

Enhancing Movie Recommendation Systems with Hybrid Collaborative Filtering, Content-based Filtering and SVD

Liheng Xu^{1,†}, Zhile Guan^{2,†}, Yu Wu^{3,†}
{1789139850@qq.com¹, kuan012@qq.com², 664812167@qq.com³}

School of Statistics and Data Science, Nankai University, Tianjin, 300071, China¹

School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, 430074, China²

School of Economics and Management, Xidian University, Xi'an, 710126, China³

[†]These authors contributed equally to this work.

Abstract. The swift advancement of Internet technology has intensified the issue of information overload. Consequently, users face difficulties in efficiently navigating extensive data and locating content tailored to their interests. To handle this issue, this paper proposes an enhanced movie recommendation system that leverages a hybrid approach combining collaborative filtering, content-based filtering, and Singular Value Decomposition (SVD). By analyzing the MovieLens dataset, we identify critical features and develop hybrid models that aim to improve upon existing methods by harnessing the strengths of each filtering technique. Our hybrid methods seek to bypass the constraints of individual filters, enhancing prediction accuracy and recommendation relevance. The experiments demonstrated that combining these techniques significantly reduces Root Mean Square Error (RMSE) and Mean Absolute Error (MAE), indicating improved performance. Our findings show that hybrid approaches improve movie recommendation systems. They offer more personalized and accurate recommendations. However, there may be limitations to these methods. Therefore, future research should investigate these limitations and work on refining the techniques. The goal is to achieve even better performance.

Keywords: Movie recommendation, Hybrid filtering algorithm, Collaborative filtering, Content-based, SVD

1 Introduction

Movie recommendation systems have become a key element in the field of personalized services, in large part due to their ability to cater to the diverse and ever-changing tastes of their users. The proliferation of online streaming platforms and the rapid growth in the number of movies have necessitated the development of sophisticated recommendation algorithms. The focus of this research is to enhance movie recommendation systems through an exhaustive examination of a hybrid approach that incorporates collaborative filtering, content-based filtering, and singular value decomposition (SVD) [1].

The main goal of this research is to design an enhanced movie recommendation system that leverages hybrid techniques to improve prediction accuracy and recommendation relevance. To achieve this objective, we pose the following research questions:

1. How does the standalone performance of collaborative filtering, content-based filtering, and SVD compare for movie recommendations using the MovieLens dataset?
2. Does combining collaborative filtering and content-based filtering improve prediction accuracy?
3. How does including SVD in hybrid models affect movie recommendation performance?
4. What is the optimal hybrid configuration combining collaborative filtering, content-based filtering, and SVD?

By integrating collaborative filtering, content-based filtering, and SVD, our study hypothesizes that the hybrid approach will significantly improve the prediction accuracy and personalization of movie recommendations compared to using individual filtering techniques in isolation. Utilizing the MovieLens dataset, which is a standard in recommendation system research, adds credibility to our study and allows for comparability with other studies. Through a detailed thematic analysis and comparison of performance across methods and studies, we aim to demonstrate the value of our work and make contributions to the development of movie recommendation systems [2].

2 Literature Review

Previous research in this area has focused on utilizing either collaborative filtering or content-based filtering approaches [3]. Collaborative filtering algorithms rely on similarities between users' past ratings or behaviors and have been effectively implemented by platforms such as Netflix and Amazon [4]. These algorithms infer user preferences for novelty items by utilizing the interests of comparable users. Meanwhile, content-based filtering techniques, which recommend movies similar to a user's previous favorites based on attributes such as genre, director, or actor, have been effectively utilized by platforms such as Hulu and YouTube. While these approaches have shown considerable promise, they all have inherent limitations [5]. Collaborative filtering struggles with data sparsity and cold-start issues, while content-based filtering may ignore collaborative signals that may lead to more diverse recommendations. To address these challenges, researchers have increasingly explored hybrid recommendation systems that combine multiple techniques to leverage their respective strengths [6].

Hybrid recommendation systems have gained significant attention in recent years due to their ability to overcome the limitations of individual filtering methods. Several studies have investigated the combination of collaborative filtering and content-based filtering, demonstrating that hybrid models can often achieve better performance than either method alone [7]. For instance, Jung et al. proposed a hybrid model that combines collaborative filtering and content-based filtering using weighted summation, showing improved recommendation accuracy compared to standalone methods [8]. Similarly, Geetha et al. developed a hybrid approach that integrates user-based and

item-based collaborative filtering with content-based filtering, reporting enhanced performance in terms of both precision and recall [5].

Moreover, the integration of matrix factorization techniques, such as Singular Value Decomposition (SVD), into hybrid recommendation systems has also been explored. SVD is a complex matrix decomposition technique that can decompose the user-item rating matrix into smaller, more manageable components, uncovering underlying patterns and relationships that are not immediately apparent. The combination of SVD with collaborative filtering or content-based filtering has been shown to further improve recommendation accuracy. For example, Afoudi et al. proposed a hybrid model that combines content-based filtering with SVD-based collaborative prediction using an artificial neural network, achieving state-of-the-art results on several benchmark datasets [1].

Inspired by these seminal works, our research endeavors to overcome the constraints inherent in individual filtering methodologies by investigating hybrid frameworks that synergize the advantages of collaborative filtering, content-based filtering, and Singular Value Decomposition (SVD). Our work builds upon the foundational research conducted by various scholars, particularly those who have studied the efficacy of hybrid recommendation systems. However, a huge gap remains in the literature regarding the optimal configuration of hybrid models that combine these three techniques [8].

3 Data and Exploratory Data Analysis

The datasets used for this study was obtained from the MovieLens website, a well-known source for movie rating data frequently used in recommendation system research [9]. The datasets, called ml-100k, contains extensive information on user interactions with movies, including datasets like user items, user-base, user-test etc.

For this project, the datasets contained the following parts that are actually used:

u1.base and u1.test: the training set and the testing set for the future work

u.data: It consists of 10,000 ratings from 943 users for 1682 movies. Each user rates at least 20 movies. Users and movies are numbered sequentially starting with number 1. The data is sorted in a random way. Label delimited list: user id — item id — rating — timestamp

So, about the dataset, it has a medium size, which enables us to increase our conclusion's reliability compared to smaller size datasets, meanwhile a modest dataset wouldn't be time-consuming [10]. As a result it is perfectly suited for our project.

Upon examining the dataset, it was observed that a small percentage of the data contained missing values. Since the missing data was minimal and did not constitute a significant portion of the overall dataset, it is better to remove the records with missing values rather than attempting to impute them with mode, because it was time data that was missing and fill in with mode would help nothing [11]. This decision was made to maintain the integrity and simplicity of our analysis, assuring that the data utilized for training and testing the recommendation algorithms was as clean and accurate as possible. Another point needing to be mentioned was that the preprocessing of the datasets was already done since it is well known that others have already been using the datasets in their papers directly without some cleaning or standardization, so it was acceptable that the data quality was good and capable for our research project.

4 Model Analysis

As embarked on the construction of the model, it was essential to first examine the standalone performance of collaborative filtering, content-based filtering, and Singular Value Decomposition (SVD). Subsequently, plans were made to explore hybrid methods by combining these techniques in various configurations. The question arose whether to hybridize each pair of methods or integrate all three together. In the research, primary emphasis was placed on the content-based method, aiming to hybridize it with both collaborative filtering and SVD separately, and ultimately to combine all three methods, in the hope of enhancing the performance of the recommendation system [12].

And it is also necessary to mention the consideration of features. This is closely linked to the choice of hybridizing the three methods mentioned above. To create a recommendation system using hybrid methods, it is crucial to identify the relevant features to work on. Working on inconsequential features may result in a model with poor performance. To select the important features, the lightGBM algorithm, which belongs to the boosting family, was used. The main theory of lightGBM, similar to XGboost, is to use the negative gradient of the loss function as an estimate of the residual of the current decision tree to form a new one. However, in this case, lightGBM outperformed XGboost due to its higher training efficiency [13]. After the iteration, the 3 most important features for the model were successfully selected.

4.1 Measurement

There are various approaches to determine the quality of a recommendation system, such as AUC or precision. In this study, the performance of the proposed hybrid recommendation system is measured using two widely recognized evaluation metrics: Root Mean Square Error (RMSE) and Mean Square Error (MSE). These metrics are particularly suitable for evaluating the accuracy of predicted ratings against actual user ratings in recommendation systems. In the proposed hybrid recommendation system, RMSE and MSE are used to assess the effectiveness of the model across different configurations, including pure collaborative filtering, content-based filtering, SVD-based predictions, and their combinations. By comparing RMSE and MSE values across these methods, the aim is to identify the optimal configuration that minimizes prediction errors and enhances overall recommendation accuracy. The goal is to achieve lower RMSE and MSE values in the hybrid model, indicating that the integration of collaborative filtering, content-based filtering, and SVD results in improved predictive accuracy compared to each method used in isolation.

4.2 Hypothesis testing

To assess the effectiveness of the proposed hybrid recommendation system, hypothesis testing was carried out to determine whether the hybrid approach (combining collaborative filtering, content-based filtering, and SVD) significantly improves prediction accuracy compared with using content-based filtering alone. It's natural that the null hypothesis is that the hybrid recommendation system does not make a significant difference, while the alternative hypothesis is that the hybrid recommendation system does make a significant difference.

4.3 Content-based Filtering

As mentioned before, other methods were hybridized with content-based filtering. This was because it is the most important feature of the dataset and it is common in daily life to receive recommendations for similar content. The core theory behind content-based filtering is based on the assumption that items can be described by a set of features or attributes, and these attributes can be used to infer user preferences. Each item (in this case, a movie) is represented by a feature vector, a collection of attributes that describe the item. In this case, a movie can be described by its genres, movieID, and ratings.

When developing the model, the most significant part of the method is the cosine similarity, a widely used tool to evaluate the similarity between two items, more specifically, the user's vector and the feature vectors of movies in the dataset [14]. The formula is below:

$$\text{CosineSimilarity} = \frac{\vec{A} \cdot \vec{B}}{\|\vec{A}\| \cdot \|\vec{B}\|} \quad (1)$$

The result of the content-based filtering is that the MAE and RMSE are 0.8532 and 1.0898 respectively. Figure 1 below clearly shows the flow of content-based filtering.

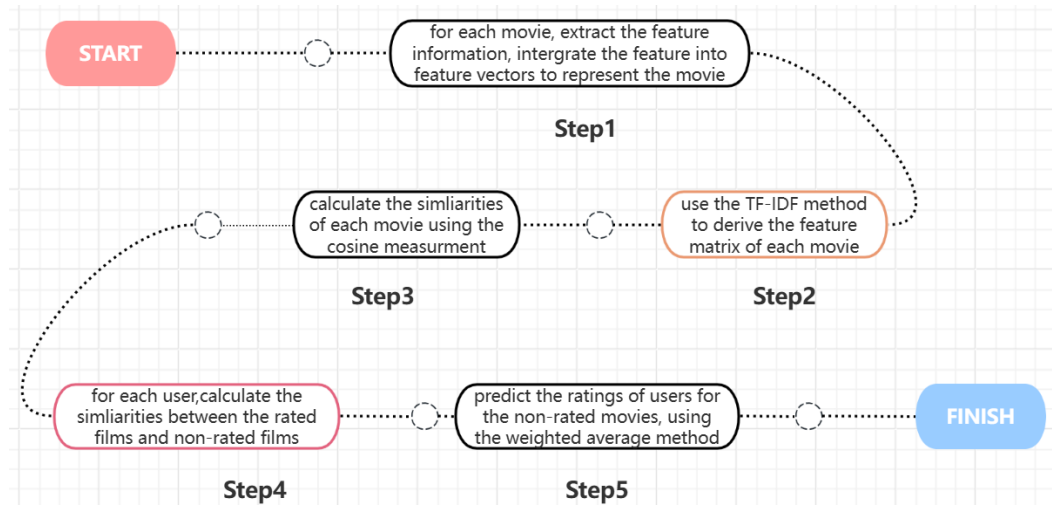


Fig. 1. Content -based Filtering Flow Chart

$$\hat{r}_{u,i} = \frac{\sum_{j \in N(u)} \text{sim}(i,j) \cdot r_{u,j}}{\sum_{j \in N(u)} \text{sim}(i,j)} \quad (2)$$

The formula (2) tells how to calculate the predicted values, the term $\text{sim}(i,j)$ refers to the cosine similarity between movie i and j while $N(u)$ denotes the set that user u has rated. So it is quite apparent that based on the known ratings $r_{u,j}$, we can obtain the desired one $\hat{r}_{u,i}$ [15].

4.4 Hybrid Collaborative Filtering and Content-Based Filtering

Before the hybridization, it was necessary to know that KNN (k-nearest-neighbors) was being used as the tool for collaborative filtering. This raises the question: what is the optimal k value? It is known that the initial k value is quite important for KNN, as an inappropriate k may cause the algorithm to not converge, let alone result in a low-quality model. Cross-validation will be applied to choose the optimal k without the need to resplit the dataset, as a train-set and a test-set are already available. A simple conclusion was first reached that the k value should be moderate. Since a k value greater than 100 or less than 5 results in a higher RMSE and MAE, the desired k value should be in the interval of 5 to 100. Then, a vector was created to store all possible k values in that interval, and a for-loop iteration code was used to select the one with the lowest arithmetic mean of RMSE and MAE. The optimal k value obtained was 70 [16].

After the k value is determined, the hybrid method can proceed. The collaborative filtering model was retrained using this k on the entire training dataset. The predictions from this optimized model were then integrated with the content-based filtering predictions to form the final hybrid recommendation. The output of the MAE and RMSE are 0.8199 and 1.0360, respectively. Figure 2, shown below, clearly depicts the flow of KNN collaborative Filtering.

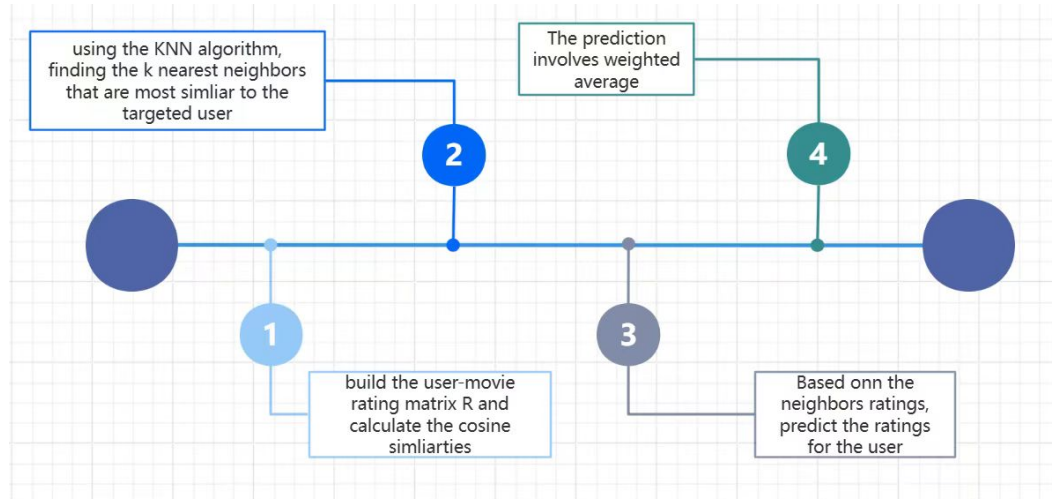


Fig. 2. KNN Collaborative Filtering Flow Chart

$$\hat{r}_{u,i}^{\text{kNN}} = \frac{\sum_{v \in N(u)} \text{sim}(u, v) \cdot r_{v,i}}{\sum_{v \in N(u)} |\text{sim}(u, v)|} \quad (3)$$

$$\hat{r}_{u,i}^{\text{ItemCF}} = \frac{\sum_{j \in N(i)} \text{sim}(i, j) \cdot r_{u,j}}{\sum_{j \in N(i)} |\text{sim}(i, j)|} \quad (4)$$

The formula (3) and (4), have many resemblances, the main idea is like formula (2), but the method to derive the rating $r_{u,j}$ and the similarity $\text{sim}(i,j)$ is different, one involves KNN algorithm, while the other one involves the item collaborative filtering.

4.5 Hybrid SVD and Content-based Filtering

Singular Value Decomposition (SVD) is a powerful matrix factorization technique used in statistics to decompose a large user-item matrix into smaller, more manageable components. Now since we have transformed abstract data into vectors and matrices, the SVD is capable to be applied in the recommendation system. The core idea behind SVD is to reduce the dimensionality of the data while preserving the most significant information. This is particularly useful in dealing with sparse matrices.

The original user-item rating matrix R can be decomposed into three multiples $R \approx U \Sigma V^T$ and here the dimension of V is $k \times k$, where k is a hyperparameter needed to be chosen carefully. However, the entire process of choosing this k is parallel to that in the KNN method. A primary interval was set, an iteration was built, and the best option was chosen based on RMSE and MAE. The optimal k value obtained was 5. The hybrid model that combines SVD with content-based filtering takes advantage of the strengths of both approaches. SVD excels at capturing latent relationships in the data that are not immediately apparent from the raw ratings, while content-based filtering ensures that recommendations are aligned with the user's known preferences based on movie attributes. The output of the MAE and RMSE are 0.8178 and 1.0344, respectively. The following figure, Figure 3, shows the flow of SVD.

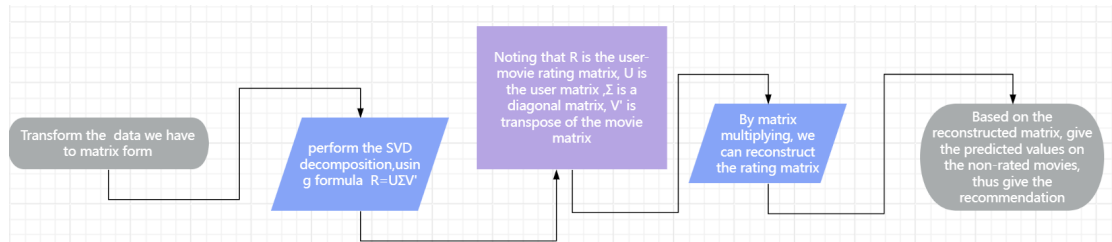


Fig. 3. SVD Flow Chart

4.6 Hybrid Collaborative Filtering, Content-Based Filtering, and SVD

The pairwise combination was tried before, and it was eager to determine if combining all three methods would yield better performance. We have tried the pairwise combination before and we are eager to find out if the three methods combined together would perform better. By integrating these methods, the hybrid model aims to compensate for the weakness of every individual technique. Collaborative filtering can struggle with sparse data, content-based filtering might overfit to user history without incorporating collaborative signals, and SVD alone might miss specific attribute-level preferences. The hybrid approach addresses these challenges by combining all three techniques. The

optimal k is 5 and 60 for SVD and KNN. The output of the MAE and RMSE are 0.8135 and 1.0284, respectively.

$$\hat{r}_{u,i} = 0.5 \cdot \hat{r}_{u,i}^{\text{kNN}} + 0.5 \cdot \hat{r}_{u,i}^{\text{ItemCF}} \quad (5)$$

Having conducted experiments with different algorithms, the hybrid model is now ready to be introduced. The final target for building the recommendation system is the rating, which is a mixed weighted prediction value considering both KNN and ItemCF.

4.7 Evaluation and comparison against other methods

When it comes to building a recommendation system, given the ml-100k dataset, there is no doubt that item-based collaborative filtering algorithm and user-based collaborative filtering algorithm are the most widely applied, since they directly work on the dataset with the fewer steps of procedure. Having been running these algorithms on the dataset, it is obtained that the RMSE and MAE of the item-based is 0.8498 and 1.0634, respectively. And the RMSE and MAE for the user-based is 0.8261 and 1.0313, respectively. It is apparent that both algorithms have outperformed the initial chosen algorithm, the content-based filtering. However, the situation has reversed. All the hybrid model have shown a decrease in MAE compared to the mentioned two algorithms. The RMSE is a little different, where only the hybrid collaborative filtering, content-based filtering and SVD is better than the user-based. But consider that our final target is the hybrid model that combines all the three methods, the advancement has been well elaborated.

5 Results and Interpretations

5.1 The Result Table

Table 1: Result Table

	MAE	RMSE
Content-based	0.8532	1.0898
Item-Based Collaborative Filtering	0.8498	1.0624
User-Based Collaborative Filtering	0.8261	1.0313
Hybrid Content-based and SVD	0.8199	1.0360
Hybrid Content-based and Collaborative Filtering	0.8178	1.0344
Hybrid Collaborative Filtering, Content-Based Filtering, and SVD	0.8135	1.0284

5.2 The Interpretations

From the table above, it is obvious that after conducting the hybrid methods, the MAE and RMSE have reduced to some extent, which proves the primary correctness of our ideas. The hybrid of content-based and SVD is close to the hybrid content-based collaborative filtering, which reduces

MAE and RMSE 4.1% and 4.3%, respectively [17]. The model of the three combined together performed the best, but the reduction of MAE is not so apparent, while it takes a further step in reducing the RMSE by 5.6%. If a rejecting threshold of 5% is taken, the null hypothesis will be rejected and the alternative one accepted for the hybrid model. It must be admitted that, from a statistical perspective, a general conclusion cannot be drawn on the meaningful improvement of the hybrid model, and the hypothesis may seem hasty. This is because the measurement is changeable, and different ways of measuring may result in different properties. The aim is to show readers an improvement in the hybrid methods, but proving statistical superiority is not part of the plan.

An important aspect of our project that must be mentioned is the hyperparameter of the final hybrid model. When combining two methods, such as KNN and SVD or content-based methods, the value of the hyperparameter differs from its optimal value when used alone. This is because there may be interactions between the methods. Efforts are being made to train an ideal value for the hyperparameters. The graphic results are presented here: Figure 4 shows the optimal k value for content-based and SVD, while Figure 5 shows the optimal k value for content-based and collaborative filtering.

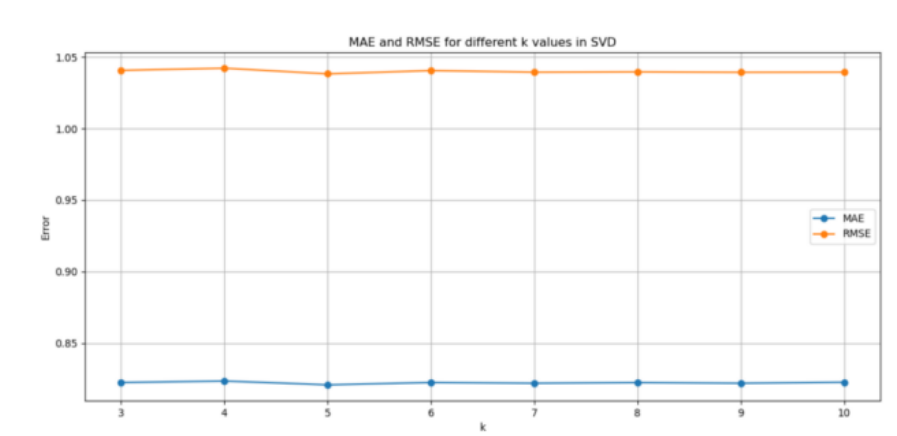


Fig. 4. Content-based and SVD The optimal k is 5

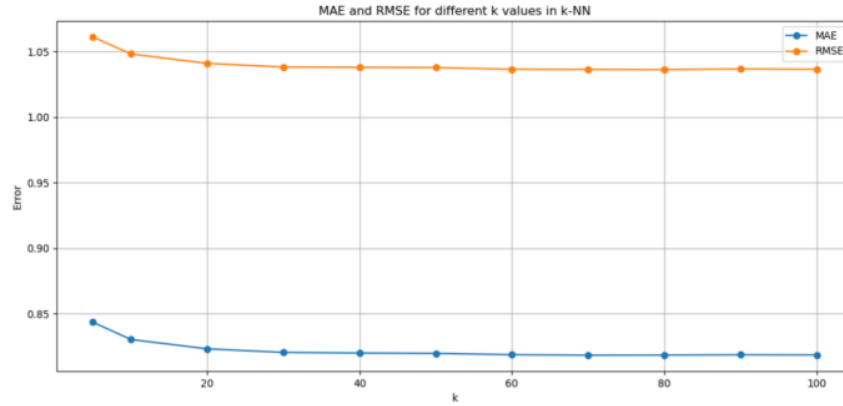


Fig. 5. Content-based and collaborative filtering The optimal k is 70

6 Conclusions and Future Developments

6.1 Conclusions

The outcomes of this study have significant ramifications for the field of recommendation systems. By demonstrating the superiority of hybrid approaches that combine collaborative filtering, content-based filtering, and Singular Value Decomposition (SVD), our research emphasizes the significance of utilizing diverse information sources and techniques to maximize prediction accuracy and user satisfaction. [18] These insights can guide practitioners in the advancement of more effective recommendation systems that meet the needs of the varying demands and desires of users.

6.2 Limitations

While our study offers useful perspectives into the capability of different recommendation techniques, it is not without limitations. One notable limitation is the methodological focus on algorithmic improvements without substantial consideration of contextual factors or user feedback mechanisms [19]. Incorporating such elements could further personalize recommendations and augment the overall user experience. Additionally, our study utilized a medium-sized dataset, and the performance of the hybrid model on much larger datasets remains to be seen.

6.3 Broader Implications

Based on this study, future research should explore the integration of deep learning techniques into recommender systems. Deep learning has shown great promise in different domains, and applying it to recommender systems may yield better performance. However, implementing deep learning methods in real-world systems also faces a number of challenges, such as the need for large amounts of labeled data, computational resources, and deep learning expertise.

To overcome these obstacles, future research may delve into the application of transfer learning and pre-trained models that utilize knowledge gained from related tasks to improve target performance. In addition, researchers could explore the use of unsupervised and semi-supervised learning methods to optimize the use of unlabeled data and reduce the reliance on labeled data.

In addition, future research should look more deeply into incorporating contextual factors and user feedback mechanisms into recommender systems. This may involve developing models that can learn from user behaviors and preferences, as well as incorporating contextual information such as time, place, and user sentiment to provide more personalized and contextually relevant recommendations [20].

In summary, while this study provides valuable insights into the performance of different recommendation techniques, there is still much room for improvement and exploration. Future research should focus more on addressing the limitations of existing approaches, exploring new strategies such as deep learning, and incorporating contextual factors and user feedback mechanisms to develop more impactful and personalized recommender systems.

7 Acknowledgements

Thanks to the guidance of professor Divakaran Liginlal and the teaching assistant. Without their help, it would be difficult for us to finish the project. And thanks to each member of our group, they equally contributed to the work and should all be considered the co-first author.

References

- [1] Yassine Afoudi, Mohamed Lazaar, and Mohammed Al Achhab. Hybrid recommendation system combined content-based filtering and collaborative prediction using artificial neural network. *Simulation Modelling Practice and Theory*, 113:102375, 2021.
- [2] Alkiviadis G. Akritas and Gennadi I. Malaschonok. Applications of singular-value decomposition (svd). *Mathematics and Computers in Simulation*, 2004.
- [3] Erion Çano and Maurizio Morisio. Hybrid recommender systems: A systematic literature review. *Intelligent data analysis*, 2017.
- [4] Sherin Eliyas and P Ranjana. Recommendation systems: Content-based filtering vs collaborative filtering. In *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, pages 1360–1365. IEEE, 2022.
- [5] G Geetha, M Safa, C Fancy, and D Saranya. A hybrid approach using collaborative filtering and content based filtering for recommender system. In *Journal of physics: conference series*, 2018.
- [6] Mahesh Goyani and Neha Chaurasiya. A review of movie recommendation system: Limitations, survey and challenges. *ELCVIA: electronic letters on computer vision and image analysis*, 19(3):0018–37, 2020.

- [7] Andreas Höcker and Vakhtang Kartvelishvili. Svd approach to data unfolding. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 1996.
- [8] Kyung-Yong Jung, Dong-Hyun Park, and Jung-Hyun Lee. Hybrid collaborative filtering and content-based filtering for improved recommender system. In *International Conference on Computational Science*, pages 295–302. Springer, 2004.
- [9] Manoj Kumar, DK Yadav, Ankur Singh, and Vijay Kr Gupta. A movie recommender system: Movrec. *International journal of computer applications*, 124(3), 2015.
- [10] Sri Hari Nallamala, Usha Rani Bajjuri, Sarvani Anandarao, D Durga Prasad, and Pragnaban Mishra. A brief analysis of collaborative and content based filtering algorithms used in recommender systems. In *IOP Conference Series: Materials Science and Engineering*, volume 981, page 022008. IOP Publishing, 2020.
- [11] Govindarajan Parthasarathy and Shanmugam Sathiya Devi. Hybrid recommendation system based on collaborative and content-based filtering. *Cybernetics and Systems*, 54(4):432–453, 2023.
- [12] Jean E Roberts and J-M Thomas. Mixed and hybrid methods. *Elsevier*, 1991.
- [13] Debadrita Roy and Arnab Kundu. Design of movie recommendation system by means of collaborative filtering. *International journal of emerging technology and advanced engineering*, 3(4):67–72, 2013.
- [14] Ramni Harbir Singh, Sargam Maurya, Tanisha Tripathi, Tushar Narula, and Gaurav Srivastav. Movie recommendation system using cosine similarity and knn. *International Journal of Engineering and Advanced Technology*, 9(5):556–559, 2020.
- [15] V Subramaniaswamy, R Logesh, M Chandrashekhar, Anirudh Challa, and Varadarajan Vijayakumar. A personalised movie recommendation system based on collaborative filtering. *International Journal of High Performance Computing and Networking*, 10(1-2):54–63, 2017.
- [16] Sharma Sunny, Rana Vijay, and Malhotra Manisha. Automatic recommendation system based on hybrid filtering algorithm. *Education and Information Technologies*, 2022.
- [17] Urvish Thakker, Ruhi Patel, and Manan Shah. A comprehensive analysis on movie recommendation system employing collaborative filtering. *Multimedia tools and applications*, 80(19):28647–28672, 2021.
- [18] Poonam B Thorat, Rajeshwari M Goudar, and Sunita Barve. Survey on collaborative filtering, content-based filtering and hybrid recommendation system. *International Journal of Computer Applications*, 2015.
- [19] Yonghong Tian, Bing Zheng, Yanfang Wang, Yue Zhang, and Qi Wu. College library personalized recommendation system based on hybrid recommendation algorithm. *procedia cirp*, 2019.

- [20] Zahra Zamanzadeh Darban and Mohammad Hadi Valipour. Ghrs: Graph-based hybrid recommendation system with application to movie recommendation. *Expert Systems with Applications*, 2022.