

Song Recommendation System Using “R”

Mounik Chinthapanti
MS in CS
mchinha@uemail.iu.edu
Indiana University Bloomington

Bhavik Thakkar
MS in CS
thakkarb@uemail.iu.edu
Indiana University Bloomington

ABSTRACT

In recent years, data mining along with its algorithms has grown to a great extent. So, in this paper we are constructing a recommendation system with the help of data mining concepts. Recommendations have become an important part in analyzing what the user is going to do next. There have been various fields in which recommendation system is of prime importance. So, in this paper, the prime focus is on recommending the user what song he/she will listen to based on their past data. This paper proposes user based collaborative filtering along with item based collaborative filtering and have even made use of cosine similarity for calculating the similarity measures. From the results, it is observed that the proposed method gives the appropriate recommendation based on the dataset which we have provided as an input to the system.

Keywords

User based collaborative filtering, Item based collaborative filtering, Cosine Similarity.

1. INTRODUCTION

Application of Data Mining in various fields has gained significance importance over the past few years. Various algorithms in data mining have increased the accuracy of prediction, diagnoses, detection and treatment providing all the fields with the state of the art software and up to date systems. Recommending systems have gained prime importance with the help of Machine Learning and data Mining Algorithms. Hence a collaborative recommendation system for song/band prediction solution has been proposed in this report.

Let us explain how recommendation system works. The basic idea here is that, if there are two users who share or have the same kind of interests, then there is a high chance or probability of those two users sharing the same choice in the future as well. For example, let us assume that one user likes a certain type of fruit and the second user too likes the same fruit. If the first user likes another type of fruit, then there is high chance of the second user to have similar set of choices as the first user than some other person or user. Therefore, the implications of using collaborative filtering is that you can predict the items first and then recommend those items to the users based on their similarity preferences. Here, in our project, we have the similarity preferences stored in the ‘lastfm-matrix-germany’ which we obtained it from an online source.

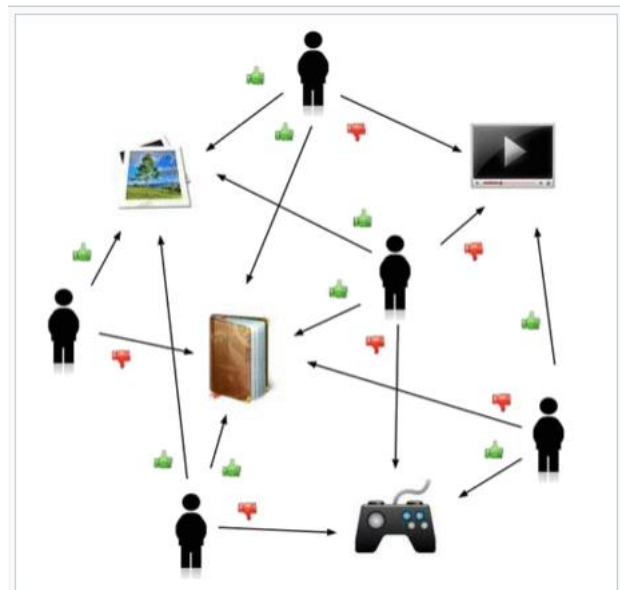


Fig. 1: System making predictions about user's rating of different items based on their likes and dislikes

Item Based Collaborative Filtering works in such a way that it takes similarities between the items which have been consumed. User Based Collaborative Filtering works in a way that it takes similarities between the consumption history of the users.

Collaborative filtering (CF) is a technique used by recommender systems. Collaborative filtering has two senses, a narrow one and a more general one. In the newer, narrower sense, collaborative filtering is a method of making automatic predictions (filtering) about the interests of a user by collecting preferences or taste information from many users (collaborating). The underlying assumption of the collaborative filtering approach is that if a person A has the same opinion as a person B on an issue, A is more likely to have B's opinion on a different issue than that of a randomly chosen person. [1].


























				
				
				
				
				
				

Fig. 2: the above image shows an example where the user's ratings are predicted with the help of Collaborative Filtering technique.

2. LASTFM-MATRIX GERMANY DATASET

The dataset which we have used is publicly available and in order to keep this recommendation system simple, we have made use of a comparatively smaller dataset which consists of 1250 rows and 285 columns.

The rows are the number of different users and columns consist of all the various bands and songs. We extracted a small dataset out of the originally very large dataset which contained about 20000 rows. You can find our attached dataset in the project zip folder which we have submitted and the file has been named as 'lastfm-matrix-germany.csv'. The dataset basically consists features which include details about the user such as the user name, their gender, their age, the artist which they like listening to, the artist which they have listed to previously,

3. ITEM-BASED COLLABORATIVE FILTERING

Item based collaborative filtering or item-based or item-to-item is a form of collaborative filtering for recommendation systems based on the similarity between items calculated using people's rating of those items. [2].

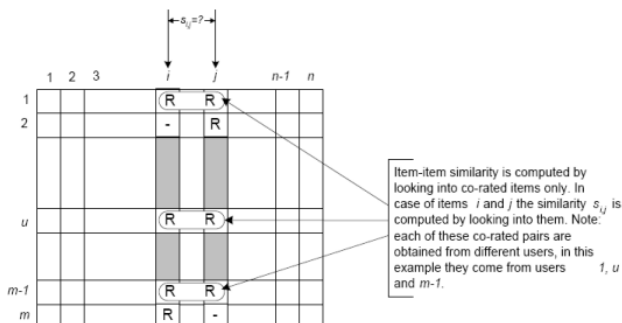


Fig. 3: This image shows the general idea behind item-based collaborative filtering.

In this type of filtering, the users are not given much importance. So, it can be conducive if we do not even consider the user column here. So, we can easily drop the user column data. In our project, we have then calculated the similarity measure of each song with every other song. This is done with the help of Cosine Similarity and we have made use of a cosine similarity function in R. So, how does Cosine similarity work? It first takes the sum of the product of the first column as well as the second column. This sum is then divided by the product which is obtained after taking the square root of the sum of the squares which are present in all the columns. The output obtained is a number which represents the similarity between the first column and the second column. The main method with all the steps in a proper flow will be explained in the proposed method section.

4. USER-BASED COLLABORATIVE FILTERING

This approach uses user rating data to compute the similarity between users or items. This is used for making recommendations. This was an early approach used in many commercial systems. It's effective and easy to implement. [3].

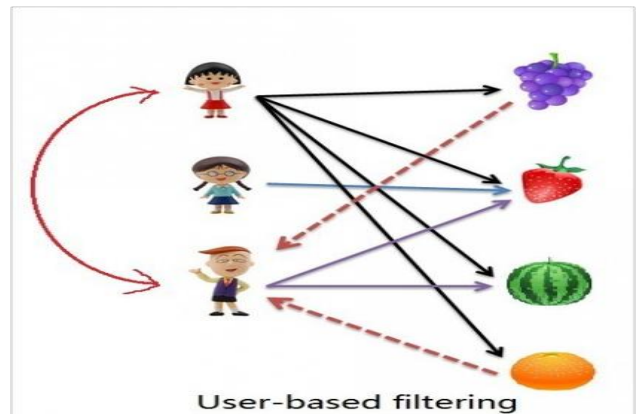


Fig. 4: The above animated image perfectly describes how a user-based filtering takes place.

This type of filtering method makes use of vector model which is based on similarity measures. Similarity matrix is made use of generally in user based recommendation system. It calculates the similarity between two users and then a prediction can be estimated by calculating the weighted average of all the ratings which were recorded. It works in such a way that after suppose n most similar users are found, a user matrix corresponding to them is generated which helps us in identifying the items which are to be recommended.

5. HIERARCHICAL CLUSTERING

In data mining and statistics, hierarchical clustering (also called hierarchical cluster analysis or HCA) is a method of cluster analysis which seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types [1]

Agglomerative: This is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.

Divisive: This is a "top down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

In general, the merges and splits are determined in a greedy manner. The results of hierarchical clustering are usually presented in a dendrogram

6. IMPLEMENTATION OF ITEM BASED COLLABORATIVE FILTERING

Step 1: First the dataset is taken as input into R and then converted into a matrix format where the first column contains user information represented as a number and the first row contains names of different songs.

Step 2: Then the cosine similarity matrix is calculated with the dataset as input matrix. This gives a measure of similarity between two non-zero vectors of an inner product space that measures the **cosine** of the angle between them

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|_2 \|\mathbf{B}\|_2} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}$$

Step 3: After calculating the similarity scores, similarity matrix for the songs are sorted so that we have the most similar first. Take the top 11 songs in the decreasing order of their similarity scores (first element will always be the same artist) and put them in a new matrix. `t()` is used to rotate the similarity matrix since the neighbour one is shaped differently.

Step 4: The above procedure is repeated for each and every song, and a recommendation matrix is constructed. This matrix actually recommends the next song if he is listening to the song which is the row name of it.

Step 5: After calculating the recommendation matrix, the distance between each of the binary datapoints is calculated using `dist` function and method used to find this distance is binary. In this method, vectors are regarded as binary bits, so non-zero elements are 'on' and zero elements are 'off'. The distance is the *proportion* of bits in which only one is on amongst those in which at least one is on

Step 6: After calculating the distance, using `hclust` function with five different methods, five different dendrograms are plotted. (methods used : Complete, Single, Ward ,Average, Centroid)

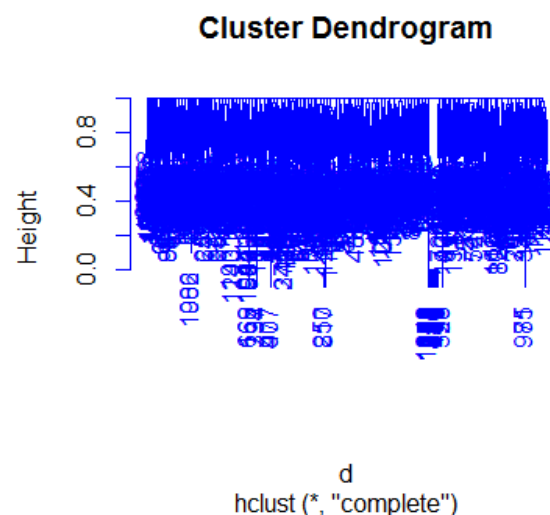


Fig. 5: This image shows the dendrogram plotted using complete linkage method in hierarchical clustering

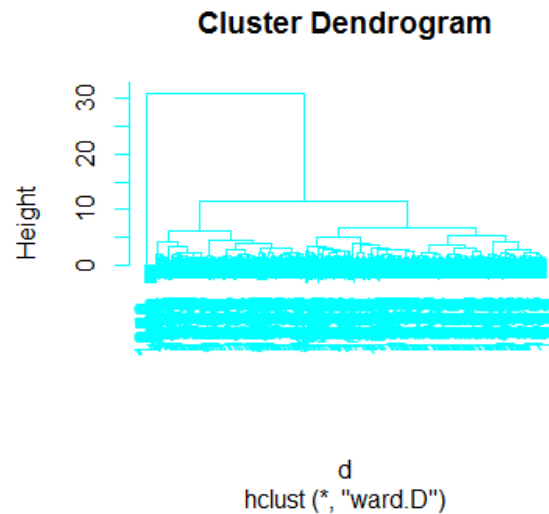


Fig. 6: This image shows the dendrogram plotted using ward distance linkage method in hierarchical clustering

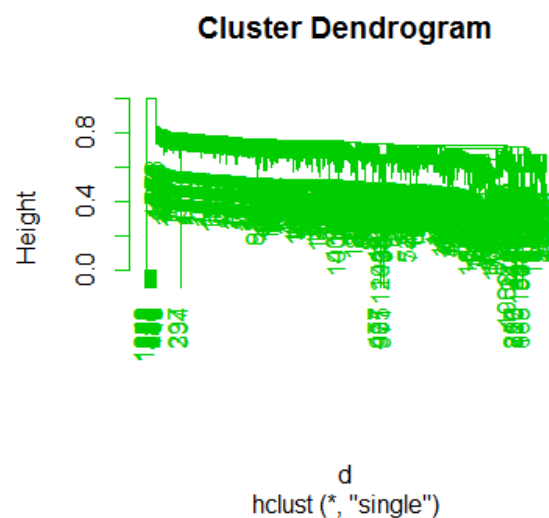


Fig. 7: This image shows the dendrogram plotted using single linkage method in hierarchical clustering

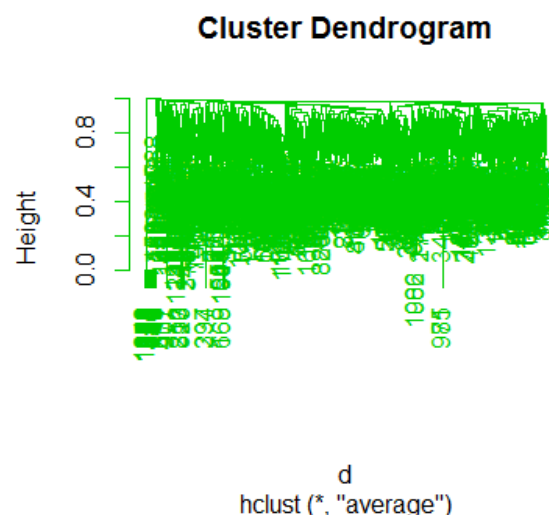


Fig. 8: This image shows the dendrogram plotted using average linkage method in hierarchical clustering

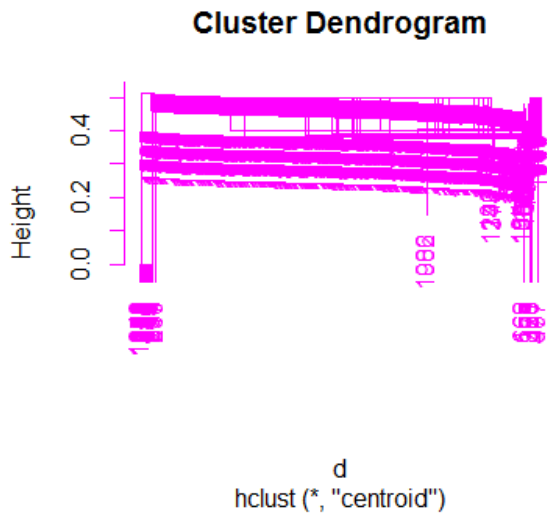


Fig. 9: This image shows the dendrogram plotted using centroid linkage method in hierarchical clustering

Step 7:After performing hierarchical clustering on data ,cut the tree using the cutree function which cuts a tree, e.g., as resulting from `hclust` , into several groups either by specifying the desired number(s) of groups or the cut height.After cutting the dendrogram based on number of clusters the different clusters are plotted as follows where the y-axis represents the cluster number and x-axis represents the index.

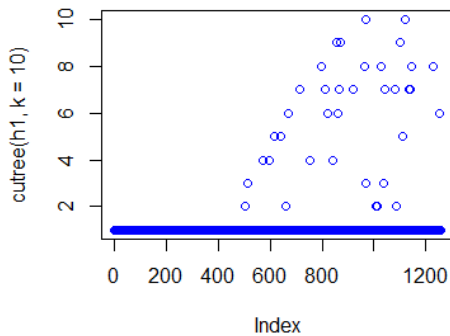


Fig. 10: This image shows 10 different clusters using complete linkage hierarchical clustering

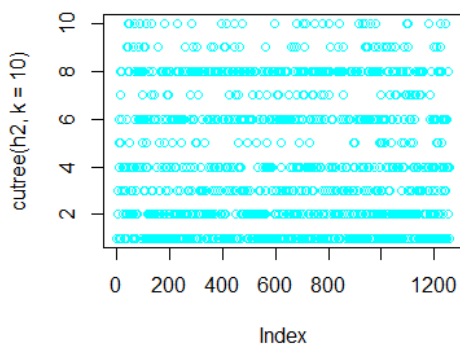


Fig. 11: This image shows 10 different clusters using ward linkage hierarchical clustering

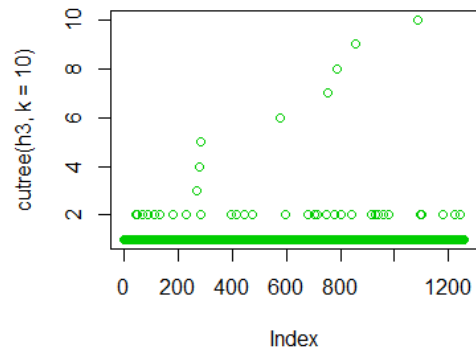


Fig. 12: This image shows 10 different clusters using single linkage hierarchical clustering

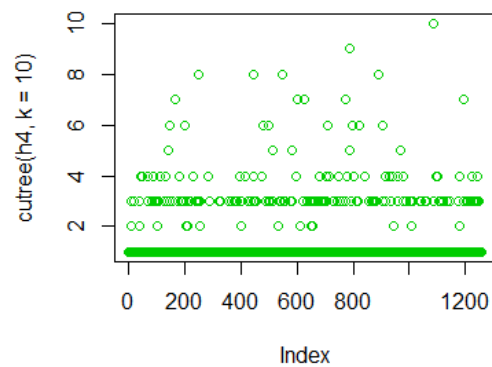


Fig. 13: This image shows 10 different clusters using average linkage hierarchical clustering

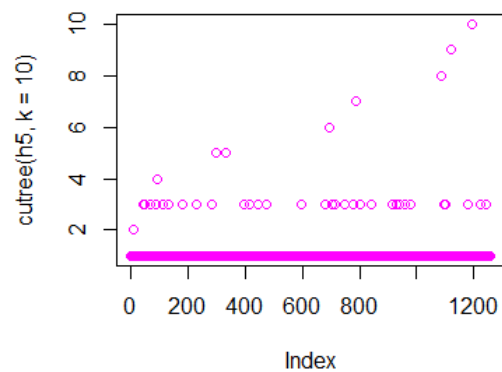


Fig. 14: This image shows 10 different clusters using centroid linkage hierarchical clustering

7. IMPLEMENTATION OF USER BASED COLLABORATIVE FILTERING

Step 1: Like the item based implementation import the dataset into R and calculate the similarity matrix using cosine similarity.

Step 2: Then loop through each user and each song and check whether the user has listened to that song or not. If yes place a empty string in the recommendation matrix, else get the next similar 10 songs based on the similarity matrix and store the similarities scores and song names.

Step 3: Find out whether the user has listened to these next similar 10 songs or not then filter out songs that match the user.

Step 4: After the loop execution, the scores for each user and each song are obtained .

Step 5: After obtaining the scores matrix, process the data and for each song, in this implementation we recommend its next 100 songs in the decreasing order of their scores.

8. RESULTS & CONCLUSION

The Clustering result with different plots have already been attached in the above explanation of our entire process behind this recommendation system which includes item based collaborative filtering, user based collaborative filtering as well as hierarchical clustering techniques along with a dendrogram representation. All the R files have been included in the upload with the recommendation results in the form of .csv files.

9. REFERENCES

[1] https://en.wikipedia.org/wiki/Collaborative_filtering

[2]https://en.wikipedia.org/wiki/Item-item_collaborative_filtering

[3]https://en.wikipedia.org/wiki/Collaborative_filtering#Memory-based