International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 04 | Apr 2022 www.irjet.net p-ISSN: 2395-0072

Overview of Movie Recommendation System using Machine learning by R programming Concepts

Mallareddy Sai Prakash[1], Harwant Singh Arri[2]

[1]*Student, Department of Computer Science and Engineering, Lovely Professional University, India.*
[2]*Assistant Professor, Department of Computer Science and Engineering, Lovely Professional University, India.*

Abstract - *Now a days, in our daily lives, the usage of internet, along with that usage of the several apps is also increased. In that apps, several software's are developed to keep the user engaged every time. In those software's, Recommendation System is one of the main software. This system is used in apps like food, delivery, shopping, over the top (ott) platforms and so on to keep the user endorse to them. This paper represents the overview of Recommendation system using Collaborative filtering approach. This paper elaborates the technique and key concepts involved in this approach.*

Keywords----Recommendation system, IBCF, Collaborative filtering, Recommenderlab, K-means clustering, Normalization.

I. Introduction

Recommendation system is a system which furnishes its users with numerous contents formed on their tastes, preferences and attachments. Machine learning algorithms will be used to implement this recommendation system.

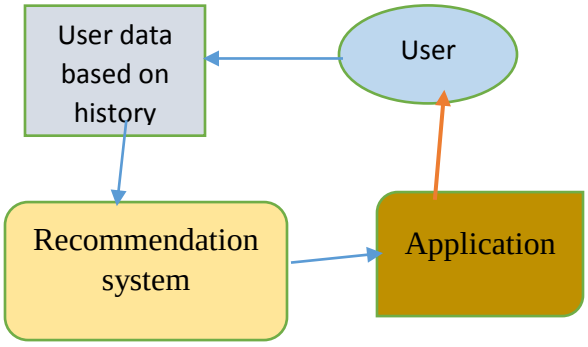
Recommendation system dispenses recommendations to its users by a filtering process which was based on user likings and history which they recently gone through or by searching. This data about the user will be considered as an input. This data which collected reflects the prior usage of the application or product as well as the assigned reviews or ratings expressed with the users.

Another consideration with the recommendation system is it finds a similar connection between the various products. Let's consider example like Prime, Aha, Netflix, and Zee5 platforms provides one user with the suggestions of the movies that are similar to the other users that they have watched previously or one searched.

Recommendation system are there of two types which are Content based recommendation system and Collaborative filtering recommendation system.

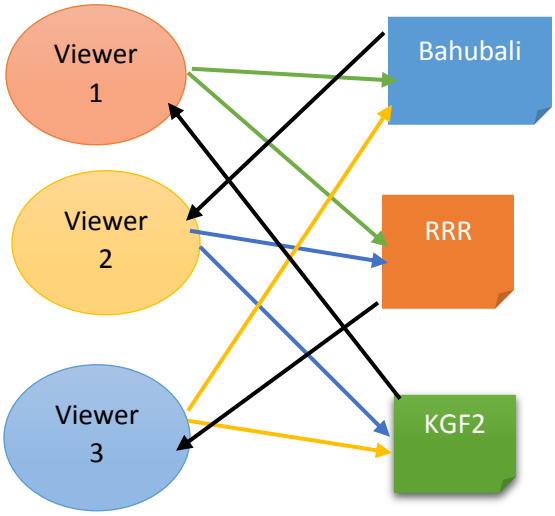
Content based recommendation system: Content based recommendation system approach is it works with user data which was collected either by ratings or by collecting from the activities they do. This approach examines some data related to user which was collected previously from user based on preferences, ratings or from his interests. Then, the recommendation system tries to match the content collected from user to content which was available. And then, it suggests the content to user. By this way, this recommendation system recommends several things to the users by using content.

The below diagram explains the working of content based recommendation system:



Collaborative filtering recommendation system: Collaborative filtering recommendation system is process which involved collaboration of data. This filtering works by collecting their habitual works, ratings and genres by several users from the movies they watch, from the things they do and draws the similarity between the user's habits, reviews and genres they prefer most. By this way, recommendation system finds the like-minded users and recommends them the things / movies which are most liked among them.

The below diagram explains the working of collaborative filtering recommendation system:



Black line denotes Recommendations done to Viewers.

II. Key Concepts:

Recommenderlab package: Recommender lab is a package which furnishes the framework to test and develop recommender algorithms.

- It helps in creating recommendations for given data base which directly extracts ratings such as ratings from 1 to 5 in established area and evaluation environment.
- This Package provides algorithms and helps the user to design and execute their own algorithms in the framework through an uncomplicated procedure.
- In R programming, Recommender lab is coined as 'recommenderlab'.

GGPlot2 package: Ggplot2 is a package of plotting that designed for plotting the representations from data.

- GgPlot2 is a package that designed for plotting commands to design complicated plots from the data in a data frame.
- This package has the graphics which boosts layer after layer by adding brand latest and crisp features to it.
- This package plays a crucial role even by reducing the work for changing from one plot to other by making lesser amount of adjustments.
- These layers helps for customizing the plots as per the user needs by using ample flexibility it provides.
- In this package, we can the plot the data using ggplot function.
- Syntax of ggplot:

  Ggplot (data, color, aes(x, y)).

Data. Table package: Data. Table package is a package where it provides to work with the tabular form of data.

- Considerably, Data. Table package is a substitute package of R programming in-built Data. Frame package.
- Main reason of this package being widely used is because of its speed. It handles the large amount of easily and responds in quick time.
- This package is also known fast reader package because of fast accumulation of large datasets.

Reshape2 package: Reshape2 is a package of R written by Hadley Wickham. Reshape2 package was designed to allow one to transform the data easily into different types of structures as one required.

- Reshape2 is one of the package which uses for manipulating the data along with tidyr package to convert wider format data to the long format data.
- Reshape2 is a very much fast version of reshape package. It was also more memory efficient.
- In this new package, cast function is remodeled by two new functions which are 'dcast' and 'acast'.
- Dcast is for producing the data frames, whereas acast for producing the arrays or matrices.
- Melting and Casting are two important functions in reshape package. Melting uses to

stretch the data in data frame by converting it into long format. Casting uses to convert long format data to aggregated form data.

- Syntax of Melt and Cast functions:
  Melt (data, na.rm, value.name)
  Cast (data, formula, fun. Aggregate)

Tstrsplit function: Tstrsplit function is literally a combination function of Transpose and strsplit functions.

- As, we can consider as Tstrsplit(x) is Transpose (strsplit(x)).
- Tstrsplit is an appropriate wrapper function which uses the transpose to the function of splitting a string.
- Syntax of Tstrsplit function:
  Tstrsplit (data, fill, type. Convert)

Recommender Registry: Recommender registry is a registry where it provides the registry to manage methods from the Registry package.

- This Recommender registry helps users to add and specify new methods to the user.
- This Recommender registry is from Recommender lab package.
- Recommender registry$get_entries () is a function which was useful for getting the entries from the registry of recommender. This is also a part of recommender lab package.
- Real Rating matrix is a matrix designed as it containing matrix of ratings of user or item based which are rated with certain ratings or stars.

Lapply function: Lapply is an inbuilt function in R language. Lapply function applies for list objects and returns same length of list objects.

- In Lapply, L means list, it returns output in only list.
- Lapply function takes inputs of Vector, list or data frame and return only list as output.
- Syntax of Lapply function:

  Lapply(x, fun)

  Where x is vector or list or data frame.

Sapply function: Sapply function is an inbuilt function in R language. Sapply function takes all the inputs like vector, list or data frame as an input argument and returns a matrix or vector.

- Sapply is a Wrapper class to Lapply in R where only difference is it returns in vector.
- Syntax of Sapply function:
  Sapply(x, fun)
  Where x can be vector or list or data frame.

Item based Collaborative filtering (IBCF): Item based collaborative filtering is one of the filtering methods in the recommendation system. This collaborative filtering will search for similar data based on the items which user have liked or interacted often will be suggested.

- Item based collaborative filtering is a process where two or more users watching similar kind of movies like same genre, theme, language, actor and so on then the movies which watched by them will be recommended to each one of them.
- This approach is developed by Amazon in 1998 and from then it has been pivotal in many recommendation systems.

Similarity & Cosine functions:

- Similarity is the function which helps to find the similarities between the any documents and features.
- Cosine similarity is the method of finding the similarity measure between any two vectors in an inner product space.
- Formula for Cosine similarity calculation is " $\sum A_i B_i / (\sqrt{\sum A_i^2}\ \sqrt{\sum B_i^2})$"
- Syntax of Cosine function:
  Cosine (Data1, Data2)
- If there are more than 2 data to find the cosine we use cbind to datasets and then we will find the Cosine function.
- To find Cosine for a numeric value, we will use cos() function. As, we know Cos is an inbuilt mathematic function in R.

Geom_bar (): Geom_bar is a function used to create or draw a bar graph in R studio. Geom_bar is a package that comes under GGplot2 package.

- Syntax of Geom_bar():
  Geom_bar (data, col)
- Here the term Geom denotes that we want to create the bar plot.
- Geom_text () is used to add the text to the representation.
- Geom_label () is used to add the labels to the representation. So, that makes the user to identify the exact outcome from the representation.

Heat Map: Heap Map is a data visualization technique that produces the graphical representation of data. This Heat Map represents the values in a matrix as colors in the visual diagram.

- Shading matrix is another title of the heat map. In this heat map, higher activity is represented by using brighter colors and lesser activity is by using darker colors.
- Syntax of heat map:
  Heatmap (data, col)
  Here data must be in rows and columns.

Quantile: Quantile is a generic function. Quantile is a function used to generate/ create sample quantiles with probability between 0 and 1 within the data present.

- Syntax of Quantile function:
  Quantile (data, probs, na.rm)
- Here probs means the probabilities.

Qplot (): The function Qplot comes under the package of GGplot2. Qplot is the function similar to "Plot" function. This qplot() is used to create and combine different types of plots.

- Syntax of qplot():
  Qplot (data1, data2, geom)
- Here geom indicates the type of representation.

Normalization: Data Normalization is the one of the best methods in Data Science. Normalization makes sure that the data in customer database is well organized and can accessed in the same kind of way across all the records in the database.

- This normalization is taken forward by standardizing the specific fields and records by transforming the formats within the database.
- Normalize() is used to perform this normalization.
- Syntax of Normalize():
  Normalize(data, method, range, margin)
- Normalization is one best way to change a bad or average machine learning model to good learning model by normalizing the data.
- For suppose, if we do not normalize the data, then that makes the data clumsy which definitely effects the model performance.

Binarize function: The Binarize function is a function which collects data in numerical and categorical form and returns the binary data.

- This binarize function is part of a preparatory step of Correlate function.
- This binarize function converts normal form of matrix into binary form of matrix.
- Syntax of binarize function:
  binarize(data, threshold)
  Here threshold is default by NA which means median is considered as threshold.

KNN Algorithm: KNN algorithm means K Nearest Neighbors algorithm. K Nearest Neighbor algorithm is a supervised machine learning algorithm used for classification and regression concepts.

- KNN algorithm assumes the similarity among the new data and accessible cases. And compare, replace the case which is most similar to the available data categories.
- KNN collects all data and classifies new point 'K' based upon similarity. Then it is classified into another category using the algorithm.
- KNN is known as Lazy Learner algorithm. Because it doesn't learn from the data.

- Instead it stores the data and performs the task of classifying.

- Steps of KNN algorithm:
  Step1: Firstly, we have to select value 'K', number of neighbors.
  Step2: By using distance formula, we have to find 'K' nearest neighbors.
  Step3: As, we know distance formula is ""$\sqrt{(X_2 - X_1)^2 + (Y_2-Y_1)^2}$.
  Step4: In each category, count the number of neighbors.
  Step5: Category which has maximum neighbors assign the new data point.

- By these steps, implementation of KNN algorithm is completed.

K- Means Clustering Algorithm: K-Means is an unsupervised machine learning algorithm used in data science and machine learning, groups the unlabeled data into different clusters.

- Main aim of this algorithm is minimizing the sum of distances b/w two data points among parallel clusters.

- In K – Means algorithm, clusters are associated with a centroid, hence it is a centroid based algorithm.

- In this algorithm, K indicates the count of pre-defined clusters that user want to made, as if K= 5, then there will be 5 clusters as it has similar properties within the clusters.

- K- Means algorithm steps:
  Step1: Choosing the number K to decide the number of clusters.
  Step2: Choose random K centroids.
  Step3: Allocating each data point to their nearest centroid, which create the pre-defined K clusters.
  Step4: Calculating the variance and putting a new centroid for each new cluster.
  Step5: Repeat the third step, for newly created clusters.
  Step6: If reallocation happens, then repeat the fourth step, else finish the steps.

- By these steps K- Means clustering implementation is completed.

Train & Test: This Train and test is a procedure demands considering a dataset and then divide them into two subsets.

- Training dataset: The 1st subset which divided was machine learning model which used to be fit in process.

- Testing dataset: The 2nd subset was used in the process to evaluate the fit machine learning, then estimations are made and start comparing with the expected values.

- Mainly, these two are used to know the accuracy of data present. It is a technique to examine the performance of the ML algorithm.

- This is used for any supervised learning algorithm and used for regression or classification.

Recommender Algorithms: There are some recommender algorithms which helps while using recommender technique.

And there some algorithms which are specified according to the ratings from 1 to 5 stars which are helpful in Recommendation systems which are:

- UBCF: User-based collaborative filtering
- IBCF: Item-based collaborative filtering
- SVD with column-mean imputation
- SVDF: Funk SVD
- ALS: Alternating Least Squares
- LIBMF: Matrix factorization with LIBMF
- AR: Association rule-based recommender
- POPULAR: Popular items
- RANDOM: Randomly chosen items for comparison
- RERECOMMEND: Re-recommend liked items
- HYBRID Recommender: Hybrid recommendation

GetModel: GetModel is a type of model which returns the representation of the data. It elaborates the kind, properties of data. This GetModel is used for analysis of the dataset which present.

- Syntax of GetModel:
  X <- data
  getModel(X)

- This gets or retrieves all the model of data from the X.

Predict Function: The Predict function (Predict()) in R language is used to estimate/predict the values based on the input data occurred from the data present.

- Predict is a generic function and it is a single method for 'lm' object. The lm method is present under the stats package in R studio.

- Syntax of predict:
  Predict (data, newdata, interval)

- Interval in prediction is a range of values which are likely to contain of a new observation for given specified things of predictions or predictors.

III. Conclusion

Recommendation system played a crucial role for many in getting succeed. Generally, as a human beings, we recommend to others or we seek recommendation from others. Like us, applications also require to recommend their features to keep the user engaged while using their application. So, this recommendation system played a crucial for the applications. There is more to explore in this recommendation system. In future, there will be eminent growth in this topic. In this paper, I have mentioned some key concepts which I used in the movie recommendation system using machine learning algorithms in R programming.

IV. References

- Geeksforgeeks.org/r-programming/
- Tutorialspoint.com/r/
- Datameter.io/r-programming
- Rdocumentation.org
- Datacarpentry.org
- Pluralsight.com/r/
- R-lang.com
- Programmingr.com
- Statology.org
- Rfunction.com
- Javatpoint.com