Enhancing Tourism Growth Through a Hybrid Recommendation System Utilizing Web Scraping and Artificial Intelligence Techniques

[1]Dr. Thirupathi Regula, https://orcid.org/0000-0002-2939-3552
[2]Anshar Ali, https://orcid.org/0009-0006-5439-2996
[3]Dr. Dhanasekar, https://orcid.org/0000- 0002-6347-337X
[1,2,3]*Department of Information Technology, College of Computing and Information Sciences, University of Technology and Applied Sciences, Muscat, Oman.*
Corresponding Author: *Dr. Thirupathi Regula (email: Thirupathi.regula@utas.edu.om)

***Abstract*— Tourism is one of the emerging areas for economic development. Tourists from all over the world are more interested in exploring the landscapes and tourist places, but they are facing so many difficulties extracting the complete information about the tourist places, hotels, flight booking and local transportation. The complexities are increased more when they are looking for a convenient schedule and budget. This research proposes a hybrid approach that uses web scraping and Artificial Intelligence (AI) cutting edge techniques to get the complete information for the tourists very easily at their convenience. The web scraping technique will extract the data from the websites of hotels, local transport, flight booking, tourist places and weather for all kinds of bookings for the tourists. This data will be given to the AI model to train the system and to get the recommendation of nearby other tourist places, hotels and transportation based on their preferences such as days and cost as a package. With this information, the tourist will enjoy the beauty of nature and explore the culture of a country. This hybrid approach will satisfy the expectation of tourist through the recommendation system and improve the growth of the tourism development of the country.**

***Index Terms*— *Web Scraping, Recommendation system, Tourism* growth*, Artificial Intelligence, economic growth*.**

I. INTRODUCTION

Every country offers enjoyable and varied tourism. It boasts numerous World Heritage Sites in addition to a variety of fascinating tourist destinations, including markets, beaches, hills, and historic homes. Web scraping techniques and an AI-based recommendation system is used in an integrated strategy to improve the discovery of rich destinations, enabling visitors to experience the country at their level of expertise.

Using the traditional method, which can be time-consuming and stressful for tourist as they are going to check the booking and place details individually. By bringing in many visitors from both inside and outside of the country, this integrated approach enhances the country's tourism industry and ensures that visitors have a seamless, pleasurable, and unforgettable trip.

Objectives:

This research paper aims to create an AI based system that:

- To help tourists to extract the details of the places and all booking information easily.
- To give real time recommendations of nearby places, hotels, flights and transportation for their suitable dates and budget.
- To explore the beautiful places, tradition, and culture of the country
- To get the satisfaction of tourists and encourage them to suggest the country’s tourist places with their friends and relatives.
- To encourage the tourism in the country.
- To develop the economy of the country.

II. LITERATURE REVIEW:

Web scraping is a technique used to extract data from websites. It involves writing code (usually in Python, but other programming languages can be used as well) to programmatically access a webpage, retrieve its HTML content, and then parse that content to extract the desired information.

Web scraping can be done manually, but it's often automated using libraries like Beautiful Soup or Scrapy in Python. These libraries make it easier to navigate through

HTML documents and extract specific elements such as text, links, images, or any other structured data.

AI techniques are often integrated into web scraping processes to enhance their effectiveness and efficiency. Some of the AI technologies commonly used in web scraping includes Natural Language Processing (NLP), Machine Learning (ML), Natural Language Generation (NLG), Deep Learning and Reinforcement Learning. This paper focuses on three algorithms such as K-nearest neighbor (KNN), Matrix Factorization and Bayesian Network.

To solve the issue of identifying trends in different web scrapers and eliminating them altogether, Kaushal et al (2018). looked at the detection of web scraping using machine learning. He developed a program that captures such attacker signatures and prevents such attacks in real time, which display such attacks in a graphical perspective for customers to quickly identify them.

One method that attempts to solve this problem is web scraping. Unstructured online data can be converted into structured data via web scraping so that it can be kept and examined in a central local database or spreadsheet. The goal of SCM de S Sri Suriya (2015) was to provide a comparative explanation of well-known online scraping technologies and approaches. Using online tools and methods for online data extraction from educational websites, he compared and evaluated several web scraping approaches and well-known web scraping applications.

Vidhi Singrodia and Anirban Mitra investigated in 2019 on the review of web scraping and its applications. The investigation concentrated on numerous aspects of web scraping, including its fundamentals and the use of software tools for online scraping. They looked over the web scraping system's working principles, advantages, and disadvantages before seeing its applications.

Using the keyword "wonderful Indonesia" as the identification of the tourist campaign, (2024) Eka Purnama Harahap1 etl.la., examined Twitter users' opinions about travel to Indonesia. They have identified the tweets' positive, neutral, or negative sentiment by applying the K-nearest neighbor (KNN) algorithm. Furthermore, 98.5% accuracy, 97.6% precision, 98.5% recall, and 98.1% F1-score were attained in the training outcomes they examined.

Firman Gazali Mahmud, Teguh Iman Hermanto, and Imam Maruf Nugroho (2024) investigated about the making a right decision for the tourists who visit to Indonesia during their seasonal time. Using the K-nearest neighbor algorithm with SMOTE, they have applied sentiment analysis to hotel evaluations. The accuracy results are displayed, and the suitable AUC value serves as a validation of the good classification category.

According to Reham Alabduljabbar's (2023) research, capturing the complex relationships between clients and restaurants can be effectively achieved using matrix factorization algorithms in a collaborative method. The evaluation's findings demonstrated the efficacy of both SVD and NMF in producing recommendations, with SVD doing marginally better in terms of RMSE and NMF doing marginally better in terms of MAE.

A case study was investigated by Galya Stateva and Marek Cierpiał-Wolan (2022) by exploring the experience with new web data sources. They have explored the new data sources and monitor the real estate market and derived early estimates of constructing activities, pertaining to both already built and planned buildings, based on real estate web portals. Furthermore, developed new indices for tourism statistics, using the data from booking portals, air traffic portals, travel agencies portals and portals related to quality of life. Also concentrated on mass web scraping, primarily for the enhancement of the quality of the business register via linking URLs of enterprises and predicting main economic activity codes (NACE) etc,

A model for semantic search and its result, which is based on users' priorities when searching the tourism domain of interest, has been examined by K. Palaniammal and Dr. S. Vijayalakshmi. This model considers the fact that current research cannot support reliable matchmaking between the information provided by the web search field and the requests made by tourists for services. They have worked on three main projects to improve search performance: semantic query, generating probability with a Bayesian network and Netica-J,

and ontology knowledge base. Their paper's findings were more precise, providing comprehensive information for the chosen topic on a single page.

A weighted average of daily and monthly lodging charges as well as room occupancy rates have been studied by Yustiar Adhinugroho et al. in 2020. Online occupancy figures were found to follow the same pattern as official statistics after being compared with data from Indonesia. The outcome of this study appears to meet the needs of both designing an online accommodation data collection system from a Web travel source and implementing the Web scraper algorithm. Based on the findings, it is possible to use big data to supplement or replace official statistics. The indication, which typically takes more time and money, can be completed more quickly and affordably by using the Web scraping technique.

Using machine learning techniques, the research created a recommendation system for tourist attractions for visitors. The K-Nearest Neighbor (KNN) method was the machine learning methodology employed. Numerous experiments were carried out, the preparation stage was previously completed based on the dataset in order to enhance the data format. This involved choosing data that had been separated based on preexisting criteria, calculating the closest distance, and figuring out the value of k in the KNN method. The recommendation results obtained by using the KNN algorithm have also been accurate and have adhered to the rating value provided by the tourists visiting the attraction. The anticipated value of visitor ratings for tourist attractions is generated with a high accuracy value of about 78% with k=1 using the system accuracy calculation formula, which was explored by Devie Rosa Anamisa, Achmad Jauhari, and Fifin Ayu Mufarroha. (2023).

T. Praveen of etl.la (2023) has looked at Today, it's essential to be able to spot phony reviews. They described a new technique, which has supervised machine learning is used to identify false reviews. A technique that tells users whether product reviews are reliable, helping them differentiate between genuine and fraudulent ones. The use of supervised machine learning is explained by this method for identifying fraudulent reviews. This methodology was created in response to flaws in the way that categorical datasets or emotion polarity ratings were used in classic fake review detection systems to classify reviews as real or false by showing the accuracy of 88% with machine learning techniques.

Using historical data, Xu Cheng and Chenyuan Zhao (2019) have studied the navel model in place of more conventional methods for analyzing the elements influencing tourist consumption to forecast the pattern of consumption in the near future. In addition to employing a neural network comparison analysis that outperformed the NN model on the same data set, they have employed the Bayesian Network model to forecast the price of airline tickets with an accuracy of over 80%.

Research has been conducted by Roxana Norouzi Isfahani et al. (2023) to assist tourism managers, policymakers, and associated organizations in determining the best course of action for achieving systemic continuous improvement. To create the best possible tourist system, they have put forth an improved Nature-based tourism (NBT) model. Additionally, they have used the Bayesian network (BN) to quantify the effects of each subsector on NBT. Based on how much each group affected the NBT status, they separated the three groups spread through three tiers.

Today, hundreds of online scraping programs are accessible, the majority of which were created with Ruby, Python, and Java. Along with commercial software, there are also some open-source web scraping programs available. The greatest resources for novices in web scraping are web scraping programs like Yahoo Pipes, Google Web Scrapers, and Outwit Firefox plugins.

III. METHODOLOGY

The automatic web crawling/ web scrapping can be implemented by automated web crawling tools like Parse Hub / Scrapy/ Selenium. These tools will extract the from various websites of the flight booking, hotel booking, taxi booking and some tourist places websites. The collected data will be stored

in CSV and Json format. To remove the noisy / unwanted data the collected data will be preprocessed. The figure 1 describes the hybrid model for AI based recommendation system.

In this preprocessing, the system will perform the Data Cleaning, Data Transformation, Data Reduction, Data Integration, Improving Model Accuracy and Enhancing Interpretability. With the help of these the unnecessary data will be removed, and some data will be enhanced. Then the system will apply subset selection or feature selection to focus on important data such as budget, number of days, family /individual trip and many others. These will apply on the AI based algorithms to train the system. The data will be trained by user-based K nearest neighbors' algorithms, matrix factorization algorithms and Bayesian network algorithm.

The K nearest neighbor's algorithm is a collaborative filtering technique which recommends User-based k-NN recommends items to a user based on the preferences of similar users. It identifies users who are most like the target user and suggests items they liked.

One effective method in recommendation systems, especially for collaborative filtering, is matrix factorization. To capture latent variables that explain observed interactions, it breaks down a huge user-item interaction matrix into lower-dimensional matrices (like ratings or preferences).

The matrix factorization

$$R \approx U \cdot V^T \quad (1)$$

Where R is the user-item matrix

U (user features) and V (item features)

Here, U is of size m×k (where m is the number of users and k is the number of latent features)

and V is of size n×k (where n is the number of items).

$$L = \sum_{(u,i)\in K} (R_{ui} - (U_u \cdot V_i^T))^2 + \lambda(||U||^2 + ||V||^2) \quad (2)$$

K is the set of observed ratings, and λ is a regularization parameter to prevent overfitting.

The predicted rating can be calculated by

$$\hat{R}_{ui} = U_u \cdot V_i^T \quad (3)$$

Bayesian networks can be effectively used for recommendation systems by capturing the probabilistic relationships between users, items, and their attributes. We can use either graphical representation or probabilistic relationships in the Bayesian networks.

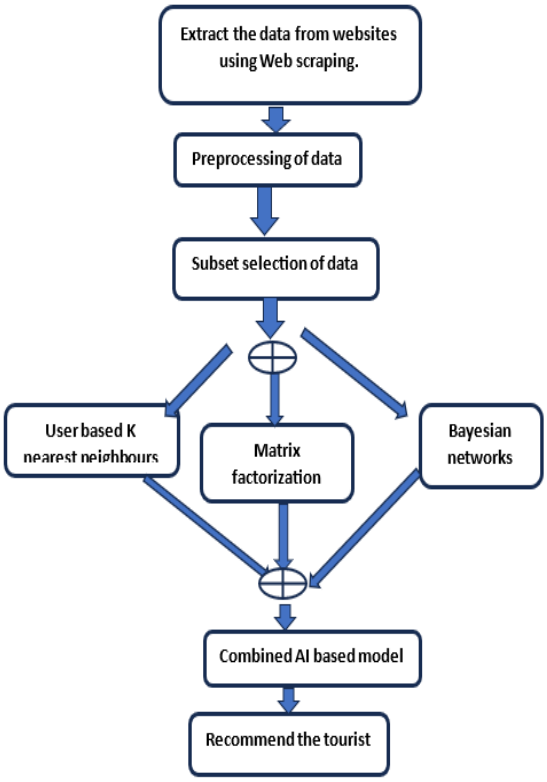

Figure 1: Hybrid model for web scraping and AI based recommendations.

The results of these three algorithms will be combined to build the proposed model.

The model will recommend the best booking details for flight, hotel, transport, weather and tourist places if necessary. Based on this the tourist can decide and plan their travel. This system will recommend the nearby tourist places and hotels and transport within their budget. The system allows the tourists to review and suggest the places and other details to their friends and relatives.

IV. CONCLUSION

Using web scraping, this AI-powered recommendation system will gather information about tourist destinations, airline, hotel, and taxi bookings. The algorithm

will then use these data to suggest specifics depending on user criteria, including spending limit, number of days, preferred locations, etc. The algorithm will suggest the best specifics to ensure a straightforward and easy travel experience. As a result, it will motivate more and more travelers to visit the various locations, boosting both travel and the national economy.

The system will consider some cutting-edge elements in the future, such as voice--based search and image-based search recommendation systems, which will help travelers by making chores easier.

REFERENCES

[1] Kaushal et al., (2018), "Detection of Web Scrapping Using Machine Learning", *Open Access International Journal of Science & Engineering*, Vol. 3.

[2] SCM de S Srirsuriya (2015), "A Comparative Study on Web Scraping", *Proceedings of 8th International Research Conference, KDU, Published November 2015.*

[3] Vidhi Singrodia and Anirban Mitra (2019), "A Review on Web Scrapping and its Applications", *International Conference on Computer Communication and Informatics (ICCCI -2019), Jan. 23 – 25.*

[4] Eka Purnama Harahap1 etl.la (2024), "Trends in sentiment of Twitter users towards Indonesian tourism: analysis with the k-nearest neighbor method", *Computer Science and Information Technologies, Vol. 5, pp. 13~22*.

[5] Firman Gazali Mahmud, Teguh Iman Hermanto, and Imam Maruf Nugroho (2024), "Implementation of K-nearest neighbor algorithm with smote for hotel reviews sentiment analysis", *Computer Science and Information Technologies, Vol. 5, No. 1, pp. 13~22*.

[6] Reham Alabduljabbar(2023)," Matrix Factorization Collaborative- Base d Recommender System for Riyadh Restaurants: Leveraging Machine Learning to Enhance Consumer Choice" , *Applied Science journal, 2023,13,9574.*

[7] Galya Stateva and Marek Cierpiał-Wolan (2022), "Exploration and experience with new web data sources. A Case Study for innovative tourism statistics", *4th International Conference on Advanced Research Methods and Analytics (CARMA2022).*

[8] K.Palaniammal, Dr. S. Vijayalakshmi (2014), "Improving Search Performance for Toursim Domain Using Semantic Web and Bayesian Network", *International Information Institute (Tokyo). Information; Koganei Vol. 17, Iss. 8, 3675-3682.*

[9] Yustiar Adhinugroho etl.la., (2020), "Development of online travel Web scraping for tourism statistics in Indonesia", *Information Research, Vol. 25, No.4, paper 885* . https://doi.org/10.47989/irpaper885.

[10] Devie Rosa Anamisa, Achmad Jauhari, and Fifin Ayu Mufarroha. (2023), "K-Nearest Neighbors Method for Recommendation system in Bangkalan's Tourism", *Computer, Mathematics and EngineeringApplications, Vol.14(1), 33-44*. https:// DOI: 10.21512/comtech. v14i1.7993

[11] T. Praveen etl.la., (2023), "A supervised Machine Learning Approach using K-Nearest Neighbor Algorithm to Detect Fake Reviews on Amazon ", International Research Journal Of Engineering and Technology, *Vol.10(2), 2395-0072* https:// DOI: https://www.irjet.net/archives/V10/i2/IRJET-V10I2123.pdf.

[12] Xu Cheng and Chenyuan Zhao (2019), "Prediction of Tourist Consumption Based on Bayesian Network and Big Data", *International Information and Engineering Technology Association, Vol.24(5), 491-496*, DOI: https://doi.org/10.18280/isi.240505

[13] Roxana Norouzi Isfahani, Ahmad Talaee Malmiri, Ahmad Bahoo Toroody and Mohammad Mahdi Abei (2023), "A Bayesian-based framework for advanced nature-based tourism model", *Journal of Asian Business and Economic Studies, Vol.30(2), 86-104*, DOI: https://doi.org/10.1108/JABES-11-2020-0119 l