Personalized Movie Recommendation System

Zeynep Meriç Aşık
Artificial Intelligence Engineering
TOBB ETU
Ankara, Turkiye
zeynepmericasik@etu.edu.tr

***Abstract*—This project aims to develop a personalized movie recommendation system using a dataset comprising user ratings for films. The primary methodology employed is Collaborative Filtering (CF), enhanced by advanced matrix factorization techniques such as Singular Value Decomposition (SVD) and Non-negative Matrix Factorization (NMF) to improve prediction accuracy and handle the inherent sparsity of user-item interactions effectively. The objective is to generate tailored film recommendations for existing users by analysing their interaction patterns within the dataset. This report details the implementation process, evaluates the performance of the employed methods through comparative analysis, and discusses the efficacy of SVD and NMF in refining the recommendations provided by the CF model. The findings aim to contribute to the ongoing development of more precise and user-specific recommendation systems in the entertainment domain.**

***Index Terms*—Collaborative Filtering, NMF, SVD, Recommendation Systems**

I. Introduction

The rise of digital media has significantly transformed how people access and engage with movies, resulting in an urgent need for efficient and user-specific recommendation systems. These systems not only enhance user experience by providing personalized content but also boost platform retention rates and user satisfaction. The current project utilizes a widely recognized movie dataset to develop a recommendation system employing Collaborative Filtering (CF), a popular technique in recommendation systems known for its effectiveness in predicting user preferences based on past interactions.

The initial phase of the project involved data preprocessing to ensure the quality and usability of the dataset for further analysis. This included cleaning the data and making necessary modifications to perform both visualization and the training process. Next step was to transform the cleaned data into two pivotal structures: the user-item and item-user matrices. These matrices are vital in implementing CF, as they allow the system to interpret and analyse the relationships between users and items effectively.

For the experiments and model training, the ratings and movies data were utilized from the dataset. This involved applying advanced matrix factorization techniques such as Singular Value Decomposition (SVD) and Non-negative Matrix Factorization (NMF). These techniques are significant for enhancing the recommendation system's capability to deal with the sparsity of the data and to improve the accuracy of the predictions.

The final goal of this project is to refine the predictive accuracy of the recommendation system, thereby making it better at discerning and aligning with individual user preferences. This report will detail the methodologies employed, the experiments conducted, and the comparative analysis of the different techniques used, ultimately evaluating their effectiveness in achieving a robust movie recommendation system.

II. Related Work

The development of efficient recommendation systems requires careful consideration of the data used. Various methods have been adopted in literature to refine and prepare data for such systems, ensuring accuracy and relevance in the recommendations generated. This section discusses several key techniques and methodologies that align with and have inspired the approaches used in this project.

The first step in any recommendation system project involves data cleaning and preprocessing. It's common to remove items with insufficient interactions to ensure the quality of the dataset. This approach helps in focusing the system on items with enough user feedback, thereby enhancing the predictiveness of the model. The method of determining a threshold for inclusion based on the cumulative sum of ratings has been widely discussed, although specifics can vary across implementations [1].

Post-cleaning, the integration of distinct datasets such as movies and ratings is essential to form a comprehensive view. This merged dataset is then transformed into matrices—specifically user-item and item-user matrices—that facilitate the application of collaborative filtering techniques. These matrices are foundational for identifying and predicting user preferences based on existing data [2].

The distinction between item-based and user-based collaborative filtering (CF) [3] forms a significant part of the literature, with studies highlighting their respective strengths and suitability depending on the context and nature of the dataset. To enhance the CF approach, advanced techniques such as Singular Value Decomposition (SVD) [4] and Non-negative Matrix Factorization (NMF) [5] are often employed. These methods are particularly valued for their ability to handle large-scale data by reducing dimensionality and uncovering latent factors in user-item interactions.

Finally, the evaluation and comparison of different recommendation models are critical to understanding their effectiveness and applicability. Common metrics used include accuracy, precision, recall, and F1-score, among others. These metrics help in assessing the performance of various models in real-world scenarios, guiding improvements and the selection of the most appropriate model [6]. This collection of work provides a robust framework for understanding and applying various techniques in the creation and evaluation of recommendation systems, particularly in the context of movie recommendations.

III. Methodology

This section describes the systematic steps taken to develop a robust movie recommendation system using collaborative filtering (CF) and matrix factorization techniques outlining the process from data preprocessing to the application of various recommendation algorithms and the evaluation of their performance.

A. Data Preprocessing

The initial step involved the utilization of a popular movie dataset. The dataset was first subjected to a thorough cleaning process to remove any null values and ensure data integrity. Important columns were visualized to understand distributions, which informed further preprocessing steps. Not all movies had enough ratings to be statistically significant; thus, a threshold for the number of ratings was established using a cumulative sum approach. This threshold helped in filtering out movies with minimal ratings, which could skew the analysis and affect the accuracy of the recommendation system.

B. Data Transformation and Matrix Formation

After cleaning, the movies and ratings data were merged, excluding columns from the movie data that did not contribute meaningful information for predictive modelling. This refined dataset was then transformed into two primary structures crucial for CF: the user-item matrix and the item-user matrix. These matrices serve as the backbone for the collaborative filtering process, where rows represent individual users and columns represent movies, with matrix values denoting the ratings given by users to movies.

C. Application of Collaborative Filtering Techniques

The project employed both user-based and item-based collaborative filtering techniques. The distinction between these methods was explored to understand their respective effectiveness based on the dataset's characteristics. User-based CF considers the preferences of similar users, while item-based CF focuses on the similarity between items rated by users. This exploration helped in determining the appropriate model based on the dataset's unique traits.

D. Advanced Matrix Factorization Techniques

To enhance the CF models, advanced matrix factorization techniques such as Singular Value Decomposition (SVD) and Non-negative Matrix Factorization (NMF) were utilized. SVD helps in decomposing the user-item matrix into component matrices that reveal latent factors associated with users and items. Similarly, NMF was applied to decompose the rating matrix into non-negative factors, facilitating the discovery of inherent patterns in user preferences and item characteristics.

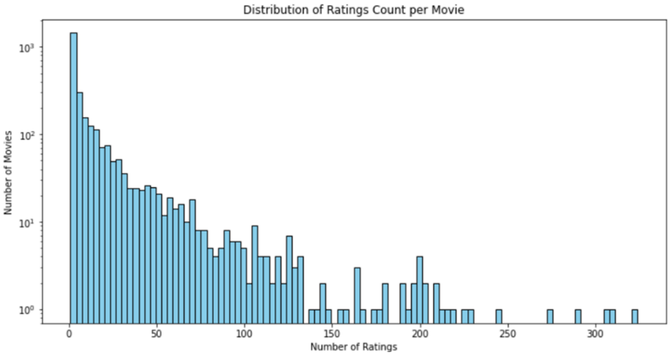

Fig. 1. Distribution of Ratings

E. Model Training and Evaluation

The CF models, along with SVD and NMF, were trained using the ratings data. The evaluation of these models was conducted by comparing their ability to predict ratings accurately. This involved calculating performance metrics such as RMSE (Root Mean Square Error) and MAE (Mean Absolute Error) to assess each model's predictive accuracy.

F. Comparison and Analysis

The final phase of the project involved a comparative analysis of all models to identify the most effective technique for the dataset. The differences in performance highlighted the strengths and limitations of each method, guiding the selection of the best approach for the movie recommendation system.

IV. Results

The project's dataset [7] is divided into different files consisting of a list of around 45,000 movies and their features and 26 million viewer ratings. Before performing the potential item-based collaborative filtering, the dataset needs to be organized. For example, the genre attribute within the movie list is given as an array that covers multiple genres. Such columns need editing and cleaning. Additionally, movies are referred to by their IDs in other files, necessitating the creation of an ID-movie list.

After these preparations, the main algorithms are applied to the data. There are four main algorithms trained in this project: Item-Based Collaborative Filtering (CF), User-Based Collaborative Filtering, Singular Value Decomposition (SVD), and Non-Negative Matrix Factorization (NMF). The models are able to recommend a specific user the top-10 movies they might like the most with their predicted rating that would be given if this user had watched the movie. However, as it is indicated in Fig. 1, the number of votes for each movie is not the same nor close to each other. So, the dataset is reduced by a threshold (8) of the cumulative distribution of ratings that is displayed on Fig. 2.

The evaluation of the models is based on the RMSE values displayed in Fig. 3. The SVD model emerges as the top

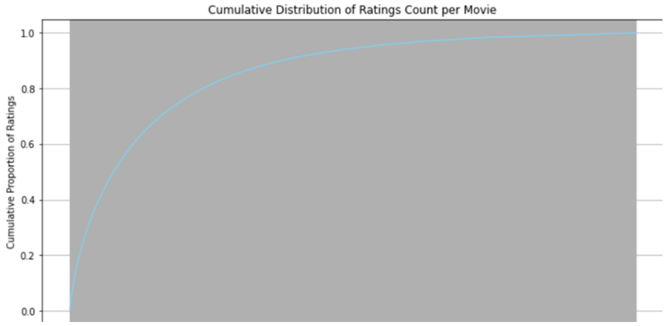

Fig. 2. Cumulative Distribution of Ratings

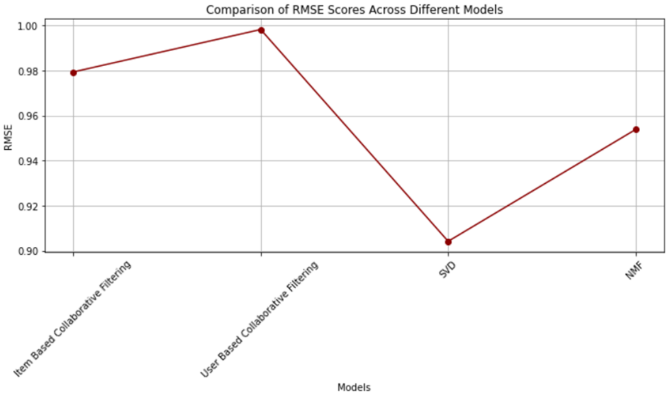

Fig. 3. Comparison of RMSE Scores

performer, exhibiting the lowest RMSE score, followed closely by the NMF model. Each type of Collaborative Filtering shows a relatively higher RMSE.

There is a sample output for all models below, showcasing the top-10 movie recommendations derived from a user's previous voting history. These recommendations are for the user with ID 1, with each movie also displaying a predicted vote next to it.

Movie ID: 54256, Movie Name: Rang Birangi, Estimated Rating: 4.18
Explanation: This movie is recommended because you liked similar movies such as: Rocky III, Greed, American Pie, My Tutor, Jay and Silent Bob Strike Back, Confidentially Yours.

For example, this is the first movie recommended to the user with ID=1 which is the highest probability. Below there is a recommendation that comes as the 10th recommended movie.

Movie ID: 8208, Movie Name: The Man Who Knew Too Much, Estimated Rating: 3.88
Explanation: This movie is recommended because you liked similar movies such as: Rocky III, Greed, My Tutor, Confidentially Yours.

As it can be observed when two of the outputs are compared, the movie with a higher predicted rating has more similar movies user watched before.

V. Conclusion

This project set out to develop a robust movie recommendation system by employing collaborative filtering (CF) techniques enhanced with advanced matrix factorization methods like Singular Value Decomposition (SVD) and Non-negative Matrix Factorization (NMF). The aim was to provide personalized movie recommendations that are accurately tailored to individual user preferences, thereby improving user engagement and satisfaction.

The methodology began with data preprocessing, including cleaning and visualization to understand and refine the data. This process ensured that the recommendation system was built on a solid foundation of reliable and relevant data. The transformation of this data into user-item and item-user matrices was crucial for effectively applying CF techniques.

Both user-based and item-based collaborative filtering methods were explored. The comparative analysis revealed distinct advantages and limitations of each approach, highlighting the importance of context when choosing the most appropriate recommendation strategy. The incorporation of SVD and NMF provided a significant enhancement in handling sparse data and extracting latent factors that influence user preferences and movie characteristics.

The evaluation of these models through performance metrics such as RMSE and MAE demonstrated varying levels of success. These metrics were useful in fine-tuning the models and ultimately selecting the most effective method for this dataset. The project emphasized the dynamic nature of recommendation systems, where different techniques can yield different results based on the specific characteristics of the data and the intended application.

In conclusion, the project successfully demonstrated the application of sophisticated machine learning techniques to solve a real-world problem in the entertainment industry. The findings contribute valuable insights into the field of recommendation systems, particularly in the context of collaborative filtering and matrix factorization. Future work could explore hybrid models that combine both user and item-based approaches, as well as incorporating additional data sources such as user demographics or contextual information to further enhance the accuracy and relevance of the recommendations.

This study not only advances the academic and practical understanding of building effective recommendation systems but also provides a framework that can be adapted to other domains requiring personalized content delivery.

References

[1] Goldberg, Ken, et al. ”Eigentaste: A constant time collaborative filtering algorithm.” information retrieval 4 (2001): 133-151.

[2] Breese, John S., David Heckerman, and Carl Kadie. ”Empirical analysis of predictive algorithms for collaborative filtering.” arXiv preprint arXiv:1301.7363 (2013).

[3] Huang, Zan, Daniel Zeng, and Hsinchun Chen. ”A comparison of collaborative-filtering recommendation algorithms for e-commerce.” IEEE Intelligent Systems 22.5 (2007): 68-78.

[4] Koren, Yehuda. ”Factorization meets the neighborhood: a multifaceted collaborative filtering model.” Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. 2008.

[5] Luo, Xin, et al. ”An efficient non-negative matrix-factorization-based approach to collaborative filtering for recommender systems.” IEEE Transactions on Industrial Informatics 10.2 (2014): 1273-1284.

[6] Herlocker, Jonathan L., et al. ”Evaluating collaborative filtering recommender systems.” ACM Transactions on Information Systems (TOIS) 22.1 (2004): 5-53.

[7] https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset?select=movies_metadata.csv