bar{u}_0\|$ and $\|\bar{u}_i\|$ represent the magnitudes of the interest vectors for the target tourist and similar tourists, respectively. Tourists are considered valid similar tourists only after their similarity reaches a certain threshold. The condition for the similarity between tourists is shown in the following formula:

$$Sim\left(u_o\right) = \left\{u_i \mid Sim\_val\left(u_o, u_i\right) \geq \mu, u_o \neq u_i\right\} \tag{6}$$

where μ is the threshold for similarity, and $Sim\left(u_o\right)$ represents the similarity that meet the threshold. The ratings of these similar tourists for attractions can be used to predict missing ratings. After selecting similar users for each visitor, based on the attraction ratings from these similar users, the missing rating for the attraction by the tourist can be obtained. Considering the different rating styles of tourists, the prediction formula for the attraction rating is as follows:

$$\hat{r}_{u_o,s_t} = \tilde{u}_o + \frac{\sum_{u_i \in Sim\left(u_o\right)} Sim\_val\left(u_o, u_i\right) \cdot \left(r_{u_i,s_t} - \tilde{u}_i\right)}{\sum_{u_i \in Sim\left(u_o\right)} Sim\_val\left(u_o, u_i\right)} \tag{7}$$

where $\hat{r}_{u_o,s_t}$ indicates the predicted rating for the unrated attraction s_t by the tourist u_o, $\tilde{u}_o$ and $\tilde{u}_i$ represent the average ratings of the attractions by the tourist u_o and the similar tourists u_i,respectively. r_{u_i,s_t} denotes the rating of the attraction s_t by the tourist u_i.

After obtaining the ratings of attractions by tourists through the aforementioned strategy, these ratings can be used to rank the recommended attractions for tourists. When a tourist clicks on an attraction, the SABTR scheme calculates the probability of the theme classification for this attraction based on $p\left(\mathrm{s_t}|\mathrm{z_k}\right)$. Generally, an attraction may belong to multiple themes. Attractions under these themes could all be of interest to the tourist. We select multiple attractions from each theme and sort them according to their predicted ratings. The top-r attractions from the sorted list are then added to the recommendation list. The recommendation list for attractions is shown in the following formula:

$$\mathrm{Re}\left(s_i\right) = \left\{ s_{z_k}^r \middle| N\left(\sum_{r=1}^{R}\sum_{k=1}^{K} s_{z_k}^r\right) = L, \left(s_i, s_{z_k}^r\right) \in \bar{z}_k, \widehat{r}\left(s_{z_k}^1\right) \geq \widehat{r}\left(s_{z_k}^r\right) \right\} \tag{8}$$

where $\mathrm{Re}\left(s_i\right)$ is the recommendation list for the attraction s_i, and $\widehat{r}\left(s_{z_k}^1\right)$ represents the attractions that belong to the topic z_k and have the highest ratings or predicted ratings. The variable r signifies the number of attractions selected from each theme. After an attraction is bookmarked, clicked, or rated by a tourist, semantic analysis is conducted on the attraction to determine the probability of its belonging to certain topics, and then the top r attractions are selected from each theme based on their ratings to be included in the recommendation list. The specific values and value ranges for the aforementioned parameters will be discussed in detail during the experimental phase.

4 Experiments

This section introduces the experimental dataset, evaluation criteria, baseline algorithms, algorithm performance comparison.

4.1 Dataset description

In the experimental phase, the required data includes the IDs of tourists, the tourist attractions they visit, and the ratings given by tourists to these attractions. Previous tourism datasets, such as dataset-tourist-attractions.csv and KG-Rec-Sys-Tourism-SG-main, either only contain information about attractions or have insufficient records of user visits to these attractions. To more effectively validate the proposed solution, this paper adapts the MovieLens (1M) dataset to the tourism recommendation scenario.

The rating.csv file can be used to simulate the rating data of attractions, while the movie.csv file can simulate the record of tourist attraction visits.

4.2 Experimental evaluation criteria

This paper uses Precision, Recall, and F1-measure to evaluate the performance of tourist attraction recommendations, uses RMSE (Root Mean Square Error) to measure the error between predicted and actual attraction ratings, and uses perplexity to assess the performance of semantic analysis models. The evaluation criteria are as follows:

$$
\begin{aligned}
&Perplexity = exp \wedge \left\{ -\left(\sum_{t=1}^{V} log\left(p\left(s_t\right)\right) \right)/(V) \right\} \\
&where \left[p\left(s_t\right) = \sum_{k=1}^{K}\sum_{m=1}^{M} p\left(s_t|z_k\right).p\left(z_k|u_m\right) \right] \\
&Precision\left(s\right) = \frac{N\left(Re\left(s\right) \cap U_{re,\neg s}\right)}{L} \\
&Recall\left(s\right) = \frac{N\left(Re\left(s\right) \cap U_{re,\neg s}\right)}{N\left(U_{re,\neg s}\right)} \\
&F1 - measure\left(s\right) = \frac{2.Precison.Recall}{Precison + Recall} \\
&RMSE = \sqrt{\frac{\sum_{u_i,s_j \in record\left(u_i\left(s\right)\right)} \left(r_{u_i,s_j} - \hat{r}_{u_i,s_j}\right)^2}{N\left[record\left(u_i\left(s\right)\right)\right]}}
\end{aligned}
\tag{9}
$$

Where V represents the total number of attractions, and the lower the perplexity value, the better the model's performance. Precision measures the accuracy of the attraction recommendations, Recall indicates the coverage rate of the attraction recommendations, and the F1-measure is a comprehensive evaluation metric of both precision and recall. $Re(s)$ represents the recommended list generated by the system after the user selects the attraction s. $U_{re,\neg s}$ represents the record of attractions visited by tourists (excluding the currently selected attraction s). r_{u_i,s_j} represents the actual rating of attraction s_j by the tourist u_i, and r_{u_i,s_j} represents the actual rating of attraction s_j by the tourist u_i.

4.3 The baseline algorithms

In this paper, we utilized high-performance experimental equipment, including an Intel Xeon 3206R CPU, 32 GB DDR4 memory, a 2x2 TB RAID hard drive configuration, and an NVIDIA RTX 3090 GPU, to ensure the accuracy and efficiency of the algorithm comparison. We conducted a comparative analysis of the recommendation system performance of the SABTR algorithm proposed in the paper with PLSA [22], LSI [23], and Skip-gram [24] algorithms.

PLSA and LSI, as fundamental topic models, can infer the distribution of users' interests by analyzing their historical records. The Skip-gram algorithm, on the other hand, is a word vector model that can convert attractions into distributed vector representations, and then recommend similar attractions to users by calculating the similarity between vectors.

To comprehensively evaluate the performance of these algorithms, we designed a series of experiments to measure from multiple dimensions, including precision, recall, and F1-score. Through these experiments, we aim to verify the advantages and limitations of the SABTR algorithm compared to existing algorithms in terms of recommendation system performance.

4.4 Parameter settings and performance comparison of the SABTR scheme

When comparing the proposed scheme with the aforementioned baseline algorithms, we first need to determine the optimal parameters for our scheme. For the semantic analysis model in the SABTR scheme (i.e., the LDA model), there are three key parameters: k (representing the number of attraction topics), α and β. In the experiment, we first set the number of attraction topics k to 50 (i.e., k = 50), and set the model's hyperparameters α and β to their default values. Next, we vary the number of model iterations from 500 to 1,350 and calculate the perplexity value after each iteration. By comparing the perplexity values at different numbers of iterations, we can find the optimal number of iterations for model convergence. Finally, we optimize the hyperparameters using the fixed iteration method (i.e., finding the optimal values of α and β while keeping other hyperparameters unchanged and fixing the number of tourist topics). The values of parameters are shown in Figure 2:

Figure 2A shows that when the number of iterations is 1,250, the value of perplexity is 3,957.26, which is the lowest during the iterative training process, indicating that the model has reached a state of convergence. Figure 2B illustrates the change in the model's perplexity value as the number of topic interests increases. It can be observed from the figure that the optimal number of interest categories for the model is 90 when the perplexity value is at its minimum (at this point, the perplexity value is 3,484.1). To find the optimal value of the hyperparameter α for tourist-interest distribution, we fix the number of interest categories (i.e., K = 90) and the value of the hyperparameter for interest-attraction distribution (i.e., β = 1/90), and increase the value of α from 0.010 to 0.024. From the series of perplexity values in Figure 2C, it can be seen that the optimal value of α is 0.022. Finally, increasing the value of β from 0.004 to 0.015, the optimal value of β can be seen in Figure 2D as 0.008. From the values of the hyperparameters, it can be inferred that the distribution of tourists' interests is relatively concentrated, while the topics to which attractions belong are more diverse.

When a tourist shows interest in an attraction, the recommendation system needs to determine the interests associated with the attraction and recommend attractions that match the tourist's interests. In terms of recommendation strategies, we need to focus on the following issues: when an attraction is associated with many themes, how many themes need to be considered to accurately meet the user's needs; among the selected themes, how many attractions should be chosen for each theme to improve the system's recommendation precision, recall

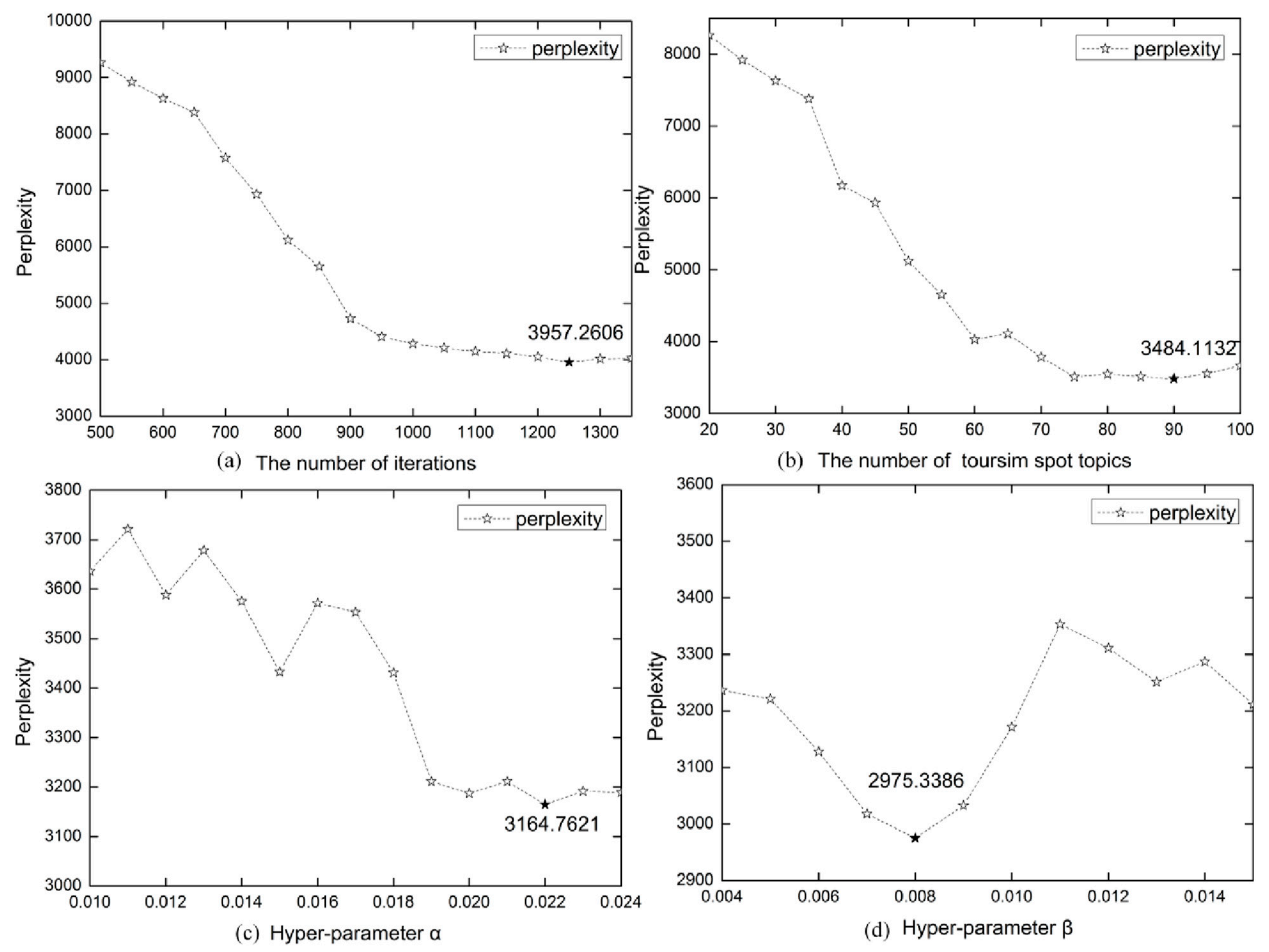

FIGURE 2
Parameters of the Semantic Analysis Mode. Figure 2A illustrates the trend of model performance as the number of iterations increases, with the core objective being to determine the optimal number of iterations for model convergence. Figure 2B presents the variation of model performance with the increase in the number of topics, aiming to find the topic count that yields the optimal model performance. Figures 2C, D respectively explore the impact of changes in model hyperparameters α and β on model performance, with the goal of identifying the optimal values for hyperparameters α and β that maximize model performance. **(A)** The number of iterations. **(B)** The number of tourism spot topics. **(C)** Hyper-parameter α. **(D)** Hyper-parameter β.

rate, and F1-measure. In order to determine the parameters of these recommendation strategies, experiments were conducted by increasing the number of topics and the number of attraction selections to compare the different results of the semantic analysis algorithm, as shown in Figure 3.

Figure 3A primarily analyzes the number of themes to which an attraction belongs. The experiment sets the range of themes from 1 to 10, and when making recommendations, five attractions are selected from each theme to be added to the recommendation list. The number of themes is continuously increased to compare the algorithm's precision, recall, and F1-measure. The results from Figure (a) show that as the number of themes to which an attraction belongs increases, the precision of the recommendations also increases. However, when the number of themes reaches 4, the accuracy of the recommendations begins to decline, and the recall does not improve, indicating that when a user is interested in an attraction, knowing the four main interests to which the attraction belongs can meet the user's needs. After determining the number of interest categories to which an attraction belongs, a series of experiments analyze how many attractions should be selected from each theme to improve the algorithm's performance. We set the number of attractions selected per theme to 5, 10, and 15 to compare the algorithm's performance, and the performance under different parameters is shown in Figures 3B–D. From the three sub-figures, it can be seen that when 10 attractions are selected from a theme, the algorithm has the best precision value, which is 0.2098, and at this time, The appropriate number for the recommendation list is 40.

The method for predicting attraction ratings is to aggregate the ratings of similar users for prediction. We need to determine two parameters: one is the similarity between similar users, and the other is the number of similar users to be selected for rating prediction. The paper first calculates the average similarity between tourists, and then identifies the optimal similarity and the optimal number of similar users based on the calculated Root Mean Square Error (RMSE) values. Next, we rank the candidate attractions based on the predicted ratings to form a recommendation list for tourists to refer to. In addition, we also compared the performance of the recommendation list obtained from the SABTR scheme with the recommendation list without ratings. As shown in Figure 4:

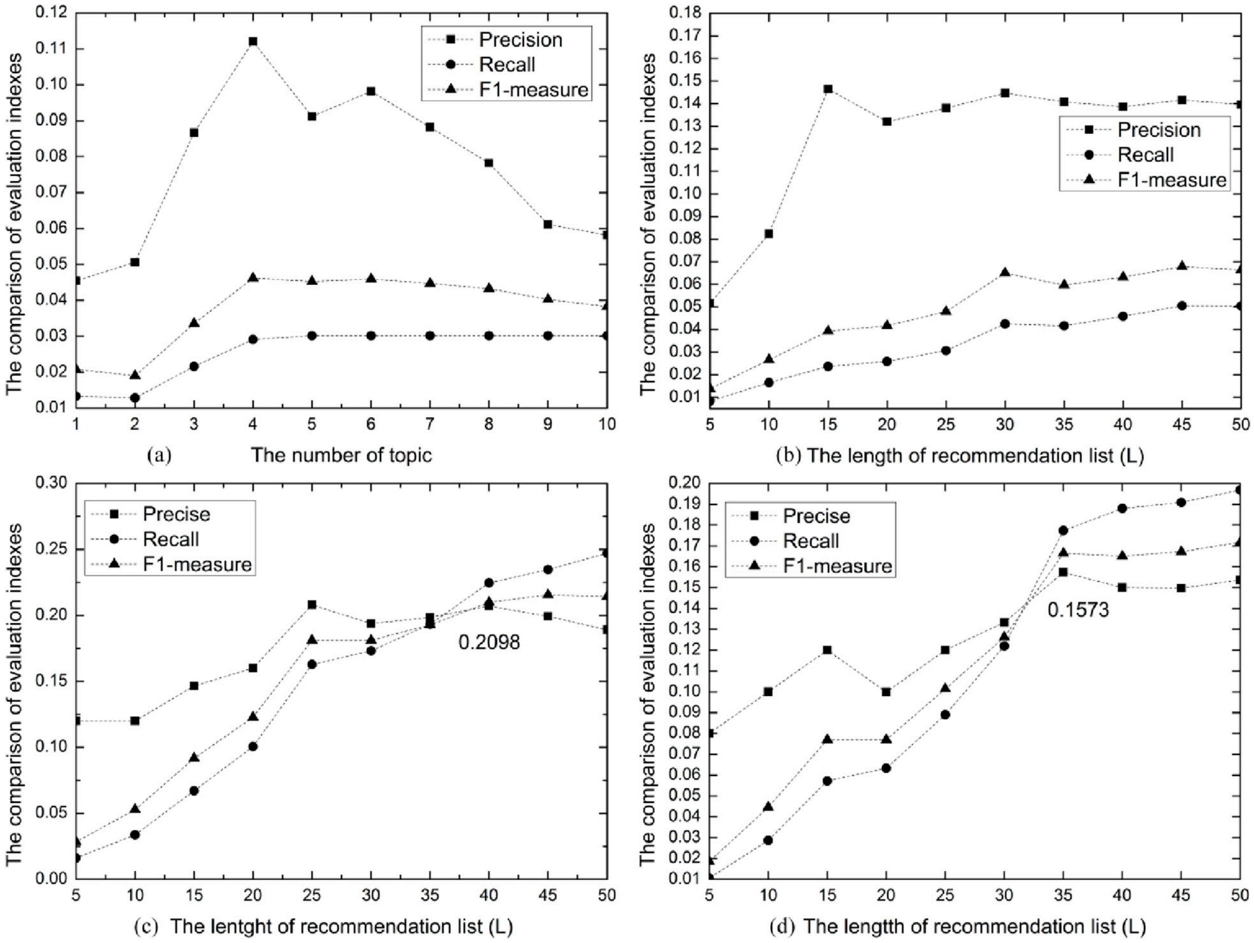

FIGURE 3
Recommendation Strategy of SABTR Approach. Figure 3A illustrates the trend of algorithm performance metrics as the number of themes to which attractions are categorized varies, with the aim of determining the optimal number of themes for achieving the best algorithm performance. Figures 3B–D further explore how algorithm performance fluctuates with the increase or decrease in the number of selected attractions within each specific theme, with the goal of identifying the optimal number of attractions per theme to maximize the overall performance of the algorithm. **(A)** The number of topic. **(B)** The length of recommendation list (L). **(C)** The lentght of recommendation list (L). **(D)** The lengtth of recommendation list (L).

Figure 4A illustrates the relationship between the number of similar users and their similarity for tourists. It can be observed that when the number of similar users is 8, the average similarity of these users exceeds 0.9; however, when the number of similar users increases to 24, the average similarity drops below 0.5. Therefore, the appropriate upper limit for the number of similar users is set to 24. Figures 4B, C use the model's RMSE to determine the optimal number of similar users and the similarity value. The results show that when the number of similar users is set to 12, the model's RMSE value is the lowest at 0.892; simultaneously, setting the similarity to 0.7 yields the best performance in rating prediction, further reducing the RMSE value to 0.861. Figure 4D compares the performance differences of the SABTR approach with and without rating sorting. When the recommendation list length is 10, the precision (precision) of the SABTR approach with rating sorting is 0.21469, while the precision of the SABTR approach without rating sorting is 0.19576, which is 9.6% higher for the former. However, as the recommendation list length increases, the system's precision decreases while the recall rate rises. When the recommendation list length reaches 45, the precision of the SABTR approach without rating sorting is 0.14675, slightly higher than the precision of the SABTR approach with rating sorting (0.14923). This indicates that including attractions with lower ratings in the recommendation list may reduce recommendation effectiveness.

After determining the optimal parameters for the SABTR approach, we compared its performance with other baseline algorithms. In this approach, 90 topics were selected, and the hyperparameters were set to 0.022 and 0.008, respectively. The length of the recommendation list was set to 40, with 4 topics chosen and 10 attractions selected within each topic. We divided the dataset into a training set and a testing set, where the proportion of the training set gradually increased from 30% to 90%, and correspondingly, the testing set proportion decreased from 70% to 10%. The performance comparison results of these algorithms are described in Table 1.

The data in Table 1 show that when the data density does not exceed 50%, the proposed SABTR method outperforms PLSA, Skip-Gram, and LSA. Especially when the data density is low (such as 30%), the precision of the SABTR algorithm is 22% higher than that of Skip-Gram. As the data density increases, the performance of

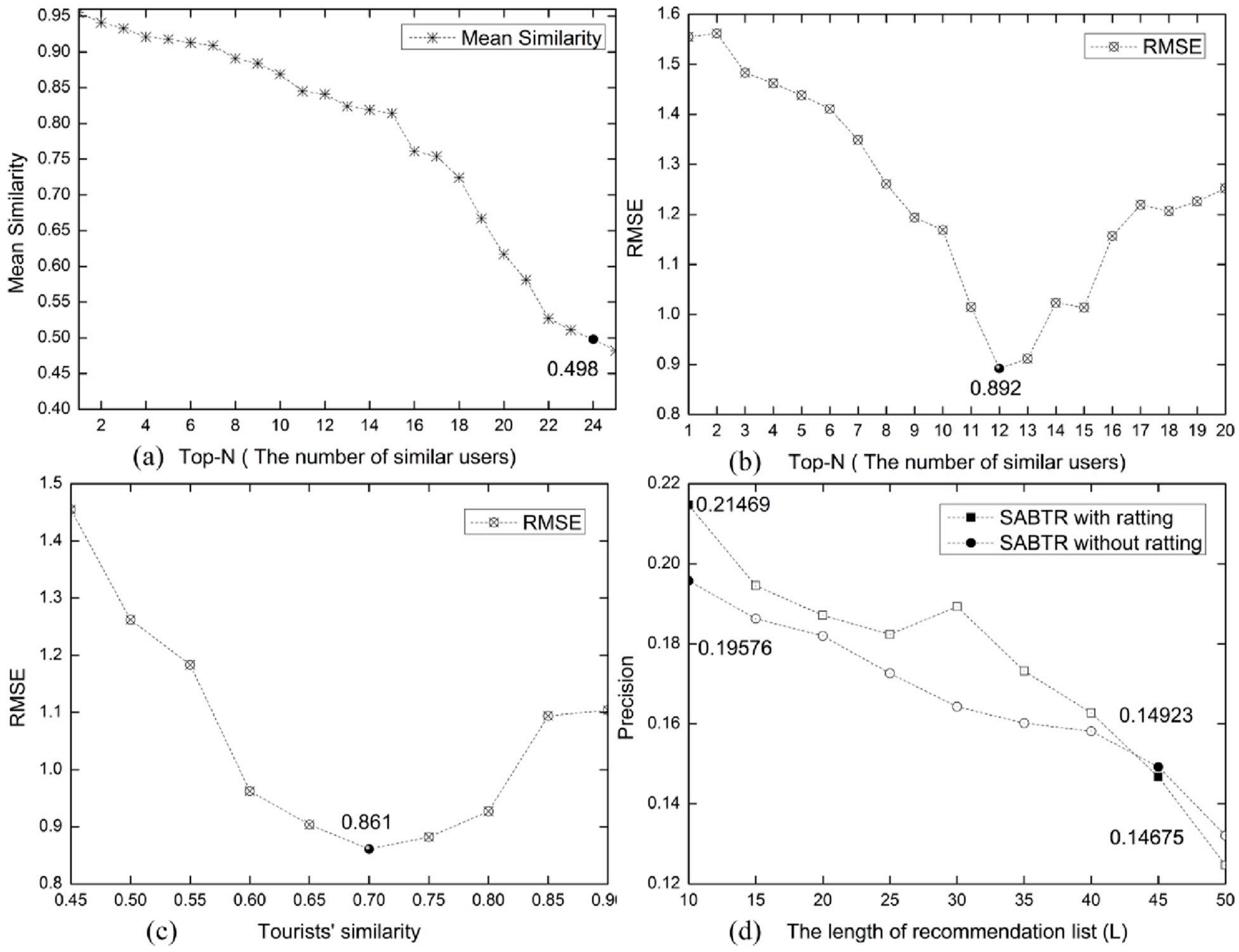

FIGURE 4
Parameters of the attraction rating prediction scheme and the ordering of attraction ratings in recommendations. Figure 4A describes the distribution of average similarity among similar users. Figure 4B explores the impact of the number of similar users on the Root Mean Square Error (RMSE). Figure 4C analyzes the effect of the similarity between tourists on the Root Mean Square Error (RMSE). Figure 4D compares the performance of the interest recommendation list generated by the SABTR method with that of a recommendation list without rating information.

both SABTR and the baseline algorithms improves; however, when the data density exceeds 50%, the precision of Skip-Gram surpasses that of SABTR, particularly when the data density reaches 90%, at which point the precision of Skip-Gram reaches 0.2502, an 18% increase compared to SABTR.

While Skip-Gram, as a word vector model, can find similar attractions by converting them into distributed vectors and using vector similarity, this does not necessarily mean it provides a better service experience for tourists. This is because it tends to find the most similar attractions, potentially leading tourists into an information echo chamber and causing interest fatigue. In contrast, SABTR analyzes the interest topic distribution of attractions through a topic model, helping to expand tourists' interests and meet their diverse needs. Especially in cases of insufficient data (e.g., when data density is 30%), SABTR performs best, effectively alleviating the cold start problem in recommendation systems, whereas other algorithms (Skip-Gram, PLSA, and LSA) exhibit overfitting in recommendations.

When the data density is 90%, the recommendation precision of the PLSA algorithm is 0.2050, close to that of SABTR (which has a recommendation precision of 0.2107), indicating that PLSA can also provide good recommendation performance when there is sufficient data. In contrast, LSA, due to the negative values in the interest factors it extracts, cannot effectively cluster attractions and tourists, performing the worst across four different data densities.

5 Conclusion and future work

In this paper, we propose a tourism recommendation scheme based on semantic analysis, aimed at recommending suitable attractions to tourists. The scheme primarily leverages semantic topic modeling for user clustering and attraction clustering. When a user expresses a preference for a particular attraction on a travel service website, other attractions similar to it enter the recommendation candidate list. Subsequently, the ratings for these candidate attractions are calculated based on the ratings given by other users who share similarities with this user. After ranking the candidate attractions according to their scores, a list of attractions tailored to the user's interests is generated and sent to the

TABLE 1 Performance comparison between SABTR and baseline algorithms.

Evaluation Metrics	Methods	Matrix Density = 30%	Matrix Density = 50%	Matrix Density = 70%	Matrix Density = 90%
Precision	SABTR	**0.1137**	**0.1412**	0.1871	0.2107
	Skip-Gram	0.0927	0.1238	**0.2325**	**0.2502**
	PLSA	0.0943	0.1134	0.1612	0.2050
	LSI	0.0794	0.0986	0.1211	0.1413
Recall	SABTR	**0.1211**	**0.1537**	0.2045	0.2247
	Skip-Gram	0.0836	0.1325	**0.2258**	**0.2487**
	PLSA	0.1132	0.1224	0.1724	0.2106
	LSI	0.0886	0.0971	0.1518	0.1753
F1-measure	SABTR	**0.1173**	**0.1471**	0.1954	0.2175
	Skip-Gram	0.0879	0.1280	**0.2291**	**0.2494**
	PLSA	0.1028	0.1177	0.1666	0.2078
	LSI	0.0984	0.0978	0.1347	0.1565

The bolded values in the table denote the most outstanding results in the performance comparison of different algorithms.

user. Experimental results demonstrate that the proposed scheme not only improves the accuracy and recall of recommendations but also saves tourists time in selecting travel resources, thereby enhancing the user's service experience.

The method we employ requires the use of tourists' travel records and rating data. Given the increasing emphasis on privacy concerns, future recommendation systems will also place greater importance on protecting user information. Therefore, in future work, we plan to adopt a federated learning framework, where instead of directly using individual user records, gradients provided by users will be utilized to analyze their interests. This approach allows us to recommend appropriate attractions while safeguarding user privacy.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://grouplens.org/datasets/movielens/.

Author contributions

JL: Supervision, Writing–review and editing. HX: Software, Writing–original draft, Writing–review and editing. QT: Investigation, Writing–review and editing. HW: Writing–review and editing. TG: Formal Analysis, Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. 1. "Research on the Teaching Model of Deep Integration of Industry and Education in E-commerce Major Based on the OBE Concept", Henan Province Undergraduate Colleges and Universities' 2023 Annual Industry-Education Integration Project. 2. "Research on the Teaching Reform of "Software Development and Practice" Course Based on OBE Education Concept under the Ecology of Open Source Software (2023xj022)". 3. The project of enterprise temporary job of Science and engineering teachers from colleges and universities of Anhui Province (2024jsqygz104). 4. The Project of Artificial intelligence serving the characteristics of ten emerging industries of Anhui province (2023sdxx078).

Acknowledgments

In this article, we employed an AI large model to scrutinize the grammar of the text. We extend our gratitude to Moonshot AI for providing the "Kimi" artificial intelligence technology, version 1.0, 2024. For more information about this technology, you can visit the website: https://kimi.moonshot.cn.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Hsieh SC. Tourism demand forecasting based on an LSTM network and its variants. *Algorithms* (2021) 14(8):243. doi:10.3390/a14080243

2. Ma D, Hu J, Yao F. Big data empowering low-carbon smart tourism study on low-carbon tourism O2Osupply chain considering consumer behaviors and corporate altruistic preferences. *Comput and Ind Eng* (2021) 153:107061. doi:10.1016/j.cie.2020.107061

3. Yao Y, Cao Y. A Neural network enhanced hidden Markov model for tourism demand forecasting. *Appl Soft Comput* (2020) 94:106465. doi:10.1016/j.asoc.2020.106465

4. Yang X, Zhang L, Feng Z. Personalized tourism recommendations and the E-tourism user experience. *J Trav Res* (2024) 63(5):1183–200. doi:10.1177/00472875231187332

5. Gasmi I, Soui M, Barhoumi K, Abed M. Recommendation rules to personalize itineraries for tourists in an unfamiliar city. *Appl Soft Comput* (2024) 150:111084. doi:10.1016/j.asoc.2023.111084

6. Ding Y, Zhang L, Huang C, Ge R. Two-stage travel itinerary recommendation optimization model considering stochastic traffic time. *Expert Syst Appl* (2024) 237:121536. doi:10.1016/j.eswa.2023.121536

7. Chen S, Tong J, Chen J. Collective tourist destination recommendation: a dynamic trust network-based fuzzy decision-making model. *Int J Fuzzy Syst* (2024) 1–17. doi:10.1007/s40815-024-01797-x

8. Liu Y, Zhou X, Kou H, Zhao Y, Xu X, Zhang X, et al. Privacy-preserving point-of-interest recommendation based on simplified graph convolutional network for geological traveling. *ACM Trans Intell Syst Technology* (2024) 15(4):1–17. doi:10.1145/3620677

9. Chen Z. Beyond boundaries: exploring the Metaverse in tourism. *Int J Contemp Hospitality Manage* (2024). doi:10.1108/ijchm-06-2023-0900

10. Ding K, Gong XY, Huang T, Choo WC. Recommend or not: a comparative analysis of customer reviews to uncover factors influencing explicit online recommendation behavior in peer-to-peer accommodation. *Eur Res Manage Business Econ* (2024) 30(1):100236. doi:10.1016/j.iedeen.2023.100236

11. Gamidullaeva L, Finogeev A, Kataev M, Bulysheva L. A design concept for a tourism recommender system for regional development. *Algorithms* (2023) 16(1):58. doi:10.3390/a16010058

12. Chen L, Cao J, Tao H, Wu J. Trip reinforcement recommendation with graph-based representation learning. *ACM Trans Knowledge Discov Data* (2023) 17(4):1–20. doi:10.1145/3564609

13. Nilashi M, Abumalloh RA, Samad S, Minaei-Bidgoli B, Thi HH, Alghamdi OA, et al. The impact of multi-criteria ratings in social networking sites on the performance of online recommendation agents. *Telematics Inform* (2023) 76:101919. doi:10.1016/j.tele.2022.101919

14. Majid GM, Tussyadiah I, Kim YR, Pal A. Intelligent automation for sustainable tourism: a systematic review. *J Sust Tourism* (2023) 31(11):2421–40. doi:10.1080/09669582.2023.2246681

15. Jain P, Singh RK, Mishra R, Rana NP. Emerging dimensions of blockchain application in tourism and hospitality sector: a systematic literature review. *J Hospitality Marketing and Manage* (2023) 32(4):454–76. doi:10.1080/19368623.2023.2184440

16. Kou G, Yüksel S, Dinçer H. A facial expression and expert recommendation fuzzy decision-making approach for sustainable business investments within the metaverse world. *Appl Soft Comput* (2023) 148:110849. doi:10.1016/j.asoc.2023.110849

17. Zheng D, Wen J, Kozak M, Phau I, Hou H, Wang W. Vulnerable populations with psychological disorders in tourism: methodological challenges and recommended solutions for empirical research. *Tourism Manage* (2023) 98(8):104760. doi:10.1016/j.tourman.2023.104760

18. Abbasi-Moud Z, Vahdat-Nejad H, Sadri J. Tourism recommendation system based on semantic clustering and sentiment analysis. *Expert Syst Appl* (2021) 167:114324. doi:10.1016/j.eswa.2020.114324

19. Esmaeili L, Mardani S, Golpayegani SAH, Madar ZZ. A novel tourism recommender system in the context of social commerce. *Expert Syst Appl* (2020) 149:113301. doi:10.1016/j.eswa.2020.113301

20. Cheng X. A travel route recommendation algorithm based on interest theme and distance matching. *EURASIP J Adv Signal Process* (2021) 2021(1):57. doi:10.1186/s13634-021-00759-x

21. Shim C, Vo BT, Vo BN, Ong J, Moratuwage D. Linear complexity Gibbs sampling for generalized labeled multi-Bernoulli filtering. *IEEE Trans Signal Process* (2023) 71:1981–94. doi:10.1109/tsp.2023.3277220

22. Duan L, Gao T, Ni W, Wang W. A hybrid intelligent service recommendation by latent semantics and explicit ratings. *Int J Intell Syst* (2021) 36(12):7867–94. doi:10.1002/int.22612

23. Gao T, Cheng B, Chen J, Chen M. Enhancing collaborative filtering via topic model integrated uniform euclidean distance. *China Commun* (2017) 14(11):48–58. doi:10.1109/cc.2017.8233650

24. Gao T, Duan L, Feng L, Ni W, Sheng QZ. A novel blockchain-based responsible recommendation system for service process creation and recommendation. *ACM Trans Intell Syst Technology* (2024) 15:1–24. doi:10.1145/3643858