2018

Destination Recommendation System

CS598

KUNAL JAWALE(U00286144)

Table of Contents

List of Figures

➢ Objective

To create a recommendation system which help users to explore new places of interest in different categories based on the likes of other user choices and ratings.

➢ Overview

As Netflix CEO Reed Hastings says, "You know, think about it, when you watch a show from Netflix and you get addicted to it, you stay up late at night. We're competing with sleep, on the margin." Same way, the important thing is to deliver right content and information to right people at the right time. The discovery of IT industry has led to a revolution which have brought an expansion in travel and tourism industry. It has given new idea to people to travel and explore new things based on the availability of destination-related information. People, now-a-days, based on rating and reviews on internet, looks for new places, restaurant etc. to plan and enjoy a quality time with their family and friends. So, looking upon above facts, I have decided to build a recommendation system which will help user to provide a recommended destination as per user interest. To build recommendation system, I have used Foursquare dataset which contains the check-in of users at different places, restaurant etc. in New York City.

For designing recommendation system, Machine learning techniques are used. To start with dataset, pre-processing is needed. Then I have used K-means clustering to divide users in seven different clusters based upon their visiting frequencies. Then to assign test user to respective clusters, decision tree classification is used. Then on that cluster, cosine distance metric is used to find users who have similar interests. After that, system will suggest test user three destinations, which user has not visited previously.

➢ Introduction

As the advancement of ever-growing technology, people are more involved with different websites and application for searching and exploring their desired interest. One of the top web activities, now a days, is looking for travel related information and services. And there is fast growth in the developing these website and apps that support a user in the selection of a destination or a travel service.

Planning is a complex problem solving and time-consuming activity. It involves the creation and deciding many activities such as destination selection, bookings etc. With this digital world, we have digital footprints of the people in the form of their check-ins at different location. Now we don't have to ask people or travelling agent about the destination to decide the best option possible.

So, to make recommendation digital, destination recommendation system is introduced and designed. For now, Foursquare dataset is used which contains check-ins of the user in New York City.

➢ All About Recommendation

Recommendation System: A recommender system or a recommendation system (sometimes replacing "system" with a synonym such as platform or engine) is a subclass of information filtering system that seeks to predict the "rating" or "preference" a user would give to an item.

Recommender systems are utilized in a variety of areas including movies, music, news, books, research articles, search queries, social tags, and products in general. There are also recommender systems for experts, collaborators, jokes, restaurants, garments, financial services, life insurance, romantic partners (online dating), and Twitter pages.

General Requirements for Recommendation Systems: To make a practical recommendation system, three things are required:

Background Information – information about the categories that system needs to begin the recommendation system

Input Information – information from the user so that recommendation system starts working.

An Algorithm – which will take the above two information and gives a suggestion to the user.

➢ Dataset Files

Dataset Source: https://archive.org/details/201309_foursquare_dataset_umn

Initial Foursquare dataset contained 227,428 check-ins from 1,083 users in the New York City area. Each of these check-in rows contained information about the user's ID, the venue ID, the venue category ID, the venue category name, the location of the venue (latitude and longitude) time of the check-in. To improve performance of recommendation system, we have removed irrelevant columns from dataset. After all the preprocessing, we left with the below files.

 # Processed_original_data.csv
The original data, but each record now contains the following attributes:
User ID
Venue ID
Venue sub Category ID
Venue sub category name
Venue main category (added by us)

 # Processed data 1 user
ID Subcategories.txt
List of check-ins
for each of the 1083 users in the 9 main categories of venues.
Used for computing similarity measures.

 # Processed data 2 user

ID Categories.csv
List of check-ins
for each of the 1083 users in the 9 main categories of venues.
Used for computing clusters for data.

 # Categories.csv
A list of the 9 main categories provided by foursquare, used while asking user what category of place he would like to visit.

➢ Data Analysis

Distribution of data across nine categories:
1. Summary:

```
> summary(grouped_data)
       X          Arts...Entertainment College...University      Food       Nightlife.Spot
 Min.   :   1.0   Min.   :  0.000       Min.   :  0.000       Min.   :  0.0   Min.   :  0.00
 1st Qu.: 271.5   1st Qu.:  2.000       1st Qu.:  0.000       1st Qu.: 22.0   1st Qu.:  2.00
 Median : 542.0   Median :  4.000       Median :  2.000       Median : 36.0   Median :  8.00
 Mean   : 542.0   Mean   :  7.755       Mean   :  7.527       Mean   : 45.4   Mean   : 15.63
 3rd Qu.: 812.5   3rd Qu.:  9.000       3rd Qu.:  6.000       3rd Qu.: 55.0   3rd Qu.: 21.00
 Max.   :1083.0   Max.   :181.000       Max.   :256.000       Max.   :804.0   Max.   :378.00
 Outdoors...Recreation Professional...Other.Places   Residence    Shop...Service   Travel...Transport
 Min.   :  0.00        Min.   :  0.00               Min.   :  0.0   Min.   :  0.00   Min.   :   0.00
 1st Qu.:  3.00        1st Qu.:  6.00               1st Qu.:  0.0   1st Qu.: 13.00   1st Qu.:   4.00
 Median :  8.00        Median : 15.00               Median :  3.0   Median : 25.00   Median :  10.00
 Mean   : 18.93        Mean   : 25.98               Mean   : 18.3   Mean   : 37.04   Mean   :  29.74
 3rd Qu.: 17.00        3rd Qu.: 35.00               3rd Qu.: 22.0   3rd Qu.: 49.00   3rd Qu.:  25.00
 Max.   :408.00        Max.   :393.00               Max.   :395.0   Max.   :566.00   Max.   :1119.00
>
```

Figure 1. Summary of Data

Analysis: The box plot shows that measure means mean will be influenced by the large number of check-ins by the few users and the rest will concentrated close to zero value, for these 9 main categories.

To improve clustering performance, we wanted to just have a measure to determine if the user is a "frequent" or "infrequent" visitor of a certain category of venues.

2. Scatter Plot

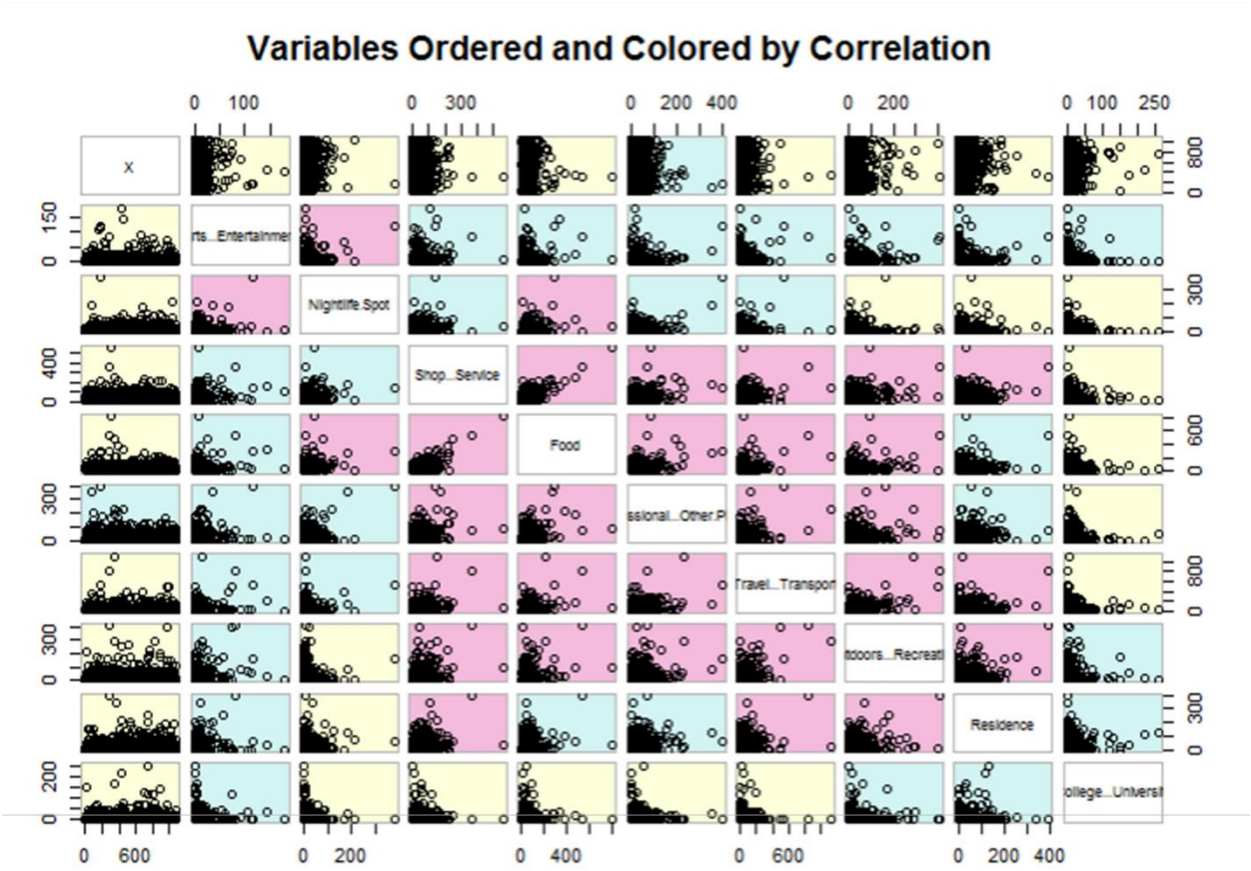

Figure 2. Scatter Plot

Analysis: The above plot shows that the range of data was not ideal for clustering algorithm as most of the data is clustered around just one point.

Solution: Mean measure is not effective because largely data is around origin. So Trimmed mean should be used to solve this problem.

A truncated mean or trimmed mean is a statistical measure of central tendency, much like the mean and median. It involves the calculation of the mean after discarding given parts of a probability distribution or sample at the high and low end, and typically discarding an equal amount of both. This number of points to be discarded is usually given as a percentage of the total number of points but may also be given as a fixed number of points.

It is calculated including 10 percentiles to 90 percentiles of the given data. According to this, all user whose check-in are below the mean is set as 0 and above the mean is set as 1.

Also, venue name is not used in the dataset. Venue IDs are present in dataset which are hash values. And to convert hash values and display the name, I have used additional Foursquare APIs.

➢ System Design

System Elements

System Elements	Details
Designing Tool	Python Spyder IDE
Programming Language	Python
Dataset Format	CSV files

➢ High Level Design

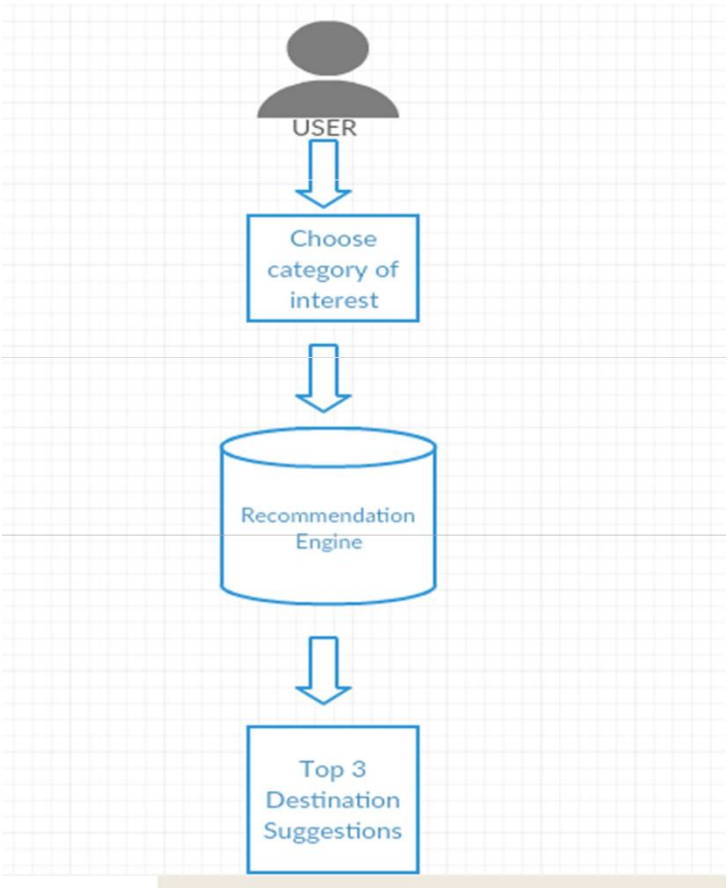

Figure 3. Destination Recommendation System

➢ Algorithms

➢ Clustering

I. SSE- Sum of Squared Error

It is difficult to find number of clusters needed to divide the users when forming user clusters. To determine the correct number of clusters, sum of squared error is used. After modeling the relationship, best relationship between these variables is with 7 clusters.

The squared error is defined as the sum of the squared Euclidean distances between each element in the cluster and the cluster centroid C_k.

The squared error is defined as

$$se_{K_i} = \sum_{i-1}^{m} \| t_{ij} - C_k \|^2$$

Given a set of clusters $K = \{K_1, K_2, \ldots K_n\}$, the squared error for K is defined as

$$se_K = \sum_{j=1}^{\kappa} se_{K_j}$$

II. K-Means

After trying different clustering algorithm, K-means was the most effective algorithm among others. K-means was most suitable for clustering similar users based on the categories of check-ins. Clusters formed by K-means were non-hierarchical and did not overlap with one another. One of the advantages of K-means was fast computation with less time complexity.

The cluster mean of $K_i = \{t_{i1}, t_{i2}, \ldots, t_{im}\}$ is defined as

$$m_i = \frac{1}{m} \sum_{j=1}^{m} t_{ij}$$

➢ Classification

Once user clusters were ready, cluster number were added as additional attributes to the user data. I compared different classification algorithm to find most accurate algorithm.

Accuracies of algorithm are as follow:
Knn → 96.5% to 60.4% with k =1 to 100 respectively.
Decision tree → 98.12%
Random forest → 97.52%

Based on above data, Decision tree classification algorithm was used.

Decision Tree:

With this approach, a tree is constructed to model the classification process. The 2 basic steps of this algorithm are: building the tree and applying the tree to the dataset. A decision tree classifier was built using the cluster number as the class labels for classification. This model assigns a class label to a test user which is the same as the class label for the most similar users. This will be helpful in reducing the search space when searching for similar users.

➢ Recommendation System:

Finally, to enhance decision making capabilities, recommendation system is used. This will help us to choose from selecting a movie to a book or to any products. The inputs to recommendation system are the user number and the categories of venues that user would like a recommendation for.

A list of all the venues in the dataset is checked sequentially and a frequency value is associated with each venue. This frequency value is defined as the count of visits to that venue in the original check in dataset. With the help of the decision tree created from the training dataset, the test user is first classified to identify the user group that the test user belongs to. Cosine similarity measure is used to identify users that are most similar to the test user.

The list of venues in the category that has been requested by test user is then checked sequentially in order to recommend the user. This list contains the venues visited by the similar users. Based on the frequencies of visits, the top venues from this list are then extracted for recommendation.

It is possible that the list of recommendations has less than 5 venues that is if the test user went to all the venues in that category that the similar users went to or if any of the similar users have not been to any venue in that category. In such case, first, search visits the original list of venues and then the top venue which is not in the list of recommendation as well as not visited by the test user are then added to the list of the recommendations.

At the end of this process, the list of recommendations has venue IDs which represent venues in the category chosen by the user. Each of the venue IDs are used to make three

calls to the Foursquare API. The API resolves the venue IDs and returns the details which are then displayed to the user.

Output:

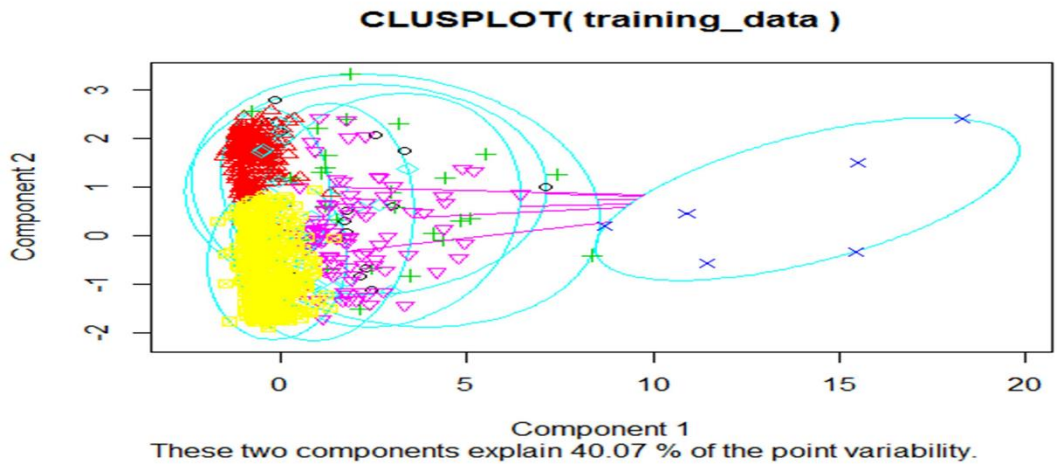

Figure 4. Cluster Plot for training data

groupedData_subset...1.	usersim	User
31	372 0.8016920	111
37	443 0.7384163	111
46	492 0.7186708	111
33	390 0.7111000	111
71	699 0.6848015	111
395	664 0.8996019	110
232	325 0.8576855	110
328	502 0.8499531	110
278	402 0.8451077	110
350	554 0.8447299	110
50	525 0.8785941	109
45	508 0.8774934	109
92	661 0.8634505	109
235	971 0.8612571	109
171	843 0.8335300	109

Figure 5. Cosine similarities of top users

```
Login ID: -
111

Enter Category Number from the list

1:  Shop & Service
2:  Outdoors & Recreation
3:  Residence
4:  Professional & Other Places
5:  Food
6:  Travel & Transport
7:  Arts & Entertainment
8:  College & University
9:  Nightlife Spot
10: Athletic & Sport
Category Number:4
Top recommendations based on User History:

[1] VaynerMedia HQ
[2] MWW
[3] MWW
[4] Jacob K. Javits Convention Center
[5] Foursquare HQ
[6] Onswipe HQ
[7] Canvas
[8] SocialChorus
[9] Milk Studios
[10] New York City Civil Court
[11] Forrest Solutions
[12] Techstars HQ
[13] Carat North America
[14] New York State DMV
[15] Spark Capital
[16] Onswipe: Publishing & Advertising in the Post-PC Era
[17] 190 East 7th Rooftop
[18] Safeguard Self Storage
[19] Queen Of Peace High School
[20] Canaan Partners
```

Figure 6. Recommendation of venues

➢ Conclusion:

Recommendation systems are very useful for users which provide personalized content and services by filtering, prioritizing and efficiently delivering relevant information in order to alleviate the problem of information overload.

There is significant progress in the research community, and many experts are trying to bring the benefits of new techniques to the end users. But there are still important gaps that make personalization and adaptation difficult for users.

That's why, I have implemented a recommendation system based on clustering and classification to help users discover and explore new places of interest in different categories based on similar users' habit. It predicts the user's likeliness to visit a place that user has never visited.

K-means clustering algorithm was employed as it practically works well and gives more accurate results than other algorithms. The clustered produced are easily interpretable. It has computational cost of O(K*n*d).

For classification using the decision tree, cosine similarity measure is used to identify three users that are most like the test user and a list containing the venues visited by the similar users is generated, out of which the top three venues are then extracted for recommendation. Advantages of the decision tree model lie is its transparent nature, specificity, and comprehensiveness. The decision tree makes explicit all possible alternatives and traces each alternative to its conclusion, allowing for easy comparison. Its ability to assign specific values to problem, decisions, and outcomes of each decision reduces ambiguity in decision-making. A decision tree also allows for partitioning data in a much deeper level, not as easily achieved with other decision-making classifiers such as logistic regression or support of vector machines.

Finally, Foursquare API is used to reference the venue IDs and is made to retrieve the details that is displayed to the user. This will show top recommended places in the selected category.

➢ Future Work

If history of results could be stored in database for each user, this will allow us to build context for future prediction. This will improve quality of result and will dynamically update the result set for every subsequent prediction.

Since we are living in ever growing technology world, where every day new algorithms are getting implemented. So, I am planning to use different algorithm on different user segment, to make system more effective and accurate.

Another extension is to assign users to several overlapping clusters by using an algorithm other than k means for clustering and use the opinion of these clusters to generate recommendations. Overlapping clusters could depict the real-world situations where users participate in different communities. That would further improve our recommendation system.

➢ References

1. Data Mining Introductory and Advanced Topics, Margaret H. Dunham

2. https://archive.org/details/201309_foursquare_dataset_umn

3. Mohamed Sarwat, Justin J. Levandoski, Ahmed Eldawy, and Mohamed F. Mokbel. LARS*: A Scalable and Efficient Location-Aware Recommender System, IEEE Transactions on Knowledge and Data Engineering TKDE

4. https://en.wikipedia.org/wiki/Recommender_system

5. https://web.stanford.edu/class/cs345a/slides/12-clustering.pdf

6. https://en.wikipedia.org/wiki/Decision_tree_learning

7. http://infolab.stanford.edu/~ullman/mmds/ch9.pdf

8. https://www.udemy.com/share/100034A0UYcV1VRXo=