Movie Recommendation System Report

Muhammad Abdullah Cheema

May 1, 2024

1 Data Exploration

During the data exploration phase, we examined the relationships between users' ratings and the ratings received by movies. The following plots (Figures 1 and 2) were generated to visualize these relationships.

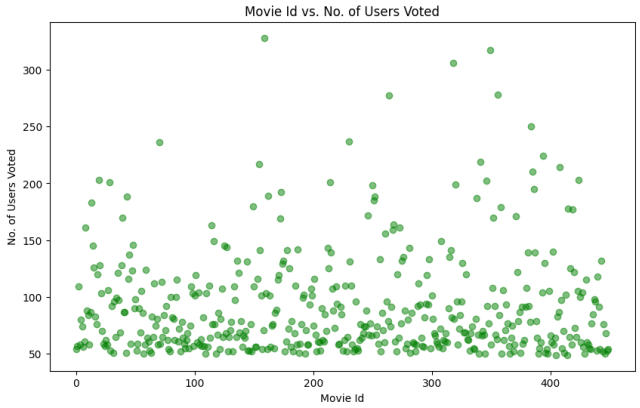

Figure 1: Movie ID vs No. of Ratings Given

2 Data Preprocessing

In the preprocessing stage, several steps were taken to prepare the data for modeling:

- Null values were replaced with 0.0 to handle missing data.
- Extra spaces were removed from the title column.
- We utilized a CSR matrix to represent the data due to its efficiency in handling sparse datasets. The CSR (Compressed Sparse Row) matrix format is advantageous in this context because it optimally stores large

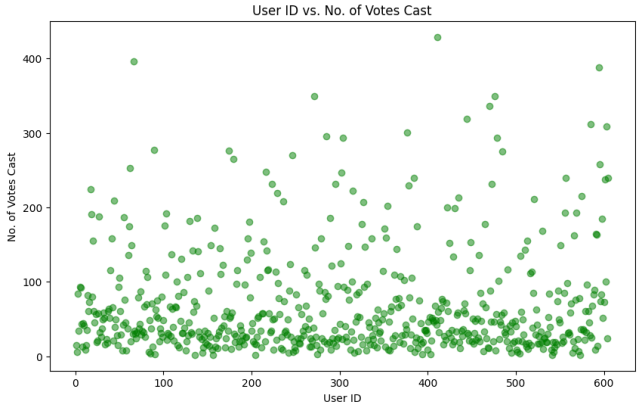

Figure 2: User ID vs No. of Ratings Given

matrices with mostly zero values, reducing memory consumption and computational overhead.

3 Model Selection

We opted for Item-based Collaborative Filtering for our recommendation system. This approach focuses on identifying similar movies based on their ratings by users. Using this technique, pairs of movies with similar user ratings are identified, and recommendations are made based on their similarity scores.

3.1 Advantages of Item-based Collaborative Filtering

- Stability: Movie preferences are more stable than individual user tastes.
- Scalability: It is computationally efficient, especially when dealing with a large number of items.
- Robustness: Less susceptible to shilling attacks compared to user-based approaches.

For implementing this technique, we employed the K-Nearest Neighbors (KNN) algorithm with a parameter of $n_{\text{neighbors}} = 5$.

4 Initial Model Evaluation

We evaluated the initial model by providing sample input-output pairs along with their respective similarity distances:

- American Pie (1999) $\rightarrow$ Austin Powers: The Spy Who Shagged Me (1999) (Distance: 0.366)

- Iron Man (2008) → Dark Knight, The (2008) (Distance: 0.329)
- Memento (2000) → Fight Club (1999) (Distance: 0.330)
- Some Like It Hot (1959) → Rear Window (1954) (Distance: 0.514)

5 Filtering Dataset to Remove Noise/Outliers

To address the sparsity of ratings and enhance the credibility of our recommendations, we applied filters to the dataset:

- Movies must have a minimum of 10 user votes to qualify.
- Users must have rated a minimum of 50 movies to qualify.

Following figures (3 and 4) represent the filtered dataset:

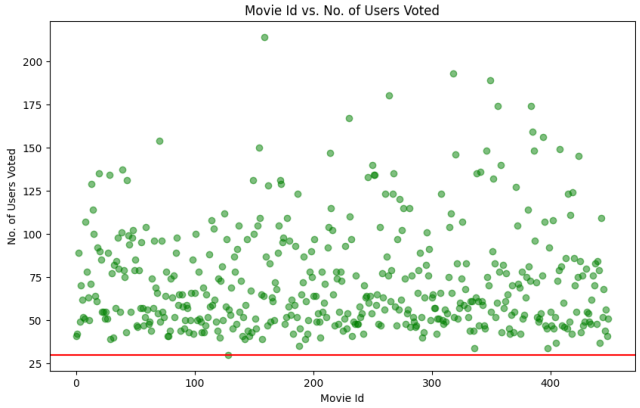

Figure 3: Movie ID vs No. of Ratings Given

6 Final Results Evaluation

After filtering the dataset, we observed a consistent reduction in similarity distances, indicating an improvement in model performance:

- American Pie (1999) → Austin Powers: The Spy Who Shagged Me (1999) (Distance: 0.327)
- Iron Man (2008) → Dark Knight, The (2008) (Distance: 0.240)
- Memento (2000) → Fight Club (1999) (Distance: 0.226)
- Some Like It Hot (1959) → Rear Window (1954) (Distance: 0.414)

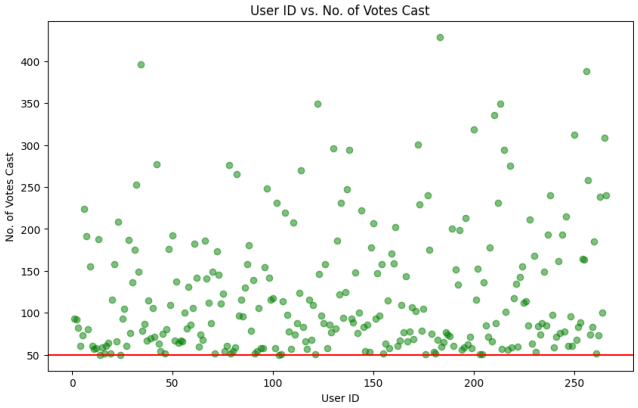

Figure 4: User ID vs No. of Ratings Given

7 Conclusion

In conclusion, our choice of using Item-based Collaborative Filtering with KNN algorithm has yielded promising results. The model demonstrates the capability to recommend similar movies based on user preferences, as evidenced by the reduction in similarity distances after filtering the dataset. Furthermore, the observed trend of recommending superhero movies for input superhero movies and thriller movies for input thriller movies aligns with our expectations, indicating the practicality and effectiveness of our approach.

Overall, the implemented movie recommendation system showcases the potential to provide relevant and personalized movie suggestions to users, thereby enhancing their movie-watching experience.