SNOTRA

Scalable Neighbor-based Online Textbook Recommendation Application

Kunaal Ahuja, Naman Arora, Abhishek Patria, Sai Abhishek Pidaparthi, Samadipa Saha

Introduction

We have built a book recommendation system (**SNOTRA**) using publicly accessible data (Goodreads API).

There are two ways it is done today, **content based** filtering and **collaborative** filtering.

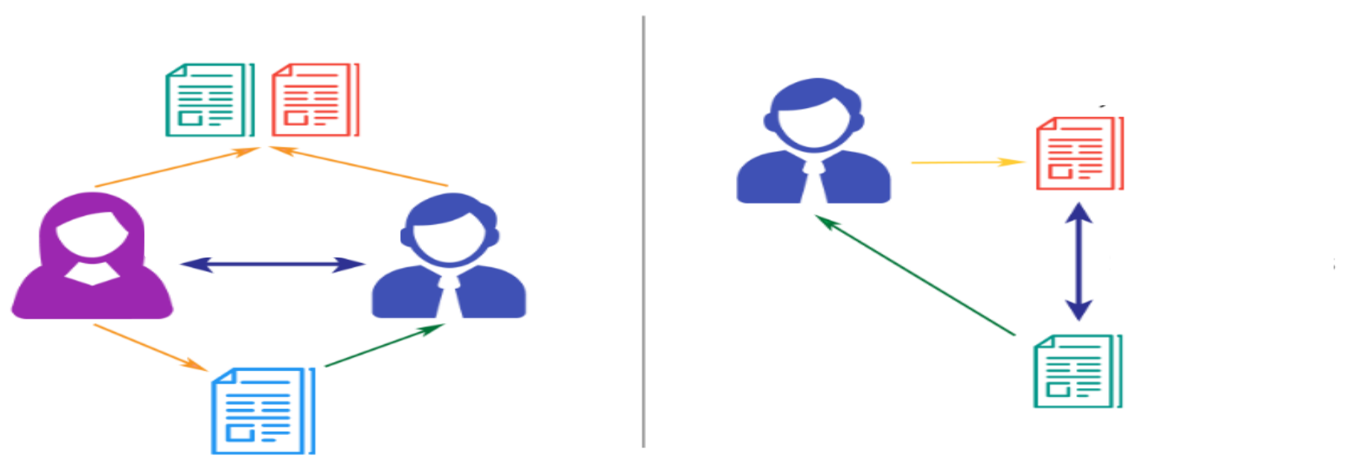


Figure 1: Collaborative (left) vs Content-Based (right) Filtering

We are integrating both the techniques to create a **hybrid** recommendation system to provide a better **personalized recommendation** to the user.

Approach

Collaborative Filtering

We performed user-to-user K-nearest neighbor collaborative filtering using cosine similarity on all the books.

We find the approximate rating that the user 'u' might give an unrated book 'i'. The formula is given below:

$$\hat{r}_{ui} = \mu_u + \sigma_u \frac{\sum\limits_{v \in N_i^k(u)} \text{sim}(u, v) \cdot (r_{vi} - \mu_v)/\sigma_v}{\sum\limits_{v \in N_i^k(u)} \text{sim}(u, v)}$$

We measured the accuracy of our model using RMSE metric.

Content Based Filtering

In this approach, we are recommending subset of those books to the users, which are similar to the user's reviewed books in the past.

To calculate similarity between 2 books, we use the book descriptions for both the books and pre-process the texts (removing any numbers and special characters, converting all text to lower case, removing stop words and lemmatizing the remaining words).

Then, we evaluate bag-of-words representation of the books and calculate cosine similarity to create similarity matrix for all books.

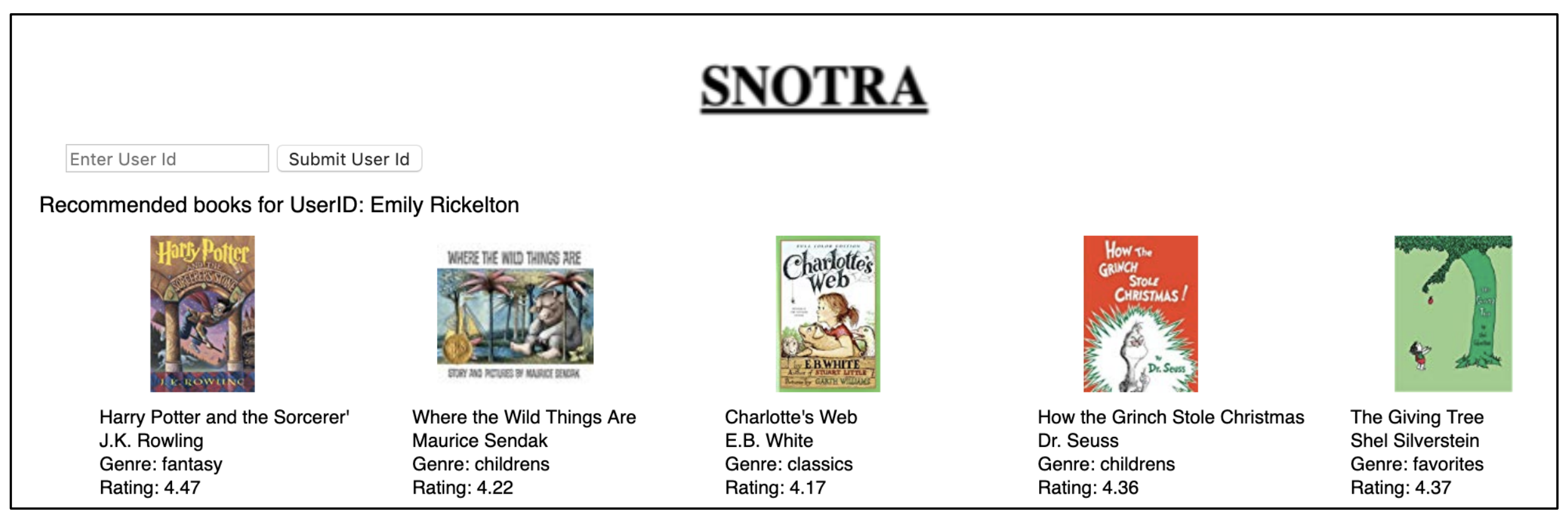
Hybrid Approach

We recommend the top books to the user as the **intersection** of books by the above methods which would give us the best of both worlds. In addition, we would give unique recommendations individually from the 2 approaches.

Visualization

Existing user

We use the hybrid model to find the recommended books for a user and display the top 5 books using D3. We show the name of the book, image of the cover, author, genre and the average rating of the book. The image of the cover is clickable and takes it to actual Goodreads page of the book on its website.



SNOTRA

Recommended books for UserID: Emily Rickelton

Harry Potter and the Sorcerer'
J.K. Rowling
Genre: fantasy
Rating: 4.47

Where the Wild Things Are
Maurice Sendak
Genre: childrens
Rating: 4.22

Charlotte's Web
E.B. White
Genre: classics
Rating: 4.17

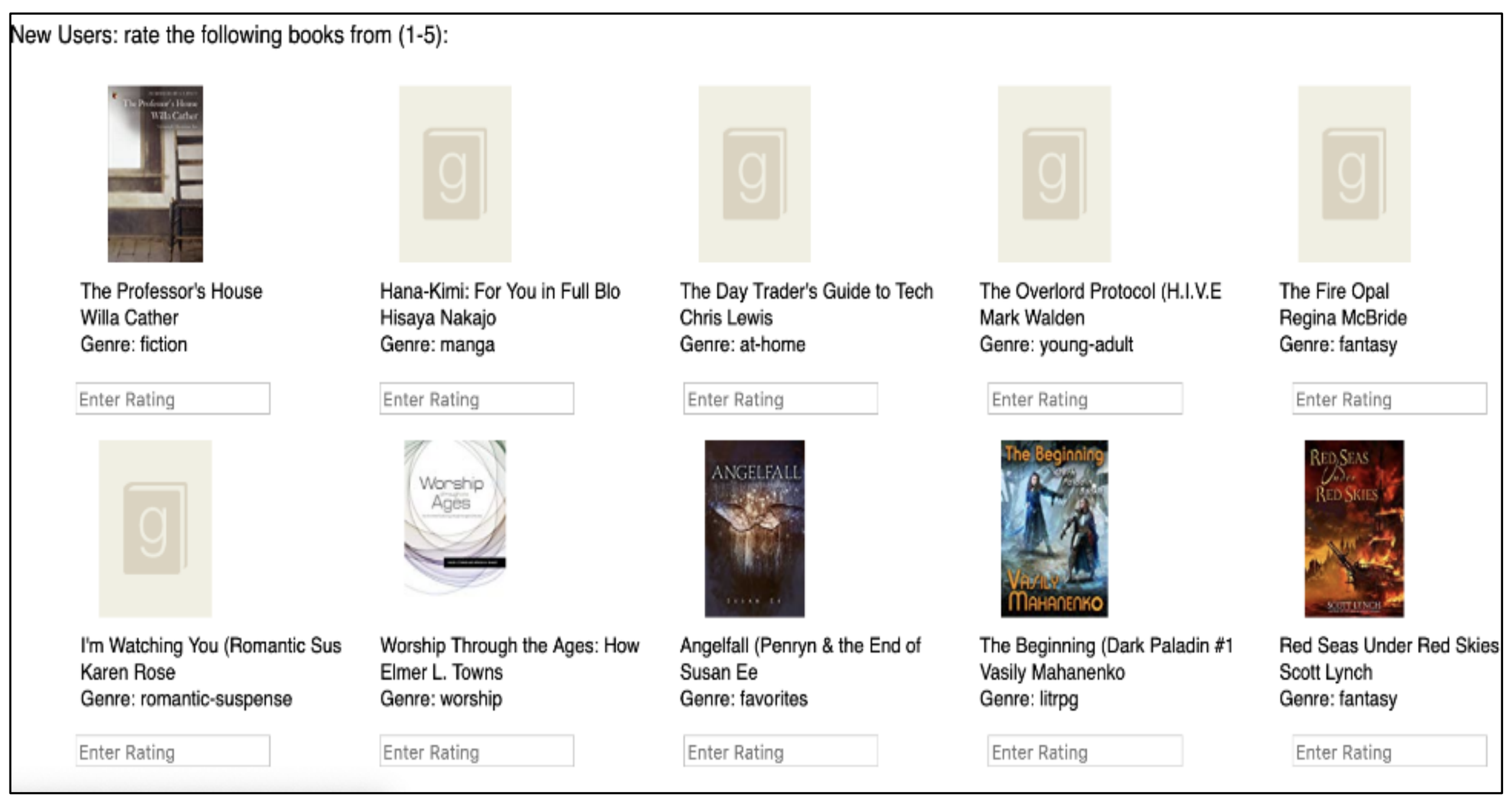
How the Grinch Stole Christmas
Dr. Seuss
Genre: childrens
Rating: 4.36

The Giving Tree
Shel Silverstein
Genre: favorites
Rating: 4.37

Figure 2: A snapshot of book recommendations for an existing user

New user

We ask a new user to rate 10 random books on the scale of 1 to 5 suggesting how likely he is to read these books. Based on his ratings, we are able to deploy our hybrid algorithm and give him the best set of recommendations.



New Users: rate the following books from (1-5):

The Professor's House
Willa Cather
Genre: fiction

Hana-Kimi: For You in Full Blo
Hisaya Nakajo
Genre: manga

The Day Trader's Guide to Tech
Chris Lewis
Genre: at-home

The Overlord Protocol (H.I.V.E
Mark Walden
Genre: young-adult

The Fire Opal
Regina McBride
Genre: fantasy

I'm Watching You (Romantic Sus
Karen Rose
Genre: romantic-suspense

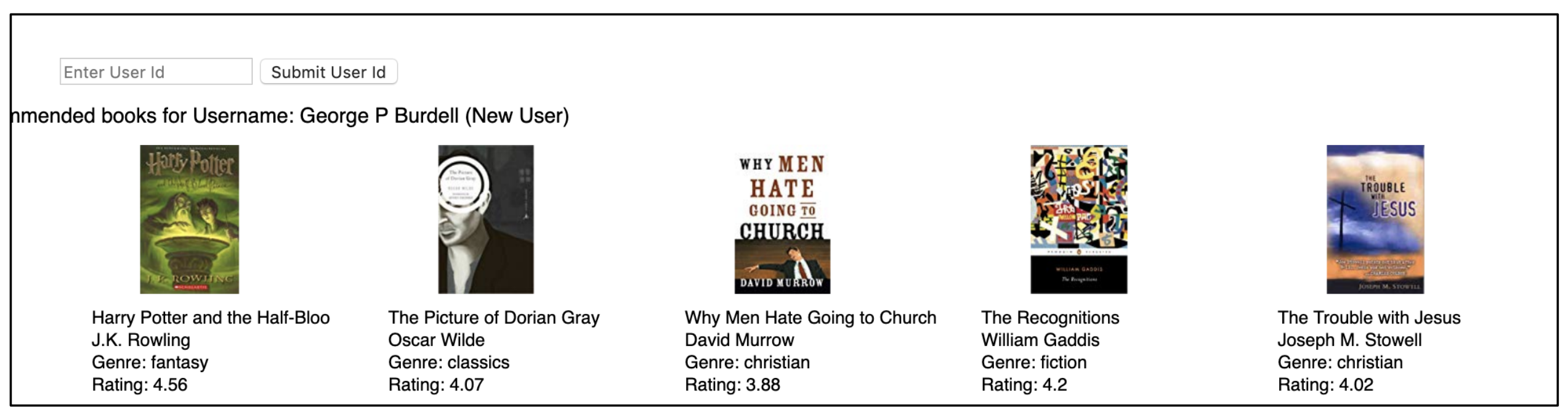
Worship Through the Ages: How
Elmer L. Towns
Genre: worship

Angelfall (Penryn & the End of
Susan Ee
Genre: favorites

The Beginning (Dark Paladin #1
Vasily Mahanenko
Genre: litrpg

Red Seas Under Red Skies
Scott Lynch
Genre: fantasy

Figure 3: A snapshot of books need to be rated by a new user



nmended books for Username: George P Burdell (New User)

Harry Potter and the Half-Bloo
J.K. Rowling
Genre: fantasy
Rating: 4.56

The Picture of Dorian Gray
Oscar Wilde
Genre: classics
Rating: 4.07

Why Men Hate Going to Church
David Murrow
Genre: christian
Rating: 3.88

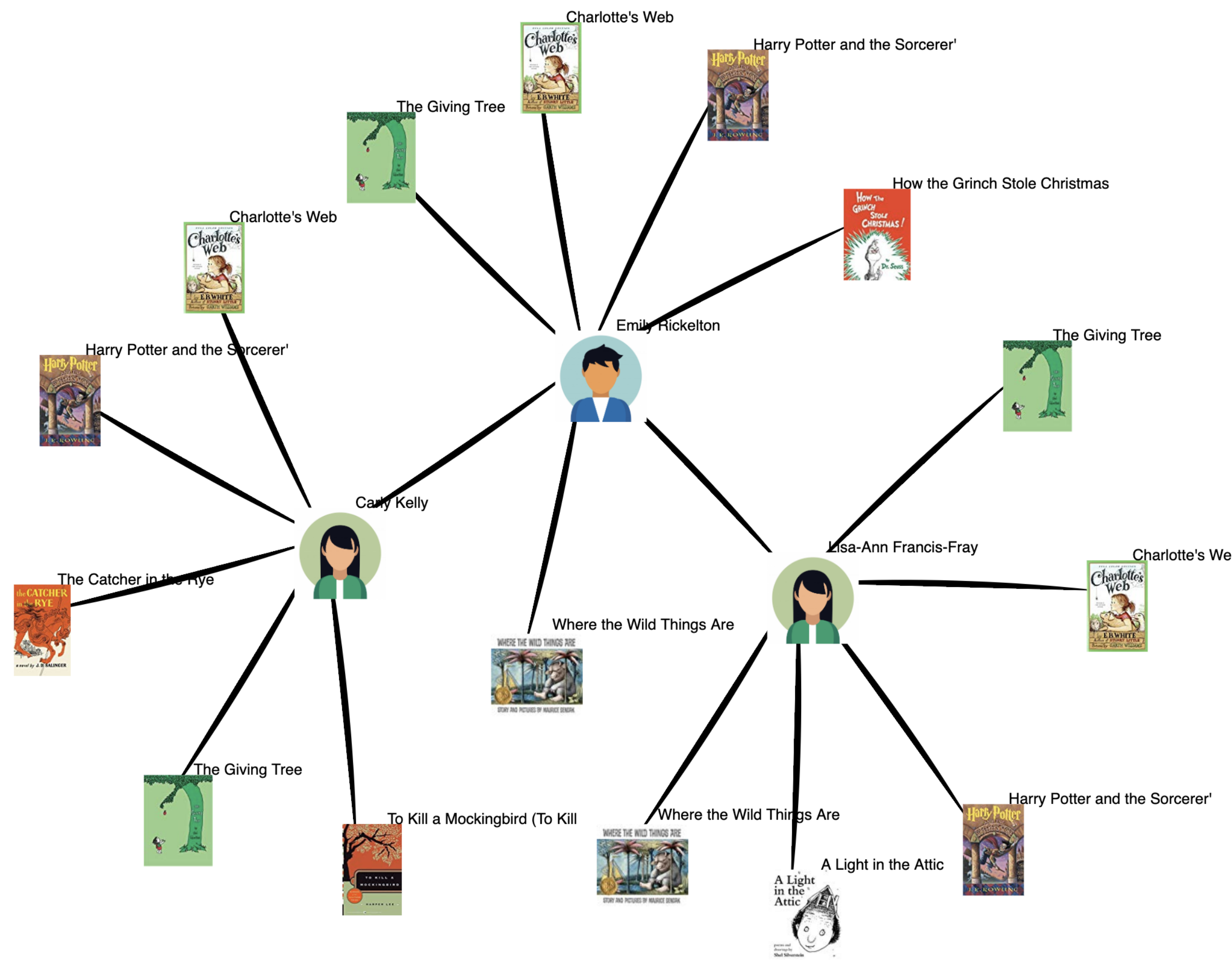
The Recognitions
William Gaddis
Genre: fiction
Rating: 4.2

The Trouble with Jesus
Joseph M. Stowell
Genre: christian
Rating: 4.02

Figure 4: A snapshot of book recommendations for a new user

Neighbor based graph

Another key visualization in our project is a force layout neighbor-based graph, which is also incorporated using D3. This graph shows a user's top 2 similar users along with top 5 rated books for the user as well as both the similar users.



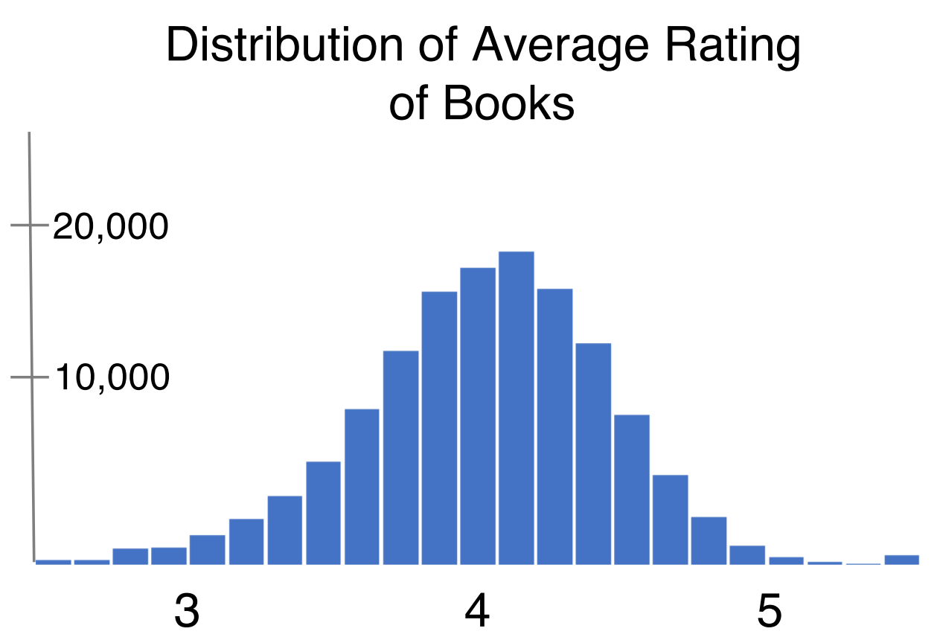
Charlotte's Web
Harry Potter and the Sorcerer'
The Giving Tree
How the Grinch Stole Christmas
Charlotte's Web
Emily Rickelton
Harry Potter and the Sorcerer'
The Giving Tree
Carly Kelly
Lisa-Ann Francis-Fray
Charlotte's Web
The Catcher in the Rye
Where the Wild Things Are
The Giving Tree
To Kill a Mockingbird (To Kill
Where the Wild Things Are
A Light in the Attic
Harry Potter and the Sorcerer'

Figure 5: A snapshot of force layout neighbor based graph

Data

Used **Goodreads** API to fetch all the data. Randomly selected users from a set of all Goodreads users and fetched their corresponding ratings of all the books they reviewed.

Dataset Size	**485 MB**
Total Reviews	***425k+***
Total Books	***150k+***



Experiments and Results

Collaborative Filtering

For collaborative filtering, we experimented with various KNN models: KNN basic, KNN with means and KNN baseline being among them. We also tried two approaches of similarity, cosine and Pearson similarity respectively. **KNN baseline** model with cosine similarity emerged as the best model.

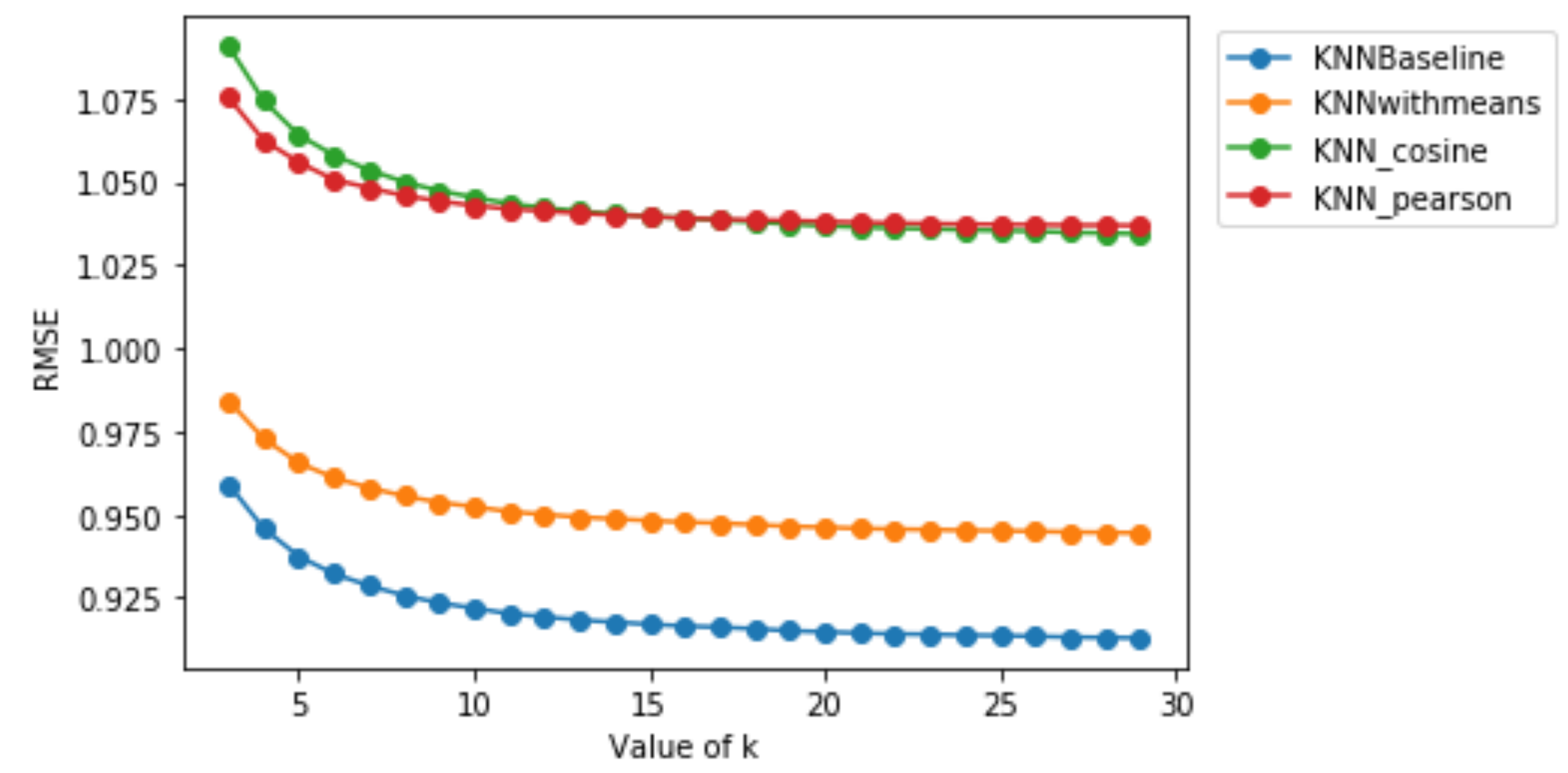


Figure 6: RMSE vs k for user-based collaborative filtering algorithm

User Survey

We conducted two user surveys with sample size 52. One for a basic collaborative filtering model and one for a hybrid model to see if there is any improvement in the recommendations.

32 out of 52 users like the recommendations of hybrid model compared to only 10 users who like the basic collaborative filtering model.

We performed t-test to compare the mean of the current ratings with that of the previous ratings, the p-value is very low suggesting means of the two ratings are statistically different.

t = 4.4908, df = 100.39, p-value = 1.899e-05
H_0: True difference in means is equal to 0
H_A: True difference in means is not equal to 0
95% confidence interval: (0.44,1.13)

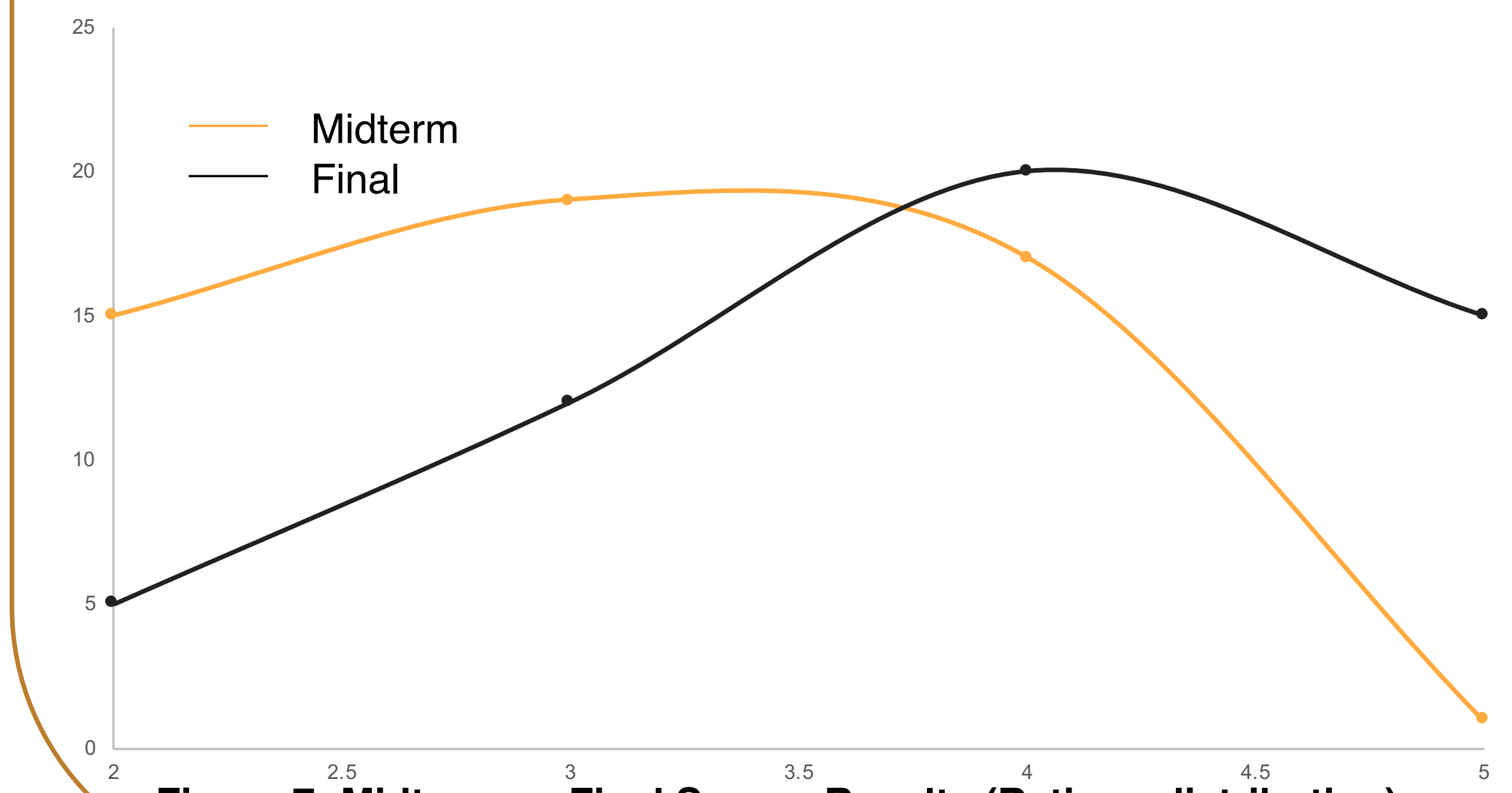


Figure 7: Midterm vs Final Survey Results (Ratings distribution)