Enhancing tourism growth using Web Scraping and KNN, Bayesian Network recommendation system-based Hybrid Mobile application

Dr. Thirupathi Regula[a], Anshar Ali[b], and Dr. Dhanasekar N.[c]

[a] Department of Information Technology, College of Computing and Information Sciences, University of Technology and Applied Science, Muscat, Oman, thirupathi.regula@utas.edu.om | regulathirupathi@gmail.com

[b] Department of Information Technology, College of Computing and Information Sciences, University of Technology and Applied Science, Muscat, Oman, anshar.ali@utas.edu.om

[c] Department of Information Technology (Math Section), College of Computing and Information Sciences, University of Technology and Applied Science, Muscat, Oman, dhanasekar.natarajan@utas.edu.om

Abstract:

Tourism is a dynamic, ever-changing industry that needs to benefit greatly from technological improvements. Mobile and mobile applications are an integral part of daily life. The tour planning is not an easy task which requires cumbersome planning with many bookings like flight bookings, hotel booking, taxi booking and restaurant booking. These bookings should be made within the user's preferences like the budget, duration, ratings and others. So, with these constraints a recommendation system is needed to ease the tour plan within the preferences. This research investigates the development of a hybrid mobile app that runs on cross platforms (Android and iOS) and is designed to help tourism grow by offering personalized recommendations. The app gathers up-to-date information from various tourism sources using web scraping methods, providing users with detailed details about places to visit, hotel stay, transport booking, restaurant booking based on reviews from other users. The app's recommendation system uses a mix of two models: K-Nearest Neighbors (KNN) and Bayesian Network (BN). KNN helps make suggestions based on what users like and how they behave. The Bayesian Network improves the decision-making process by understanding the likelihood of certain features being connected. Together, these methods help the app offer better, more tailored recommendations to users. The research applied on the booking.com website to scrap the data and for tour recommendation. The performance metrics are calculated for KNN, Bayesian and hybrid algorithms. Based on the RMSE, MAE, Precision and Recall Metrics values, the hybrid approach provides the best result among all other algorithms. So, this proposed approach significantly improves the tour recommendation with mobile application and hybrid recommendation approach.

Keywords: **Hybrid Mobile Application, Recommendation system, KNN algorithm, Bayesian Network and Tourism growth**

Introduction

The global economy relies a lot on tourism, which helps people from different cultures connect, boosts economic growth, and supports many service industries. However, as the world becomes more digital, the way people plan, and experience travel has changed. Today, tourists use mobile devices and online platforms to make decisions based on recommendations from others, their own preferences, and the latest information. Because of this, the biggest challenge for the tourism industry is to provide tourists with personalized and meaningful experiences. To meet this need, combining online data collection, hybrid mobile apps, and advanced recommendation systems could greatly improve travel experiences and make trip planning faster and easier.

Hybrid mobile application combines web technology with native mobile features to work on many platforms such as Android and iOS. These apps help tourism businesses connect with more people and offer the same experience on different devices. Tourists can use these apps to get personalized suggestions for places to visit, stay, eat, and things to do based on their interests and travel plans. However, providing these tailored services requires advanced systems that can understand user preferences and behavior from large and varied data.

Web scraping, which is the automatic process of collecting data from websites, is very important here. By scraping public tourism-related information online, such as reviews, ratings, prices, and social media posts for flight booking, taxi booking and hotel booking. we can gather a lot of real-time, external data using this web scraping. When this data is mixed with user preferences, it helps create more personalized and accurate recommendations. This ensures that tourists are not only shown popular spots but also unique experiences and lesser-known places they might otherwise miss.

To make the recommendation process better, we need to use advanced machine learning methods to analyze the large amounts of data collected from web scraping and hybrid applications. Three popular algorithms that can significantly improve personalization are K-Nearest Neighbors (KNN) and Bayesian Networks. KNN uses past user data to suggest items that similar users have liked. MF works by breaking down large tables of user-item interactions to find hidden patterns that affect preferences. Bayesian Networks, which are based on probability, can include factors like weather or seasonal trends, as well as uncertainty, to make recommendations more accurate.

This research aims to explore how using web scraping, hybrid mobile apps, and advanced recommendation algorithms (like KNN, MF, and Bayesian Networks) can make travel experiences better. By combining these technologies into a smart tourism platform, the system will provide travelers with personalized suggestions that change based on their preferences, real-time data, and the situation. The goal is to improve the field of smart tourism by using the latest technology to enhance user experiences and help the travel and tourism industry grow.

Literature Review:

Ponciano, et al. (2021, June) [1] developed the mobile application for inclusive tourism. A mobile application that uses Google Maps and an algorithm to classify the accessibility of each tourist attraction. The App informs the person with disabilities right away whether a particular point is inaccessible due to their features.

The customers review data collection using multithreaded web scraping approach proposed by Nariman, D.et al. (2024, April) [2]. The study suggests using multiple threads to collect data. Create a system that uses multiple threads for scraping and test how well it works to show that it is faster and more efficient. This study will talk about the proposed system and share the results of tests on how well it gathers data. Lastly, the article will look at the benefits of using multiple threads, the advantages of the tool, and how it could be used in different industries in the future.

Huda et al. (2024) [3] introduced a new way to build a tourist recommendation system - RecSys by combining User-Based Collaborative Filtering (UBCF), Demographic Filtering (DF), Aspect-Based Sentiment Analysis (ABSA), and Content-Boosted Collaborative Filtering. The dataset was created by combining data from TripAdvisor and Google Maps to get more detailed user information. The system's performance was tested using two measures, MAE and RMSE, over 5 rounds of 10-Fold Cross Validation. The results showed an improvement of 84.7% and 82.3% compared to using just UBCF with a fully filled user-item matrix. This approach avoids common problems like the cold-start issue and sparse data by using artificially created data to make the system work better.

For the industrial recommender system, Liu, H.et al. (2021) [4] proposed an efficient deep matrix factorization (EDMF) method that works together with learning from customer reviews. They found two key features in customer reviews. First, EDMF uses convolutional neural networks and a word-attention mechanism to understand the interaction aspects mentioned in a single review. Second, since review information is often sparse (not detailed), They used L0 norm to limit the review data. Additionally, they created a loss function using maximum posteriori estimation theory, where the interaction and sparsity features are turned into two prior probability functions.

Santoso, A. J et al. (2017) [5] developed a M-guide app that mixes location-based services with a hybrid recommendation system to suggest places for tourists. The app uses two methods: collaborative filtering and content-based filtering, which recommend tourist spots based on past visitor data. It also uses the k-nearest neighbor (k-NN) algorithm to rank these tourist spots. Since smartphones are widely used, this research explores how M-Guide can help by recommending tourist attractions and guiding visitors to places in East Timor.

Kbaier et al. (2017) [6] proposed a combined recommendation system that uses collaborative filtering (CF), content-based filtering (CB), and demographic filtering (DF). It uses different machine learning techniques, such as K-nearest neighbors (K-NN) for CB and CF, and decision trees for DF. To make the recommendations more accurate, two combination methods are used: switching and weighted. A new linear programming model is applied to find the best weights for the weighted method. The system is tested using TripAdvisor data, and many evaluation metrics are used to check its performance.

Al Ghobari and his team (2021) [7] created a recommendation system called LAPTA (Location-Aware Personalized Traveler Assistance) that uses GPS and user preferences to give personalized and location-based suggestions. LAPTA organizes Google location data into name and category tags. The system uses the K-Nearest algorithm to match these tags with what the user is looking for, providing personalized recommendations. To make the system more user-friendly, it suggests popular nearby attractions using Google's point of interest feature. LAPTA was found to be more reliable and accurate compared to other recommendation systems they reviewed.

Zhang, Q. (2022) [8] demonstrated a method for first clustering users on the user item rating matrix and then locating the nearest neighbors in the clusters with high similarity to the target user, which successfully decreases the query space and improves recommendation. this method provided a fuzzy-improved K-means algorithm to cluster items in the product attribute matrix and then fuses the similarity of item belongingness to clusters in the fuzzy clustering.

Fahrizal. D [9] proposed an automated application that recommends attractions and events mainly based on user's preferences, actions, context over time and demographic data. This approach applied the context modeling and advanced machine learning approaches to overcome the drawbacks of traditional recommendations systems, especially for newcomers. The system adaptation and optimization thus prove to supplement and better destination branding and travel experience significantly. It has immense consequences not only for the scholarly masses but also for the marketing experts.

Methodology:

The proposed methodology using the Hybrid mobile application that runs on Android and iOS platforms. The Selenium WebDriver and BeautifulSoup are used for web scraping that will extract data from the websites and the data can be stored in any format. For example, Json format. This extracted data will be used in the mobile application database for tour recommendation. Our methodology uses the KNN and Bayesian Network recommendation algorithms and hybrid recommendation also for best performance. Since it is a hybrid mobile application with limited resources in the mobile device, the simple recommendation algorithms like KNN, Bayesian Network and hybrid algorithms are more suitable than the Deep learning recommendation algorithms.

KNN:

KNN is a non-parametric, instance-based learning technique. The algorithm identifies the "K" closest data points to the input and predicts the outcome based on their majority (classification) or average (regression). The following steps are involved in the KNN algorithms. Building a User-Item Matrix, Similarity Calculation, Finding the K Nearest Neighbors, Generating Recommendations, Location-Aware and Real-Time Recommendations

Bayesian Network:

A Bayesian Network is a graphical model that represents variables and their conditional connections using a directed acyclic graph (DAG). The model is based on Bayesian probability and considers probabilistic correlations between variables. The network uses Bayes' Theorem to calculate probabilities and make predictions. In Bayesian network the node represents variables in recommendation system such as user preferences, item features, rating, location and others. The edges define the conditional or normal dependencies between the variables(edges). Each node is associated with a Conditional Probability Tables (CPTs) that specifies the probability of the node given its parent node. The following figure 1 shows the BN in tree structure.

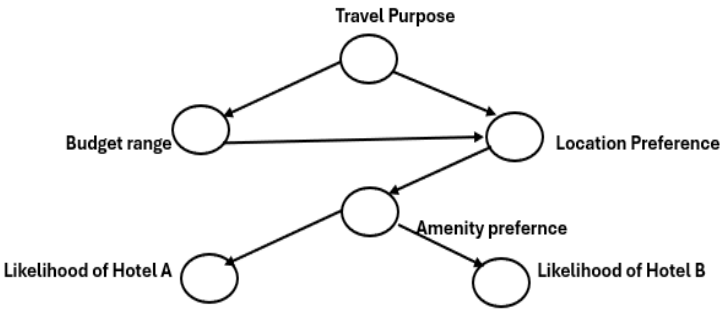

Figure 1: Bayesian Network recommendation in tree structure

The Bayesian Network is constructed in two phases:

Structure Learning: Identify dependent variables to define the graph structure. Use domain knowledge or data-driven methods (such as constraint- or score-based learning).

Parameter Learning: Estimate the conditional probabilities in CPTs using methods such as Maximum Likelihood Estimation (MLE) or Bayesian Estimation.

Conditional Probability Representation is calculated as follows.

$$P(H, A, L) = P(H|A, L) \cdot P(A|L) \cdot P(L) \quad \text{.....1}$$

Where: H is the Probability of booking a hotel, A is the Attributes like amenities or price and L is the Location preference.

Inference for Recommendation is calculated by using the formula below

$$P(H|E) = \frac{P(E|H) \cdot P(H)}{P(E)} \quad \ldots\ldots\ldots\ldots 2$$

Where: E is the preference (e.g., location, budget, ratings).P(H|E) is the Probability of booking given the evidence.

Aspect	KNN	Bayesian Network
Model Type	Instance-based (lazy learning)	Probabilistic (graphical model)
Complexity	Simple to implement but computationally expensive during prediction	More complex due to graph construction and probability calculations
Training Process	Minimal (stores the training data)	Requires learning the graph structure and probabilities
Prediction Speed	Slower, as it calculates distances for each query	Faster once the network is trained
Feature Relationships	Does not explicitly model relationships	Explicitly models relationships using conditional dependencies

Table 1: Comparison of KNN and BN

Hybrid Recommendation:

The Hybrid recommendation that combines the KNN and Bayesian Network algorithms to handle the varieties of dataset and huge data. It gives the best performance than the individual recommendation algorithm (KNN and Bayesian Network).

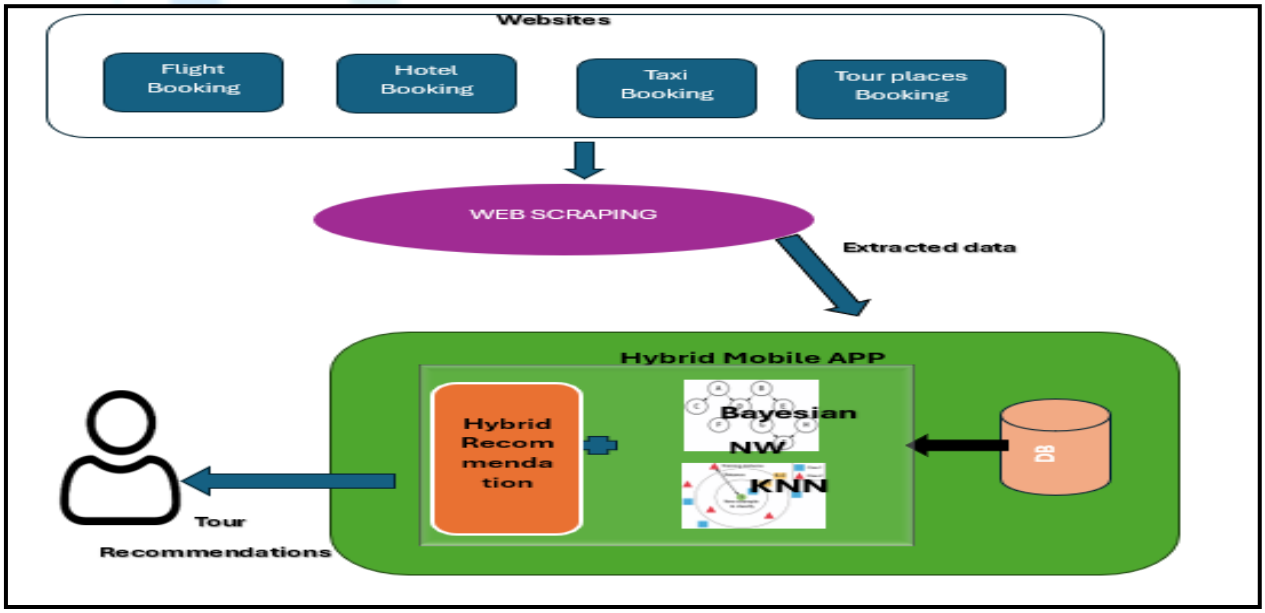

Figure 2: Architecture of Proposed Recommendation system using

Results and Discussion:

This approach is applied on the **booking.com** website and scraped the data for flights, taxis and hotels booking. The recommendation algorithms are applied on the scraped data to recommend the best for the tourists. The performance evaluation of the algorithms using the following things:

Root Mean Squared Error (RMSE)	It measures the average difference between values predicted by a model and the actual values.
Mean Absolute Error (MAE)	Measures the average absolute difference between predicted and actual ratings. For better performance the value should be small.
Precision	Proportion of recommended items that are relevant.
Recall	Proportion of relevant items that are recommended.

Table 2: Performance metrics of Recommendation

The table 3 below shows the scraped data from booking.com website for flight booking.

airline	price	duration	departure_time	arrival_time
Oman Air	601.77	8h 15m	14:10	18:25
Etihad Airways, Wizz Air Abu Dhabi, Wizz Air UK	330.94	33h 20m	18:50	00:10
Etihad Airways	499.08	10h 45m	11:40	18:25
Etihad Airways	499.08	10h 25m	23:20	06:45
Etihad Airways	523.54	10h 45m	11:40	18:25
Etihad Airways	499.08	10h 40m	04:25	12:05
Etihad Airways	523.54	10h 45m	11:40	18:25
Etihad Airways	499.08	10h 25m	23:20	06:45
Etihad Airways	523.54	11h 25m	23:20	06:45
Etihad Airways	523.54	10h 40m	04:25	12:05
Etihad Airways	523.54	11h 25m	23:20	06:45
Etihad Airways	523.54	10h 40m	04:25	12:05
Etihad Airways	499.08	11h 25m	23:20	06:45
Etihad Airways	499.08	10h 40m	04:25	12:05
Gulf Air	552.46	10h 40m	23:55	06:35

Table 3: Real-time web scraped data from booking.com for flight booking

KNN Algorithm: Table 4 shows the KNN recommendation for flight booking from booking.com.

airline	price	duration	Departure_time	arrival_time
Etihad Air	499.08	10h 45m	11:40	18:25
Etihad Air	499.08	11h 25m	23:20	06:45
Etihad Air	499.08	11h 25m	23:20	06:45
Etihad Air	499.08	11h 40m	04:25	12:05
Etihad Air	499.08	10h 45m	11:40	18:25

Table 4: KNN recommendation for flight booking

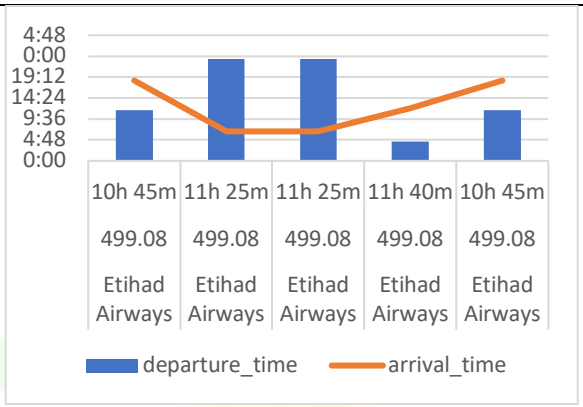

Figure 3: Flight booking recommendation using KNN

Bayesian Network Algorithm:

Figure 4 shows the recommendation of flight booking using Bayesian network.

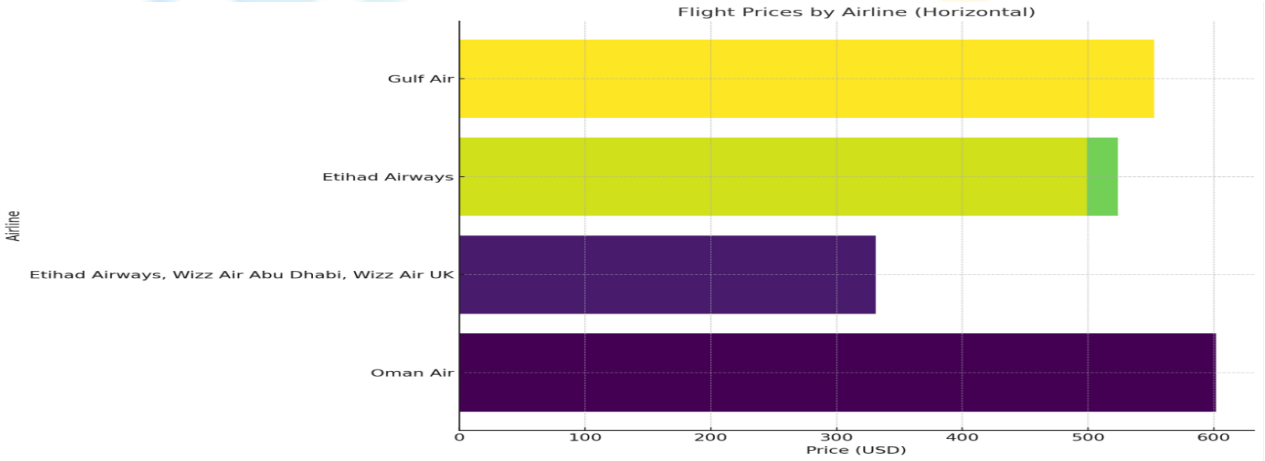

Figure 4: Bayesian Network recommendation for flight booking

Hybrid recommendation:

Figure 5 shows the recommendation of flight booking using hybrid approach.

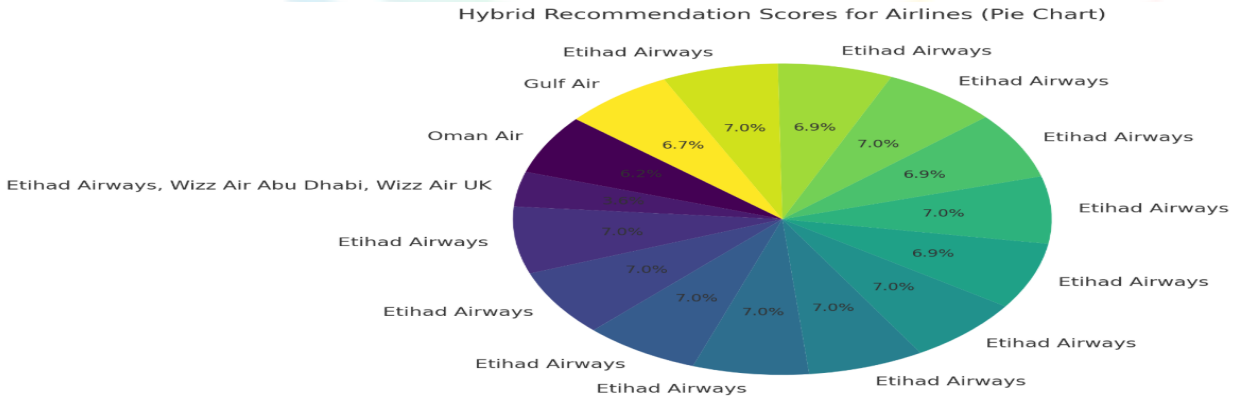

Figure 5: Hybrid recommendation for flight booking

Figure 6 shows the recommendation of flight booking using KNN, BN and hybrid approach.

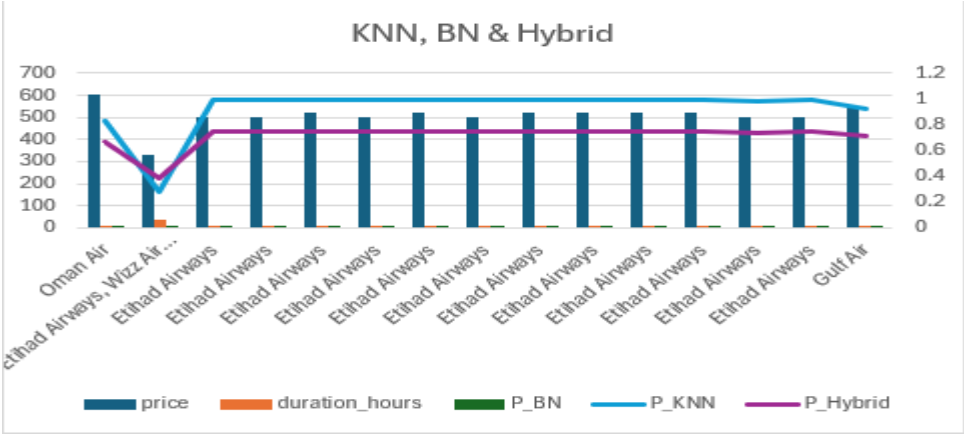

Figure 6: KNN, Bayesian and Hybrid recommendation for flight booking

Hotel Booking:

Table 5 gives the scraped data from booking.com website for hotel booking. Similarly, the KNN, BN and Hybrid algorithms are applied and metrics values are analyzed.

name	price	location	rating	reviews
ibis London City - Shoreditch	0	Tower Hamlets, London	7.9	7216
Pestana Chelsea Bridge Hotel & SPA	952	Wandsworth, London	8.3	4786
ibis budget London Barking	249	Barking	6.6	4716
High Quality Lovely Full Flat Near the station in central London	336	Tower Hamlets, London	8.8	2200
Toby Carvery Beckenham by Innkeeper's Collection	336	Bromley	8.2	2290
Ramada London North	214	Barnet	7.6	1911
The Selwyn, Richmond	686	Richmond Town, Richmond upon Thames	8.7	1941
ibis budget London Hounslow	194	Hounslow	6.7	1911
ibis budget London Heathrow Central	194	Cranford, Hounslow	7	1111
Ridgemount Hotel	956	Camden, London	8.7	3070
Dolphin House Serviced Apartments	954	Westminster Borough, London	8.5	1096
Vertus Edit Canary Wharf	457	Tower Hamlets, London	8.7	5023
The Columbia	416	Westminster Borough, London	7.6	1491
St Giles London - A St Giles Hotel	564	Camden, London	7.8	15289
ibis London Barking	89	Barking	6.6	3000
The Manor Elstree	451	Elstree	7.5	768
Chester Hotel	74	Westminster Borough, London	7.2	536
Holiday Inn Express London - Dartford, an IHG Hotel	405	Dartford	7.7	1877
City Sleeper at Royal National Hotel	661	Camden, London	7.7	7569
Park Grand Heathrow	746	Hounslow	7.6	6777
The Westminster London, Curio Collection by Hilton	879	Westminster Borough, London	8	7356
Hampton by Hilton London Croydon	106	Croydon	8.1	2512
Lords Hotel	70	Westminster Borough, London	7.5	1596
Chessington Hotel	459	Chessington	8.2	4589
Novotel London Brentford	708	Brentford	8.3	7019
Vancouver Hotel and Studios	816	Westminster Borough, London	8.2	2466

Table 5: Scraped data for hotel booking from booking.com

Table 6 shows the performance analysis of KNN, BN and Hybrid algorithms.

Metric	KNN	Bayesian Network	Hybrid Recommendation System
RMSE	Moderate	Low	Very Low
MAE	Moderate	Low	Very Low
Precision	Moderate	High	Very High
Recall	Moderate	High	Very High

Evaluation Metrics	Bayesian Network Only	KNN Only	Hybrid (KNN+ MF)
Root Mean Square error (RMSE)	**0.72**	**0.83**	**0.70**
Mean Absolute Error (MAE)	**0.6**	**0.72**	**0.61**
Precision	**0.74**	**0.68**	**0.75**
Recall	**0.68**	**0.70**	**0.74**

Table 6: Performance analysis of KNN, Bayesian Network and Hybrid recommendation system

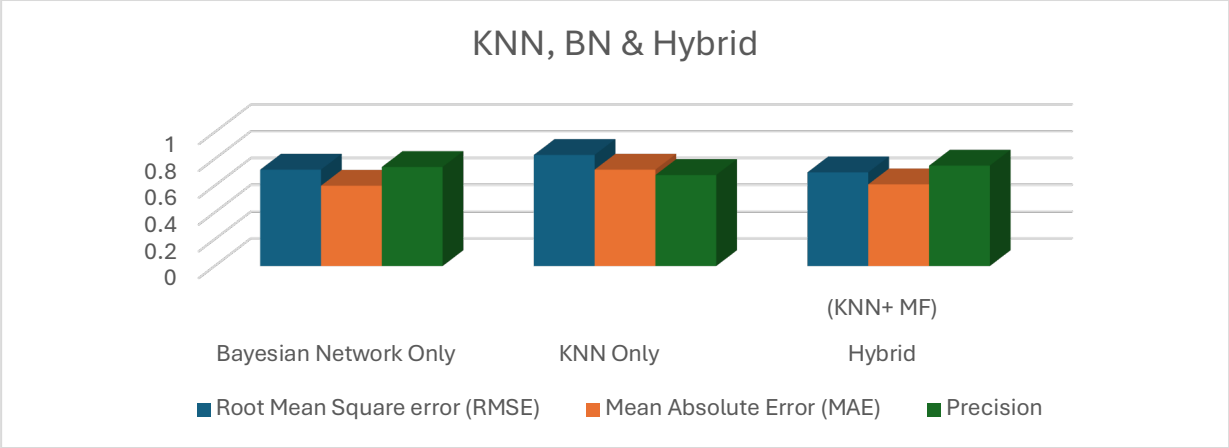

Figure 7: Performance analysis of KNN, Bayesian and Hybrid recommendation system

Metric	Web Scraping	KNN	Bayesian Network	Hybrid Recommendation
Processing Speed	**6**	**4**	**6**	**8**
Scalability	**8**	**4**	**6**	**8**
Accuracy	**0**	**8**	**9**	**10**
Resource Usage	**5**	**3**	**5**	**5**
Real-Time Efficiency	**4**	**3**	**6**	**7**

Table 7: Components of performance

Conclusion:

Compared to separate KNN and Bayesian Network models, the combined KNN and Bayesian Network recommendation system gives the best performance. Generally, the mobile phones are having very low resources such as RAM, processor speed, storage and battery. So, applying the latest deep learning-based recommendation systems in this mobile application will need more resources. But KNN and BN recommendation algorithms are less complex and need less resources compared to deep learning algorithms. The hybrid system provides more precise, pertinent, and customized tour recommendations by utilizing the advantages of both approaches. This hybrid approach overcomes the cold start concerns, data sparsity problems effectively.

Future Enhancement:

In the future the latest deep learning-based algorithms and hybrid algorithms will be used for recommendation. Recommendation system will be applied on diverse dataset to achieve the best performance on recommendation.

Acknowledgments

We extend our heartfelt gratitude to the University of Technology and Applied Sciences (UTAS), Muscat, for their generous support and funding of this research project through the Internal Research Funding Program under grant No: UTAS-IRFP-24-MCT/13.

We are also deeply appreciative of the invaluable assistance and encouragement provided by our colleagues, mentors, and the administrative staff at UTAS-Muscat. Their insights and guidance have greatly contributed to the successful completion of this research.

References:

[1] Ponciano, V., Pires, I. M., Ribeiro, F. R., & Garcia, N. M. (2021, June). Mobile application for inclusive tourism. In *2021 16th Iberian Conference on Information Systems and Technologies (CISTI)* (pp. 1-5). IEEE. . doi: 10.23919/CISTI52073.2021.9476276.

[2] Nariman, D. (2024, April). Enhancing Effectiveness and Efficiency of Customers Reviews Data Collection Through Multithreaded Web Scraping Approach. In International Conference on Advanced Information Networking and Applications (pp. 282-291). Cham: Springer Nature Switzerland.

[3] Huda, C., Heryadi, Y., & Budiharto, W. (2024). Smart Tourism Recommender System Modeling Based on Hybrid Technique and Content Boosted Collaborative Filtering. *IEEE Access*. doi: 10.1109/ACCESS.2024.3450882.

[4] Liu, H., Zheng, C., Li, D., Shen, X., Lin, K., Wang, J., ... & Xiong, N. N. (2021). EDMF: Efficient deep matrix factorization with review feature learning for industrial recommender system. IEEE Transactions on Industrial Informatics, 18(7), 4361-4371.doi: 10.1109/TII.2021.3128240

[5] Santoso, A. J., & Soares, J. D. C. L. (2017, September). M-guide: hybrid recommender system tourism in east-timor. In 2017 International Conference on Soft Computing, Intelligent System and Information Technology (ICSIIT) (pp. 303-309). IEEE. **doi:** 10.1109/ICSIIT.2017.16

[6] Kbaier, M. E. B. H., Masri, H., & Krichen, S. (2017, October). A personalized hybrid tourism recommender system. In 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA) (pp. 244-250). Ieee. DOI: 10.1109/AICCSA.2017.12

[7] Al-Ghobari, M., Muneer, A., & Fati, S. M. (2021). Location-Aware Personalized Traveler Recommender System (LAPTA) Using Collaborative Filtering KNN. Computers, Materials & Continua, 69(2). DOI:10.32604/cmc.2021.016348

[8] Zhang, Q. (2022). Personalized hybrid recommendation for tourist users based on matrix cluster apriori mining algorithm. Mathematical Problems in Engineering, 2022(1), 8299761. https://doi.org/10.1155/2022/8299761

[9] Fahrizal, D., Kustija, J., & Akbar, M. A. H. (2024). Development tourism destination recommendation systems using collaborative and content-based filtering optimized with neural networks. *Indonesian Journal of Artificial Intelligence and Data Mining*, *7*(2), 285-298.

Author Biographies

Dr. Thirupathi Regula is a faculty member in the Department of Information Technology at the University of Technology and Applied Sciences (UTAS), Muscat. A distinguished academic with a PhD, he possesses extensive expertise in research and teaching. Currently, he is pursuing a Post-Doctorate in Artificial Intelligence and Machine Learning at the Singapore Institute of Technology, Singapore. He has published extensively in journals, conferences, and book chapters, earning accolades like the "Outstanding Scientist Award" and "Best Researcher Award" from IJIEMR-ELSEVIER and a "Quarterly Franklin Membership" from the London Journals Press, UK.

Dr. Regula serves as an active reviewer for prestigious research journals, TRC-Oman projects, and international conferences. As a member of esteemed professional organizations including IEEE, ACM, and AAAI, his research interests span Data Analytics, Data Mining, Artificial Intelligence, GeoFencing, and IoT. He has actively participated in and organized numerous national and international conferences and teaches courses such as Big Data, Big Data Analytics, and database programming.

Anshar Ali is a faculty member in the Department of Information Technology at the University of Technology and Applied Sciences (UTAS), Muscat. He holds a Master of Engineering in Computer Science and Engineering and is currently pursuing a PhD specializing in Deep Learning. His research focuses on applying Deep Learning techniques to image classification and recommendation systems.

As a member of esteemed professional organizations, including IEEE and many others, he remains actively engaged in advancing his field. At UTAS, he teaches a variety of courses, including programming, software design, and software testing.

Dr. Dhanasekar received his Ph.D. in February 2016 from Manonmaniam Sundaranar University, Tamil Nadu, India. He completed his M.Sc. in Mathematics in 2002 at the University of Madras, Chennai, Tamil Nadu. Presently, he serves as a faculty member in the Mathematics Section of the IT Department at the University of Technology and Applied Sciences, Muscat, Sultanate of Oman.