A method classifying the domestic tourist destination base similarity measuring

Nguyen Thi Hoi, Tran Thi Nhung, Bui Quang Truong, Nguyen Quang Trung
Faculty of Economic Information System and E-commerce, Thuongmai University, Hanoi, Vietnam

Article Info

Article history:

Received Oct 6, 2024
Revised May 10, 2025
Accepted Jun 8, 2025

Keywords:

Domestic tourist
Historical search data
Machine learning
Recommendation system
Similarity measures

ABSTRACT

The classification problem is crucial in business, providing an effective method for supporting search activities in areas such as e-commerce, education, and marketing. This has become especially important in the wake of the COVID-19 pandemic, which has increased the need to promote and stimulate domestic tourism. This research focuses on recommending tourist destinations based on historical search data related to domestic tourism. The study uses techniques like term frequency-inverse document frequency (TF-IDF) weight vector analysis and similarity measures to calculate recommendation scores. Data was collected from various tourism websites, covering destinations across all 63 provinces and cities in Vietnam. Experiments were conducted using three approaches: cosine similarity, the brute force algorithm, and long short-term memory (LSTM) for long-text processing. The results indicate that similarity-based methods produce recommendations that closely match user preferences. For full-sentence queries, the brute force algorithm delivers more accurate results, while LSTM provides faster processing times. These findings offer businesses multiple strategies for improving recommender systems in practical applications.

Corresponding Author:

Nguyen Thi Hoi
Faculty of Economic Information System and E-commerce, Thuongmai University
No. 79, Ho Tung Mau Street, Cau Giay District, Hanoi, Vietnam
Email: hoint@tmu.edu.vn

1. INTRODUCTION

With the rapid development of information and communications technology (ICT) and the widespread availability of the internet, users are increasingly relying on digital platforms, such as search engines and online entertainment services, to access a wide range of information and utilities. This digital transformation has significantly reshaped consumer behavior, facilitating easier access to travel information, the ability to compare destinations, and the convenience of booking travel services with just a few clicks [1]–[4].

Tourism remains a vital engine of global economic growth. According to the World Travel & Tourism Council (WTTC) [5], the travel and tourism industry contributed 9.1% to global gross domestic product (GDP) in 2023. International tourism expenditures increased by 33.1%, reaching USD 1.63 trillion, while domestic expenditures rose by 18%, surpassing USD 5 trillion [5], [6]. Leisure-related spending saw a growth of 21.2%, and business travel spending rose by 22.4%, reflecting a strong post-pandemic recovery [5], [6]. In terms of direct economic impact, the travel and tourism sector added USD 3,059.3 billion to global GDP (2.8%), primarily driven by industries such as hospitality, travel agencies, and passenger transportation services [6], [7]. From 2024 to 2034, the sector's contribution is projected to grow at an annual rate of 3.7%, reaching USD 4,865.7 billion (3.5% of GDP) [5]–[7].

In Vietnam, the tourism industry has demonstrated impressive post-pandemic recovery, especially in the domestic sector. According to the General Statistics Office of Vietnam [8], international tourist arrivals reached 12.6 million in 2023, approximately 3.5 times higher than in 2022, recovering to 70% of pre-pandemic levels [8], [9]. This surpassed Asia's average recovery rate of 65%. Domestic tourism was even more robust, recording 110 million domestic trips in 2023 [9]. Infrastructure improvements, enhanced service quality, and strategic investment in cultural preservation have contributed to this growth [8]–[11].

This growth has generated significant economic spillover effects across various industries, including hospitality, food and beverage services, transportation, and retail. In addition, cultural festivals and sightseeing activities contribute to the development of the creative and entertainment industries while supporting local businesses and traditional craft villages [12]–[15]. Tourists now benefit from real-time information access, easy price comparisons, and instant booking features. However, the volume of available information has become overwhelming, often hindering travelers from efficiently identifying destinations that align with their preferences [1], [2], [16]–[20].

To address this, tourism recommendation systems (TRSs) have become increasingly important. These systems apply machine learning techniques to analyze user behavior, historical data, and preferences to offer personalized travel suggestions. By utilizing both static and dynamic (real-time reviews, social media, and ratings) information, TRSs can deliver context-aware recommendations that improve user satisfaction [1]–[3], [13], [16]–[19], [21], [22]. Such systems are already in use across various industries including e-commerce (Amazon), streaming (Netflix and YouTube), recruitment, and tourism [21]–[24]. Despite the economic potential of tourism, participation among local communities in tourism-related activities remains limited [2], [12]–[14], [25]. Most Vietnamese travel websites provide basic lists of attractions, but lack personalized recommendations tailored to individual behaviors or previous interactions. This leads to decision fatigue and inefficient travel planning [10], [13]–[15], [25]–[28].

To overcome these limitations, this study proposes the development of a personalized TRS tailored to the domestic travel market in Vietnam. By leveraging users' browsing behavior and search patterns, the system will suggest lesser-known destinations that align with individual preferences. This not only enhances the user experience but also promotes equitable tourism development across regions. The proposed system integrates advanced computational techniques, including cosine similarity, brute force matching, and long short-term memory (LSTM) models [1], [18], [19], [26], [27], [29]–[31]. These techniques are implemented on a prototype website to assess recommendation accuracy and user relevance.

By integrating these algorithms, the system aims to offer more accurate, real-time, and interest-based travel suggestions, while contributing to the visibility and economic uplift of underexplored destinations across Vietnam. The article includes the following sections: problem statement, related works, research model, research results, and some implications for localities with many tourist destinations in Vietnam.

2. RELATED WORK

Recommender systems for domestic travel destinations have emerged as a vital tool in the digital transformation of the tourism industry, offering personalized travel suggestions based on individual user preferences. These systems aim to generate tailored lists of tourist sites that align with user expectations, thereby improving decision-making and user satisfaction. As travelers increasingly demand convenient and customized planning solutions, particularly in the post-pandemic context, tourism recommender systems have become widely adopted not only in Vietnam but globally [1]–[4], [12], [13], [15], [17], [20], [21], [29]–[33].

Current research in travel recommendation systems is typically structured around two main methodological approaches. The first approach involves clustering tourist destinations based on shared attributes such as natural landscapes (beach and mountains), cultural and historical value, or urban experiences [3], [4], [12], [13], [15], [18], [19], [26], [29]–[36]. These categorizations are achieved using unsupervised learning techniques, allowing the system to classify destinations into meaningful groups that reflect user preferences. For example, [1], [2], [20], [30], [34] demonstrated the application of K-means or classification-based association (CBA)-fuzzy algorithms for destination clustering, yielding promising results in automated recommendations. Similarly, [23], [24], [35] applied support vector machines (SVM) and Bayesian algorithms to analyze large-scale datasets derived from user behavior and survey feedback, enabling improved segmentation of user travel intentions.

Significantly, [2] and [23] introduced an innovative deep learning framework based on EfficientNet-lite, optimized for mobile deployment, achieving over 80% user satisfaction with more than 50% of recommendations meeting expectations. This model underscores the potential of deep learning in enhancing recommendation systems, particularly in real-time and user-centered mobile applications. The second approach focuses on supervised and semi-supervised learning methods for generating recommendations [1]–[4], [16], [20]–[24], [26], [29]–[33], [35], [36]. These models are trained on extensive labeled datasets to optimize the relevance and accuracy of destination suggestions. For instance, [35] evaluated six machine learning algorithms naïve Bayes (NB), SVM, logistic regression, neural networks, decision trees, and random forests on tourism data, with SVM, logistic regression, and neural networks demonstrating superior performance in predicting user preferences. Furthermore, [29] compared Doc2Vec embeddings with traditional techniques such as item-k-nearest neighbors (KNN) and neural collaborative filtering, finding that Doc2Vec significantly outperformed conventional models, particularly in handling sparse datasets.

Building on these foundations, our study proposes a personalized travel recommender system designed specifically for the Vietnamese domestic tourism context. This system analyzes user input, such as search queries and browsing patterns, and compares them with structured destination profiles to generate accurate and relevant recommendations. To facilitate this, we utilize N-gram analysis and a Wikipedia-based dictionary to extract semantic features from user input and destination metadata. To further enhance recommendation accuracy, we apply cosine similarity and Pearson correlation to measure alignment between user interests and destination characteristics.

Expanding the dataset is another key objective to improve system robustness and accuracy. By including more diverse user profiles and regional destination data, the system's adaptability can be enhanced. Beyond tourism, the proposed architecture holds potential for cross-domain applications in education, entertainment, and healthcare, where personalized information delivery is equally valuable.

3. PROPOSED RESEARCH MODEL

3.1. Proposed model

The problem of suggesting domestic tourist destinations can be briefly described as based on the user's request, the system provides suggestions as a list of tourist destinations that are considered closest to the user's requests. From this problem, the research proposed a model as follows:

- Input: question or query with a list of keywords that describe the user's wishes.
- Output: a list of tourist destinations close to the user's wishes.
- Perform: i) step 1: analyze the question or query; ii) step 2: compare based on similarity between the question and database by weight or based on training; iii) step 3: perform evaluation and sorting based on similarity results; and iv) step 4: provide a list of locations.

3.2. Similarity measures and correlation

To operationalize the proposed research model, a comprehensive, multi-stage methodology was adopted. This approach integrates descriptive data analysis with advanced text normalization techniques, tailored for both structured and unstructured data within tourism systems. The primary objective was to construct a robust recommendation architecture capable of accurately interpreting user-generated queries and effectively matching them to relevant destination metadata.

The initial phase involved rigorous preprocessing of textual data collected from various sources, including destination descriptions, user reviews, and behavioral logs. The preprocessing steps encompassed tokenization, lowercasing, stop-word removal, and lemmatization, aiming to normalize linguistic features across datasets. Subsequently, the cleaned textual data was transformed into numerical form using SVM, primarily through term frequency-inverse document frequency (TF-IDF) weighting schemes. TF-IDF has been shown to be effective in reducing noise and emphasizing semantically significant terms in sparse textual datasets [18], [19], [27], [29], [37]–[40]. Calculate weight is calculated based on keyword TF, the IDF, and the TF×IDF is calculated as (1).

$$w_{ij} = \begin{cases} TF * IDF = (1 + log(f_{ij})) * log\left(\frac{N}{1+df_i}\right) if\ tf_{ij} \geq 1 \\ 0\ if\ tf_{ij} = 0 \end{cases} \tag{1}$$

Where w_{ij} is the weight of w_i in text d_j, f_{ij} is the frequency of appearance of w_{ij} in text d_j.

This transformation enables the application of various computational techniques such as cosine similarity, SVM, and K-means clustering, which operate effectively in high-dimensional feature spaces [17]–[19], [26], [27], [30]–[36], [40]. To identify semantic alignment between user queries and destination profiles, the cosine similarity metric was employed. This measure evaluates the cosine of the angle between two non-zero vectors in a multi-dimensional space, making it highly effective in text mining and recommendation systems.

Such cross-domain adaptability is made possible by the system's reliance on generalized deep learning and information retrieval techniques, rather than domain-specific rules or hardcoded logic.

$$cosine\ (\vec{A}, \vec{B}) = \frac{\vec{A} \times \vec{B}}{|A| \times |B|} \qquad (2)$$

In which: $\overrightarrow{A} \times \vec{B}$ is the dot product of two vectors; |A|, |B| is the magnitude of the two vectors.

Cosine similarity is particularly advantageous for high-dimensional TF-IDF vectors, where traditional distance metrics like Euclidean distance are less reliable due to the curse of dimensionality [17], [27], [31], [35], [40]. A brute force algorithm was also implemented to serve as a baseline method. This approach exhaustively compares each user profile vector with all available destination vectors. While computationally intensive, brute force ensures maximum matching precision, making it a useful benchmark for evaluating more scalable algorithms [2], [17], [25], [36].

In smaller datasets or controlled experimental scenarios, brute force methods are highly interpretable and often outperform more complex models in terms of raw accuracy. To capture the temporal and contextual dynamics of user behavior, an LSTM network, a variant of recurrent neural networks (RNNs), was integrated into the recommendation engine. LSTM networks are particularly well-suited for sequence modeling due to their ability to maintain long-term dependencies via gated mechanisms [1], [2], [16]–[19], [39]. The LSTM model used in this study comprises three key gates: i) forget gate: filters irrelevant information from the previous cell state; ii) input gate: integrates new input into the current state; and iii) output gate: determines which information contributes to the output. This architecture enables the system to distinguish between short-term and long-term user preferences (e.g., temporary interest in beach resorts vs. consistent preference for historical landmarks) [13], [21], [23], [25], [26], [29], [30], [37]–[39].

To overcome the sparsity and high dimensionality problems inherent in TF-IDF matrices, enhancements to the cosine similarity measure were adopted. These included adaptive term weighting, soft cosine similarity, and local similarity thresholds, which together increased both recall and precision in sparse datasets [1], [13], [19], [21], [23]–[26], [35]. Such enhancements allowed the system to better handle semantically similar but lexically different queries, especially in multilingual or domain-specific contexts. The integration of vector-based similarity, brute force baselines, and sequence-aware LSTM models results in a hybrid methodology that is not only robust but also scalable across domains. Although initially implemented for Vietnamese domestic tourism, the underlying architecture is transferable to adjacent sectors such as education, entertainment, and healthcare [4], [13], [15], [25], [26], [28].

4. EXPERIMENT RESULTS AND DISCUSSION

4.1. Data collection

In this study, a comprehensive dataset of tourist destinations across all 63 provinces and centrally governed cities in Vietnam was constructed to support the training, evaluation, and validation of the proposed TRS. The data acquisition process was conducted through a semi-automated pipeline leveraging publicly available information retrieved from multiple prominent travel service platforms, including Traveloka, Vinpearl, DiscoveryTravel, VietTravel, Vntrip.vn, and VietnamBooking. These platforms are widely used in tourism data analytics due to their rich, diverse, and regularly updated repositories of destination metadata [12]–[15], [25], [26], [28].

An initial corpus of 8,940 raw descriptive records was collected from these platforms, encompassing various types of tourist locations, metadata categories, and user-generated textual content. To ensure data quality and eliminate redundancies, a comprehensive preprocessing and cleaning phase was applied. This process included: i) removal of duplicate entries; ii) normalization of naming conventions across sources; and iii) filtering out entries with missing, ambiguous, or non-informative content.

Following this refinement, a curated dataset of 2,100 high-quality and unique destination profiles was obtained, deemed suitable for experimental modeling and machine learning-based training purposes. The finalized dataset captures critical attributes for recommendation modeling, such as: i) location name and province; ii) geographic region; iii) category (e.g., nature, heritage, entertainment); iv) descriptive summary; and v) user-generated tags and text reviews. These features were designed to support both text vectorization and semantic similarity computation in downstream recommendation algorithms. The statistical distribution and structural summary of the dataset are presented in Table 1 and serve as the basis for clustering, classification, and deep learning models applied in subsequent phases [1], [2], [13], [18]–[21], [23], [24], [27], [29]–[31], [33]–[35], [38], [39].

To further enhance the dataset's textual quality, a rigorous preprocessing pipeline was implemented. This phase aligned with state-of-the-art methodologies proposed in [1], [2], [13], [18]–[21], [23], [24], [27], [29]–[31], [33]–[35], [38], [39], recognizing that real-world travel descriptions often exhibit: i) lexical inconsistencies (e.g., spelling errors); ii) redundancy and repetition; and iii) non-standardized terminology and abbreviations. To mitigate these issues, the study adopted a Wikipedia-based lexical dictionary as a semantic reference source. This dictionary was employed to normalize key terms, identify alternative expressions, and enrich sparse content. The use of Wikipedia for semantic expansion is consistent with prior research that highlights its capacity to extract domain-agnostic, semantically rich representations from noisy text data [1], [27], [38], [39].

In parallel, a comprehensive Vietnamese stop-word list was compiled from both Wikipedia's linguistic tools and the Vietnamese Language Research Project, as discussed in [4], [13], [25]–[27], [40]. These resources were used to identify and eliminate non-informative terms (e.g., "ở", "rất", "một nơi") that commonly appear in user reviews but offer limited semantic contribution to destination profiling. Once cleaned, the textual data was subjected to N-gram analysis, enabling the extraction of co-occurrence patterns that reflect contextual and syntactic relationships. Experiments were conducted using N values ranging from 1 to 4, with bigrams (N=2) yielding the most coherent results for Vietnamese sentence structures, where compound concepts often span two or more syllables [27]. Trigrams (N=3) also provided promising performance, especially for complex location descriptors. Consequently, the final feature extraction focused on bigrams and trigrams, which were then transformed into TF-IDF vectors. This transformation assigns each term a weight that reflects its importance within a specific document relative to the entire corpus, ensuring that both common and distinctive terms are appropriately captured.

Table 1. Sample list of websites

Name	Volume	Sample
Traveloka *(https://www.traveloka.com/vi-vn)*	1,890	550
Vinpearl *(https://vinpearl.com/vi)*	1,760	500
Du lịch khám phá *(https://dulichkhampha.com.vn/)*	1,575	350
Du lịch Việt *(https://dulichviet.com.vn/)*	1,260	250
Vntrip.vn *(https://www.vntrip.vn/en)*	1,510	250
VietNam Booking *(https://www.vietnambooking.com/)*	945	200
Total	8,940	2,100

(Source: authors' statistics from resources in Vietnam)

4.2. Experimental data sample

To implement the algorithms and measurements, the study built a sample data set consisting of: no., name of destination, city, description, and multimedia. After processing the data, the researchers selected a set of 2,100 locations. They include beach tourism, mountain tourism, ecotourism, historical relics, churches, pagodas, and entertainment areas. Table 2 shows the structure of a sample of data about a tourist destination.

Table 2. A sample of data for the experiment

No.	Name	City	Description	Multimedia
1	Khu du lịch Tam Cốc-Bích Động	Ninh Bình	*Khu du lịch Tam Cốc–Bích Động được ví như "vịnh Hạ Long cạn" với nhiều cảnh đẹp như Tam Cốc, đền Thái Vi, chùa Bích Động, động Tiên, hang Bụt, thung Nắng, thung Nham, vườn chim... Nổi tiếng với danh xưng 'Nam thiên đệ nhị động', Tam Cốc Bích Động sở hữu cảnh sắc làng quê yên bình cùng hệ thống hang động núi đá vôi ấn tượng. Là một phần trong Quần thể danh thắng Tràng An, Tam Cốc Bích Động là điểm đến hoàn hảo dành cho những ai muốn khám phá trọn vẹn vẻ đẹp non sông.* *Địa chỉ: Ninh Hải, Hoa Lư, Ninh Bình (đồng bằng sông Hồng, miền Bắc Việt Nam)* *Phân loại du lịch: du lịch sinh thái, du lịch văn hóa, du lịch nghỉ dưỡng và du lịch khám phá*	https://static-images.vnncdn.net/files/publish/2023/7/8/ninh-binh-anh-so-du-lich-cc-1021-864.jpeg

(Source: from data sample set built by authors)

4.3. Experimental scenario

Research carried out by scenario as follows: i) step 1: build a set of user queries; ii) step 2: build a word set for each query, preprocessing, N-gram separation, stop-word removal; iii) step 3: calculate the weight vector by TF. IDF. Calculate the correlation of each query with the locations; iv) step 4: conduct manual matching from volunteers; v) step 5: conduct search using models and evaluate the search performance of the model.

The detailed steps of the scenario are as follows: to rigorously evaluate the performance of the proposed personalized TRS, the research team developed a structured dataset of 100 user queries, each simulating a real-world search intent in the tourism domain. These queries varied in word length, lexical complexity, and thematic focus, covering interests such as ecotourism, cultural heritage, coastal destinations, and urban exploration. Each query was stored in an Excel spreadsheet with two primary columns: i) serial number (query ID) and ii) query text. This structure ensured traceability and reproducibility throughout the experimental workflow. An additional column was later added to log the expected number of matching destinations, determined through human annotation.

The query evaluation process followed a multi-stage architecture, consisting of the following steps:

i) Stage 1: query design and ground truth annotation: each query was constructed to reflect natural language expressions of travel interest. A group of human evaluators was recruited to manually associate each query with a relevant set of destinations from the curated dataset of 2,100 tourist locations, previously collected across all 63 provinces and municipalities in Vietnam. These associations served as the gold standard reference set for evaluating algorithmic outputs [12], [25]–[27], [31], [33], [37].

ii) Stage 2: text preprocessing and normalization: each query was tokenized and normalized using standard natural language processing (NLP) techniques: tokenization and stop-word removal based on linguistic resources from Wikipedia and the Vietnamese Language Research Project [13], [26], [27], [39]. N-gram decomposition with values of N=1 to 4, to capture the frequent occurrence of compound and multi-syllabic expressions in Vietnamese [13], [15], [26], [27]. Lowercasing and morphological normalization to ensure uniform representation across datasets.

iii) Stage 3: query vectorization via TF-IDF: following preprocessing, both the user queries and the textual metadata of each tourist destination were transformed into high-dimensional vector representations using the TF-IDF model. This weighting scheme effectively quantifies term significance within the corpus, enabling accurate semantic comparison [4], [19], [28], [38].

iv) Stage 4: algorithmic evaluation and similarity computation: three algorithms were employed to assess semantic similarity between queries and destination descriptions: cosine similarity: computes the cosine of the angle between TF-IDF vectors, facilitating a scalable and interpretable matching mechanism [17], [27], [31], [35], [40]; brute force matching: performs exhaustive pairwise comparisons between query vectors and all destination vectors [23], [25], [30], [37]–[39].

Although computationally intensive, it guarantees maximum recall and serves as a baseline evaluation method [13], [21], [23]–[26], [35]; LSTM neural network: utilizes a gated RNN architecture to capture temporal and contextual dependencies in text sequences. This method is particularly well-suited for interpreting nuanced patterns in user intent [23]–[25], [30], [31], [33], [35], [37]. Experiment setup: all implementations were carried out in Python, using Visual Studio Code (v2020) as the integrated development environment. Experiments were executed on a Windows 11 machine powered by an Intel(R) Core i7-1185G7 @ 3.00 GHz CPU and 32 GB RAM, ensuring stable and reproducible computations. To establish an empirical performance benchmark, the research team conducted a manual evaluation process involving a group of trained volunteers. Each participant was instructed to select a set of relevant destinations from the master dataset for each of the 100 queries. These manually selected results were consolidated to form the ground-truth relevance set for each query.

For each algorithm, the system-generated recommendations were compared to the ground truth using a binary scoring mechanism: let N denote the number of correct locations (from human annotation), let M represent the number of locations retrieved by the algorithm; for each match between the retrieved and reference sets, the accuracy score was incremented by one unit. This per-query matching accuracy was then aggregated across all 100 queries to assess overall system effectiveness [13], [15], [27]. In addition to binary accuracy scoring, the system's predictive correctness was evaluated using the mean squared error (MSE) metric. MSE measures the average squared deviation between the number of expected and retrieved results per query, offering a quantitative indicator of over- or under-prediction [1], [2], [19], [26], [27], [31], [37].

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(p_i - r_i)^2 \tag{3}$$

With each $request_i, i = 1 \dots 100,$ $location_j, j = 1 \dots 2100,$ total as $(100 \times 2100) = 210000$ pairs, and n=210000. The MSE formula states that accuracy increases as the value of MSE gets closer to zero, and vice versa. Acquired experimental findings as in (4).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(p_i - r_i)^2 = \frac{1}{210000}\sum_{i=1}^{210000}(p_i - r_i)^2 = 0{,}1251 \tag{4}$$

Corresponding to the precision of the experiment with cosine is (5).

$$CR = (1 – MSE) \times 100\% = (1 - 0{,}1251) \times 100\% = 87{,}49\% \tag{5}$$

This evaluation framework ensured that both the precision of matching and volume of relevant results were systematically assessed across all models and queries.

4.4. Discussion and limitation

The following is the query search results table of the models: cosine similarity, brute force algorithm, and LSTM model in Step 3. Accuracy in modes is illustrated in Table 3. Following the completion of experimental trials, the study identified that the cosine similarity algorithm outperformed other models in terms of recommendation accuracy, achieving an accuracy rate of 87.49%, followed by the brute force matching method at 84.55%, and the LSTM model at 82.18%. As detailed in Table 3, cosine similarity not only yielded the highest average accuracy but also demonstrated the lowest standard deviation, indicating strong stability and consistency across various query types. Conversely, the brute force model exhibited the highest variability, suggesting sensitivity to input query structure and vector sparsity [23]–[25], [30], [31], [33], [35], [37]–[39].

Table 3. The results of the three modes

No.	Brute force (%)	Cosine similarity (%)	LSTM (%)
1	84.55	87.49	82.18

The observed superiority of cosine similarity is primarily attributed to its computational efficiency in measuring angular distance between TF-IDF vectors derived from user queries and destination profiles [1], [19], [20]. This approach facilitates the extraction of semantically aligned results without relying on historical user interaction data. The transformation of unstructured text into high-dimensional structured feature vectors enabled the model to match user intent with destination metadata in a purely content-based framework [2], [25]–[27], [29].

In operational terms, the most relevant destinations, those with the highest cosine similarity scores, were automatically ranked at the top, providing users with suggestions that closely mirrored their intent. The brute force matching model, though computationally expensive, was shown to be particularly effective in handling full-sentence queries where semantic nuance and context were essential. This model conducted exhaustive pairwise comparisons between every query vector and the destination corpus, thus maximizing coverage [17], [27], [30], [31].

Meanwhile, the LSTM model, despite being slightly less accurate in top-1 recommendation ranking, demonstrated notable strengths in processing speed and scalability, especially when deployed in environments requiring real-time responses or handling large query volumes [17], [19], [20], [23], [33], [35], [38]. To evaluate the practical applicability of the proposed models, a web-based TRS was developed using the FastAPI framework, deployed in a Python environment, and executed via Visual Studio Code. The system interface enables users to input free-text queries and receive an interactive output interface that presents: i) the number of matched results; ii) a ranked list of destination suggestions per algorithm; and iii) real-time visual comparison of outputs from cosine similarity, brute force, and LSTM methods. This feature allowed researchers and test users to qualitatively assess algorithm outputs and validate model behavior under different linguistic scenarios.

Several key insights derived from system experimentation are that the brute force matching achieved marginally higher accuracy for long-form queries, aligning well with user-selected destination sets in manual validation. LSTM, while slightly trailing in accuracy, offered faster runtime and better resource optimization, suggesting potential for real-time or mobile deployment. Cosine similarity, especially when fine-tuned with optimized TF-IDF parameters and bigram modeling, consistently delivered semantically rich and accurate results across all query types.

These findings reaffirm the strength of textual similarity-based models, particularly when combined with robust preprocessing and vector space transformation techniques, in capturing user preferences with high fidelity. When benchmarked against leading recommender models from recent literature, the proposed hybrid framework demonstrated several notable advantages: in contrast to [18], which primarily relies on user reviews and historical behavior, the proposed system mitigates popularity bias by prioritizing semantic relevance between query content and destination profiles. Compared to [20], which emphasizes image-based and activity recommendations, our model offers stronger textual processing capabilities but could be extended with image metadata and caption embedding for multimodal integration. Unlike [2], which activates recommendations only after the user selects a destination, our system enables pre-decision exploration, supporting discovery of non-mainstream or under-visited locations. While [25], [32], and [35] present a hybrid recommendation framework, the accuracy metrics achieved through our TF-IDF+cosine similarity+LSTM ensemble demonstrated superior performance and lower error rates.

Based on empirical findings, several strategic enhancements are recommended to elevate the system's accuracy and applicability:

i) Dataset expansion: increasing the number of curated destination entries will improve model generalization, reduce vector sparsity, and enhance recommendation diversity [1], [18], [22], [29], [30].
ii) Multimodal data integration: enriching the dataset with user reviews, GPS traces, and visual data (images, videos) would improve context awareness, enabling emotionally intelligent and situationally relevant suggestions [20], [26], [31], [38].
iii) Cross-domain transferability: the current framework's modular design enables easy transfer to domains such as education (course recommendation), entertainment (movie or event suggestions), and healthcare (wellness services), where personalized content delivery can drive user engagement [23], [33], [35], [39].

By incorporating these enhancements, the proposed recommendation engine is expected to evolve into a scalable, domain-independent, and highly adaptive system capable of delivering personalized and contextually meaningful recommendations in both tourism and broader knowledge-based application areas.

5. CONCLUSION

In conclusion, this study presents a practical and effective solution for aligning user preferences, expressed in natural language, with structured destination metadata. By applying content-based filtering grounded in semantic similarity, the system delivers personalized domestic tourism recommendations. Core techniques such as N-gram decomposition, TF-IDF vectorization, and the use of a domain-specific dictionary allow for deeper interpretation of user intent. The integration of similarity metrics like Cosine Similarity and Pearson Correlation ensures accurate query matching. Experimental results demonstrate the system's effectiveness in capturing user expectations through relevant suggestions. The findings highlight the importance of dataset scale and semantic diversity in optimizing recommendation performance. This approach also proves to be flexible and scalable, with the potential for extension to other application domains. Additionally, the framework supports multilingual capabilities with minimal adaptation. Ultimately, this research contributes a robust foundation for intelligent, user-centric recommendation systems in the field of digital tourism.

ACKNOWLEDGMENTS

We would like to express our sincere appreciation to the student group from Thuongmai University for their valuable support in data collection and preprocessing.

FUNDING INFORMATION

This research is funded by Thuongmai University, Hanoi, Vietnam.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Nguyen Thi Hoi	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓
Tran Thi Nhung	✓		✓	✓	✓	✓	✓	✓	✓	✓				
Bui Quang Truong	✓		✓	✓	✓	✓	✓	✓	✓					
Nguyen Quang Trung	✓	✓	✓	✓	✓	✓		✓	✓					

C :	**C**onceptualization	I :	**I**nvestigation	Vi :	**Vi**sualization
M :	**M**ethodology	R :	**R**esources	Su :	**Su**pervision
So :	**So**ftware	D :	**D**ata Curation	P :	**P**roject administration
Va :	**Va**lidation	O :	Writing - **O**riginal Draft	Fu :	**Fu**nding acquisition
Fo :	**Fo**rmal analysis	E :	Writing - Review & **E**diting		

CONFLICT OF INTEREST STATEMENT

The authors declare that there are no conflicts of interest related to this study.

INFORMED CONSENT

We have obtained informed consent from all individuals included in this study.

ETHICAL APPROVAL

The research related to human use has been compiled with all the relevant national regulations and institutional policies in accordance with the tenets of the Helsinki Declaration and has been approved by the authors' institutional review board or equivalent committee.

DATA AVAILABILITY

All data utilized for the testing phase were systematically collected from reputable travel websites featuring domestic destinations across Vietnam. A comprehensive list of these data sources is presented in Table 1 of this paper.

REFERENCES

[1] C. Huda, Y. Heryadi, Lukas, and W. Budiharto, “Smart tourism recommender system modeling based on hybrid technique and content boosted collaborative filtering,” *IEEE Access*, vol. 12, pp. 131794–131808, 2024, doi: 10.1109/ACCESS.2024.3450882.

[2] D. Fahrizal, J. Kustija, and M. A. H. Akbar, “Development tourism destination recommendation systems using collaborative and content-based filtering optimized with neural networks,” *Indonesian Journal of Artificial Intelligence and Data Mining*, vol. 7, no. 2, p. 285, Apr. 2024, doi: 10.24014/ijaidm.v7i2.28713.

[3] N. Khan and Dr. M. Haroon, “Trends and techniques used in tourist recommender system : a review,” *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 9, no. 3, pp. 33–39, May 2023, doi: 10.32628/CSEIT23902105.

[4] W.-C. Lin, C. K. Wu, T. K. T. Le, and N. A. Nguyen, “Assessment of Vietnam tourism recovery strategies after COVID-19 using multi-criteria decision-making approach,” *Sustainability*, vol. 15, no. 13, Jun. 2023, doi: 10.3390/su151310047.

[5] World Travel & Tourism Council, “Travel & tourism economic impact research (EIR),” World Travel & Tourism Council. Accessed: Mar. 10, 2025. [Online]. Available: https://wttc.org/research/economic-impact

[6] World Travel & Tourism Council, “Vietnam’s travel & tourism set for a record 2024,” World Travel & Tourism Council. Accessed: Mar. 10, 2025. [Online]. Available: https://wttc.org/news/vietnams-travel-and-tourism-set-for-a-record-2024

[7] Trading Economics, “Vietnam GDP annual growth rate,” Trading Economics. Accessed: Aug. 12, 2024. [Online]. Available: https://tradingeconomics.com/vietnam/gdp-growth-annual

[8] General Statistics Office of Vietnam, “Socio-economic situation report in July and 7 months of 2023,” 2023. [Online]. Available: https://www.nso.gov.vn/en/data-and-statistics/2023/08/socio-economic-situation-report-in-july-and-7-months-of-2023/#:~:text=State budget revenue and expenditure&text=Total state budget expenditure in, the same period last year.

[9] Vietnam National Authority of Tourism, “Vietnam tourism statistics 2023,” Tourism Information Technology Center. Accessed: Aug. 12, 2024. [Online]. Available: https://vietnamtourism.gov.vn/en/statistic/international

[10] Ministry of Culture, Sports and Tourism, “Symposium ‘digital transformation in the field of culture’ (in Vietnamese: *hội thảo chuyên đề ‘chuyển đổi số trong lĩnh vực văn hóa’*),” Ministry of Culture, Sports and Tourism. Accessed: Mar. 10, 2025. [Online]. Available: https://bvhttdl.gov.vn/hoi-thao-chuyen-de-chuyen-doi-so-trong-linh-vuc-van-hoa-20250604093709418.htm

[11] Ng. D. Thang, “Cultural resources promote tourism development in Vietnam,” *Journal of Information Systems Engineering and Management*, vol. 10, no. 47s, pp. 206–214, May 2025, doi: 10.52783/jisem.v10i47s.9248.

[12] H. D. T. Anh, H. N. K. Giao, and H. T. H. Lan, “The impact of smart tourism ecosystem on tourists’ return intention: the case of Hochi Minh City, Vietnam,” in *The 4th International Joint Conference on Hospitality and Tourism: Transformative Trends Shaping the Future of Sustainable Tourism*, Nha Trang, Vietnam: International Joint Conference on Hospitality and Tourism 2024, 2024. [Online]. Available: https://www.researchgate.net/publication/386244589_The_impact_of_smart_tourism_ecosystem_on_tourists%27_return_intention_The_case_of_HoChi_Minh_City_Vietnam

[13] T. Ho *et al.*, “A new model for collecting, storing, and analyzing big data on customer feedback in the tourism industry,” *Interdisciplinary Journal of Information, Knowledge, and Management*, vol. 18, pp. 225–249, 2023, doi: 10.28945/5107.

[14] H. T. Hoi, “Advertising Vietnam’s tourism products in the technology age,” in *Proceedings of the 7th International Conference on Management of e-Commerce and e-Government*, New York, NY, USA: ACM, Jul. 2020, pp. 11–15. doi: 10.1145/3409891.3409892.

[15] P. Do, T. H. V. Phan, and B. B. Gupta, “Developing a Vietnamese tourism question answering system using knowledge graph and deep learning,” *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 20, no. 5, pp. 1–18, Sep. 2021, doi: 10.1145/3453651.

[16] Z. A.-Moud, H. V.-Nejad, and J. Sadri, “Tourism recommendation system based on semantic clustering and sentiment analysis,” *Expert Systems with Applications*, vol. 167, p. 114324, Apr. 2021, doi: 10.1016/j.eswa.2020.114324.

[17] A. Banerjee, A. Satish, F. N. Aisyah, W. Wörndl, and Y. Deldjoo, “SynthTRIPs: a knowledge-grounded framework for benchmark data generation for personalized tourism recommenders,” in *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, USA: ACM, Jul. 2025, pp. 3743–3752. doi: 10.1145/3726302.3730321.

[18] M. Badouch and M. Boutaounte, “Personalized travel recommendation systems: a study of machine learning approaches in tourism,” *Journal of Artificial Intelligence, Machine Learning and Neural Network*, vol. 3, no. 3, pp. 35–45, Apr. 2023, doi: 10.55529/jaimlnn.33.35.45.

[19] A. Pramarta and A. Baizal, “Hybrid recommender system using singular value decomposition and support vector machine in Bali tourism,” *Jurnal Ilmiah Penelitian dan Pembelajaran Informatika*, vol. 7, no. 2, pp. 408–418, May 2022, doi: 10.29100/jipi.v7i2.2770.

[20] I. Fathima and B. Kotaiah, “Deep learning based tourism recommendation system,” *SSRN Electronic Journal*, 2022, doi: 10.2139/ssrn.4286575.

[21] B. H. C, Vijay, S. M, Y. V, and S. H. G, “Personalized travel recommendation system using machine learning,” *International Journal of Advance Research and Innovative Ideas in Education*, vol. 9, no. 3, 2023, [Online]. Available: https://ijariie.com/AdminUploadPdf/Personalized_Travel_Recommendation_System_Using_Machine_Learning_ijariie20353.pdf ?srsltid=AfmBOood-mSKZipaJT3qTG4AEFPC2MUihS2VF2czgjaDEOg7Ks5mTp9k

[22] J. Karthiyayini and R. J. Anandhi, “Personalized travel recommendations system using hybrid filtering and deep learning.” Research Square, Nov. 27, 2024. doi: 10.21203/rs.3.rs-5408442/v1.

[23] X. Xiao, C. Li, X. Wang, and A. Zeng, “Personalized tourism recommendation model based on temporal multilayer sequential neural network,” *Scientific Reports*, vol. 15, no. 1, Jan. 2025, doi: 10.1038/s41598-024-84581-z.

[24] Y. Moshfeghi, B. Piwowarski, and J. M. Jose, “Handling data sparsity in collaborative filtering using emotion and semantic based features,” in *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, New York, NY, USA: ACM, Jul. 2011, pp. 625–634. doi: 10.1145/2009916.2010001.

[25] N. L. Ho, R. K.-W. Lee, and K. H. Lim, “SBTREC - a transformer framework for personalized tour recommendation problem with sentiment analysis,” in *2023 IEEE International Conference on Big Data (BigData)*, IEEE, Dec. 2023, pp. 5790–5798. doi: 10.1109/BigData59044.2023.10386486.

[26] M. H. Nguyen, T. T. Nguyen, M. N. Ta, T. M. Nguyen, and K. V. Nguyen, “RRS: review-based recommendation system using deep learning for Vietnamese,” *SN Computer Science*, vol. 5, no. 5, Apr. 2024, doi: 10.1007/s42979-024-02812-6.

[27] T. H. Nguyen, D. Q. Tran, G. M. Dam, and M. H. Nguyen, “Estimating the similarity of social network users based on behaviors,” *Vietnam Journal of Computer Science*, vol. 5, no. 2, pp. 165–175, May 2018, doi: 10.1007/s40595-018-0112-1.

[28] N. P. Hung and B. T. Khoa, “Applying smart tourism management in Vietnam after COVID-19 pandemic: a qualitative research,” *Quality - Access to Success*, vol. 24, no. 194, Jan. 2023, doi: 10.47750/QAS/24.194.01.

[29] G. K. Murugananda, T. S. S. Angel, S. Kumar, N. Snehalatha, and S. S. Manipaul, “A real-time tourism recommender system using KNN and RBM approach,” in *2023 International Conference on Data Science, Agents and Artificial Intelligence*, IEEE, Dec. 2023, pp. 1–5. doi: 10.1109/ICDSAAI59313.2023.10452558.

[30] S. Kongpeng and A. Hanskunatai, “Tourist destination recommendation system based on machine learning,” in *2024 9th International Conference on Big Data and Computing*, New York, NY, USA: ACM, May 2024, pp. 58–67. doi: 10.1145/3695220.3695229.

[31] T. A. W. Tyas, Z. K. A. Baizal, and R. Dharayani, “Tourist places recommender system using cosine similarity and singular value decomposition methods,” *Jurnal Media Informatika Budidarma*, vol. 5, no. 4, Oct. 2021, doi: 10.30865/mib.v5i4.3151.

[32] P. Zhang, J. Wang, and R. Li, “Tourism-type ontology framework for tourism-type classification, naming, and knowledge organization,” *Heliyon*, vol. 9, no. 4, Apr. 2023, doi: 10.1016/j.heliyon.2023.e15192.

[33] D. Shrestha, T. Wenan, D. Shrestha, N. Rajkarnikar, and S.-R. Jeong, “Personalized tourist recommender system: a data-driven and machine-learning approach,” *Computation*, vol. 12, no. 3, Mar. 2024, doi: 10.3390/computation12030059.

[34] G. Özdemir and V. A. Arzık, “Segmentation of social media users for destinations,” *Tourism*, vol. 70, no. 1, pp. 53–66, Dec. 2021, doi: 10.37741/t.70.1.4.

[35] J. Yoon and C. Choi, “Real-time context-aware recommendation system for tourism,” *Sensors*, vol. 23, no. 7, Apr. 2023, doi: 10.3390/s23073679.

[36] Y. Sui, “Question answering system based on tourism knowledge graph,” *Journal of Physics: Conference Series*, vol. 1883, no. 1, Apr. 2021, doi: 10.1088/1742-6596/1883/1/012064.

[37] J. C. Cuizon and C. G. Agravante, “Sentiment analysis for review rating prediction in a travel journal,” in *Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval*, New York, NY, USA: ACM, Dec. 2020, pp. 70–74. doi: 10.1145/3443279.3443282.

[38] N. Malik and M. Bilal, “Natural language processing for analyzing online customer reviews: a survey, taxonomy, and open research challenges,” *PeerJ Computer Science*, vol. 10, Jul. 2024, doi: 10.7717/peerj-cs.2203.

[39] M. Á. Á.-Carmona *et al.*, “Natural language processing applied to tourism research: a systematic review and future research directions,” *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 10, pp. 10125–10144, Nov. 2022, doi: 10.1016/j.jksuci.2022.10.010.

[40] S. Qaiser and R. Ali, “Text mining: use of TF-IDF to examine the relevance of words to documents,” *International Journal of Computer Applications*, vol. 181, no. 1, pp. 25–29, Jul. 2018, doi: 10.5120/ijca2018917395.

BIOGRAPHIES OF AUTHORS

Nguyen Thi Hoi is received a Ph.D. in Information Systems from the Postal Institute of Technology (PTIT) in 2021. Currently a lecturer at the Faculty of Information Technology and Economics, Thuongmai University (TMU), Vietnam. Areas of interest: economic information systems, e-commerce systems, and social computing. She can be contacted at email: hoint@tmu.edu.vn.

Tran Thi Nhung is a Ph.D. student at the Information System Management in National Economic University (NEU). Currently a lecturer at the Faculty of Information Technology and Economics, Thuongmai University (TMU), Vietnam. Areas of interest: economic information systems, e-commerce systems, optimization computing. She can be contacted at email: nhung.tt@tmu.edu.vn.

Bui Quang Truong is a Ph.D. student at the Information System Management in National Economic University (NEU). Currently a lecturer at the Faculty of Information Technology and Economics, Thuongmai University (TMU), Vietnam. Areas of interest: economic information systems, e-commerce systems, and digital transformation. He can be contacted at email: truongbq@tmu.edu.vn.

Nguyen Quang Trung is a Ph.D. student in Information Technology at National University (NEU). Currently a lecturer at the Faculty of Information Technology and Economics, Thuongmai University (TMU), Vietnam. Areas of interest: economic information systems, e-commerce systems, and quality assurance. He can be contacted at email: trungnq@tmu.edu.vn.