ENSAI

IT Tools 2

Building a movie recommendation system based on collaborative filtering.

Authors

Robbert Driehuijs

Gonem Lau

November 6, 2023

1 Introduction

Recommendation systems play a crucial role in discerning and predicting users' preferences or interests, turning these insights into suggestions for new products or services based on their previous behavior or similarities with other users. Leveraging machine learning techniques, these systems manage extensive user data to deliver personalized recommendations. These systems use various types of user data, such as metadata, ratings, reviews, user profiles, and network data, to provide personalized recommendations to enhance the overall user experience.

These systems generally fall into three categories: collaborative filtering, content-based filtering, and hybrid methods. Collaborative filtering relies on the behavior of other users, assuming that users who exhibited similar rating behavior in the past will do the same in the future. Content-based filtering suggests items similar to those a user has liked before based on various item features. Hybrid methods combine both collaborative and content-based filtering.

In this report, we will concentrate on collaborative filtering, commonly used in the industry and effective when item content is complex or unavailable. Our objective is to create a movie recommendation system based on collaborative filtering and evaluate its performance by analyzing the accuracy of predictions using different model complexities.

2 Methodology

In this section, the methodology for building a recommendation system is outlined together with the metric to evaluate the recommendations.

2.1 Data

For the recommendation system, two datasets are utilized that describe movie ratings. Both datasets originate from MovieLens and differ in the number of ratings, users, and movies covered. The summary statistics are displayed in Table 1.

We had limited time and resources, so we mostly used a smaller dataset for our computations and model tuning. However, we still wanted to test the scalability and effectiveness of our approach on a larger scale and applied one model to the large dataset.

Table 1: Descriptive statistics of movie ratings datasets.

Variable	Small dataset	Large dataset
Ratings	100,836	20,000,263
Users	610	6,085
Movies	9,742	27,278
Average	3.50	3.52

2.2 Method

The recommendation model predicts a user's rating for a movie by analyzing a subset of users who exhibit similar rating behavior.

The system is built using the collaborative filtering approach. The basic idea behind this technique is to predict a user's rating for a target movie by averaging the ratings of similar users. The top 10 most similar users are considered for prediction. Similarity between users is assessed by computing the cosine similarity between their ratings. The subset of users with the highest similarity scores, relative to the user of interest, forms the neighborhood for constructing a new rating.

To make predictions on a personal level, the collaborative filtering algorithm adjusts the ratings of similar users by incorporating the global average, user bias, and item bias. The global average rating represents the average rating given to the movies across all the users in the dataset. The user bias represents the average difference in movie ratings between a user and the global average. The movie bias represents the average difference between a movie's ratings and the global average. The final predicted rating for a particular movie and user is computed as a weighted sum of the ratings by similar users, adjusted by the global average, user bias, and movie bias.

2.3 Models

To examine the effects of different components on prediction accuracy, several heuristic models are constructed. These models are summarized in Table 3.

Table 2: Movie Rating Models.

Model	Model description
Global Average	Using the global average μ only
User Average	Combining the global average μ with the user-specific bias
Movie Average	Combining the global average μ with the movie-specific bias
Individual Model	Combining the global average μ with the user and movie bias
Neighborhood Model	Combining the individual model with selected k-nearest neighbors

The following equations describe the construction of the Individual model and the Neighborhood model.

$$b_{xi} = \mu + b_x + b_i,$$
$$\hat{r}_{xi} = b_{xi} + \frac{\sum_{j \in N(i;x)} S_{ij} \cdot (r_{xj} - b_{xj})}{\sum_{j \in N(i;x)} S_{ij}},$$

where b_{xi} represents the baseline estimate for the rating by user x for movie i, μ is the global average rating, b_x is the user bias, and b_i is the movie bias.

2.4 Neighborhood size experiment

The examination of the recommendation system is broadened by examining the effect of the neighborhood size on the predictions.

The estimates of the Neighborhood model include the influence of the k-nearest neighbors. This model consists of two main components: an individual component, which includes the global average, user bias, and movie bias, and a local component, which includes a weighted average based on a predetermined number of closest similar neighbors. Through testing two different values for the number of neighbors, the influence of the user similarity network on the RMSE is explored. We experiment with the Neighborhood model using different values for k, specifically, $k = 10$ and $k = 100$.

2.5 Accuracy

To measure the accuracy of the predictions, the Root Mean Squared Error (RMSE) is utilized. The RMSE measures the differences between the predictions and the actual ratings for a test dataset.

3 Results

After evaluating the recommendation system using the global average model, it was found that the RMSE was almost identical for both the small and large datasets. The RMSE values were approximately 1.037 and 1.035, respectively. This result indicates that the small dataset is a representative subset of the larger one. It suggests that the smaller subset maintains the overall distribution found in the full dataset, making it a reliable proxy for larger-scale experiments when computational resources are limited.

A focused analysis of the small dataset provided showed that the accuracy of predictions varies across different collaborative filtering model complexities, as demonstrated by employing RMSE as a performance benchmark. The results are displayed in Table 3.

Table 3: Comparative Analysis of RMSE Across Different Movie Rating Models.

Model	Model description
Global Average	1.037
User Average	0.920
Movie Average	0.923
Individual model	0.869
Neighborhood Model Small	0.959
Neighborhood Model Large	0.880

The most basic model, which only used the overall average rating, gave an RMSE of 1.037. While this model is computationally efficient, it lacks personalization as it only considers the general popularity of movies. Models incorporating user bias (RMSE of 0.920) and movie bias (RMSE of 0.923) separately showed improved accuracy. These inclusions mean that personalized biases play a substantial role in understanding user preferences. The individual model, which combined the global average with user and movie biases, further decreased the RMSE to 0.869. This experiment highlights the effectiveness of a multi-dimensional approach that incorporates various bias factors.

During our experimentation, we discovered that increasing the number of similar users from a smaller group of 10 ($k = 10$) to a larger group of 100 ($k = 100$) resulted in a better RMSE outcome. The Neighborhood Model, employing a small neighborhood ($k = 10$), yielded an RMSE of 0.959, while a larger neighborhood ($k = 100$) provided an RMSE of 0.880. These RMSE results are still lower than for the Individual model. This can be attributed to the inclusion of Neighbors into the predictions, which may introduce noise and overfitting.

However, it is crucial to consider the trade-off between accuracy and computational cost, as the increase in neighborhood size also leads to an increase in computational cost. The smaller neighborhood size, despite its efficiency, lacked the prediction refinement seen with the larger group.

Extending our analysis to the large dataset, the resulting RMSE for the Neighborhood Model Small ($k = 10$) is 0.932, lower than the RMSE obtained on the smaller dataset. This decrease in Neighborhood Model RMSE can be attributed to the higher likelihood of finding more similar users in the larger dataset. However, this comes at the expense of much higher computational costs due to the increased number of ratings and connections between users and movies. Due to the substantial computational costs, the experiment was not performed for the Neighborhood Model Large ($k = 100$).

4 Conclusion

The study confirms the importance of incorporating multiple factors in collaborative filtering to refine the accuracy of movie recommendations. While a model based solely on the global average is computationally efficient, it fails to provide personalized recommendations and only reflects general trends. The Individual Model is more accurate than the Global Average model due to the inclusion of user and movie biases. However, it still relies on trends, providing a shifted baseline rather than authentic personalization.

Although the collaborative filtering approach has been successful, there is still room for further enhancements. Future research can explore the potential for temporal dynamics, like the changing preferences of users and the evolving popularity of movies over time. Additionally, the study emphasizes the significance of efficient computation techniques to manage large neighborhoods without compromising performance. The small dataset's comparable performance to the large one on the global average model confirms its representativeness. While the Neighborhood Model with fewer neighbors showed better results in the large dataset, it must be balanced against the higher computational requirements, making a case for careful consideration of model complexity and efficiency in large-scale recommendation systems.

To sum up, while collaborative filtering is an effective methodology, it has limitations when simplistic models are employed. It is vital to balance accuracy and personalization in recommendations and computational feasibility. In the future, systems should aim to capture complex user preferences and temporal trends using scalable and efficient algorithms.