A Content-Based Smart Tourist Recommendation System Using TF-IDF and Cosine Similarity

Pooja Suresh Thombare
Department of MSc Information Technology
D.G Ruparel College of Arts, Science and Commerce
Mumbai, Maharashtra, India
Academic Year: 2025-2026

ABSTRACT

Tourism recommendation systems have become essential tools for assisting travelers in making informed decisions from a large volume of available travel information. Traditional tourism platforms often provide generic suggestions that fail to adapt to individual preferences. This paper presents a Smart Tourist Recommendation System based on content-based filtering techniques using TF-IDF vectorization and cosine similarity. The proposed system analyzes attraction descriptions, categories, and related metadata to generate personalized recommendations without relying on historical user data. A multi-page web interface is implemented using Streamlit to support attraction-wise, city-wise, and comparison-based recommendations. Experimental evaluation shows that the system efficiently delivers relevant recommendations with low computational complexity, making it suitable for real-world smart tourism applications.

Keywords: Smart Tourism, Recommendation System, Content-Based Filtering, TF-IDF, Cosine Similarity, Machine Learning

INTRODUCTION

The rapid growth of digital tourism platforms has significantly increased the availability of travel-related information. While this abundance of data benefits travelers, it also creates challenges in selecting suitable destinations and attractions. Intelligent recommendation systems provide a practical solution by filtering large datasets and presenting personalized suggestions based on user interests.

Machine learning techniques play a crucial role in modern recommender systems. Among them, content-based filtering is widely used due to its ability to generate recommendations without requiring explicit user feedback or large historical datasets. This research focuses on developing a Smart Tourist Recommendation System that leverages textual information of tourist attractions to recommend similar points of interest.

The proposed system aims to assist travelers by offering attraction-wise and city-wise recommendations through an intuitive interface. By combining natural language processing techniques with similarity-based learning, the system demonstrates how AI-driven methods can improve tourism decision-making.

LITERATURE REVIEW

The application of recommendation systems in the tourism domain has gained significant attention over the past decade due to the increasing demand for personalized travel experiences. Early tourism recommendation systems primarily relied on static rule-based methods and popularity-driven rankings, which often failed to capture individual traveler preferences. As digital tourism platforms expanded, researchers began exploring intelligent approaches using machine learning and artificial intelligence to improve recommendation accuracy and relevance.

Several studies have investigated the use of content-based filtering techniques for tourism recommendations. Content-based approaches analyze item attributes such as attraction descriptions, categories, and metadata to generate personalized suggestions. These methods are particularly effective in scenarios where user interaction data is sparse or unavailable. Research by various authors highlights that content-based filtering is well-suited for tourism applications because attractions inherently contain rich textual information that can be exploited for similarity analysis. Techniques such as TF-IDF vectorization and cosine similarity have been widely adopted to convert unstructured text into numerical representations, enabling effective comparison between tourist attractions.

In contrast, collaborative filtering approaches focus on leveraging user behavior patterns, such as ratings and reviews, to recommend items. While collaborative filtering has demonstrated success in e-commerce and entertainment domains, its application in tourism faces challenges related to cold-start problems and limited user data availability. To overcome these limitations, hybrid recommendation systems combining content-based and collaborative filtering have been proposed. Studies conducted in recent years show that hybrid models improve recommendation quality by balancing personalization with collective user behavior, although they introduce increased system complexity and computational overhead.

Context-aware tourism recommendation systems represent another significant advancement in this field. These systems incorporate contextual factors such as location, time, weather, and user constraints to generate dynamic recommendations. Research indicates that context-aware models enhance user satisfaction by adapting suggestions to real-world conditions. However, such systems often require access to real-time data sources and complex integration mechanisms, making them less suitable for lightweight or academic prototype implementations.

More recent studies emphasize the role of explainable and user-friendly interfaces in tourism recommendation systems. Interactive dashboards and web-based platforms allow users to explore recommendations transparently and compare multiple attractions easily. Tools such as Streamlit have been recognized for enabling rapid development of interactive AI applications with minimal overhead. The integration of machine learning algorithms with intuitive interfaces significantly enhances user engagement and practical usability.

Based on the reviewed literature, it is evident that content-based filtering remains a reliable and effective approach for smart tourism recommendation, especially in systems with limited user history. The existing research also highlights a gap in lightweight, explainable, and easily deployable tourism recommender systems that combine efficient algorithms with user-centric interfaces. The proposed Smart Tourist Recommendation System addresses this gap by employing a content-based filtering approach supported by TF-IDF and cosine similarity, along with a multi-page interactive interface designed for ease of use and clarity.

METHODOLOGY AND SYSTEM ARCHITECTURE

This section describes the overall methodology adopted for designing and implementing the Smart Tourist Recommendation System. The proposed system follows a structured pipeline consisting of data acquisition, preprocessing, feature extraction, similarity computation, and recommendation generation. The methodology is designed to ensure simplicity, efficiency, and scalability while maintaining recommendation accuracy.

System Architecture

The architecture of the Smart Tourist Recommendation System is illustrated in Figure 1. The system is organized into distinct functional layers that interact seamlessly to generate personalized tourist recommendations. At the core of the system lies a structured tourism dataset that contains information about various attractions, including city name, attraction name, category, rating, visit duration, and textual description.

The data processing layer is responsible for cleaning and preparing the dataset. This includes handling missing values, removing formatting inconsistencies, and standardizing column names. Once the data is preprocessed, relevant textual attributes; such as attraction name, category, and description are combined to form a unified textual representation.

The feature extraction layer applies TF-IDF (Term Frequency–Inverse Document Frequency) vectorization to convert textual data into numerical vectors. TF-IDF assigns higher importance to terms that are frequent within a document but rare across the dataset, enabling meaningful differentiation between attractions. The generated vectors are stored in a matrix that represents all tourist attractions in a high-dimensional feature space.

The similarity computation layer uses cosine similarity to calculate the degree of similarity between attraction vectors. Cosine similarity measures the angular distance between vectors and provides a normalized similarity score ranging between 0 and 1. Attractions with higher similarity scores are considered more relevant and are ranked accordingly.

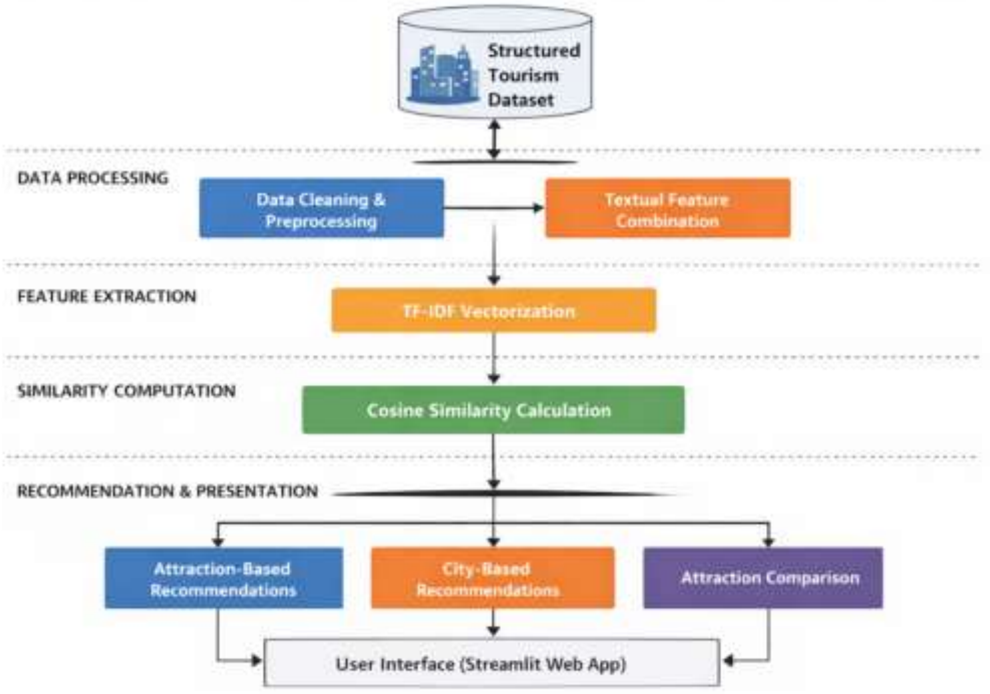


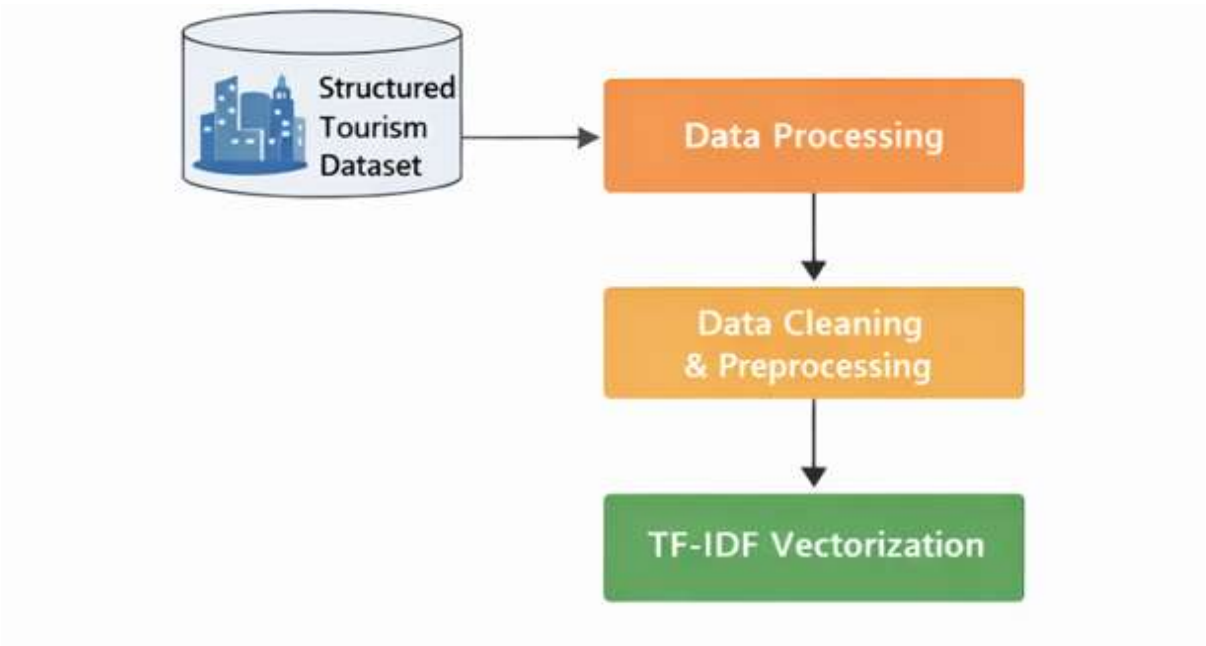
Figure 1: Architecture of Smart Tourist Recommendation System

Data Flow and Processing

The data flow within the Smart Tourist Recommendation System follows a structured, sequential, and modular design, as illustrated in Figure 2. The process begins with loading the tourism dataset into the system, where it is forwarded to the preprocessing module. This module ensures data consistency by cleaning missing values, standardizing text formats, and preparing relevant attributes for further processing. After preprocessing, key textual features such as attraction name, category, and description are combined and transformed into numerical form using TF-IDF vectorization.

Once vectorization is completed, the system computes pairwise cosine similarity scores between all attractions. When a user selects a specific attraction, the corresponding feature vector is retrieved and compared with other attraction vectors to generate attraction-wise recommendations. For city-wise recommendations, the system filters attractions based on the selected city and ranks them according to their ratings. This dual recommendation workflow improves system flexibility and allows users to explore tourist destinations through multiple personalized recommendation modes.

Furthermore, the modular data flow design improves system scalability and maintainability. Each processing stage operates independently, allowing future enhancements such as incorporating user preferences, real-time data, or additional recommendation algorithms without disrupting the existing workflow. This separation of concerns ensures efficient data handling, faster computation, and easier system updates, making the recommendation framework adaptable to evolving tourism and user requirements.

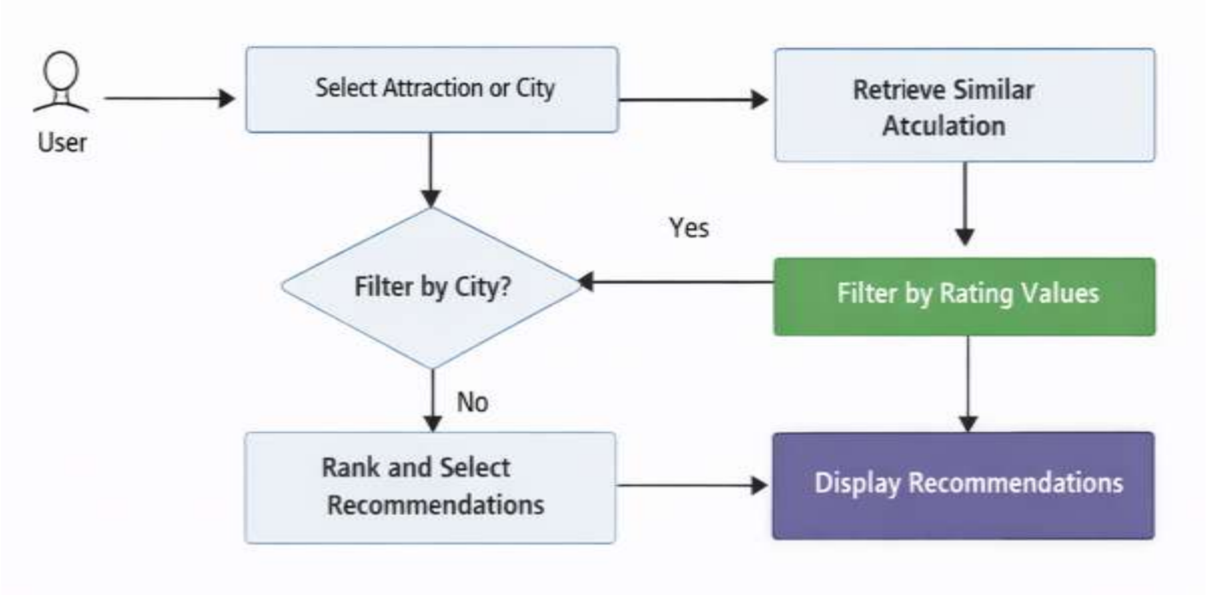


Recommendation Workflow

The overall recommendation workflow is designed to ensure both efficiency and accuracy in generating tourist suggestions. Initially, the system loads the preprocessed dataset and constructs a similarity matrix using TF-IDF vectors and cosine similarity. This matrix is computed once during system initialization and stored in memory, eliminating the need for repetitive similarity calculations during user interaction. As a result, when a user selects an attraction or city, the system can retrieve recommendations instantly, ensuring low response time and smooth real-time interaction.

Once a user input is received, the workflow follows a decision-based path depending on the selected recommendation mode. For attraction-wise recommendations, the system identifies the corresponding attraction vector and retrieves the most similar attractions based on cosine similarity scores. In contrast, for city-wise recommendations, the system filters attractions by the selected city and ranks them using rating values. The final results are then formatted and displayed through the user interface, allowing users to explore, compare, and analyze recommended destinations efficiently.

The modular workflow structure enhances system flexibility and extensibility. Each stage—input handling, similarity computation, ranking, and result presentation—operates independently, allowing individual components to be modified or replaced without affecting the entire system. This design supports future improvements such as hybrid recommendation models, user preference learning, or real-time data updates, making the system robust and adaptable for advanced smart tourism applications.


User
Select Attraction or City
Retrieve Similar Atculation
Yes
Filter by City?
Filter by Rating Values
No
Rank and Select Recommendations
Display Recommendations

ALGORITHM DESIGN AND DESCRIPTION

The proposed Smart Tourist Recommendation System employs a content-based filtering algorithm to generate personalized attraction recommendations. Content-based filtering focuses on analyzing the intrinsic attributes of items rather than relying on user interaction data. In the tourism domain, attractions naturally contain rich descriptive information, making content-based approaches particularly suitable for recommendation tasks. The algorithm is designed to identify attractions that are similar in nature based on their textual descriptions and categorical attributes.

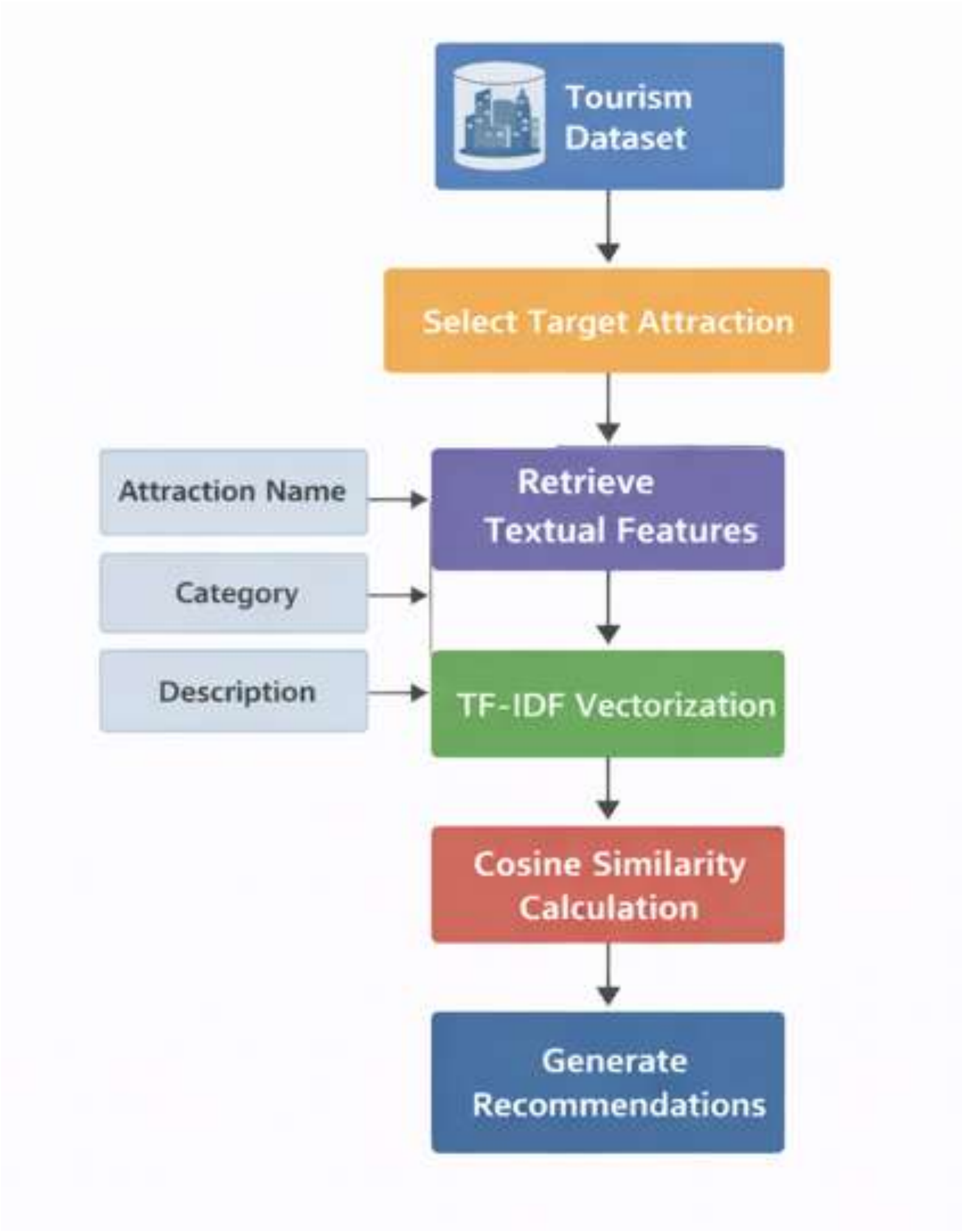
The algorithm begins by extracting relevant textual features from the dataset, including attraction names, categories, and descriptive summaries. These textual fields are merged into a single composite document for each attraction. This combined representation captures the semantic characteristics of the attraction and serves as the foundation for similarity analysis. Before further processing, the text is normalized by converting it into a consistent format to ensure uniformity across the dataset.

To transform textual information into a numerical format that can be processed by machine learning algorithms, TF-IDF (Term Frequency–Inverse Document Frequency) vectorization is applied. TF-IDF assigns a weight to each term based on its frequency within a document and its rarity across the entire dataset. As a result, common but uninformative words receive lower weights, while distinctive terms that better describe an attraction receive higher importance. The outcome of this step is a high-dimensional vector representation for each attraction, where each vector captures the semantic content of the corresponding tourist location.

Once the TF-IDF vectors are generated, cosine similarity is used to compute the similarity between attraction pairs. Cosine similarity measures the cosine of the angle between two vectors, producing a similarity score that ranges from 0 to 1. A higher score indicates greater similarity between attractions. By computing similarity scores between all attraction vectors, the system constructs a similarity matrix that serves as the core of the recommendation engine.

When a user selects a particular attraction, the system retrieves its corresponding vector from the similarity matrix and identifies other attractions with the highest similarity scores. These attractions are ranked and presented as recommendations. For city-wise recommendations, the algorithm filters attractions based on the selected city and ranks them using rating values. This dual strategy allows the system to support both similarity-based exploration and city-based discovery, enhancing overall usability.

The algorithm is computationally efficient, as the similarity matrix is precomputed and reused during runtime. This design minimizes processing overhead and enables real-time recommendation delivery. The simplicity and interpretability of the algorithm also make it suitable for academic and practical applications in smart tourism environments.



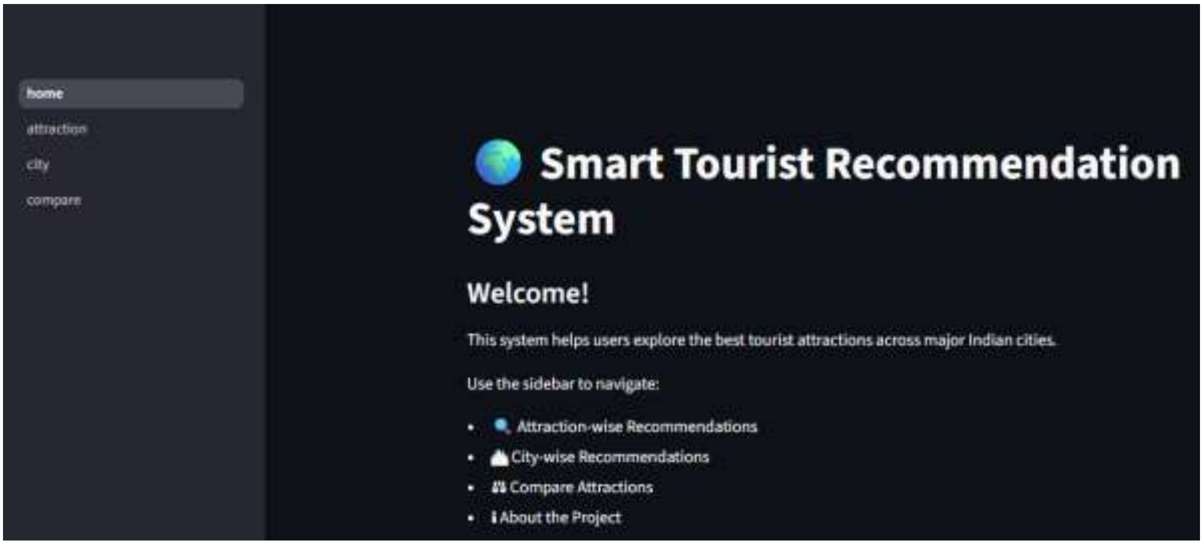
IMPLEMENTATION AND EXPERIMENTAL RESULTS

The Smart Tourist Recommendation System is implemented using Python, leveraging widely adopted libraries such as Pandas for data handling and Scikit-learn for machine learning operations. The system's user interface is developed using Streamlit, which enables rapid creation of interactive, web-based applications. This implementation approach ensures ease of deployment, platform independence, and smooth interaction between the recommendation engine and the end user.

During implementation, the dataset is loaded dynamically at runtime and preprocessed to ensure consistency. Textual features are vectorized using the TF-IDF model, and a cosine similarity matrix is computed once during initialization. This precomputation significantly reduces response time during user interaction, allowing recommendations to be generated instantly when a user selects an attraction or city. The system supports attraction-wise recommendations based on similarity scores, city-wise recommendations based on rating rankings, and a comparison feature that allows users to evaluate multiple attractions side by side.

The application follows a multi-page structure that includes a landing page, recommendation pages, a comparison page, and an informational about page. Custom CSS styling is applied to improve layout consistency, readability, and overall visual appeal. The interface design emphasizes simplicity and clarity, enabling users with minimal technical expertise to navigate the system effortlessly.

Experimental evaluation was conducted using a dataset of tourist attractions across major Indian cities. The results demonstrate that the system consistently produces meaningful recommendations that align with the thematic and descriptive similarities of attractions. Attraction-wise recommendations successfully identify related points of interest, while city-wise recommendations effectively highlight top-rated attractions. The system exhibits low computational overhead and stable performance, validating its suitability for real-time tourism recommendation scenarios.



🌍 Smart Tourist Recommendation System

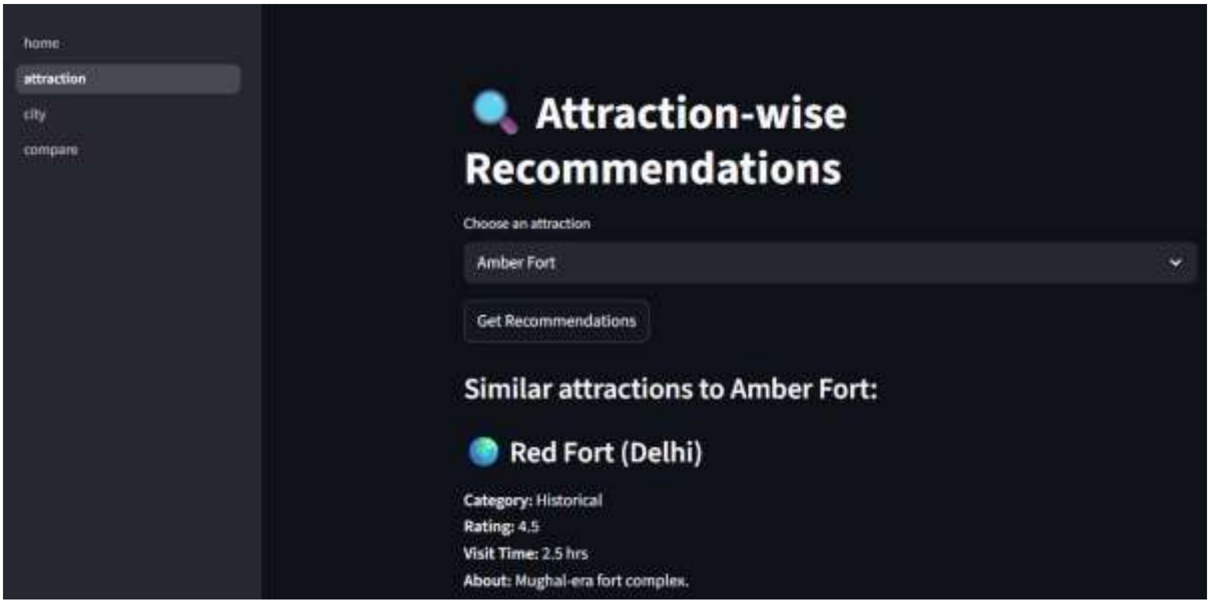
Welcome!

This system helps users explore the best tourist attractions across major Indian cities.

Use the sidebar to navigate:

- 🔍 Attraction-wise Recommendations
- City-wise Recommendations
- Compare Attractions
- ℹ About the Project

Landing Page



🔍 Attraction-wise Recommendations

Choose an attraction

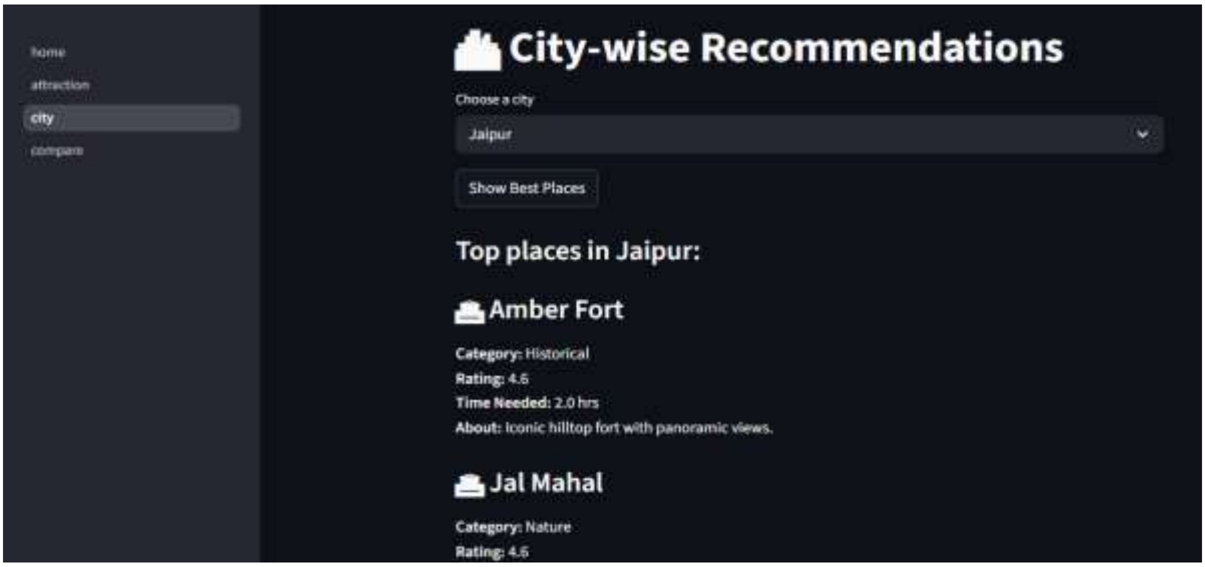
Amber Fort

Similar attractions to Amber Fort:

🌍 Red Fort (Delhi)

Category: Historical
Rating: 4.5
Visit Time: 2.5 hrs
About: Mughal-era fort complex.

Attraction & City Page



🏙 City-wise Recommendations

Choose a city

Jaipur

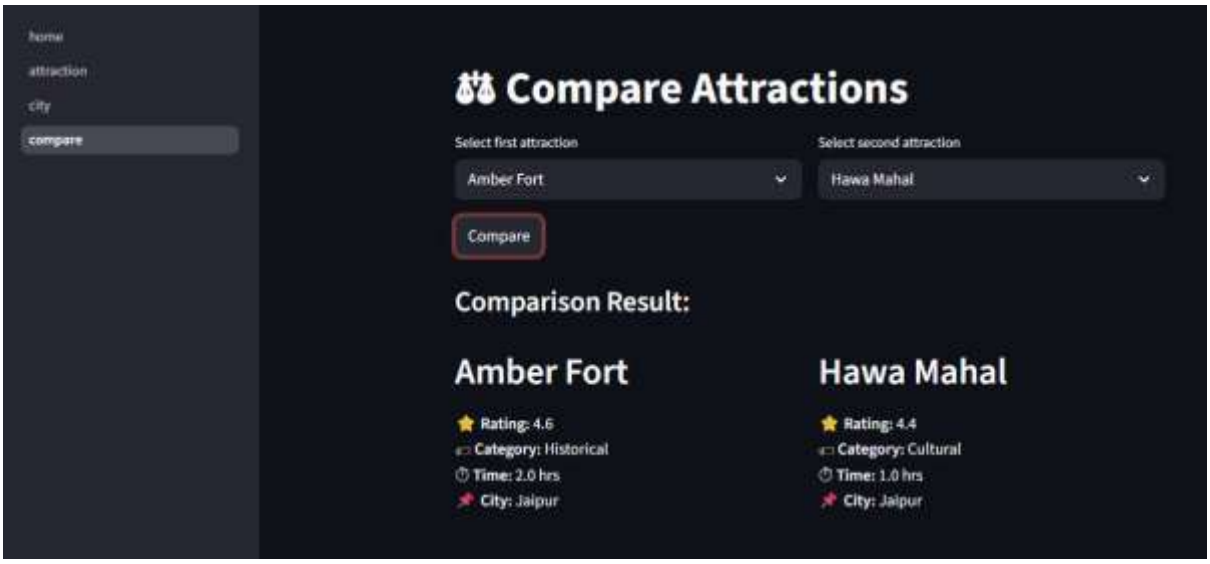
Top places in Jaipur:

🏰 Amber Fort

Category: Historical
Rating: 4.6
Time Needed: 2.0 hrs
About: Iconic hilltop fort with panoramic views.

🏰 Jal Mahal

Category: Nature
Rating: 4.6



⚖ Compare Attractions

Select first attraction: Amber Fort

Select second attraction: Hawa Mahal

Comparison Result:

Amber Fort	Hawa Mahal
⭐ **Rating:** 4.6	⭐ **Rating:** 4.4
🏷 **Category:** Historical	🏷 **Category:** Cultural
⏱ **Time:** 2.0 hrs	⏱ **Time:** 1.0 hrs
📌 **City:** Jaipur	📌 **City:** Jaipur

Compare Page

CONCLUSION AND KEY ACHIEVEMENTS

This research presented an AI-based Smart Tourist Recommendation System that utilizes content-based filtering techniques to provide personalized tourist attraction recommendations. By leveraging TF-IDF vectorization and cosine similarity, the system effectively analyzes attraction descriptions and categories to identify similarities between points of interest. Unlike traditional recommendation systems that rely heavily on user interaction history, the proposed approach performs efficiently even in the absence of prior user data, making it suitable for first-time users and small-scale deployments.

The system successfully integrates machine learning algorithms with an intuitive, multi-page web interface developed using Streamlit. The inclusion of attraction-wise recommendations, city-wise ranking, and attraction comparison features enhances user engagement and supports informed travel planning. The project demonstrates the practical applicability of artificial intelligence and natural language processing techniques in smart tourism environments, providing a strong foundation for further research and development.

REFERENCES

1. Personalized Travel Recommendation Systems: A Study of Machine Learning Approaches in Tourism (2023) – HM Journals (Mohamed Badouch & Mehdi Boutaounte)
https://www.researchgate.net/publication/370274670_Personalized_Travel_Recommendation_Systems_A_Study_of_Machine_Learning_Approaches_in_Tourism

2. Smart Tourism Using IoT and AI: A Review - ScienceDirect, 2020 (Zahra Abbasi-Moud, Hamed Vahdat-Nejad & Javad Sadri)
https://www.sciencedirect.com/science/article/abs/pii/S2211973625000455

3. An Extensive Study on the Evolution of Context-Aware Personalized Travel Recommender Systems 2020 - (Shini Renjith, A. Sreekumar & M. Jathavedan)
https://www.sciencedirect.com/science/article/abs/pii/S0306457319300111

4. Mobile Recommender Systems in Tourism 2014 – (Damianos Gavalas, Charalampos Konstantopoulos, Konstantinos Mastakas & Grammati Pantziou)
https://www.researchgate.net/publication/260113086_Mobile_Recommender_Systems_in_Tourism