Hybrid Movie Recommendation System using Collaborative and Content-Based Filtering

Thanapoom Phatthanaphan,[1] Chandini Rayadurgam,[2] Shreya Daniel Jacob[3]
Stevens Institute of Technology
tphattha@stevens.edu,[1] crayadur@stevens.edu,[2] sjacob4@stevens.edu[3]

Abstract

The research focuses on movie recommendation systems as its application domain, aiming to create a hybrid recommendation system that blends collaborative filtering and content-based filtering techniques. By doing so, the system can provide personalized movie recommendations to users, which is a common practice among movie streaming platforms and other entertainment providers. This research aims to enhance users' movie-watching experience by suggesting movies that match their preferences, leading to greater satisfaction and engagement.

Introduction

Movie recommendation systems have gained popularity with the emergence of streaming platforms like Netflix and Amazon Prime. These systems suggest movies to users based on their past preferences. However, traditional recommendation systems using collaborative filtering (CF) or content-based filtering (CBF) techniques have limitations. CF faces the cold start problem for new users or movies without historical data, while popularity bias leads to recommendations favoring popular movies. CBF relies on movie features, limiting recommendation diversity and exposure to new types of movies.

To address these limitations, this research proposes the development of a hybrid movie recommendation system that combines CF and CBF techniques that can address the limitations of traditional recommendation systems, evaluate the performance of the proposed hybrid system, and compare it to traditional CF and CBF techniques. [7] The MovieLens 1M dataset from GroupLens Research will be used to train and evaluate the system. This dataset contains 1,000,209 anonymous ratings of approximately 3,900 movies made by 6,040 MovieLens users who joined MovieLens in 2000, making it a rich source of data for building a movie recommendation system.

The proposed system combines Collaborative Filtering (CF) and Content-Based Filtering (CBF) techniques to enhance recommendation accuracy and diversity. CF analyzes user behavior to identify similar users and movies, while CBF examines movie attributes such as genre, director, and cast. By integrating these approaches, the system aims to address the limitations of each method and offer a comprehensive and personalized recommendation system.

Problem Statement

The problem statement for building a movie recommendation system is that the system may encounter challenges such as data sparsity, cold start problem, and lack of diversity. These issues can make it difficult to accurately capture user preferences and provide recommendations that are relevant and diverse.

Data sparsity

Data sparsity occurs when there is a limited amount of data available for some movies, which can make it challenging to provide accurate recommendations.

Cold start problem

The cold start problem arises when there is a lack of data available for new users, which can make it difficult to provide personalized recommendations.

Lack of diversity

The lack of diversity issue can arise when the system tends to recommend similar movies, which can limit the user's exposure to new and different content.

Relate work

This research paper explores six related works in the field of movie recommendation systems. [1] The first study focuses on collaborative filtering, proposing an approach that utilizes user similarity for movie recommendations. [2] The second research highlights content-based filtering, employing movie features to suggest movies based on user preferences. [3] The third research discusses the integration of Collaborative Filtering (CF) and Content-Based Filtering (CBF) techniques to improve the recommendation accuracy. It explores how CF and CBF can be combined to leverage

user preferences and movie features to provide more effective movie recommendations. [4] The fourth work emphasizes the integration of cosine similarity-based collaborative filtering to enhance the accuracy of recommendations. [5] The fifth study delves into matrix factorization, specifically single value decomposition, as an effective technique for personalized movie recommendations. Lastly, [6] the sixth study compares and analyzes movie recommendation systems using evaluation metrics such as MAPE and RMSE. By examining these related works, this research aims to gain insights into various techniques and approaches used in movie recommendation systems, contributing to the development of more accurate and personalized recommendation algorithms.

Artificial Intelligence Techniques

Collaborative Filtering

[1] Collaborative Filtering is a method that involves gathering information through user ratings and identifying similarities between users as shown in Figure 1. This technique is used to provide personalized recommendations to users. As with any algorithm, there are advantages and disadvantages to this approach.

Advantages

Collaborative filtering offers personalized recommendations based on user behavior, leading to improved user engagement and satisfaction. It is also scalable, capable of handling large user and item volumes, making it efficient for online marketplaces and e-commerce platforms.

Disadvantages

However, there are also some disadvantages to collaborative filtering. One of the primary concerns is the **Cold-start problem**. Collaborative filtering algorithms require data on user's behavior to make recommendations, which can be a challenge for new users or new items with limited data. Collaborative filtering can also suffer from popularity bias, where it tends to recommend popular items even if they are not a good fit for the user's preferences. **Data sparsity** can also be a problem, especially if users have not rated enough items or if there are too many items with few ratings.

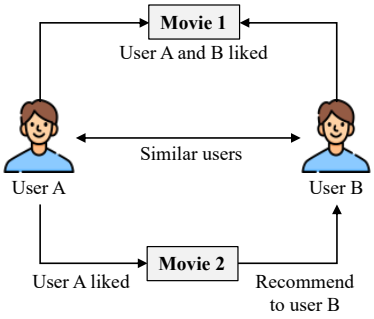

Figure 1: Collaborative Filtering

Content-based Filtering

[2] Content-based filtering is another popular technique for making recommendations to users. This approach analyzes the content of the items to recommend similar items to users as shown in Figure 2. Like collaborative filtering, there are both advantages and disadvantages to content-based filtering.

Advantages

Content-based filtering has notable advantages in the realm of recommendation systems. It does not rely on user preferences, instead analyzing item content to suggest items similar to those previously liked or interacted with by the user. This feature is particularly beneficial for new users or newly introduced items with limited data. Moreover, content-based filtering enables personalized recommendations tailored to specific user interests since it operates based on item content. Additionally, it performs well in handling the cold-start problem by recommending items even in the absence of extensive user data.

Disadvantages

However, there are also some disadvantages to content-based filtering. One of the main concerns is that it can suffer from overspecialization. If the algorithm only recommends items that are too similar to the user's past preferences, it can limit the diversity of the recommendations. Content-based filtering can also have the **lack of diversity**. Since it relies on the content of the items, it can only recommend items that are similar in content to items the user has previously liked or interacted with.

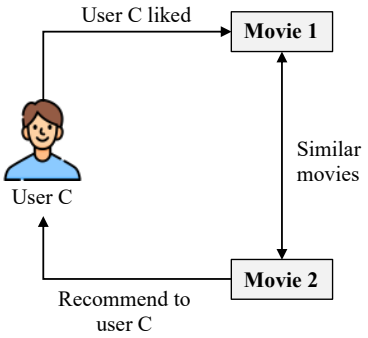

Figure 2: Content-Based Filtering

Hybrid system

To address the challenges of data sparsity, cold start problem, and lack of diversity effectively, a hybrid system that combines collaborative filtering and content-based filtering can be used as shown in Figure 3.

Advantages

One of the primary advantages of hybrid system is that it can provide more diverse recommendations than content-based filtering alone. By incorporating collaborative filtering, the algorithm can recommend items that may not be similar in

content but are still a good fit for the user's preferences. Another advantage of hybrid system is that it can handle the cold-start problem better than either approach alone. By combining the two approaches, the algorithm can make recommendations based on both user behavior and movie content, even if there is limited data on either.

Disadvantages

However, there are also some disadvantages to hybrid system. One concern is that it can be more complex and difficult to implement than either approach alone. The algorithm requires both collaborative filtering and content-based filtering components, which can be challenging to integrate. Another disadvantage of hybrid filtering is that it can still suffer from popularity bias. If the collaborative filtering component relies heavily on user behavior, it may still recommend popular items even if they are not a good fit for the user's preferences.

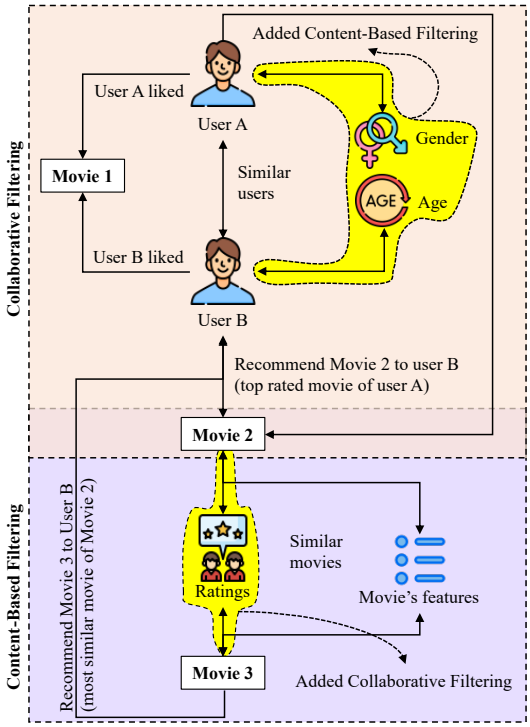

Figure 3: Hybrid system

According to the dataset of MovieLens 1M that contains ratings data of each user, movie's information, and user's preferences, developing the recommendation system by using either Collaborative Filtering technique or Content-Based Filtering technique offers different advantages and disadvantages. By combining these two techniques to develop the hybrid system can offers the advantages and overcome the disadvantages of those 2 techniques.

Step of implementation

Step 1 Data Proprocessing

In the initial stage, an investigation was conducted on the MovieLens dataset comprising three distinct files: ratings.dat, users.dat, and movies.dat. These files contain essential data for the creation of a hybrid recommendation system. To ensure a comprehensive dataset, the data from these separate files were merged into a unified file as shown in Figure 4, consolidating all the required information for the development of the system.

	UserID	MovieID	Ratings	Timestamp	Gender	Age	Occupation	Zixp-code	Title	Genres
0	1	1193	5	978300760	F	1	10	48067	One Flew Over the Cuckoo's Nest (1975)	Drama
1	2	1193	5	978298413	M	56	16	70072	One Flew Over the Cuckoo's Nest (1975)	Drama
2	12	1193	4	978220179	M	25	12	32793	One Flew Over the Cuckoo's Nest (1975)	Drama
3	15	1193	4	978199279	M	25	7	22903	One Flew Over the Cuckoo's Nest (1975)	Drama
4	17	1193	5	978158471	M	50	1	95350	One Flew Over the Cuckoo's Nest (1975)	Drama

Figure 4: Merged dataset

Once the merged file was obtained, we proceeded to extract a dataset specifically tailored for the development of the Collaborative Filtering, Content-based Filtering, and Hybrid recommendation systems. This dataset was created by selecting and including only the relevant features essential for each of these systems. By refining the dataset to contain the necessary attributes, we ensured a focused and efficient implementation of Collaborative Filtering, Content-based Filtering, and the Hybrid system.

Dataset for Collaborative Filtering

To establish the dataset that captures the ratings data between users and movies, we constructed a matrix where each row corresponds to a unique user ID and each column represents a distinct movie ID. The values within the matrix denote the ratings provided by users for the respective user-movie pairs. The visual representation of this dataset structure can be observed in Figure 5.

MovieID	1	2	3	4	5	6	7	8	9	10	...
UserID											
1	5	0	0	0	0	0	0	0	0	0	...
2	0	0	0	0	0	0	0	0	0	0	...
3	0	0	0	0	0	0	0	0	0	0	...
4	0	0	0	0	0	0	0	0	0	0	...
5	0	0	0	0	0	2	0	0	0	0	...
...	...	...	...	...	...	...	...	...	...	...	...

Figure 5: Dataset for Collaborative Filtering system

Dataset for Content-Based Filtering (User's attributes)

To capture user attributes such as age and gender, we developed a matrix with rows representing unique user IDs and columns representing the user's attributes. The matrix entries are numerical values, with 0 indicating "No" and other non-zero values indicating "Yes" Figure 6 visually

represents this matrix structure and the corresponding representation of user attributes.

Gender	F	M	0-10 years	11-20 years	21-30 years	31-40 years	41-50 years	>50 years
UserID								
1	2	0	1	0	0	0	0	0
2	0	1	0	0	0	0	0	6
3	0	1	0	0	3	0	0	0
4	0	1	0	0	0	0	5	0
5	0	1	0	0	3	0	0	0
...	...	...	...	...	...	...	...	...

Figure 6: Dataset for Content-Based Filtering (User's attributes)

Dataset for Content-Based Filtering (Movie's features)

To capture movie features such as genres, we developed a matrix with rows representing unique movie IDs and columns representing the movie's genres. The matrix entries are numerical values, with 0 indicating "No" and other non-zero values indicating "Yes" Figure 7 visually represents this matrix structure and the corresponding representation of user attributes.

	Action	Adventure	Animation	Children's	Comedy	Crime	Documentary	Drama	Fantasy	F
MovieID										
1	0	0	1	1	1	0	0	0	0	
2	0	1	0	1	0	0	0	0	1	
3	0	0	0	0	1	0	0	0	0	
4	0	0	0	0	1	0	0	1	0	
5	0	0	0	0	1	0	0	0	0	
...	...	...	...	...	...	...	...	...	...	

Figure 7: Dataset for Content-Based Filtering (Movie's features)

Step 2 Normalization

Upon obtaining the datasets required for developing each system, we conducted a data normalization step to ensure that all datasets shared a consistent range between 0 and 1. The formula utilized for this normalization was as follows:

$$Normalized\ value = \frac{Actual\ value - Min\ value}{Max\ value - Min\ value}$$

This normalization process aimed to bring different features or variables to a similar scale, thereby mitigating the dominance of any features due to its larger magnitude. To achieve this standardization, we employed the Max-Min normalization technique, also referred to as min-max scaling. By applying this approach, we transformed the values within the datasets while preserving their relative relationships. This normalization technique allowed us to effectively compare and analyze the datasets, facilitating fair comparisons and preventing any bias arising from varying scales or units across features or variables.

Step 3 Matrix Factorization

[4] To address the challenges posed by the high dimensionality and sparsity of the dataset for developing the Hybrid system, we employ the Singular Value Decomposition (SVD) technique. This approach enables us to effectively reduce the dimensionality of the matrix by identifying and capturing the most significant latent features. The formula utilized for SVD was as follows:

$$SVD(A_k) = U_k \Sigma_k V_k^T$$

Where,

$\boldsymbol{U_k}$ and $\boldsymbol{\Sigma_k}$ are $\boldsymbol{m\ x\ k}$ and $\boldsymbol{n\ x\ k}$ matrices composed by the first $\boldsymbol{k}$ columns of matrix $\boldsymbol{U}$ and the first $\boldsymbol{k}$ columns of matrix $\boldsymbol{V}$ respectively. Matrix $\boldsymbol{\Sigma_k}$ is the $\boldsymbol{k\ x\ k}$ principle diagonal sub-matrix of $\boldsymbol{\Sigma}$. $\boldsymbol{A_k}$ represents the closest linear approximation of the original matrix $\boldsymbol{A}$ with reduced rank $\boldsymbol{k}$

Step 4 Hybrid system development

As previously discussed, the Hybrid movie recommendation system developed in this research paper is a fusion of Collaborative Filtering and Content-Based Filtering methodologies. The hybrid system is composed of two sub-hybrid systems, each integrating Collaborative Filtering and Content-Based Filtering techniques.

Within the constructed Hybrid movie recommendation system, one of the sub-hybrid systems functions by recommending movies to a specific user through the utilization of similarity among users. This approach considers various factors including ratings data, age, and gender to identify users who exhibit similarities with the target user and leverages their movie recommendations as a basis for generating personalized suggestions.

The second sub-hybrid system operates by recommending movies based on similarities among movies themselves. By analyzing genres and ratings score of each movie, this approach identifies movies that share similar ratings score, and genres. Recommendations are then made by considering the preferences and viewing history of users who have shown interest in those similar movies.

Hybrid system based on similar users

To construct the hybrid system, we merged the dataset for Collaborative Filtering (Figure 5) with the dataset for Content-Based Filtering (Figure 6). This integration allowed us to leverage both user ratings data and movie features. Subsequently, we employed a similarity measure to identify the most similar users within this combined dataset. To determine the similarity between users, the Cosine similarity technique was employed. This technique allowed us to calculate the similarity score between users based on their shared interests, age, and gender. By leveraging the Cosine similarity measure, we were able to identify the users who exhibited the highest degree of similarity to the target user, thereby providing a foundation for generating personalized

and relevant movie recommendations within the hybrid system. The formula of cosine similarity was as follows:

$$cosine(x, y) = \frac{x^T \cdot y}{||x|| \cdot ||y||}$$

Where,

- $\boldsymbol{x}$ is a vector representing the user preferences.
- $\boldsymbol{y}$ is a vector representing the target user preferences.

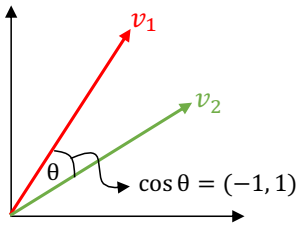

Figure 8: Cosine similarity

The cosine similarity formula is used to measure the similarity between two vectors, with a range of -1 to 1. A similarity score of 1 indicates perfect similarity, while a score of -1 represents perfect dissimilarity. In Figure 8, the cosine similarity is visualized, showcasing the relationship between two vectors. The cosine similarity scores helped us identify the most similar users who exhibited the highest cosine similarity scores. The comparison was based on various factors, including ratings data, age, and gender, ensuring a comprehensive evaluation of similarity between users and the target user.

Hybrid system based on similar movies

The process of developing the hybrid system based on similar movies follows a similar approach to the hybrid system based on similar users but involves different datasets. In the case of the hybrid system based on similar movies, we utilize two datasets: the Collaborative Filtering dataset (Figure 5) and the Content-Based Filtering dataset (Figure 7). To construct the hybrid system based on similar movies, we transpose the Collaborative Filtering dataset, interchanging the rows and columns. This transformation allows us to compare the similarities between movies rather than users. Simultaneously, we incorporate the Content-Based Filtering dataset to consider the attributes and characteristics of movies. By merging these two datasets, we calculate the cosine similarity to identify the most similar movies. The cosine similarity measure provides insights into the degree of similarity between movies within the hybrid system based on similar movies. This comparison aids in establishing meaningful connections between movies and enhancing the effectiveness of the hybrid recommendation system.

Recommendation movies

The two sub-hybrid systems within our research, based on similar users and similar movies, offer distinct movie recommendations for the target user. The former system suggests movies based on the preferences of the most similar user to the target user. By analyzing the movie choices of this highly similar user, the system provides recommendations that align with their tastes and preferences.

On the other hand, the latter system recommends movies that are similar to the top-rated movie of the target user. By identifying the movie that the target user rates most highly, the system utilizes this information to find movies that exhibit similar attributes, genres, or themes. This approach ensures that the recommended movies align with the specific preferences and interests of the target user.

For instance, in the case of user ID 5, the former system would suggest movies liked by the user who shares the highest similarity in terms of preferences, age, and gender. The latter system, on the other hand, would recommend movies that are similar to the top-rated movie of user ID 5. By combining these two sub-hybrid systems, the hybrid movie recommendation system offers a diverse range of personalized movie suggestions tailored to the unique preferences of individual users as shown in Figure 9.

```
Recommended movies from the most similar user of user ID 5:
1. Three Kings (1999)
2. Bridge on the River Kwai, The (1957)
3. Four Days in September (1997)
4. Malcolm X (1992)
5. Raging Bull (1980)

Recommended movies that are the most similar movies to movie ID 2890:
1. Pulp Fiction (1994)
2. Fargo (1996)
3. Good Will Hunting (1997)
4. Truman Show, The (1998)
5. Usual Suspects, The (1995)
```

Figure 9: Example of Hybrid Movie Recommendation System

Step 5 Evaluation

To evaluate the performance of our movie recommendation system, we used two metrics: Mean Absolute Percentage Error (MAPE) and Root Mean Squared Error (RMSE). These metrics were used to measure the accuracy of our system in predicting the average ratings of each user and movie.

The purpose of this evaluation was to determine the potential benefits of combining content-based filtering and collaborative filtering. By doing so, we aimed to create a hybrid system that could recommend movies to both old and new users, with a variety of preferences. Specifically, we wanted to determine whether the inclusion of content-based filtering into traditional collaborative filtering, and vice versa, would result in an improvement, decrease or no change in the performance of our system in predicting the average ratings of each user and movie.

Data preparation for evaluation

In preparation for evaluating the system, we enhanced the dataset for the hybrid system, which relies on similar users, by incorporating the average ratings of each user. Additionally, we augmented the dataset for the hybrid system based on similar movies by including the average ratings of each

movie. These average ratings served as the target variable for prediction in both datasets.

Implement Linear Regression model for predictions

For implementing the machine learning model, we divided each dataset into two parts: a training dataset comprising 70% of the data and a testing dataset comprising 30% of the data. Subsequently, we applied Singular Value Decomposition (SVD) to the training dataset before implementing the machine learning model. The machine learning model chosen for predictions was Linear Regression, which is a statistical modeling technique used to establish a relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the variables, meaning that the change in the dependent variable is proportional to the change in the independent variable(s). The goal of linear regression is to find the best-fit line that minimizes the difference between the predicted values and the actual values of the dependent variable. Linear regression can be extended to multiple independent variables, resulting in multiple linear regression. The equation becomes:

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n$$

Where,

- y represents the dependent variable.
- $x_1, x_2, \ldots, x_n$ represent the independent variables.
- b_0 is the y-intercept.
- $b_1, b_2, \ldots, b_n$ are the coefficients corresponding to each independent variable.

The results of the MAPE and RMSE metrics for the average ratings predictions of both hybrid systems, achieved through the utilization of the Linear Regression model, are presented in Figure 10.

	MAPE	RMSE
CF	0.094009	0.181991
CF with added CBF	0.095074	0.184299

Hybrid system based on similar users

	MAPE	RMSE
CBF	0.178820	0.389887
CBF with added CF	0.118153	0.221191

Hybrid system based on similar movies

Figure 10: Result for both hybrid systems

Conclusion

In conclusion, our evaluation of the hybrid system for movie recommendations demonstrates the effectiveness of incorporating Collaborative Filtering (CF) and Content-Based Filtering (CBF) techniques. The addition of CF to CBF significantly improved the accuracy of predicting average movie ratings compared to the traditional CBF system alone, as evidenced by a reduction in the MAPE of approximately 0.06, and a reduction in the RMSE of approximately 0.17.

These findings suggest that a hybrid recommendation system combining CF and CBF can effectively recommend diverse types of movies to both new and existing users. By leveraging collaborative filtering and incorporating user ratings data, the hybrid system can provide more accurate and comprehensive recommendations that cater to individual preferences, leading to higher user satisfaction.

To enhance the effectiveness of movie recommendations further, it is important to continue exploring optimal combinations of recommendation techniques and further refine the hybrid system. Continued research in this area will contribute to the development of advanced recommendation algorithms that better serve the needs and preferences of users.

Reference

[1] Pradhan, R., Swami, A.C., Saxena, A., & Rajpoot, V. 2021. A Study on Movie Recommendations using Collaborative Filtering. IOP Conference Series: Materials Science and Engineering (Vol. 1119, p. 012018). 10.1088/1757-899X/1119/1/012018

[2] Reddy, SRS, Nalluri, S., Kunisetti, S., Ashok, S., & Venkatesh, B. 2019. Content-Based Movie Recommendation System Using Genre Correlation: Proceedings of the Second International Conference on SCI 2018, Volume 2. 10.1007/978-981-13-1927-3_42.

[3] Geetha, G.P., Safa, M., Fancy, C., & Saranya, D. 2018. A Hybrid Approach using Collaborative filtering and Content-based Filtering for Recommender System. Journal of Physics: Conference Series, 1000. National Conference on Mathematical Techniques and its Applications (NCMTA 18), 5–6 January 2018, Kattankulathur, India. Published under licence by IOP Publishing Ltd. 10.1088/1742-6596/1000/1/012101.

[4] Singh, R., Maurya, S., Tripathi, T., Narula, T., & Srivastav, G. 2020. Movie Recommendation System using Cosine Similarity and KNN. International Journal of Engineering and Advanced Technology. 9. 2249-8958. 10.35940/ijeat.E9666.069520.

[5] Bokde, D., Girase, S., & Mukhopadhyay, D. 2015. Matrix Factorization Model in Collaborative Filtering Algorithms: A Survey. Procedia Computer Science. 49. 10.1016/j.procs.2015.04.237.

[6] Herlocker, J., Konstan, J., Terveen, L., Lui, J.C.s., & Riedl, T. 2004. Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems. 22. 5-53. 10.1145/963770.963772.

[7] Harper, F. M., & Konstan, J. A. 2015. The MovieLens Datasets: History and Context. ACM Transactions on Interactive Intelligent Systems (TiiS) 5, 4 (December 2015), 19 pages. http://dx.doi.org/10.1145/2827872