Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 8299761, 12 pages
https://doi.org/10.1155/2022/8299761

Research Article

Personalized Hybrid Recommendation for Tourist Users Based on Matrix Cluster Apriori Mining Algorithm

Qian Zhang

School of Jiaozuo Teachers College, Jiaozuo 454000, Henan, China

Correspondence should be addressed to Qian Zhang; jzszzq@jzsz.edu.cn

Received 18 January 2022; Revised 22 February 2022; Accepted 24 February 2022; Published 13 April 2022

Academic Editor: Gengxin Sun

With the rapid development of Internet technology and the arrival of the era of big data, the rapid expansion of network information resources has formed massive information. Massive information resources have brought great convenience to people's lives. However, it becomes more and more difficult to find the content that interests you, which is the phenomenon of "information overload." In order to solve this problem, a solution based on personalized recommendation technology is proposed. In personalized recommendation technology, collaborative filtering algorithm is the most widely used technology. Clustering technology can effectively divide objects into groups, so that the similarity of attributes between objects in the same group is high, and the similarity of objects in different groups is low. The core step of the filtering recommendation algorithm is to find the similar neighbors of the target user by calculating the similarity. Applying the clustering technology to the recommendation can effectively improve the performance of the recommendation system. Aiming at the real-time problem of collaborative filtering recommendation, this paper introduces a method of firstly clustering users on the user item rating matrix, and finding the nearest neighbors in the clusters with high similarity with the target user, which effectively reduces the query space and improves the recommendation. This paper proposes a method to measure the user's preference for item attributes, which is used in the above clustering process to improve the recommendation accuracy while retaining the advantage of reducing the query space. Aiming at the problem of poor recommendation accuracy, this paper proposes a fuzzy-improved K-means algorithm to cluster items in the product attribute matrix, and then fuses the similarity of the belongingness of items to clusters in the fuzzy clustering. The similarity calculated on the score matrix shows that this method is better than the traditional hybrid recommendation in accuracy.

1. Introduction

In recent years, with the continuous development of technology, a large amount of data has been accumulated in various industries, and there is an urgent need for a highly efficient data mining technique that can mine the relevant data from many relevant data sets that are relevant to users. As data sets are expanding, traditional data mining techniques are unable to quickly and effectively mine and analyze massive data sets, and it is a hot issue to discover the rules and hidden connections contained in the data sets with high efficiency and performance. At the same time, cloud computing technology is developing rapidly and has become a hot topic in computer research [1]. Because mining frequent itemsets requires obtaining the support of itemsets from massive transaction records in various ways. This includes processes such as generating candidate item sets, traversing transaction records and counting candidate itemset support. The process of statistical confidence is after mining frequent itemsets. In 2004, Google published a paper on parallel computing framework, Hadoop is an open-source framework for parallel computing of massive data on large-scale clusters. As the technology continues to mature, the Hadoop distributed framework is being introduced as a subproject to daily R&D with increasing proficiency. The current combination of the Hadoop platform with data mining, using the parallelization algorithm of Reduce model in Hadoop, is also an important application in data mining. Association rules can be divided into several categories according to the category of variables, the level of abstraction of data, and the dimensionality of data. There are two types of concatenation rules, multidimensional and unidimensional, using the dimensionality of the data set as the separate criterion. A major part of concatenation analysis, namely, multidimensional association rule mining, is applied to various aspects and is one of the main goals of data mining research.

With the increase in the number and types of points of interest recommended by platforms, it is difficult for people to choose high-quality locations based on their preferences; it becomes extremely difficult for people to extract effective information, which makes the feedback of information increasingly decreasing [2]. Researchers are eager to find a way to extract information effectively, Recommendation System was born, and now it has become one of the important means of information filtering, an effective way to solve the problem of information overload at present. A personalized recommendation system uses the user's previous behavioral information, such as purchase history, to push the products of interest to the user, to reduce the time spent by the user in finding the required information. The recommendation system can generate great economic benefits, so personalized recommendation systems have been widely used in many fields such as movies, music, personalized advertising, and location-based services [3].

With the progress of society and the rapid sharing of information resources, the age span of tourism groups is getting bigger and bigger, and more people put tourism on the top of the list to improve the quality of life. At the same time, tourists pursue more initiative in the process of tourism and hope that scenic spots can provide comprehensive and detailed personalized services, while the traditional tour guide services can no longer meet the increasingly diverse individual requirements of different tourists. So, if the support of X is greater than the minimum support, the support IM of the subset of itemset X must also be greater than the minimum support, because the number of transaction records containing the subset of X is greater than or equal to the transaction records containing the itemset X quantity. There is a large gap between the quality of scenic tourism services and tourists' expectations, and the contradictions in tourism are becoming increasingly prominent. Therefore, to narrow this gap between supply and demand and reduce the negative impact on tourism, the tourism industry must build an information service platform based on the Internet, and provide targeted services for tourists through data collection and analysis. The widespread use of mobile Internet, especially mobile Internet, has laid the groundwork for meeting the real-time information needs of tourists; however, the explosive growth of information has caused an information overload, which has caused problems for tourists in choosing from the vast amount of tourism information [4]. On the other hand, the characteristics of tourism activity itself make travelers encounter various temporary or unexpected problems in the process, so tourism service providers should focus their service quality improvement on solving these problems and devote themselves to proposing various personalized solutions.

2. Related Works

Apriori algorithm is one of the most classic association rule mining algorithms, which was proposed by American scholar R. Agrawal. The Apriori algorithm consists of two main steps, the first step is to obtain the set of frequent items from the transaction records; the second step is to obtain the association rules based on the set of frequent items [5]. The Apriori algorithm is not without drawbacks, it has multiple iterations of recording things, resulting in I/O that cannot be done quickly, and can create multiple candidates sets as well as the problem of too frequent itemsets, which can make this algorithm less useful. For the problems described above, Chung et al. proposed the FP-growth algorithm without generating candidate mining frequent itemsets in the early twentieth century, but this algorithm has limitations when the amount of data is relatively large [6]. After that, association rule mining algorithms based on data partitioning, hashing, sampling, and dynamic itemset techniques were proposed one after another. To address the shortcomings of the Apriori algorithm, many scholars have made various changes based on the core of the algorithm, to name a few, implementing a reduction method with the records of the dataset to reduce the number of scans, or the combination of MapReduce and Apriori algorithm discovered by Kim and Chung, and then this genetic algorithm-based rule mining algorithm discovered by Kabir to Different analytics need to be integrated with other techniques as a benchmark [7]. In addition, the transformation of the transaction database into another data is also an effective method of progress, for example, the Apriori algorithm for mining multiple item sets can be used in a matrix, chain tables, etc. Therefore, it can be concluded that the Apriori algorithm has the defects of low time efficiency, and the strong association rules that may be brought about by the improper setting of the minimum support threshold and the minimum confidence threshold have no obvious practical significance.

Currently, the largest search engines are Google and Yahoo from 2004 yahoo's MyWeb to 2005 Google's search history, personalized search function by analyzing specific user search needs to limit the scope, and use this filter search results to give relevant results. In addition, there are also personalized search engines for the public, in which Google proposes an active analysis method and yahoo uses post-search restructuring for personalized search services [8]. Palmer has calculated the relevance by analyzing the hierarchical relationship between information to achieve query expansion for user search, and Fellbaum has realized the application of ontology in the field of personalized search by building WordNet, an ontology library. When users search, the structure will calculate the distance value between search results and user model tags for filtering and sorting, and realize a personalized search function [9]. The large search engine also includes Baidu and Sogou, among which, Baidu provides personalized search by providing users with a platform for search settings and access to result collection, which will be presented according to the history of previous searches in the next search, and combined with recommendation technology, applied to advertising, by analyzing user Taobao data, customized advertising for different users to improve efficiency [10]. Sogou has provided users with more accurate and personalized search information or goods by launching an exploration engine that values and amplifies users' personalized needs. In addition to search engines, search technology research experts have also proposed improvement methods, such as Professor Zhang, Huang establishing ontology knowledge and semantic analysis based on it, establishing user models and resource text libraries, supplementing the keyword deficiencies caused by the lack of user knowledge architecture, and realizing personalized resource information systems [11].

3. Construction of Personalized Search Model for Travel Users Based on Matrix Clustering Apriori Algorithm

3.1. Matrix Clustering Apriori Algorithm Model Design. The computational effort in mining frequent itemsets is much larger than that in mining association rules. This is because mining frequent itemsets require obtaining the support of itemsets from many transaction records in various ways. This includes, for example, generating candidate itemsets, traversing transaction records to count the support of candidate itemsets, and so on. The process of counting the confidence is after mining the frequent itemsets [12]. The Apriori algorithm obtains the confidence of the association rules simply by iterating through the sets, knowing the support of the frequent itemsets. By reducing the number of candidates itemsets through the two properties used in this algorithm, avoids generating too many candidate itemsets and scanning the database too much to count the support of the candidate itemsets. If an itemset X is a frequent itemset, then all subsets of this itemset satisfy the condition $x \in X$. And the contained transaction records must satisfy $X \subseteq t$, so $x \subseteq t$. That is, a transaction record containing an itemset X must contain a subset of the itemset X. So, if the support of X is greater than the minimum support, then the support of the subset of the item set X must also be greater than the minimum support because the number of transaction records containing a subset of X is greater than the number of transaction records containing the item set X. For example, if the itemset $\{b, c, d\}$ is frequent, then the itemsets $\{b]$. $\{c]$, $\{d]$, $\{b, c\}$, $\{b, d]\}$. $\{c, d]$ must all be frequent itemsets as well. As shown in Figure 1, the gray-filled nodes are frequent itemsets.

Let us analyze the classic "beer and diapers" example. If there are 10,000 purchase transactions in a supermarket, we know that 4500 transactions contain both beer and diapers, 7000 transactions contain beer and 5000 transactions contain diapers, and we set a minimum support threshold of 30% and a minimum confidence threshold of 40%. According to the definition, the following association rules can be mined.

(a) Buy beer > buy diapers (48% support, 60% confidence)

(b) Buy beer > no diapers (32% support, 40% confidence)

However, both association rules satisfy minimum support and minimum confidence, and both are strong association rules, but the rule (a) and rule (b) contradict each other.

Therefore, it can be concluded that the Apriori algorithm suffers from time inefficiency and the shortcoming that the strong association rules may not have obvious practical significance due to improperly set minimum support threshold and minimum confidence threshold. It is necessary to introduce the degree of interest to improve the accuracy of association rule extraction. The degree of correlation PI between itemsets in association rules $X \longrightarrow Y$ represents the degree of interest of association rules. Under the premise of statistical independence, interest Degrees represent the ratio of the true intensity to the desired intensity.

The traditional Apriori algorithm uses support as the determinant for generating candidate item sets, and some of the generated association rules have meaningless "pseudo" association rules which are easy to cause "illusion" [13]. Therefore, it is necessary to introduce the degree of interest to improve the accuracy of association rule extraction. The degree of interest of an association rule is represented by the PI of the correlation between the sets of items in the association rule $X > Y$. Under the assumption of statistical independence, the degree of interest represents the ratio of the true strength to the desired strength. From the above, when $\mathrm{PI} > 1$, it means that X is positively correlated with Y. The larger the PI is, the higher the correlation between X and Y, that is, the correlation rule is more relevant in practice and more practical and realistic. The effect of introducing the interest degree on the extraction of association rules is that the introduction of the interest degree can filter the association rules with no correlation or negative correlation, further reduce the number of candidate association rules (the candidate association rules here refers to the association rules just generated by the frequent item set that have not yet been compared with the minimum confidence level), and ensure that the association rules judged by the confidence threshold are all with realistic.

The parallelization method in the multidimensional association rule mining algorithm based on the Apriori property, first dividing, and processing multidimensional data sets is implemented by HDFS, where small data files are split by slightly larger data modules. Map chunking processes the input most primitive data sets, using the parallel programming model of MapReduce for key values for characteristics, which are distributed in chunks by the head process to the chunked data is then processed in parallel algorithms on each computer. Because of the MapReduce mechanism, the chunk candidate sets are calculated by the Map process on each computer in the above process, and then the k-item candidate set support numbers of all chunks obtained by the Map process on each computer are combined to obtain the global k-item candidate set support numbers, which is implemented by the Reduce process operation in the MapReduce programming model [14]. A frequent itemset is calculated from the support of the global candidate set. Each computer then starts the next Map process to process the second data chunk, which is done after

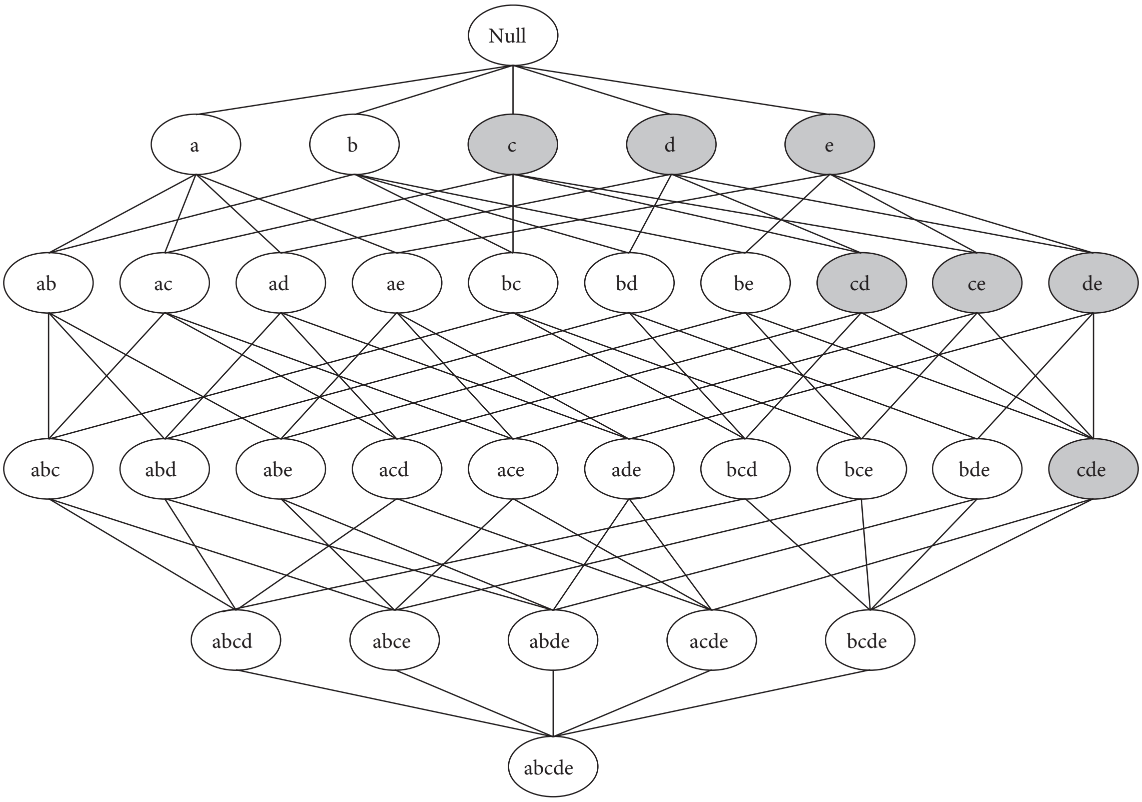

Figure 1: The nodes filled in gray are all frequent itemsets.

the global k-item frequent set of a data chunk is computed, and so on until all the data chunks are processed, as shown in Figure 2.

The LDA topic model cannot accurately model the latent topic, so it cannot accurately analyze and obtain its score on the hidden topic: when $K > 25$, the relationship between topics is too high, which leads to Decreased accuracy of rating predictions on. At the same time, the greater the number of topics, the less the user's rating results on each potential topic will be. To improve the timeliness of association rule mining and the value of the generated strong association rules, two aspects of work are done to address some shortcomings of the Apriori algorithm. We introduce the degree of interest to ensure the value of strong association rules and parallelize the improved algorithm using the Spark platform to improve its timeliness. Designing an improved Apriori algorithm MC-Apriori parallelization scheme based on Spark platform. The deployment of Spark is described in detail, and the efficiency of the algorithm is tested using the dataset generated by the data generator. The experimental results demonstrate that the parallelization scheme of the MC-Apriori algorithm based on the Spark platform can effectively improve the efficiency of association rule mining in the massive data processing.

Clustering belongs to unsupervised learning, the purpose of which is to assign samples that are close (similar) to the same class and samples that are not close (similar) to different classes, this method has been widely used in image retrieval, user classification, and other fields. In clustering, the most central thing is similarity or sample distance, and many different methods to calculate sample similarity or sample distance have emerged during the development of clustering [15]. Non-parametric estimation (non-parametric density estimation) does not specify a mathematical model (e.g., normal distribution, binomial distribution), but uses learning samples to estimate the mathematical model. Nonparametric does not mean that the method does not require parameters, but indicates that the parameters are variable and can be considered as individualized parameters in a recommendation system. The common nonparametric estimation includes a histogram, Kernel Density Estimation (KDE), and K-nearest neighbor estimation, and nonparametric estimation has been widely used in the field of risk prediction such as stock and finance. In this paper, the kernel density estimation method is used in the establishment of the user consumption model, the consumption characteristics of tourists cannot be constructed using a uniform model, and one of the advantages of taking this method is that it can reflect the personalization of user consumption, as shown in equations (1) and (2).

$$f_n(x) = \frac{1}{nh}\sum_{i=1} kh\left(\frac{x + x_i}{h}\right), \tag{1}$$

$$\text{MISE}(h) = d f_n(x) - x f(x)^2. \tag{2}$$

3.2. Travel User Personalized Search Model Construction. Explicit interest information, implicit interest information, and geographic context interest information can be obtained

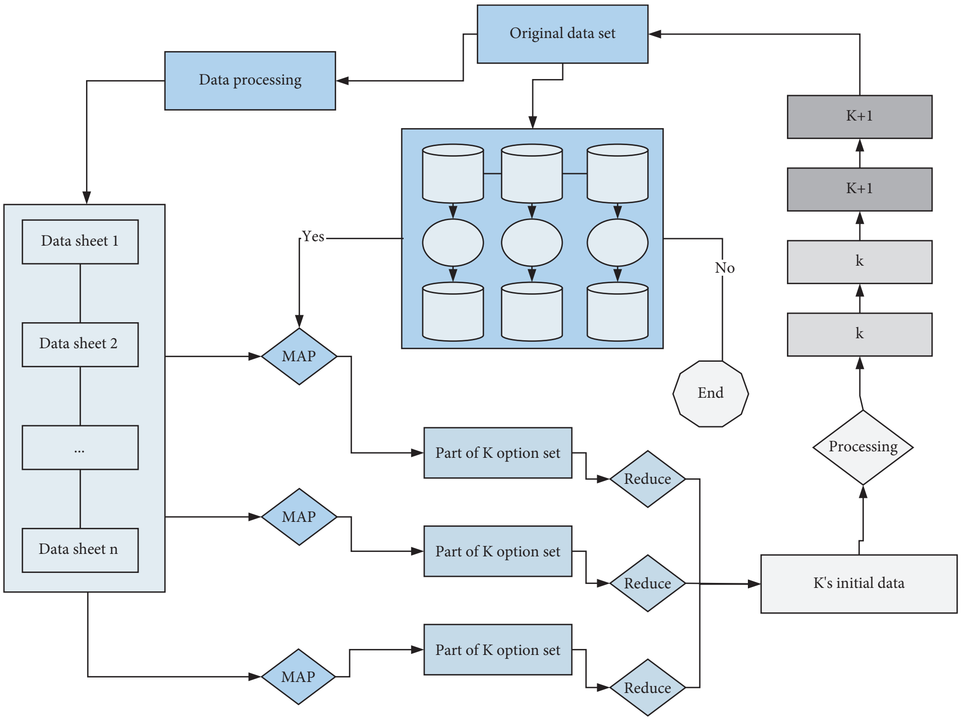

Figure 2: Flow chart of MC-Apriori algorithm parallelization.

from three aspects of user interest information, explicit information including personal information submitted by the user when registering for a web page, and other information such as user interest keywords, while implicit information is based on user browsings behavior information, such as user favorites, downloads, printing, and other behavioral information, the user's time spent browsing a web page and other information. In addition, the geographic context information mainly includes the user's surrounding geographic environment, the user's current geographic location, and the current geographic scene, and other related context information, and the above information is used to mine the user's comprehensive preference characteristics [16]. Users' interests and preferences do not always remain the same, users' interests are endlessly changing, and users' interests are divided in different ways. Users' interests can be classified into explicit interests and implicit interests according to the different ways of acquiring interests; users' interests can be classified into long-term fixed interests and short-term interests according to the length of time they stay interested in things; users' interests can also be classified into stable interests and immediate interests according to the persistence of certain types of interests. In the calculation process, the correlation degree r of the parent topic S and S and the distance d of the area where the instance pair is located are mainly considered. (1) Relevance r between parent topics S and S: where S is the parent node of I, and S is the parent node of I. The higher the correlation of the topic nodes to which the two instances belong, the higher the corresponding correlation of the instance. This article mainly uses user interest acquisition methods to classify user interests and divides user interests into explicit interest and implicit interest. The explicit interest is mainly based on the user's multidimensional tags, and the implicit interest is based on the user's browsing behavior. On this basis, geographic contextual interest is introduced, and the structure diagram of the user personalized search adaptation model is shown in Figure 3.

The number of different topics K directly determines the accuracy of the recommendation algorithm. This experiment K takes the value range of [10, 50], and the experiment starts from K taking 10 and increments 5 each time to end at 50, for a total of 8 times, setting the number of nearest neighbors $N = 20$ and the distance weight = 0.5. It can be seen from the figure that the algorithm achieves the best accuracy, coverage, and recall when $K = 25$. The reason for this is that when $K < 25$, the LDA topic model cannot accurately model the potential topics, so it cannot accurately analyze their scores on the hidden topics: when $K > 25$, the tightness of the connection between the topics is too high, which leads to a decrease in the accuracy of the score prediction on each topic. Also, the higher the number of topics, the lower the user's rating on each potential topic. The experimental results are shown in Table 1.

The line graphs of accuracy, recall, and coverage for the different number of topics K are generated from the experimental results as shown in Figure 4.

Dwell time refers to the time that a user spends at a location and is $T_s(p_x)$ expressed using a photo dataset. The photo dataset is used, so its method of calculating the user's dwell time at a location is the time difference between the photos taken. In this paper, it is not possible to extract the dwell time of users in each location from the dataset, so this paper adds the basic dwell time to each location (the dwell

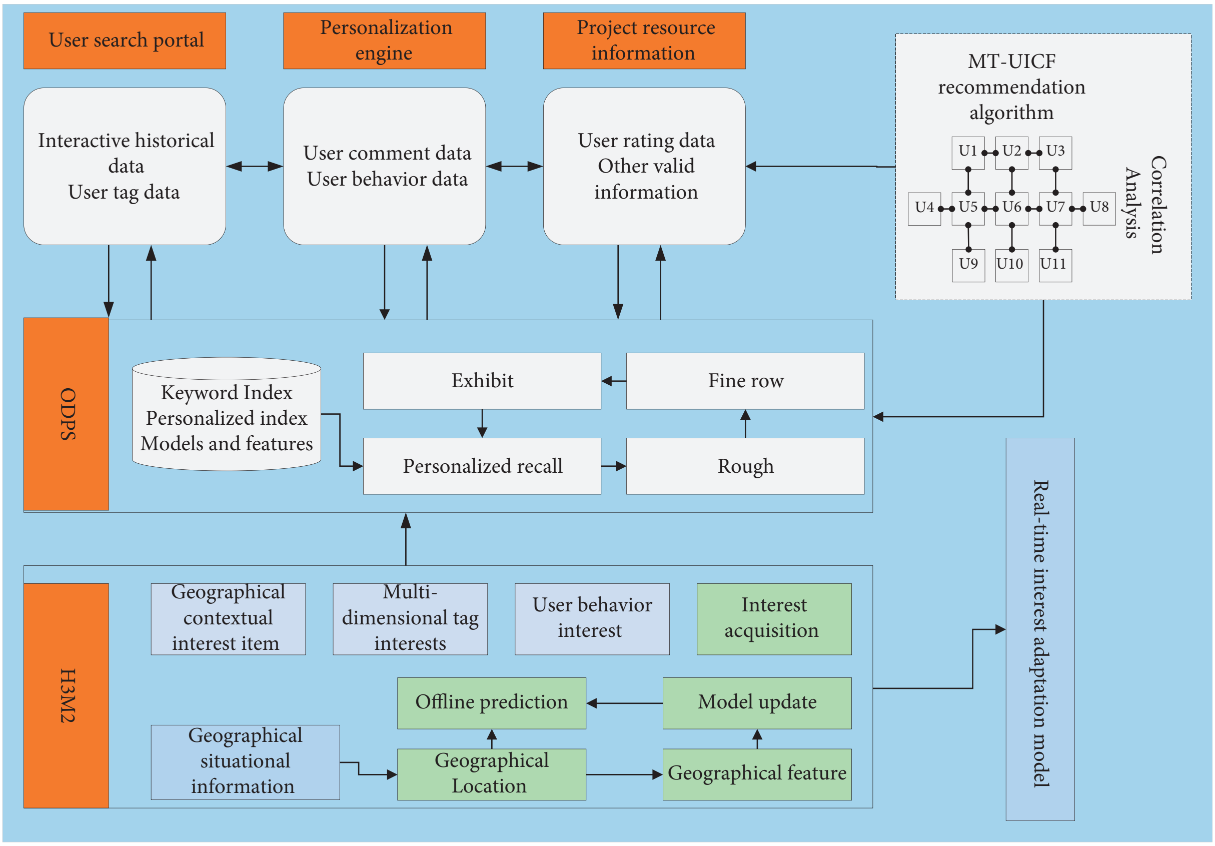

Figure 3: User personalized search model.

Table 1: Comparison of experimental results at different K.

K	10	15	20	25	30	35	40	45	50
Accuracy	23.54	23.58	28.93	33.87	31.82	31.07	27.53	27.12	30.25
Recall	11.03	11.09	11.64	13.54	14.25	12.17	12.21	11.47	13.78
Coverage	36.35	38.37	42.31	47.42	47.72	37.87	37.51	35.64	45.15

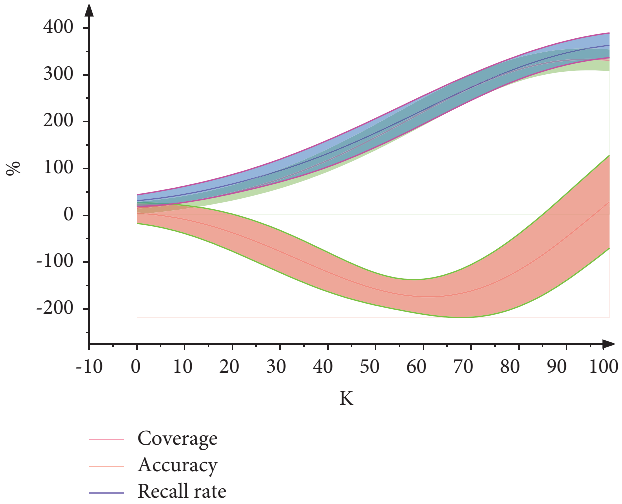

Figure 4: Comparison of experimental results at different K.

time is generated after arriving at an attraction, regardless of whether the attraction attracts users or not), and both the user's preference and the popularity of the location itself also affect the dwell time, so this paper takes the popularity of the location and the user's preference as the factors affecting the dwell time, and its mathematical expression is as follows:

$$T_s(p_x) = \frac{\max\left(\mathrm{Score}\left(p_y, \in p_{x\cdot c}\right)\right)}{\mathrm{Score}_{\left(p_{x\cdot c}\right)}} - T_c. \tag{3}$$

In equation (4), the dwell time is obtained in hours, $Max(Score_{(p_y,\in p_{x.c})})$ and the maximum value of the attraction score in the category to which the location p_x belongs ($Score_{(p_y,\in p_{x.c})}$) is the location score calculated according to the method. T represents the base time of the dwell time, and according to practical considerations, the dwell time at the beginning and end of the route is set to 0 in this paper. $Pop(l)$ is the popularity of the location, and $p_{\mathrm{re}_c}(u, l.c)$ is the user's preference for the category $l.c$. Trip time refers to the total time required by tourists to follow the itinerary, i.e., the total time required for the travel route is the sum of all the dwell time and the transition time of all the attractions from the user's starting location during the trip.

$$D_{ij} \approx \min_k \left| x_{kj} + x_{ki} \right|, \tag{4}$$

$$\sin \alpha = \frac{e_{ij}}{\left|e_i\right| \cdot \left|f_j\right|}. \tag{5}$$

Tourists are the main users of scenic intelligent tour guide systems, and tourists' experience reflects the quality of scenic tourism services to a certain extent and determines the length of the survival cycle of mobile tourism applications. Although there are thousands of tourists using the scenic intelligent tour guide system, their respective personalities, travel preferences, and spending power are different, and through the behavior-preference model data mining algorithm, some commonalities of different tourists' experiences can be obtained [17]. In the Node-Apriori algorithm, even if the transaction records are stored in binary encoding, it will still occupy a large amount of memory in the process of mining frequent itemsets. The Node-Apriori algorithm uses a limited breadth method when mining frequent itemsets. Nodes at the same level contain transaction records that may be repeated, which requires more memory. The data obtained from the analysis can get detailed information on tourists' purchase of attraction mouth tickets and pre-coincident scenic services. The basic information such as gender, age, education, income, etc. of visitors is obtained through the visitor ID account, and then pre-processing operations such as data cleaning and sorting are carried out to initially obtain the basic information that visitors to scenic spots are mainly between 22 and 35 years old and come from travel enthusiasts and Chinese culture lovers from all over the world. Combined with the visitors' historical query, browsing, and consumption records, we can dig out the consumption ability level of the tourists in the scenic spot, and we can find that the consumption-ability of the tourists in the scenic spot is mainly concentrated in the middle level. Similarly, the mining of tourists' data can analyze that the preference of "romantic" tourists visiting Tianyahaijiao scenic spot is significantly higher than that of "literary" tourists. As shown in Figure 5, the first search for neighboring groups of tourists will divide them into romantic, adventurous, literary, and other tourist target groups, and then analyze the enthusiasm of different tourist target groups for the attractions based on their browsing traces, transaction details, and other related contents. So, that businesses can make favorable decisions based on the results of data analysis, better grasp the psychology of tourists, explore more target tourists, and locate target tourists.

4. Analysis of Results

4.1. Matrix Clustering Apriori Algorithm Model Results. The relationship between instances is different from that of topics, which contains not only hierarchical relationships but also the degree of influence of regions. In Chapter 3, the weight map has been constructed and the information of the regions where the attractions are located between them has been recorded, so the regional influence factor is added in this part of the calculation process [18]. In the calculation process, we mainly consider the correlation degree r of the parent class theme S and S and the distance d of the instance to the region where it is located. The constructed user interest model and model need to continuously supplement and expand the relevant instance and attribute information with the development of the user's application and attractions, so as to continuously ensure and improve the precision and recall rate requirements in the experimental results of the system.

(1) The correlation degree r of the parent class theme S and S: where S is the parent node of I and S is the parent node of I. The higher the correlation degree $r = 1$ when I and coincidentally belong to the same instance of a parent class theme, and $r = 1$ when they belong to two themes respectively when they belong to two topics respectively, we get the correlation $r = R$ (S_1, S_2), the larger this correlation is, the higher the correlation between the instance and the instance will increase with it.

(2) The distance of the instance pair's region d: according to the region weight map derived in Chapter 3, the distance of the instance pair's region is calculated, and the closer the distance is, the higher the instance pair's relevance is, and vice versa, the lower it is.

Considering the above two parts together, the instance-to-instance correlation is calculated as shown in equation (3), where β is the same as the adjustment coefficient.

$$R(I_1, I_2) = r\beta \frac{r}{d}. \tag{6}$$

In the user interest model constructed according to the attraction area weights, the user's query process is divided into a precise query and fuzzy query. When the user makes a

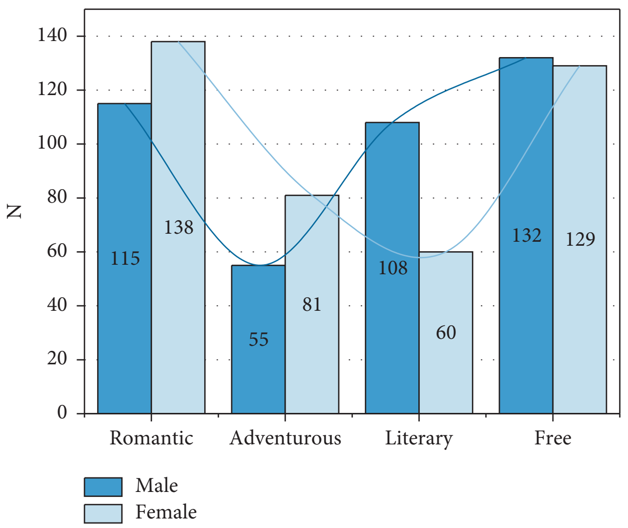

Figure 5: Trend of users' enthusiasm for travel.

precise query, to a certain extent, it indicates that the user is searching for a certain attraction, and at this time, the user's unknown attraction is solved and the gap of the user's travel knowledge system is filled by recommending relevant and high-priority attractions for the user as functional extensions. When the user makes a fuzzy query, the traditional meaning of keyword expansion is carried out. With the continuous expansion of data sets, traditional data mining technology has been unable to quickly and effectively mine and analyze massive data sets. How to efficiently and efficiently discover the rules and hidden connections contained in the data set is a hot issue. That is, according to the keywords entered by the user, query all the attractions that meet their semantic information, and through the designed algorithm to calculate the priority level of these attractions for the user, to sort and search the results from the document library to display to the user. For example, if you type in "climbing," the website will normally rank and display the results according to the hotness of the relevant tourist attractions [19]. However, the user has informed the system of his purpose, that is, climbing mountains, if the user lives in Haidian District, the closer Xiangshan Mountain is a higher priority than the distant Shilin Gorge, then all the attractions of the mountain theme category in Beijing should be displayed for the user because they are all mountains, i.e., the same theme relevance weight, so the priority of attractions depends mainly on the regional weight, at this time, according to the different regional attractions of different, according to different regional attractions different regional weights sorting, for different users to come up with different attractions sorting and display results.

Step 1. Judge the correlation value, if it is less than 0 or more than 2, it is illegal, go to step 7;

Step 2. Analyze the keyword type and find the matching node level according to the topic and instance information structure diagram, if it is a topic node turn to step 3, if it is an instance node then turn to step 6;

Step 3. Judge the relevance value, if it is less than 1, move to step 4; otherwise, move to step 5;

Step 4. Get the instance nodes under the topic, and return the sequence of nodes whose relevance to the keyword is greater than the user input value, move to step 7;

Step 5. Get all the instance nodes under the topic and its sibling nodes, and return the sequence of instance nodes whose correlation with the keyword is greater than r ($r = 2$-user input value), and turn to step 7;

Step 6. If the correlation is less than 1, return the keyword instance node and move to step 7, otherwise, get the sibling node of the instance node and return the instance node whose correlation with the keyword is greater than r ($r - 2$-user input value) and move to step 7;

Step 7. End of query expansion.

After the above steps, the number of failed records is 121 and the number of qualified records is 4732, and the percentage of failed records in the total number of results is 2.5%. When the support degree is greater than 2.5%, the association rules for the records of unqualified sampling cannot be mined. To find the results related to the nonconformity, it is necessary to set the support degree less than 2.5%. The minimum support set in this experiment is 0.3% and the confidence level is 30%. The number of frequent items mined by the Tree-Apriori algorithm in the first step is 3208, and the number of association rules mining in the second step is 16,082, where the scatter plot of support and confidence is as follows, with support as the horizontal coordinate and confidence as the vertical coordinate.

The observation of the graph shows that the association rules decrease with the increase of support, but the association rules with higher confidence distributed in all support levels. The number of frequent itemsets mined in this data mining process is 3208, and the number of association rules mined based on frequent itemsets is 16,082. Some of the data have less support but have higher confidence (Figure 6).

The Tree-Apriori algorithm combines the generation of frequent itemsets with transaction records and directly uses transaction records to generate frequent itemsets instead of isolating frequent itemsets from transaction records as in Apriori, FIMST, or Node-Apriori. The number of transaction records that need to be traversed for a candidate itemset is expressed in the sense that each frequent itemset contains transaction records behind it. Then it is possible to mine the frequent itemsets directly by processing the transaction records. This avoids the need for the algorithm to traverse transaction records frequently when counting the support of candidate itemsets, even though the Node-Apriori algorithm reduces the number of transaction records that need to be traversed to count the candidate itemsets. The Node-Apriori algorithm does not need to store potentially duplicate transaction records in the nodes as the Node-Apriori algorithm does, but even if the transaction records are stored by binary encoding, it still takes up a large amount of

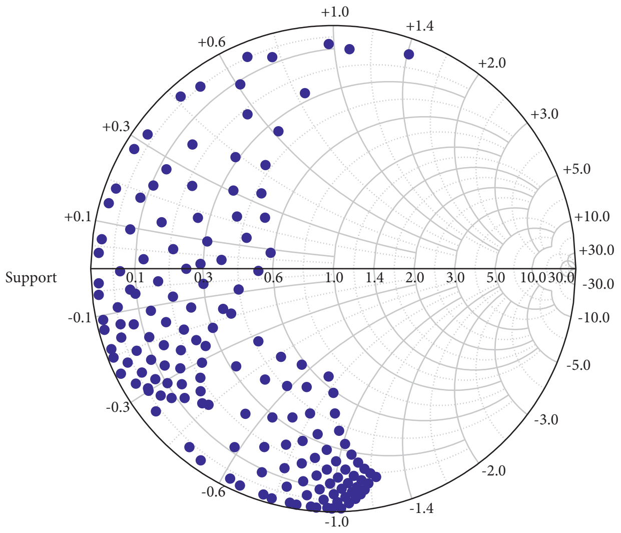

Figure 6: Scatter plot of support and confidence of association rules.

memory in the process of mining frequent itemsets. As the technology continues to mature, the Hadoop distributed framework is introduced as a sub-project into daily R&D with increasing proficiency. At present, combining the Hadoop platform with data mining and using the parallelization algorithm of the Reduce model in Hadoop is also an important application in data mining. The Node-Apriori algorithm mines the frequent itemset in a breadth-limited way and the nodes of the same level contain transaction records that may be repeatedly sent, which is more memory intensive. In contrast, the Tree-Apriori algorithm uses a depth-limited approach, and a merged tree can be deleted after mining the frequent itemsets to avoid the need for large memory.

4.2. Tourism User Personalized Search System Implementation. With the increase in the number and types of points of interest recommended by the platform, it is difficult for people to choose high-quality locations according to their own preferences; it becomes extremely difficult for people to extract effective information, which makes the feedback of information decrease day by day. The key techniques for information search include similarity matching techniques for user interests, semantic analysis techniques for attraction topics, and query expansion techniques. These techniques are all methods closely related to the attraction hierarchy, attraction attributes, and attraction classification in the tourism domain [20]. Therefore, the constructed user interest models and models need to be continuously supplemented and extended with the information of relevant instances and attributes with the development of user applications and attractions, to continuously ensure and improve the accuracy and recall requirements in the experimental results of this system. The implementation of the system includes, on the one hand, the application of user interest model, model, and query expansion techniques to improve the search results to provide users with higher recall and precision, as well as more optimized sorting of information search results; on the other hand, the maintenance and expansion functions of the knowledge base should be designed to ensure that user feedback can be collected promptly while expanding the node information of the attractions. After the analysis of the requirements overview, to better realize the function of personalized search, this system will be divided into two major functions: offline information processing and online personalized search. The comparison results of the recall and F-measure values can be seen in Figure 7, which shows the comparison results and the changing trend of both. Personalized recommendation systems use users' previous behavior information, such as purchase records, to push products that users are interested in to users, so as to reduce the time users spend looking for the information they need. The recommendation system can generate great economic benefits, so the personalized recommendation system has been widely used in many fields such as movies, music, personalized advertisements, and location-based services.

In the user's precise search, the system provides the user with the nearest and topic-related tourist attraction information recommendation to meet the user's potential expectation that the tourist attraction is closer to the user's location, and the attraction is the type of attraction he wants to visit. Take a user who lives in Pinggu and searches for tourist attractions as an example, in the spring season, she wants to climb mountains to get fresh air and exercise, and the mountains are famous for "Fragrant Mountain," so the user searches for the keyword "Fragrant Mountain," and it can be seen that the user searches for Fragrant Mountain. When the user searched for "Fragrant Mountain," she was given personalized recommendations on the left side of the page, among which "Jingdong Grand Canyon" and "Shilin Gorge" are both scenic mountain canyons in Pinggu, which can satisfy her desire to climb mountains and are close enough not to consume too much of the user's time on the way there and back. At the same time, tourists are more active in the travel process, hoping that scenic spots can provide comprehensive and detailed personalized services, while traditional tour guide services can no longer meet the increasingly diverse individual requirements of different tourists. The distance is close enough that it does not take too much time for the users to travel to and from the mountain, which is perfectly in line with the users' mood to take a day for a weekend outing. By comparison, it can be found that the search results of the system implemented in this thesis are more in line with users' needs, firstly making up for the search knowledge, and secondly providing users with closer travel information on similar topics, satisfying their potential travel expectations. In this section, the system results are visualized by inviting users to use the system and by a user satisfaction comparison chart. Five users are invited to use the system for precise search and score the recommended results respectively to get an overall satisfaction result from the users.

The most obvious performance test of the system is to examine the system carrying capacity, by allowing more users to access the system at the same time and carry out

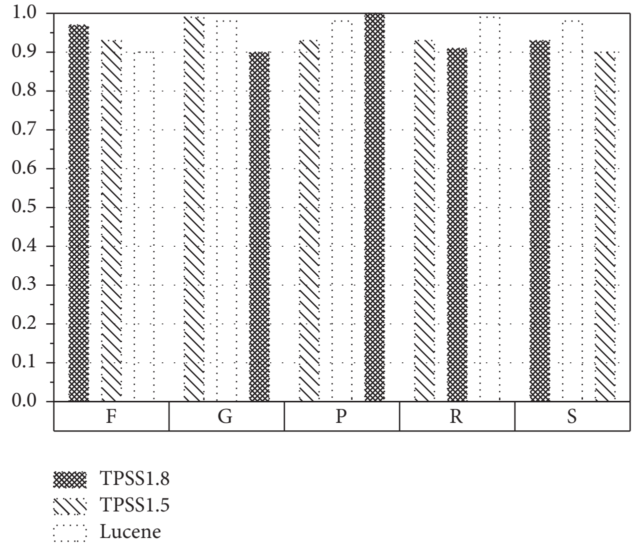

Figure 7: Accuracy, recall, and metric values.

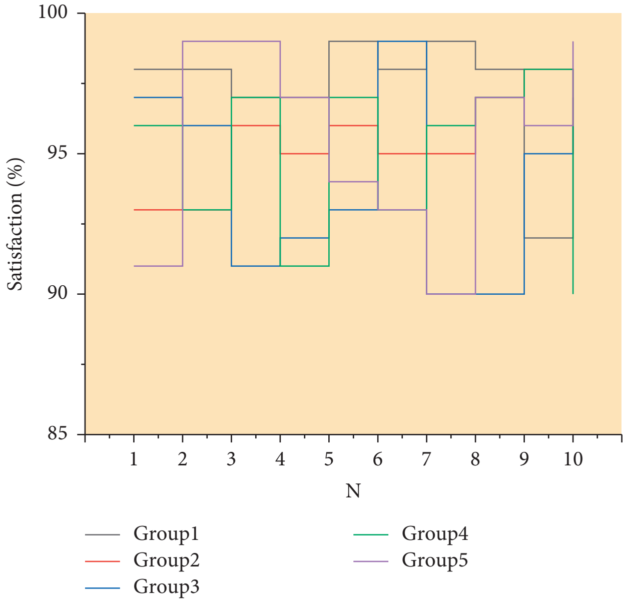

Figure 8: Search recommendation result satisfaction.

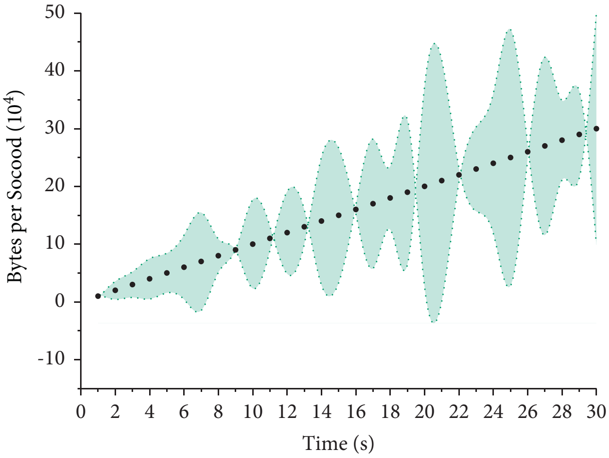

Figure 9: Model performance test results.

some work processes, the system's response efficiency is expected to process the efficiency of the process to examine, through the actual data obtained from the examination of the system performance calculation and analysis, if the calculated data can meet the corresponding standards, it means that the system performance test passed. In the system performance test, we need to examine the stability of the system operation, and analyze the data and information standards obtained from the test process to see if they can meet the test standards, which is used to determine the degree of perfection of the system development. In this paper, Load Runner, a common system performance testing software, is used to simulate the user's environment and user's operation behavior, and based on this state, the system parameters are counted and the changes are analyzed to get the final test results (Figure 8).

"Hits per Second" is the number of requests sent by the client to the server per second, which is proportional to the "Average Throughput." This chapter mainly tests the function, performance, platform page, and user satisfaction of the system. The system function test mainly includes the function completion and fault tolerance test; the system performance test is conducted from the number of clicks per second, throughput rate, response time, etc.; the system interface test detects the correct link from all sectors; the user satisfaction of the recommendation result is obtained by issuing questionnaires. Through system testing and result analysis, the security and applicability of the system are ensured (Figure 9). The Apriori algorithm is not without its shortcomings. It has to record things repeatedly, resulting in the inability to complete the IO quickly, and there will be a variety of candidate sets and the problem of too many frequent itemsets, which will reduce the effect of this algorithm.

The continuous development of Internet technology has made users publish and share information resources more and more freely, and a large amount of travel-related information has emerged on various social media platforms. The acquisition of travel information is the cornerstone of making travel recommendations, and it is crucial to obtain accurate, effective, and useable travel information. In this paper, we consider three types of tourism information: tourism user information, tourism attraction information, and tourism user evaluation information about tourism attractions. The above three types of information can be further subdivided, for example, in personalized recommendation, tourism user information includes user static information such as name, gender, etc., user preference information such as like autumn travel, like natural landscape type attractions, etc., user behavior event information such as climbing mountains, swimming lake, shopping, etc., tourist attraction information includes the name of the attraction, attraction address, attraction star, attraction business hours The information of tourist attractions includes the name of tourist attractions, address of tourist attractions, star rating of tourist attractions, opening hours of tourist attractions, etc. The evaluation information of tourist attractions by tourist users includes the rating information of tourist attractions by tourist users, the comment information of tourist attractions by tourist users, etc.

5. Conclusion

This thesis first analyzes the importance of personalized search and user needs, and then introduces theories and key technologies related to personalized search and commonly used algorithms, etc. Examples are given to analyze its distance attribute, topic attribute, extended concept attribute, and instance attribute, and based on such information applied to the creation of user interest models and models, a dual-model personalized search system for the tourism field is finally realized to meet the different needs of users. In this paper, for the special problems faced by personalized recommendation in the tourism field, a method of MC-Apriori improvement by matrix clustering is proposed, which only needs to scan the database once and generate a series of different clustering matrices, while only partial clustering matrices need to be calculated to produce frequent itemsets. The introduction of the user's reading content not only reflects the user's current interest tendency, but also ensures the accuracy, variety, and individuality of the recommended content, and at the same time improves the search efficiency and reduces the workload by using the improved MC-Apriori algorithm. Through extensive testing of the collected corpus data, it is demonstrated that the travel recommendation model proposed in this paper has a certain application value compared with the traditional recommendation model. The recommendation algorithm is to be further studied for better recommendation results.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the Key Research Project of Social Science in Henan Province, Research on the integration of "Two mountains and two fists" in Henan Province under the background of the three-dimensional integration of "Literary style and travel" (SKL—2020—2646).

References

[1] S. Rahimipour, M. Ghatee, and S. M. Hashemi, "A hybrid of neuro-fuzzy inference system and hidden Markov Model for activity-based mobility modeling of cellphone users," *Computer Communications*, vol. 173, pp. 79–94, 2021.

[2] G. Qiu, R. Song, S. He, W. Xu, and M. Jiang, "Clustering passenger trip data for the potential passenger investigation and line design of customized commuter bus," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 9, pp. 3351–3360, 2018.

[3] G. Hong and G. Nan, "Research and application of a multidimensional association rules mining algorithm based on Hadoop," in *Proceedings of the 2021 IEEE International Conference on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)*, pp. 636–643, New York, NY, USA, September 2021.

[4] G. H. Lee and H. S. Han, "Clustering of tourist routes for individual tourists using sequential pattern mining," *The Journal of Supercomputing*, vol. 76, no. 7, pp. 5364–5381, 2020.

[5] C. L. Hsu, "A multi-valued and sequential-labeled decision tree method for recommending sequential patterns in cold-start situations," *Applied Intelligence*, vol. 51, no. 1, pp. 506–526, 2021.

[6] K. Chung, H. Yoo, and D. E. Choe, "Ambient context-based modeling for health risk assessment using deep neural network," *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, no. 4, pp. 1387–1395, 2020.

[7] J. C. Kim and K. Chung, "Associative feature information extraction using text mining from health big data," *Wireless Personal Communications*, vol. 105, no. 2, pp. 691–707, 2019.

[8] C. Bhadane and K. Shah, "Context-aware next location prediction using data mining and metaheuristics," *Evolutionary Intelligence*, vol. 14, no. 2, pp. 871–880, 2021.

[9] F. Mannhardt, A. Koschmider, N. Baracaldo, M. Weidlich, and J. Michael, "Privacy-preserving process mining," *Business & Information Systems Engineering*, vol. 61, no. 5, pp. 595–614, 2019.

[10] Z. Cui, X. Xu, X. U. E. Fei et al., "Personalized recommendation system based on collaborative filtering for IoT scenarios," *IEEE Transactions on Services Computing*, vol. 13, no. 4, pp. 685–695, 2020.

[11] O. Ben-Assuli, T. Heart, J. R. Vest, R. R. Gonen, N. Shlomo, and R. Klempfner, "Profiling readmissions using hidden Markov model-the case of congestive heart failure," *Information Systems Management*, vol. 38, no. 3, pp. 237–249, 2021.

[12] W. F. Shih, C. W. Lin, W. F. Wang, and H. H. Wu, "Association rule mining of care targets from hospitalized dementia patients from a medical center in Taiwan," *Journal of Statistics & Management Systems*, vol. 21, no. 7, pp. 1299–1310, 2018.

[13] A. P. Phyu and E. E. Thu, "Short survey of data mining and web mining using cloud computing," *International Journal of Advanced Networking and Applications*, vol. 12, no. 5, pp. 4725–4731, 2021.

[14] J. C. Kim and K. Chung, "Neural-network based adaptive context prediction model for ambient intelligence," *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, no. 4, pp. 1451–1458, 2020.

[15] Y. Sun, H. Qiang, J. Xu, and G. Lin, "Internet of Things-based online condition monitor and improved adaptive fuzzy control for a medium-low-speed maglev train system," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 4, pp. 2629–2639, 2019.

[16] J. C. Kim and K. Chung, "Discovery of knowledge of associative relations using opinion mining based on a health platform," *Personal and Ubiquitous Computing*, vol. 24, no. 5, pp. 583–593, 2020.

[17] R. Angmo, N. Aggarwal, V. Mangat, A. Lal, and S. Kaur, "An improved clustering approach for identifying significant locations from spatio-temporal data," *Wireless Personal Communications*, vol. 121, no. 2, pp. 985–1009, 2021.

[18] L. D. Xu and L. Duan, “Big data for cyber physical systems in industry 4.0: a survey,” *Enterprise Information Systems*, vol. 13, no. 2, pp. 148–169, 2019.

[19] E. Fisher and Y. Mehozay, “How algorithms see their audience: media epistemes and the changing conception of the individual Media,” *Culture & Society*, vol. 41, no. 8, pp. 1176–1191, 2019.

[20] N. Ali, A. Fatima, and H. Shahzadi, “Online reviews & ratings inter-contradiction based product’s quality-prediction through hybrid neural network,” *Journal of the Institute of Electronics and Computer*, vol. 3, no. 1, pp. 24–52, 2021.