Tourist Attraction Recommendation Model Based on RFPAP-NNPAP Algorithm

Jun Li
Department of Tourism Management, Zhengzhou Tourism College, Zhengzhou, 450000, China

***Abstract*—Driven by globalization and digitization, the tourism industry is facing new challenges and opportunities brought about by big data and artificial intelligence. The recommendation of tourist attractions, as an important part of the industry, has a direct influence on the tourist experience. However, with the diversification and personalization of tourism demand, traditional recommendation methods have shown shortcomings: weak processing ability for complex nonlinear data, affecting recommendation accuracy and personalization, and insufficient efficiency and stability when processing large-scale data. Faced with this challenge, this study proposed a hybrid tourist attraction recommendation model with random forest, artificial neural network, and frequent pattern growth. This model utilized the powerful classification and regression capabilities of random forests, as well as the complex nonlinear mapping ability of artificial neural networks, to predict tourist attraction preferences. And on this basis, the frequent pattern growth algorithm was introduced to mine the associated attractions of tourist preferences, thereby achieving accurate recommendation of tourist attractions. In experimental verification, the proposed model demonstrated superior performance. It not only surpassed traditional tourist attraction recommendation methods in accuracy and personalization, but also exhibited efficient and stable characteristics when processing large-scale data. After about 16 iterations, the MAPE value of the mixed model decreased to 0.44%. After about 39 iterations, the MAPE value of the mixed model decreased to 0.40%. The average accuracy, recall rate and F-value of the proposed model are 92.26%, 82.11% and 84.43%, respectively, which are superior to the comparison algorithm. Its error correction accuracy fluctuates around 90%. This study provides a new solution to the problem of personalized recommendation of tourist attractions, providing theoretical guidance for the tourism applications of random forests and artificial neural networks, and improving the tourist experience, promoting the development of the tourism industry.**

***Keywords*—Tourist attractions; recommendation model; RF; ANN; FP-Growth**

I. INTRODUCTION

In the context of globalization and digitization, the tourism industry is undergoing unprecedented development and changes. The information technology growth, especially the popularization of big data and artificial intelligence, has provided new possibilities and challenges for the growth of the tourism industry [1]. Among them, Tourist Attraction Recommendation (TAR) is a crucial part of the tourism industry, which directly affects the experience and satisfaction of tourists. However, with the diversification of tourist destinations and the increasing demand for personalization, traditional TAR methods cannot meet the demands of tourists [2].

The existing TAR algorithms mainly have two major problems: firstly, their ability to handle complex and nonlinear data patterns is insufficient, which affects the accuracy and personalization of recommendations. Secondly, when processing large-scale data, the computational efficiency and stability of algorithms are insufficient, making it difficult to meet the needs of the big data era [3]. Faced with this challenge, how to provide accurate and personalized tourist TAR has become a focus of attention in academia and industry. In recent years, advanced machine learning technologies such as Random Forest (RF), Artificial Neural Network (ANN), and Frequent Pattern Growth (FP-Growth) have achieved significant results in many fields, including tourism recommendations [4-5]. However, research that combines the three, especially in the field of TAR, has not yet existed.

In order to enhance the ability of tourist attraction recommendation system to process complex data and better meet the individual needs of tourists, this study combines RF, ANN and FP-Growth to propose a hybrid TAR algorithm. The research aims to improve the accuracy, computational efficiency and personalized experience of travel recommendations to meet the development needs of the modern tourism market. The advantages of this method are that by combining the advantages of the three algorithms, it not only optimizes the ability to process large-scale complex data, improves the accuracy and efficiency of the recommendation system, but also enhances the response ability to the personalized needs of tourists, thus significantly improving the satisfaction of tourists and promoting the innovation and development of the tourism industry.

The innovation of this study contains the following aspects: for the first time, the combination of RF, ANN, and FP-Growth is applied to TAR. Aiming at the disadvantage that RF may overly rely on certain features when processing a large number of features, the ANN algorithm is introduced to construct a more accurate tourist preference attraction prediction model with better predictive performance. In response to the drawbacks of FP-Growth, which consumes a lot of computation time and suffers from memory overflow, parameter optimization is carried out on its support and confidence.

The main contribution of the research is that the proposed hybrid tourist attraction recommendation algorithm integrating RF, ANN and FP Growth can effectively solve the shortcomings of traditional algorithms in personalized

recommendation and big data environment by optimizing the ability of the algorithm to process complex data and improving the computational efficiency. In addition, by combining the advantages of different algorithms, the new model has significant advantages in improving the accuracy and response speed of the recommendation system, which can provide a more efficient and personalized tourist attraction recommendation solution for the tourism industry, thus enhancing the experience and satisfaction of tourists, and promoting the sustainable development of the tourism industry.

The study is divided into six sections: Literature review IN Section II discusses existing technologies. Methodology in Section III used in the research. The proposed model is experimentally validated in Section IV. Discussion is given in Section V and finally, Section VI concludes the paper.

II. Literature Review

The intelligent recommendation function has been widely applied in the selection of tourist attractions, and its practicality in daily life is significant. In recent years, many researchers have also made important contributions to the development of this field. Researchers such as C. Si conducted an in-depth study on TAR models based on vehicle movement data. This study adopted an advanced prediction model, which is unique in that it utilizes tensor decomposition technology to predict and process possible missing values, greatly improving the accuracy of recommendations. Tourists can obtain a more satisfactory travel experience, thereby improving the overall quality of tourism to some extent [6]. Scholar R H ö singer et al. proposed an innovative model called TR-DNNMF to provide TAR to users. The matrix factorization model is mainly responsible for handling the linear part in this model, which can cut down the complexity of the data and enable the model to more accurately grasp the linear relationship between different scenic spots. Meanwhile, deep neural network models are responsible for handling the nonlinear part, revealing the deep level features and patterns of each attraction. This model can not only accurately recommend known attractions, but also discover and recommend some new attractions that have not been discovered by the public, providing users with a richer and more personalized travel experience [7]. Scholar L Wen et al. conducted in-depth research on the radial basis function (RBF) neural network algorithm and successfully constructed an accurate model that can predict popular tourist attractions using advanced parameter optimization techniques. The model uses complex parameter optimization techniques to finely adjust the parameters of the RBF neural network, greatly improving the prediction accuracy and effectiveness of the model. The predictive ability of the model provides great convenience for the tourism industry, allowing tourists to plan and prepare in advance [8]. B. Cao et al. developed a context aware personalized recommendation model for mobile tourism e-commerce, with the main goal of addressing the sparsity and low accuracy issues encountered by current recommendation models in personalized recommendation data. The construction of this model is based on situational awareness technology, which can understand and analyze the user's current actual environment. By accurately understanding and grasping these contextual information, the model can better understand the needs and preferences of users, thereby providing more personalized recommendations [9]. Y. Zhang et al. put forward a new recommendation system method that fused human Particle Swarm Optimization (PSO) with fuzzy Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) method, mainly used for recommendation systems in the tourism industry. To address the potential inefficiency of PSO in dealing with complex decision-making problems, the fuzzy TOPSIS method was introduced into this system to effectively handle various uncertain factors in tourism recommendations. The performance verification results showed that this new recommendation method performed well in practical applications, not only improving the accuracy of recommendations, but also improving the efficiency of recommendations [10].

RF improves the accuracy and stability of predictions by building a large amount of decision trees and voting or averaging their prediction results. RF can handle a large number of input variables, effectively prevent overfitting, and can be utilized for regression and classification issues, making it widely used in fields such as financial prediction and medical diagnosis. Numerous scholars have proposed improvements to make it more applicable to the field of study. A. Hill and other researchers used the RF model to predict severe weather. This study selected the spatiotemporal evolution simulated near the prediction point throughout the entire prediction period as the input variables for the model, which includes a series of climate parameters such as temperature, humidity, wind speed, pressure, etc. By using these input variables with temporal and spatial dynamic changes, weather changes can be described and predicted more comprehensively and accurately. After training with the RF model, the experiment outcomes indicated that the use of the RF model can effectively raise the prediction of severe weather throughout the entire prediction period [11]. J. Yoon proposed a unique method for predicting real GDP growth using machine learning models. This study mainly focused on Japan's real GDP growth and conducted predictive analysis of data from 2001 to 2018. The research results indicated that between 2001 and 2018, the prediction accuracy of the fusion model exceeded the benchmark prediction, mainly due to the powerful performance of the model, which can capture and learn a large number of complex nonlinear relationships, thereby improving the accuracy of prediction [12]. S. Chen and other researchers were committed to improving the accuracy of runoff prediction for cascade hydropower stations and have chosen the RFR model for modeling medium and long-term runoff prediction. To ensure the fairness of the results, the researchers compared the prediction results of the RFR model with those of Support Vector Machine (SVM) and Integrated Autoregressive Moving Average Model (IARMA). Through comparison, it was found that the Mean Square Error (MSE) of the RFR model was the smallest, which proves that it has better prediction accuracy than other models, and has higher reliability and practicality [13]. T. Wang et al. innovatively combined RF with Bayesian optimization techniques for quality prediction of large-scale dimensional data. The model first selects the key factors that affect production through information gain, and then applies

sensitivity analysis to maintain the stability of product quality. The experimental results showed that a small number of key features processed through RF Bayesian optimization can significantly reduce computational time while ensuring prediction accuracy, thus having good cost-effectiveness. This provides a new perspective and operational strategy for product quality prediction and control in the process industry [14]. Y. Shi and other scholars have innovatively proposed a prediction model based on Genetic Particle Swarm Optimization Algorithm (GAPSO) and RF regression (RFR) to raise the accuracy of prediction and effectively reduce the losses of flood disasters for predicting mine water inrush. The experiment iteratively trained 34 samples to find the optimal parameters. After testing, the outcomes have proved the merits of the GAPSO-RFR model in improving prediction accuracy and reducing generalization errors, providing strong technical support for the prevention of mine water inrush disasters [15].

In summary, current TAR models have shown certain shortcomings in accuracy and personalization, such as data sparsity issues and challenges in handling complex decisions and nonlinear relationships. RF has certain applicability in this field, as it can learn and reveal complex nonlinear relationships, effectively improve the accuracy and stability of recommendations, and is expected to provide new solutions for TARs. Therefore, this study innovatively raised a hybrid model based on RF, ANN, and FP-Growth to achieve more accurate TAR results.

III. METHODS

In order to accurately predict and recommend the top attractions for tourists, this section first combines RF and MLP models. After that, in order to further strengthen the ability of mining the data related to scenic spots based on tourists' preferences, the parameters of the FP-Growth algorithm were introduced and optimized. In this process, the adjustment of FP-Growth algorithm is mainly aimed at the automatic setting of support and confidence, so as to improve the efficiency and accuracy of data processing. Finally, these three technical means are integrated to form a TAR model using hybrid algorithms, which can achieve a deep understanding of tourist behavior and preferences, provide support for the tourism industry, and promote the development of personalized tourism services.

A. Construction of a Tourist Preference Attraction Prediction Model Based on RFPAP-NNPAP Algorithm

The prediction model for tourist attraction preferences combines research results from multiple disciplines such as big data, artificial intelligence, sociology, and psychology, which is significant for the tourism industry growth [16]. Research in this field mainly focuses on predicting tourist preferences for different tourist attractions. The demand and preferences of tourists continue to change, requiring predictive models to have adaptability and flexibility [17]. For this purpose, the study adopts two machine learning algorithms, RF and ANN, to integrate and construct a prediction model for tourist preference for scenic spots. RF is an ensemble learning method that creates multiple decision trees and combines their outputs to obtain accurate and stable prediction results [18]. ANN can address nonlinear problems and learn and extract deep level features from data. By integrating these two algorithms, the effectiveness of the prediction model can be improved and have a positive impact. This will help tourism enterprises to effectively position themselves in the market, design products, and optimize services, providing tourists with a more personalized and satisfactory travel experience [19]. RF is composed of numerous CART trees, which improve classification accuracy by integrating multiple decision results. The implementation steps include: using Bootstrap sampling method to extract k training sets with replacement from the original data, constructing k trees, and generating k out of bag data. m features at each node are randomly selected, and the feature with the strongest classification is selected. A threshold is set, and no pruning. Multiple trees are combined to form an RF, and the classification result of the new data is determined by the voting of the tree classifier [20]. It assumes that there are n tourists, each with p features, a matrix of $n \times p$ can be formed, as shown in Eq. (1).

$$A = \begin{bmatrix} a_{1f_1} & a_{1f_1} & \cdots & a_{1f_p} \\ a_{2f_1} & a_{2f_2} & \cdots & a_{2f_p} \\ \vdots & \vdots & \cdots & \vdots \\ a_{nf_1} & a_{nf_1} & \cdots & a_{nf_p} \end{bmatrix} \quad (1)$$

In Eq. (1), $f_1, f_2, \ldots, f_p$ means the selected p factors. a_{if_j} means the measured value of the j th characteristic factor of the i th tourist, as shown in Eq. (2).

$$X = \{X_1, X_2, \cdots, X_n\},\ X \in A \quad (2)$$

The expression for the predicted value is shown in Eq. (3).

$$Y = f(X) = \{y_1, y_2, \ldots, y_n\} \quad (3)$$

In Eq. (3), X_n represents the feature vector of the n th tourist. y_n represents the tourist attraction that is predicted to be preferred by the n th tourist. $f(X)$ represents the RF classification function. The application process of using RF to establish a tourist preference attraction prediction model is shown in Fig. 1.

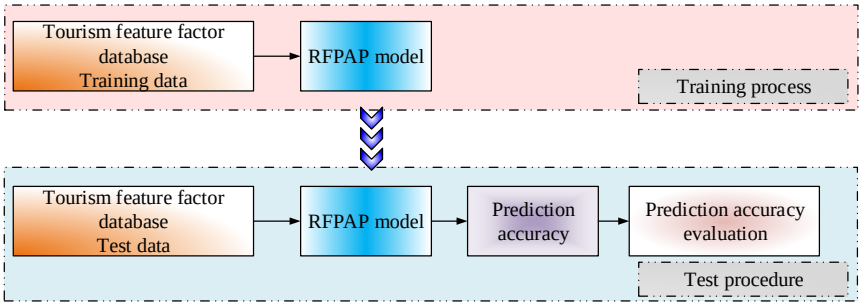
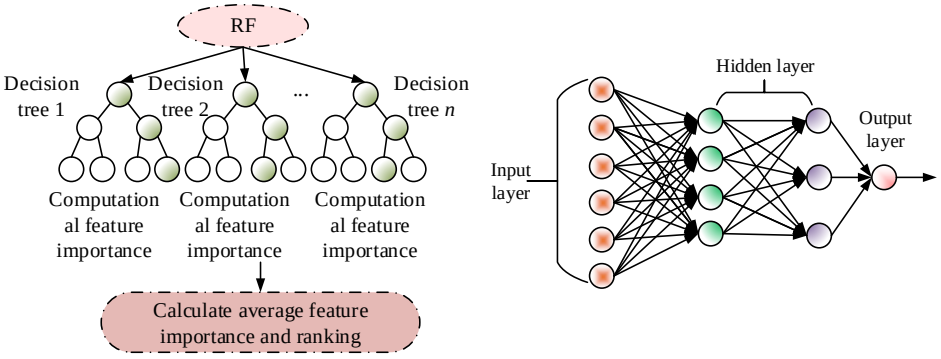


Fig. 1. RFPAP Tourist preference prediction model.

Although the RF-based tourist preference attraction prediction model has certain applicability, there are also some limitations. For example, when processing a large number of features, RF may overly rely on certain features and ignore other relevant and informative features, which may affect the predictive effect of the model. In predicting tourist attraction preferences, for example, there may be certain correlations between characteristics such as age, occupation, and income of tourists, and failure to handle them properly may affect the results [21]. In addition, RF is difficult to handle nonlinear relationships. RF has limited processing capabilities for complex nonlinear and high-dimensional data. In predicting tourist preferences for attractions, tourist behavior and preferences may be influenced by multiple factors, and there may be complex nonlinear relationships between these factors, which RF may find difficult to fully capture. ANN simulates the connectivity patterns of human brain neurons, and through massive training data for learning, it can effectively extract high-dimensional feature information from the data, and even recognize complex and nonlinear patterns. This makes it perform well in various tasks, including data classification, object detection, target tracking, etc. [22]. MLP is a specific ANN architecture. The layers are connected by weights and nonlinearity is introduced through activation functions, allowing MLP to learn and process complex data models. This study integrates RF with MLP in ANN to construct the final tourist preference attraction prediction model. The structure of RF and MLP is denoted in Fig. 2.



(a) Importance selection of RF features (b) MLP model

Fig. 2. Structure of RF and MLP.

The process of integrating RF and MLP in this study includes data preprocessing, grouping, building neural network models, obtaining uncertainties, and improving prediction results. The main goal of the data preprocessing stage is to eliminate noise and improve analysis efficiency. The methods include data cleaning, integration, and transformation to reduce analysis costs [23]. The processed dataset consists of two subsets of data, as shown in Eq. (4).

$$\begin{cases} S = \{S_1, S_2, \cdots, S_n\}, \ S_i \in [0,1], \ i \in \{1,2,\cdots,n\} \\ Z = \{Z_1, Z_2, \cdots, Z_n\}, \ Z_i \in R, \ i \in \{1,2,\cdots,n\} \end{cases} \quad (4)$$

The dataset S is analyzed using MLP and trained to obtain the output of MLP, and the dataset Z is predicted using RF. After obtaining the output of RF, uncertainty can be obtained by comparing the different outputs between them. After obtaining the set of uncertain items, it is passed to the logistic regression layer of the MLP model for parameter updates. By using the uncertainty in the training set to fit the logistic regression layer, the uncertainty in the test set can be predicted, and combined with the previous output results, the final prediction can be obtained. The process of training logistic regression layers is similar to the process of obtaining differential terms [24]. The Random Forest Preferred Attraction Prediction-Neural Networks Preferred Attraction Prediction (RFPAP-NNPAP) model constructed is shown in Fig. 3.

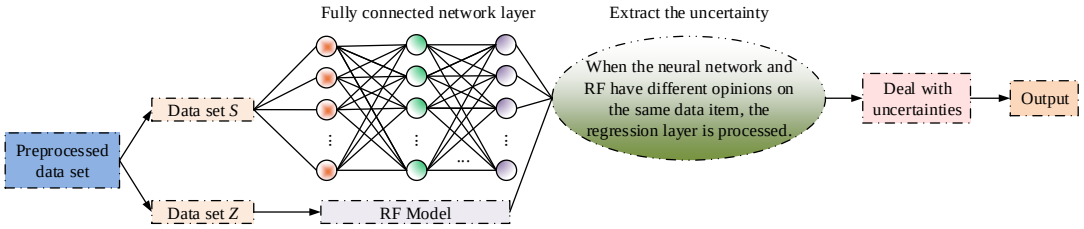


Fig. 3. RFPAP-NNPAP Model structure.

In Fig. 3, this study added an uncertainty extraction algorithm between the MLP layer and the logistic regression layer of the original MLP. The hidden layer state of MLP or RF was obtained through the Sigmoid activation function to obtain the output, which is then optimized by the logistic regression layer. The Sigmoid binary classification algorithm is based on conditional probability, with a threshold of 0.5 for classification, and can be extended to multi-dimensional feature space binary classification. The Sigmoid function in the multidimensional feature space is indicated in Eq. (5).

$$h_\theta(X) = g(\theta^T X) = \frac{1}{1+e^{-\theta^T x}} \tag{5}$$

In Eq. (5), θ represents multidimensional parameters. X represents the feature space matrix. For the binary classification problem, the conditional probability formula for the sample and parameter θ is shown in Eq. (6).

$$P(y|X;\theta) = (h_\theta(X))^y (1-h_\theta(X))^{1-y} \tag{6}$$

In Eq. (6), y represents the output of the binary classification problem. After obtaining the probability function, maximum likelihood estimation is performed, as shown in Eq. (7).

$$\rho(\theta) = \log L(\theta) = \sum_{i=1}^{m} y^{(i)} \log h(X^{(i)}) + (1-y^{(i)}) \log(1-h(X^{(i)})) \tag{7}$$

The derivative of parameter θ is calculated for Eq. (7) and the parameter gradient iteration function is obtained as expressed in Eq. (8).

$$\theta_j := \theta_j + \alpha(y^{(i)} - h_\theta(X^{(i)}))X_j^{(i)} \tag{8}$$

The training set is continuously iterated to obtain the approximate extremum of the loss function gradient. During each iteration, the model parameters are updated based on the current gradient direction to maximize the objective function. This optimization process will continue until the preset stopping criteria are met. After the termination conditions are met, the obtained model parameter θ is considered the optimal parameter, and the model can fit the training data to the maximum extent possible, while also being suitable for predicting new data.

B. Construction of a Tourist Attractions Recommendation Model Integrating FP-Growth and RFPAP-NNPAP Algorithms

To enhance the personalized level of tourism experience and services, this study applied association rule algorithms to explore the association relationships between tourist attractions and establish a tourist attractions association model. Association rule algorithm is a method for finding relationships between features in large-scale datasets, widely used in the field of market analysis [25]. In this study, the association rule algorithm was used to explore the preference patterns of tourists when choosing tourist attractions, as well as the co-occurrence relationships between different attractions. Through this method, potential behavioral patterns of tourists when choosing tourist attractions can be revealed, providing a basis for providing personalized tourism recommendation services [26]. Meanwhile, by analyzing the correlation between tourist attractions, it can further understand the characteristics and values of each attraction, which has important reference value for tourism planning and management. The data mining process model is indicated in Fig. 4.

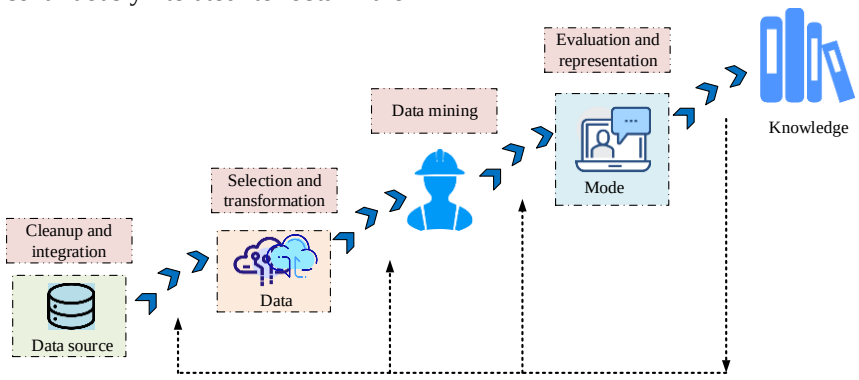


Fig. 4. Data mining process model.

The FP-Growth algorithm is a data mining method that belongs to association rules. It mainly generates frequent itemsets from the frequent pattern tree FP-Tree, divides the scanned database into numerous conditional datasets, and then mines association rules from them. The purpose of association rule mining is to find some trustworthy rules from massive data, which can help relevant personnel make judgments and decisions based on the situation to a certain extent [27]. The association rule mining system is based on two minimum thresholds to find association rules, namely the minimum support threshold $\min\_sup$ and the minimum confidence threshold $\min\_conf$ [28]. The work of association rule mining is mainly divided into two stages: The first is to find all itemsets that are not less than the minimum support threshold $\min\_sup$, that is, frequent sets. The second is to search for association rules that are not less than the minimum confidence threshold $\min\_conf$ for each frequent set. If the association rules $A \Rightarrow B$, $A = \{a_1, a_2, \ldots, a_i\} \subseteq I$, $B = \{b_1, b_2, \ldots, b_j\} \subseteq I$, and $A \neq \varnothing$, $B \neq \varnothing$ are defined, then the support of $A \Rightarrow B$ can be expressed as Eq. (9).

$$Support(A \Rightarrow B) = Support(A \cup B) = P(AB) \quad (9)$$

In Eq. (9), A represents the antecedent, B represents the consequent, and B appears with the appearance of A. At the same time, the confidence of rule $A \Rightarrow B$ is the ratio of $A \cup B$'s support to A's support, and its function expression is Eq. (10).

$$Confidence(A \Rightarrow B) = P(B|A) = \frac{Support(A \cup B)}{Support(A)} \quad (10)$$

In Eq. (10), $P(B|A)$ represents the ratio of the probability of event A and event B occurring simultaneously to the probability of event A occurring. The degree of improvement $lift$ refers to the ratio of the likelihood of both containing B under the condition of A to the likelihood of B sets occurring under unrestricted conditions [29]. This indicator is basically consistent with the confidence function and can be used to measure the reliability of rules. It is a supplementary explanation of confidence, and its calculation method is shown in Eq. (11).

$$\begin{aligned} Lift(A \Rightarrow B) &= \frac{P(B|A)}{P(B)} \\ &= \frac{Confidence(A \Rightarrow B)}{P(B)} \\ &= \frac{Support(A \cup B)}{Support(A)^{*} Support(B)} \end{aligned} \quad (11)$$

The meaning of Eq. (11) is to measure the independence of itemset A and itemset B. When the improvement degree of $A \Rightarrow B$ rule is 1, it indicates that event A and event B are independent of each other. If the improvement is less than 1, it indicates that event A and event B are mutually exclusive. In general, only when the improvement degree is greater than 3, can the association rules obtained in data mining be considered valuable. The traditional FP-Growth algorithm is suitable for situations with small databases, as as the database continues to expand, the FP-Tree established by traditional FP-Growth will occupy a large amount of memory, consume a lot of computation time, and there is a possibility of memory overflow, which reduces the efficiency of data mining [30]. Therefore, when using the FP-Growth algorithm in practice, it is necessary to optimize its support and confidence. The minimum support and confidence levels are automatically set based on the characteristics of the data itself, in order to avoid subjective randomness in manually setting parameters. The transaction set D is defined, the support numbers of each item in D are sorted in descending order, and the polynomial curve fitting function is calculated based on the corresponding numbers in the order table. The expression is Eq. (12).

$$y = f(x) = \sum_{i=0}^{t} m_i \times x^i \quad (12)$$

In Eq. (12), t means the amount of samples, and x expresses the ordinal value. A quadratic differentiation is performed on the function of Eq. (12) to obtain the second-order derivative function $f''(x)$, which is expressed as Eq. (13).

$$y'' = f''(x) = \sum_{i=2}^{t} i \times (i-1) \times m_i \times x^{i-2} \quad (13)$$

In Eq. (13), the value of x for the first occurrence of $f''(x) = 0$ in the interval (1, m) of $f''(x)$ is denoted as x_0, and $\lfloor f(x_0) \rfloor$ rounded down from $f(x_0)$ is used as the algorithm parameter. It is also necessary to improve the mining process of FP-Growth after optimizing its parameters. This study used the method of adding constraints to mine association rules based on the adaptive adjustment of minimum support and minimum confidence, forming a tourist attraction rule library [31]. Correlation coefficients are used to eliminate highly correlated redundant data and constraints are formed to reduce unnecessary data and simplify calculations, thereby improving the efficiency of data processing. The calculation method for correlation coefficients can be expressed as Eq. (14).

$$\rho = \frac{Cov(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}} \quad (14)$$

Eq. (14) represents the correlation coefficient between data X and Y, with a value range of [-1,1]. When $\rho = 0$ is used, it indicates that X and Y are not correlated. When $|\rho| = 1$, it means that X and Y are completely correlated, and one of the data needs to be removed; When $0.8 < |\rho| < 1$

occurs, changes in X will cause partial changes in Y, indicating that X and Y are highly correlated, and one of the data needs to be removed. When $|\rho| < 0.3$ is used, it indicates that X and Y are low correlated, and it is necessary to consider removing one of them as appropriate. When building an FP-tree, the parent node pointer and child node pointer are combined into one pointer, and the sibling node pointer and the node pointer with the same name are combined into one pointer to construct an OFP-tree. This operation can save space and simplify the process. The results of OFP-Tree establishment are shown in Fig. 5.

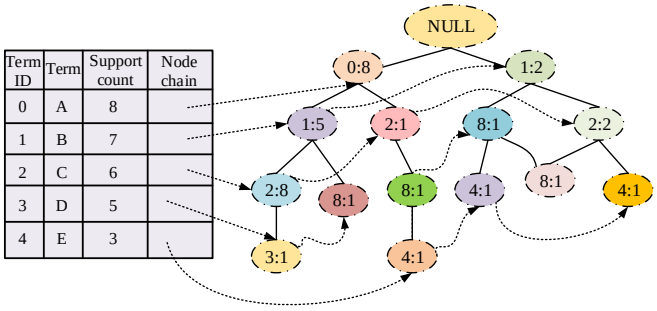


Fig. 5. OFP-Tree building results.

This study proposed a hybrid recommendation method that combines FP-Growth algorithm and RFPAP-NNPAP algorithm. This method aims to improve the diversity and richness of attraction recommendations by jointly constructing a preference attraction prediction model and attraction association model. The FP-Growth algorithm is used to mine frequent itemsets and reveal association rules between scenic spots. The RFPAP-NNPAP algorithm is applied to predict tourist attraction preferences. The fusion of this algorithm can not only provide recommendations that meet the personal preferences of tourists, but also reveal the correlation between attractions, providing tourists with more diverse choices. The specific recommendation method process is shown in Fig. 6.

In this study, a comprehensive method was used to select tourism characteristic factors, covering three key dimensions: tourist attractions, individual tourists, and contextual perception information. A total of 13 key tourism characteristic factors were selected, including scenic spot location, scenic spot ticket prices, season, gender, etc. These feature factors not only cover the basic information of tourist attractions, but also include the individual characteristics of tourists and contextual information of the tourism environment. By selecting these factors, a rich library of tourism feature factors was constructed, providing comprehensive and in-depth feature references for the problem of recommending tourist attractions. The construction of this feature factor library helps to deepen the understanding of tourist behavior patterns and reveal the key driving factors for tourist attractions selection. The specific tourism characteristic factor library is shown in Fig. 7.

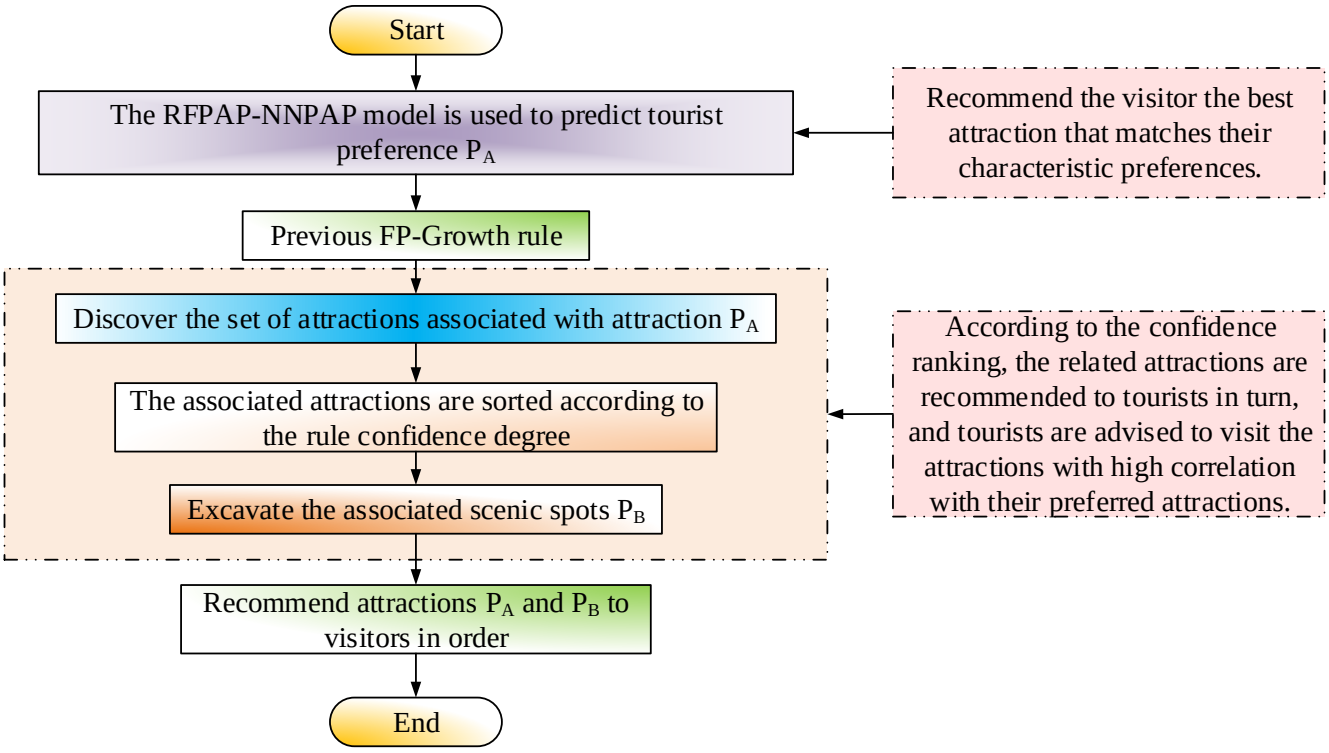


Fig. 6. Operation flow of the hybrid tourist attractions recommendation model.

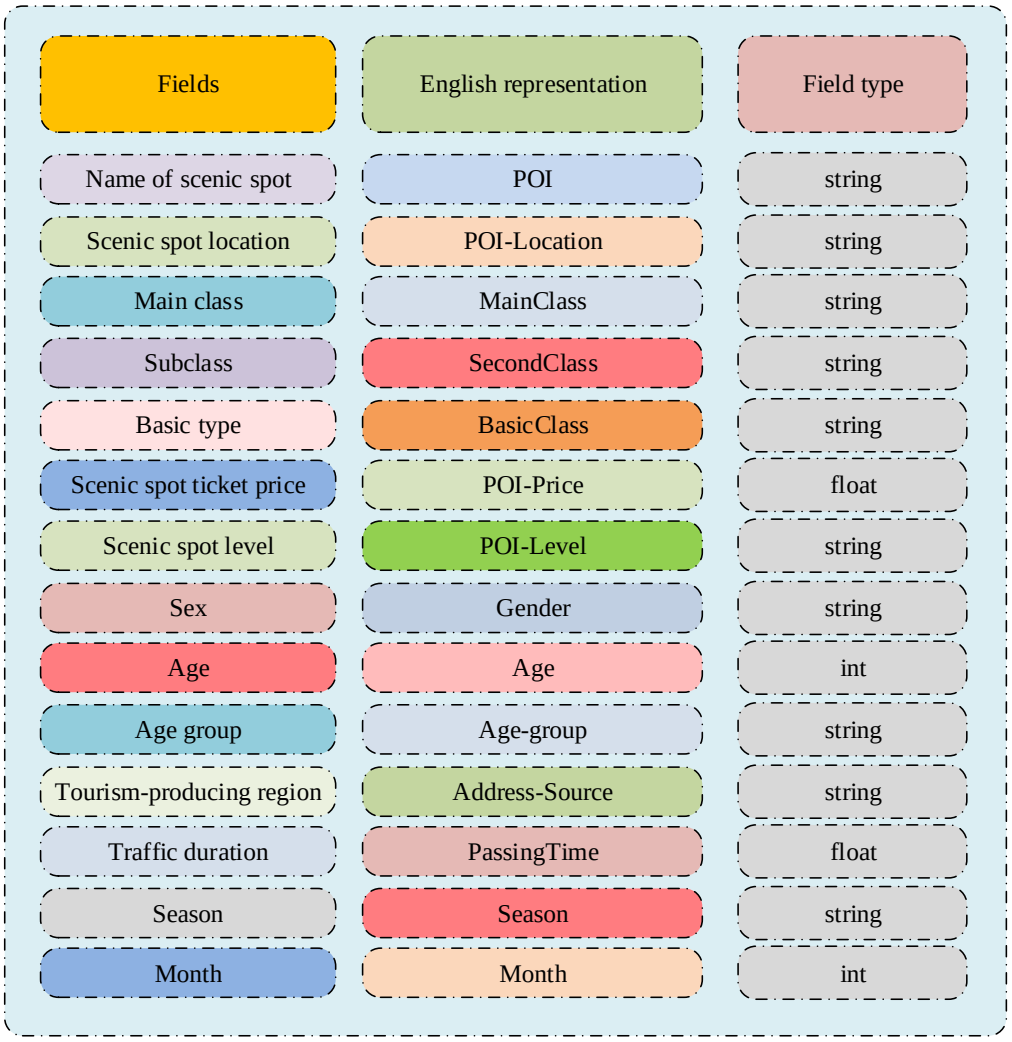


Fig. 7. Tourism feature factor database.

IV. Performance Verification of Tourist Attraction Recommendation Model Based on Hybrid Model

To confirm the practicality of the proposed algorithm, this study first conducted in-depth exploration and analysis of the effect of the RFPAP-NNPAP model through experiments. Afterwards, a detailed evaluation and analysis of the recommendation effect of the hybrid model in practical scenarios was conducted, to better understand and evaluate the practical application value and potential of this hybrid model in TAR.

A. Performance Verification of RFPAP-NNPAP Model

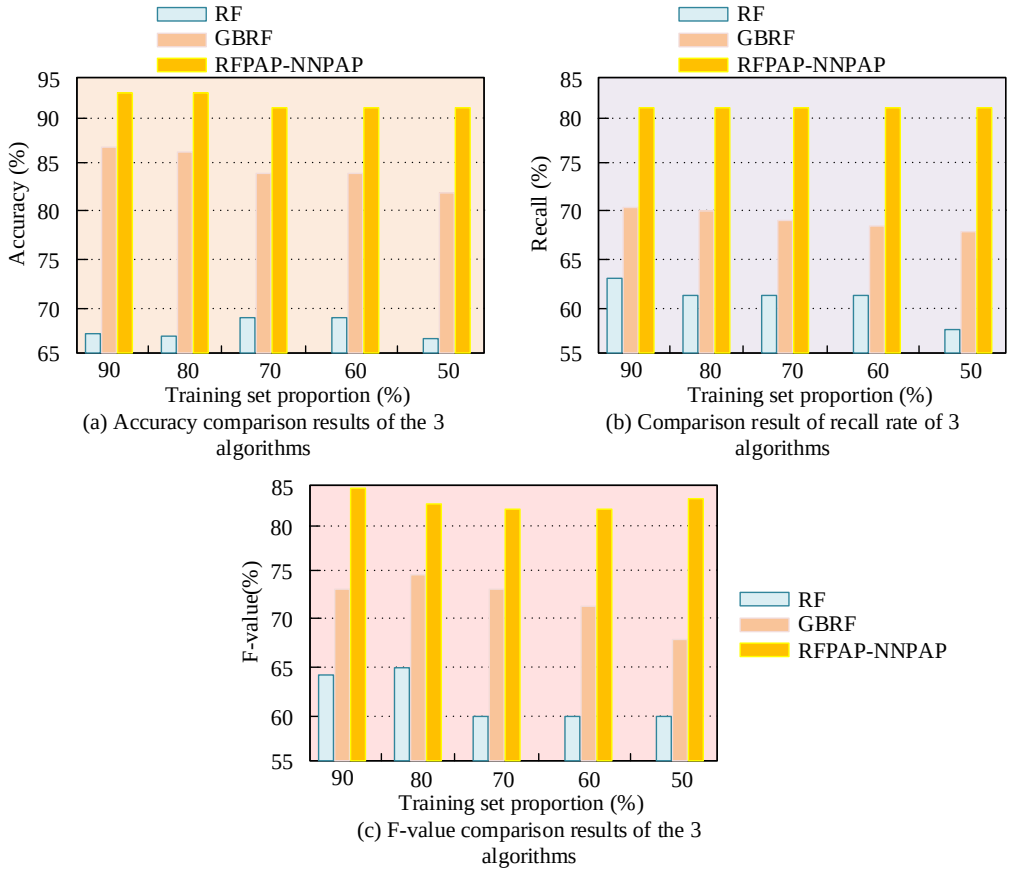
This study used an i7-6500U processor, a 16GB memory computer, and a Windows 10 64 bit system. The experimental data came from the Sina Weibo tourism dataset, which includes a large amount of tourism information. The experimental environment was the Spyder integrated development environment, and the Scikit-learn library was utilized to convert the data into numerical values. The study set five gradient percentages for sampling the test dataset, and the corresponding training dataset was also five gradient percentages. Dataset D was randomly split into training data and test data. When verifying the performance of RFPAP-NNPAP, in addition to comparing it with traditional RF, Gradient Boosting Random Forest model (GBRF) was also selected for comparative verification.

The study compared the performance of RFPAP-NNPAP, GBRF, and RF models on different segmentation ratio datasets, as denoted in Fig. 8. The outcomes denoted that the average accuracy of RFPAP-NNPAP, GBRF, and RF was 92.26%, 84.12%, and 66.41%, respectively. The average Recall value of RFPAP-NNPAP, GBRF, and RF was 82.11%, 69.11%, and 60.12%, respectively. The average F-value of RFPAP-NNPAP, GBRF, and RF was 84.43%, 71.11%, and 61.11%, respectively. RFPAP-NNPAP had higher accuracy, Recall, and F-value than the GBRF model by 8.14%, 13.00%, and 13.32%, respectively. Thus, the superiority of RFPAP-NNPAP was validated.

Fig. 9 shows the error correction comparison results of two models. The experiment findings indicated that the prediction results of GBRF were not ideal, and the accuracy rate was mostly less than 20%. RFPAP-NNPAP used all uncertain terms in the training set as the training set for the logistic regression layer, trained and updated the parameters of the

logistic regression layer, with accuracy fluctuating around 90%. The results indicated that the logistic regression layer with updated parameters had good prediction results.

Fig. 10(a) shows the confidence and accuracy of the 30 selected association rules. The confidence of the rules themselves had a similar trend to the confidence of the rules in the test set, and the accuracy fluctuates around 95%, indicating that the mined association rules are universal. Fig. 10(b) shows the experimental comparison curves of FP-Growth and FP-Growth algorithms after parameter optimization. In Fig. 10(b), when processing the same data, the optimized FP-Growth algorithm significantly outperformed the traditional algorithm in runtime. Especially when the support was smaller, the advantages of improving the FP-Growth algorithm became more apparent, indicating that the performance of the optimized algorithm has been improved.

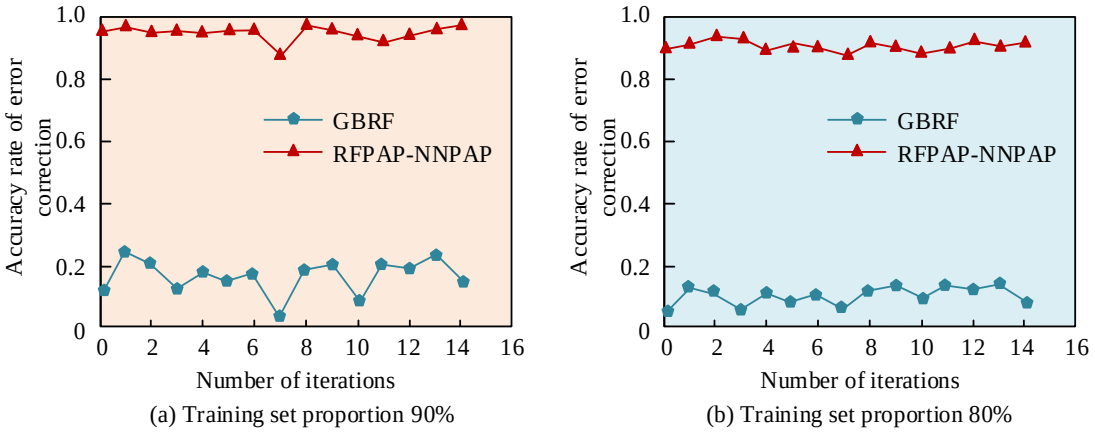


(a) Accuracy comparison results of the 3 algorithms

(b) Comparison result of recall rate of 3 algorithms

(c) F-value comparison results of the 3 algorithms

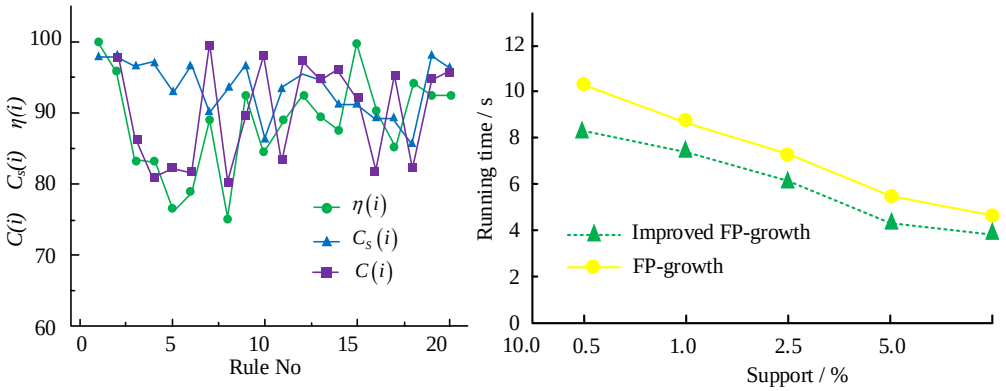
Fig. 8. Comparison results of accuracy, recall rate and f-value of the three algorithms.



(a) Training set proportion 90%

(b) Training set proportion 80%

Fig. 9. Comparison of error correction results of two models.



(a)Accuracy of association rules

(b)Comparison of execution time between two algorithms

Fig. 10. Accuracy of association rules and running time of two algorithms.

To deeply identify the effect of the optimized FP-Growth algorithm, the experiment chose to compare the algorithm before and after optimization with the AprioriTid algorithm. Fig. 11 shows the comparison curve of the time taken by the FP-Growth algorithm before and after optimization, as well as the AprioriTid algorithm with the change of minimum support. As shown in the figure, with the increasing minimum support, the overall time effect of the optimized FP-Growth algorithm was better than that of the AprioriTid algorithm. After calculation, the improved FP-Growth algorithm saved an average of 23.4 seconds in running time compared to the original FP-Growth algorithm. Compared to the AprioriTid algorithm, it had an average reduction of 12.3 seconds. The FP-Growth algorithm mined association rules based on adjusted support and confidence, and could obtain all the rules that meet the requirements without leaving any omissions. The experimental results in Fig. 11 showed that as the minimum support increased, the number of rules decreased. It can be calculated that the optimized FP-Growth algorithm has a maximum elimination rate of 38% for invalid rules.

This study sorted the data items in descending order based on the obtained support numbers and performed curve fitting. A total of three polynomial curve fitting was performed, and the support numbers and curve fitting results of the data items are illustrated in Fig. 12(a). At the same time, the confidence of the association rules was sorted in descending order, and the results obtained by fitting the cubic polynomial curve are expressed in Fig. 12(b). In the figure, the fitting degree of the curve was relatively high, indicating that the predicted results are basically consistent with the actual situation, and the algorithm has operability and practicality.

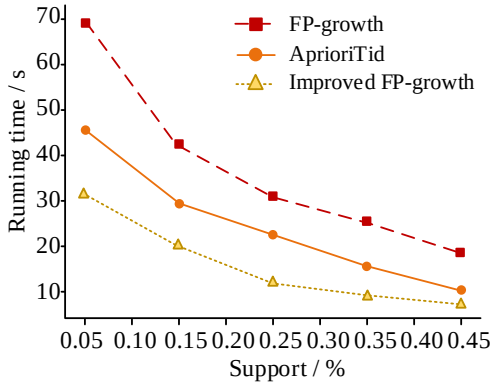
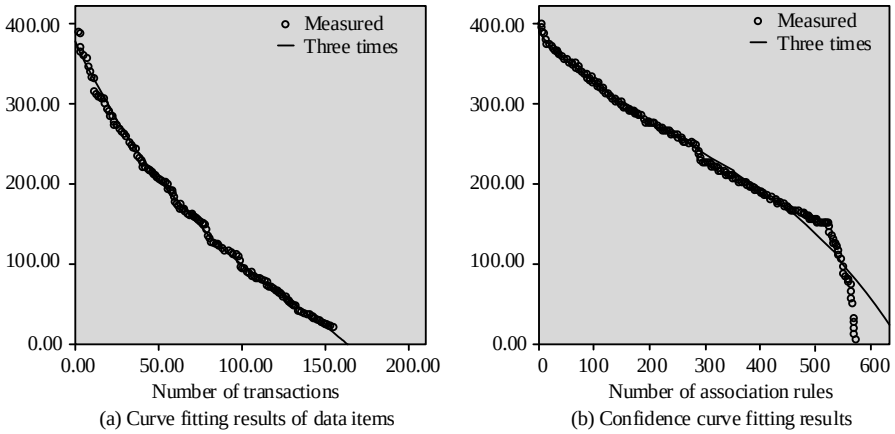


Fig. 11. Running time comparison of three algorithms.



(a) Curve fitting results of data items

(b) Confidence curve fitting results

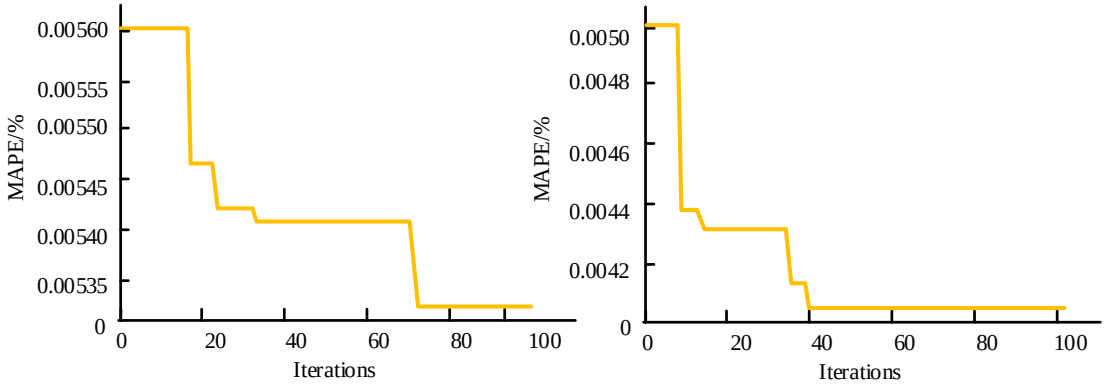
Fig. 12. Curve fitting degree diagram.

B. *Performance Verification of a Hybrid Attraction Recommendation Model Combining Optimized FP-Growth and RFPAP-NNPAP Algorithms*

This experiment selected two representative popular tourist cities, Beijing and Yunnan. The experiment utilized a constructed mixed model to predict tourist attraction preferences. The evaluation indicators for the experiment include Mean Absolute Error (MAE), MSE, Normalized Root Mean Square Error (NRMSE), and Mean Absolute Percentage Error (MAPE) to evaluate the predictive effect of the model. Fig. 13 shows the error gradient trend of AprioriTid and mixed model in estimating the preference of tourist attractions in Beijing. The initial iteration of the hybrid model was around 0.49%, while AprioriTid was around 0.55%. The main reason was that the hybrid model had faster local search ability and iteration speed, ultimately achieving better convergence. In addition, after approximately 16 iterations, the MAPE value of the mixed model decreased to 0.44%. After about 39 iterations, the MAPE value of the mixed model decreased to 0.40%.

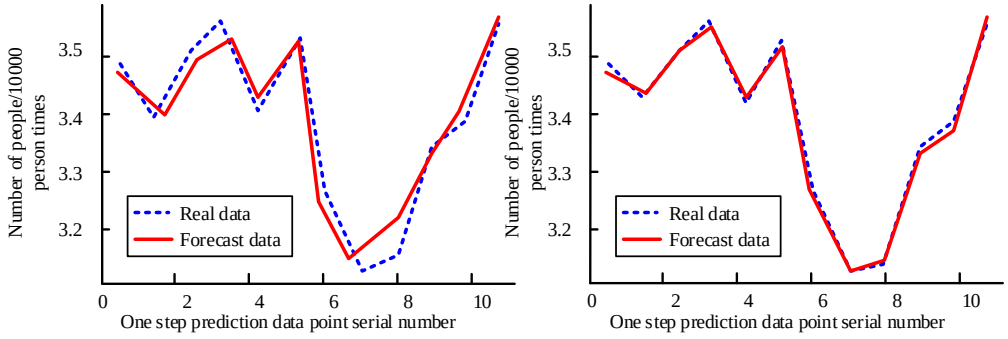
Fig. 14 shows the estimation results of AprioriTid and mixed model on the number of tourists preferred by Beijing's tourist attractions. From the graph, the estimated values of both algorithms tended to be consistent with the true values, indicating that they both had good predictive performance. Compared to the hybrid model, AprioriTid had slightly lower prediction accuracy, which was reflected in the data sequence numbers between 0-6. The degree of overlap between AprioriTid's predicted values and the true values was not as significant as that of the hybrid model, indicating that its prediction error was greater than that of the hybrid model. Therefore, the estimation of preferences for tourist attractions in Beijing also confirmed that the hybrid model had superior predictive performance compared to AprioriTid.

Fig. 15 shows the estimation results of tourist numbers for Yunnan and Beijing tourist attractions using the AprioriTid model and a hybrid model. From the data from Yunnan, AprioriTid's predicted value curve deviated significantly from the true value curve and had few overlapping points, indicating a significant estimation error. Observing that the predicted value curve of the hybrid model basically coincided with the true value curve could also prove that the estimation error of the hybrid model was less than AprioriTid. The above results further confirmed that the hybrid model had a good optimization effect at the initial position, resulting in a higher convergence speed and prediction accuracy of the overall prediction model compared to the AprioriTid model.



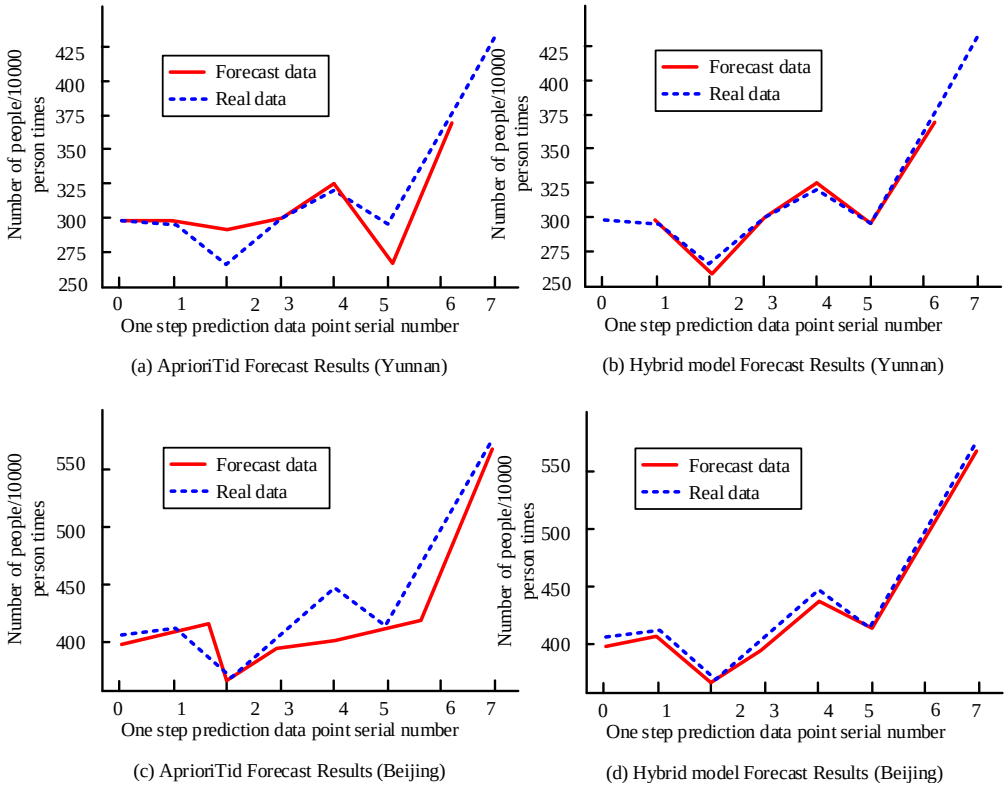
(a) AprioriTid Error gradient trend chart (b) Hybrid model Error gradient trend chart

Fig. 13. Error gradient of the two models in estimating the number of tourists preferred by tourist attractions in Beijing.



(a) AprioriTid Predicted number of tourists in Beijing (b) Hybrid model Predicted number of tourists in Beijing

Fig. 14. Estimate results of AprioriTid and hybrid model on the number of tourists with preference for tourist attractions in Beijing.



(a) AprioriTid Forecast Results (Yunnan)

(b) Hybrid model Forecast Results (Yunnan)

(c) AprioriTid Forecast Results (Beijing)

(d) Hybrid model Forecast Results (Beijing)

Fig. 15. Prediction results of two algorithms on tourist attraction preferences of two cities.

V. Discussion

The RFPAP-NNPAP model constructed in this study has a good effect in predicting tourist attraction preference, with an average accuracy of 92.26%, an average recall value of 82.11%, and an average F value of 84.43%, all of which are better than GBRF and RF models. In addition, the optimized FP-Growth algorithm significantly improves the running time and rule mining efficiency, and shows higher performance than the traditional algorithm and AprioriTid algorithm. The main reason is that the proposed model integrates FP-Growth algorithm and RFPAP-NNPAP algorithm to form a hybrid TAR model, which effectively improves the ability to process complex data and significantly optimizes the operational efficiency and accuracy, so it is effective in predicting the number of preferred tourists of tourist attractions. Compared with the study of Huang et al. [2], although they used an optimized neural network algorithm to predict tourist hotspots, this study not only improved the accuracy of prediction, but also enhanced the universality and adaptability of the model by integrating the two algorithms. P. Nitu et al. [3] proposed a personalized travel recommendation system considering timeliness in his research. This study further optimized the real-time response ability and accuracy of the recommendation system by combining a variety of algorithms to process complex data, and the two are consistent. In addition, the integrated model not only optimizes route selection, but also deeply analyzes user behavior and preferences through data mining technology, which has similar significance to C. Chen et al. [4] Personalized travel route recommendation model based on improved genetic algorithm. To sum up, this study not only improves the accuracy and efficiency of the scenic spot recommendation system through the combination of multiple algorithms, but also proves the robustness of the model in different data sets. Although the proposed model performs well in terms of performance and application scope, there are some limitations. The complexity of the model can lead to a large demand on computing resources, and future research needs to explore more efficient algorithm implementation ways to mitigate hardware requirements. At the same time, with the increase of data volume and dimension, the scalability and stability of the model need to be further verified. Future studies can test the validity of the model on more regions and different types of tourism data, further explore the optimal configuration and practical application scenarios of the algorithm, and provide scientific decision support tools for the tourism industry.

VI. Conclusion

In the tourism industry driven by globalization and digitization, TAR faces challenges. Existing algorithms have limitations in handling complex, nonlinear, and large-scale data, and there is an urgent need for new solutions to meet personalized needs. Therefore, this study proposed a hybrid TAR model that integrates optimized FP-Growth and RFPAP-NNPAP algorithms. The study conducted performance validation on the proposed model, and the outcomes indicated that the average accuracy of RFPAP-NNPAP was 92.26%, the average Recall value was 82.11%, and the average F value of RFPAP-NNPAP was 84.43%, all of which were better than the comparison algorithms. The error correction accuracy of RFPAP-NNPAP fluctuated around 90%. The optimized FP-Growth algorithm had significantly better runtime than

traditional FP-Growth, and its elimination rate for invalid rules could reach up to 38%. The actual verification results of the hybrid model showed that after about 16 iterations, its MAPE value decreased to 0.44%. After about 39 iterations, the MAPE value of the hybrid model decreased to 0.40%. The estimation results of the number of tourists preferred by the hybrid model for tourist attractions in Yunnan and Beijing indicated that the predicted value curve of the hybrid model basically overlapped with the true value curve. Thus, the effectiveness of the hybrid model was validated. The main contribution lies in providing a new solution to meet the personalized needs of TAR. However, there are still shortcomings in the research, such as the need for further optimization of the model's performance in specific types of data or specific scenarios. In the future, efforts will be made to raise the universality and stability of the model, to offer better TARs in a wider range of application scenarios.

REFERENCES

[1] A. Alsharef, K. Aggarwal, Sonia, M. Kumar and A. Mishra, "Review of ML and AutoML solutions to forecast time-series data," Arch. Comput. Method E., vol. 29, no. 7, pp. 5297-5311, November, 2022, DOI: https://doi.org/10.1007/s11831-022-09765-0.

[2] X. Huang, V. Jagota, E. Espinoza-Muñoz and J. Albornoz, "Tourist hot spots prediction model based on optimized neural network algorithm," Int. J. Syst. Assur. Eng., vol. 13, no. 1, pp. 63-71, March, 2022, DOI: https://doi.org/10.1007/s13198-021-01226-4.

[3] P. Nitu, J. Coelho and P. Madiraju, "Improvising personalized travel recommendation system with recency effects," in Big Data Mining and Analytics, vol. 4, no. 3, pp. 139-154, September, 2021, DOI: 10.26599/BDMA.2020.9020026.

[4] C. Chen, S. Zhang, Q. Yu, Z. Ye, Z. Ye and F. Hu, "Personalized travel route recommendation algorithm based on improved genetic algorithm," J. Intell. Fuzzy Syst., vol. 40, no. 3, pp. 4407-4423, March, 2021, DOI: 10.3233/JIFS-201218.

[5] H. Huang, A. V. Savkin and C. Huang, "Reliable Path Planning for Drone Delivery Using a Stochastic Time-Dependent Public Transportation Network," in IEEE Transactions on Intelligent Transportation Systems, vol. 22, no. 8, pp. 4941-4950, August, 2021, DOI: 10.1109/TITS.2020.2983491.

[6] X. Wang, Z. Dai, H. Li and J. Yang, "Research on Hybrid Collaborative Filtering Recommendation Algorithm Based on the Time Effect and Sentiment Analysis," Complexity, vol. 2021, no. 2, pp. 1-11, March, 2021, DOI:10.1155/2021/6635202.

[7] R. Hössinger, F. Aschauer, S. Jara-Díaz, S. Jokubauskaite, B. Schmid, S. Peer, K. Axhausen and R. Gerike, "A joint time-assignment and expenditure-allocation model: value of leisure and value of time assigned to travel for specific population segments," Transportation, vol. 47, no. 3, pp. 1439-1475, June, 2020, DOI: https://doi.org/10.1007/s11116-019-10022-w.

[8] L. Wen, C. Liu, H. Song and H. Liu, "Forecasting tourism demand with an improved mixed data sampling model," J. Travel Res., vol. 60, no. 2, pp. 336-353, March, 2021, DOI: https://doi.org/10.1177/004728752090622.

[9] B. Cao, J. Zhao, Z. Lv and P. Yang, "Diversified Personalized Recommendation Optimization Based on Mobile Data," IEEE T INTELL TRANSP, vol. 22, no. 4, pp. 2133-2139, April 2021, doi: 10.1109/TITS.2020.3040909.

[10] Y. Zhang and Z. Tang, "PSO-weighted random forest for attractive tourism spots recommendation," Future Gener. Comp. Sy., vol. 127, pp. 421-425, February, 2022, DOI: https://doi.org/10.1016/j.future.2021.09.029.

[11] A. Hill, G. Herman and R. Schumacher, "Forecasting severe weather with random forests," Mon. Weather Rev., vol. 148, no. 5, pp. 2135-2161, May, 2020, DOI: 10.1175/MWR-D-19-0344.1.

[12] J. Yoon, "Forecasting of real GDP growth using machine learning models: Gradient boosting and random forest approach," Comput. Econ., vol. 57, no. 1, pp. 247-265, January, 2021, DOI: https://doi.org/10.1007/s10614-020-10054-w.

[13] S. Chen, Q. Wei, Y. Zhu and G. Ma, "Medium-and long-term runoff forecasting based on a random forest regression model," Water Supply, vol. 20, no. 8, pp. 3658-3664, September, 2020, DOI:10.2166/ws.2020.214.

[14] T. Wang, X. Wang, R. Ma, X. Li, X. Hu, F. Chan and J. Ruan, "Random forest-bayesian optimization for product quality prediction with large-scale dimensions in process industrial cyber–physical systems," IEEE Internet Things, vol. 7, no. 9, pp. 8641-8653, May, 2020, DOI: 10.1109/JIOT.2020.2992811.

[15] D. Yun, B. Zheng, B. Gu, X. Gao and R. Behnaz, "Predicting the CPT-based pile set-up parameters using HHO-RF and PSO-RF hybrid models," Struct. Eng. Mech., vol. 86, no. 5, pp.673-686, May, 2023, DOI: 10.12989/sem.2023.86.5.673.

[16] H. Pan and Z. Zhang, "Research on context-awareness mobile tourism e-commerce personalized recommendation model," J. Signal Process. Sys., vol. 93, no. 2, pp. 147-154, March, 2021, DOI: https://doi.org/10.1007/s11265-019-01504-2.

[17] O. Västberg, A. Karlström, D. Jonsson and M. Sundberg, "A dynamic discrete choice activity-based travel demand model," Transport Sci., vol. 54, no. 1, pp. 21-41, October, 2020, DOI: https://doi.org/10.1287/trsc.2019.0898.

[18] M. A. Guillermo, M. C. Rivera, K. Lucas, A. Bandala, R. Billones, E. Sybingco, A. Fillone and E. Dadios, "Strategic Transit Route Recommendation Considering Multi-Trip Feature Desirability Using Logit Model with Optimal Travel Time Analysis," J. Adv. Comput. Intell., vol. 26, no. 6, pp. 983-994, December, 2022, DOI: https://doi.org/10.20965/jaciii.2022.p0983.

[19] B. Balcik and İ. Yanıkoğlu, "A robust optimization approach for humanitarian needs assessment planning under travel time uncertainty," Eur. J. Oper. Res., vol. 282, no. 1, pp. 40-57, April, 2020, DOI: https://doi.org/10.1016/j.ejor.2019.09.008.

[20] V. Shinkarenko, S. Nezdoyminov, S. Galasyuk and L. Shynkarenko, "Optimization of the tourist route by solving the problem of a salesman," J. Geol. Geogr. Geoecol., vol. 29, no. 3, pp. 572-579, March, 2020, DOI: 10.15421/112052.

[21] A. Koushik, M. Manoj and N. Nezamuddin, "Machine learning applications in activity-travel behaviour research: a review," Transport Rev., vol. 40, no. 3, pp. 288-311, January, 2020, DOI: https://doi.org/10.1080/01441647.2019.1704307.

[22] G. Assaker, "Age and gender differences in online travel reviews and user-generated-content (UGC) adoption: extending the technology acceptance model (TAM) with credibility theory," J. Hosp. Market Manag., vol. 29, no. 4, pp. 428-449, August, 2020, DOI: https://doi.org/10.1080/19368623.2019.1653807.

[23] P. Yochum, L. Chang, T. Gu and M. Zhu, "Linked Open Data in Location-Based Recommendation System on Tourism Domain: A Survey," in IEEE Access, vol. 8, pp. 16409-16439, 2020, August, DOI: 10.1109/ACCESS.2020.2967120.

[24] R. Pop, Z. Săplăcan, D. Dabija and M. Alt, "The impact of social media influencers on travel decisions: The role of trust in consumer decision journey," Curr. Issues Tour., vol. 25, no. 5, pp. 823-843, March, 2022, DOI: https://doi.org/10.1080/13683500.2021.1895729.

[25] F. Huang, J. Xu and J. Weng, "Multi-Task Travel Route Planning With a Flexible Deep Learning Framework," in IEEE Transactions on Intelligent Transportation Systems, vol. 22, no. 7, pp. 3907-3918, July, 2021, DOI: 10.1109/TITS.2020.2987645.

[26] D. M. Lemy, J. Amelda Pramezwary, R. Pramono and L. Nabila, "Explorative Study of Tourist Behavior in Seeking Information to Travel Planning," Planning, vol. 16, no. 8, pp. 1583-1589, August, 2021, DOI: 10.18280/ijsdp.160819.

[27] C. Archetti, D. Feillet, A. Mor and M. Speranza, "Dynamic traveling salesman problem with stochastic release dates," Eur. J. Oper. Res., vol. 280, no. 3, pp. 832-844, February, 2020, DOI: https://doi.org/10.1016/j.ejor.2019.07.062.

[28] D. Samara, I. Magnisalis and V. Peristeras, "Artificial intelligence and big data in tourism: a systematic literature review," J. Hosp. Tour. Technol., vol. 11, no. 2, pp. 343-367, September, 2020, DOI: 10.1108/jhtt-12-2018-0118.

[29] W. Wang, N. Kumar, J. Chen, Z. Gong, X. Kong, W. Wei and H. Gao, "Realizing the Potential of the Internet of Things for Smart Tourism with 5G and AI," in IEEE Network, vol. 34, no. 6, pp. 295-301, November/December, 2020, DOI: 10.1109/MNET.011.2000250.

[30] G. Mehdi, H. Hooman, Y. Liu, S. Peyman and R. Arif, "Data Mining Techniques for Web Mining: A Survey," AIA, vol. 1, no. 1, pp. 3-10, October, 2022, DOI: https://doi.org/10.47852/bonviewAIA2202290.

[31] H. Cao, Y. Wu, Y. Bao, X. Feng, S. Wan and C. Qian, "UTrans-Net: A Model for Short-Term Precipitation Prediction," AIA, vol. 1, no. 2, pp. 106-113, September, 2023, DOI: https://doi.org/10.47852 /bonviewAIA2202337.