Book recommender system

Clarice, Anne-Marijn, Tereza & Andrea

Text Mining Project

Date of submission: 02/06/24

Abstract

This study evaluates the genre relevance of book recommendations generated by three distinct recommender systems—description-based, rating-based, and a combined model—comparing their outputs against the widely-used Goodreads recommendation system. Utilizing a comprehensive dataset from Goodreads, comprising over 10 million book entries, we developed and tested each model to assess how closely they align with user genre preferences reflected in Goodreads' recommendations. Through rigorous preprocessing, including data merging, cleaning, and transformation, we prepared the dataset for analysis. The models utilized techniques such as TF-IDF vectorization for book descriptions and cosine similarity for rating comparisons, with an innovative approach of integrating both features in the combined model. Initial results indicate varying levels of genre alignment across the models, with the combined model showing a promising balance between descriptive relevance and user ratings. This comparison not only highlights the potential of hybrid recommender systems in enhancing user experience but also provides insights into the critical features that influence genre-specific book recommendations. The findings suggest directions for future research, particularly in optimizing recommender systems for enhanced accuracy and user satisfaction in digital book platforms.

1. Introduction

In the digital era, recommender systems have become significantly important in guiding user choices across various platforms, especially in sectors where the selection is wide and varied, such as in the book industry. As the volume of available online content continues to increase, the need for efficient and accurate recommendation systems becomes more critical. These systems not only contribute to better user experience by providing personalized suggestions but also play a crucial role in increasing user engagement and satisfaction.

This study focuses on the genre relevance of book recommendations by developing and comparing three distinct models: a description-based model, a user rating-based model, and a combined model that integrates both descriptive and user rating data. The effectiveness of these models is evaluated against the recommendations made by Goodreads, a widely used platform with over 10 million book entries. This comparison aims to determine whether there is a significant difference in genre relevance between the recommendations generated by our models and those from Goodreads.

The motivation for this research stems from the desire to align our recommendation system closely with Goodreads' system, given its broad adoption and proven user engagement. Using the Goodreads dataset from Kaggle, this study develops and assesses each model's ability to mirror user genre preferences, as reflected in Goodreads' recommendations. By doing so, we not only aim to enhance the recommendation accuracy of our systems but also seek to contribute valuable insights into the development of more sophisticated, genre-sensitive recommender systems for digital book platforms.

Our hypothesis is that the combined model will outperform the individual description and rating-based models. We anticipate that integrating descriptive data, which is crucial for genre alignment, with user ratings, which reflect book popularity and user satisfaction, will result in a more robust and effective recommender system.

2. Related Work

This section discusses previous studies on book recommendation systems that inspired our research and identifies literature gaps our project seeks to explore.

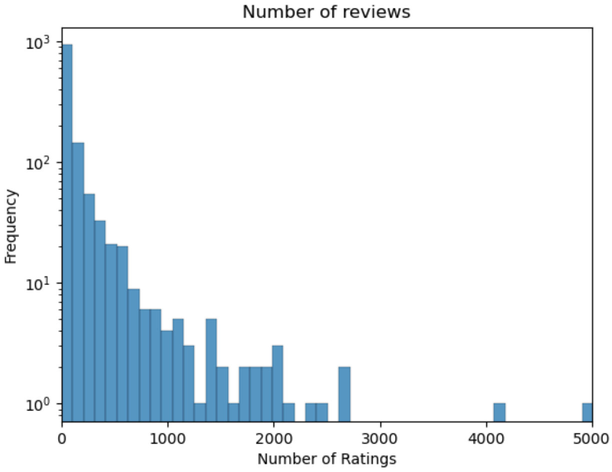
One notable study is the "Web Based Book Recommendation System Using Collaborative Filtering," which emphasizes the use of recommender systems by online bookstores to improve customer engagement and increase revenue (Ketaki Mankar et al., 2023). This research implements a machine learning model using the K-nearest neighbors (KNN) algorithm to adapt recommendations based on user ratings and preferences. The system's architecture facilitates a dynamic feedback loop where user interactions continually refine the recommendation accuracy.

Our thoughts on the use of KNN for collaborative filtering were mainly positive; we recognized it as a clever and innovative approach. However, we observed that this research was limited to user ratings and preferences, omitting a critical feature when assessing a book’s appeal—the book's content. This oversight prompted us to develop a model based on book description, another one based on user ratings and a hybrid model that combines both elements. Our intention is to explore whether a combined approach can outperform the traditional methods in aligning recommendations more closely with users' genre preferences.

Another significant study is "Personalized Book Recommendation System using Machine Learning Algorithm," which addresses the challenges posed by the expansion of online book availability amplified by the COVID-19 pandemic (Sarma et al., 2021). This system employs a content-based method using K-means clustering and cosine similarity to enhance the relevance of recommendations. By focusing on book content rather than just user ratings, the system effectively filters out less engaging books, as indicated by its high specificity scores across multiple datasets. This study incorporates a content-based approach and suggests that recommendations grounded in the actual content of books are more effective and satisfying for users than approaches solely based on user preferences. This finding resonated with our project objectives, prompting us to integrate a similar approach in our models. Furthermore, the successful use of cosine similarity in this study influenced our decision to employ this method in our models.

3. Data Collection (Brief)

Before starting preprocessing and the development of our models, it was important to select appropriate data that was fitting for the objective of our research. We acquired our data from Kaggle, the biggest online data science platform with a wide variety of both tools and resources (*Goodreads Book Datasets with User Rating 2M*, n.d.). When considering different datasets, we were focused on finding a dataset with a high number of reviews. Given this criterion, we finally chose the data files uploaded by Bahram Jannesar and an unspecified collaborator. The csv files have a usability rating of 9.7 and were uploaded three years ago. Initially, we started with one file regarding all the books, which had 19 columns and 55156 rows (referred to as 'book data' from now on). After some short preprocessing, the ratings file (referred to as 'ratings data' from now on) consisted of 362596 rows and 3 columns. The rows that we used from the books data for the development of our models were: Name (str), CountsofReview (int), Language (str) and Description (str). The main row that was used from the ratings data was: Rating (str). Furthermore, we performed a brief summary analysis of the reviews to gain insight in its distribution regarding the books. More information about consequent preprocessing will be discussed later on in this report.



4. Methods with Description of Main Algorithms

The preprocessing of our dataset was a multi-step process, designed to clean and structure our data for effective analysis in developing book recommender systems. Initially, we imported multiple datasets into pandas dataframes, specifically loading book details into 'books_df' and user ratings into 'ratings_df' from CSV files. Given the extensive number of books in 'books_df' and the comparatively smaller number of ratings across multiple files, we concataned several CSV files containing user ratings into a single dataframe.

We merged 'books_df' and 'ratings_df' dataframes based on book names which automatically excluded books that did not have corresponding entries in both datasets. We then renamed the 'Id' column in the merged dataset to 'bookID' and the user ID to 'userID' to clarify identifiers and prevent confusion. We proceeded to filter non-English books which reduced the dataset to 1273 rows while maintaining 21 columns as our target audience was english speaking readers. Therefore, we considered non english books irrelevant to our analysis.

We assessed the missing values in our dataset and discovered that they were only present in the ISBN, Publisher, and Language columns. Given the focus on genre alignment for our recommendation system, missing values in Publisher and ISBN were ignored as they were not critical. However, the Language column, crucial for book recommendations, was treated with mode imputation to maintain dataset integrity and reflect user preferences accurately. Textual data processing involved creating a function to tokenize and lemmatize text from the Name, Authors, Publishers, and Rating_y columns, which helped normalize and simplify the textual data for better analysis and machine learning model performance. Finally, our dataset was split into training and validation sets to evaluate model performance and generalizability.

For model 1, we calculated similarity scores based on book descriptions. We initiated this process by developing a function that uses TfidfVectorizer to transform book descriptions into a TF-IDF matrix. In this matrix, each row represents a book, while each column corresponds to the TF-IDF score of a specific word. Following this transformation, we computed the cosine similarity between every pair of books using their TF-IDF vectors, resulting in a similarity matrix. This matrix was then converted into a pandas DataFrame to facilitate more straightforward manipulation. To make practical use of this data, we implemented a recommendation function that takes book title as input. This function retrieves the similarity scores for the chosen book from the matrix, sorts these scores to identify the books with the highest similarity, and finally outputs the top N recommendations, complete with their similarity scores.

For model 2, we calculated the similarity scores for the “book rating” feature and implemented the number of “written reviews” feature as a weight to the vectors. Initially, we converted text-based ratings into numeric values and constructed a cosine similarity matrix from these user ratings to evaluate the similarity across books. We then applied a weighting scheme that considers the number of reviews each book has received to adjust the similarity scores, aiming to improve the reliability of recommendations by favoring books with more user feedback. A recommendation function was employed to again sort and select the top similar books to a given title, based on their similarity scores, and to generate a list of book recommendations.

For model 3, we created a combined model including “book description” and “book rating”.

We began by normalizing two distinct similarity matrices—one based on book descriptions and the other on user ratings—using MinMaxScaler. These normalized matrices were then combined into a single matrix using a weighted average method, where the balance between the two sources of similarity could be adjusted via an alpha parameter. We set this alpha parameter to 0.5 to ensure both matrices had the same weight, therefore book description and rating had the same importance in the recommendations. Lastly, our recommendation function first verifies the presence of the book in the dataset. If found, it retrieves the book's similarity scores, sorts these scores, and identifies the top N books for recommendation.

5. Results and Findings

To examine the performance of our three models, we evaluated the results of two books in particular. For the first book, we ran a short code block that chose a random book from the books dataset. This was the book *The Devil's Dictionary* written by Ambrose Bierce. For the second book, we chose the book *Harry Potter and the Philosopher's Stone* by J.K. Rowling. We intended to include this book as we could examine whether our models would recommend other books from the series or not.

Evaluation of Model Performance

Firstly, we used Goodreads to acquire the top five genres describing the input book. For instance, Goodreads provided the following five genres for *The Devil's Dictionary*: humor, nonfiction, classics, reference and language. Afterwards, we inspected the top five recommended books for *The Devil's Dictionary* and their corresponding provided genres. In case a book had a fewer number of genres displayed, the maximum number of genres was notes. For instance, the first Goodreads recommendation for *The Devil's Dictionary* was *The most popular president who ever lived (so far)* by Nancy Omeara and had four genres: contemporary, drama, novels, fiction.

Consequently, we ran our three models. This provided us with five recommended books per model, thus 15 books. For each book, we also indicated its top five genres as specified by Goodreads. For instance, the description model's first recommended book for *The Devil's Dictionary* was *The Describer's Dictionary: A Treasury of Terms & Literary Quotations, written by David Grambs* and had the corresponding five genres: writing, reference, nonfiction, language and dictionary.

We generated results for four recommendation models: the Goodreads recommender, our description recommender, our ratings model and our combination model. In total, we yielded twenty recommended books with their corresponding genres as indicated by Goodreads. In order to evaluate the performance of our models, we compared our models' proposed genres to the genres of the test book and the genres of its recommended books. We conducted this evaluation for both books. An overview of our results can be found in the appendix of this report.

Results 1: The Describer's Dictionary

Our description model recommended books with very similar genres in comparison to the Goodreads recommendations. However, the genres recommended by the ratings model were very different from those of the input's book and its corresponding Goodreads recommendations. It is interesting to highlight the functioning of the combination model. Here, we see that the first book as recommended by the description model is also recommended by the combination model, but is the fourth recommendation instead. This demonstrates that the combination model incorporates the description, but the recommended title drops to fourth place due to the influence of the ratings model.

Results 2: Harry Potter and The Philosopher's Stone

For the Harry Potter book, we were mainly interested in examining whether our models would propose other books of the series. Indeed, both the description model and the combination model recommended the three other books that were in the data set. It is interesting to note that both five recommendations were exactly equal for these two models. Lastly, the ratings model did not recommend any other books of the series, nor did it really provide any similar genres to the input book.

6. Conclusion

Our study set out to examine the genre relevance of book recommendations generated by three distinct models—description-based, rating-based, and a combined model—compared against the established Goodreads recommendation system. Contrary to our initial hypothesis, which posited that the combined model would offer superior results by integrating textual descriptions with user ratings, our findings indicated that the description model actually performed the best. This outcome suggests that the textual content of a book provides more reliable indicators of genre, aligning more closely with genre-specific user preferences than ratings alone or a hybrid approach. The ratings model was less effective in recommending books that matched the genres of the input model, highlighting a potential limitation in using user ratings as the sole basis for recommendations. While the combined model did show promise, particularly with books like "Harry Potter and the Philosopher’s Stone" where it mirrored the description model's recommendations exactly, it did not surpass the performance of the description model.

The primary limitation of our study was that our models were specifically developed to work on the training data and did not generalize well to the validation data. This specialization restricted the reproducibility and scalability of our results, as performance might degrade when applied to new, unseen data. Additionally, the evaluation of recommendations was largely based on personal assessments, which may not universally reflect user satisfaction or genre accuracy.

To address these limitations and enhance the practical utility of our recommender systems, future research should focus on conducting user satisfaction surveys to gather direct feedback on the accuracy and user satisfaction of our models compared to established systems like Goodreads. This feedback would allow us to refine our models to better meet user expectations and preferences in real-world scenarios. Further research should also aim to expand the dataset to include a wider range of genres and books and integrate more sophisticated natural language processing techniques to better capture genre details from descriptions.

7. Author Contributions

The division of tasks was made to ensure each team member contributed in an equal manner. We all sought feedback from each other and supported one another throughout every stage of this project. Our collaborative effort and teamwork was key to the project's success.

Preprocessing (Stage 1):

Andrea was in charge of loading the data, making necessary adjustments, and ensuring that the dataset was clean, merged appropriately, and ready for analysis. This included data exploration, handling missing values and textual data processing.

Building the Recommender Systems (RSs) (Stage 2):

Model 1→ Description-based Recommender System:
Clarice & Anne Marijn were responsible for developing the algorithm that calculated similarity scores based on book descriptions, employing techniques such as TF-IDF vectorization and cosine similarity.

Model 2 → Rating-based Recommender System:
Tereza developed a model that used user ratings, employing cosine similarity as the core technique. She integrated the number of written reviews as weights within the cosine similarity calculations.

Model 3 → Combined Recommender System:
Andrea developed the hybrid model that integrated the scores from both the description-based and rating-based systems, adjusting equal weight to both features.

Comparison and Evaluation (Stage 3):

Clarice, with the assistance of Anne Marijn,Tereza and Andrea, was tasked with comparing the outputs from our recommender systems with those from Goodreads. This involved collecting and manually analyzing the genres of recommended books from both our systems and Goodreads to evaluate alignment and effectiveness.

References

GoodReads. (2022). *Goodreads*. Goodreads. https://www.goodreads.com/

Goodreads Book Datasets With User Rating 2M. (n.d.). Www.kaggle.com. Retrieved June 2, 2024, from https://www.kaggle.com/datasets/bahramjannesarr/goodreads-book-datasets-10m/data

Ketaki Mankar, Pawar, S., Agarwal, H., Tejas Sangale, & Kulkarni, S. (2023). Web Based Book Recommendation System Using Collaborative Filtering. *International Conference on Emerging Smart Computing and Informatics*, 1–6. https://doi.org/10.1109/esci56872.2023.10099750

Sarma, D., Mittra, T., & Shahadat, M. (2021). Personalized Book Recommendation System using Machine Learning Algorithm. *International Journal of Advanced Computer Science and Applications*, *12*(1). https://doi.org/10.14569/ijacsa.2021.0120126

Appendix 1: Evaluation of Models

Book 1: Random Book

Book 1: The Devil's Dictionary by Ambrose Bierce
Genres: **Humor, nonfiction, classics, reference, language**

Genres from Goodreads
1: The most popular president who ever lived (so far)
2: Our dumb world: the onion's atlas of the planet earth
3: Coyote Blue
4: When will Jesus bring the pork chops?
5: Cake wrecks: when professional cakes go hilariously wrong

Title	**Genres**
Recommendation 1	Contemporary, drama, novels, fiction
Recommendation 2	Humour, comedy, nonfiction, reference, geography
Recommendation 3	Humour, fiction, fantasy, comedy, urban fantasy
Recommendation 4	Humour, nonfiction, comedy, (essays, audiobook), religion, American
Recommendation 5	Humour, nonfiction, food, photography, comedy

Genres from model 1 (description)
1: The Describer's Dictionary: A Treasury of Terms & Literary Quotations, written by David Grambs
2: A Concise Chinese-English Dictionary for Lovers, written by Xiaolu Guo
3: Blithe Spirit, written by Noël Coward
4: The Importance of Being Earnest, written by Oscar Wilde
5: Essays, written by George Orwell

Title	**Genre**
Recommendation 1	Writing, reference, nonfiction, language, dictionaries
Recommendation 2	Fiction, China, Romance, contemporary, asia
Recommendation 3	Plays, drama, classics, humour, fiction
Recommendation 4	Classics, plays, fiction, drama, humour

Recommendation 5	(essays) Nonfiction, essays, politics, classics, philosophy

Genres from model 2 (rating)

1. Dancing Carl, written by Gary Paulsen
2. Seven Gothic Tales, written by Isak Dinesen
3. Pawn of Prophecy (The Belgariad, #1), written by David Eddings
4. The Real All Americans: The Team That Changed a Game, a People, a Nation, written by Sally Jenkins

Title	**Genre**
Recommendation 1	Young adult, fiction, middle grade, war, childrens
Recommendation 2	Short stories, fiction, gothic, classics, horror
Recommendation 3	Fantasy, fiction, epic fantasy, science fiction fantasy, high fantasy
Recommendation 4	Sports, history, nonfiction, football, american history
Recommendation 5	Graphic novels, comics, young adult, fiction, asia

Model 3: combined model

1. Grendel, written by John Gardner
2. The Real All Americans: The Team That Changed a Game, a People, a Nation, written by Sally Jenkins
3. The Greatest Generation Speaks: Letters and Reflections, written by Tom Brokaw
4. The Describer's Dictionary: A Treasury of Terms & Literary Quotations, written by David Grambs
5. Seven Gothic Tales, written by Isak Dinesen

Title	**Genre**
Recommendation 1	Fiction, fantasy, classics, mythology, school
Recommendation 2	Sports, history, nonfiction, football, american history
Recommendation 3	Nonfiction, history, WWII, war, biography
Recommendation 4	Writing, reference, nonfiction, language, dictionaries
Recommendation 5	Short stories, fiction, gothic, classics, horror

Book 2: Harry Potter

Book 1: Harry Potter and the Philosopher’s Stone
Genres: Fantasy, Fiction, Young Adult, Magic, Childrens

Recommendations from Goodreads

1. **The Hunger Games (Suzanne Collins)**
2. **Twilight (Stephanie Meyer)**
3. **The Hobbit (J.R.R. Tolkien)**
4. **Divergent (Veronica Roth)**

5. The Fault in Our Stars (John Green)

Title	**Genre**
Recommendation 1	Young Adult, Fiction, Fantasy, Science Fiction, Teen
Recommendation 2	Fantasy, Young Adult, Romance, Fiction, Vampires
Recommendation 3	Fantasy, Classics, Fiction, Adventure, Young Adult
Recommendation 4	Young Adult, Dystopia, Fantasy, Fiction, Science Fiction
Recommendation 5	Young Adult, Fiction, Contemporary, Realistic Fiction, Teen

Model 1: Description
1: Harry Potter and the Chamber of Secrets (Harry Potter, #2), written by J.K. Rowling
2: Harry Potter and the Prisoner of Azkaban (Harry Potter, #3), written by J.K. Rowling
3: Harry Potter and the Half-Blood Prince (Harry Potter, #6), written by J.K. Rowling
4: Wish List, written by Lisa Kleypas
5: Essays, written by George Orwell

Title	**Genre**
Recommendation 1	Fantasy, Fiction, Young Adult, Magic, Childrens
Recommendation 2	Fantasy, Fiction, Young Adult, Magic, Childrens
Recommendation 3	Harry Potter and the Half-Blood Prince
Recommendation 4	Historical romance, Romance, Anthologies, Historical, Christmas
Recommendation 5	Nonfiction, essays, politics, classics, philosophy

Model 2: Ratings
1. The Illearth War (The Chronicles of Thomas Covenant the Unbeliever, #2), written by Stephen R. Donaldson
2. A Canticle for Leibowitz (St. Leibowitz, #1), written by Walter M. Miller Jr.
3. Modern Man in Search of a Soul, written by C.G. Jung
4. Critical Discourse Analysis: The Critical Study of Language, written by Norman Fairclough
5. Cancer Ward, written by Aleksandr Solzhenitsyn

Title	**Genre**
Recommendation 1	Fantasy, Fiction, Science Fiction Fantasy, Epic Fantasy, Science Fiction
Recommendation 2	Science Fiction, Fiction, Post Apocalyptic, Classics, Dystopia
Recommendation 3	Psychology, Philosophy, Nonfiction, Spirituality, Science
Recommendation 4	Linguistics, Nonfiction, Language, Research Methods, Academic

Recommendation 5	Fiction, Classics, Russia, Russian Literature, Literature

Model 3: Combination

1. Harry Potter and the Chamber of Secrets (Harry Potter, #2), written by J.K. Rowling
2. Harry Potter and the Prisoner of Azkaban (Harry Potter, #3), written by J.K. Rowling
3. Harry Potter and the Half-Blood Prince (Harry Potter, #6), written by J.K. Rowling
4. Wish List, written by Lisa Kleypas
5. Essays, written by George Orwell

Title	**Genre**
Recommendation 1	Fantasy, Fiction, Young Adult, Magic, Childrens
Recommendation 2	Fantasy, Fiction, Young Adult, Magic, Childrens
Recommendation 3	Fantasy, Fiction, Young Adult, Magic, Childrens
Recommendation 4	Historical romance, Romance, Anthologies, Historical, Christmas
Recommendation 5	Nonfiction, essays, politics, classics, philosophy