Traveler's Guiding System Using Data Mining Techniques.

Prof. Sagar Birje*

Asst. Prof & HOD. Department of Computer Science & Engineering, Angadi Institute of Technology and, Management, Belagavi, India

Pramod Patil[1], Kshitija Desai[2], Malatesh Patil[3], Malaprabha Patil[4]

Department of Computer Science & Engineering, Angadi Institute of Technology and Management, Belagavi, India

---***---

Abstract — Nowadays, Recommender Systems are being used in multiple different domains. The application recommends to a tourist the best attractions in a particular place according to his preferences, his profile and his appreciation to previous visited places. This paper proposes a hybrid recommender system that combines the three most known recommender methods which are: the collaborative filtering (CF), the content-based filtering (CB) and the demographic filtering (DF). In order to implement these recommender methods, we have applied different machine learning algorithms which are the K-nearest neighbors (K-NN) for both CB and CF and the decision tree for the DF. The hybridization is a good choice to make the best of their advantages and to overcome the cold start problem. To enhance the recommendation accuracy, we use two hybridization techniques : switching and weighted.

Keyword- Recommendation, KNN clustering algorithm, Apriori classification algorithm, Content based filtering.

I. INTRODUCTION

Data mining is nothing but the process of identifying patterns in order to extract useful data from large datasets using methods of machine learning, statistics and database systems. As a huge amount of data get generated from various organizations websites, social media sites, etc. Hence it's very essential to extract required data which will useful for taking future decisions. Sometimes past history/data becomes useful for future predications. Main goal of data mining process is to extract data from large datasets and convert it into understandable or useful format.

Travelling is an important part of our life. [1] Hence planning of it is also important. Nowadays, lots of travel agencies are there which helps tourists to plan their vacations according to their packages. Hence, sometimes user's needs to adjust their plans according to their agencies generalized plans. Lots of Websites provides us travelling options. Some websites helps us to plan our trips. They recommend us places if we specify a particular location. But these systems are more generalized and also they may suggest us same places repeatedly.

In this paper, we propose to develop a new hybrid tourism rec- ommender systems that combines three recommender filtering methods (CF, CB and DF) while using two hybridization tech- niques: switching and weighted. For the weighted technique, we propose an automatic approach to set the weights' values by applying a novel linear programming model

II. EXISTING SYSTEM

When we want to plan a trip for holidays or general visit, very first we take a help from travel agencies then we need to plan according to travel agencies.[2]But, because of this we face some difficulties like our vacation get start but travel agency package date is at the end of our holiday or in our working time.

Existing system is generalized system, i.e. travelling recommendation might be same for some of tourists. It provides plans according to travel agencies, which is not match with tourists need and interest. Sometime travel agencies promises good quality service to tourist, but that does not happen actually and tourist face many problems.

III. RECOMMENDATIONSYSTEM

We propose a system in which tourist will define his/her interest, type of place in which user is really interested then system will provide some recommendations like best places to visit according to season, route, hotels, start time and end time, address, website(if available), reviews of other users, etc. based on his/her need, past history and interest. Then tourist will choose place and other things according to his/her need.Tourist will first need to fill the details then the system will analyze the data entered by tourist. These details will include information like users current location, distance range in Km and also types of places user is interested in. Here, user can choose multiple types of places. After this system will analyze data entered by user and will recommend the number of places which are within the specified range as well as which are of that specified type. After clicking on that particular place user will be able to view all the information about that place along with photos and reviews of that place. Also user can be able to view frequently visited places which are visited by most of the users even though it's not within that specified range. Content-based filtering is used in review module of recommendation system. System takes the reviews in text format and generates the numeric value for each review. These numeric values get displayed in star ratings to the tourist users and these star ratings help the user to decide which place to visit. For filtering data of reviews it uses dictionary of words which contains positive and negative words. These words help to generate numeric values for text review entered by user. This process uses **Text Mining** approach for filtering the words present in review.

IV. SYSTEM ARCHITECTURE

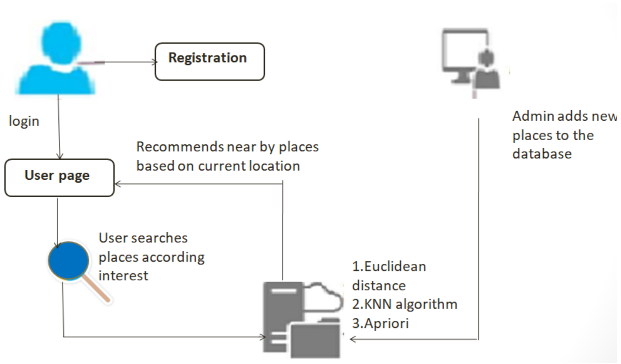


Fig (1) : System architecture .

Fig. 1 shows the architecture of our proposed system, it contains three modules: the data set, the recommender engine and the user interface. When a user log in the system, he must choose what type of activities to search (i.e. hotels, restaurants and destinations). Once he chooses the type, the hybrid engine generates a personalized recommendations according to the user's interests. The engine applies the three approaches (CF, CB and DF) and combines their results in order to provide a good prediction of the activities' rates. As an output, the user receives a final ranked list of recommended activities based on their predicted rates.

Euclidian Distance formula is used to find distance between two places i.e. distance between users current location and places stored in dataset.

KNN algorithm is used to find the nearest attraction and location in this project. It uses Euclidean distance formulae to find distance between two places. KNN finds out the places which are within that range.

Apriori algorithm is used for classification. In this project, we are recommending the most visited places by tourists as we are storing history of tourists.

V. IMPLEMENTATION

1. User Registration and Login Module

Tourists firstly need to register to the system. In registration user needs to provide some personal details such as email id and password. This data will get stored into database in encrypted format to protect user's data. After registration user can be able to log into the system.

2. Recommendation Module

In this module, user searches for places in which user needs to provide his/her current location and range within which user wants to search for the places. System provides list of places based on user's interest. System filters the places based on parameters like current location of user, range and type of places in which users are interested.

3. External Modules

System provides hotel link which is redirected to the pre-existing hotel websites. It also provides a route link and by clicking on it user gets a route from source to destination with the help of google map api.

4. Admin Module

Admin can add new places into the databases. Admin needs to add all related information and admin can also add images of places.

A. *Collaborative filtering*

In our work, we have used the NReco Recommender framework which is port of Apache Mahout CF engine, also we have chosen the user-based CF. This method aims to predict the active user's rates on the non rated products basing on the user' interest and the neighborhood interest. There are many similarity measure to compute the distance between two users

Where f_i is the set of item for which user i express preference, f_j is the set of item which user j express preference and f_i f_j is the intersection set of preferred item for user i and preferred item for user j.

The user-based collaborative filtering suffers from cold-start problem which appears when we have a new user that has no ratings enters into the system or when we have a new item that is non rated yet by users. In those cases the CF algorithm cant predict user's rates. This method contains two main steps:

- Computing interest similarity between the active user and all other users by using Tanimoto coefficient measure in Mahout.
- Searching neighbors users by the K-NN algorithm, here

  K is equal to 50.

B. Content-based filtering

In our work, we have used the C# language to implement the CB. In order to predict the rate of the user on one specific product, we used the nearest neighbor algorithm which computes the distance between the current item with all the rated item by the user. Each item is represented as a vector of features.

$$dE(x,y) = \sum_{i=1}^{N} \sqrt{(xi2 - yi2)}$$

C. Demographic filtering

The decision tree try to classify a user according to his information profile (age, gender, region, travel style) in order to obtain his rate about one specific activity, so we assign the demographic information as the nodes and the ratings as the leafs.

There is many decision trees algorithm such as ID3, C4.5,

CART, CHAID and GUIDE [17]. We have chosen the ID3 decision tree proposed by Quinlan in 1986 [18] because it mostly used when we have discrete attributes and because it builds the fastest tree. ID3 suffers from the over-fitting or overclassification problem if a small sample is tested that's why we have pruned the constructed decision tree.

The DF approach overcomes the new user cold start problem that appears in the CF and CB approaches but it still suffers from the new item cold start problem where it cannot predict the rate of a new item that has no previous rates.

E. Hybrid engine

In this paper, we propose a hybrid model that combines the three RS approaches: the CF, the CB and the DF. This method aims to overcome the drawbacks of each recommender approach used separately and especially the cold start problem.

Moreover, this hybrid method tries to find the best combination of the cited approaches in order to increase the accuracy of prediction.

Each algorithm predicts the rate of one user u on one item i separately and the hybrid method combines all of them.

This method takes into account the advantage of each one, for example, if we are on the case of a new activity where no one rated it yet then the CB will perform better than the other methods, and if we are on the case of a new user then the DF will perform better than the other methods.

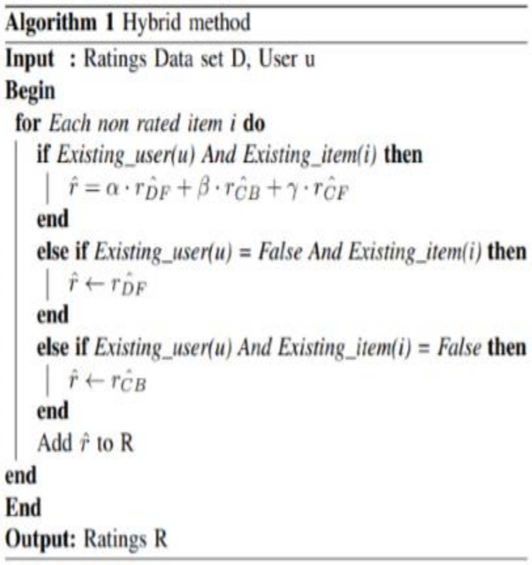
To avoid the cold start problem and to take into account the advantages of each recommender approach, we have realized a particular hybrid method (Fig. 2) that uses a double hybridization techniques, a weighted hybridization combines the rating of users by the following formula (3):

$$r\hat{}w=\alpha \cdot rD\hat{}F+\beta \cdot rC\hat{}B+\gamma \cdot rC\hat{}F$$

Where, *rD̂F* is the predicted rating using the DF approach,

rĈB is the predicted rating using the CB approach and *rĈF* is the predicted rating using the CF approach. *α, β* and *γ* are fractions that represent the weight of each method.

A switching hybrid techniques switches between different recommender results in order to take advantage of each type at different situation and to take the best rating result. The switching techniques uses the weighted hybrid recommender result in the case of an existing user and an existing item situation, in the case of an existing user and a novel item situation, it uses the CB recommender result and in the case of a novel user and an existing item, it uses the DF recommender result



```
Algorithm 1 Hybrid method
Input : Ratings Data set D, User u
Begin
  for Each non rated item i do
    if Existing_user(u) And Existing_item(i) then
      $\hat{r} = \alpha \cdot r_{\hat{DF}} + \beta \cdot r_{\hat{CB}} + \gamma \cdot r_{\hat{CF}}$
    end
    else if Existing_user(u) = False And Existing_item(i) then
      $\hat{r} \leftarrow r_{\hat{DF}}$
    end
    else if Existing_user(u) And Existing_item(i) = False then
      $\hat{r} \leftarrow r_{\hat{CB}}$
    end
    Add $\hat{r}$ to R
  end
End
Output: Ratings R
```

The Algorithm 1 summarizes the hybrid method. It finds the rates of all the non rated items by the active user in order to make recommendation. At every item it switches to the optimal solution as follows:

- Check whether no cold start situation is detected. If so, it use the average weighted sum of DF, CB and CF results.

- Check whether a new user cold start situation is detected.

If so, it use the DF recommender result.

- Check whether a new item cold start situation is detected.

If so, it use the CB recommender result.

In order to find the optimal and stable coefficient of the weighted technique, we did many experiments using the cross validation procedure. We proposed a new linear programming

model (4) described as follows :

minimize _

$$\frac{\sum_{k=1}^{n} |\alpha \cdot rD\hat{}Fi + \beta \cdot rC\hat{}Bi + \gamma \cdot rC\hat{}Fi - yi|}{n}$$

subject to $\alpha + \beta + \gamma = 1$;

α_ 0;

β_ 0;

γ_ 0.

This linear programming problem minimizes the difference between the the weighted sum result $\alpha \cdot rD\hat{}F + \beta \cdot rC\hat{}B + \gamma \cdot rC\hat{}F$ and the real value of prediction yi, where n is the number of tested instances.

Using the cross validation, we got 10 subsets of training and testing sets. For each subset, we run our three recommender systems separately to compute the predicted values. Given the obtained results for each subset, the linear programming problem is then executed to generate the optimal values of the weights. For each subset, we run our three recommender systems separately to compute the predicted values. Giventhe obtained results for each subset, the linear programming problem is then executed to generate the optimal values of the weights. The final coefficients to be included in our system, are the average of the optimal decision variables values of the 10 subsets. The final retained values are as follows

$\alpha = 0, 02$, $\beta = 0, 83$ and $\gamma = 0.15$.

VI. EXPERIMENTAL RESULTS

In this section, we describe the implementation of our tourism recommender system and some scenarios of it.

A. Data set generation and pre-processing

We extracted data set from Trip Advisor website as the experimental data and we chose Paris as a destination because it's one of the most famous touristic destination in the world and because attractions in Paris have very large number of reviews in Trip Advisor.

Our data set contains 11, 737 reviews rated by 6576 users on 160 attractions.

In order to use the activities' attributes by the CB approach and to compute the Euclidean distance between the activities features, we have transformed the features to a vector with numeric value and the categorical attribute "activity category" is encoded into several binary attributes. Each activity can contains more than one category value from this selection (Points of Interest & Landmarks, Churches & Cathedrals, Historic Sites, Architectural Buildings, Monuments & Statues, Sacred & Religious Sites, Fountains, Educational sites). So we have transform each category value to one binary attribute where 1 represents the existence of the category and 0 represents the nonexistence of the category.

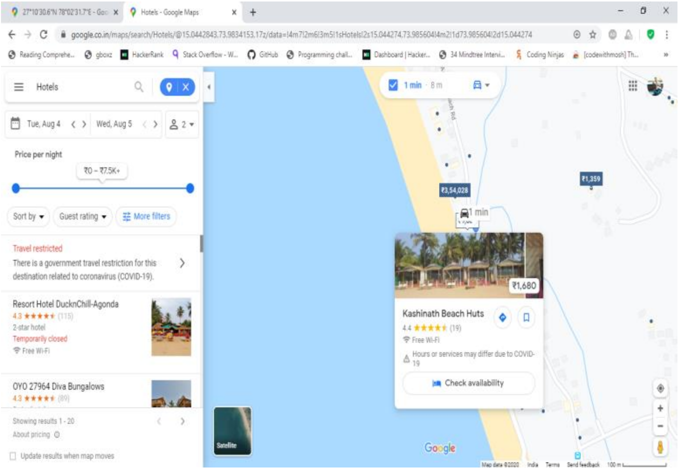
In TripAdvisor, there 20 possible styles for each activity, which are: Foodie, Beach Goer, Nature Lover, History Buff, Vegetarian,

60+ Traveler, Backpacker, Eco-tourist, Like a Local,

Luxury Traveler, Trendsetter, Thrifty Traveler, Urban Explorer,

Family Vacationer, Thrill Seeker, Art and Architecture Lover,

Peace and Quiet Seeker, Shopping Fanatic and Nightlife Seeker. Each user can have more than one travel style. We merged these values into five groups in order to reduce the sparsity of our data set. The five groups are the following: style 1 (Foodie and Vegetarian), style 2 (Beach Goer, Nature Lover, Eco-tourist and Backpacker), style 3 (History Buff, Art and Architecture Lover, Peace and Quiet Seeker), style 4 (Urban Explorer, Like a Local, Family Vacationer) and style 5 (Thrill Seeker, Shopping Fanatic, Nightlife Seeker, Trendsetter and Luxury Traveler). This five travel style attribute are transformed also into binary attribute. transportation means, nearby attraction, interface for booking hotels from existing websites. In Future, it can be further extended to provide hotel booking system instead of redirecting to existing hotel booking sites. Also instead asking user for his/her interest system can fetch it from user's social networking profiles.



google.co.in/maps/search/Hotels/@15.0442843,73.9834153,17z/data=!4m7!2m6!3m5!1sHotels!2s15.044274,73.985604!4m2!1d73.985604!2d15.044274

Hotels

Tue, Aug 4 | Wed, Aug 5 | 2

Price per night: ₹0 – ₹7.5K+

Travel restricted
There is a government travel restriction for this destination related to coronavirus (COVID-19).

Resort Hotel DucknChill-Agonda
4.3 ★★★★★ (115)
2-star hotel
Temporarily closed
Free Wi-Fi

OYO 27964 Diva Bungalows
4.3 ★★★★★ (89)

Showing results 1 - 20

Kashinath Beach Huts
₹1,680
4.4 ★★★★★ (19)
Free Wi-Fi
Hours or services may differ due to COVID-19

Fig (2) : Result based on the nearest Hotel and its per day Rate .

VII. REFERENCES

[1] "Smart traveller guide: A model for guiding traveller with imagematching algorithm" J. Sindhu Sri; N. V. Sri Sravani; P. Suresh Kumar.

[2] "KAMO - mobile guide for the city traveller "J. Liikka; J. Lahti; P. Alahuhta; M. Rosenberg"

[3] Route choice decision-marking analysis based on congestion charging "Zhenggang Li; Jian Wang; Qiu Yan; Ling Zhou

[4] "A Model of Risk-Sensitive Route-Choice Behavior ad the Potential Benefit of Route Guidance "J. Illenberger; G.Flotterod; K. Nagel

[5] "Urbis: A touristic virtual guide" Ivaldir de Farias; Nelson Leitão; Marcelo M. Teixeira

[6] C. Bettini; X. S. Wang; S. Jajodia, "Protecting Privacy Against Location-Based Personal Identification", In: SECOND VLDB WORKSHOP SECURE DATA MANAGEMENT (SDM), 2005, Trondhein, Noruega.

[7] Barry Brown & Mathew Chalmers, "Tourism and Mobile Technology", University of Glasgow, Glasgow, 2012

[8] D. Buhalis &, R. Law, "Progress in information technologyand tourism management: 20 years on and 10 years after theInternet - The state of e Tourism research". 2008, Tourism Management, 29, 609–623.

[9] "Traveler's Guide": A personalized recommendation system for tourists, Prof. Shrikant Kokate, Ashwini Gaikwad, Manisha Gutte, Pranita Patil, Kalyani Shinde, published in IJISRT.