CSC478 Final Project Report
Jun Tae Son | Yuchen Wu | Tzu Hao Peng

Movielens Recommender System Analysis

Project goals:

A great movie recommendation system saves our time to choose movies and enhance our entertainment and lifestyle. We are using Movielens dataset to see if there is any interesting patterns among users, ratings, movies and algorithms behind movie recommender system. Furthermore, we will explore how the traditional methods such as collaborative filtering, content based item based, svd matrix,etc composed a great engine.

Method used:

We explored general characteristics of the movielens data and visualized them through histogram and bar graphs. We performed PCA to find the optimal number of principal components that explain 90% of the total variance and use k-means to do cluster analysis.

We proposed three different methodology of recommender algorithm and compared the results of recommended items in this paper. First model is a content-based recommender system that gives recommendations that are similar in content of the items. Second and third models is based on collaborative filtering with KNN and SVD approach. These item-based collaborative recommender systems recommend movies that is similar to the items that a user has shown interest.

5-folds cross validation is used for the evaluation of the models. The result shows that KNN and SVD algorithms give the best performance. After implementation, we tested the models with a sample of user id=1. Content-based filtering model predicted that user 1 is more likely to watch comedy, drama, and romance genre of movies in the future. Two collaborative filtering models with KNN and SVD algorithm also return movies with similar genre, but the lists of 5 recommended movies from each model were different one another.

Conclusion:

Overall, the average movie rating score is 3. There are 10 movies are rated at 5; however, there are only few users rated those movies. Star Wars is the most popular rated movie, it has 583 users rating it and average rating score is 4.4. There are more male participants than females in the dataset. Age from 20-29 is the group which gives the most rating; however, it gives the lowest average rating score. One potential reason is the millennials are savvy about mass media and they are more harsh on the rating score. Students are the most people who participate in the survey and doctors are the least. People who don’t have an occupation tend to give the highest average rating score and people who work in the healthcare field give the lowest average rating score.

We proposed content-based and two item-based recommendation algorithms in this paper, but the actual recommended movie results from each model were different one another. Since each model is based on different approaches, it is reasonable that they do not provide the same result. We expect that implementation with weighting method suggested by Manoj Kumar et al. and larger dataset with less sparsity may help the recommender systems achieve a better performance.