Book Recommendation System: A Hybrid Approach

Anupam Garg
anupam20555@iiitd.ac.in

Saharsh Dev
saharsh20572@iiitd.ac.in

Saksham Singhi
saksham20463@iiitd.ac.in

Vishal Kumar
vishal20154@iiitd.ac.in

Vanshika Goel
vanshika20413@iiitd.ac.in

Vanisha Singh
vanisha20347@iiitd.ac.in

ABSTRACT

This project aims to develop a novel book recommender system that considers users' reading history, preferences, and behavior. The system utilizes a combination of collaborative filtering and content-based filtering techniques to provide personalized and accurate recommendations. The proposed system outperforms existing systems in terms of recommendation accuracy and user satisfaction. Overall, the project presents an effective solution to the challenges faced by existing book recommender systems.

1. INTRODUCTION

1.1 Problem Statement

Book Recommender Systems are becoming increasingly popular with the rise of e-commerce platforms and digital libraries. These systems are designed to recommend books to users based on their reading history, preferences, and behavior. However, despite their growing popularity, many book recommender systems are not effective enough to meet users' needs. The problem with existing book recommender systems is that they often fail to provide personalized and accurate recommendations. Many of these systems rely on simple recommendation algorithms that do not take into account users' reading history, preferences, or behavior. As a result, users are often presented with irrelevant and uninteresting book recommendations, leading to dissatisfaction and reduced engagement with the platform. Therefore, there is a need for more effective book recommender systems that can provide personalized and accurate recommendations to users. Our project aims to develop and evaluate a novel book recommender system that takes into account users' reading history, preferences, and popularity trends. The goal is to improve the overall user experience of book recommendation systems and increase user engagement with digital libraries and e-commerce platforms.

1.2 Motivation

Reading is a popular and enjoyable leisure activity for many people. However, with the vast number of books available today, finding books that match an individual's taste and interests can be a daunting task. Book recommender systems have been developed to address this problem by providing personalized book recommendations to users. However, existing systems suffer from various limitations, such as low accuracy, inadequate consideration of users' behavior and preferences, and lack of user-friendliness. This project is motivated by the need to develop a novel book recommender system that overcomes these limitations and provides a more accurate, personalized, and user-friendly solution to the challenge of book recommendation.

1.3 Literature Review

[1] In 2010 Choi et.al. proposed RecSys based on HYRED, a hybrid algorithm using both content and collaborative filtering where they used an altered Pearson Coefficient based Collaborative filtering and distance-to-boundary (DTB) Content filtering.

[2] Kurmashov et al. in their paper[1] tried to propose a book recommendation service by taking inputs of the preference like the most favorite genre and ratings on different books to further narrow down the space of search for recommendations while a user logs into the system.

[3] In 2016 Mathew et.al. proposed a system that saves details of books purchased by the user. From these Book contents and ratings, a hybrid algorithm using collaborative filtering, content-based filtering and association rule generates book recommendations.

[4] In 2021, Sarma, Mittra, Hossain proposed a clustering-based book recommendation system that

uses different approaches, including collaborative, hybrid, content-based, knowledge-based, and utility-based filtering.

[5] In 2022, Mishra, Asthana tried to outline the limitations of the content and the collaborative based filtering and proposed an effective solution in this regard. While content filtering works on the traditional methods, the hybrid filtering tries to combine them and employs a collaborative networking approach, which compares the results with a wider audience and produces accurate results.

1.4 Novelty

The novelty of our book recommender system lies in the evaluation of our content-based approach, which is done using the estimated ratings generated from collaborative filtering. Using these calculated ratings (a user might give to each book), we defined our true positives of CB recommendations as those books which have their rating greater than the average rating that the user has given. And using these TP, FP, TN and FN values, we calculated the accuracy and specificity. Also, to cater to new users who have not yet established their preferences, we use a popularity-based system that recommends books based on their popularity and overall appeal, which provides a starting point for users to explore different types of books before our system begins to personalize recommendations based on their individual preferences.

2. DATABASE

Two datasets were used in this Information Retrieval project, one for content-based filtering and the other for collaborative filtering. The first dataset used for the content-based filtering system is the Goodreads 2M dataset. This dataset contains information on books, such as their name, authors, rating, publish year, publisher, counts of reviews, language, page number, and description. The dataset has books in different languages, including English, Spanish, French, German, Portuguese, Italian, and Japanese. The second dataset used for the collaborative filtering system is called "final users" and has been generated by the researchers for this project. This dataset contains information on users' ratings of books. The dataset has columns for the user ID, book name, book ID, and rating. The dataset has 11,057 unique users and 3,534 unique book IDs, which were specifically selected based on the research needs of this project. Overall, the datasets used in this project provide a rich source of information for content-based and collaborative filtering systems. The Goodreads 2M dataset provides a variety of book attributes that can be used for content-based filtering, while the final users dataset provides user ratings that can be used for collaborative filtering.

3. METHODOLOGY

The hybrid book recommender system proposed in this project offers several features to address common challenges associated with book recommendation systems.

- One of the significant challenges in book recommendation systems is the problem of sparsity, where users have rated only a small fraction of the available books. To address this challenge, our system integrates collaborative filtering and content-based filtering techniques to leverage both the user's past behavior and the book's content features to provide accurate recommendations.
- Another challenge is the cold start problem, where the system struggles to provide recommendations for new users who have not yet established their preferences. Our system uses a popularity-based approach to recommend books that are popular and have broad appeal to help new users discover new types of books.
- The system also provides scalable recommendations based on the user's query. By identifying the most relevant similar entities from a large dataset, the system can recommend books that match the user's interests and preferences.
- Lastly, the system provides qualitative recommendations to users automatically based on their past preferences without asking for more information. By leveraging collaborative filtering and content-based filtering techniques, the system can analyze the user's behavior and recommend books that are relevant and of interest to them.

4. WORKFLOW

The book recommendation system developed in this project employs three methods to generate book recommendations - Collaborative Filtering, Content-Based Filtering, and Hybrid Filtering. Each method has its own unique approach to generating recommendations, and when used in combination, they provide a more comprehensive and accurate

recommendation system. In this section, we describe the workflow of each method, its strengths, and how we have integrated them into our hybrid model.

- Collaborative filtering is a key method used in our recommendation system. We compute the normalized mean absolute error (NMAE) for evaluating the accuracy of the model. The NMAE is computed using a function that creates a user-item matrix, which is then split into training and testing sets. We use the Multiplicative Updates algorithm of latent factor models, to factorize the training set and generate predictions. To enhance the scalability of the system, we recommend using the most relevant similar entities from a large dataset based on a user's query. The system also provides qualitative recommendations to users automatically based on their past preferences without asking for more information.

- Content-based filtering is a recommendation technique that involves analyzing the features or characteristics of items and recommending similar items to users based on their preferences. In our report, we utilized content-based filtering by preprocessing book descriptions and vectorizing them using TF-IDF to generate a matrix of cosine similarities between books. By using this technique, we were able to recommend books to users based on the books they have already read and their similarity to other books in our dataset. The code snippet provided shows how we preprocessed the book descriptions, generated the cosine similarity matrix, and mapped user preferences to the most similar books in our dataset.

- Hybrid-based Recommendation Model works by combining the results of both Collaborative-based and Content-based book recommendations systems. It is a powerful way of improving the accuracy and robustness of a recommendation system as it overcomes the limitations of both the approaches. The Content-based approach can suffer from cold start, where new items with limited or no metadata are difficult to recommend, while Collaborative-based filtering can suffer from the sparsity problem, where not all users have enough data to form accurate recommendations. Hybrid-based models minimize their limitations by combining these two approaches, including weighted averaging, cascading, or switching. We used a weighted averaging scheme, by giving the ⅓ of the recommended books from the results from collaborative filtering and the rest from content-based filtering.

- In the popularity based approach, we used the Bayesian average ranking system. In this we used the number of reviews of each book and the ranking of the book into account.

- The Liking Model uses Content-based recommendation to show similar books to the user using the users' highest rated or the most liked books.

5. CODE

In this section, we will explain how our code works and how it is implemented.

Each of the model is explained below:-

1. For the content based filtering , we first did the preprocessing of the dataset and then made the TF-IDF vector of the book descriptions, and then using those vectors, we calculated the cosine similarity of each book description to another. Then taking the previously read books of the user which they liked, we recommended the books in order of highest cosine similarity values with those of previously read books.

2. For the collaborative based filtering, we first created the user vs books matrix of book ratings using the dataset and assigned zero to the missing values. Then, we divided it into a ratio of 85:15 for the train & test data. Then we integrated the Multiplicative Updates algorithm for matrix factorization.

   We used latent factors as 70 and calculated the Normalised Mean Absolute Error (NMAE).

3. To create a hybrid model, we integrated content and collaborative-based filtering. In our hybrid filtering, we first give the highest priority to books (x) that are common in both content and collaborative-based

filtering. Then, as we are proposing a total of 10 books, we divided 10-x into 2:1 ratios of content and collaborative recommendations respectively to finally show the hybrid recommendations.

4. Popularity-based Recommendation Model works by Bayesian average ranking value which takes into account two features: average rating of the book and, the number of reviews given to the book.

   Formula used: **wr = (r*v + k*m) / (v+m)**

   Here **r** stands for average rating of the book, **v** stands for number of reviews, **k** stands for average rating across the dataset, and **m** stands for the weight given to the global average rating.

6. EVALUATION

6.1 Comparison with baselines and the system's performance on existing data

For the baseline, we implemented the Collaborative filtering approach, particularly a Model-based approach where we used a kNN model and cosine similarity matrix, to generate the results. We returned the top 6 nearest books for the recommendations. We used the Normalised Mean Absolute Error as a metric for testing our model performance, where we found that the error value was quite large for the kNN model (NMAE = 0.763) as compared to our own implemented collaborative technique, where we have used deep latent factor models like Matrix Factorization and Multiplicative Updates to obtain the predicted values for the item ratings.

Using this model, our loss value has been significantly reduced (NMAE = 0.204).
The system only displayed the books which were liked by the other users who were similar. The system was not able to provide more personalized recommendations to the user as there were no criteria to compare the books which the user has already reviewed

6.2 SOTA on different evaluation metrics and how the system performs on new data / handles different cases

The RMSE value achieved from our collaborative filtering model is 0.90350 which beats 5 state of the art models as shown in the picture below.

	Model_Name	Grid Search (Y/N)	Paramters	RMSE
4	SVD	Y	{'rmse': {'n_factors': 50, 'n_epochs': 50, 'lr_all': 0.005, 'reg_all': 0.05, 'biased': True}}	0.888873
0	KNNWithMeans	N	k=50, name: pearson_baseline, user_based: True,min_support = 1)	0.895775
2	KNNWithMeans	Y	{'k': 50, 'sim_options': {'name': 'msd', 'min_support': 3, 'user_based': True}}	0.909761
6	NMF	Y	{'n_factors': 10, 'n_epochs': 40, 'biased': True}	0.915269
3	SVD	N	n_factors=20, n_epochs = 30, biased=False	0.920035
5	NMF	N	n_factors=20, n_epochs = 30, biased = True	1.299336
1	Normal Predictor	N	-	1.338852

The results are from the paper published on Sep 1 2019. The NMAE value of the collaborative filtering model came out to be 0.207 and the accuracy of our content based model using the values from the collaborative filtering model came out to be 96.37%.

For new users who have not yet established their preferences, we are using a popularity-based system that recommends books based on their popularity, which provides a starting point for users to explore different types of books before our system begins to personalize recommendations based on their individual preferences.

For new books, we are providing a section for books that are newly released so that the users can explore them. Also, using this we are solving the problem of recommendation of books that have not been read and rated by people.

7. FUTURE WORK

- We can suggest books on the basis of the language
- We can modify our dataset in order to integrate the genre of the book and then recommend based on genre
- We can also use the data about average number of pages user reads and suggest them books that are close to amount of pages he usually reads

8. REFERENCES

[1] Choi, S.H., Jeong, Y.S. and Jeong, M.K. (2010). A Hybrid Recommendation Method with Reduced Data for Large-Scale Application. IEEE Transactions on Systems, Man and

Cybernetics-Part C: Applications and Reviews, Vol. 40, No.5, September 2010.

[2] Nursultan Kurmoshov, Konstantin Latuta, Abay Nussipbekov. Online Book Recommender System Conference: 2015 Twelve International Conference on Electronics Computer and Computation (ICECCO)

[3] Mathew, P., Kuriakose, B. And Hegde, V. (2016). Book Recommendation System through content based and collaborative filtering method. Proceedings of International Conference on Data Mining and Advanced Computing (SAPIENCE)

[4] Dhiman Sarma,Tanni Mittra, Mohammad Shahadat Hossain. Personalized Book Recommendation System using Machine Learning Algorithm. (IJACSA) International Journal of Advanced Computer Science and Applications,Vol.12, No.1,2021 Pages 212 - 219

[5] Himanshu Mishra, Ashish Asthana. Book Recommendation System. International Research Journal of Engineering and Technology (IRJET)

[6]https://shamafarabi.github.io/2019/09/01/recommendation