Book Recommendation System

Helly Champaneri
Dept. of Computer Science
University of Central Florida
helly_champaneri@ucf.edu

Sindhu Priya Davuluri
Dept. of Computer Science
University of Central Florida
sindhu_davuluri@knights.ucf.edu

Amrutha Nagaraj
Dept. of Computer Science
University of Central Florida
amruthanagaraj@knights.ucf.edu

Abstract : **With large number of products available on the e-commerce websites, Recommendation Systems have found importance in many domains. As a result of this, Book Recommendation Systems have been increasingly used by most of the online book retailers. These systems aim to predict users' interests based on their past interactions with the books, and recommend similar books that they would find interesting. These help the readers to discover new books, save time, and enhance their reading experience. Moreover, the online retailers also benefit from these systems, as recommending right book for a particular user leads to increase in their book sales, while also giving them an edge over their competitors in the market. In this paper, two models have been implemented in order to recommend books to the readers: Item-based Collaborative Filtering using KNN (k-Nearest Neighbor) and Singular Value Decomposition (SVD) using Matrix Factorization, on the Book Crossing Dataset. The dataset has been analyzed and the performance of both the models have been evaluated using RMSE and MAE metrics.**

I. INTRODUCTION

A. Background

With large number of books being available on the internet and increase in the popularity of online shopping, Book Recommendation Systems have become an essential tool for the e-commerce websites. These aim at providing suitable suggestions to the readers in order to improve sales and increase customer satisfaction. There are three ways in which a book recommendation system can be developed : Content-Based Approach, Collaborative Filtering-Based Approach and Hybrid Approach.

1. *Content-Based Approach* : These recommend items based on their features or attributes. They focus on the characteristics of the items and attempt to match them with the user's preferences.
2. *Collaborative Filtering (CF) Based Approach* : These make recommendations based on past interactions of users and target items. That is, these try to search for the look-alike customers (user-based CF) or look-alike items (Item-based CF). The user-based collaborative filtering involves finding other users who have similar preferences to the target user and recommending items that those similar users have liked. In item-based collaborative filtering, items which have received high ratings from the same set of users are considered as neighbors. A new user who has rated one of these items high, is likely to like the remaining items which were liked by the same group of users. Example: Item 1 and Item 3 are considered neighbors if they were positively rated by both User 1 and User 2. So, Item 1 can be recommended to User 3, if User 3 has already shown interest in Item 3.
3. *Hybrid Approach* : This method combines recommends the books based on a combination of content-based and collaborative filtering. It combines the user's past behavior and preferences for an item, with the book's content and features to provide a more personalized and accurate recommendation.

B. Problem Statement

By providing readers with personalized recommendations, a book recommendation system can enhance the overall user experience, making it more enjoyable and engaging for readers. Hence, this paper aims to develop an efficient book recommendation system which can help to solve the problems like, waste of reader's time in searching for a book and loss in retailer's sales. The objective of the paper is to develop efficient Book Recommendation Systems and to compare different models, in order to estimate the most robust algorithm. This is accomplished by generating ML models which can determine the missing rating values in the user-item matrix generated from the dataset. This would enable the model to predict the ratings of the books that the user has not rated yet. Using the predicted rating values, the model recommends the books which are most similar to the book rated highly by the user, hence generating the most relevant recommendations for any book given as input to the system by the user.

C. Importance

Book recommendation systems have increasingly found importance in many fields. This is because, they help readers to discover new books that they may not have otherwise known about. By using data on a reader's past reading habits, a book recommendation system can provide personalized recommendations that are tailored to the reader's interests. In addition to these, they can help online book retailers to increase book sales by promoting books to readers who are likely to enjoy them. Another benefit of using these is that they help readers save time by narrowing down their search for a new book as they no longer need to search through a large collection of books to find a right book for them.

D. Existing Literature

In [1], the proposed book recommendation system combines two methods, collaborative filtering and interest degree, to provide accurate and relevant book recommendations to users. The collaborative filtering algorithm is used to calculate the cosine similarity between users, while the interest degree considers various attributes of the book, such as search times, borrowing time, and renewing times. They demonstrate the potential of combining different methods and techniques to improve the performance of book recommendation systems.

In [2], recommendation systems utilize different approaches to provide relevant recommendations to users. The methods used are collaborative filtering and content-based filtering. Content-based filtering involves learning the content of an item, such as a book, and categorizing it based on a user's preferences learned from their profile. Collaborative filtering, on the other hand, matches items with users based on the idea that those who agreed in the past will agree in the future, without relying on content. The data about user preferences is collected through the ratings they give on the items.

E. Overview

In this project, two models have been implemented in order to recommend books to the readers : Item-based Collaborative Filtering using KNN (k-Nearest Neighbor) and Singular Value Decomposition (SVD) using Matrix Factorization algorithms on the Book Crossings dataset. The kNN algorithm in CF is used to identify similar items based on their proximity in a feature space. Whereas, SVD (Singular Value Decomposition) algorithm is a Matrix

Factorization method used to decompose the original user-item interaction matrix into three lower-dimensional matrices, in order to extract and analyze the latent (hidden) features from the dataset.

F. Data Collection and Description

The dataset used is 'Book-Crossing Dataset', which was collected by C. Ziegler [3] in 2004.The data on books are obtained from various sources, such as online bookstores, public libraries, book databases. and from online Amazon Web Services The dataset used for this paper's book recommendation system consists of three datasets: BX-Users, BX-Books, and BX-Books-Ratings. The BX-Users dataset contains information about users, including a unique 'user ID' that has been anonymized and mapped to integers. Demographic data such as the user's 'location' and 'age' are also available. The BX-Books dataset contains information about books, including a unique identifier called the 'ISBN.' Finally, the BX-Book-Ratings dataset contains information on how users have rated books on the scale of 0 to 10.

G. Components of the Proposed System

In this paper, two models have been implemented in order to recommend books to the readers : Item-based Collaborative Filtering using KNN (k-Nearest Neighbor) and Singular Value Decomposition (SVD) using Matrix Factorization. Using these, comparison will be made to determine the better approach.

H. Experimental Results

On implementation, it was observed that in both the models, both the models gave relevant recommendations for an input book. To evaluate the performance of the models, RMSE and MAE metrics were used and it was concluded that SVD algorithm gives lower RMSE and MAE values, and hence has lesser errors. Therefore, SVD using Matrix Factorization is more robust algorithm for book recommendation using Book Crossing dataset.

II. IMPORTANT CONCEPTS

A. Evaluation Metrics :

The metrics used in this paper to evaluate the performance of the Recommendation System are as follows.

1. *Mean Absolute Error (MAE)* : MAE calculates the value of absolute difference between the predicted ratings and the actual ratings given by a user for a particular item. Lower the MAE, better is the performance of the model. It is measured using equation (1), where N is the number of user-item pairs in the test dataset.

$$MAE = (1/N) * \Sigma |\text{actual rating} - \text{predicted rating}| \quad (1)$$

2. *Root Mean Squared Error (RMSE)* : This metric penalizes larger errors more severely as compared to MAE. Lower the RMSE, better is the performance of the model. That is, that the ratings predicted by the model are closer to the actual ratings for an item given by a user. It is measured using the equation (2), where N is the number of user-item pairs in the test dataset.

$$RMSE = \text{sqrt}(1/N * \Sigma(\text{actual rating} - \text{predicted rating})^2) \quad (2)$$

B. Choice Of Baseline Method for Comparison:

In this paper, Item-based CF using kNN algorithm is used as the Baseline Method to be utilized for comparison. Its performance will be compared with SVD-based method in order to estimate the model having better accuracy for the Book-Crossing dataset.

III. SYSTEM OVERVIEW

In this paper, two models have been implemented in order to recommend books to the readers : Item-based Collaborative Filtering using KNN (k-Nearest Neighbor) and Singular Value Decomposition (SVD) using Matrix Factorization. Collaborative Filtering (CF) based recommender systems are based on past interactions of users and target items. We try to search for the look-alike customers and offer products based on what their look-alike has chosen. CF has two approaches: user-based and item-based. In this paper, two Item-Based Collaborative Filtering approaches have been implemented. In item-based collaborative filtering, items which have received high ratings from the same set of users are considered as neighbors. A new user who has rated one of these items high, is likely to like the remaining items which were liked by the same group of users.

The kNN algorithm in CF is used to identify similar items based on their proximity in a feature space. Each item is represented by a set of features or attributes such as Ratings of that item given by various users. Hence, each item is represented by a Rating Vector. Similarity is calculated between pairs of items based on their Rating Vectors, using a distance metric such as cosine similarity. Then, the 'k' most similar items to a given item are recommended to the user. On the other hand, SVD (Singular Value Decomposition) algorithm is a Matrix Factorization method used to decompose the original user-item interaction matrix into three lower-dimensional matrices, representing : the users' preferences for latent features, the singular values that indicate the importance of each feature, and the items' attributes for each feature. By using these matrices, SVD can predict the unknown ratings for a user and item pair by finding the dot product of the corresponding row and columns.

IV. TECHNICAL DETAILS OF PROPOSED APPROACHES

This section includes the detailed Methodology and ML Pipeline followed for developing the two models for the Book Recommendation System.

A. Data Pre-processing

After importing the required modules and loading the dataset, the missing ratings values were filled with 0, since it will be useful while calculating distance between the rating vectors of items. The rows having null values in the BookTitle column were dropped from the dataset.

B. Data Visualisation

To understand feature space, data visualisation is performed. The distribution of Ratings for the books in the dataset were plotted. The age distribution of the readers was analysed. Since the Users dataset contains information about the user's location in terms of city, state, and country in a single column, these three components of location are split into 3 separate features. Plots are then used to realise the distribution of users around the geographic

locations. World-wide graphs were analysed to realise the average ratings given by readers of various nations and the average age distribution of users in different countries.

C. Feature Extraction

Ratings and books dataset were merged over ISBN feature. The duplicated ISBN feature was dropped from the merged dataset. Books are grouped by BookTitles and a new column for total rating count is created. The rating data is combined with the total rating count data. The resulting data helps to determine which books are popular and filter out the less popular ones in order to have a relevant dataset for recommendation. Books having total rating count of at least 50 are retained, to develop a more accurate model. Though only 1% of the books were retained after the threshold of 50 ratings was applied, the large dataset makes it sufficient for us to make accurate predictions using these books. From Data Visualisation, it is observed that maximum users are from USA, Canada and UK. So, users belonging to these countries are filtered and others are dropped from the dataset. This helps to avoid consuming large amount of memory and to save processing time.

D. Predictive Modelling

The details on developing the models for the two approaches is outlined in this section.

I. Item-Based Collaborative Filtering using KNN

A 2D user-item matrix (row : item and column : user), having ratings as values in the matrix, is generated from the dataset table. The missing values in the table are filled with zeros. Distance is calculated between the rating-vectors for each Items. To find the distance between the rating vectors, KNN algorithm is used with Cosine Similarity measure. The equation used to evaluate the cosine similarity between two items A and B is given by Equation (3). The value of 'k' is the hyper-parameter that needs to be set to an optimum value. For out code, 'k' was set to 10, to recommend 10 similar books to the user. KNN is hence used to find groups of similar items, and prediction are made based on the average rating of top-k nearest neighbours. Finally, the KNN model is fit on dataset. The 'k' nearest items to an item are recommended for a particular item.

$$\text{sim}(A,B) = \cos(\theta) = \frac{A.B}{||A||||B||} \tag{3}$$

II. Singular Value Decomposition (SVD) using Matrix Factorization

A user-item matrix, also known as utility matrix, containing ratings is generated, like in case of KNN. Its missing ratings are filled with 0s. The 'bookTitles', that is items are in rows and User-IDs are in columns. Using TruncatedSVD algorithm from sklearn library, and choosing the optimum value of the hyper-parameter 'n' (number of latent components to be considered), the truncatedSVD model is then fit on the matrix. For this dataset, n=12 was considered. This leads to leads to decomposition of the utility matrix into three lower-dimensional matrices. The resulting matrices can be interpreted as embeddings of users and items in a n-dimensional (n is lower than original dimensions) latent space, where the dot product of a user embedding and an item embedding approximates the user's rating for the item. After learning these embeddings, we can predict user ratings for items that the user has not yet interacted with by taking the dot product of their embeddings. Pearson's correlation coefficient is used to calculate the similarity between all of the book pairs. The books which have highest value

of Pearson's correlation coefficient (R), that is, greater than or equal to 0.9, are recommended for a particular item entered by that user.

V. RESULTS

A. Data Visualisation Results

From the experimental results, following inferences were made. The ratings in dataset were randomly distributed between 0 to 10 (10 being the highest), and most of them were rated as '0' as shown in Figure [1]. As per the Figure [2], age distribution plot of the users depicts that most active users are about 20 to 30 years old. In Figure [3], the pie chart representing the percentage of books published by various Publishers represents that Harlequin, Silhouette and Pocket publishers published the maximum number of books.

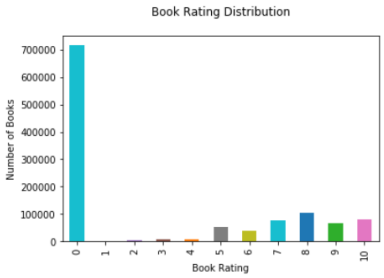

Figure 1. Book Ratings

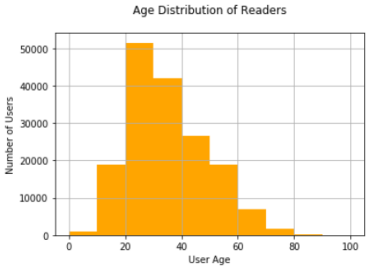

Figure 2. Age Distribution

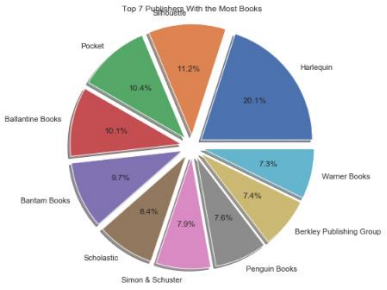

Figure 3. Publishers' Contribution

The Author VS Number of Books Written by Author plot given by Figure [4], shows that Agatha Christie wrote the maximum number of books. Most users belong to the city of London, while lesser belong from the city of Seattle as observed in Figure [5]. Figure [6] depicts the state of California has the maximum count of the readers. The country-wise distribution of users depicts that USA had most users, followed by Canada as in Figure [7].

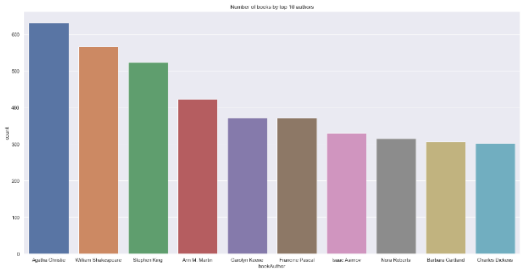

Figure 4. Number of books VS Authors

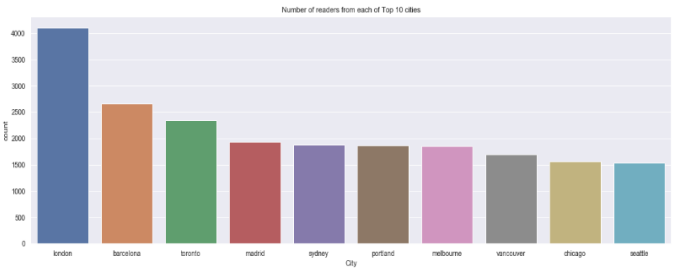

Figure 5. City-wise readers' distribution

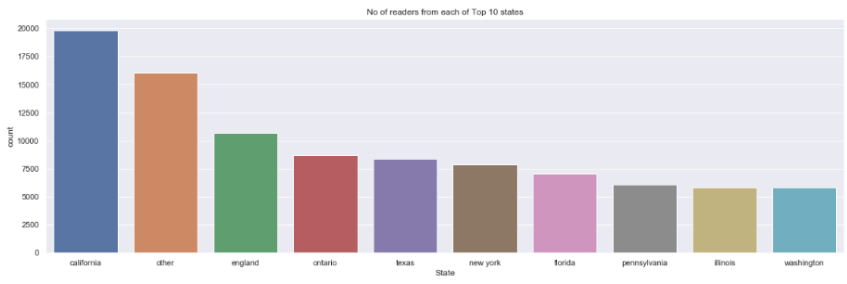

Figure 6. State-wise readers' distribution

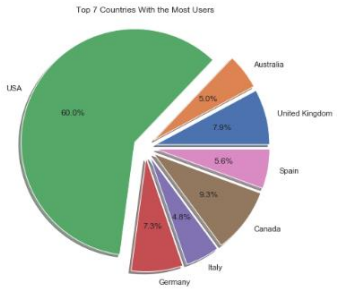

Figure 7. Country-wise readers' distribution

The world-wide distribution of Ratings by users depicted n Figure [8], shows that readers of Mongolia have given the highest average ratings, which is equal to 10. Whereas, the average ratings given by readers in Russia was lowest and was equal to 2. The world-wide distribution of average age of the Users given by Figure [9], represents that the average age of the readers in USA is 47 years, while most of the readers from all other nations are in the age group of 20 to 30 years.

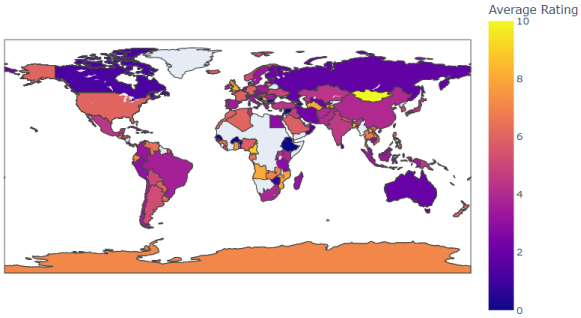

Figure 8. World-wide Average Ratings Distribution

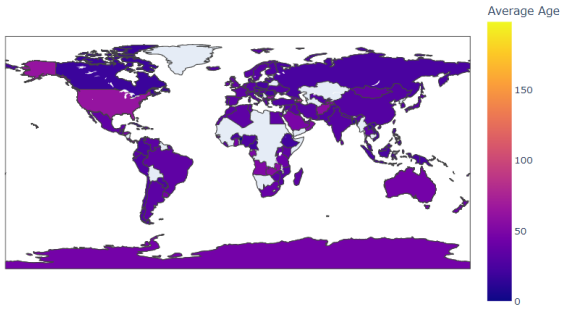

Figure 9. World-wide Average User Age Distribution

B. Results for Overall Performance of the Models

The overall performance of the models was tested and evaluated using different metrics. Using Item-Based Collaborative Filtering with kNN, top 10 books recommended for the input book : Insomnia, along with their similarity measures are listed by the model as shown in Figure [10]. The Figure [11] depicts that in SVD using Matrix Factorization algorithm, books having highest value of correlation coefficients are suggested to the user. For both the models, the performance estimation is done using cross-validate module of the surprise library, as depicted in Figure [12] for kNN-based CF model and in Figure [13] for SVD-based model. During cross-validation, the MAE and RMSE are calculated for each fold, and the average of these values across all folds is used to determine the performance of the recommender system.

```
Recommendations using KNN for the book Insomnia :

1. Rose Madder, distance = 0.483498922434937
2. Desperation, distance = 0.503110460489385
3. The Dark Half, distance = 0.518209443320151
4. Pet Sematary, distance = 0.5372411090157219
5. Nightmares &amp;amp; Dreamscapes, distance = 0.5429811312206052
6. Gerald's Game, distance = 0.5452422155613498
7. Four Past Midnight, distance = 0.5685038165512579
8. The Tommyknockers, distance = 0.5779543554590016
9. Misery, distance = 0.5978979260421607
10. The Regulators, distance = 0.6214865705904027
```

Figure 10. Recommendations by kNN-based CF

```
Recommendations Using Matrix Factorization for the Book : Insomnia
Bag of Bones
Carrie
Cujo
Desperation
Different Seasons
Dolores Claiborne
Everything's Eventual : 14 Dark Tales
Four Past Midnight
Gerald's Game
Hearts In Atlantis
Insomnia
Nightmares &amp;amp; Dreamscapes
Pet Sematary
Rose Madder
The Dark Half
The Dead Zone
The Regulators
The Tommyknockers
```

Figure 11. Recommendations by SVD

```
Evaluating RMSE, MAE of algorithm KNNBasic on 3 split(s).

                  Fold 1  Fold 2  Fold 3  Mean    Std
RMSE (testset)    1.7843  1.7715  1.7882  1.7813  0.0071
MAE (testset)     1.3238  1.3192  1.3290  1.3240  0.0040
Fit time          199.54  275.79  23.48   166.27  105.66
Test time         16.35   430.98  47.82   165.05  188.48

Item-Based CF using kNN RMSE :  1.7813136751150358
Item-Based CF using kNN MAE :  1.3240320899478637
```

Figure 12. kNN-based CF Accuracy

```
Evaluating RMSE, MAE of algorithm SVD on 3 split(s).

                  Fold 1  Fold 2  Fold 3  Mean    Std
RMSE (testset)    1.5987  1.6071  1.6020  1.6026  0.0034
MAE (testset)     1.2383  1.2429  1.2388  1.2400  0.0021
Fit time          7.21    6.49    7.04    6.91    0.31
Test time         0.53    0.48    0.47    0.49    0.03

SVD using Matrix Factorization RMSE :  1.6025953763356748
SVD using Matrix Factorization MAE :  1.2400076493773469
```

Figure 13. SVD Accuracy

RELATED WORK

In [7], recommendations were based on combined features of content-based filtering (CBF), collaborative filtering (CF) and association rule mining to produce efficient and effective recommendation. For this, a hybrid algorithm was proposed in which multiple algorithms were combined, so that it helps the recommendation system to recommend the book based on the buyer's interest. In [8], the proposed system enables both frequent-readers and rare-readers to easily get results that reflect their interests with their own content of interest as queries. The method identifies recommended books based on the similarity of the vectors of contents and emotions, contained in tweets about the content of user interests and book reviews. This was accomplished using content-based filtering (CB), collaborative filtering (CF), and hybrid systems that combine these two.

CONCLUSION

It is observed that the recommended books in case of both : Item-Based Collaborative Filtering using kNN and Singular Value Decomposition (SVD) using Matrix Factorization, are closely related to the input book : Insomnia. As observed in the results, the recommended books are based on the titles : Nightmares, Darkness, Dreamscapes and Midnight – which are related to Insomnia. Hence, both the models give accurate recommendations for an input book. To evaluate the performance of the models, RMSE and MAE metrics were used. It was concluded that SVD algorithm gives lower RMSE (=1.60) and MAE (=1.24) values than kNN-based CF's RMSE (=1.78) and MAE (=1.32). Therefore, SVD using Matrix Factorization is a more accurate algorithm as compared to kNN-based Collaborative Filtering algorithm.

REFERENCES

[1] Z. Wang, D. Hou , "*Research on Book Recommendation Algorithms Based on Collaborative Filtering and Interest Degree* ," Wireless Communications and Mobile Computing Conference, 2021, pp. 1-7

[2] N. Kurmashov, K. Latuta and A. Nussipbekov, "*Online book recommendation system*," 2015 Twelve International Conference on Electronics Computer and Computation (ICECCO), Almaty, Kazakhstan, 2015, pp. 1-4.

[3] C. Ziegler, "*Book-Crossing Dataset*", 2004, Institute for Informatics, University of Freiburg. Available : http://www2.informatik.uni-freiburg.de/~cziegler/BX/ [Online].

[4] P. Devika, K. Jyothisree, P. Rahul, S. Arjun and J. Narayanan, "*Book Recommendation System*," 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2021, pp. 1-5.

[5] M. Hussien, M. Khafagy, and M. Ibrahim, "*Recommender systems challenges and solutions survey*." 2019 International Conference on Innovative Technology.

[6] M. D. Ekstrand, "*Collaborative Filtering Recommender Systems*," Trends Human–Computer Interaction Conference 2010, Vol. 4, pp. 81– 173.

[7] A. Sachan and V. Richariya, "Survey on Recommender System based on Collaborative Technique", *Department of Computer Science And Engineering international Journal of Innovations in Engineering and Technology(IJIET) ISSN: 2319–1058*, vol. 2, no. 2, pp. 1-7, april 2013.

[8] T. Fujimoto and H. Murakami, "A Book Recommendation System Considering Contents and Emotions of User Interests," *2022 12th International Congress on Advanced Applied Informatics (IIAI-AAI)*, Kanazawa, Japan, 2022, pp. 154-157.

Project Code : https://drive.google.com/file/d/1i8Bequ-C8imbAC64_rLMaZ4Q8awOCUNz/view?usp=sharing