International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 03 | Mar 2021 www.irjet.net p-ISSN: 2395-0072

MOVIE RECOMMENDATION SYSTEM USING MACHINE LEARNING ALGORITHMS

Sandhiya G[1], Kaushika S[2], Sujithra R[3], Sathyavathi S[4]

[1-3]*Department of Information Technology, Kumaraguru College of Technology [autonomous], Coimbatore, India*
[4]*Assistant Professor(SRG), Department of Information Technology, Kumaraguru College of Technology [autonomous], Coimbatore, India*

Abstract - *Everyone loves movies no matter age, gender, race, color, or geographical location. We tend to all in the simplest way are connected to every different via this wonderful medium, nonetheless what most attention-grabbing is that the undeniable fact that however distinctive our selections and combos are in terms of picture show preference.Some individuals like genre-specific movies be it a thriller, romance, or sci-fi. Whereas others specialize in lead actors and administrators. After we take all the under consideration, it's astoundingly troublesome to generalize a movie and say that everybody would love it. However, with all that said, it's still seen that similar movies are liked by a selected part of the society. So here's whether we tend to as information scientists get play and extract the juice out of all the behavioral patterns of not solely the audience however conjointly from the films themselves. Thus, while not additional ruction let's jump right into the fundamentals of a recommendation system. This paper is planned a machine learning approach to suggest movies to the users using K-Means clustering algorithm, K Nearest neighbours algorithm and Affinity propagation clustering algorithm to recommend movies to the users.*

Key Words: Machine learning; k-means clustering algorithm; k-nearest neighbour; affinity propagation.

1. INTRODUCTION

Machine Learning is that field of study that offers computers the aptitude to find out while not being expressly programmed. ML is one of the foremost exciting technologies that one would have ever stumbled upon because it is obvious from the name, it provides the pc that creates it additional like humans: the power to learn. Machine Learning is actively getting used these days, maybe in many places than one would expect. Machine Learning is employed in net search engines, email filters to delineated spam, websites to create individualized recommendations, banking software systems to sight uncommon transactions, and much of apps on our phones like voice recognition.

Recommender systems are systems that are designed to suggest things to the user that support many alternative factors. These systems predict the foremost possible product that the users are possibly to buy and are of interest to firms like Netflix, Amazon etc. use recommender systems to assist their users to spot the right product or movies for them. The recommender system deals with an outsized volume data present by filtering the foremost necessary information supported by the information provided by a user's preference and interest. It finds out the match between user and item and imputes the similarities between users and ratings for recommendation. Both the users and the services provided have benefited from these sorts of systems, the standard and decision-making method has additionally improved through these sorts of systems.

In our project, by exploring different Machine learning algorithm such as K-Means clustering algorithm, K Nearest Neighbors algorithm and Affinity propagation clustering algorithm, we recommend top 20 movies to users based on the rating given by users to the movies

2. DATASET

The dataset (ml-latest-small) describes 5-star rating and free-text tagging activity from [Movie Lens](http://movielens.org), a film recommendation service. It contains 100004 ratings, 9125 movies and 671 users. This dataset was generated on October seventeen,2016. Users were chosen randomly for inclusion. All the chosen users had rated a minimum of 20 movies. The Movie Lens dataset primarily has two files. The primary file contains data regarding movies it's: movie id , movie name and list of its genres. The Movie Lens dataset contains a movie list of nineteen genres. The opposite file consists of: user id, movie id, ratings. These two files are pre-processed and manipulated therefore to produce our system.

3. SYSTEM DESIGN

In our system, the primary and foremost is knowledge gathering. We'll be grouping the dataset that's prepared, it ought to be pre-processed since it's real-world data. There could be a possible ton of missing data, mismatching entries etc… exploitation effective functionalities the dataset are pre-processed and provided to following steps. Now, the data is ready for applying the desired algorithmic rule, here we use K-Means clustering algorithm, K nearest Neighbors algorithm and Affinity propagation clustering algorithm. Once after applying the algorithm separately for the dataset we will predict and recommend 20 movies for the users who are yet to watch the movie:

3.1. DATA PREPROCESSING

Data Pre-processing may be a method of getting ready the data and creating it appropriate for a machine learning model. It's the primary and crucial step whereas making a machine learning model. A real-world data usually contains noises, missing values, and perhaps in an unusable format that cannot be directly used for machine learning models. Data preprocessing is needed tasks for cleansing the data and creating it appropriate for machine learning models that additionally will increase the accuracy and potency of a machine learning model. The process involves getting the dataset, importing libraries, importing datasets, finding missing data etc…

3.2. ALGORITHMS

3.2.1.K-MEANS CLUSTERING ALGORITHM

It is an unsupervised learning algorithmic program that's used to solve the cluster issues in machine learning or data science. It permits us to cluster the information into completely different groups and a convenient way to discover the classes of groups within the untagged dataset on its own without the requirement for any coaching. It's a centroid-based algorithmic program, wherever every cluster is related to a centre of mass. The most aim of this algorithmic program is to reduce the total of distance between the data point and their corresponding clusters. The algorithmic program takes the untagged dataset as input, divides the dataset into k-number if clusters, and repeats the method till it doesn't notice the most effective clusters. The value of k should be planned during this algorithm. Once after finding the optimal number of cluster, we will be plotting those cluster in a heatmap and we choose a random cluster, in that we will choose a movie which a user haven't watched yet and calculate the average score of the movie using all other users ratings for that movie, to predict whether the user will like the movie or not and finally we will recommend the most rated 20 movies to the users who are yet to watch the movie.

3.2.2. K-NEAREST NEIGHBOURS ALGORITHM

We will be using an unsupervised learning algorithm known as NearestNeighbors. Since this algorithm calculates distance between two points we will pivot our dataset into an item user matrix and an empty cell with 0. Now once a movie name is given a s input we'd like to seek out any such movie present in our dataset or not. It's not present then we have a tendency to can't recommend anything. Thus for string matching we have a tendency to use fuzzy matching, a list of recommendations are going to be generated. The function can come from the movie_id of the movie title that best matches the input string. It additionally prints all matches.

3.2.3. AFFINITY PROPAGATION CLUSTERING ALGORITHM

Affinity propagation could be a graph suppositional bunch technique developed by Frey and Dueck (2007). Affinity propagation creates clusters by sending messages between pairs of samples until convergence. It works in three ways: similarity matrix, responsibility matrix, availability matrix. Unlike a bunch of algorithms like k-means or k-medoids, affinity propagation doesn't need the quality of clusters to be determined or calculable before running the algorithmic program.

3.3. RECOMMENDATION

After applying these three algorithms, the system will recommend 20 movies to the users based on the rating given by the users and genre of the movie.

4. RESULT

The dataset used is Rate.csv and Movie.csv which is taken from MovieLens. This dataset consists of 100004 ratings given by 671 users for 9077 movies, each user rated nearly minimum of 20 movies. The below image shows the movies recommended to the users. After the result we manually check which system is more efficient for the recommendation by comparing the result.

```
avg_ratings.sort_values(ascending=False)[:20]

Out[56]: Trainspotting (1996)                                  4.714286
         Shawshank Redemption, The (1994)                      4.693878
         Star Wars: Episode IV - A New Hope (1977)             4.615385
         Fargo (1996)                                          4.562500
         Emma (1996)                                           4.500000
         Priest (1994)                                         4.500000
         Shanghai Triad (Yao a yao yao dao waipo qiao) (1995)  4.500000
         Nobody's Fool (1994)                                  4.500000
         Taxi Driver (1976)                                    4.500000
         Schindler's List (1993)                               4.444444
         Smoke (1995)                                          4.444444
         Eat Drink Man Woman (Yin shi nan nu) (1994)           4.400000
         Blade Runner (1982)                                   4.400000
         Fish Called Wanda, A (1988)                           4.375000
         Crumb (1994)                                          4.375000
         Silence of the Lambs, The (1991)                      4.368421
         In the Name of the Father (1993)                      4.363636
         Usual Suspects, The (1995)                            4.294118
         Wallace & Gromit: The Best of Aardman Animation (1996)  4.250000
         Adventures of Priscilla, Queen of the Desert, The (1994)  4.250000
         Name: 0, dtype: float64
```

5. CONCLUSION

We have used 3 machine learning algorithms and built three separate systems which recommend movies to the users. This proves that our system is a valid one for recommending movies. By comparing the result we can conclude that each algorithm is efficient in each way to build a recommendation system.

REFERENCES

[1] Amulya Ahmed, Muyeed, et al. "TV Series Recommendation Using Fuzzy Inference System, K-Means Clustering and Adaptive Neuro Fuzzy Inference System." 2017, pp. 1512–1519.

[2] Park, Moon-Hee, et al. "Location-Based Recommendation System Using Bayesian User‟s Preference Model in Mobile Devices." Ubiquitous Intelligence and Computing Lecture Notes in Computer Science, pp. 1130–1139., doi:10.1007/978-3-540-73549-6_110.

[3] Kumar, S. Gupta, S. K. Singh and K. K. Shukla, "Comparison of various metrics used in collaborative filtering for recommendation system," 2015 Eighth International Conference on Contemporary Computing (IC3), Noida, 2015, pp. 150-154. Publisher: IEEE.

[4] J. Lund and Y. Ng, "Movie Recommendations Using the Deep Learning Approach," 2018 IEEE International Conference on Information Reuse and Integration (IRI), Salt Lake City, UT, 2018, pp. 47-54. Publisher: IEEE.

[5] M. K. Kharita, A. Kumar and P. Singh, "Item-Based Collaborative Filtering in Movie Recommendation in Real time," 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC), Jalandhar, India, 2018, pp. 340-342. Publisher: IEEE.

[6] M. Ilhami and Suharjito, "Film recommendation systems using matrix factorization and collaborative filtering," 2014 International Conference on Information Technology Systems and Innovation (ICITSI), Bandung, 2014, pp. 1-6. Publisher: IEEE.

[7] Huang, Yao-Chang, and Shyh-Kang Jenor. "An audio recommendation system based on audio signature description scheme in MPEG-7 Audio." 2004 IEEE International Conference on Multimedia and Expo (ICME) (IEEE Cat. No.04TH8763), doi:10.1109/icme.2004.1394273.