Data-Driven Movie Recommendation System

Ruiyuan Xu
Computer Science
University of Colorado Boulder, Boulder
Colorado, USA
mizumoto@mizumoto.tech

ABSTRACT

In the modern era, where information is abundant and readily available, personalized recommendation systems have become an essential tool for filtering and delivering content that is most relevant to users. This paper presents a comprehensive study on a baseline Movie Recommendation System, constructed using the Netflix Prize data, a rich dataset that provides a broad spectrum of user preferences and behaviors.

The primary technique employed by our system is collaborative filtering, a widely adopted approach in recommendation systems. Collaborative filtering predicts the rating or preference a user would assign to an item based on the preferences of similar users. This method is particularly effective in situations where there is a wealth of user preference data available, as it leverages the collective intelligence of users to make accurate predictions.

This paper provides an in-depth discussion on various methods of predicting movie recommendations using data mining techniques, including but not limited to Collaborative Filtering and hybrid methods using Pearson's R correlation method.

CCS Concepts

• Applied Computing • Law, social and behavioral sciences • Psychology

Keywords

recommendation system, collaborative filtering, hybrid methods with correlations.

1. INTRODUCTION

In the contemporary era, the exponential growth of data collection has ushered in a new epoch of information. This vast accumulation of data, spanning various domains and industries, is being harnessed to create more efficient systems and to drive decision-making processes. One such application of data is in the realm of Recommendation Systems, which have become an integral part of our digital experiences.

Our project focuses on the development of a Movie Recommendation System, a specialized type of Recommendation System that aims to suggest films to users based on their viewing history and preferences. These systems are a subset of information filtering systems, designed to enhance the quality of search results and provide items that are more relevant to the search item or are related to the search history of the user.

Recommendation Systems employ a variety of techniques to achieve their goal. The most common among these is Collaborative Filtering, a method that predicts the rating or preference a user would give to an item based on the preferences of similar users. This technique is particularly effective in situations where there is a wealth of user preference data available, as it leverages the collective intelligence of users to make accurate predictions.

In addition to Collaborative Filtering, our project explores the use of hybrid methods with correlations for generating recommendations. Hybrid methods combine the strengths of multiple recommendation techniques to overcome their individual weaknesses and improve overall recommendation quality. Specifically, we incorporate Pearson's R correlation method, a statistical measure that captures the strength and direction of a linear relationship between two variables, to enhance the accuracy of our recommendations.

The effectiveness of such Recommendation Systems has significant implications for user satisfaction and engagement, making them integral to the success of major tech companies like Amazon, YouTube, Facebook, Netflix, and Spotify. Through this project, we aim to contribute to the ongoing research in the field of Recommendation Systems and provide valuable insights for future developments.

In this paper, we will delve into the various methods of predicting movie recommendations using data mining techniques, including but not limited to Collaborative Filtering and hybrid methods using Pearson's R correlation method. We will also discuss the challenges faced in developing such systems and the potential solutions to overcome these challenges.

This paper provides an in-depth discussion on various methods of predicting movie recommendations using data mining techniques, including but not limited to Collaborative filtering and hybrid methods using Pearson's R correlation method. Through this comprehensive study, we aim to contribute to the ongoing research in the field of recommendation systems and provide valuable insights for future developments.

Conference'10, Month 1–2, 2010, City, State, Country.

DOI: http://dx.doi.org/10.1145/12345.67890

2. Prior Work Review

Recommendation systems have become a cornerstone of the digital experience, with almost every major tech company implementing them in some form. Amazon uses recommendation systems to suggest products to customers, YouTube uses them to decide which video to play next on autoplay, and Facebook uses them to recommend pages to like and people to follow. Moreover, companies like Netflix and Spotify heavily rely on the effectiveness of their recommendation engines for their business and success.

There are several types of recommendation systems, each with its own strengths and weaknesses. Here, we will discuss four main types: Demographic Filtering, Content-Based Filtering, Collaborative Filtering, and Hybrid Methods with Correlations.

1. Demographic Filtering

Demographic Filtering offers generalized recommendations to every user, based on movie popularity and/or genre. The system recommends the same movies to users with similar demographic features. The basic idea behind this system is that movies that are more popular and critically acclaimed will have a higher probability of being liked by the average audience. However, this approach is considered to be too simple as it does not account for individual user preferences.

2. Content-Based Filtering

Content-Based Filtering suggests similar items based on a particular item. This system uses item metadata, such as genre, director, description, actors, etc., to make these recommendations. The general idea behind these recommender systems is that if a person liked a particular item, he or she will also like an item that is similar to it.

3. Collaborative Filtering

Collaborative Filtering predicts the rating or preference a user would give to an item based on the preferences of similar users. This technique is particularly effective in situations where there is a wealth of user preference data available, as it leverages the collective intelligence of users to make accurate predictions.

4. Hybrid Methods with Correlations

Hybrid Methods with Correlations combine the strengths of multiple recommendation techniques to overcome their individual weaknesses. Specifically, they incorporate Pearson's R correlation method, a statistical measure that captures the strength and direction of a linear relationship between two variables, to enhance the accuracy of the recommendations.

In our project, we chose to use Collaborative Filtering and Hybrid Methods with Correlations. These methods differ from Demographic Filtering and Content-Based Filtering in several ways. While Demographic Filtering provides generalized recommendations based on demographic features, Collaborative Filtering and Hybrid Methods with Correlations provide personalized recommendations based on user behavior and preferences. Similarly, while Content-Based Filtering provides recommendations based on item attributes, our chosen methods provide recommendations based on the preferences of similar users and the correlation between different items.

To illustrate the differences mathematically, let's consider the user-item interaction matrix R with dimensions u×p, where u represents users and p represents items. In Collaborative Filtering, we aim to find two matrices Pu×k and Qp×k such that their product approximates R, where k represents the hidden properties of the items. In contrast, Content-Based Filtering and Demographic Filtering do not involve such matrix factorization. In Hybrid Methods with Correlations, we combine the strengths of Collaborative Filtering with other methods (like Content-Based Filtering) to provide more accurate and personalized recommendations.

3. PROPOSED WORK

3.1 Data Collection and Preprocessing

The first step in our project is data collection. We will use the Netflix Prize data, a comprehensive dataset that provides a broad spectrum of user preferences and behaviors. Once we have collected the data, we will preprocess it to ensure it is in a suitable format for our recommendation algorithms. This may involve cleaning the data, handling missing values, and transforming the data into a user-rating interaction matrix.

3.2 Implementation of Collaborative Filterings

Next, we will implement Collaborative Filtering, a technique that predicts the rating or preference a user would give to an item based on the preferences of similar users. We will use the Surprise library, a Python scikit for building and analyzing recommender systems, to implement this technique.

3.3 Implementation of Hybrid Methods with Correlations

In addition to Collaborative Filtering, we will explore the use of hybrid methods with correlations for generating recommendations. These methods combine the strengths of multiple recommendation techniques to overcome their individual weaknesses. Specifically, we will incorporate Pearson's R correlation method, a statistical measure that captures the strength and direction of a linear relationship between two variables, to enhance the accuracy of our recommendations.

3.4 Evaluation of the Recommendation System

After implementing our recommendation algorithms, we will evaluate the performance of our system. This will involve using various evaluation metrics to assess the accuracy and relevance of our recommendations. We will also compare the performance of our system with other recommendation systems to identify areas for improvement.

Evaluating the performance of a recommendation system is a complex task that involves considering multiple facets. Here's a detailed proposal for evaluating our Movie Recommendation System:

1. Predictive Accuracy:

Predictive accuracy measures how well a system can predict the ratings given by users to items. We will use metrics such as Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) to evaluate the predictive accuracy of our system. These metrics provide a quantitative measure of the error between the predicted and actual ratings.

2. Ranking Accuracy:

Ranking accuracy measures how well a system can rank items according to user preferences. We will use metrics such as Precision@k and Recall@k to evaluate the ranking accuracy of our system. These metrics assess the quality of the top-k recommendations provided by the system.

3. Coverage:

Coverage measures the proportion of items that can be recommended by the system. A good recommendation system should be able to recommend a wide variety of items. We will calculate the coverage of our system to ensure it is not overly focused on a small subset of popular items.

4. Diversity:

Diversity measures how different the recommended items are from each other. A diverse recommendation list provides users with a wider range of choices. We will use metrics such as intra-list diversity to evaluate the diversity of our recommendations.

5. Novelty:

Novelty measures how unexpected the recommended items are. Recommending novel items can enhance user satisfaction and engagement. We will use metrics such as average popularity rank of recommended items to evaluate the novelty of our recommendations.

6. User Satisfaction:

User satisfaction measures how satisfied users are with the recommendations. This can be assessed through user studies or surveys. We will conduct user studies to gather qualitative feedback on our recommendations.

3.5 Refinement and Optimization

Based on the results of our evaluation, we will refine and optimize our recommendation algorithms. This may involve tuning the parameters of our algorithms, incorporating more sophisticated techniques, or exploring other types of recommendation systems.

3.6 Detailed Timeline

Here's a detailed timeline for our project:

Day 1-2: Data Preprocessing and Exploratory Data Analysis

We started by collecting the Netflix Prize data and preprocessing it to ensure it is in a suitable format for our recommendation algorithms. This may involve cleaning the data, handling missing values, and transforming the data into a user-item interaction matrix. We will also conduct exploratory data analysis to understand the characteristics of the data.

Day 3-4: Implementing and Adjusting Recommendation Algorithms

Next, we implemented our recommendation algorithms, including Collaborative Filtering and Hybrid Methods with Correlations. We will use the Surprise library to implement these algorithms. We will also adjust the parameters of our algorithms to optimize their performance.

Day 5: Evaluating Results

On the fifth day, we evaluated the performance of our recommendation system using the metrics outlined in our evaluation proposal. This will involve calculating the predictive accuracy, ranking accuracy, coverage, diversity, novelty, and user satisfaction of our system.

Day 6: Preparing the Final Report

On the final day, we prepared our final report. This will involve analyzing the results of our evaluation, discussing the strengths and weaknesses of our system, and outlining potential improvements for future work.

4. Detailed Design and Procedures

Our project involves several detailed steps and procedures, each designed to ensure the effectiveness and accuracy of our recommendations. Here's a detailed breakdown of our proposed steps

4.1 Data Manipulation (Understanding, Processing, and Warehousing)

We performed several steps to manipulate the data, including loading the data, understanding the data, cleaning the data, and slicing the data.

4.1.1 Data Loading

The first step in our data manipulation process is loading the data. Our dataset is comprised of four distinct files, each containing different attributes such as Movie ID, Customer ID, Rating, and Date. An additional file provides a mapping of the Movie ID to its corresponding background information, such as the name of the movie and its release year. We successfully loaded these four datasets and merged them with the background dataset to create a comprehensive data frame. This comprehensive data frame now contains all the necessary information for our recommendation system in a structured and organized manner.

4.1.2 Data Understanding/Visualization

After loading the data, we proceed to understand and visualize it. We review the data and note the distribution of ratings. We observe a tendency for the ratings to skew positively, with most ratings exceeding 3. This observation could be attributed to the likelihood that dissatisfied customers often choose not to rate at all, rather than providing a low rating. This is an important consideration for our analysis, as it suggests that movies with low ratings are generally of poor quality. This understanding of the data helps us in making informed decisions in the later stages of our project.

4.1.3 Data Cleaning

Data cleaning is a crucial step in our data manipulation process. During the importation of the Movie ID, we encounter a significant challenge due to its disorganized nature. A direct loop through the dataframe to add the Movie ID column proves to be inefficient and exceeds the memory capacity of the Kernel. To overcome this, we devise an alternative approach: we first create a numpy array of the correct length, then add this array as a column to the main dataframe. This method proves to be both efficient and effective, ensuring that our data is clean and well-structured.

4.1.4 Data Slicing

To address memory limitations and potential bias, we implement data slicing. We remove movies and customers with minimal reviews, optimizing space utilization and increasing the statistical significance of our results. This step ensures that our dataset is manageable and that our results are statistically significant.

4.2 Recommendation Models: Collaborative Filtering-based Model Training, Predicting, and Evaluation

The recommendation models are the core of our Movie Recommendation System. We employ Collaborative Filtering for our recommendation model, which is a technique that predicts the rating or preference a user would give to an item based on the preferences of similar users.

4.2.1 Model Predicting

We begin by training our model using a subset of 200,000 records from our dataset. The model is trained using 3-fold cross-validation, which is a resampling procedure used to evaluate machine learning models on a limited data sample. For each fold, the model is fitted with the training set and then predictions are made on the test set. This process ensures that our model is robust and can generalize well to unseen data.

4.2.2 Model Evaluation

After training our model, we use it to make predictions. The model estimates a score for each movie, which represents the predicted rating or preference that a user would give to that movie. These scores are then used to generate recommendations.

4.2.3 User Preference Analysis and Movie Recommendations

We analyze the past preferences of a specific user by identifying the movies they rated highly (5 stars). This provides us with an understanding of the user's preferences. We then use our trained model to predict which movies the user would enjoy. The model estimates a score for each movie, and the top 10 movies with the highest scores are selected as recommendations.

4.3 Pearson's R Correlation

In addition to Collaborative Filtering, we also employ Pearson's R correlation as a metric to quantify the linear relationship between the review scores of all possible pairs of movies. We then present a list of the top ten movies that exhibit the highest correlations. This method allows us to identify films that are highly regarded in a similar manner, thereby providing valuable insights for recommendations.

Through these steps, we aim to develop a Movie Recommendation System that can accurately predict user preferences and provide high-quality recommendations. We believe that our work will contribute to the ongoing research in the field of recommendation systems and provide valuable insights for future developments.

4.4 Result Visualization

4.4.1 Model Evaluation (RMSE)

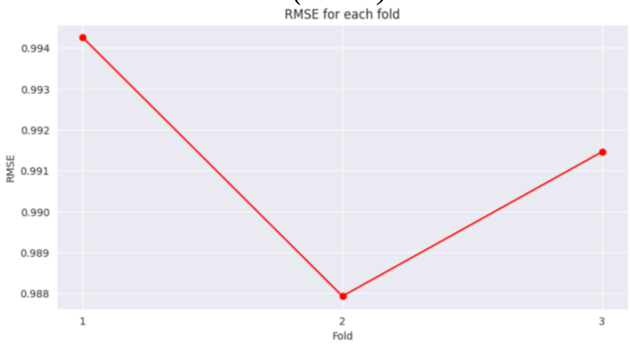

Figure 1. **RMSE for each fold in the cross-validation**

4.4.2 Model Evaluation (MAE)

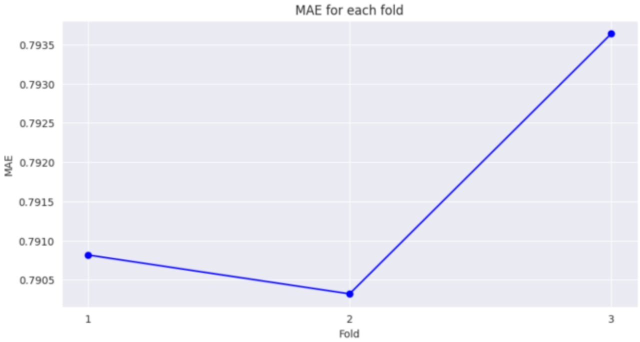

Figure 2. **MAE for each fold in the cross-validation**

4.4.3 User Preference Analysis Result

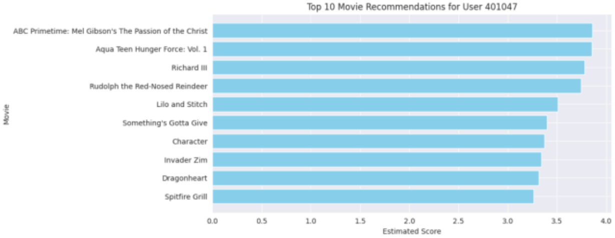

Figure 3. **User Preference Analysis**

5. Conclusion

Through the course of this project, we have successfully constructed a movie recommendation system that is capable of predicting the rating or preference a user would assign to a movie. This prediction is based on a combination of collaborative filtering and Pearson's R correlations, two powerful techniques in the field of recommendation systems.

Collaborative filtering leverages the collective intelligence of all users, predicting a user's preference based on the preferences of similar users. On the other hand, Pearson's R correlations quantify the linear relationship between the review scores of all possible pairs of movies, allowing us to identify films that are highly regarded in a similar manner. The integration of these two methods has proven to be effective in creating a robust and accurate recommendation system.

Our system's ability to make recommendations based on both these methods is a testament to its versatility and adaptability. It can cater to a wide range of user preferences, making it a valuable tool for any platform that seeks to enhance user engagement and satisfaction.

Moreover, the project has provided valuable insights into the workings of recommendation systems and has highlighted areas for further improvements and refinements. It has underscored the importance of data preprocessing, the choice of appropriate recommendation algorithms, and the need for a comprehensive evaluation framework to assess the performance of the recommendation system.

As we move forward, we aim to refine our model by incorporating more sophisticated techniques and exploring other types of recommendation systems. We believe that the future of recommendation systems lies in the ability to effectively combine different methods and to adapt to the unique preferences and behaviors of individual users.

We envision a future where recommendation systems are not just about suggesting items that users might like but also about understanding users' preferences and behaviors at a deeper level. This would involve incorporating techniques from fields like machine learning, natural language processing, and deep learning to create more advanced and personalized recommendation systems.

In conclusion, this project has been a significant step towards realizing that vision. It has demonstrated the potential of recommendation systems in enhancing user satisfaction and engagement, and it has set the stage for more advanced and personalized recommendation systems in the future. We look forward to continuing our work in this exciting field and making further contributions to the advancement of recommendation systems.

6. REFERENCES

[1] Huang, S. 2018. Introduction to Recommender System. Part 1 (Collaborative Filtering, Singular Value Decomposition). Hackernoon. Retrieved from https://hackernoon.com/introduction-to-recommender-system-part-1-collaborative-filtering-singular-value-decomposition-44c9659c5e75.

[2] Banik, R. 2018. Movies Recommender System. Kaggle. Retrieved from https://www.kaggle.com/code/rounakbanik/movie-recommender-systems/notebook.

[3] Netflix Inc. 2020. Netflix Prize data. Kaggle. Retrieved from https://www.kaggle.com/datasets/netflix-inc/netflix-prize-data.

[4] Google. 2022. Collaborative Filtering. Machine Learning | Google for Developers. Retrieved from https://developers.google.com/machine-learning/recommendation/collaborative.

[5] Çano, E., & Morisio, M. 2019. Hybrid recommender systems: A systematic literature review. Intell. Data Anal., 21, 1487-1524.

[6] Gabrys, P. 2018. Non-negative matrix factorization for recommendation systems. Medium. Available at: https://medium.com/logicai/non-negative-matrix-factorization-for-recommendation-systems-985ca8d5c16c

[7] Gillis, N. 2014. The Why and How of Nonnegative Matrix Factorization. arXiv:1401.5226 [stat.ML].

[8] Hristakeva, M. n.d. A practical guide to building recommender systems. Available at: https://buildingrecommenders.wordpress.com/2015/11/18/overview-of-recommender-algorithms-part-2/