Addressing the Cold Start Challenge through the Implementation of Content-based Filtering and Hybrid Filtering

Priyanka Vivek, Amrita Vishwa Vidyapeetham, Bengaluru, India.
Nalini Sampath, Amrita Vishwa Vidyapeetham, Bengaluru, India.
L D Mukil, Amrita Vishwa Vidyapeetham, Bengaluru, India.
Mahi Kolli, Amrita Vishwa Vidyapeetham, Bengaluru, India.
Mahi Kolli, Amrita Vishwa Vidyapee
Mahi Kolli, Amrita Vishwa Vidyapee

***Abstract*—This study presents a hybrid recommender system leveraging collaborative filtering and content-based filtering methodologies. Collaborative filtering, implemented through the Surprise library, employs a user-based approach to predict movie ratings. Concurrently, content-based filtering utilizes movie genres as features to compute item-item similarity matrices. A collaborative filtering model is trained on user-item interactions, while content-based features are used to enhance movie recommendations. A hybrid scoring mechanism combines the strengths of both techniques, providing users with personalized and diverse recommendations. Evaluation metrics such as RMSE are employed to assess the accuracy of collaborative filtering predictions. The proposed hybrid recommender system aims to address the limitations of individual recommendation techniques, offering an effective solution for personalized and diversified movie recommendations.**

***Index Terms*—Hybrid recommender system, Collaborative filtering, Content-based filtering, Surprise library, User-based approach, Movie rating prediction, Item-item similarity matrices, User-item interactions.**

I. Introduction

Reinventing how customers reach their recruit hub. The experiences range from cheap hot yoga sessions to sushi dined in 5-star restaurants that you would never get to experience otherwise. Trying something new nevertheless can be a bit frightening if you are not sure what it might turn out to be and how costly it could be. PP uses this model because it predicts coupons that would excite customers during certain times based on their purchasing and browsing history data. With data-driven predictions, we build machine learning models that predict consumer preferences and improve Ponpare's recommendation engine. Customers will always look forward to their much sought after experience or piece of favourite clothing not to mention improve both online customer shopping experience and merchants sales.With the exponential growth of digital content, recommender systems have emerged as essential tools to alleviate information overload by offering users personalized and relevant suggestions.

Collaborative filtering, a prominent recommendation technique, harnesses user-item interactions to predict preferences and generate personalized recommendations. Despite its widespread adoption, collaborative filtering faces challenges such as the cold start problem, which hinders its effectiveness in scenarios with sparse user-item interactions or new users. To overcome these challenges, this study proposes a hybrid recommender system that seamlessly integrates collaborative filtering and content-based filtering techniques. In this research, collaborative filtering is implemented using the Surprise library, employing a user-based approach to predict movie ratings. Concurrently, content-based filtering utilizes movie genres as features to compute item-item similarity matrices. The synergistic fusion of these methodologies aims to enhance recommendation accuracy and mitigate the limitations associated with individual approaches. One of the significant contributions of this hybrid recommender system lies in its scoring mechanism, which combines collaborative and content-based scores. By strategically blending these scores, the system provides users with not only personalized recommendations based on collaborative filtering insights but also diversified suggestions grounded in content-based analysis.

This approach addresses the shortcomings of traditional collaborative filtering methods, ensuring robust and adaptable recommendations for users with varying preferences. To evaluate the performance of the collaborative filtering component, standard metrics such as Root Mean Squared Error (RMSE) are employed. The integration of collaborative and content-based filtering holds the promise of offering a comprehensive and effective solution for personalized and diversified movie recommendations. This study contributes to the ongoing evolution of recommender systems, aiming to enhance user experience and satisfaction in navigating the vast landscape of digital content.

II. Related Works

Huan Liu et.al [1] examine the role that such promotions have on buyers' intention to make new purchases and the quantities in which customers choose to buy. The paper is enlightened by the analysis of the core parameters that reveal the impact of these promotional approaches to consumption behavior change among consumers providing relevant infor-

mation to the retailers whose aim is to perfect their marketing plans. Jorge Gonza´lez et.al [2] discusses practical approaches to encouraging customers' loyalty via the use of coupons in supermarkets. Retailers may maximize their marketing vestments by determining the leading product categories and brands. Findings offer practical solutions for supermarket managers striving to retain customers and boost sales using strategic coupons. Xinyao Li et.al [3] paper shows Customer retention and effective couponing have become essential as O2O e-commerce grows at a quick pace making it competitive. Association rules, classification, clustering, and regression models are some of the data mining techniques used for understanding customer behavior as well as prediction in relation to coupon usage. Optimization of coupon distribution strategies would be highly dependent on feature engineering and gradient-boosting algorithms like XGBoost. These can be of very great importance to the e-commerce platforms as they seek better ways of engaging their customers and improving loyalty levels among them. Chaitanya Pathak et.al [4] explores a generic model for dynamic pricing in internet retailing within an inventory-led organization. Adaptive pricing utilizes different types of data such as buyer attributes, buyer purchase history, and internet-web data for the prediction of purchase decisions. Industries such as retail, mobile communication, and automotion explore dynamic pricing, applying different pricing systems like agent-based modeling, data-driven models, and game theory models. The study incorporates machine learning as well as statistics in order to provide a better prediction. T. Vafeiadis et.al [5] This work presents a solution for optimizing time-limited coupons targeted at a very large consumer market, working together with a national bank acting as a coupon provider at the third-party level. A consumer market, working together with a national bank acting as a coupon provider at the third-party level. A predictive model called RUSH! employs anonymized transaction records in de-terminating the category as well as the timeframe for customers' future buying. It involves time-based analysis, buying categories correlation, and scalability for billions of transactions. It aims at raising the effectiveness of personalized, data-driven coupon delivery improving the outcome of partnering banks as well as end users. The proposed method gives a greater anticipated value than baseline models using all transaction transactions to target promotions. Yue-Hua Ren et.al [6] argues in favor of developing a predictive model of consumer coupon usage behavior and considers customer segmentation in precision marketing. An enhanced RFM model and the k-means clustering method are integrated into this approach for categorizing customers' coupon sensitivity levels. Thereafter, one applies XGBoost prediction on a complete data set and all customers' categories for a coupon write-off situations prediction. To be more precise, the approach is checked with real results, using them as a benchmark against the results obtained from GBDT and LightGBM algorithms. The results show that the suggested model performs better than other algorithms and serves as a benchmark for accurate coupon delivery in practical applications, emphasizing the importance of customer segmentation in coupon use prediction. Suvash Sedhain et.al [7] solved the cold-start recommendation issue in online retail, focusing on users and not using earlier purchases or meaningful interactions. Using constrained side information like demographics or social neighborhood-based methods in a matrix algebra framework. In actual experiments with 30,000 users and diverse information, leveraging Facebook page likes for cold start recommendations shows great improvements. Ziwei Zhu et.al explores equity in recommending new items, a scenario frequently neglected in prior research that generally focuses on biased person-feedback history. It identifies and formalizes unfairness in cold-start recommender structures, introducing two equity standards: identical opportunity and Rawlsian Max-Min fairness. The study proposes a singular learnable submit-processing framework, showcasing two models that effectively enhance equity without compromising advice utility across more than one dataset.Subramaniyaswamy Vairavasundaram et.al [9] address de-manding situations in collaborative filtering recommender systems, that specialize in information sparsity and the cold begin problem for new entities. It introduces models: RS-LOD, leveraging Linked Open Data (LOD) for dealing with cold start troubles by obtaining statistics approximately new entities, and MF-LOD, a Matrix Factorization version with LOD to address records sparsity. Yashraj Bharambe et.al [10] paper uses Logistic Regression, Support Vector Machine (SVM), Random Forest, and XGboost algorithm to predict customer churn in the telecommunication industry, obtaining accurate and insightful results for strategic decision-making. Jesu´s Bobadilla et.al [10] address the vital new users could begin trouble in recommender systems, emphasizing the inability to lack of customers due to imprecise recommendations before they provide sufficient remarks. It introduces a unique similarity degree optimized via neural, surpassing present metrics' overall performance. Tested on Netflix and Movielens databases, the proposed degree demonstrates extensive improvements in accuracy, and precision, and takes into account, particularly cold start begin eventualities for new customers. The paper includes mathematical formalization for first-class measures.

The papers mentioned above focus on the ways to grow the customer loyalty and sales in retail and e-commerce by using the promotions, coupons and dynamic pricing and they utilize the data mining, machine learning and optimization techniques. They tackle issues like user segmentation, cold start recommendations and data bias in recommendation systems.

Some research gaps include the complexity and cost of implementing these strategies, potential issues with customer privacy and data security, and the challenge of accurately predicting customer behavior despite using advanced techniques.

III. Proposed System Methodology

A. Data Description

GroupLens Research has collected data sets from the MovieLens website. The data sets were collected over various periods, depending on the size of the set. The data consists of

105339 ratings for 10329 movies. The average rating is given as 3.5 and the minimum and maximum ratings are 0.5 and 5. 668 users gave their ratings for 149532 movies

B. Preprocessing

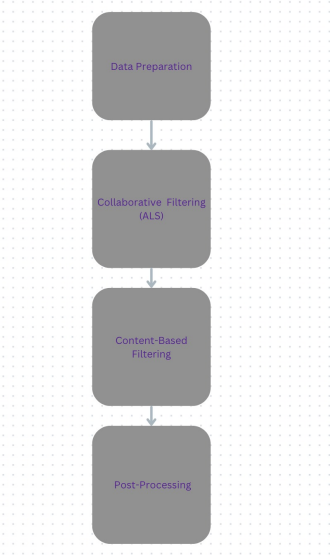

Fig. 1. Proposed System Architecture

C. Data Analysis

First Import the Dependencies. The attribute represents the features and the entity represents the items. Two plotting packages matplotlib and seaborn are used for visualisation and numpy and pandas. Second, Load the data. Use pandas and find the mean. Find the lowest rating and also find the highest average rating. It is noticed that some highly rated movies only have a handful of ratings thus placing them higher on the rating scale. To overcome this problem we apply the Bayesian average to get a fair rating. The Bayesian Average: The formula for Ri is given by: The formula for R_i is given by:

$$R_i = \frac{(C \times m + \sum \text{reviews})}{(C + N)}$$

D. Collaborative Filtering

Now, let's transform the data using collaborative filtering to generate user recommendations. Collaborative filtering is an unsupervised learning technique that allows us to infer a user's interests by leveraging the preferences of a group of users. The first step involves transforming the dataset into a utility matrix, where rows represent users and columns represent items. Notably, collaborative filtering does not require personal information about individual users. The sparsity of the X matrix is calculated as the ratio of non-zero elements by the total number of elements.

E. Content-based filtering

Content-based filtering is used because the new items are not included in collaborative filtering. This is the way to handle cold-start problems by using Content-based filtering. NumPy: for scientific computing pandas: for data manipulation scikit-learn: for machine learning matplotlib, seaborn: for data visualization. Data Cleaning and Exploration is done here. Genres are expressed as a string with a —, convert this into a list. Create a new title for the year. By using Python's Counter there are 20 genre labels. Some movies don't have any genres, where the label is (no genres listed). Use the Counter to know the most popular genres. The top 5 genres are Drama, Comedy, Thriller, Action, and Romance. The most popular decade of movie release is also found. Transform the data where "1" indicates that the movie is under a given genre and "0" is not. Use a similarity metric called cosine similarity. The expression for cosine of the angle between vectors A and B is given by: The expression for cosine of the angle θ between vectors $\mathbf{A}$ and $\mathbf{B}$ is given by:

$$\cos\theta = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$$

Then we create a movie finder.

```
Because you watched Jumanji (1995):

53                    Indian in the Cupboard, The (1995)
109                    NeverEnding Story III, The (1994)
1618    NeverEnding Story II: The Next Chapter, The (1...
8719                  The Cave of the Golden Rose (1991)
9565                          Gulliver's Travels (1996)
1357                              Borrowers, The (1997)
1565                                   Tall Tale (1995)
2539                We're Back! A Dinosaur's Story (1993)
5624    Kirikou and the Sorceress (Kirikou et la sorci...
5975    Asterix & Obelix vs. Caesar (Astérix et Obélix...
Name: title, dtype: object
```

Fig. 2. Recommendation for Jumanji(1995)

F. Implicit Feedback

Use the implicit package. The cells of the utility matrix are given by the user's degree of preference towards an item. Two types: explicit feedback: direct feedback like movie ratings implicit feedback: and indirect behavior search behavior. Here say that movie rating is nothing but the no. of times a person watches a movie. Transform the data. Create Movie Title Mappers where you can map the ID to the movie using a function that converts a movie title to a movie index. title. Now the data has been prepared to create a feedback recommender system. Use linear algebra technique known as matrix factorization. Divide the original matrix into "taste" dimensions. Use Alternating Least Squares which it will solve one factor at a time unlike SVD. keep the user-factor matrix fixed and solve for the item-factor matrix keep the item-factor matrix and solve for the user-item matrix keep doing till the dot product of the item-factor matrix and user-item matrix is approximately equal to the original X matrix. Then find the most relevant movie related to the movie that a user keeps watching. This code implements a hybrid recommendation system that combines collaborative filtering and content-based filtering to recommend movies for a given user. The function takes the user ID, predicted user ratings, content similarity

matrix, movie information data frame, and rating data frame as inputs. It computes a hybrid score for each movie by blending collaborative and content-based filtering. The function then excludes movies already watched by the user, sorts the remaining movies based on their hybrid scores, and returns the top-recommended movies. The code demonstrates the usage of this function for a specific user (user ID 2) and evaluates its performance by calculating the Root Mean Squared Error (RMSE) between the predicted ratings and the actual ratings. Finally, it prints the RMSE values for collaborative filtering and hybrid filtering, along with the recommended movies for the user. Note that there is a potential issue in the code where the variable is referenced but not defined; it appears that it should be replaced within the code. Additionally, the variable is referenced, but it is not defined in the provided code snippet.

IV. RESULT AND ANALYSIS

A. *item-item recommender system*

TABLE I
MEAN OF THE DATA AND MATRIX SPARSITY OF THE DATA

Mean of the Data	157.69
Matrix Sparsity	1.53%

Use a bar plot to visualize the distribution of movie ratings: Bayesian Average: Shawshank Redemption, The Godfather,

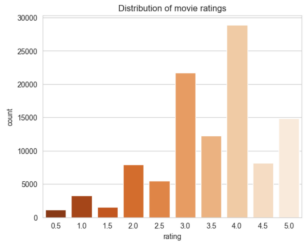

Fig. 3. Distribution of Movie Ratings

and Fight Club are the best-rated movies. Only 1.7 percent of

TABLE II
BAYESIAN AVERAGE BASED ON MOVIEID

Movie ID	Bayesian Average
318	4.454545
858	4.336534
50	4.279618
923	4.253498
527	4.252150

cells in our user-item matrix are populated with rankings. But don't be discouraged via this sparsity! User-object matrices are generally very sparse.A well-known rule of thumb is that your matrix sparsity has to be no decrease than 0.5 percent to generate first-rate outcomes. Similar movies can be recommended using the k- nearest neighbor: The results show the 10 most similar movies to Toy Story. Most movies here are family movies from the 1990s.

B. *Content-based filtering*

The plot shows that Drama and Comedy are the most popular genres and the least popular genres are Westerns,IMAX, and Film-Noir. There are 100 years of release in the dataset. What was the most popular decade? First, remove the null year values and then remove the movies that do not have the year of release. the most common decade is the 2000s as seen in Fig. 7. To get applicable recommendations for Jumanji, locate its index within the cosine similarity matrix. To become aware of which row needs to be searched at, create a movie index mapper that maps a film title to the index that it represents in our matrix. For the feature selection process, only columns containing less than six unique elements were considered categorical or binary data. Other object-type columns were also considered and this emphasizes the crucial role played by categorical variables in forecasting churn. This is important because it helps reduce the data set, with regard to the features that possibly affect customer dissatisfaction. This allows the model to focus on important factors related to customer behavior and subscription details. For visualization of the columns selected after feature selection, a pair plot with kernel density estimation on the diagonal subplots was used. Such software is very useful in studying relationships among some of the characteristics, particularly for predicting churn. Kernel density estimation creates a graphical illustration of the pairwise relationships presenting the distribution of the features and the ability to differentiate them. It was hoped that such visualization would reveal differences between churn and non-churn. The 'Churn' column, which allowed for the clear distinction between two classes, featured in the analysis. The pairplot was very helpful since it involved a general investigation of the linkage that exists between features and their ability to differentiate the categories.

C. *Implicit Feedback*

. Give recommendations for a particular movie that someone repeatedly keeps watching. Based on their alternatives above, we will get an experience that person 95 likes motion and crime movies from the early 1990s over light-hearted American comedies from the early 2000s. Let's see what recommendations our version will generate for consumer 95 as shown

TABLE III
HIGHEST RATED BY USER 95

Rating	Title
5.0	Phantom, The (1996)
4.0	If Lucy Fell (1996)
4.0	Godfather, The (1972)
4.0	Boot, Das (Boot, The) (1981)
4.0	Star Wars - A New Hope (1977)

D. *Hybrid Filtering*

It finds the RMSE value for hybrid filtering and gives the recommendation as follows:

TABLE IV
LOWEST MOVIE RATINGS BY USER 95

Rating	Title
2.0	Dracula: Dead and Loving it (1995)
2.0	Star Wars - Return of the Jedi (1983)
2.0	Lawnmower Man 2: Beyond Cyberspace (1996)
1.0	Theodore Rex (1995)

TABLE V
RMSE COMPARISON

Method	RMSE
Collaborative Filtering	0.4058849270215696
Hybrid Filtering	3.1300275287374774

V. CONCLUSION

A comprehensive recommender system that integrates collaborative filtering, content material-based filtering, and hybrid filtering methodologies. The report starts off with the identification, authors, and abstract, supplying a top-level view of the have a look at's targets and methods. The keywords section highlights the key elements of the machine, including collaborative filtering, content-primarily based filtering, the Surprise library, user-based totally technique, film score prediction, object-object similarity matrices, and consumer-item interactions. The creation segment sets the level for the take a look at, emphasizing the importance of recommender structures in coping with information overload and introducing collaborative filtering demanding situations just like the bloodless start trouble. The proposed hybrid recommender system pursuits to cope with these challenges by way of seamlessly combining collaborative and content-based filtering techniques. The related works segment gives a literature review, bringing up relevant research articles that discover the impact of promotions, couponing, consumer retention, and dynamic pricing in the retail and e-trade industries. The technique segment delves into the information collection technique, analyzing information from the MovieLens internet site. It covers information cleansing, exploration, and modifications, which include the introduction of utility matrices and content-based total features. Collaborative filtering is implemented with the use of the Surprise library, and content-based filtering leverages movie genres. The segment on implicit remarks discusses the usage of the implicit package deal to address indirect consumer conduct statistics. The hybrid filtering section outlines the code implementation for combining collaborative and content-primarily-based filtering to endorse movies. It includes the hybrid recommendation feature, which takes user-related facts and computes hybrid rankings for film guidelines. The file highlights capability troubles with variable names and mentions the calculation of Root Mean Squared Error (RMSE) for performance evaluation.

REFERENCES

[1] H. Liu, L. Lobschat, P. C. Verhoef, and Z. Hong, "The effect of permanent product discounts and order coupons on purchase incidence, purchase quantity, and spending," *Journal of Retailing*, vol. 97, no. 3, pp. 377-393, 2021. https://doi.org/10.1016/j.jretai.2020.11.007

[2] I. O. Soto, J. P. Gonzalez, and M. Capizzani, "Which Categories and Brands to Promote with Targeted Coupons to Reward and to Develop Customers in Supermarkets," *Journal of Retailing*, vol. 92, no. 2, pp. 236-251, 2016. https://doi.org/10.1016/j.jretai.2015.12.002

[3] Research on e-commerce coupon User Behavior Prediction Technology based on Decision Tree algorithm, *Airiti Library*, 2019. https://doi.org/10.6919/ICJE.201908_5(9).0008

[4] R. Gupta and C. Pathak, "A Machine Learning Framework for Predicting Purchase by Online Customers based on Dynamic Pricing," *Procedia Computer Science*, vol. 36, pp. 599-605, 2014. https://doi.org/10.1016/j.procs.2014.09.060

[5] Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, *ACM Conferences*. https://dl.acm.org/doi/abs/10.1145/3097983.3098104

[6] Y. Ren, P. Fu, and Y. Wang, "Prediction of Coupon Usage Behavior Based on Customer Segmentation and XGBoost algorithm," in *2021 2nd International Conference on Big Data Economy and Information Management (BDEIM)*. https://doi.org/10.1109/bdeim55082.2021.00016

[7] N. I. A. Razak and M. H. Wahid, "Telecommunication Customers Churn Prediction using Machine Learning," in *2021 IEEE 15th Malaysia International Conference on Communication (MICC)*, Malaysia, pp. 81-85, 2021. https://doi.org/10.1109/MICC53484.2021.9642137

[8] Fairness among New Items in Cold Start Recommender Systems, Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, *ACM Conferences*. https://dl.acm.org/doi/abs/10.1145/3404835.3462948

[9] S. Natarajan, V. Subramaniyaswamy, S. Natarajan, and A. H. Gandomi, "Resolving data sparsity and cold start problem in collaborative filtering recommender system using Linked Open Data," *Expert Systems With Applications*, July 1, 2020. https://doi.org/10.1016/j.eswa.2020.113248

[10] J. Bobadilla, F. Ortega, A. Hernando, and J. Bernal, "A collaborative filtering approach to mitigate the new user cold start problem," *Knowledge-Based Systems*, February 1, 2012. https://doi.org/10.1016/j.knosys.2011.07.021