MOVIE RECOMMENDATION SYSTEM

Arnob Ghosh(2017-3-60-058)
Sk Mohammad Asem(2017-3-60-068)
Fahmida Jahan(2017-3-60-047)
Sumaiya Sultana Akhi(2017-3-60-063)

Abrtsact

September 13, 2021

A movie recommendation is vital in our social life due to its quality in giving enhanced entertainment. Such a system can suggest a set of movies to users based on their interest or the polarities of the movies. Nowadays, the recommendation system has made getting the things easily that we require .The main reason of Movie recommendation systems is to assist movie devotees by recommending what movie to watch without the hassle to have to go through the time-consuming process of choosing from a large collection of movies which go up to millions is monotonous and confusing. In this paper, we point to limit the human effort by suggesting movies based on the user's interface and preferences. To handle such issues, we presented a demonstration based on a content-based approach, collaborative filtering and demographic. This system recommends movies by coordinating examples provided by the user to movie substance, which system determines from the movie director, cast, genre assembled from movie records, without using any human-created metadata moreover shows in case the reviews are great or terrible..

1 Introduction

The recommendation system is an application that's used for prediction in different spaces throughout the internet. A large amount of information streams through the internet and it gives away a parcel of data with respect to the client-looking activity. The data extricated from the design of previously searched data can be molded into the prediction of important information for the user. The implementation of the system can be performed by different techniques. In this paper, we have discussed Content-Based Filtering, Collaborative Filtering Hybrid Content-Collaborative Based Filtering, KNN Based techniques. A recommendation system is a sort of data filtering system which attempts to predict the preferences of a user and make recommends based on these preferences. There are a wide variety of applications for recommendation systems. These have become increasingly popular over the final few years and are now utilized in most online platforms that we use.

2 Methodology

Our first task was to collect real data of movies. After collecting several data-set we have to find relevant data-set, and we preprocess the data. After that, we analyze the importance of the features. When we finalize the features we used three kinds of filtering (Content-based, collaborative, demographic, KNN) to build our recommendation system.

2.1 Content-Based Filtering

A common approach when designing recommerder systems is content-based filtering. Content-based filtering methods are based on a desecration of The item and a profile if the user's preferences these methods are best suited to situation whether there is known information in an item (title, location, description, etc.), but not on the user.

Content-based recommenders treat recommendation as a use-specific classification issue and learn a classifier for the users likes and dislikes based on items features. It this system, techniques are based on a single criterion value, the overall inclination of user for the item these systems try to predict a rating for unexplored items of exploiting preference information on different criteria that affect this in general preference value.

2.2 Collaborative filtering:

Collaborative filtering approaches build a model from a user's past behavior as well as comparable decisions made by other users. Collaborative filtering is based on the assumption that individuals who agreed within the past will agree with in the future, which they will like similar sorts of items as they liked within the past. The system produces proposals utilizing only data about rating profiles for different users or items. By finding peer users/items with a rating history similar to the current user or item, they generate recommendations utilizing this neighborhood. Collaborative filtering methods are classified as memory-based and model-based[1].

2.3 Finding similar user:

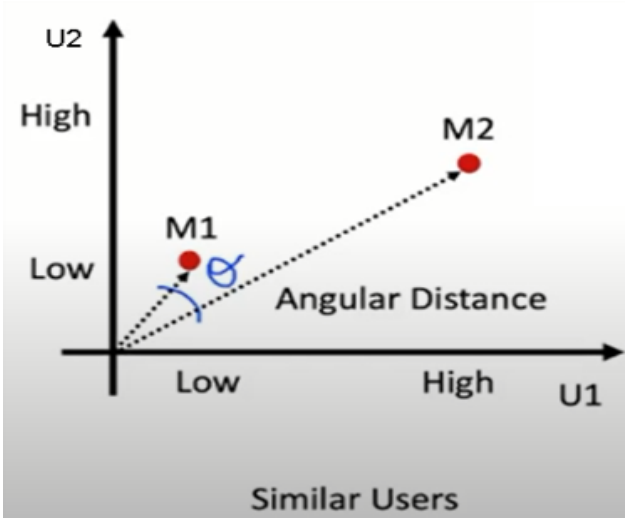


Figure 1: Similar Users

From this figure above, user1(U1) and user2(U2) are similar as per the taste of movies (M1 and M2). Pearson Correlation Coefficient is to find the similarity between these 2 users and recommend them the movies as per the genres they liked and disliked[2].

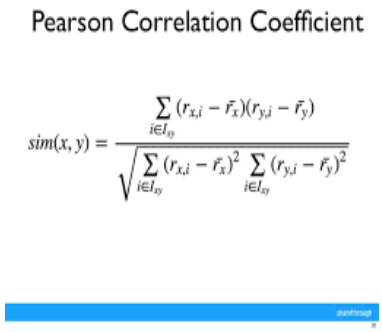
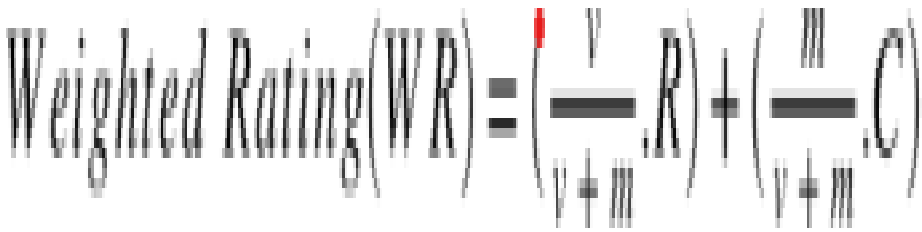


Figure 2: Pearson Correlation Coefficient

2.4 Demographic Filtering:

Demographic filtering (DF) classifies users according to their statistical data and suggests administrations accordingly. In DF the user profiles are made by classifying clients in stereotypical descriptions, representing the highlights of classes of users. Statistic data recognizes those clients that like related administrations. Semi-trusted third parties utilize DF to suggest administrations by utilizing information on individual clients. DF makes categories of clients who have similar demographic characteristics and then the cumulative buying behavior or preferences of users within these categories are being tracked. For a new user, recommendations are made by first finding which category he falls in and then the total buying preferences of previous users are applied to that category in which he belongs. Like collaborative techniques, demographic techniques moreover form “people-to-people” correlations but use dissimilar information. A collaborative and content-based strategy requires a history of client evaluation[3]

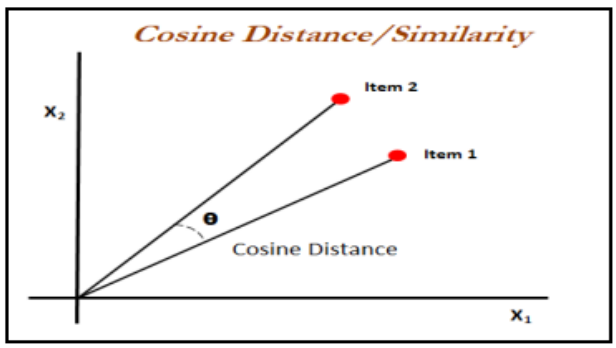


2.5 KNN Based Movie Recommendation System

K-Nearest Neighbors (KNN) is one of the simplest algorithms used in Machine Learning for regression and classification problems. KNN algorithms use data and classify modern data focuses based on similar measures. Classification is done by a larger vote to its neighbors. The data is assigned to the lesson which has the nearest neighbors. As you increase the number of nearest neighbors, the value of k, accuracy might increase.

2.6 Cosine Similarity

Similarity Score is a numeric value which ranges between Zeros to one. Which is used to determine the similarity of two items to each other on a scale of zero to one? This score is obtained by measuring the similarity between texts of both the documents. Therefore, similarity score can be defined as the measure of similarity between given text details of two given items. This can be done by- Cosine similarity. Cosine similarity is a measure used to determine how similar the texts are despite their size. To calculate the cosine angle between two vectors projected in a multi-dimensional space cosine similarity is used.[4].



$$similarity(A,B) = \frac{A \cdot B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Figure 3: Cosine Similarity

3 Implementation

3.1 Data collection

We have collected our data from Kaggle. We found several data-sets on Movies. For collaborative filtering we have two datasets for our recommendation system that are movies and ratings. In the "movies" dataset we have the attributes movie ID, title, genres. In the "ratings" dataset we have user ID, rating and timestamp.For other filtering we choose tmdb5000$_m ovies, tmdb_5000_c reditsandtofitinoursystembecausethisdata - setprovidesalotof features.Moviesdata_s ethastwentyattributesi.e'budget',' genres',' homepage',' id',' keywords',' original_l a setha sfiveattributesi.e'id',' title',' cast',' crew'.$

3.2 Data processing

At first we had to merge movies and credits by id and we handpicked some data that were imbalance data and then we had to do different types of processing in different filtering methods for different filtering such as we used cosine similarity for content-based filtering , Weighted rating for demographic filtering. For collaborative filtering we had to merge movies and ratings datasets by id and then we dropped the attribute 'timestamp' from our work as we don't need this as of now in our recommendation system.

3.3 Flow chart

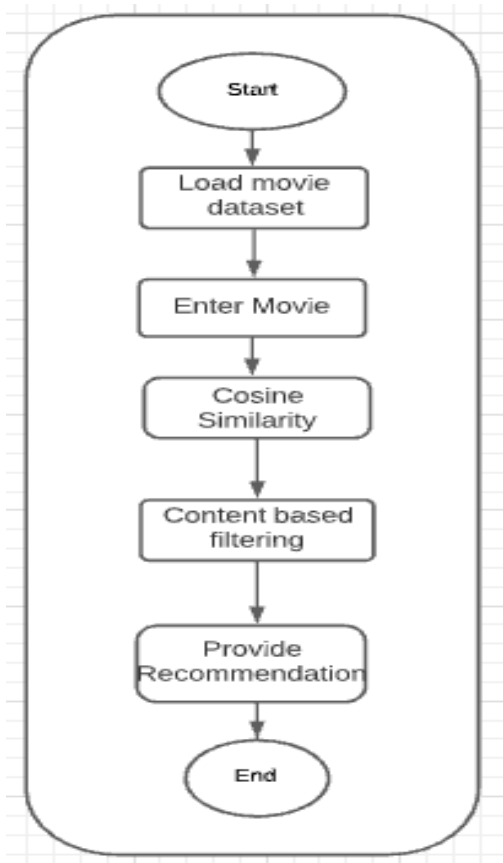


Figure 4: Content based

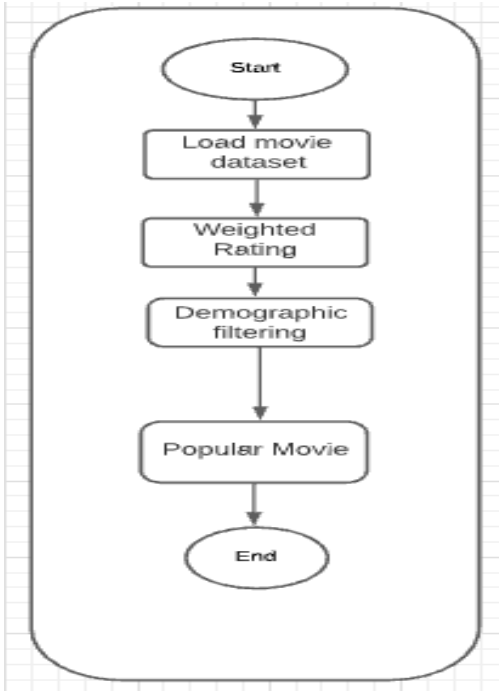


Figure 5: Demographic filtering

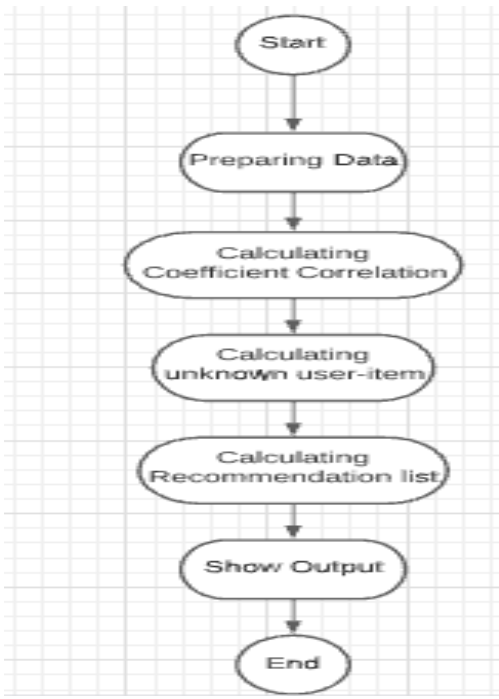


Figure 6: Collaborative

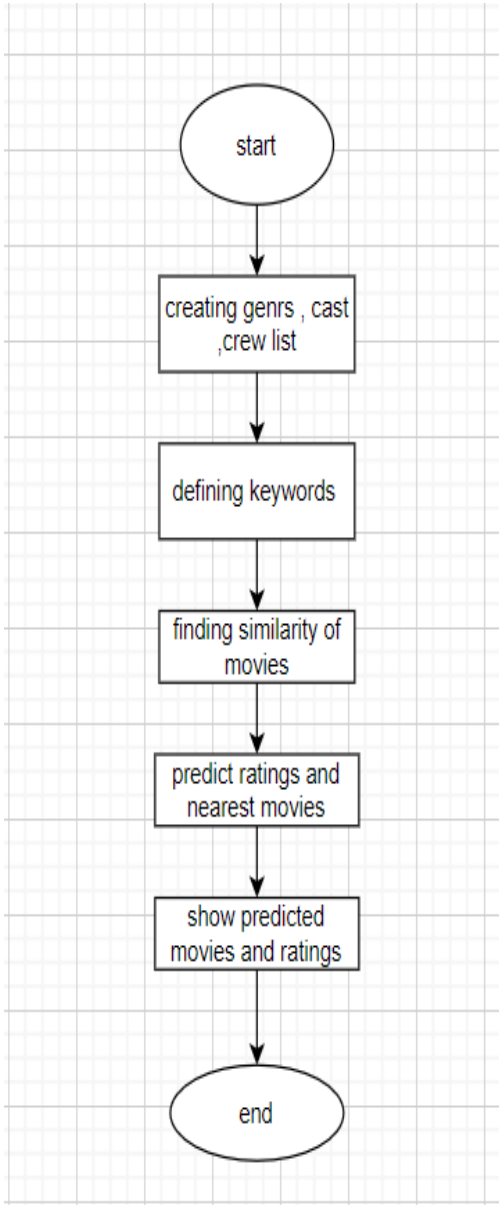

start
creating genrs , cast ,crew list
defining keywords
finding similarity of movies
predict ratings and nearest movies
show predicted movies and ratings
end

Figure 7: KNN

4 Result Analysis

4.0.1 Result of Content-based Analyses:

```
: get_recommendations('Avatar',10)

: 3604                    Apollo 18
  2130                  The American
  634                     The Matrix
  1341           The Inhabited Island
  529                Tears of the Sun
  1610                         Hanna
  311     The Adventures of Pluto Nash
  847                       Semi-Pro
  775                       Supernova
  Name: title_y, dtype: object
```

Figure 8: Content-based

```
get_recommendations('Tangled',10)

2309                      Out of Inferno
39                          TRON: Legacy
330     The Lord of the Rings: The Two Towers
4714              An American in Hollywood
1470                                Stolen
1484                    Snakes on a Plane
256                              Allegiant
1984             The Thief and the Cobbler
986                         Your Highness
```

Figure 9: Content-based

4.0.2 Result of Collaborative filtering:

Amazing Spider-Man, The (2012)	3.233134
Mission: Impossible III (2006)	2.874798
2 Fast 2 Furious (Fast and the Furious 2, The) (2003)	2.701477
Over the Hedge (2006)	2.229721
Crank (2006)	2.176259
Mission: Impossible - Ghost Protocol (2011)	2.159666
Hancock (2008)	2.156098
The Amazing Spider-Man 2 (2014)	2.153677
Hellboy (2004)	2.137518
Snakes on a Plane (2006)	2.137396
Jumper (2008)	2.129716
Chronicles of Riddick, The (2004)	2.121689
Tron: Legacy (2010)	2.111843
Fantastic Four (2005)	2.083022
X-Men: The Last Stand (2006)	2.077530
Wreck-It Ralph (2012)	2.067907
Kung Fu Hustle (Gong fu) (2004)	2.067457
Godzilla (2014)	2.061653
Incredible Hulk, The (2008)	2.050104
Quantum of Solace (2008)	2.016189
Captain America: The First Avenger (2011)	2.008849
World War Z (2013)	2.005536
Thor (2011)	2.002272

Figure 10: Collaborative filtering

4.0.3 Result of Demographic filtering

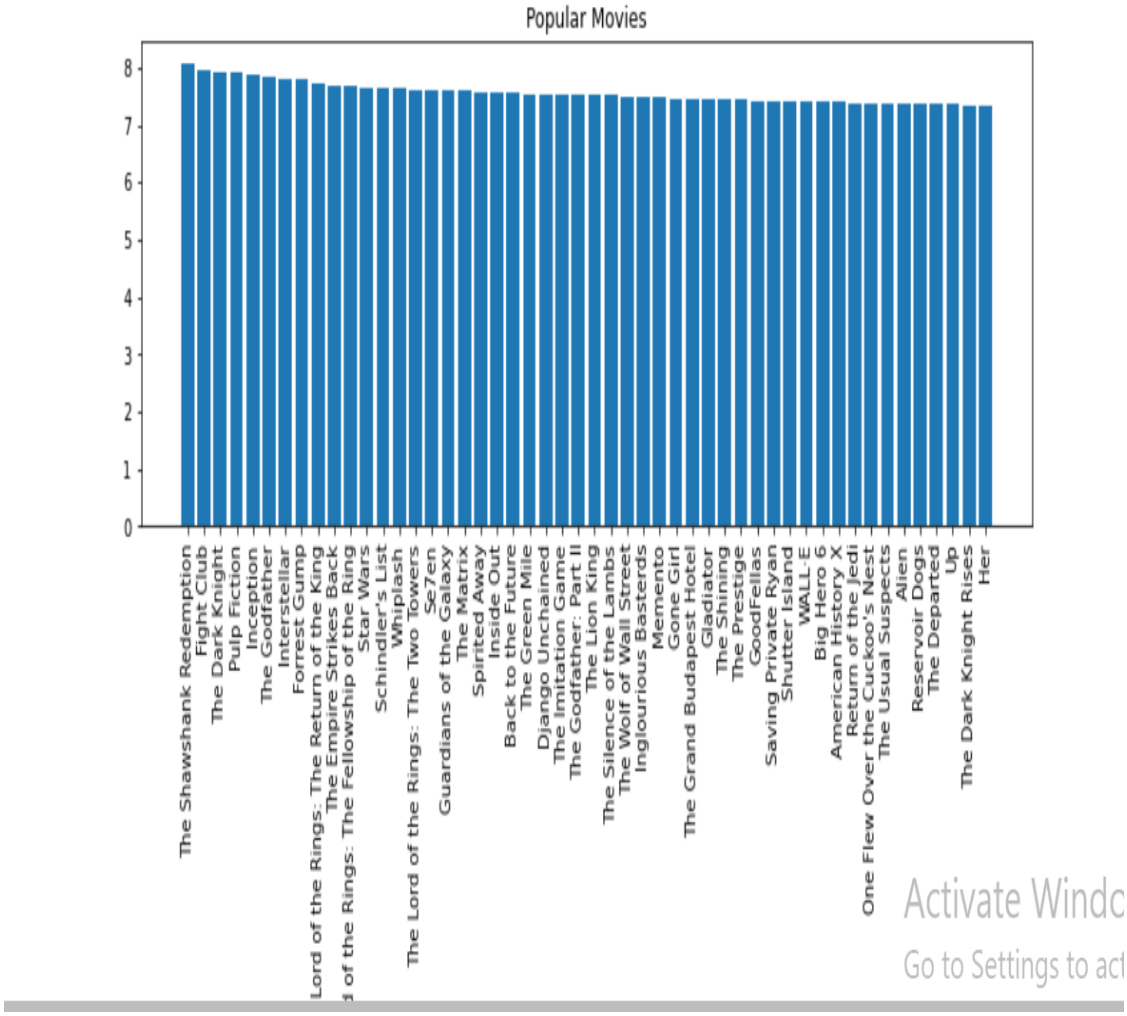


Figure 11: Demographic filtering

4.0.4 Result of K-Nearest Neighbor:

```
predict_score('Donnie Darko')
```

```
Selected Movie:  Donnie Darko

Recommended Movies:

Southland Tales | Genres: 'Action','Adventure','Comedy','Drama','ScienceFiction','Thriller' | Rating: 5.2
The Box | Genres: 'ScienceFiction','Thriller' | Rating: 5.4
Ghost | Genres: 'Drama','Fantasy','Mystery','Romance','Thriller' | Rating: 6.9
Meet Joe Black | Genres: 'Drama','Fantasy','Mystery' | Rating: 6.9
Lady in the Water | Genres: 'Drama','Fantasy','Mystery','Thriller' | Rating: 5.3
The Deep End of the Ocean | Genres: 'Drama','Mystery' | Rating: 5.9
The Lovely Bones | Genres: 'Drama','Fantasy' | Rating: 6.6
Eyes Wide Shut | Genres: 'Drama','Mystery' | Rating: 7.1
The Legend of Bagger Vance | Genres: 'Drama','Fantasy' | Rating: 6.3
Heavenly Creatures | Genres: 'Drama','Fantasy' | Rating: 7.0


The predicted rating for Donnie Darko is: 6.260000
The actual rating for Donnie Darko is 7.700000
```

Figure 12: K-Nearest Neighbor

```
predict_score('Avatar')
```

```
Selected Movie:  Avatar

Recommended Movies:

The Abyss | Genres: 'Action','Adventure','ScienceFiction','Thriller' | Rating: 7.1
The Terminator | Genres: 'Action','ScienceFiction','Thriller' | Rating: 7.3
Terminator 2: Judgment Day | Genres: 'Action','ScienceFiction','Thriller' | Rating: 7.7
Aliens | Genres: 'Action','Horror','ScienceFiction','Thriller' | Rating: 7.7
True Lies | Genres: 'Action','Thriller' | Rating: 6.8
The Wolverine | Genres: 'Action','Adventure','Fantasy','ScienceFiction' | Rating: 6.3
Titanic | Genres: 'Drama','Romance','Thriller' | Rating: 7.5
Superman Returns | Genres: 'Action','Adventure','Fantasy','ScienceFiction' | Rating: 5.4
Man of Steel | Genres: 'Action','Adventure','Fantasy','ScienceFiction' | Rating: 6.5
X-Men: Days of Future Past | Genres: 'Action','Adventure','Fantasy','ScienceFiction' | Rating: 7.5


The predicted rating for Avatar is: 6.980000
The actual rating for Avatar is 7.200000
```

Figure 13: K-Nearest Neighbor

5 Conclusion

All the algorithms described in this paper are compared. This comprehensive analysis depicts the strength and the weakness of each one of them in different forms of the Movie Lens data-set. The experiment performed is the witness the scarcity taking care of by these algorithms.

6 Future Works

With this paper, we have accomplished encouraging results from all these algorithms. In real-time sophisticated recommendation systems, there is a need for high accuracy. Such

systems still have space for improvement. There are several machine-learning algorithms that can be applied to these real-time systems. It is beneficial to look at those other calculations to progress the accuracy further.

7 References

[1] “A Simple Introduction to Collaborative Filtering,” Built In. https://builtin.com/data-science/collaborative-filtering-recommender-system (accessed Sep. 13, 2021).[2] L. Sheugh and S. H. Alizadeh, “A note on pearson correlation coefficient as a metric of similarity in recommender system,” in 2015 AI Robotics (IRANOPEN), Apr. 2015, pp. 1–6. doi: 10.1109/RIOS.2015.7270736.[3] I. Ryngksai and L. Chameikho, “Recommender Systems: Types of Filtering Techniques,” Int. J. Eng. Res., vol. 3, no. 11, p. 4.[4] “Cosine Similarity - an overview | ScienceDirect Topics.” https://www.sciencedirect.com/topics/computer-science/cosine-similarity (accessed Sep. 13, 2021).[5] “TMDB 5000 Movie Dataset.” https://kaggle.com/tmdb/tmdb-movie-metadata (accessed Sep. 13, 2021).

September 13, 2021