TOURIST PLACE RECOMMENDER SYSTEM USING MACHINE LEARNING ALGORITHMS

[1]Yuvraj Singh, [2]Siddharth Chauhan,[3]Sanchit
[1,2,3]students
Dept. of Computer Science and Engineering
Maharaja Agrasen Institute Of Technology

Abstract

Today's travellers struggle with decision- making and have a number of uncertainties and queries before visiting a location. Before coming, the visitor is more interested in learning about the destination and gathering information about it, such as reviews and ratings. He also needs assistance determining whether the location is right for him. Thus, the traveller attempts to gather information from travel websites and his or her personal contacts, which takes a lot of time. In recent years, a large number of people have used the internet, technology, and social networking to travel and explore, which has resulted in the creation of a vast amount of data. Today's travellers have a hard time making decisions and are filled with doubts and questions before they travel anywhere. The tourist is more interested in finding out information about the place before they arrive, such as reviews and ratings. He also needs help deciding if the place is the correct one for him. As a result, the traveller spends a lot of time trying to obtain information from travel websites and his or her personal relationships. An enormous amount of data has been produced in recent years as a result of people using the internet, technology, and social networking sites to travel and explore.

Index Terms: Hybrid Filtering, DBScan, Tanimoto, Haversine, Silhouette Score, Canberra

1. Introduction

On blogs, forums, websites for attractions, and other places, tourists can find information about travel. On the internet, though, information overload is possible. As a result, there has been a lot of research on recommender systems in recent years because of how important they are in real life, social networks, e-commerce, movies, and tourism. Today, the Internet has supplanted other search engines as the primary resource used by travellers to find information. They use the Internet to book accommodations, purchase services, communicate their experiences, and plan vacations. However, despite the fact that online knowledge is expanding at an exponential rate, users frequently express frustration at how challenging it is to get the appropriate information quickly. This problem is called information overload. The recommendation system, which aids in addressing the issue of information overload, enters the picture at this point. For the recommendation system to work, it is crucial for users to receive tour information recommendations.

Therefore, the project focuses on creating a tourism recommendation system for the tourists who find it difficult to choose which place to visit. So, by using this system, they can easily learn about the tourist places suitable for them and enjoy visiting the places. The final output of the system will be the recommended tourist place for the users and

the score of how much they will like that place.

2. RELATED WORK

In the literature, there are several studies about the tourism recommender systems, hybrid recommender systems and how they were implemented for location based recommendation. “Pravinkumar Swamy" et al.[1] have proposed a “Tourist Place Recommendation System” in which he used Collaborative Filtering and Content Based Filtering algorithms to recommend tourist places to users. At first, the datasets were extracted using Web Scraping technique. The extracted data were then cleaned and pre-processed. Two items are considered similar when they receive almost identical ratings from a certain user. This is the case in item based collaborative filtering. Then, a weighted mean of ratings on the items that are more comparable to the item that the user wants to rate was calculated and used as rating. This process was divided into steps of three. First, the depiction of user information, where the travel history and reviews/ratings of the user are checked. Secondly, the similarities between the tourists can be calculated on the basis of their past visiting history data and the CF algorithm proposed. Third step is to create a list of recommended places. After the CF, content-based filtering techniques are used which emphasize on the user’s preferences and the properties of the item for recommendation. Here, they use the cosine similarity which gives the result in the range between 0 and 1. “Prof. P. A. Manjare et al.[2] have proposed a “Recommendation System Based on Tourist Attraction” which generates a recommendation list of tourist attractions and hotels based on the city using the data mining techniques, then analyses user details and filters the profiles using Collaborative Filtering. The system also city wise re-ranks the tourist places depending upon their ratings and reviews. Data mining techniques are used which are of different types. The first one is the User’s Profile wise Collaborating Filtering in which a database of user preferences for the items is created. During Registration, the tourist will specify about their likings, his past visited places and search history. While the tourist searches his tourist attraction, the algorithm tracks his interest through the available data. Users will rate the tourist places. The second type is Collaborative Filtering Recommendation which evaluates the collaborative technique depending on the items which are rated by particular users. This Application does not utilise the climate of a location that is the geographical conditions and the time of day for recommendation. Does not tackle issues like location positioning etc. “Hela Masri” et al.[3] have proposed “A Personalized Hybrid Tourism Recommender System” that utilises the three most popular recommendation techniques - The collaborative(CF), the demographic (DF) and the content-based(CB). Various machine learning algorithms have been executed to implement and combine these recommendation techniques. The techniques are the decision tree for the DF and the K nearest neighbors (KNN) for both Collaborative and Content-based filtering. Firstly, data set was crawled from e-tourism website TripAdvisor.com using a web crawler called Web Harvy. The information presented in HTML format is converted to the structured data. Now, to determine the rating of the user on one particular product, the algorithm of nearest neighbor was used which computationally calculates the distance between the current item whose rating is needed and all the items that were previously rated by the user. The distance measure used was Euclid’s Algorithm. This was the content based approach. The demographic approach tries to make a user belong in a particular class based on his personal details like age, occupation gender and region to get his rating on one specific spot, so they commissioned the nodes as representative of the demographic details and the leaves as representative of the ratings. The ID3 decision tree algorithm was chosen because of its speed

and the available discrete features. User based Collaborative method works by predicting the active user's ratings on the items that are not yet rated by the user. The key idea is to use the particular
user and his friends' interests. Tanimoto was chosen to be the measure of similarity because it was the most accurate one for the dataset. Then, a weighted and switching hybrid technique was utilised as it alters between different recommender techniques' results to benefit from each type at different situations and to get the most accurate rating result. This system aims to provide a systematic tour plan given the recommendations in the future. "Akbar Etebarian " et al.[4] have proposed "A hybrid recommender system based-on link prediction for movie baskets analysis" which uses a method on the basis of link prediction so that it meets the limitations of individually used techniques. The recommended solution contains four phases. The first phase is the Content Based Recommendation System. This phase makes sure that all the users are clustered using Density-based spatial clustering of applications with noise algorithms (DBScan). After the clustering of the existing users, the new users are classified using the Deep Neural Network algorithm (DNN). The second phase consists of the Collaborative Recommender System (CRS) which uses the Hybrid Similarity Criterion. This criterion calculates the similarities between the newly registered users and the existing users who are in the particular cluster which was identified according to a threshold. The criteria consists of gender, age and occupation. Phase three utilises an improved Friendlink algorithm so as to calculate the similarity between the users. The final phase combines the collaborative recommender system's result and the improved algorithm of Friendlink. This method faces the limitation of longer processing and execution time than existing solutions. Riteshwari Ganjare" et al.[5] have proposed a "Hybrid Recommendation System For Tourism Based Social Network, and AI" which is the implementation and combination of different types of recommender systems utilised in the tourism field. The main aim is to design a framework that clarifies the working of the Hybrid recommendation systems. The Online Social Networking (OSN) system module was developed in the first module for user registration and login. The process of user data collection involves collecting the ratings of the items based on its usage and the appreciation that it was provided, and the Demographic attributes of the user, such as occupation, age, gender, socio-professional category, location, personal status. Implementations of recommendations for new users were executed with the help of this.
The cold start problem of users was solved by this solution. For the content-based technique, the past visited places of the user are in the form of keyword vectors that are generated following an indexing phase. The social module is the collaborative phase consisting of the rating data of the consulted items by the other users. Finally, the user is recommended with the items that are considered relevant for him by the system based on his context. This system identifies the requirements of the tourists and indicates the resources best suited for them even if their behaviour changes in the future.

Serial No.	Title	Advantages	Limitations
1.	Tourist Place Recommendation System [1]	The user interface developed is interactive, responsive and user adaptable. Issues of Distributed systems were considered.	The Recommendation system uses the traditional techniques and thus suffers the limitation of these algorithms like user and item cold start problems, not providing broader suggestions etc.
2.	Recommendation System Based on Tourist Attraction [2]	Uses user preferences and broader suggestions and perspective to the users to recommend nearby hotels and attractions.	This Application does not utilise the climate of a location that is the geographical conditions and the time of day for recommendation. Does not tackle issues like location positioning technologies, Query processing, social media upload
3.	A Personalized Hybrid Tourism Recommender System [3]	Accurate predictions are given by using the hybrid recommendation as it solves the limitations of individually processed systems.	Does not provide a systematic tour plan given the recommendations in the future.
4.	A hybrid recommender system based-on link prediction for movie baskets analysis [4]	Solves the cold start problems in online movie systems and generates correct movie recommendations to new users with efficient accuracy.	This method faces the limitation of longer processing and execution time than existing solutions.

TABLE I: Taxonomy on the Recommender models

From Table I, it is observed that the model building for the tourist recommendation system is expensive, time consuming and the model suffers from the cold start problem.

3. PROPOSED SOLUTION

This system is using the Hybrid recommendation techniques for tourism recommendation. In that, it uses the DBScan method for Content based recommendation system and Tanimoto Coefficient for Collaborative filtering. DBScan model is used to classify the tourist places and get the recommendations for a user according to his past visited places or places of interest. Tanimoto Coefficient is used for getting the similarities between the places. The user can get the likeability score for the places that are recommended to him.

3.1 DBScan Algorithm

Clusters are regions with more density in the data space. These are separated by the boundaries of lower dense regions of points. The

DBSCAN algorithm works on the basis of noise and clusters. The main idea is that the neighborhood of a given radius of a cluster point must contain at least a minimum number of points. There are two main attributes. First is the minPts (The minimum number of points a cluster point needs to have inside its region to be considered dense). Second is the eps (ε) (radius for a cluster point).

3.2 Tanimoto Measure

The Tanimoto coefficient is calculated by identifying the number of features that are similar in both data (the intersection of the data) with respect to the number of features that are in both the data (the union of the data).

3.3 Hybrid filtering

To create a more reliable model, hybrid filtering systems combine two or more different filtering approaches. By doing this, we can replace one type's drawbacks with another type's benefit. In paper [1] the author has created a model by combining both demographic and collaborative filtering techniques to create a movie recommendation system. In this way the recommendation quality was significantly improved. They were able to recommend by utilizing both user's demographic data as well as based on the user ratings on the movies.

3.4 Architecture of the proposed solution

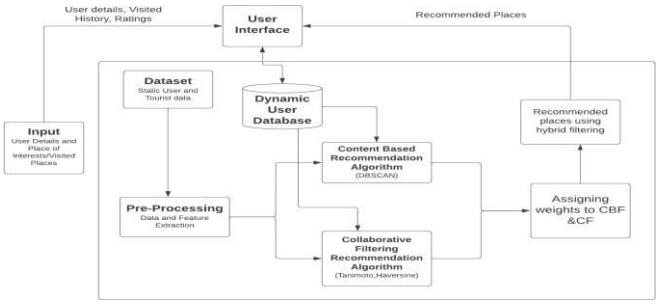

Fig 1: Architecture for Tourism Recommender System using ML

The architecture layout followed for this work is laid down in Fig. 1.

Firstly, the user_checkins dataset, user_friendship dataset, locations dataset were extracted from the gowalla website and stored in the drive. Then, the null values were removed and the outliers were replaced. The city_state attribute from the locations dataset is splitted into city and its abbreviation. Here, only certain users were retained from the checkins dataset who have visited more than 5 places and less than 50 places. This was done to minimise the computing time. For that purpose, the groupby() function was used and the users were grouped to view their count. The filtered users were then merged with locations dataset to see which users have gone to which places. The merging was done using the parameters "placeid" of filtered_checkins and "id" of the locations data. After dropping the redundant data, "frequencies" of the visited places by each user were calculated. Finally, the frequencies were normalized on the scale of (1,10) using the MinMaxScaler() to get the ratings.

For the content based recommendation, a graph was plotted to analyse the locations with latitudes as x axis and longitudes as y axis. The DBSCAN() was imported and the model was created with eps distance and minimum points parameters. The locations were clustered using the above model and the labels were stored. The silhouette_score() was used to evaluate the model. If the score is negative, then there are too many outliers. If the score is positive and close to zero, there exist overlapping clusters. Score being closer to 1 indicates the model being more accurate. For clustering a particular location, classify() function was created. All the past visited places of the particular user were extracted. The near locations of all those visited places were stored. The near locations for the visited places were extracted using the DBscan clustering. The places within the same clusters were taken. The rating for all the recommended places were calculated by getting each visited place's rating and dividing it with distance from that recommended place. The rating was then normalized.

Also, the places were recommended using location aware recommendations. The user enters the particular location similar to which he wants the recommendation to be. The latitude and longitude of that place are found. The locations dataset is searched and the places having similar latitudes and longitudes are extracted. The similarity of latitudes and longitudes are calculated through different similarity measures like Cosine similarity, Canberra similarity, Tanimoto similarity and Haversine similarity. **Tanimoto** : $(\Sigma x.y) / (\Sigma x^2 + \Sigma y^2 - \Sigma x.y)$

Haversine :

$\triangle lat = x1 - x2$

$\triangle long = y1 - y2$

$a = sin^2(\triangle lat/2) + cos(x1) + cos(x2) + sin^2(\triangle long/2)$

$c = atan(\sqrt{a}, \sqrt{(1-a)})$

$haver\_dist = R.c$

where x1,x2 are the latitudes and y1,y2 are the longitudes, c is the central angle and R is the radius of the Earth.

Canberra :

$[\ |x1 - x2|\ /\ (x1 + x2)\ ] + [\ |y1 - y2|\ /\ (y1 + y2)\ ]$

Cosine : $x.y\ /\ |x|.|y|$

The similarity ratings of all the places with that particular place was calculated through these algorithms. According to the efficient similarity measure among the four i.e. tanimoto, the places were sorted. The efficient measure was found by calculating mean absolute error and root mean square error of all measures and finding which one gives least error. Thus, The top places are filtered and the other users that have rated these places checked. Among these places, those which are highly rated by the other users are recommended to this user. Finally, the weighted hybrid filtering technique was used on the places that were recommended to users by content based and collaborative methods. The weights alloted to both were 0.5.

4. EXPERIMENTAL RESULTS

In this section, we first introduce the dataset. Then we present the measures of the proposed model.

4.1 Dataset

The dataset used was gathered from the social media Gowalla. It consists of 36,001,959 check-ins by 407,533 users in 2,724,891 Places of Interests. It includes the datasets of user check-ins information, locations information, and users information. But only users who have more than 5 check-ins and less than 50 were selected to reduce the computing time. The attributes that we used after collecting the datasets are userid, placeid, place name, latitude of the location, longitude of the location, city name, frequency of place and rating. The check-ins data was first grouped according to the users to get the count of visited places by each user. The check-ins data was merged with the location dataset. This gave us the details of places that each user visited. Then, the frequency for each location was calculated as shown in figure 6.1. It indicated how many times each user visited each place. After that, the frequencies were normalised in the range of 1 - 10 to give us the rating.

4.2 Measures

The following measures are used to evaluate the performance of the proposed system: - Root Mean Square Error and Mean Absolute Error.

MAE Formula: $(1/k).\ \Sigma_u\ \Sigma_i\ |P_{u.i} - A_{u.i}|$

RMSE Formula: $\sqrt{((1/k).\ \Sigma_u\ \Sigma_i\ (P_{u.i} - A_{u.i})^2)}$

where, i is the item id; u is the user id; P is the predicted rating; A is the actual rating; k is the number of iterations.

DBScan clustering was used in content based filtering, the locations were clustered according to their latitudes and longitudes. The DBScan model was evaluated using the silhouette score. The model that was used here gave the score of 0.23.

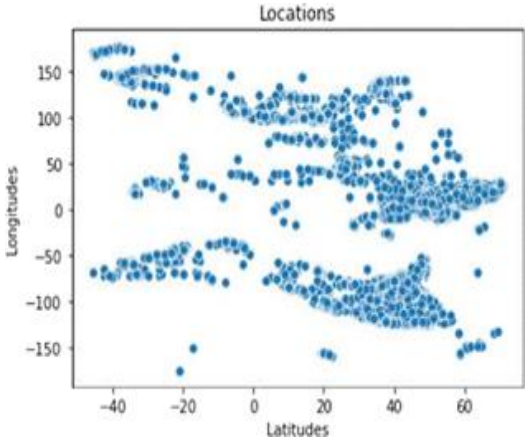

Fig 2: Tourist places without clustered

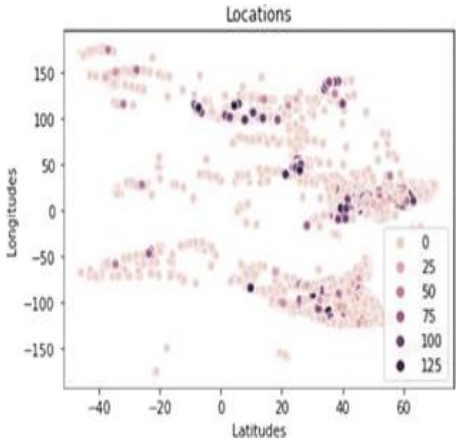

Fig 3: Tourist places clustered by DBScan

Fig 3 shows the tourist locations belonging to different clusters.

```
from sklearn import metrics
metrics.silhouette_score(coor, db.labels_)
```

```
0.2387617641660904
```

Fig 4: Silhouette Score

If the value of the score is negative (-1), it means that there are too many outliers. If the score value comes too close to 0, then it indicates that most of the clusters are overlapping. The closeness of the value to 1 gives the perfection of the model.

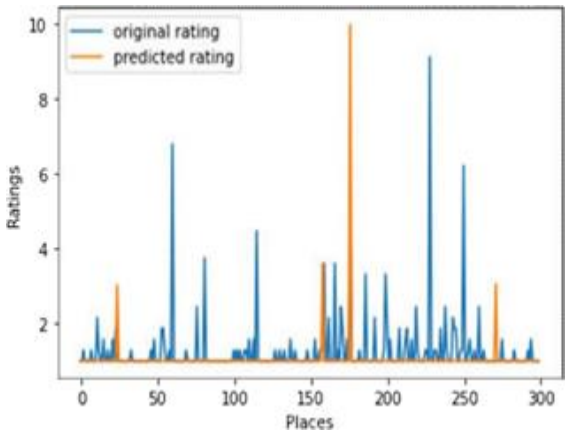

Fig 5: Observed and Predicted ratings for locations

DBSCAN Model - Content Based Filtering	
Dataset Size - 3000 locations	
MAE	0.21
RMSE	0.68

Table II: Performance Comparison of DBScan model using MAE and RMSE

From table II and figure 5, it can be inferred that the mean absolute error and root mean square of the DBScan technique for 3000 locations are 0.21 and 0.68 respectively. More the value being closer to 0, lesser the error i.e. difference between actual and predicted rating and more the model is perfect.

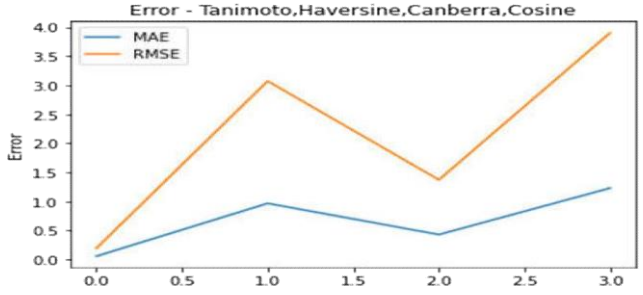

Fig 6: Linear Graph for MAE and RMSE of different

similarity measures

Dataset Size - 10 locations (5000 nearby locations)				
Metrics	**Tanimoto**	**Haversine**	**Canberra**	**Cosine**
MAE	0.047	0.993	0.509	1.23
RMSE	0.105	2.221	1.13	2.76

Table III: Performance Comparison of similarity measures using MAE and RMSE

From table III and figure 6, it is inferred that the tanimoto measure gives the least mean absolute error and root mean square error among the similarity measures and the cosine similarity measure gives the most mean absolute error and root mean square error among them. Four similarity measures are taken and evaluated for 10 locations and their 5000 nearby locations. These measures work perfectly in the order tanimoto, canberra, haversine and cosine as shown in the figure 6.4. Thus, the tanimoto similarity measure has been used for collaborative filtering.

Finally, the weighted hybrid filtering technique is used on the places that are recommended by both the content based and collaborative filtering. The weights used on the predicted ratings of content based and collaborative techniques are 0.5. This hybrid filtering technique solves the cold start problem of new users and new items.

5. CONCLUSION

In this project, Machine Learning algorithms have been used which is able to help in recommending attractive tourist places for the users who suffer from Information Overload. The users cannot decide which place to choose and visit and how to decide. Thus, this system helps the user to overcome this problem and decide their next visiting place. The user is able to get the details of the recommended places and the rating other users have given and this system also predicts the rating that the current user may give to the places.

The machine learning model running in the background allows users to provide them with a satisfactory and wholesome experience by ensuring that the places recommended to them are up-to the mark.
The machine learning model is equipped with a weighted hybrid filtering technique enabling the user to use the system in a working environment. The DBScan model is used to recommend the places using the user's past visited places that is content based filtering and Tanimoto similarity measure is used for the collaborative filtering. The usage of the hybrid technique solves the cold start problem.

The mean absolute error and root mean square error for DBScan algorithm are 0.21 and
0.68 respectively and for Tanimoto measure are
0.047 and 0.10 respectively.

Overall, the Tourist recommender system can help recommend various places that the user may like and want to visit according to his past visit history as well as other user's preferences using machine learning. The future scope of this project contains more algorithms to be used for content based and collaborative filtering and getting less error value compared to now.

6. REFERENCES

[1] "Pravinkumar Swamy", "Sandeep Tiwari","Kunal Pawar" and "Prof.Bharati Gondhalekar", "2021", "Tourist Place Recommendation System" , No. "February"/"2021"
[2]"Prof. P. A. Manjare", "Miss P. V. Ninawe", " Miss M. L. Dabhire" et al, "2016", "Recommendation System Based on Tourist Attraction", No. "April"/"2016"
[3]"Mohamed Elyes Ben Haj Kbaier", "Hela Masri" and "Saoussen Krichen" , "2017" , "A Personalized Hybrid Tourism Recommender System" , No. "October"/"2017"
[4]"Mohammadsadegh Vahidi Farashah", "Akbar Etebarian", "Reza Azmi" and "Reza Ebrahimzadeh Dastjerdi", "2021", "A hybrid recommender system based-on link prediction for movie baskets analysis" , No. "February"/"2021"
[5]"Riteshwari Ganjare", "Riya Sahu", "Pratidnya Kharate", "Vaishnavi Lohakare", "Gomati Sharnagat", "2022", "Hybrid Recommendation System For Tourism Based Social Network, and AI", No. "May"/"2022"
[6]"Saman Forouzandeh", "Mehrdad Rostami","Kamal Berahmand", "2022", "A Hybrid Method for Recommendation Systems based on Tourism with an Evolutionary Algorithm and Topsis Model" , No. "January"/"2022"
[7]"Xixi Li", "Jiahao Xing", "Haihui Wang","Lingfang Zheng","Suling Jia","Qiang Wang","2017", "A Hybrid Recommendation Method Based on Feature for Offline Book Personalization", No. "October"/"2017"
[8]"Sambhav Yadav", "Vikesh", "Shreyam" and "Sushama Nagpal", "2018", "An Improved Collaborative Filtering Based Recommender System using Bat Algorithm", No. "December'/"2018"
[9]"Hoang L. HU-FCF", "2015", "A novel hybrid method for the new user cold-start problem in recommender systems", No, "May"/"2015"
[10]"TaehyunHa", "SangwonLee", "2017", "September"/"2017"
[11]"Maryam Khanian Najafabadi", "Azlinah "November"/"2017"
[12]"Thiago Silveira", "Min Zhang", "Xiao Lin", "Yiqun Liu", "Shaoping Ma", "2017" "How good is your recommender system is? A survey on evaluations in recommendation", No. "December"/"2017"