Project Design Report

Book Recommendation System

Group 8
Dwisha Baviskar
Jacob Harper
Cagri Isilak
Julia Matys
Manik Sehgal
Cameron Zhang

Wilfrid Laurier University
Department of Computer Science
CP421: Data Mining
Dr. Yang Liu
December 8, 2024

Table of Contents

Book Recommendation System

1. Introduction

Over the past few decades, recommender systems have become an integral part of the digital ecosystem, shaping how users interact with platforms such as YouTube, Amazon, and Netflix. Recommender systems enhance user experience by tailoring content recommendations to individual preferences, bridging the gap between users and the sea of available options. From suggesting movies and articles to recommending products and services, recommender systems have transformed industries like e-commerce, online advertising, and entertainment. At their core, these systems use algorithms to analyze user behaviour, preferences, and patterns to provide personalized suggestions.

In e-commerce, platforms like Amazon rely on these systems to suggest products that align with a user's browsing history, purchase patterns, and preferences. Similarly, Netflix utilizes algorithms to recommend shows and movies, boosting user engagement and retention. Overall, leveraging data-driven insights streamlines a user's decision-making. In this project, our focus is to develop a book recommendation system using concepts taught in CP421-Data Mining to help user's streamline their decision making process for choosing books.

Leveraging the Book Recommendation dataset available on Kaggle, our project aims to explore various approaches for enhancing the efficiency and accuracy of recommendation models. This dataset, curated from the Book-Crossing community, includes 278,858 anonymized users, 1,149,780 ratings, and 271,379 books. The dataset is segmented into three files: *Users* (demographic data), *Books* (content-based information), and *Ratings* (explicit (1-10 scale ratings) and implicit feedback on books).

Millions of books are readily available on the market, making finding the perfect book to match an individual's taste and preferences increasingly challenging. With an estimated 700,000 to 1,000,000 new titles released globally each year, the sheer volume of options can overwhelm readers, making the need for effective recommendation systems more crucial than ever (Errera 2023). These systems not only streamline the discovery process but also enhance the overall reading experience by presenting personalized suggestions tailored to a reader's preferences.

The 2020 COVID-19 pandemic brought significant lifestyle changes to many people's lives, one of which was a global surge in reading and book sales (Errera 2023). With lockdowns and social distancing, many people turned to books for entertainment, learning, and solace, driving a revival in the popularity of reading. Additionally, social media platforms brought this revival in reading, with social media influencers on TikTok coining their book specific content "BookTok" (Thapa 2022). Apps like Goodreads, StoryGraph, Basmo, and TBR - Bookshelf allow readers to set reading goals, track the books they have read and books they want to read. This renewed interest in reading emphasizes the importance of tools that can help readers navigate the endless options available to them, keeping the momentum going for reading.

Our book recommendation system is a valuable asset for readers and publishers. For readers, our system simplifies their decision making, enabling them to discover titles that are of interest to them. For publishers, our system provides a mechanism to reach target audiences with ease, driving sales and engagement. Overall, our project seeks to harness the power of recommendation systems to create a system that enriches the reading experience for users.

2. Related Work & Research

Recommender systems have evolved significantly over time, with foundational research laying the foundation for methods such as collaborative filtering, content-based filtering, and hybrid models. These methods aim to overcome challenges like sparsity, cold-start problems, and scalability, which are particularly relevant for large datasets like our Book Recommendation Dataset.

Bobadilla et al. (2013) provide an extensive survey of recommender system methodologies, emphasizing collaborative filtering as one of the most widely used techniques. The survey outlines how user-to-user

and item-to-item collaborative filtering work by identifying similarities in user preferences or item features. However, Bobadilla et al. (2013) also highlight limitations and challenges such as sparsity and cold-start problems, which occur when user-item interaction data is limited. The cold-start problem occurs when it is difficult to provide personalized recommendations due to new items added with little or no interaction history or when a new user joins a platform with no data on their preferences and interests. Hybrid systems, which combine collaborative filtering with other techniques such as content-based filtering, are presented as a promising solution that mitigates these challenges.

The HRS-IU-DL model proposed by Sami et al. (2024) integrates advanced techniques, including user-based collaborative filtering, item-based collaborative filtering, neural collaborative filtering, and content based filtering using term frequency-inverse document frequency (TF-IDF). Similar to Bobadilla et al. (2013), this model addresses challenges of recommender systems like cold-start issues, sparsity, and scalability, by combining sequential user behaviour analysis via RNNs and item metadata. The combination of these techniques enables more accurate and personalized recommendations.

Jain et al. (2017) proposes a hybrid recommender system that incorporates contextual information, such as user demographics and preferences into the recommendation process. Their content aware hybrid systems study highlighted pre-filtering datasets based on contextual features, followed by collaborative filtering to generate recommendations. Their approach reduces the density of the data as well as enhancing recommendation relevance.

Based on all the work by Bobadilla et al. (2013), Sami et al. (2024), Jain et al. (2017), our implementation adopts their mentioned principles by integrating collaborative filtering into our pipeline through user-item interaction matrices. We aim to address sparsity using matrix factorization techniques, such as incremental principal component analysis and truncated SVD, to reduce the dimensionality of sparse matrices while preserving the meaningful patterns found in the data. Additionally, we want to normalize explicit ratings to ensure consistency and enable the system to handle missing or sparse data efficiently. While our implementation will not incorporate deep learning models like RNNs, it aligns with Sami et al. 's (2024) hybrid philosophy by combining collaborative filtering and content based filtering. For instance, we will use TF-IDF to vectorize item metadata, such as book authors and publishers, creating a content-based feature matrix. Similarities between items are computed using cosine similarity, while SVD reduces feature dimensions to improve efficiency. The work done by Sami et al. (2024) informs our strategy by merging several approaches to ensure diverse and accurate recommendations, even for new users or items with limited interactions. Additionally, the approach by Jain et al. (2017) in context-aware hybrid systems inspired us to encode demographic features such as user age and location into our user profiles, filling missing values with the median or placeholders for categorical data. Label encoding is also used to transform categorical attributes into numerical representations, making them suitable for model inputs.

Drawing from these studies, our system combines elements from traditional and advanced recommendation system methodologies such as hybrid filtering, context integration, and dimensionality reduction. These integrations ensure our system is scalable, adaptable, and capable of handling large and complex datasets such as the Book Recommendation Dataset.

3. Solution/Methodologies

Building our recommendation system involved multiple stages of preprocessing, design, data cleaning, and implementation. Our approach draws on the well-established principles discussed in class as well as research in recommender systems, incorporating hybrid strategies to address challenges such as scalability and cold starts. This section details our exploration of chosen methods in our systems, aligned with insights from the literature discussed previously.

3.1 Data Preprocessing

Data preprocessing is the foundation of any recommender system. It ensures the data is of good quality, facilitates efficient computation, and enhances model accuracy. Drawing from Jain et al. (2017), which highlighted the importance of reducing dataset density for collaborative filtering, we implemented the following preprocessing techniques.

3.1.1 Data Cleaning

The data cleaning process in our system aims to address inconsistencies, missing values, and outliers, ensuring the dataset is accurate and reliable for our analysis and recommendation system. Data cleaning ensures our recommender system does not have errors that propagate and amplify errors in recommendations. In our data cleaning process, we addressed three main issues: unrealistic user demographics, incomplete or ambiguous locations, and invalid publication years.

First, the user age column contained unrealistic entries, such as ages below 5 or above 100, which were replaced with NaN. Missing age values were imputed with the median age, ensuring the distribution remained representative of the population without introducing bias from extreme values.

Similarly, the location column was split into city, state, and country, with missing countries labelled as "Unknown". Splitting the location into city, state, and country fields enables better geographic profiling, following the emphasis on contextual features highlighted by Sami et al. (2024).

Finally, book metadata had some invalid publication years. These were converted to NaN and replaced with the median. Inconsistencies in item attributes can lead to poor results in collaborative filtering and content based filtering models. By ensuring clean and complete metadata, we enhance the system's ability to generate accurate recommendations.

3.1.2 Normalization

Normalization is crucial in bringing data scales into a common range, ensuring that features contribute proportionally during similarity calculations and model training. For the year of publication, min-max scaling transformed values to a [0, 1] range. This method, prevents features with larger numerical ranges from disproportionately influencing similarity metrics, such as cosine similarity, that are critical for content-based filtering. Normalizing publication years is especially important for a book recommendation system, as it avoids biasing the system toward recent publications simply because they have a higher numerical value.

Explicit user ratings were normalized at the user level by subtracting the mean rating for each user. This approach aligns with collaborative filtering strategies described by Bobadilla et al. (2013), where normalized ratings improve the model's ability to detect relative preferences. Without normalization, users who consistently give higher or lower ratings than average might skew similarity computations, reducing the effectiveness of collaborative filtering models. By transforming ratings into relative scores, our recommendation system highlights patterns of user-item interactions over absolute values, improving the accuracy of our recommendations.

3.1.3 Label Encoding

Label encoding was applied to categorical variables, including user countries, book authors, and publishers, to convert them into numerical formats suitable for machine learning. This choice in our system aligns with Jain et al. (2017), where demographic and contextual attributes are used to enrich user profiles for personalized recommendations. Encoding categorical data such as countries enables collaborative filtering to incorporate user context when determining similarities, addressing one of the limitations of traditional collaborative filtering models as discussed by Bobadilla et al. (2013).

Additionally, encoding book metadata, such as authors and publishers, ensures that these features can be effectively utilized in content-based filtering and hybrid models. In our implementation, the encoded author and publisher attributes are included in the combined feature space for TF-IDF vectorization. Encoding provides a scalable way to manage large categorical datasets, ensuring compatibility with both matrix-based and neural algorithms without introducing interpretability issues. By encoding these features, our recommendation system gains the ability to capture meaningful relationships between users and items while also maintaining computational efficiency.

3.2 Implementation Strategy

Our implementation strategy combines content-based filtering, collaborative filtering, and a hybrid approach to create a robust recommendation system. Content-based filtering utilizes Singular Value Decomposition (SVD) to extract latent features from the user-item interaction matrix, focusing on item attributes and user preferences. Collaborative filtering identifies patterns in user interactions, leveraging a sparse user-item matrix and Incremental Principal Component Analysis (PCA) to enhance scalability and accuracy. Finally, the hybrid approach integrates these two methods using linear regression, merging their strengths to generate comprehensive and personalized book recommendations while mitigating their individual limitations.

3.2.1 Content Based Filtering

In our content-based filtering, Singular Value Decomposition (SVD) was used to extract latent features from the user-item interaction matrix. The goal was to identify the most important elements that represent user preferences and book properties. In order to maximize computing efficiency, we successfully decreased the interaction matrix's dimensionality by truncating it using SVD and concentrating on the most important elements. Our strategy reflects the principles discussed by Sami et al. (2024), where dimensionality reduction techniques like SVD are crucial in extracting latent features that represent meaningful patterns in a user's preferences and item properties. The algorithm was able to produce recommendations based on user-item similarities since each user and book was represented in this latent feature space. In order to guarantee consistency and improve the effectiveness of content-based suggestions, normalization techniques were also used in the ratings. Normalization of user rating was vital for consistency in similarity calculations as discussed by Bobadilla et al. (2013).

3.2.2 Collaborative Filtering

Collaborative filtering was employed to identify trends in the way users engaged with books and to emphasize the connections between users and items. We compared people and items to identify trends, and created a sparse user-item matrix to depict these interactions. Incremental principal component analysis (PCA) was introduced to manage the matrix's sparsity and enable scalability, as it is effective for large-scale datasets. By transforming the matrix into a reduced latent space, we identified hidden user preferences and item similarities. An addition to the content-based approach, this model solely depends on user behavior.

3.2.3 Hybrid Filtering

The hybrid filtering approach combined the advantages of collaborative and content-based filtering techniques to improve the overall caliber of suggestions. Drawing from the principles of hybrid systems, we used linear regression as an integration mechanism, which combined the outcomes of the two techniques to produce a single recommendation score. In order to overcome the limitations of separate methods, this hybrid model combined the feature-driven insights of content-based filtering with the user interaction data from collaborative filtering. Regression techniques were utilized to fill in the gaps in the

predictions after the results from each strategy were combined. This resulted in a system that balanced accuracy and diversity of suggestions. Our hybrid model enhances recommendation quality and mitigates cold-start and sparsity issues. This tactic made sure that the model was thorough and reliable enough to provide highly customized book recommendations.

4. Evaluation and Results

Our recommendation system was evaluated using key metrics such as precision, F1-score, recall, Root Mean Square Error (RMSE), and Mean Absolute Error (MAE), across different parameter configurations for both our content-based and collaborative filtering approaches. These metrics allowed us to assess the trade-offs between accuracy, diversity, and computational efficiency.

Figure 1: Content Based Filtering Evaluation

Components	Precision	Recall	F1-Score
100	0.076190	1.0	0.141593
400	0.933333	1.0	0.965517
700	1.000000	1.0	1.000000
900	0.885714	1.0	0.939394

For our content based filtering approach, we tested different levels of feature truncation using Singular Value Decomposition (SVD). The evaluation shows that using 700 truncated features provides the best results, with precision, recall, and F1-score all reaching 1.0. This indicates that the model effectively captures the latent relationships between user preferences and book attributes when using this configuration. Lower feature counts, such as 100, yielded significantly lower precision and F1-scores, suggesting inefficient dimensionality to represent the complex relationships in the data. Additionally, while configurations with 900 features still performed well, they introduced a slight reduction in precision, which could have been caused by overfitting or redundancy in the latent feature space. Sami et al. (2024) discusses the importance of dimensionality reduction for improving computational efficiency without compromising the accuracy of our model. The use of TF-IDF for content vectorization and SVD for truncation ensures that our recommendation system is both scalable and capable of generating high-quality recommendations. Our results also corroborate that content-based methods are most efficient when combined with robust feature extraction and dimensionality reduction techniques. Similarly, for collaborative filtering, we tested a range of latent factors to evaluate their impact on RMSE and MAE. The optimal value was determined to be 100 latent factors, as this configuration provided the lowest RMSE (4.8107) and a competitive MAE (2.8770). Lower latent factors, such as 20 and 30, resulted in slightly better RMSE and MAE values, but these configurations are likely to oversimplify the user-item interaction matrix, missing the nuanced and more subtle relationships. Configurations with higher latent factors, such as 200, lead to comparable RMSE and MAE values, however, there is an increased computational complexity without important performance gains.

Figure 2: Collaborative Filtering Evaluation

	user_item_reduced_train.shape	Vt.T.shape	RMSE	MAE
k = 20	(105283, 20)	(340556, 20)	4.8105	2.8765
k = 30	(105283, 30)	(340556, 30)	4.8104	2.8764
k = 50	(105250, 50)	(340556, 50)	4.8106	2.8766
k = 100	(105200, 100)	(340556, 100)	4.8107	2.8770
k = 200	(105200, 200)	(340556, 200)	4.8104	2.8770

Our results indicate that collaborative filtering benefits from fine-tuning latent factors to balance model complexity and accuracy. Additionally, our use of incremental principal component analysis (PCA) to handle sparsity and scale the user-item matrix reflects best practices for large-scale recommendation systems.

Overall, our evaluation demonstrates that our hybrid approach that combines content-based and collaborative filtering is an effective solution for our book recommendation system. By selecting 700 features for SVD in content-based filtering and 100 latent factors in collaborative filtering, we built a system that balances computational efficiency and high quality recommendations.

5. Conclusion

Overall, our book recommendation system effectively combines content based filtering and collaborative filtering techniques into our hybrid model. Our system achieves personalization, scalability, and computational efficiency. Through data preprocessing, data cleaning, normalization, and label encoding, we ensured our model was built on high-quality data. This step was critical in addressing data sparsity and inconsistencies.

Our content based filtering component utilized Singular Value Decomposition (SVD) to extract meaningful latent features, with an optimal truncation of 700 features delivering sound precision, recall, and F1-scores. This approach demonstrated the importance of dimensionality reduction in maintaining accuracy and efficiency for large-scale datasets. Collaborative filtering utilized a user-item matrix and incremental PCA, with 100 latent factors providing the lowest RMSE and MAE values. Our hybrid approach integrated content based and collaborative filtering methodologies through linear regression, leveraging the strengths of each model while mitigating their limitations. Our evaluation metrics validate the robustness of our recommendation system. The hybrid model achieves high accuracy and demonstrates scalability for large datasets, which ultimately, makes our model suitable for deployment. As reading enjoys a global resurgence fueled by the pandemic, the rise of platforms like Goodreads, and social media trends such as “BookTok”, the importance of tools such as our book recommendation system becomes clear. Our system curates meaningful connections between readers and stories, ensuring that the right book reaches the right hands.

6. References

Bobadilla, J. (2013). Hybrid recommender systems: Survey and experiments. *Expert Systems with Applications, 40*(4), 1396-1407. https://doi.org/10.1016/j.eswa.2012.08.018

Errera, R. (2023, October 4). *Eye-popping book and reading statistics [2023]*. Toner Buzz. https://www.tonerbuzz.com/blog/book-and-reading-statistics/#:~:text=Since%20the%20onset%20of%20the,book%20sales%20rising%20by%208.9%25.

Jain, R., Tyagi, J., Singh, S. K., & Alam, T. (2017). Hybrid context aware recommender systems. *AIP Conference Proceedings, 1897*(1), 020028. https://doi.org/10.1063/1.5008707

Sami, A., El Adrousy, W., Sarhan, S., & Elmougy, S. (2024). A deep learning-based hybrid recommendation model for internet users. *Scientific Reports, 14*(29390). https://doi.org/10.1038/s41598-024-79011-z

Thapa, A. (2022, April 30). *Anuja Thapa*. The Muse. https://themuse.ca/how-tiktoks-booktok-revived-reading/