Web Mining: Tour Recommendation System

Rith Chanthyda[1], Phun Sreypich[2], Phai Ratha[3], Kry Senghort[4], Meng Heabvathanak[5] and Rithy Vira[6]

[1, 2, 3, 4, 5, 6] *Institute of Technology of Cambodia, Department of Applied Mathematics and Statistics, Cambodia*

Reception date of the manuscript: 06/05/2023

Abstract— The tourism industry has seen significant growth in recent years, with an increasing number of people traveling for leisure and business purposes. However, finding suitable travel options can be overwhelming for many tourists due to the abundance of information available on the web. To address this challenge, we propose a tour recommendation system that utilizes web mining techniques to generate personalized tour itineraries. The system incorporates data preprocessing, web scraping, and data analysis to evaluate user preferences to provide customized tour recommendations. Our research presents a comprehensive framework for web mining and recommends tour generation to personalize travel experience. We evaluated our system using a real-world dataset from a popular travel website and achieved promising results. The system generated personalized tour recommendations that improved the user's travel experience by recommending tours that aligned with their interests. Our research contributes to the development of intelligent tourism systems that provide personalized recommendations to tourists.

Keywords— Recommendation Algorithms , Data science, Data mining, Machine learning, Tourism

I. Introduction

The internet is a vast source of information, including travel-related data that can be used to enhance the tourism experience. However, this data is often unstructured and difficult to access, making it challenging to extract valuable insights. Web scraping, the process of automatically extracting data from websites, has emerged as a crucial technique for collecting and analyzing travel-related data. This technique is particularly useful for the tourism industry, as it can be used to generate personalized tour recommendations based on historical data and user preferences.

With web scraping, data engineers can collect data using a personal computer connected to a hotspot or WiFi network. The process involves identifying the relevant websites, selecting the data to be extracted, and developing a script to automate the extraction process. Web scraping provides a means of collecting data in real-time, allowing users to obtain up-to-date information that can be used for analysis and decision-making[1]. And it's really important for data science as it is:

- Less time consuming
- Cost effectiveness
- Accuracy result and high quality
- Unable to access a wide range of data
- Real-time data
- Customizable and greater insights

The purpose of this manuscript is to contribute to the development of intelligent tourism systems that provide personalized recommendations to tourists. The proposed tour recommendation system incorporates web scraping techniques to collect real-time data from relevant websites, which can be analyzed using EDA and data analysis techniques to generate personalized tour recommendations. This approach provides tourists with valuable insights into the most popular tourist destinations, accommodations, and activities, among other factors, thereby enhancing the overall tourism experience.

II. History of Web Scraping

The brief history of web scrapping data has five generations, involving:

- **First web browser**: was written in 1990. Tim Berners-Lee created the first web server and graphical web browser in 1990 while working at CERN, the European Organization for Nuclear Research, in Switzerland.
- **The Wanderer and Wandex**: In 1993, the World Wide Web Wanderer developed by Matthew Gray, Perl-based web crawler whose sole purpose was to measure out the size of the web. It was used to generate an index called the Wandex. Wanderer with Wandex had the potential to become the first general-purpose World Wide Web search engine[2].
- **BeautifulSoup**: In 2004, BeautifulSoup - HTML parser, a library of commonly used algorithms written in Python programming language. It helps to grasp the sense of site structure and parse the contents within the HTML containers. We are easily searchable. Sometime websites did not prohibit the ability to download the content of their sites. However, slowly that changed, and for the amount of data that was getting downloaded - simply manually copy-pasting was not an option; therefore, other ways of obtaining the information

were bound to be developed[2].

- **Visual web scrapping**: Visual scrapers are tools in which the user can visually select the elements to extract, and the logical order to performing a sequence of extractions. They require little or no code, and assist in designing XPath or CSS selectors. Its tools vary in how flexible they are, how easy to use, to what extent they help you identify and debug scraping problems, how easy it is to keep and transfer your scraper to another service, and how costly the service is[2].

Web scraping has grown immensely in recent years, and almost guaranteed to continue upward growth. Currently, the commercial web scraping scene is mostly for gaining a competitive advantage by collecting leads, scraping competitors, price monitoring, etc. However, as technology develops, such as Artificial Intelligence, and data becomes even more accessible and crucial to different aspects of life, web scraping will advance with it and produce various new and remarkable applications that we are only looking forward to experimenting with.[3]

There are three types of web scraping:

- **Report mining**: Programs pull data from websites into user-generated reports. It's a bit like printing a page, but the printer is the user's report.
- **Screen scraping**: The tool pulls information on legacy machines into modern versions.
- **Web scraping**: Tools pull data from websites into reports users can customize.

III. METHODOLOGY

The process in scraping data for tour recommendation system is really challenging as the methodology for data collection would involve identifying relevant travel websites, choosing a web scraping tool and studies in web scraping. Each steps involve different technique in order to obtained the data that we are needed.

1. **Travel Website**:the most important part of data collection is to choose a good travel website which is qualified to the following condition:

   - Popularity: the website need to be popular locally and globally.
   - Consistency of data: the information in website need to be regularly update so our recommendation system have an accurate result.
   - Website structure: websites with a clear and organized structure are generally easier to scrape. If the website has a well-defined HTML structure with consistent class names and IDs, it will be easier to extract the data that is needed.

   In our case, we choose Tripadviser websites and Viator websites as the source of our data collection.

2. **Web-Scraping Tool**: when it comes to web scraping, choosing the right tool for the job can make all the difference and decided the outcome of the scraping. We first studies on how each tool effect and benefit to our scraping technique, hence, we decided to choose these Scraping Tool:

   - **Selenium**: with Selenium, we can automate the process of loading the page, interacting with the various elements on the page (such as filling out forms or clicking buttons), and even scrolling through the page to load additional content. This makes it a powerful tool for scraping dynamic websites that might be difficult to scrape with other tools.
   - **BeautifulSoup**: since we are dealing with a static website and just need to extract information from the HTML, BeautifulSoup is a simpler and more lightweight option. It provides a simple and intuitive interface for parsing HTML documents and extracting the data you need. We can use it to extract specific tags or attributes from the HTML, or even to search for text patterns using regular expressions.

3. **Studies in web scraping**: The most studied issue for web scraping is to determine the extraction patterns automatically. On the other hand, these patterns are created manually by developers/experts in the implementation of web scrapping tools that are used HTML parsers. Hence, studies the these can be time consuming in order to obtained the data that is needed. For example, the following opening tags can be employed to eliminate the unnecessary content and to extract necessary content such as:

   • : All necessary content
   • <h1 itemprop="headline">: Title of the web page
   • <h2 itemprop="Summary">: Summary of the web page
   • : Main text of the web page
   • <p>: Inner texts of the web page

   A web developer can prepare an extraction pattern by using the tag of an element and specifying desired attributes. Moreover, this pattern is used to produce different web pages for a website. Therefore, when required patterns are resolved for a website, the extraction process can be easily applied by using resolved elements [4].

4. **Programming**: In the programming step of web scraping, the goal is to apply the extraction patterns identified in the previous step to the actual code. This is where knowledge of programming languages such as Python, Java, or R come into play. And in our case, we choose Python programming language to performing the scraping. Once the extraction patterns have been identified, we write code to parse the HTML or XML code of the website by using **BeautifulSoup** and **requests** library and extract the desired information.

   During this step, it is important to be careful with the code to prevent errors while scraping. A common technique that we used is to include **error exception handling** code in case the scraper encounters unexpected issues, such as a broken link or missing data. We also

recommended to test the scraper on a small amount of data before applying it to the entire website to ensure it is working properly.

Furthermore, programming in web scraping can involve not only extracting data but also processing it. In our case, a web scraper is collecting data from multiple pages of a website and the collected data will need to be merged and formatted into a usable format. Hence, we have required additional programming knowledge and tools such as such as **Pandas, NumPy,** and **BeautifulSoup** in Python, which allow for data cleaning, manipulation, and analysis. These libraries can help with merging data from multiple web pages, converting data types, filtering data, and much more. .

IV. Data Collection

In order to build a tour recommendation system using web scraping, we can collect data from various sources such as Tripadviser websites and Viator websites. Type of data collection include I texts, links, html code and any other information on the site which is hidden. These variables that scraped includes features such as:

- **Tour Name**(text): defined as the name of each tour
- **Tour Type**(text): defined as the category of the tours
- **Number of Reviewer**(float): defined as the number of people that rate the tours
- **Rating**(float): defined as the rating of each tour which scale from 1 to 5
- **Duration**(text): duration of the each tour which scale from minute to hour
- **Price**(float): defined as a total cost per group or per individual
- **Highlight**(text): defined the as the description plus activity that include in the tours
- **location**(text): defined as the location that the tour take place

However, there are several limitations to consider when collecting data through web scraping. Firstly, some of the websites have anti-scraping measures in place that make it difficult to access data. Additionally, data quality may vary across different websites, and there may be missing or inaccurate information in the collected data. Furthermore, there may be legal and ethical considerations to keep in mind when scraping data from third-party websites.

To mitigate these limitations, it is important to carefully select the sources of data, and to ensure that data is cleaned and pre-processed before being used in the recommendation system. This can involve removing duplicates, handling missing data, and ensuring that data is in a consistent format which will be explained in detail in the next section. By taking these steps, we can create a high-quality dataset that can be used to build an effective tour recommendation system that provides users with personalized and relevant tour recommendations.

V. Data Preprocessing

In our tour recommendation system, the data collected through web scraping went through several preprocessing steps to clean and transform it before analyzing it.

a. Data Cleaning

Data cleaning was done to remove any duplicates and handle missing data. In our case, a tour package had missing values for certain features such as tour price, rating, number of reviewer and tour highlight.To deal with the missing value, we imputed the missing values with **KNN Imputation** known as a technique used to impute missing values in a dataset by finding the k nearest neighbors to each missing value where k is a specified number of neighbors. The missing value is then imputed using the average value of its k nearest neighbors. In other words, for each missing value, the algorithm looks for the k nearest data points based on the available features in the dataset. It then takes the average value of the feature for those k nearest neighbors and uses it to fill in the missing value.

b. Data transformation

Data transformation was performed to ensure consistency in data format. This included converting textual data such as descriptions or location names into numerical data through techniques such as **one-hot encoding** or **term frequency-inverse document frequency (TF-IDF) vectorization**. Additionally, we created new features such as tour duration by calculating the difference between the start and end dates of the tour package.

c. Data normalization

Data normalization was also done to ensure that all features were on the same scale. This involved scaling numerical features such as price and tour duration to a range of 0 to 1 using min-max scaling.

d. Data Reduction

Data reduction techniques such as principal component analysis (PCA) were applied to reduce the dimensionality of the dataset while retaining most of the variance in the data. This helped to improve computational efficiency and reduce the risk of overfitting.

By implementing these preprocessing steps, we were able to transform the raw data into a clean, normalized, and structured format that was suitable for analysis and modeling in our tour recommendation system.

VI. Data Analysis

In this analysis, we have examined a dataset on tours with the aim of building a tour recommendation system utilizing content-based filtering. The analysis involved utilizing various techniques and tools to gain a deeper understanding of the data, such as exploratory data analysis, visualization, and statistical analysis. We identified the data types of each column, checked for outliers and anomalies, and examined correlations and patterns between the columns. Furthermore, we identified the most popular tours based on the number of reviewers and analyzed common themes or tags in the "Highlight" column. This analysis provided us with valuable insights into the dataset, which will aid in the development of an effective tour recommendation system utilizing content-based filtering.

a. Rating Analysis

Rating analysis is a critical aspect of tour recommendation systems. By analyzing ratings given by users, we can determine the overall quality of the tours and identify any potential issues with low-quality tours. Based on the histogram plot in Fig. 1), it is evident that the distribution of ratings is right-skewed, indicating that a majority of tours have a higher rating.

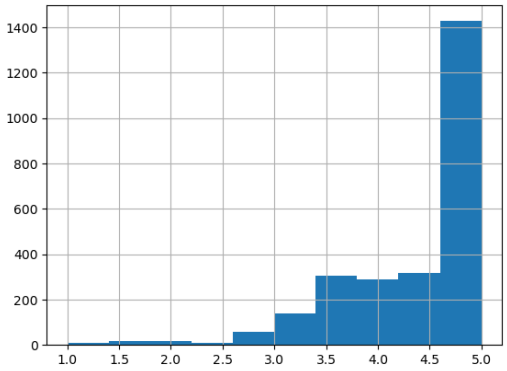

Fig. 1: Histogram of Tour Rating

b. Correlation Analysis

Correlation analysis is an essential tool in constructing an effective tour recommendation system. It helps to identify significant relationships between variables in the dataset, which can inform the recommendation algorithm and improve its accuracy.

The heatmap in Figure 2 presents a visual representation of the correlation between tour price and other features in the dataset. This heatmap allows us to observe to what degree, in which direction, and alerts us to potential issues with our variables.

In particular, the correlation coefficient between tour price and tour duration is 0.59, indicating a moderately strong positive relationship between the two variables. This means that as the tour price increases, so does the tour duration, on average. This correlation is not entirely linear, as shown by the heatmap. However, the moderately strong positive relationship suggests that tour duration can be used to predict tour price to some extent.

On the other hand, the correlation coefficients for other features, such as the number of reviewer and rating are relatively weak, indicating a weak correlation with tour price. While these features may still be useful in the tour recommendation system, they may not be as informative as tour duration in predicting tour prices.

It is worth noting that correlation does not necessarily imply causation. Although a correlation between two variables may exist, it does not necessarily mean that one variable causes the other. Therefore, it is important to exercise caution when interpreting correlation coefficients and to consider other factors that may affect the relationship between variables.

Overall, the heatmap in Figure 2 provides valuable insights into the relationships between tour price and other features in the dataset. These insights can inform the development of a more accurate tour recommendation system, which can ultimately improve the customer experience and satisfaction.

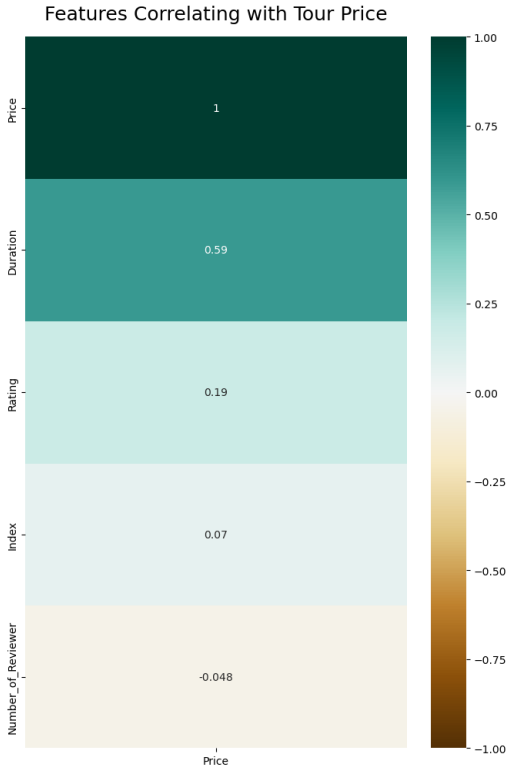

Fig. 2: Scatter Plot of Tour Price and Tour Duration to observe their correlation

c. Distribution

By analyzing the distribution of tour features such tour category, we can gain insights into the overall behavior in the data.

	Tour_Type	Counts	Percent
0	Bus Tours	911	35.19
1	Full-day Tours	736	28.43
2	Adventure Tours	266	10.27
3	Half-day Tours	51	1.97
4	Multi-day Tours	50	1.93
...	...	...	...
63	Mixology Classes	1	0.04
64	Family-friendly Shows	1	0.04
65	Motorcycle Tours	1	0.04
66	Circuit Tours	1	0.04
67	Sightseeing Tours	1	0.04

TABLE 1: TABLE OF THE TOUR CATEGORY DISTRIBUTION

Our dataset consists of 68 unique Tour Type values. Among all the types, Bus Tours have the highest number of packages, with a total of 911 packages, which covers 35.19% of the entire package portfolio. Full-day Tours are

the second-best package type, with 736 products, covering 28.43% of the total package portfolio. In contrast, some Tour Types, including Sightseeing Tours, have only one package, which covers 0.04% of the total package portfolio. Understanding the distribution of tour packages among different types can provide valuable insights into the popularity of tour types and their demand among customers.

d. Popularity Analysis

By understanding which tours are popular among customers, we can recommend similar tours to other customers and improve the overall satisfaction of our users. Popularity analysis helps us identify which tours are in high demand and which factors contribute to their popularity. This information can inform our recommendation algorithm and help us tailor recommendations to individual customers' preferences.

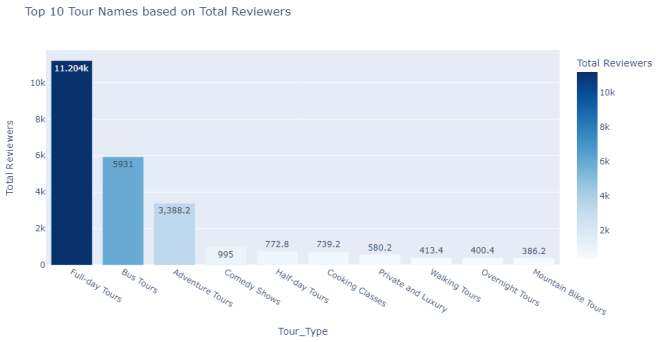

Fig. 3: Bar Chart of the Top 10 Tour Type Base on Number of Reviewer

From the Bar Chart in Fig. 3,The tour category "Full-day Tours" has received 11.204K reviews, representing 39.62% of the total reviews. Among the top 10 tours, "Bus Tours" has the highest number of reviews, with 5931 reviews, accounting for 20.97% of the total reviews. On the other hand, "Mountain Bike Tours" has the lowest number of reviews among the top 10 tours. These findings suggest that these tours are popular among users and may indicate their high satisfaction with these tour categories. It is reasonable to assume that these popular tours are the most desirable and satisfying packages, making them more likely to be recommended to other customers.

VII. Conclusions

In this article, through web-scraing and data analysis , we can highlights the importance of popularity analysis and data distribution for developing effective tour recommendation systems. We observed a positive correlation between tour price and duration, indicating a preference for longer tours at higher prices. Additionally, our analysis identified **Bus Tours** and **Full-day Tours** as the most popular tour types, covering a substantial proportion of the total package portfolio.

Companies like Amazon and Netflix can drive a lot of revenue using a recommendation system. They can engage and retain existing customers by recommending relevant content to them. These results we have obtained have important implications for future research and the travel industry as a whole. Further research could explore the different types of recommendation systems. We then developed a content-based recommendation system as here are several metrics we can use to identify similarities between products. We used cosine similarity for our recommendation system.

VIII. Acknowledgment

The author would like to express their deepest appreciation to Professor Chan Sophal, for his guidance and support throughout this research project. His insightful comments and feedback have been invaluable in shaping this paper.

References

[1] Jaiswal, N. L., Gupta, and S. K., *Web scraping: A review of techniques, tools and related ethics.* , 2017, vol. 8, no. 5.

[2] Liu, B. Li, and X., "Web Data Extraction: A Review. In Proceedings of the International Conference on Artificial Intelligence and Big Data," pp. 107–114, 2021.

[3] Shealy and M., "The Future of Web Scraping in the AI Age," *The Astrophysical Journal*, 2021.

[4] Ghosh, S. Mihailidis, and P., *Web scraping: An introduction to data mining on the internet*, 2018, vol. 8, no. 1.