Hybrid movie recommender

Vinay Reddy, Siddharth Sachdeva, Sasha Trubetskoy – CS 122

Introduction

There are two types of recommendation systems. Suppose we have a particular user, Matt, who likes the movie topics crime and romance. Collaborative filtering involves finding users who have rated movies similarly to Matt, and then recommending movies that these users like, but Matt has not seen. A content-based algorithm would recommend movies that are about crime and/or romance to Matt. Both of these algorithms have certain pros and cons.

Collaborative filtering does not require any content data and it can suggest "serendipitous" items that would not match on content, but are liked by similar users. However, it does not work well for new users or for new movies.

Content-based algorithm is useful for recommending new content that hasn't been rated by many users. However, it may not recommend movies outside the Matt's narrow profile, and it may not capture all of Matt's movie interests.

We will develop a hybrid model that benefits from the best aspects of both algorithms. Unlike previous hybrid methods, our model analyzes the full movie script to look at content.

Goal

Our project will culminate in a simple movie recommendation website. The site will present the user with 100 random movies, and the user must then rate at least 10 of those movies. After submitting their ratings, the website will give the user a list of recommendations, using the hybrid algorithm.

The most important indicator of success is the root-mean-square error (RMSE). This is essentially the standard deviation of the residuals of predicted vs. actual ratings. Our goal is to minimize this number, ideally below a benchmark of 1.0. We can compare various sub-models and visualize certain errors that our model tends to make.

Datasets

MovieLens

MovieLens, assembled by GroupLens Research, is a very large set of movie ratings and tags. The full version of MovieLens contains 20 million ratings by 138,000 users, though a "small" version is available with only 100,000 ratings. The dataset also contains metadata for each film in the form of tags and tag relevance scores.

Springfield! Springfield!

SpringfieldSpringfield.co.uk contains over 16,000 full movie scripts. We will use Beautiful Soup to scrape these and will pair the script in string format with the associated film's database entry.

Timeline

30 January – Have read relevant literature, become familiar with the model
5 February – Have cleaned data sets
13 February – Have pseudocode of model and partial models
25 February – Have working full model completed
6 March – Django site completed, full model finalized
13 March – Final software due and presentations