Department of Computer Science

Missouri State University

CSC735 - Data Analytics

Movie recommendation system with clustering algorithm using Movie lens dataset

Project Proposal

Submitted by

Section: 1 Group number: 4

Ayesha Siddiqua	ID: M03526026
Shusmoy Chowdhury	ID: M03386286

Submitted on September 17, 2023

1 Introduction

The concept of recommendation systems are an active platform for researchers. The system could be built to recommend a book, movies, songs, venues, sites and many more depending on the user needs. There are multiple techniques to build a recommendation system, they are: Content-based filtering, Memory-based filtering, and Model-based filtering. Most recommendation systems rely on user input, ratings, preferences and similarities. The collected information is analyzed for these to produce recommendations. There have been many proposed recommendation systems that used KNN, Neural networks and other deep learning algorithms with either content-based, memory-based or model-based filtering techniques.

Movie recommendation system is one of the most popular applications of big data analysis using machine learning. The recommendation is done by studying the users' past ratings and observed behaviors.

In this project, we will be using the Movielens dataset[1] for the movie recommendation system. GroupLens Research has collected Rating data of movies from users over periods of time and made them available in the MovieLens website.The movielens latest dataset were created by 330975 users between January 09, 1995 and July 20, 2023. This dataset was generated on July 20, 2023 and it contains 33832162 ratings and 2328315 tag applications across 86537 movies. The dataset contains the following files: genome-scores.csv, genome-tags.csv, links.csv, movies.csv, ratings.csv, tags.csv. This project mostly utilises data from ratings.csv for extracting the features populartiy and average rating of a specific movie, and data from movies.csv for extracting the feature genre and movie specific metadata.

2 Problem Statement

he movie recommendation system filters and predicts only those movies that a corresponding user is most likely to want to watch. In our project, we are planning to use a content-based filtering approach for movie recommendation. It is a type of recommendation system that tries to guess what a user would like based on the user's activity or prior liking. The recommendation system will be created using clustering algorithms such as (LDA, K means, GMM) unsupervised machine learning algorithms that can be used to solve recommendation problems. In this project, We have will the clustering algorithms using a set of extracted features from the MovieLens dataset. We will apply various clustering algorithms on the dataset for movie recommendation and compare the results.

3 Objective and Goals

Our goal is to make a movie recommendation system for the users that can achieve the following goals:

- Cluster the similar kinds of movies based on the movie features
- Provide better suggestions of movies based on users' choice
- Maintain user satisfaction by suggesting similar movies

- Compare the recommendation results of various ML algorithms

4 Methodology

R. Singh et al[2] describes an approach which offers generalized recommendations to every user, based on movie popularity and/or genre. They have illustrated the modelling of a movie recommendation system by making the use of content based filtering in the movie recommendation system. The KNN algorithm is implemented in this model along with the principle of cosine similarity.

B.-B. Cui et al[3] designed and implemented a movie recommendation system prototype combined with the actual needs of movie recommendation through researching of KNN algorithm and collaborative filtering algorithm

1. We will load the move lens dataset for training Apache Spark
2. We will train our K means model with movie lens dataset
3. We will tune the model hyper parameters
4. Similarly, we will implement LDA and GMM clustering algorithms and train the models
5. We will compare the performance of K means, LDA and GMM algorithms
6. We will also compare the performance of our clustering algorithms with the ALS algorithm

5 Timeline

Goal	**Timeline**
Training the K-means model	2 weeks
Training the LDA Model	2 weeks
Training the GMM Model	2 weeks
Tuning Hyper Parameter	1 week
Writing Reports	1 week

References

[1] F. M. Harper and J. A. Konstan, “The movielens datasets: History and context,” *Acm transactions on interactive intelligent systems (tiis)*, vol. 5, no. 4, pp. 1–19, 2015.

[2] R. Singh, S. Maurya, T. Tripathi, T. Narula, and G. Srivastav, “Movie recommendation system using cosine similarity and knn,” *International Journal of Engineering and Advanced Technology*, vol. 9, pp. 2249–8958, 06 2020.

[3] B.-B. Cui, “Design and implementation of movie recommendation system based on knn collaborative filtering algorithm,” in *ITM web of conferences*, vol. 12. EDP Sciences, 2017, p. 04008.