Personalized Travel Recommender using Collaborative and Content-based Filtering Techniques

Rawan Osama
Dept. of Software Engineering
Faculty of Computer and Information Sciences,
Ain Shams University
Cairo, Egypt
rwan20191702026@cis.asu.eu.eg

Nancy Ahmed
Dept. of Software Engineering
Faculty of Computer and Information Sciences,
Ain Shams University
Cairo, Egypt
nancy20191702075@cis.asu.edu.eg

Salma Magdy
Dept. of Software Engineering
Faculty of Computer and Information Sciences,
Ain Shams University
Cairo, Egypt
salma20191702032@cis.asu.edu.eg

Ahmed Sobhy
Dept. of Software Engineering
Faculty of Computer and Information Sciences,
Ain Shams University
Cairo, Egypt
ahmed20191702003@cis.asu.edu.eg

Mina Maged
Dept. of Software Engineering
Faculty of Computer and Information Sciences,
Ain Shams University
Cairo, Egypt
mina20191702074@cis.asu.edu.eg

Salah Ashraf
Dept. of Software Engineering
Faculty of Computer and Information Sciences,
Ain Shams University
Cairo, Egypt
salah20191702039@cis.asu.edu.eg

***Abstract*—This paper aims to discuss building a personalized travel application that provides recommendations to users based on their interests, with a focus on Egypt. The application uses two popular Collaborative Filtering techniques; Matrix Factorization Alternating Least Squares (MF-ALS) and Restricted Boltzmann Machine (RBM), as well as Content Based Filtering using Cosine Similarity to generate relevant recommendations for hotels, attractions, and restaurants. Places data was collected by scraping TripAdvisor and Google Maps. This paper also includes an itinerary feature that allows users to create a personalized travel plan based on the trip's number of days, average start and end times, city, starting location and preferences. The Collaborative Filtering models were evaluated using the Root Mean Squared Error (RMSE) metric, and MF-ALS was found to perform better than RBM. Overall, this work aims to enhance the travel experience for users by providing tailored recommendations that match their preferences in Egypt.**

***Keywords*—** Collaborative Filtering, RBM, MF-ALS, Content-Based Filtering using Cosine Similarity, User Profiling, Personalization, Itinerary Plan.

I. Introduction

Recently, with more people traveling than ever before, the travel and tourism sector has experienced a considerable increase in demand. It can be overwhelming for travelers to navigate and filter through the vast amount of data to find the most relevant and personalized recommendations with the extensive amount of information available on the internet. The proposed travel application fills in this need by offering individualized recommendations based on the user's preferences. The significance of this research comes from its potential to benefit the travel and tourism industries by helping users make more informed travel decisions by giving them recommendations that are more accurate and relevant to their needs. This paper discusses different recommendation algorithms to analyze user data and produce recommendations that are relevant to the user's interests.

Collaborative filtering is a type of recommendation technique that analyzes behavior and preferences of a group of users similar to the user receiving the recommendation [3]. Cold start is a problem in collaborative filtering that the recommendation system faces as it cannot recommend to a new user or a new item with neither history nor other users in the system. Content based filtering is a type of recommendation technique that suggests items to a user based on their preferences and previous interactions with similar items [4].

This paper explores the challenges faced in the recommendation system. Starting with the unavailable dataset, which was solved by Web Scraping. Another challenge was the cold start problem, which was solved using User profiling and Surveying.

This paper is structured into the following sections: Section II presents the research efforts for implementing different recommendation techniques, while Section III discusses the proposed methods implemented by this paper. Section IV reviews the results of the different methods applied and Section V discusses the conclusion and suggested future work to fill the research gaps.

II. Related Work

One of the most popular recommendation approaches used in research papers is collaborative filtering. Wang, Z. et al. [6] created an android application for recommendation using collaborative filtering, collecting their data by web crawling. The authors highlighted the challenges faced by users in choosing from a large amount of tourist information available on travel portal websites and how tourism recommendation systems can help solve this problem. The paper discussed the different recommendation systems and problems such as cold start, which refers to the difficulty in making recommendations when the users or the items are new. In this paper, data mining technology and collaborative filtering algorithms were applied to the travel recommendation system. Similarities with this paper and our paper are collaborative filtering and web scraping techniques, the difference is that this paper was focused on China's Tourism industry.

978-1-6654-4076-9/21/$31.00 ©2021 IEEE

The aim of Chaitra, D. et al. [7] was to create an itinerary plan for their user. The authors used collaborative based filtering and collected their data using web scraping and Google Forms. The authors highlighted the usage of collaborative based recommendation systems and how the output was used to create an itinerary plan based on these places. The similarities is using web scraping and Google Forms to collect data and itinerary algorithms using recommendation output as an itinerary algorithm input. The difference is it focuses on Karnataka Tourism while this paper focuses on Egypt Tourism . Moreover, the differences between this paper and what Chaitra, D., et al [7] created is the basis of Indian tourism, and not offering a full plan for the user but rather it recommends based on time (morning or night) although separating them into hotel, attraction, and restaurant categories.

Venkateswaran, S. et al. [8] presented the Intelligent Travel Recommender System (ITRS), which was an innovative solution to simplify the process of planning a vacation. The system uses data science tools and techniques to generate a tailor-made travel plan that includes accommodation, attractions, and dining options based on the user's travel details and preferences. The paper discussed the analysis that led to the decisions made by the system, including sentiment analysis on user reviews and user-tip reviews. The algorithms used were hybrid recommendation techniques mixed together: RBM, MF-ALS, K-Means and KNN. Similarly this paper is using RBM and MF-ALS as collaborative models, and user profiling.

III. Methods

The proposed system pipeline as illustrated in Fig.1. shows the components and techniques of the system which are web scraping to collect places data, user preferences collection to perform user profiling. Moreover, RBM is shown to generate attractions recommendations, MF-ALS to generate hotels and restaurants recommendations and cosine similarity to generate similar places recommendations. The output of the previous processes is then entered to the itinerary plan to generate a personalized plan.

Different methods were applied as follows:

A. Data Collection

There were no datasets of places in Egypt that are readily available to be used in the project. Therefore, the available data on TripAdvisor and Google Maps was collected by Web Scraping.

B. Applied Methods to solve Cold Start Problem

A major obstacle for Collaborative Filtering is the cold start problem, which refers to the difficulty in making recommendations when the users or the items are new (new user). Moreover, it occurs when you have a new system with no existing users (new system). Two techniques were used to tackle this problem which are User Profiling and Surveying.

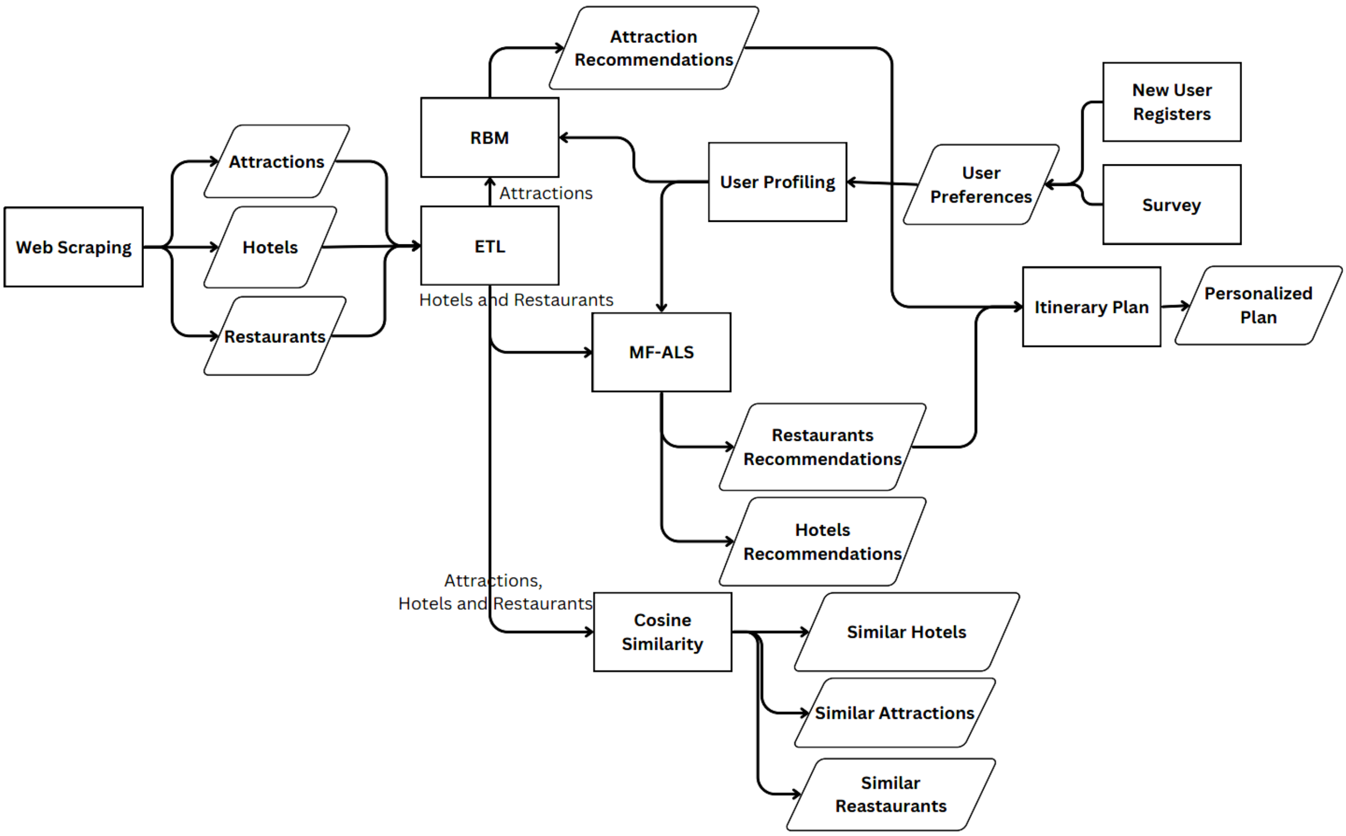

Fig. 1. System Pipeline

1) User Profiling

It is a technique used to solve the new user problem. A user profile is typically used for user modeling and customization. It reflects demographic data, such as name, age, education level, nation, etc., as well as personal information or facts about a specific user.The data collected about users can be Explicit Feedback or Implicit Feedback or Hybrid Feedback [1]. Explicit Feedback is the most reliable way to provide accurate data where the user enters their preferences manually but requires more effort from the user. Explicit Feedback was used in this system where the user is asked to select at least five of their preferences in hotel amenities, restaurant cuisine types and attraction types when they first register into the system. The top 15 amenities, cuisine types and attraction types that were most repeated in the dataset were provided to the user to choose from. This data was used to create a profile for each user by giving an estimated rating to the places according to the availability of the user's preferred amenities in hotels, preferred cuisines in restaurants and preferred attraction types in attractions.

2) Surveying

To solve the new system cold start, a survey was conducted through Google Forms in order to collect real user preferences. Users were asked to choose at least 5 amenities, attraction types and cuisine types from 15 choices that had the most occurrences in the dataset. So they can be used later in user profiling, where we create a user profile for each of them, therefore a new user can be matched with those who have similar preferences as the new user. 133 responses were collected from the survey.

C. Collaborative Filtering Techniques

The idea that users with similar preferences are likely to have similar opinions on items is the foundation of the popular method of collaborative filtering used to create recommendation systems. Collaborative filtering can be implemented using a variety of methods. This paper experimented with two popular collaborative techniques which are Matrix Factorization Alternating Least Square (Mf-ALS) and Restricted Boltzmann Machine (RBM).

1) Matrix Factorization Alternating Least Square

Matrix Factorization is a method used to decompose a matrix into lower dimensions matrices. It is used to uncover latent factors from a given matrix of data. One of matrix factorization's most popular methods is alternating least squares. MF-ALS is an iterative algorithm that works on minimizing the differences between the original matrix and the product of the two lower rank matrices. It is called alternating as it alternates between optimizing the two matrices.

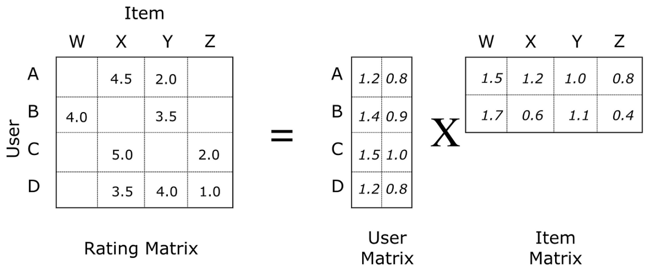

Fig. 2. User-item Matrix

In MF-ALS, the user-item interaction matrix is decomposed into two lower rank matrices: one representing the user matrix and the other item matrix. The two are multiplied together to obtain an approximation of the original matrix, which can predict user-item interactions.

MF-ALS is iterative where it updates the user matrix and item matrix to minimize the difference between predicted and actual ratings. The updates are done using ALS, which means that each iteration one of the matrices will be fixed and the other updated. This process is repeated till the difference between predicted and actual is minimized.

The model is evaluated using root mean square error (RMSE), with each different rank (latent factors) until we reach the smallest RMSE value and that is the best model to recommend data with. The model is then sorted by ratings descendingly and filtered by the city the user chose.

2) Restricted Boltzmann Machine

Restricted Boltzmann Machines (RBMs) are a type of artificial neural network used for unsupervised learning. They were introduced by Geoffrey Hinton and his colleagues [11]. RBMs are used in this paper for the purpose of recommending places to a user.

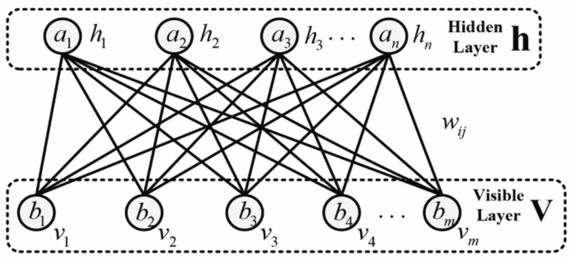

Fig. 3. The basic structure of RBM [12].

RBMs consist of two layers; hidden layer consists of hidden nodes that captures the underlying features of the rated places whereas visible layer consists of visible nodes which correspond to the places.

The training set is modeled through hidden layers and visible layers which are represented in the following energy function:

$$E(v, h) = -\sum_{i\epsilon visible} a_i v_i - \sum_{j\epsilon hidden} b_j h_j - \sum_{i,j} v_i h_j w_{ij} \quad (1)$$

During training, RBM learns a set of weights that minimize the difference between the predicted ratings and the actual ratings. This is done through a process called Contrastive Divergence, which adjusts the weights to make the predictions more accurate. The data used for training was a merge of places data and user's preferences data that was obtained from user profiling.

Once the RBM has been trained, it is used to make recommendations to users. This is done by feeding in the user's past ratings as input to the visible layer, and then using the RBM to predict which places the user is most likely to enjoy. The predicted places then are sorted by their recommendation scores, filtered by the city the user has chosen and then displayed to the user for a better personalized experience.

RMSE was used to evaluate the model with different epochs (iterations) until the model reached the lowest RMSE value possible as well as considering not to overfit the model.

D. *Content-based Filtering using Cosine Similarity*

Content-based filtering is used in the system to recommend places similar to the places selected by the user, represented in a section called "More Like This" below each place the user views. The vector space method is used, cosine similarity is a popular technique that can find similarity between two vectors as shown in equation (2).

$$cosine\ similarity = S_C(A, B) := cos(\theta) = \frac{A \cdot B}{||A||\ ||B||} \quad (3)$$

Cosine Similarity is used in this system to determine relevant hotels, restaurants, and attractions based on similarities between:

- Keywords description of the attractions
- Amenities for hotels
- Cuisine types for restaurants

Steps for implementation is as follows:

- The initial step is to build a "bag of words" model—a list of terms used to describe the items—from the input data.These terms or words are used to characterize the places (restaurants, attractions and hotels)
- Turn the bag (of words) into a feature vector, The outcome is a vector matrix with columns for features and records for items.The textual content in our case is simple, and we want to recommend items based on keyword matching, so we used CountVectorizer which counts the number of times a particular word appears in the document.
- Measure item similarity and assign similarity scores using cosine similarity equation as in (3).
- Sort the items based on similarity scores and suggest the items that are the most similar to the given place [2].

E. *Itinerary Plan Method*

Fig.4 illustrates the itinerary plan that was used to create the personalized trip plan for the user [8].

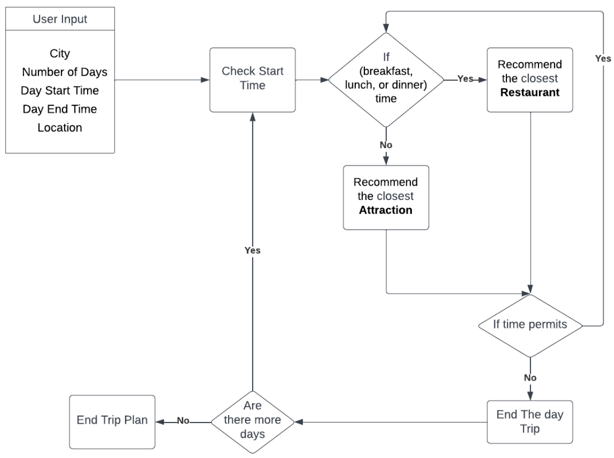

Fig. 4. Itinerary Plan Flowchart

Steps for implementing the plan is as follows:

- User enters the following data: destination city, number of days of trip, the average start time of each day of the trip, the average end time of each day and the starting location— user enters their current address as a starting point or any other address they wish to start the trip from.
- In the city the user chose, the recommendation models of attractions and restaurants are run to return the personalized recommendations for the plan to be constructed.
- If it is breakfast, lunch or dinner time, the closest restaurant with the highest recommendation score is recommended.
- Then the closest two attractions to the user and are open at that time, by checking open hours, are recommended.
- By examining the average end time of the day, the same procedure is repeated until the day is over.
- The process is repeated for all the days of the trip and no places are repeated (each day has different places to visit).
- The current location and time are updated after a place is recommended. One hour is added after a

restaurant while two hours are added after an attraction [9][10].

For calculating the distance between places, Haversine Formula was used as shown in equation (4). The latitude and longitude of two points are used to calculate the distance between two places on the surface of a sphere using the Haversine Formula, which is a very accurate method [13].

$$a = \sin^2(\Delta\varphi/2) + \cos\varphi 1 \cdot \cos\varphi 2 \cdot \sin^2(\Delta\lambda/2),$$

$$c = 2 \cdot atan2(\sqrt{a}, \sqrt{(1-a)}),\ d = R \cdot c \quad (4)$$

Where φ is latitude, λ is longitude, R is earth's radius (mean radius = 6,371km).

IV. Results

A. Data Collection Results

Data was scraped for attractions, hotels and restaurants for the six most popular Egyptian cities (Cairo, Alexandria, Giza, Luxor, Hurghada and Sharm El-Sheikh). Around four thousand places were collected.

- For each attraction the following was scraped: the attraction name, rating, attraction type, description, location, keywords, image and open hours.
- For each hotel the following was scraped: the hotel name, rating, price range, amenities, location and image.
- For each restaurant the following was scraped: the restaurant name, phone number, location, open hours, prince range, rating, cuisine types and image.

B. Comparing the Collaborative Filtering Techniques

RBM and MF-ALS were used as collaborative methods for this recommendation system. The differences between attraction, hotel, and restaurant recommendations were examined. Since the attractions data size is relatively small, the two models ended up having similar error rates for attractions recommendations. Whereas for hotels and restaurants data, MF-ALS had a lower RMSE than RBM, indicating greater accuracy as seen in TABLE 1. RBM performed significantly faster than MF-ALS, taking only ten seconds compared to MF-ALS's one minute and twenty seconds as shown in TABLE 1.

TABLE 1.

Model	RMSE in restaurants	RMSE in Hotels	Speed (seconds)
MF-ALS	0.195	0.22	80 seconds
RBM	0.42	0.32	10 seconds

V. Conclusion

This paper aimed to create a personalized recommendation system for the users where they can get personalized recommendations for hotels, attractions or restaurants and create a tailored plan based on the user's preferences with a focus on Egypt to encourage tourism.

The project used two different collaborative filtering models. MF-ALS and RBM that were evaluated using RMSE. In addition to MF-ALS and RBM, Cosine Similarity was also used to calculate the similarity between places. The Cold Start Problem in Collaborative Filtering was solved using User Profiling and Surveying.

The project also developed an itinerary feature that allows users to create a personalized travel plan based on their starting point and preferences, saving the user time and effort in organizing their trip, while ensuring that they get the most out of their travel experience in Egypt.

Since there is no one-size-fits all solution and what works best for one use case may not work for another, MF-ALS and RBM were compared using RMSE evaluation metric, MF-ALS was shown to outperform RBM in our case as it resulted in a smaller RMSE.

Suggested future work can be:

- Include more hidden gems in Egypt's cities.
- Collecting accurate data implicitly on the average time users spend in each place and the average travel time between places to provide more accurate scheduling in the plan.
- Integrate with websites like Booking.com and different places websites to allow the user to book directly from the system.

Acknowledgment

We thank God for enabling us to successfully complete this project.

We would like to thank Dr. Salsabil Amin and TA Yomna Ahmed, our supervisors, for their constant guidance and support.

References

[1] M. Farid, R. Elgohary, I. Moawad, and M. Roushdy. (2019) *User Profiling Approaches, Modeling, and Personalization*, SSRN.

[2] F. Follonier (2023) *Create a personalized movie recommendation engine using content-based filtering in Python*, relataly.com. Available: https://www.relataly.com/content-based-movie-/recommender-using-python/4294/ (Accessed 5/7/2023)

[3] X. Su and T. M. Khoshgoftaar (2009) *A Survey of Collaborative Filtering Techniques*, Hindawi Publishing Corporation.

[4] Z.B. Ayman (2022) *Recommendation systems: Content-based filtering, Medium*. Available: https://medium.com/mlearning-ai/recommendation-systems-content-based-filtering-e19e3b0a309e (Accessed 5/7/2023)

[5] K. Liao (2018) *Prototyping a Recommender System Step by Step Part 2: Alternating Least Square (ALS) Matrix Factorization in Collaborative Filtering*, Medium. Available:https://towardsdatascience.com/prototyping-a-recommender-system-step-by-step-part-2-alternating-least-square-als-matrix-4a76c58714a1 (Accessed 5/7/2023)

[6] Z. Wang and B. Liu (2019) *Tourism recommendation system based on data mining*. IOPScience.

[7] Chaitra, D., Prasad, V.R.B. and Vinay, B.N. (1970) *A comprehensive travel recommendation system*, SpringerLink.

[8] S. Venkateswaran, S. Senthilkumar and S.P. Thandapani (2020) *Intelligent Travel Recommendation System*. Available: https://github.com/sachinnpraburaj/Intelligent-Travel-Recom

mendation-System/blob/master/report.pdf (Accessed 5/7/2023)

[9] A.L. Pagán (2022) *This is the ideal amount of time to spend at a restaurant*, The Takeout. Available: https://thetakeout.com/the-ideal-amount-of-time-to-spend-at-a-restaurant-1848823547 (Accessed 5/7/2023)

[10] J.G.G. Gabriel, V. Statzu, M. Meleddu and M. Pulina (2020) *How much time do visitors spend in a cultural site? An Empirical Investigation*. ResearchGate.

[11] G. Hinton (2010) *A practical guide to training restricted boltzmann machines*. University of Toronto.

[12] Y. Chu, X. Zhao, Y. Zou, W. Xu, J. Han and Y. Zhao (2018) A *Decoding Scheme for Incomplete Motor Imagery EEG With Deep Belief Network*. ResearchGate.

[13] S. Kettle (2021) *Distance on a sphere: The Haversine formula, Esri Community.* Available: https://community.esri.com/t5/coordinate-reference-systems-blog/distance-on-a-sphere-the-haversine-formula/ba-p/902128#:~:text=For%20example%2C%20haversine(%CE%B8),longitude%20of%20the%20two%20points. (Accessed 5/7/2023)