Journal of Information & Systems Management

DLINE JOURNALS

ISSN: 2230 – 8776

JISM 2025: 15 (4)
https://doi.org/10.6025/jism/2025/15/4/161-168

Transport Route Recommendation Using LDA Topic Modeling and Apriori Association Rules

Jinbo Li
Physical Education Department
Henan Vocational College of Logistics
Xinxiang, Henan, 453500. China
lijinbo19790221@163.com

ABSTRACT

The paper explores the application of data mining techniques specifically the LDA (Latent Dirichlet Allocation) topic model and the Apriori association rule algorithm to enhance personalized tourism route recommendations. As tourism shifts from standardized group itineraries toward individualized experiences, the study addresses inefficiencies in current travel planning methods, which often over look user preferences. By analyzing user generated content such as reviews, browsing histories, and click behaviors, the LDA model uncovers latent thematic interests and sentiment trends within tourism related text data. Meanwhile, the Apriori algorithm identifies frequent associations among tourist attractions and services, enabling the construction of optimized, preference aligned itineraries. The proposed recommendation system features a three tier architecture (application, logic, and data processing layers) that integrates real time user data to refine suggestions dynamically. The research demonstrates that combining LDA for topic and sentiment analysis with Apriori for association mining improves the accuracy, relevance, and personalization of travel recommendations. This approach not only enhances user satisfaction but also boosts competitiveness for tourism enterprises by aligning offerings with actual traveler needs. The study concludes that these data-driven methods effectively address information overload and preference ambiguity in modern tourism, marking a significant step toward intelligent, personalized travel planning in the big data era.

Keywords: Personalized Recommendation, Tourism Route Planning, LDA Topic Model, Apriori Algorithm, Data Mining, Sentiment Analysis, Association Rules, User Preferences

Received: 2 July 2025, Revised: 20 September 2025, Accepted: 1 October 2025

1. Introduction

As living standards have improved, individuals' expectations for life have transitioned from focusing on material desires to seeking spiritual fulfillment. An increasing number of people are taking an interest in leisure and entertainment activities beyond their everyday routines. [1] The tourism sector has emerged as a key area for enhancing quality of life. With rising competition within the tourism industry, businesses are becoming more focused on their competitive edge and tourism performance. [2] Delivering more suitable travel itineraries to customers, decreasing users' confusion regarding travel routes, and improving personalized services and customer awareness have become vital elements. Surveys indicate that many tourism companies often implement several fundamental strategies when designing travel routes: one involves creating business travel itineraries in collaboration with companies to offer users organized visits to attractions. Customers are unable to choose attractions based on personal preferences freely but must follow the itineraries provided by the tourism firms and businesses. Another strategy is the development of themed travel routes, catering to groups with specific age demographics, allowing tourism companies to craft tours based on themes such as red tourism. Additionally, travel routes tailored for various age groups, including teenagers and children, are available to clients. The final basic strategy employs data mining and statistical analysis to identify the most effective travel paths. [3] By leveraging market research and the characteristics of attractions, along side the actual start up capital provided by users, the most engaging tourist sites are linked to offer customers the most economical travel routes.

These suggested travel itineraries have not fully leveraged modern information technology. With advancements in data information and the establishment of databases, numerous information technologies, including data mining, have been extensively applied across various sectors. Analyzing and processing data information has also offered diverse support to different businesses. In conjunction with the growth of the tourism industry, data mining offers numerous opportunities for travel route planning. [4] It has the potential to mitigate the existing challenges in the tourism market and address the issue of ineffective user information that hinders industrial advancement. Relevant tourism companies can utilize the insights gained from tourism data mining to develop more rational and appealing travel plans for users, thereby enhancing their competitiveness within the industry. Online tourism and the personalized recommendations for travelers have become significant sources of data for tracking shifts in the tourism market. On modern information platforms, various WeChat public accounts, travel recommendation articles, hotel reviews, and scenic spot evaluations can all influence tourists' decisions. [5] Hence, data mining and association analysis have become primary tools for addressing tourism challenges. Through precise data analysis, fluctuations in the popularity of tourist attractions and various initiatives, such as creating guides for scenic areas, can be assessed to improve travelers' experiences during their trips.

2. Current Status of Tourism Industry Development under Data Mining Algorithms

With the flourishing growth of the Internet and social media, the tourism sector is encountering numerous new demands while continuously refining its services. Offering suitable support for the creation of travel itineraries based on accurate visitor information and traveler preferences has become a significant area of interest for researchers in the tourism field. As society progresses, digital media platforms and online services provide the public with an abundance of information sources. A majority of young individuals prefer utilizing We Chat public accounts and other cultural tourism platforms, along with hotel reviews, in depth scenic details, and travel guides to curate the best options for their travel itinerary planning. However, when faced with scattered information, users struggle to derive effective assistance, resulting in relatively low efficiency in gauging the popularity of various attractions and in obtaining real time data updates. After many social platforms and tourism professionals raised these concerns, they conducted comprehensive studies and proposed several solutions. For instance, employing Bayesian classification algorithms, support vector machin- es, and logistic regression, along with integrating multiple neural networks to create machine learning models aimed at resolving data collection and topic classification challenges. Text analysis and dictionary classification techniques are utilized for sentiment prediction, while weight calculations are employed to mitigate the potential influence of sentiment analysis on users. This approach aids the tourism industry in aligning with the genuine needs of users during the allocation of scenic area plans and thematic development to achieve personalized suggestions.

Under the influence of data mining algorithms, the pace of development for tourism industry models varies. The primary factors impacting this are largely derived from fluctuations in the popularity of tourist attractions on online platforms and shifts in the public's tourism content demands. The majority of travelers tend to utilize online searches to discover network information that assists them in planning their journeys. Nevertheless, the vast amount of online data often leads to information overload, hindering tourists from obtaining precise assistance during their information searches and attraction selections. Numerous international scholars have explored new research avenues regarding the implementation of personalized recommendation systems within the tourism sector. They have discovered that effectively employing data mining and association rule algorithms can extract relevant decision making information from tourism data. This predictive method of relevance not only aids tourists in discovering travel routes that align with their needs and preferences but also minimizes costs for tourism enterprises and enhances their profitability. Moreover, the application of various topic models can facilitate personalized travel itinerary design and boost the overall fault tolerance of the tourism industry. Tourism theme recommendation models possess distinct characteristics that set them apart from other recommendation initiatives: Firstly, the fluctuations in user tourism data are intricate, making it difficult for travelers to articulate their needs regarding tourist attractions accurately. Secondly, the system model encounters challenges in acquiring user preference data for tourism. Lastly, data on tourists' travel is relatively limited, which increases the complexity of practical predictive itinerary design. This work outlines the advancements in the digitalization of the tourism sector both domestically and internationally. It suggests employing LDA topic models and Apriori algorithms to identify and track tourism users, ultimately providing information to enhance tourism recommendation models by analysing historical keywords.

3. Personalized Tourism Data Mining and Route Recommendation based on LDA Topic Model and Apriori Algorithm

3.1 Tourism Data Mining based on LDA Topic Model and Apriori Algorithm

In our research, we extracted keywords related to tourism expressions and conducted sentiment and popularity analyses in Chinese. By combining user click counts and browsing frequencies, we generated a tourism product association model. The LDA topic model can display information about topic content in text as a probability distribution. It is an unsupervised learning algorithm that discovers relevant content about hidden tourism topics. It forms the core topic structure of documents by identifying common features among words and sentences. At the same time, it is a classical phrase based model that treats text as a frequency vector of phrases. There is no ranking of words, and the distance of each topic word to the document follows a multinomial distribution. Assuming there are multiple texts related to tourism information, each containing a different number of topic contents, the distribution of topic words across texts follows a multinomial distribution. By setting different parameter letters, the distribution of parameter letters in the word corpus is similar to the distribution structure of topics in the text. That is, after introducing multi dimensional vectors, simultaneous transformations can be performed. Using data mining sampling, all the words related to tourism topic data are obtained, and by calculating the number of topic words, the effective data amount in the text is statistically obtained, and the distribution form of topic samples is simplified as follows:

$$D(\vec{p},\vec{a}) = \frac{1}{\square a} \times \prod_{k=1}^{k} p_k^{a_{k-1}} \tag{1}$$

Calculate the distribution condition of vocabulary samples:

$$p(\vec{z},\vec{a}) = \frac{1}{\square a} \times \int \prod_{k=1}^{k} p_k^{a_{k-1+n(k)}} d\theta_d = \frac{\square(\vec{n_d} \mid \vec{a})}{\square(\vec{a})} \tag{2}$$

In the formulas, represents the parameter sample. Represents the multinomial distribution count corresponding to the topic words in different text documents. Combining with the actual user tourism data set, the LDA topic model is used to determine the optimal distribution of tourism routes. Then, based on the vector features transformed by the topic distribution, the similarity of user tourism topics is calculated to perform reasonable classification. Data mining technology can extract valuable information from large scale, noisy, a-nd fuzzy data, even in the presence of interference. Apriori association rule calculation is also a major method for handling hidden relationships between various features in large data sets. We present the algorithm flow of Apriori association rule calculation as in figure 1.

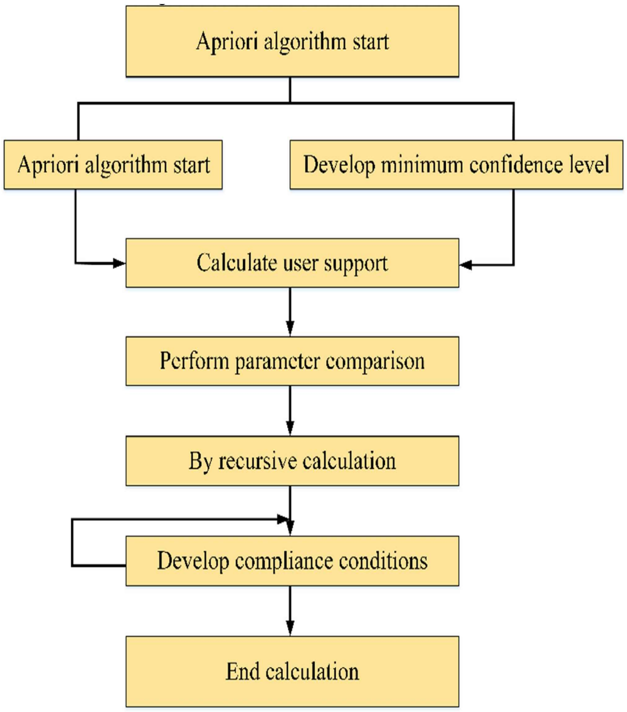

Figure 1. Apriori Association Rule Calculation Process

From Figure 1, it can be observed that at the initial stage of the algorithm, the support for candidate tourism routes is calculated using minimum support and confidence. By comparing support and confidence, recursively generated conditions that meet the requirements are obtained. Assuming certain correlations between multiple attributes, each item's data can be judged based on other data samples. Factors affecting the correla tion rules and correlation coefficients of the topic content include the frequency of data item occurrences, support, confidence, effectiveness, minimum support, and minimum confidence. The frequency with which data items appear in the calculations indicates the number of valid data records, which supports the count based confidence calculation. From browsing records and search frequencies of tourism users, frequent sets of tourism information that appear iteratively are obtained layer by layer. This type of information data serves as effective content to judge users' preferences for tourism routes and occupies a certain proportion in the tourism data recommendation model.

3.2 Personalized Recommendation of Data Mining Tourism Routes based on LDA Topic Model and Apriori Algorithm

Traditional tourism recommendation models focus only on travel routes, entertainment activities, shopping experiences, food choices, and hotel accommodations. These recommendations primarily focus on cooperation among tourism companies, related merchants, and tourist attractions, without considering tourists' individual preferences. To accurately recommend diverse and personalized tourism products for tourists, optimization needs to be carried out in the following four aspects: user needs, tourist preferences, constraints, and information on tourist attractions. User needs refer to tourists' subjective intentions in judging whether the current tourism route and products meet their requirements. This involves subjective conditions, such as users' specific demands for hotel accommodations and sightseeing. Using the LDA topic model to analyse target requirements can increase the likelihood that the tourism plan will be selected. Tourist preferences can be calculated by analyzing historical data traces of users, such as evaluations and ratings of the tourism industry. Constraints are objective conditions that influence tourists' decisions in the tourism industry, such as their inome and age. Finally, information on tourist attractions is valuable tourism data and one of the main elements in generating recommendations for the recommendation system.

Currently, most tourism routes and plans are designed based on group tourism patterns. In such patterns, handling the varying internal demands of different tourists is essential. When conducting tourism data mining, it is important to determine whether tourism information texts belong to the same category. Categorization based on the LDA model can facilitate relevant personnel in tourism companies to obtain detailed descriptions of specified travel information and arrange attractions that meet the demands of most people based on topic over lapping. To select popular tourist attractions, a popularity analysis model is developed using tourism attraction review data. Using the Apriori association analysis algorithm, similar attractions are judged and identified, and the characteristics and differentiating information of highly similar attractions are statistically compared. Finally, the frequency of appearances of tourist attractions in the information network is weighted, while the relationships between this attraction and other joint attractions are described to provide reasonable route plans for group tourism patterns.

4. Personalized Recommendation of Tourism Data Mining Based on LDA Topic Model and Apriori Algorithm

With the development of socio economics and internet technology, the tourism industry has witnessed an abundance of various travel routes and recommendation plans. In the era of big data, offline tourism enterprises have gradually evolved into a hybrid of online and offline business models. Compared to traditional offline tourism, online tourism platforms have richer personalised features and advantages, such as broad services and high communication efficiency. However, faced with massive amounts of tourism information and basic data on tourist attractions, the main challenge lies in helping tourists independently select travel routes that suit their needs. To address this issue, this study conducts tourism data mining using the LDA topic model and Apriori algorithm to generate a corresponding tourism industry recommendation system, providing more personalized tourism services to tourists.

Firstly, the LDA topic model is used to perform sentiment analysis of users' historical traces, automatically determining tourists' attitudes or emotional inclinations towards specific tourism topics in the text data. The sentiment information extraction, classification, and summarization are utilized to calculate the degree of tourists' preference for travel plans. Simultaneously, the topic model is trained on the dataset, annotating relevant features across the vast comment dataset. The Apriori association analysis algorithm is applied to compute the correlation coefficient between tourism attractions and plans, using weighted calculations based on the support and confidence of tourist products. The architecture of the personalized tourism recommendation system is as in figure 2.

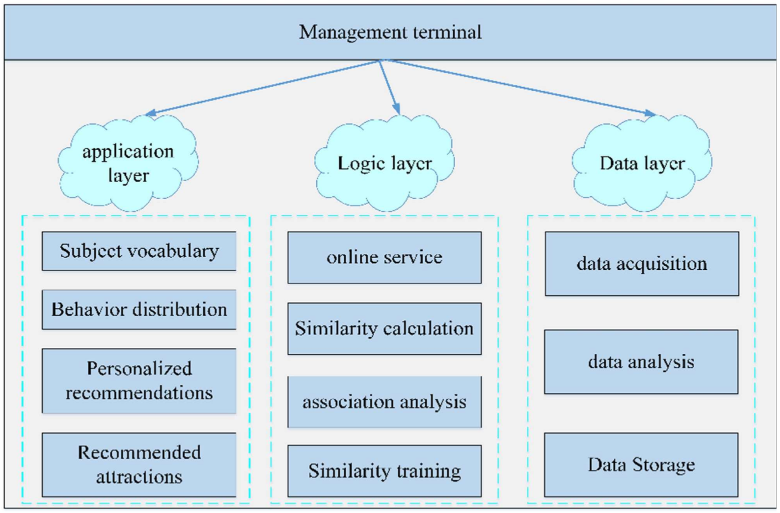

Figure 2. Architecture of Personalized Tourism Recommendation System

As illustrated in Figure 2, the tourism route recommendation system is structured into three tiers application, logic, and data processing following the traditional client server separation framework. Users access the tourism recommendation system platform to sift through relevant thematic terms linked to their preferred travel experiences and receive tailored route suggestions. Within the data processing tier, user browsing

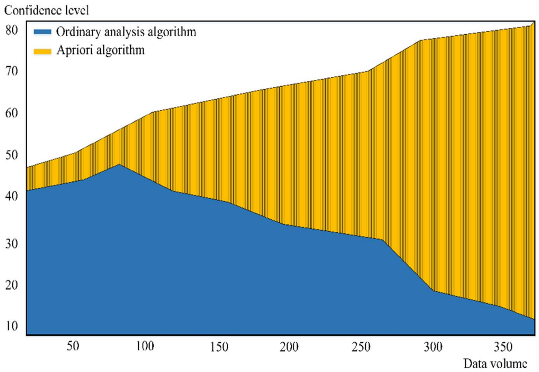

Figure 3. Confidence Comparison of Different Data Mining Algorithms

behaviors and click statistics are sent back to the system's central processing interface, allowing tourism indu- stry leaders to offer personalized scenic area suggestions based on user behavior data. We have selected the Apriori algorithm as the optimal approach for assessing association strength, comparing it with traditional data processing techniques in the realm of dynamic tourism data management.

As depicted in Figure 3, the Apriori algorithm yields a greater confidence level for tourism data insights and can deliver customized service recommendations to travelers. Given that internal elements such as weather conditions at tourist sites can be unpredictable, when analyzing related tourism data, it is also essential to dynamically account for the influences of temporal and spatial factors. Suppose it is established that the choice of a scenic area is more affected by external influences than by time. In that case, the data associated with this influence will be omitted from the calculations.

5. Conclusion

With the rise in socio economic conditions, individuals' daily living standards have consistently improved. The desire for both material and cultural fulfillment is increasingly apparent. As a sector rich in entertainment value within cultural pursuits, the tourism industry accounts for a significant share of people's daily expenditures. Offering personalized services and travel itinerary suggestions to users is a crucial focus in this field. Utilizing the LDA topic model and Apriori algorithm, this research performs data mining on tourism information to identify the primary factors affecting the public's choice of travel routes and plans. Initially, the LDA topic model is applied to examine tourists' browsing patterns and click data on online tourism platforms. After generating behavioral trace text data, relevant thematic terms connected to tourism are extracted to assess users' emotional tendencies and requirements. The Apriori algorithm is then utilized to evaluate the weights of user behavioral traits, identifying the needs of travelers to generate recommended tourism routes that fulfill their expectations. Findings indicate that the LDA topic model and Apriori algorithm demonstrate significant breadth and accuracy in tourism data mining, offering personalized recommendation strategies for the tourism sector.

References

[1] Jelodar, H., Wang, Y., Yuan, C., et al. (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 78, 15169-15211.

[2] Chauhan, U., Shah, A. (2021). Topic modeling using latent Dirichlet allocation: A survey. *ACM Computing Surveys (CSUR),* 54 (7), 1-35.

[3] Hajjem, M., Latiri, C. (2017). Combining IR and LDA topic modeling for filtering microblogs. *Procedia Computer Science*, 112, 761-770.

[4] Mohammed, S. H., Al-augby, S. (2020). Lsa lda topic modeling classification: Comparison study on e-books. *Indonesian Journal of Electrical Engineering and Computer Science,* 19 (1), 353-362.

[5] Hagen, L. (2018). Content analysis of e-petitions with topic modeling: How to train and evaluate LDA models? *Information processing management, 54(6), 1292-1307.*

[6] Wang, X., Yang, X., Wang, X., et al. (2020). Evaluating the competitiveness of enterprise's technology based on LDA topic model. *Technology Analysis Strategic Management,* 32 (2), 208-222.

[7] Hegland, M. (2007). The apriori algorithm a tutorial. Mathematics and computation in imaging science andinformation processing 209-262.

[8] Singh, J., Ram, H., Sodhi, D. J. (2013). Improving efficiency of apriori algorithm using transaction reduction. *International Journal of Scientific and Research Publications,* 3 (1) 1-4.

[9] Aflori, C., Craus, M. (2007). Grid implementation of the Apriori algorithm. *Advances in engineering software,* 38(5). 295-300.

[10] Hanguang, L., Yu, N. (2012). Intrusion detection technology research based on apriori algorithm. *Physics Procedia,* 24, 1615-1620.