Project Proposal for NTDS
Movie Recommendation Engine

Group members: Luo Yaxiong, Deng Wenlong, Qiao Qianqian, Wang Pei.

Project Description

We aimed to model an engine for recommendation of movies based on several aspects such as popularity, similarity. We would like to find similar items by using a similarity metric. Once we have the matrix, we use it to determine the best recommendations for a user based on the movies he has already liked and watched.

Data acquisition

Movielens Dataset contains 26,000,000 ratings and 750,000 tag applications applied to 45,000 movies by 270,000 users, including tag genome data with 12 million relevance scores across 1,100 tags and was updated on 8/2017. This dataset consists of:

Genome-scores	Genome-tags	Links
Movies	Ratings	Tags

Data exploration

As with nearly any real-life dataset, we need to do some cleaning first. For various keywords of movies, it should be reduced to several main words which can describe most movies. Then in consideration of calculation times and our evaluation process, select a subset of users according to their ratings.

Data Exploitation

Our job is to design a algorithm which could output the best similar ones with the movies having been rated by users as inputs.

- Know the data and get a network on movies.
- Calculate the similarity is a big job. We will try Collaborative Filtering Model to find similarities between recommendations. There are three types which we might use:
  1. Jaccard Similarity (based on the number of users which have rated item A and B divided by the number of users who have rated either A or B); 2. Cosine Similarity; 3. Pearson Similarity (the pearson coefficient between the two vectors).
- Consider inner connections between actors and directors. i.e. directors have their own preferable type of movies and they would like to use specific actors. Those movies are liked by some fixed groups.

Evaluation

For evaluating recommendation engines, we can use the concept of precision-recall. Our aim is to maximize both precision and recall. From the dataset, we model the engine using only a subset of users. When we finished to model the recommendation engine, we use it to recommend movies to the users according to part of their rating list and compare the rest ratings and the results we recommend taking precision-recall as standard.