JCSE Journal of Computing Science and Engineering

Regular Paper

Journal of Computing Science and Engineering,
Vol. 19, No. 3, September 2025, pp. 81-93

Design of Smart Tourism Data Mining Technology Based on Optimized Apriori Association Rule

Fen Ma*

Department of Cultural Tourism and Wellness, Chongqing College of Finance and Economics, Chongqing, China
xm20161231@126.com

Abstract

With the widespread application of information technology, smart tourism has emerged as an important direction for the development of the tourism industry. To improve the technological level of the smart tourism industry, a technology based on optimized Apriori association rule mining algorithm is designed. The Spark in-memory computing framework is chosen as the distributed programming framework to store elastic distributed datasets in distributed memory. The support is used to measure the probability of two scenic spots appearing simultaneously, and the order of tourists visiting the scenic spots is considered. A complete data mining process is designed. In the test of the association rules, the research method generated 4,271 association rules at a confidence level of 0.056. When the minimum support threshold of the research method was 30%, the number of invalid frequent itemsets generated was only 831. In the analysis of operational accuracy, the research method maintained an accuracy of over 98.8% when analyzing a data volume of 10,000. This indicates that the research method has good data mining performance and faster computation speed during runtime. The research method can provide good technical support for smart tourism.

Category: Databases / Data mining

Keywords: Apriori; Data mining; Smart tourism; Parallelization; Frequent itemsets

I. INTRODUCTION

In the era of informatization, global economic integration and rapid development of information technology have brought about significant changes in people's lifestyles. Tourism is an important form of leisure, which is growing in size and the competition. Practitioners in the tourism industry are beginning to seek innovative solutions to meet the diverse and personalized needs [1, 2]. Smart tourism has emerged, integrating various information technologies to provide tourists with more convenient, comfortable, and intelligent travel experiences [3]. The core technologies of smart tourism include cloud computing, Internet of Things, artificial intelligence, etc. The Internet of Things technology enables tourism facilities and services to be interconnected in real-time. Cloud computing provides powerful data storage and computing capabilities. Big data analysis helps managers gain insights into tourist behavior and optimize services. Artificial intelligence technology provides personalized recommendations and intelligent customer service through machine learning, natural language processing, and other means [4, 5]. Data mining can extract key information about tourists, providing effective data support for the development of scenic spots, activity arrangements, and on-site management. However, the current mainstream

Open Access http://dx.doi.org/10.5626/JCSE.2025.19.3.81 http://jcse.kiise.org

Received 23 October 2024; **Revised** 19 September 2025; **Accepted** 21 October 2025
*Corresponding Author

 pISSN: 1976-4677 eISSN: 2093-8020

data mining techniques have high computational complexity and low efficiency when processing large datasets. The Apriori association rule mining algorithm can extract the connections between data and analyze the hidden meanings behind different data relationships. Under this background, the research attempts to innovatively propose a smart tourism data mining technology based on the Apriori association rule mining algorithm. The core innovation of the research method lies in significantly reducing the disk overhead in the process of generating frequent itemsets by residing the data set in distributed memory. And embed the temporal characteristics of tourists' behaviors into the computing process to reduce the set of invalid candidate items while maintaining parallel efficiency. The massive tourist behavior data accumulated by the intelligent tourism platform naturally contains rich associated information. Association rule mining techniques, especially the Apriori algorithm, aim to discover interesting and frequent co-occurrence relationships among itemsets from large-scale transaction data. These two technologies are complementary in terms of goals: the intelligent tourism system provides the data basis and application scenarios, while data mining based on Apriori offers the core analytical methods to reveal potential patterns (such as strong correlations between scenic spots and combinations of tourists' preferences) from the data. Therefore, they are not orthogonal but highly collaborative. The data mining method designed in this study is precisely oriented towards the intelligent tourism scenario. By optimizing the Apriori algorithm, it aims to efficiently mine the frequent patterns and strong association rules in tourists' behavior data, providing underlying support for the core functions of intelligent tourism.

II. RELATED WORKS

With the rapid development of the tourism industry, tourism information technology is also advancing rapidly. Taking relevant data and algorithms to design tourist visit prediction models has become a popular research topic. Scholars have conducted relevant research on smart tourism technology. Ballina [6] built a smart tourism research method on the basis of commercial information and communication technology to address the smart component factors of tourism companies. Based on studying limitations and sampling procedures, the tourism formula database was expanded. The proposed method had a positive effect on smart tourism. Rahmadian et al. [7] designed a method in software engineering that combined systematic literature review and meta-analysis to address technical issues in sustainable tourism. The feasibility of big data was demonstrated by investigating its sources, methods, purposes, and background. The proposed method could effectively promote the development of sustainable tourism industry. Femenia-Serra et al. [8] built a hybrid approach of semi-structured interviews and surveys to address privacy issues in smart tourism. In the study, a combination of qualitative and quantitative analysis was used to develop strategies to address tourism privacy issues. The results indicated that the method was of great significance to tourism organizations and policy makers. Orden-Mejia and Huertas [9] conducted research on emerging technologies in the smart tourism industry and proposed a chatbot attribute application method based on data analysis. Exploratory factor analysis and stratified regression were used to perform statistical analysis on the data during the process. The results indicated that emerging robot technology had good interactivity for tourists. Adamis and Pinarbasi [10] designed a network design method on the basis of visual communication for the dissemination of smart tourism destinations. During the process, functions such as subject analysis, object detection, and text mining were used to retrieve sample data. The proposed method had stronger analytical performance.

Some scholars have conducted related research on Apriori association rule mining algorithm. Faquetti et al. [11] used the Apriori algorithm to search for frequent itemsets in a database for identifying and quantifying combinations of combination drugs, in order to investigate harmful multi drug patterns and identify patient drug combinations. The proposed method could effectively analyze the efficacy of medication. Peng et al. [12] proposed an improved Apriori algorithm to understand the development trend of maritime education practitioners. Employment preferences was explored by identifying employability and status. The results indicated that the proposed method had good flexibility. Hassan et al. [13] used an Apriori algorithm to extract important rules for analyzing drug addiction. Data were collected to analyze the severity index of addiction. The results indicated that the Apriori algorithm could effectively analyze drug addiction status. Syafariani [14] designed an optimized Apriori association rule mining algorithm to explore the dependencies of various projects. Descriptive results were obtained through prior qualitative analysis of associations. The results indicated that the method had good stability. Dhinakaran and Joe Prathap [15] conducted frequent itemset mining on medical data and combined it with Apriori algorithm to implement fruit fly optimization algorithm, which provide solutions by reducing the low performance of big data models. The proposed method had good robustness.

In summary, although some scholars have conducted relevant research on smart tourism technology and Apriori association rule mining algorithm, there is still a lack of research combining the two. Therefore, the study attempts to use the Apriori association rule mining algorithm to optimize the smart tourism technology, mining the association relationships between frequent itemsets. It is

expected to provide technical support for the high-quality development of the tourism industry.

III. SMART TOURISM DATA MINING TECHNOLOGY COMBINED WITH APRIORI ASSOCIATION RULE MINING

A. Platform Construction based on Apriori Association Rule Algorithm

With the development of information technology, the tourism industry has gradually begun to use information processing technology to assist work. Smart tourism data mining technology can provide personalized recommendations and services to tourists through accurate data analysis, thereby enhancing their satisfaction and loyalty [16, 17]. Moreover, data mining can help tourism managers better understand tourist behavior and preferences, thereby optimizing the allocation and management efficiency of tourism resources [18, 19]. The Apriori algorithm can rely on powerful data mining capabilities to conduct accurate analysis of tourist data access in scenic spots. Data mining techniques for smart tourism are designed by combining Apriori association rule mining. To ensure the quality of data mining for smart tourism, the study adopts frequent itemsets based on Apriori algorithm. Then the Spark framework is taken as the distributed programming framework. The construction mode of Spark framework based on Hadoop is shown in Fig. 1.

In Fig. 1, the construction of the Sparta platform for includes three modes. The independent deployment mode reflects the characteristics of cluster management, specifically manifested in Spark occupying the top position of the Hadoop distributed file system and allocating reasonable space for the Hadoop distributed file system. In Spark on yarn mode, Spark runs on top of yarn. Spark on MapReduce is an open-source software in SIMR that allows users to run Spark directly on licensed Hadoop MapReduce v1. Meanwhile, Spark's efficiency and convenience in interactive queries make the distributed

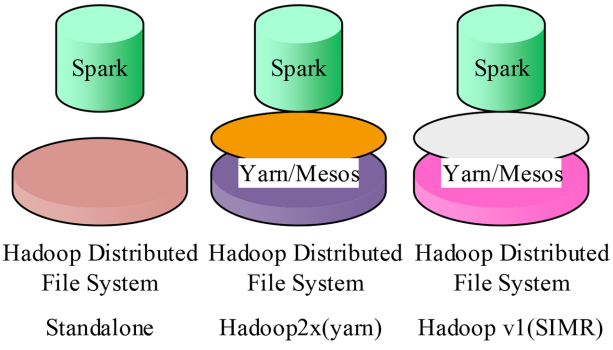

Fig. 1. Spark build mode based on Hadoop.

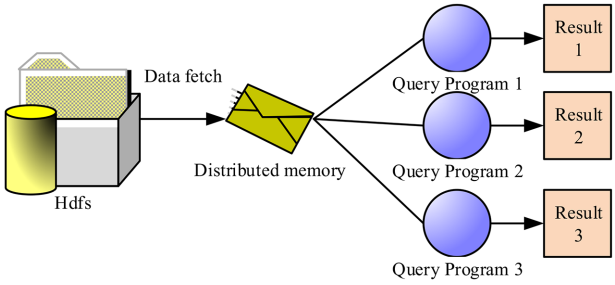

Fig. 2. Spark data interaction process.

computing process more convenient. The Spark data interaction process is shown in Fig. 2.

As shown in Fig. 2, the study adopts a memory sharing form and an elastic distributed dataset, which can transform the elastic distributed dataset at any operation execution time, and also save the elastic distributed dataset in distributed memory, making the next access more efficient. Taking Spark on yarn for platform construction can achieve rationalization in ensuring data security, cluster stability, and computational efficiency. When conducting data mining for smart tourism, a large number of correlation relationships need to be considered. When using the Apriori algorithm to analyze correlation relationships, the computational cost of discovering association rules and the credibility of association rules need to be taken into account [20]. The solution for calculating the total number of rules requires setting up a binary attribute transaction set to record the visit status of the scenic area. The constructed binary attribute transaction set is shown in Table 1.

As shown in Table 1, in the binary attribute transaction set, different scenic spots only record whether they have been visited, with 1 representing visited and 0 representing unvisited. Due to the different visit records of each scenic spot, the binary vector of the visited scenic spot becomes an asymmetric binary vector. When a scenic spot appears in two itemsets at the same time, the study uses support to measure the probability of both scenic spots appearing simultaneously. The support calculation is shown in Eq. (1):

Table 1. Example of binary attribute transaction set

Tid	Scenic area					
	X	Y	Z	A	B	C
1	1	0	1	0	1	1
2	1	1	0	0	0	1
3	0	1	1	1	0	0
4	1	1	1	0	1	1
5	1	1	1	1	0	0

$$S(X \to Y) = \frac{N(X,Y)}{N} \tag{1}$$

In Eq. (1), X represents X scenic area. Y represents Y scenic area. $S(\cdot)$ represents itemset support. $N(\cdot)$ represents the items that appear simultaneously. N represents the total items. The frequency of Y element appearing in the set where X is located can be represented by confidence, and the confidence operation is shown in Eq. (2):

$$c(X \to Y) = \frac{N(Y|X)}{N(X)} \tag{2}$$

In Eq. (2), $c(X \to Y)$ represents the confidence level of $X \to Y$. In practice, there is usually no absolute standard for the selection of support thresholds. Empirical settings and adjustments need to be made in combination with specific application scenarios, data scale and data distribution characteristics. The research is mainly based on historical data analysis, observing the number of rules generated and business relevance under different support degrees, and selecting a threshold that can balance the number of rules, significance and business interpretability. Finally, it was optimized through multiple experiments and business evaluations. According to the proposed equation, the total number of rules is shown in Eq. (3):

$$R = 3^d - 2^{d+1} + 1 \tag{3}$$

In Eq. (3), d represents the terms and R represents the total rules. Then the Apriori algorithm is used to generate frequent itemsets, introducing the laws of superset and subset. The current frequent itemset is used to generate the next potential frequent itemset, and then scan the dataset again to verify whether these candidate itemsets meet the minimum support requirement. This process is repeated until a new frequent itemset cannot be generated. If an itemset is frequent, then all its non-empty subsets must also be frequent. When filtering from a candidate set, the candidate itemset support is counted. Due to the large number of candidate frequent sets, the study applies a mixed column structure counting method. Establishing the hierarchical association (Ha) coefficient to scan the dataset is an important task in data mining to evaluate the correlation between itemsets. Ha coefficient is an important indicator for measuring the strength of rules, reflecting the conditional probability of the occurrence of the latter term in the case where the former term appears. For each itemset, its Ha coefficient is calculated, which is the relative frequency of the occurrence of the posterior term when the itemset is used as the antecedent. In this study, the Ha coefficient, as a domain-specific indicator proposed for the smart tourism scenario, aims to quantify the temporal correlation intensity between scenic spots. It is based on the frequency with which a certain scenic spot appears as the "first pre-visited scenic spot" before another scenic spot during a tourist's itinerary. For the Ha coefficient, first set T_i and T_j as two different scenic spots. For the visiting sequence of a certain tourist, if T_j is visited and T_i is the first scenic spot visited before T_j, then record a temporal correlation from T_i to T_j. The Ha coefficient is denoted as $Ha(T_i \to T_j)$, and its definition formula is shown in Eq. (4):

$$Ha(T_i \to T_j) = \frac{count(T_i \to T_j)}{count(T_j)} \tag{4}$$

In Eq. (4), $count(T_i \to T_j)$ represents the number of times T_i is the first pre-visited attraction of T_j, and $count(T_j)$ represents the total number of times T_i is visited. This coefficient measures the probability that tourists have visited T_i as their first attraction before visiting T_j, and is used to capture the weak temporal dependencies in tourists' behaviors, playing a key role in filtering out invalid candidate sets. The higher the Ha coefficient value, the stronger the tourists' preference for the visiting sequence from T_i to T_j. The mining rule is shown in Eq. (5):

$$\frac{S(X-X',\ Y+X')}{S(X-X')} = \frac{S(X,\ Y)}{S(X)+k} < \frac{S(X,\ Y)}{S(X)} \tag{5}$$

In Eq. (5), X' represents a subset of X and a represents the minimum confidence level. The support degree of X' must be greater than or equal to Y. k represents additional correction count.

B. Improved Data Mining Technology Combining Apriori Association Rule Algorithm

When the smart tourism data mining technology runs on a platform combined with Apriori association rule algorithm, due to the special nature of the tourism industry, it involves multiple data types and requires high resource consumption. Therefore, the operating structure is optimized to improve the data mining efficiency. The designed association extraction rule for frequent itemsets is shown in Fig. 3.

From Fig. 3, the association extraction of frequent itemset is divided into two steps. Firstly, a term that satisfies the confidence rule is set. Then, a new candidate set is generated and the reliability is judged. Finally, the rules that satisfy the confidence rule are integrated and filtered out. The Apriori algorithm requires multiple scans of data during runtime, resulting in excessive load on the database and operating system, thereby reducing overall mining efficiency. When mining data that meets the conditions of association rules but is not suitable for application scenarios, it can cause interference to the database and result in incorrect results. To make data

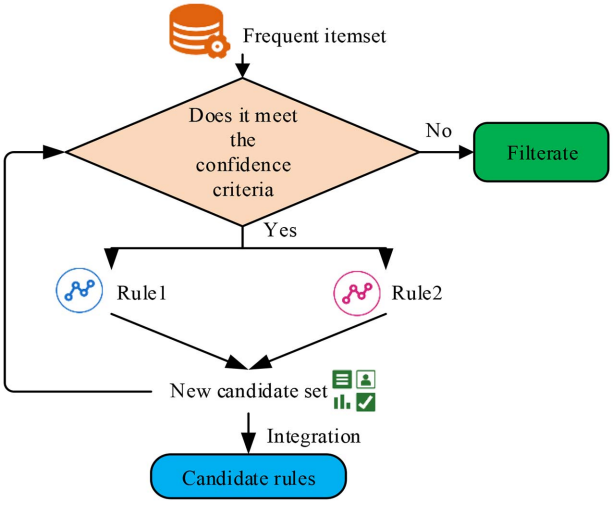

Fig. 3. Association extraction rule graph for frequent itemset.

mining more accurate, the study considers the order in which tourists visit scenic spots to reflect the priority of visits and achieve the goal of predicting visits. The statistical algorithm for the collection of tourists first visit scenic spots is shown in Eq. (6):

$$V_{lac} = S_{lac},\ V_{ci} = S_{ci} \tag{6}$$

In Eq. (6), V represents tourists, S represents the scenic area, *lac* represents the position code, and *ci* represents sector. The collected data after statistical analysis provides a data foundation for subsequent mining calculations. When introducing the first visit scenic spot based on actual scenarios, if any item in the item set is not the first visit scenic spot, it will not be calculated. The calculation conditions for the itemset are shown in Eq. (7):

$$\begin{cases} k_{item} = \{X_1, X_2 \ldots X_k\} \\ first\_v\_\sec = \{S_1, S_2 \ldots S_n\} \end{cases}, if\ X_i\left(1 \le i \le k\right) \notin first\_v\_\sec, k_{item} \ne k+1 \tag{7}$$

In Eq. (7), k_{item} represents the frequent itemset and $first\_v\_\sec = \{S_1, S_2, ..., S_n\}$ represents the first visit to the scenic area. These do not conform to the first visit scenic spots is shown in Eq. (8):

$$T\left(X_1\right) = T\left(X_2\right) = \ldots T\left(X_k\right) \tag{8}$$

In Eq. (8), $T(\cdot)$ represents the label attribute of the scenic spot, and data mining can continue after meeting the attribute conditions. To improve the strong association rules of the Apriori algorithm, the concept of relevance is introduced in conjunction with first visit scenic spots to optimize the Apriori algorithm. The judgment of the first visit scenic spot is shown in Eq. (9):

$$k_{item} \exists X_i \left(1 \le i \le k\right) \in first\_v\_sev \vee T\left(X_1\right) = T\left(X_2\right) = \ldots = T\left(X_k\right) \tag{9}$$

In Eq. (9), *first_v_sev* represents the prerequisite for the first visit to the scenic area. When generating frequent itemsets, they must be included in the list of first visited scenic spots and meet the existence conditions of first visited scenic spots. The association rule satisfies Eq. (10):

$$k_{item} \exists A_i \left(1 \le i \le n\right) \in first\_v\_sev \vee T\left(A_1\right) = \ldots = T\left(A_n\right) = T\left(B_1\right) \ldots = T\left(B_n\right) \tag{10}$$

In Eq. (10), A_i represents any element of the itemset. If the visits to two scenic spots are correlated events, the probability correlation is as shown in Eq. (11):

$$P\left(A \to B\right) = \frac{P\left(AB\right)}{P\left(A\right)P\left(B\right)} \tag{11}$$

In Eq. (11), $P(\cdot)$ represents the correlation probability. When $P(\mathrm{A} \to B) > 1$ is present, itemsets A and B are positively correlated, and the larger the correlation, the stronger the correlation. The correlation judgment is shown in Eq. (12):

$$\frac{P\left(AB\right)}{P\left(A\right)P\left(B\right)} > 1 \wedge \left(A_i \in first\_v\_\sec \vee T\left(A_1\right) = \ldots = T\left(A_n\right) = \ldots = T\left(B_1\right) = \ldots = T\left(B_n\right)\right) \tag{12}$$

When the data satisfies the condition of Eq. (12), it is judged as strongly correlated. Otherwise, it will be filtered. The temporal constraint intensity of the research method is a weak constraint. It only needs to mark the first visited scenic spot, avoiding the high cost of full-sequence pattern mining. The overall computational complexity is only O(k·n), and it can maintain good operational quality in discontinuous recommendations that rely on the first scenic spot. The timing constraint intensity of standard sequential patterns or ordered association mining techniques like PrefixSpan is a strict sequence, and the computational complexity can reach $O(n^2)$-O(n!) (see Appendix). And it is only applicable to continuous behavior analysis. The research method, by contrast, offers better operational efficiency and fewer operable scenarios with constraints. However, the optimized Apriori algorithm may still have low computational efficiency. Therefore, the parallel computing is introduced to improve the algorithm's computational efficiency. The schematic diagram of parallel computing and serial computing operation is shown in Fig. 4.

As shown in Fig. 4, there is a clear difference between parallel computing and traditional serial computing.

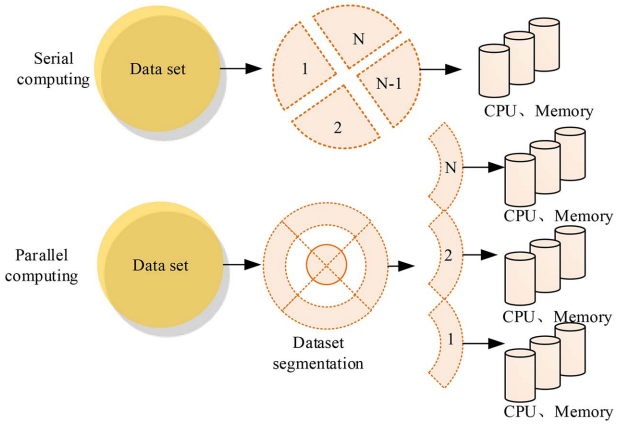

Fig. 4. Parallel operation diagram.

Compared with the overall nature of serial computing, parallel computing divides data into different clusters for computation, and finally merges the results. Parallel computing greatly improves data processing speed and reduces memory burden by splitting data. The re-optimization of Apriori algorithm is based on a parallelization computing framework, combined with a distributed programming architecture, to solve the long computation time when generating frequent itemsets. The re-optimization flowchart of Apriori algorithm is shown in Fig. 5.

As shown in Fig. 5, the designed optimized Apriori has two stages. The first stage is to input the raw dataset into distributed memory to obtain the in memory dataset. The memory dataset is quickly scanned and the support for each itemset is calculated. Support refers to the frequency of itemsets appearing in all transactions, and only itemsets above a preset threshold are considered frequent. The second stage is to generate k+1 itemsets through frequent k-itemsets, load and transform them, count transaction elements, and broadcast the candidate itemsets to various work nodes in the system. When using the research method for smart tourism data mining, key data such as passenger behavior data and scenic spot visit records should be collected first. After cleaning and preprocessing the data, binary attributes are obtained and support and confidence thresholds are defined to evaluate the effectiveness of association rules. Distributed memory systems are used to process datasets, generate association rules, and evaluate their strength by calculating the confidence level of the rules. Then the mined association rules are applied to the smart tourism system to achieve specific tasks such as personalized recommendation and resource optimization allocation, enriching the tourism experience and improving the efficiency of resource utilization. When the research method is applied in practice, the system analyzes the high-frequency combination rules of scenic spots (such as "from scenic spot A to scenic spot B") mined by historical tourists' visit records, and can recommend scenic spot B in real time when tourists plan their trips or arrive at scenic spot A. The excavated associations of travel preferences can be used to personalize the push of relevant service information or coupons. This can not only significantly enhance the convenience and satisfaction of tourists, but also help scenic area managers optimize the planning of scenic spot tour routes, precisely allocate service resources and conduct precise marketing activity planning, thereby creating significant economic and social value.

IV. EFFECTIVENESS ANALYSIS OF DATA MINING TECHNOLOGY FOR SMART TOURISM COMBINED WITH APRIORI

A. Performance Testing of Smart Tourism Data Mining Technology Combined with Apriori

To test the performance of the smart tourism data mining technology combined with Apriori in practical operation, the Ctrip scenic spot dataset and Meituan travel dataset are used as the testing datasets. After screening, the final

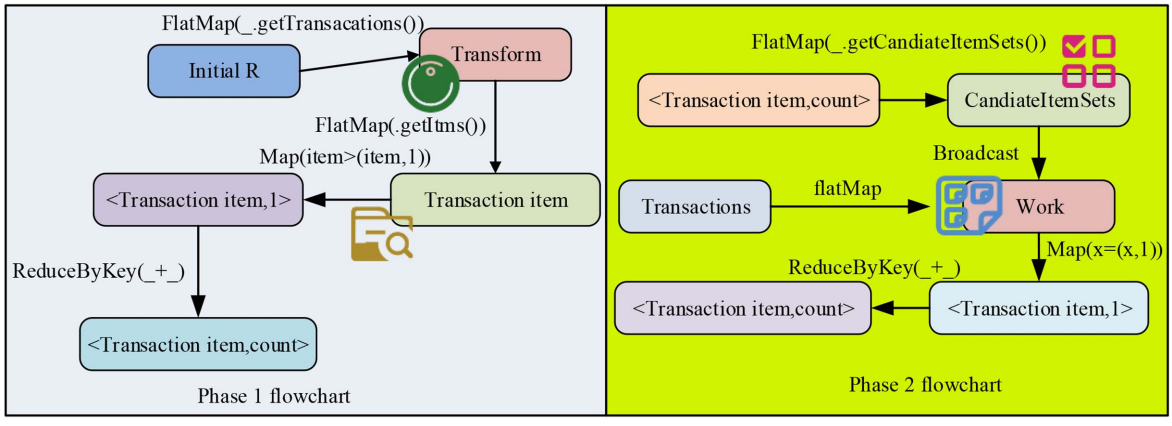

Fig. 5. Process diagram for re-optimization of Apriori algorithm.

Table 2. Dataset parameter information

Dataset	Transactions (A)	Attractions (B)	Avg. B/A	Sparsity	Preprocessing steps
Ctrip	86,742	217	4.3	0.980	1. Remove records with stay time <5 min
					2. Merge entries of same attraction
Meituan	112,569	184	3.8	0.979	1. Filter bot accounts
					2. Standardize attraction aliases
Alpha (Weekdays)	42,815	158	3.1	0.981	1. Aggregate by date
					2. Remove outliers (>10 attractions/day)
Bravo (Holidays)	73,926	172	5.2	0.970	1. Aggregate by date
					2. Impute missing location data

data for testing is obtained. To further determine the effectiveness of the research method during operation, a large natural tourist attraction is selected for practical application. When conducting application analysis, due to the significant differences in tourist travel characteristics between weekdays and holidays, the tourist data for weekdays is merged into one dataset called Alpha, and the tourist data for holidays is merged into one dataset called Bravo. The parameter information of the data set is shown in Table 2.

The preprocessing steps listed in Table 2 were designed to ensure data quality and consistency for mining meaningful association rules. For instance, removing records with very short stay times (<5 minutes) helps filter out accidental or invalid visits, while aggregating data by date and imputing missing locations enhance the completeness and temporal coherence of tourist behavior sequences. The parameters are presented in Table 3.

The experiment was carried out on a distributed cluster configured as Spark 3.3.1. The cluster consists of one driver node and four worker nodes, with each worker node running one Executor instance. Each Executor is allocated 4 CPU cores and 8 GB of memory (including in-heap execution memory and management overhead). The driver node allocates 4 GB of memory. Spark application default parallelism (spark.default.parallelism) and Shuffle operation on the number of partitions (spark.sql.shuffle.partitions) are set to 200, to match the cluster computing power efficiency and optimize the Shuffle. Radio variable transmission enabled compression (spark.sql.broadcast.compress = true) and set the timeout is 1,200 seconds (spark.sql.broadcastTimeout) to enhance the exchange performance of candidates. The original data set is hashed and partitioned based on the transaction ID during loading. The initial number of partitions is set to three times the total number of CPU cores in the cluster (16 cores), that is, 48 partitions, to promote computing load balancing. All experiments in YARN cluster resource management model (master YARN), and enable the dynamic resource allocation (spark.dynamicAllocation.enabled = true) to adapt to the demand of different calculation stages. When conducting experiments, the research method is referred to as Smart Apriori. It is compared with the frequent pattern growth (FPG) method and the equivalent class clustering and bottom-up lattice traversal (ECLAT) method. The generated association rules during the execution are tested, as shown in Fig. 6.

Table 3. Hardware and software configuration

Component	Description
Power supply power	900 W
Processor	Intel Core i9-12900K
Hard disk	1 TB NVMe SSD + 2 TB HDD
Operating system	Windows 11 Pro
Internal memory	32 GB DDR4 RAM
Graphics card	NVIDIA GeForce RTX 4080

As shown in Fig. 6, the quantity generated by different methods decreased with increasing confidence and support. Fig. 6(a) shows that the association rules generated by ECLAT at a confidence level of 0.004 were 31,213. The association rules generated at a confidence level of 0.056 decreased to 8,869. The association rules generated by FPG decreased to 5,978 when the confidence level was 0.056. The association rules generated by Smart Apriori at a confidence level of 0.004 was 19,451. The association rules generated at a confidence level of 0.056 decreased to 4,271. As shown in Fig. 6(b), the association rules generated by ECLAT decreased to 17,912 when the support level was 0.064. The association rules generated by FPG at a support level of 0.012 was 29,841. The association rules generated at a support level of 0.064 decreased to 10,776. The association rules generated by Smart Apriori at a support level of 0.012 was 14,847. The association rules generated at a support level of 0.064 decreased to 4,633. The association rules operated by the research method are more concise. The number of invalid frequent itemsets generated during runtime is tested, as

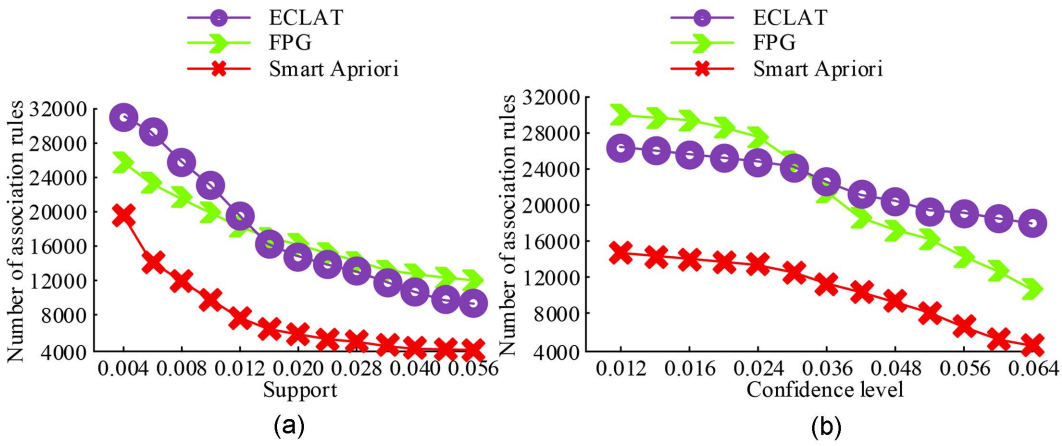

Fig. 6. Number of association rules generated under different levels of support (a) and different levels of confidence (b).

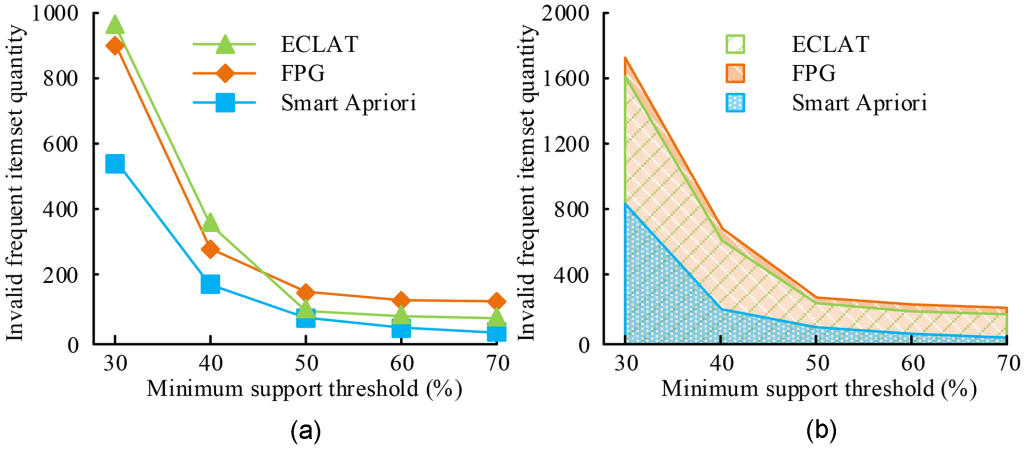

Fig. 7. Invalid frequent itemset quantity test: (a) Ctrip scenic spot dataset and (b) Meituan travel dataset.

shown in Fig. 7.

As shown in Fig. 7, the invalid frequent itemsets generated by different methods decreased as the minimum support threshold increased. In Fig. 7(a), in the Ctrip scenic spot dataset, ECLAT generated 964 invalid frequent itemsets when the minimum support threshold was 30%. When it increased to 70%, the invalid frequent itemsets decreased to 82. The FPG method generated 898 invalid frequent itemsets at 30%. The invalid frequent itemsets decreased to 123 at 70%. Smart Apriori generated 543 invalid frequent itemsets when the minimum support threshold was 30%. When the minimum support threshold increased to 70%, the invalid frequent itemsets decreased to 25. As shown in Fig. 7(b), in the Accidents dataset, ECLAT generated 1,617 invalid frequent itemsets when the minimum support threshold was 30%. When the minimum support threshold increased to 70%, the invalid frequent itemsets decreased to 192. The FPG method generated 1,745 invalid frequent itemsets at 30%. The invalid frequent itemsets decreased to 206 at 70%. Smart Apriori generated 831 invalid frequent itemsets when the minimum support threshold was 30%. The invalid frequent itemsets decreased to 40 at 70%. Under the 30% threshold of the Ctrip dataset, Smart Apriori generated an average of 545±15 invalid itemsets, which was significantly less than that of FPG (900±20, $p<0.001$) and ECLAT (965±22, $p<0.001$). The research method can effectively reduce the invalid frequent itemsets in the data mining process.

B. Practical Application Analysis of Data Mining Technology for Smart Tourism based on Apriori

The data mining time under different support thresholds is compared, as shown in Fig. 8.

In Fig. 8, the data mining time of different methods decreased with the increase of the minimum support threshold. In Alpha dataset, the data mining time of FPG was 613 ms when the minimum support threshold was increased to 70%. The data mining time of ECLAT when

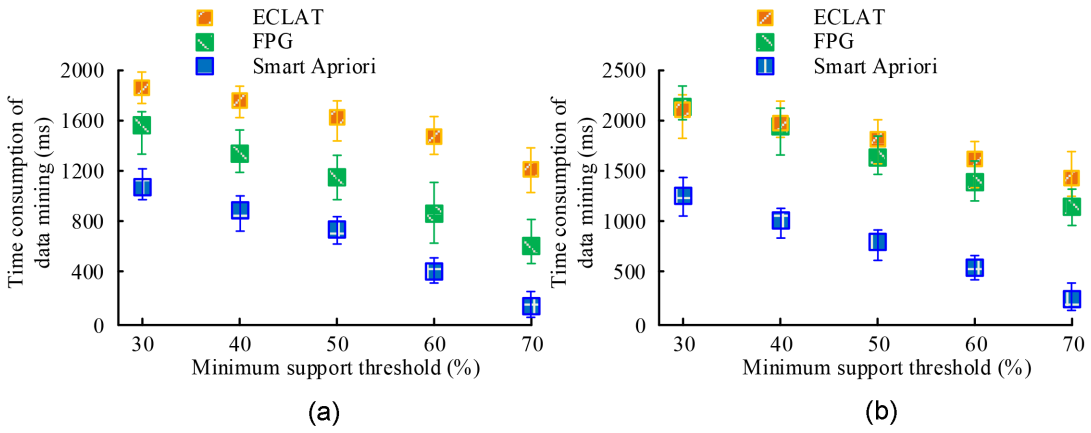

Fig. 8. Time consumption analysis of data mining: (a) Alpha dataset and (b) Bravo dataset.

the minimum support threshold was increased to 70% was 1,217 ms. The data mining time of Smart Apriori at a minimum support threshold of 30% was 689 ms. The data mining time when the minimum support threshold was increased to 70% was 174 ms. As shown in Fig. 8(b), in Bravo dataset, the data mining time of FPG when the minimum support threshold was increased to 70% was 1,191 ms. The ECLAT was 1,443 ms. The Smart Apriori at a minimum support threshold of 30% was 1,256 ms, and the data mining time was 247 ms at minimum support threshold of 70%. The average time of Smart Apriori at the 70% minimum support threshold in the Alpha dataset was 175±5 ms (mean±standard deviation), which was significantly lower than that of FPG (623±28 ms, $p<0.001$) and ECLAT (1,220±40 ms, $p<0.001$). This indicates that the research method has a faster actual running speed. The low latency is the result of the combined effect of algorithm optimization to reduce the amount of ineffective computations and the Spark parallel framework to accelerate effective computations. Especially when the minimum support threshold is relatively low, the effect of algorithm pruning is more significant. The processor utilization of the research method during actual operation is analyzed, as shown in Fig. 9.

As shown in Fig. 9, the processor utilization of different methods fluctuated within a certain range during runtime. In Fig. 9(a), the processor utilization of ECLAT fluctuated between 48% and 82% when generating frequent itemsets. The processor utilization of FPG fluctuated between 26% and 68%. The processor utilization of Smart Apriori fluctuated between 31% and 40%. As shown in Fig. 9(b), the processor utilization of ECLAT fluctuated between 38% and 72% when generating association rules. The processor utilization of FPG fluctuated between 61% and 76%. The processor utilization of Smart Apriori fluctuated between 41% and 47%. The fluctuation range of CPU usage rate during the operation of the research method is smaller and the overall level is lower. On the one hand,

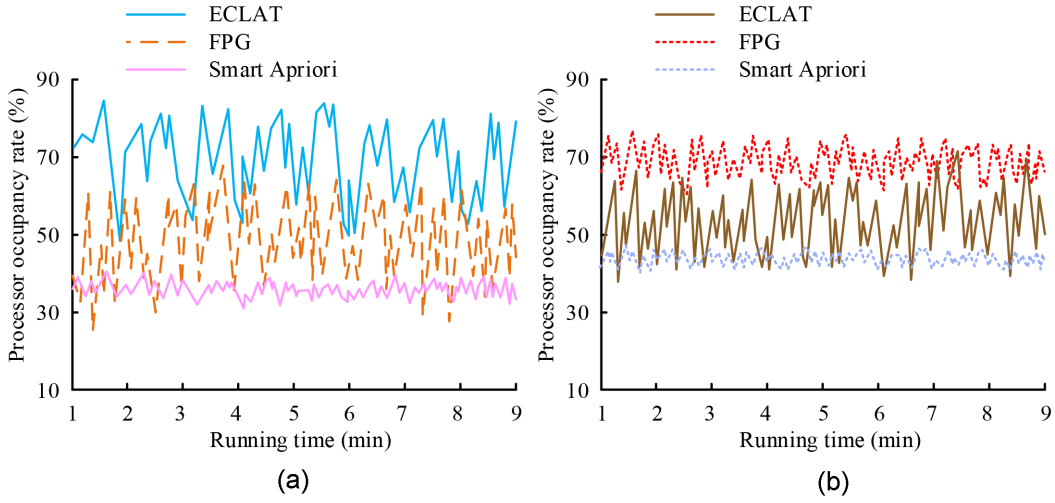

Fig. 9. Processor occupancy rate: (a) generate frequent itemsets and (b) generate association rules.

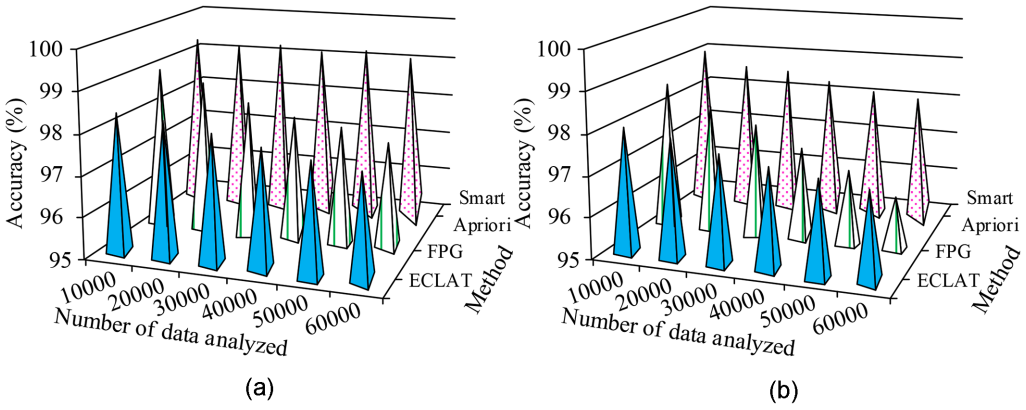

Fig. 10. Accuracy analysis of operation: (a) Alpha dataset and (b) Bravo dataset.

this is attributed to the algorithm optimization that reduces the amount of redundant computing tasks. On the other hand, it is due to Spark's efficient resource scheduling and memory management, which alleviates the pressure of single-point computing. The accuracy of the research method during runtime is tested, as shown in Fig. 10.

As shown in Fig. 10, the data accuracy of different methods during runtime was affected by the amount of analyzed data. In Alpha dataset, the accuracy of ECLAT reached 98.1% when the analyzed data volume was 10,000, but decreased to 97.4% when the analyzed data volume increased to 60,000. The accuracy of FPG reached 98.7% when analyzing data of 10,000, but decreased to 97.1% when analyzing data of 60,000. The accuracy of Smart Apriori was not significantly affected by the amount of analyzed data, and there was no obvious downward trend, remaining within the range of 99.0%–99.2%. As shown in Fig. 10(b), in Bravo dataset, the accuracy of ECLAT reached 97.8% when the analysis data volume was 10,000, but decreased to 96.9% when the analysis data volume increased to 60,000. The accuracy of FPG reached 98.2% when analyzing data of 10000, but decreased to 95.9% when analyzing data of 60,000. The accuracy of Smart Apriori also began to decrease with the increase of analyzed data volume, reaching 98.8% when the analyzed data volume was 10,000, and decreasing to 97.9% when the analyzed data volume increased to 60,000. When analyzing 10,000 data points in the Alpha dataset, the average accuracy rate of Smart Apriori was 99.1%±0.1%, which was significantly higher than that of FPG (98.7%±0.1%, $p<0.001$) and ECLAT (98.1%±0.2%, $p<0.001$). This indicates that the research method has better data accuracy in practical processing of smart tourism data mining tasks. All t-tests ($p<0.001$) and χ^2 tests ($p<0.001$) confirmed that the performance differences were statistically significant. In order to further verify the superiority of the research method, the study compared the research method in the Alpha dataset with the advanced "Seq2Seq+Attention" and "Improved Random Forest" in recent years [21, 22]. The ECLAT and FPG algorithms were implemented using the MLxtend library (v0.28.0) in Python, with minimum support values searched over {0.01, 0.05, 0.1, 0.15, 0.2} and confidence thresholds set to {0.3, 0.5, 0.7, 0.8, 0.9} for rule generation. The Seq2Seq+Attention model was implemented in PyTorch (v1.12.0) using a two-layer LSTM with Bahdanau attention, hidden and embedding dimensions of 128 and 64, respectively, a learning rate of 0.001, a batch size of 64, and early stopping for training. The Improved Random Forest was implemented using scikit-

Table 4. Comparison of advanced methods

Metric	Smart Apriori	Seq2Seq+Attention	Improved Random Forest
Precision@5	0.723	0.781	0.692
Recall@5	0.658	0.621	0.589
F1-score@5	0.688	0.692	0.636
NDCG@5	0.742	0.813	0.705
Coverage (%)	86.2	72.5	78.3
User satisfaction	4.2	3.8	3.5
Latency (ms)	175	320	210

learn (v1.2.0), with key hyperparameters including the number of trees (100–300) and maximum depth (10–20) determined via grid search. All experiments were independently run five times, with results reported as mean values±standard deviation; a fixed random seed (42) was used to ensure reproducibility. The advanced method examples are shown in Table 4.

In Table 4, all the statistical data have practical statistical significance. The measurement methods for user satisfaction and coverage reported in Table 4 are as follows. User satisfaction was evaluated through a user study, with a total of 30 participants (n=30) using the recommendation system interface over a period of 7 days. Each participant filled out a standardized questionnaire after completing the system interaction tasks and rated their satisfaction with the recommendation results on a 5-point Likert scale. Among them, 1 indicates very dissatisfied and 5 indicates very satisfied. The final satisfaction score is the average of all participants' and task ratings. Coverage is defined as the proportion of unique items that appear in at least one recommendation rule in the test set, and the calculation formula is shown in Eq. (13):

$$P_{\text{coverage}} = \frac{\left| \bigcup_{r \in R_{rec}} \left(antecedent(r) \cup consequent(r) \right) \right|}{|\Gamma_{test}|} \quad (13)$$

In Eq. (13), P_{coverage} represents the coverage rate, R_{rec} represents the set of recommended rules, Γ_{test} represents the set of all items in the test dataset, and *antecedent*(*r*) and *consequent*(*r*), respectively represent the sets of all items contained in the antecedent and consequent parts that constitute rule *r*. Seq2Seq+Attention was optimal in terms of precision (Precision@5=0.781) and ranking quality (NDCG@5=0.813), but had a lower recall rate (0.621), indicating accurate recommendations for mainstream attractions but ignoring niche ones. The Improved Random Forest has a processing time of more than 200 ms due to the complex feature engineering (requiring the construction of tourist portraits + scenic spot attributes). The research method leads in recall rate (0.658) and coverage rate (86.2%), and is suitable for diverse recommendation scenarios. Meanwhile, the User Satisfaction of the research method reached 4.2, which was significantly higher than that of other methods. It is proved that the research method indeed has stronger recommendation quality and performance.

V. CONCLUSION

A data mining technique combining association rule mining algorithm and parallel computing was developed to support the high-quality operation of smart tourism technology. During the process, yarn was used as the central component for resource management, and a Sparta platform with three modes was designed. An example of a binary attribute transaction set for scenic spots was constructed, and the Apriori was applied to generate frequent itemsets. The searched itemsets were iteratively collected multiple times, and the concept of relevance was introduced in conjunction with the first visit to the scenic spot. Parallel computing was introduced to optimize computational efficiency, and the data were divided and calculated in different clusters. Finally, the effectiveness of the research method was analyzed. When conducting analyzing the number of association rules, the research method generated 14,847 association rules with a support degree of 0. 012. When conducting data mining time analysis, the research method kept the data mining time below 247 ms when the minimum support threshold was increased to 70%. When the research method was running, the processor utilization rate for generating frequent itemsets fluctuated between 31% and 40%, and the processor utilization rate for generating association rules fluctuate between 41% and 47%. The research method for data mining not only has a simpler operating process, but also requires lower equipment requirements, making it smoother to run in real-world environments. However, the research has not yet considered the interference caused by short-term distortion of tourism data due to social hot trends. In the future, more special scenarios will be added to test and optimize the research method to expand its applicability.

CONFLICT OF INTEREST

The author has declared that no competing interests exist.

ACKNOWLEDGMENTS

The research is supported by the Chongqing Higher Education Society Project, Exploration and Practice of the Path of Smart Cultural and Tourism Learning Workshop in Vocational Colleges under the Background of Industry-Education Integration (No. cqgj23257C); Chongqing Municipal Educational Science Planning Project 2024, Empowerment in Constructing Higher Vocational Tourism and Culture Majors Clusters (No. K24ZG3210098).

APPENDIX

The time complexity of the two-stage method under weak time constraints is described as follows: n represents the number of transactions in the dataset, and k represents the average number of "first visit" item candidates involved in each transaction after the initial pruning.

Assuming that the item set size is bounded after support filtering and the time window is fixed, this method avoids exhaustiveness for all possible sequence patterns. The entire process is divided into two stages: candidate generation and pairing matching. In the candidate generation phase, frequent items with timestamps are extracted for each transaction. Due to the size limitation of the item set, the processing time for each item is O(1). In the pairing and matching phase, the items in each transaction are matched with the "first visit" candidate items in the previous transactions within the time window. On average, each transaction only involves O(k) candidates. Since each transaction is traversed only once and the average complexity of each matching operation is O(k), the total time complexity is O(k·n). In contrast, strict sequence pattern mining methods require searching the combination space of ordered itemsets.

REFERENCES

1. P. Sedarati, F. Serra, and T. Jere Jakulin, "Systems approach to model smart tourism ecosystems," *International Journal for Quality Research*, vol. 16, no. 1, pp. 285-306, 2022. https://doi.org/10.24874/IJQR16.01-20
2. D. Buhalis, A. Papathanassis, and M. Vafeidou, "Smart cruising: smart technology applications and their diffusion in cruise tourism," *Journal of Hospitality and Tourism Technology*, vol. 13, no. 4, pp. 626-649, 2022. https://doi.org/10.1108/JHTT-05-2021-0155
3. I. Sustacha, J. E. Banos-Pino, and E. del Valle, "Research trends in technology in the context of smart destinations: a bibliometric analysis and network visualization," *Cuadernos de Gestión*, vol. 22, no. 1, pp. 161-173, 2022. https://doi.org/10.5295/cdg.211501is
4. N. Bano and S. Siddiqui, "Consumers' intention towards the use of smart technologies in tourism and hospitality (T&H) industry: a deeper insight into the integration of TAM, TPB and trust," *Journal of Hospitality and Tourism Insights*, vol. 7, no. 3, pp. 1412-1434, 2024. https://doi.org/10.1108/JHTI-06-2022-0267
5. G. Bandewad, K. P. Datta, B. W. Gawali, and S. N. Pawar, "Review on discrimination of hazardous gases by smart sensing technology," *Artificial Intelligence and Applications*, vol. 1, no. 2, pp. 86-97, 2023. https://doi.org/10.47852/bonviewAIA3202434
6. F. J. Ballina, "Smart business: the element of delay in the future of smart tourism," *Journal of Tourism Futures*, vol. 8, no. 1, pp. 37-54, 2022. https://doi.org/10.1108/JTF-02-2020-0018
7. E. Rahmadian, D. Feitosa, and A. Zwitter, "A systematic literature review on the use of big data for sustainable tourism," *Current Issues in Tourism*, vol. 25, no. 11, pp. 1711-1730, 2022. https://doi.org/10.1080/13683500.2021.1974358
8. F. Femenia-Serra, A. Ioannou, and I. P. Tussyadiah, "Is smart scary? A mixed-methods study on privacy in smart tourism," *Current Issues in Tourism*, vol. 25, no. 14, pp. 2212-2238, 2022. https://doi.org/10.1080/13683500.2021.1987399
9. M. Orden-Mejia and A. Huertas, "Analysis of the attributes of smart tourism technologies in destination chatbots that influence tourist satisfaction," *Current Issues in Tourism*, vol. 25, no. 17, pp. 2854-2869, 2022. https://doi.org/10.1080/13683500.2021.1997942
10. E. Adamis and F. Pinarbasi, "Unfolding visual characteristics of social media communication: reflections of smart tourism destinations," *Journal of Hospitality and Tourism Technology*, vol. 13, no. 1, pp. 34-61, 2022. https://doi.org/10.1108/JHTT-09-2020-0246
11. M. L. Faquetti, A. M. D. la Torre, T. Burkard, G. Obozinski, and A. M. Burden, "Identification of polypharmacy patterns in new-users of metformin using the Apriori algorithm: a novel framework for investigating concomitant drug utilization through association rule mining," *Pharmacoepidemiology and Drug Safety*, vol. 32, no. 3, pp. 366-381, 2023. https://doi.org/10.1002/pds.5583
12. F. Peng, Y. Sun, Z. Chen, and J. Gao, "An improved apriori algorithm for association rule mining in employability analysis," *Tehnički vjesnik*, vol. 30, no. 5, pp. 1435-1442, 2023. https://doi.org/10.17559/TV-20230327000481
13. M. M. Hassan, S. Zaman, S. Mollick, M. M. Hassan, M. Raihan, C. Kaushal, and R. Bhardwaj, "An efficient Apriori algorithm for frequent pattern in human intoxication data," *Innovations in Systems and Software Engineering*, vol. 19, no. 1, pp. 61-69, 2023. https://doi.org/10.1007/s11334-022-00523-w
14. R. F. Syafariani, "Association analysis with Apriori algorithm for electronic sales decision support system," *International Journal of Informatics, Information System and Computer Engineering (INJIISCOM)*, vol. 3, no. 1, pp. 61-72, 2022. https://doi.org/10.34010/injiiscom.v3i1.8089
15. D. Dhinakaran and P. M. Joe Prathap, "Protection of data privacy from vulnerability using two-fish technique with Apriori algorithm in data mining," *Journal of Supercomputing*, vol. 78, pp. 17559-17593, 2022. https://doi.org/10.1007/s11227-022-04517-0
16. Q. Hao, W. J. Choi, and J. Meng, "A data mining-based analysis of cognitive intervention for college students' sports health using Apriori algorithm," *Soft Computing*, vol. 27, no. 21, pp. 16353-16371, 2023. https://doi.org/10.1007/s00500-023-09163-z
17. D. T. Tran and J. H. Huh, "Forecast of seasonal consumption behavior of consumers and privacy-preserving data mining with new S-Apriori algorithm," *The Journal of Supercomputing*, vol. 79, no. 11, pp. 12691-12736, 2023. https://doi.org/10.1007/s11227-023-05105-6
18. S. Chen, Y. Xue, and X. Cui, "Information literacy of college students from library education in smart classrooms: based on big data exploring data mining patterns using Apriori algorithm," *Soft Computing*, vol. 28, no. 4, pp. 3571-3589, 2024. https://doi.org/10.1007/s00500-023-09621-8
19. Y. Song and Y. He, "Toward an intelligent tourism recommendation system based on artificial intelligence and IoT using Apriori algorithm," *Soft Computing*, vol. 27, no. 24, pp. 19159-19177, 2023. https://doi.org/10.1007/s00500-023-09330-2
20. O. Dogan, F. C. Kem, and B. Oztaysi, "Fuzzy association rule mining approach to identify e-commerce product association considering sales amount," *Complex & Intelligent Systems*, vol. 8, no. 2, pp. 1551-1560, 2022. https://doi.org/10.1007/s40747-021-00607-3
21. C. Zhang, H. Zhang, T. Pu, and J. Pan, "Supply chain demand forecasting based on data mining algorithm and Seq2Seq," *International Journal of Control, Automation and Systems*, vol. 23, no. 1, pp. 89-104, 2025.

https://doi.org/10.1007/s12555-024-0141-8
22. H. Yang, S. Zhang, J. Zhang, and C. Wang, "Evaluating the performance of athletes in various sports using data mining and big data analytics," *Soft Computing*, vol. 28, no. 4, pp. 2875-2890, 2024. https://doi.org/10.1007/s00500-023-09620-9

Fen Ma

Fen Ma received her Ph.D. in Tourism Management from Sichuan University, China, in 2018. She joined Chongqing College of Finance and Economics in September 2018. Her research interests focus on tourism big data and sustainable tourism.