Recommendation Systems using Netflix Movies and TV shows Dataset

Dhvani Shah
CSE
PES University
Bangalore, India
dhvani.pranav.shah@gmail.com

Geethika Kommineni
CSE
PES University
Bangalore, India
kommineni.geethika@gmail.com

Manali Tanna
CSE
PES University
Bangalore, India
manalitanna29@gmail.com

Milinda KN
CSE
PES University
Bangalore, India
milimilindakn@gmail.com

Abstract **-** Providing useful suggestions and recommendations to their customers to increase their consumption of products is the goal of many companies nowadays. Some of the best recommendations are achieved by observing user behavior online and leveraging user data from all over the internet. In recent times this approach has been viewed negatively by the public and the protection of users' has become a priority. To build an equally powerful recommendation system without leveraging user data and behavior is the goal of this project. In this paper, a recommendation system using two approaches of contest-based modeling will be built and analyzed. The dataset used here is that of "Netflix TV shows and movies" consisting of records of over 6000 movies and 3000 TV Shows. Both approaches have been introduced in this paper. While one model explores various features like director, country of origin, cast, and genre the other leverages keywords from the description provided by Netflix to recommend a TV Show/Movie to a user. The assumption is that the input to the model is a TV show/movie that the user has already watched and likes. We utilize concepts of cosine similarity that best fit the algorithms and the data to implement the models and compare them. We see that recommendations given based on features like director, country of origin, cast, and genre are more fruitful to a user.

Keywords - Recommendation Systems, Content-Based Filtering, Cosine Similarity, Netflix TV Shows and Movies, User Data Privacy

I. Introduction

Recommendation algorithms are at the heart of Netflix products. They provide the users with personalized suggestions, reducing the time and frustration of finding great content to watch. A recommender system is an algorithm that aims to suggest related items to users (movies to watch, texts to read, products to buy, etc., depending on the industry).

Why is it important? They are very important in some industries because they generate huge revenues if they are efficient or clearly stand out from the competition. Recommender systems process large amounts of existing information by filtering the most important information based on the data provided by the user and other factors that take into account the user's preferences and interests. In recommendation systems, we examine the correspondence between users and articles, and infer the similarities between users and articles and recommend them.

Why are content-based algorithms important? Content-based filtering is a machine learning technique that uses function similarity to make decisions. This technique is widely used in recommender systems, which are algorithms designed to promote or recommend users based on the knowledge they have gathered. The model does not need data about other users because the recommendations are specific to that user. This makes it easy to scale up to a large number of users. This model can capture a user's specific interests and recommend niche items that other users have little interest in.

What is the specific problem you seek to solve? The main purpose of the recommendation systems is to provide the users with the most accurate results of what the users may want to watch. Considering the current data security threats, our purpose is to provide the users with an efficient recommendation system without breaching the user's privacy and information. The recommendation system comes up with the top 5 recommendations for the user. In this paper, we explored and compared different types of content-based algorithms. As there are already so many recommendation systems in the market, our model also helps in comparing the content-based algorithms and choosing the best fit.

What is the problem area? The main drawback of content-based filtering is that the feature representations should be known and understood as a prerequisite. This technique requires a lot of domain knowledge and the

model can only make recommendations based on the user's existing or current interests only. In other words, this model has limited capabilities in understanding the user's existing interests.

II. Related Work

Traditional content-based filtering recommendation algorithm. Content-based filtering makes recommendations by using keywords and attributes assigned to objects in the database and then comparing them to user profiles. User profiles are created based on data derived from user actions such as click-on purchases, reviews (likes and dislikes), downloads, items searched on the website, items in your shopping cart, and product links. Content-based filtering is based on assigning attributes to database objects, so the algorithm knows something about each object. These attributes largely depend on the recommended product, service, or content. Content-based filtering can be used because it does not require data from other users to make recommendations. In contrast to collaborative filtering, content-based filtering does not require data from other users to make recommendations. After a user has searched and browsed some items and later made some purchases, the content-based filtering system now can start to make a few recommendations based on the user's activity. The recommendations are very relevant to the user. Because the process relies on matching the characteristics or attributes of database objects with the user's profile, content-based recommendations can be highly tailored to the user's interests, including recommendations for niche articles. Recommendations are transparent to the user. Relevant recommendations give users a sense of openness and increase their credibility with the recommendations provided. By comparison, collaborative filtering often makes it difficult to understand why users are displaying specific recommendations. Collaborative filtering creates potential cold-start scenarios when a new website or community has few new users and no user connections.

A brief review of only the most relevant predecessor work. Users can use recommender systems to deal with information overload and identify objects that are relevant to them. The existing CBF method is over-specialized because it lacks the appropriate data to determine article similarity. In terms of accuracy and resilience, the system outperformed traditional methods. In contrast, CBF-MN improves system performance by allowing users to recommend a variety of items based on network analysis. CBF – MN solves the problem of over-specialization because many attributes are used as criteria to characterize an item. Ultimately, recommender systems are highly desirable to recommend a variety of items that take into account different criteria, rather than items that are too similar. In addition, CBF-MN performs network analysis that examines the relationships between all items and examines the structural and indirect relationships between them. CBF-MN improves system performance by allowing users to recommend a variety of products based on network analysis using MovieLens data. More text features in future research to solve the problem of over-specialization of recommender systems.

What limitations have you identified that you seek to address in your work? Data breaches involve a large number of parties (joint users, service providers, Or outsiders) and in some cases intentional actions (sniffing, hacking) or monitoring (mismanagement, remaining data). Depending on the confidentiality of the relevant information, incidents can have serious consequences. *Data Collection:* Many users are unaware of the amount and extent of information. That service providers can collect and what they can derive from this information. *Data Possession:* According to service providers, it is often difficult to delete online information You can even knowingly prevent or even prevent data deletion. That's it: The commercial value of user information. Analysis and/or data sales. Also, information that appears to have been deleted can place it from one location to another in your system. *Data Trading:* The abundant information stored in the online system is valuable to third parties and in some cases sold. user review, All tastes, and purchase history are potentially interesting to marketing Purposes. As a rule, data sales go against your expectations of data protection. Data is often anonymized before it is sold to protect the privacy of its users, but re-identification is a threat that is often overlooked or ignored. So our goal is to find a solution to build a recommendation model without using user behavior or historical data.

What are the assumptions you have made about the data? The assumptions that we`ve made about the data are that our dataset is up to date and latest and that it is also correct and reliable. One of the main features is that the user`s data is not present. Our model also behaves in a way where it asks the user what they **currently** like watching and no history of the user`s data is recorded.

III. Proposed Solution

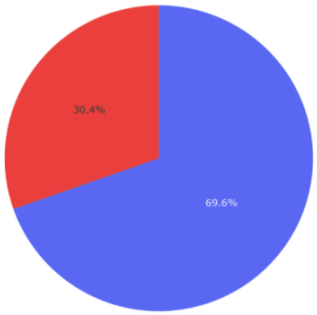
Our aim is to find a solution to build a recommendation model without using the user’s behavior and past data. To achieve this we propose the method of content-based filtering. Content-based filtering uses features to recommend other items similar to what the user likes, based on their previous actions or explicit feedback. In our case, We will not analyze the user’s previous actions and instead explicitly ask the user what items they liked and take that as our input.

3.1 Dataset

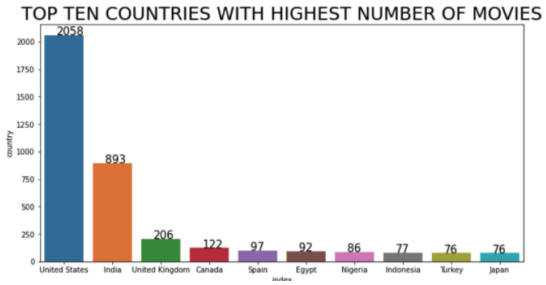
The dataset we used has been taken from Kaggle under the CC0: Public Domain License. The “Netflix Movies and TV Shows” dataset is regularly updated with records of all Movies and TV shows on Netflix. The dataset is approx 3.4 MB in size with 8807 records and 12 features. Some features include Show-ID, type, title, director, cast, country, date added, rating, release year, duration, and the genre.

3.2 Preprocessing and EDA

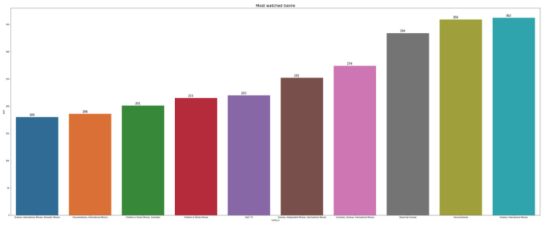
Exploratory data analysis was performed to understand user behavior in order to properly recommend movies and TV shows. During preprocessing, we detected the existing NULL values present and removed all the corresponding records for the same in order to avoid any discrepancies. Post this, our dataset was ready to use. Our EDA suggests that most of the content in Netflix is movies, of which, most of it is generated from the United States. The top genres are Drama and International Movies. Also, the majority of viewers use Netflix to watch Movies or TV-Shows on Fridays.



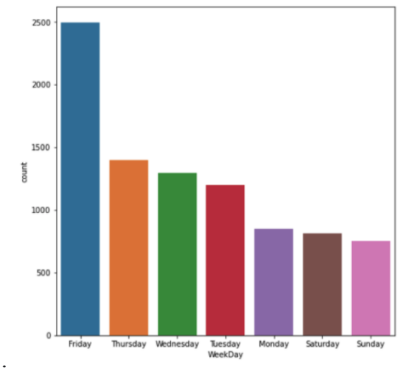
69.6% content is Movies and 30.4% Content is TV-Shows



The United States has the highest contribution to movies.



Drama and International Movies are the most-watched Genres

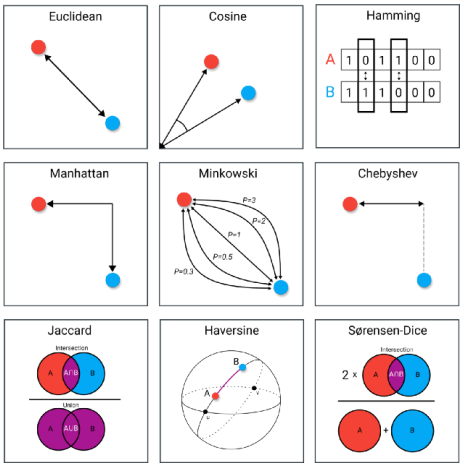


Most of the users use Netflix to watch TV-Show or Movies on Fridays.

3.3 Approaches

To build a recommendation model using content-based filtering we have used two approaches. The first approach leverages various features like the director of the movie/TV show, the country of origin of the movie/TV show, the cast acting in the movie/TV Show, and the genre that it is listed in. The second approach leverages some keywords from the description of the movie/TV show provided by Netflix itself to recommend a movie/TV Show to the user. As stated earlier, the main aim of this paper is to only use the current data provided by the user and not the past history or exploit the user's data in any way. Input for both the approaches will be a single movie/TV show that is currently liked by the user and according to that, we will recommend five other items that the user might want to watch next.

In order to do the recommendation, content-based filtering makes use of similarity measures between two items such that the recommended item is one with the Highest similarity measure. This similarity measure basically tells us how alike the two given items are. In scientific terms, it can be thought of as a dimensional distance of the features of the instance. The smaller the distance, the higher the degree of similarity. There are multiple measures of similarity as shown in the diagram. For our purpose, we will be using cosine similarity as it is good for measuring similarity between two vectors even if there is duplicate data.



Cosine similarity is a way to measure the distinction between two non-zero dimensional vectors of an inner product space. This measures the similarity between the two given vectors which is an estimation of how similar the two features are.

Similarity = cos(theta) = a.b / ||a|| ||b||

a.b => Is the product of two vectors
||a|| ||b|| =>Is the product of the vector's magnitude

3.3.1 First Model Approach

There are five features of the dataset that we will make use of in the approach director, country, cast, and listed in which is nothing but the genre. The Input is a single movie or TV show that the user currently likes and the output is a list of five other movies or TV shows that he can watch next based on these features that we've considered. The idea is to convert the data of these into binary data frames such that these combined values of the features can be compared for each movie. We can then compute a similarity between them and select the top five similarities as our result. The similarity measure that we made use of is cosine similarity which was explained previously.

3.3.2 Second Model Approach

Netflix and countless other product-based companies also give content descriptions. In the case of Netflix, this content is usually the summary of the movie or the tv show. These Netflix descriptions or synopses are concise and usually something that captures the eye of the audience. In some cases, this is taken positively by the public and in other cases, fans are offended as these descriptions are not apt for the said movie/TV show. Our second approach takes advantage of these descriptions.

We extract all the words from the description and convert them into binary data frames for easier comparison. This can be easily achieved using natural language processing libraries such as Punkt from nltk package. This is done for each record. We then compare these and find the movies that have the most words in common, this again is done using a similarity measure called cosine similarity.

Our proposed solution makes use of python3 and related libraries to preprocess our dataset, visualize it and build the recommendation model.

IV. Experimental Results

After implementing the proposed methods of building the recommendation models, we tested both models using the same inputs. The results of which can be seen below in figures 4.1 and 4.2 below.

For both models, the input is the Bollywood movie "Kal Ho Na Ho".

Using the first model [refer to figure 4.1] i.e recommendation based on features like director, country, cast, and genre we get five other Bollywood movies originated from India with very similar genres. This is a satisfactory result based on our personal experience and public opinion.

Using the second model [refer to figure 4.2] i.e recommendation based on the keywords in the description we get five movies, some of which are Bollywood and some are Hollywood. This result is not satisfactory according to our personal experience and public opinion. The reason behind this is that the descriptions may have similar keywords but not be related at all. as seen in this example.

V. Conclusion

Data breaches involve a large number of parties and in some cases intentional actions (sniffing, hacking) or monitoring (mismanagement, remaining data). Depending on the confidentiality of the relevant information, incidents can have serious consequences. We executed or modeled using content-based filtering using functional similarities to make decisions. The model works by asking the user what they are looking at and not recording a history of user data. Our recommendation model was made using features of the dataset such as show_id, type, title, director, cast, country, date_added, release_year, rating, duration, listed_in, and description.

```
[1] recommender_model_1('Kal Ho Naa Ho')
```

	title	director	cast	country	rating	listed_in	cos_sim
1057	Soldier	Abbas Mustan	Rakhee Gulzar, Bobby Deol, Preity Zinta	India	TV-14	Comedies, Dramas, International Movies	0.553399
3619	Chal Dhar Pakad	Aatmaram Dharne	Nagesh Bhonsle	India	TV-14	Comedies, Dramas, International Movies	0.530330
4079	Irada Pakka	Kedar Shinde	Smita Jaykar, Siddarth Jadhav	India	TV-14	Comedies, Dramas, International Movies	0.527645
4184	Kya Kehna	Kundan Shah	Preity Zinta, Saif Ali Khan, Anupam Kher, Fari...	India	TV-PG	Dramas, International Movies, Romantic Movies	0.505181
718	AK vs AK	Vikramaditya Motwane	Anil Kapoor, Anurag Kashyap	India	TV-MA	Comedies, Dramas, International Movies	0.500000

Figure 4.1

```
recommender2('Kal Ho Naa Ho')
```

	title	description	description_filtered	cos_sim
161	Team America: World Police	In this musical satire, an all-marionette police force takes on the challenging role of keeping the peace on a troubled planet.	in musical satire , all-marionette police force takes challenging role keeping peace troubled planet .	0.410993
577	Mariposa	New student Acha falls for Iqbal, a high-achieving student who's torn between love and family pressure.	new student acha falls iqbal , high-achieving student 's torn love family pressure .	0.410242
202	Midnight Sun	Born with a fatal sensitivity to sunlight, a sheltered teen girl falls for her neighbor, but hides her condition from him as their romance blossoms.	born fatal sensitivity sunlight , sheltered teen girl falls neighbor , hides condition romance blossoms .	0.407223
1157	Chashme Baddoor	When pretty new neighbor Seema falls for their shy roommate Sid, jealous womanizers Omi and Jai plot to break up the new lovebirds.	when pretty new neighbor seema falls shy roommate sid , jealous womanizers omi jai plot break new lovebirds .	0.403757
3629	Chashme Buddoor	When pretty new neighbor Seema falls for their shy roommate Sid, jealous womanizers Omi and Jai plot to break up the new lovebirds.	when pretty new neighbor seema falls shy roommate sid , jealous womanizers omi jai plot break new lovebirds .	0.403757

Figure 4.2

VI. References

[1] mygreatlearning.com/blog/matrixfactorizationexplained/#contentbasedfiltering
[2] upwork.com/resources/what-is-content-based-filtering
[3] Jieun Son, Seoung Bum Kim: Content-Based Filtering for Recommendation Systems Using Multiattribute Networks - 2017
[4] towardsdatascience.com/9-distance-measures-in-data-science-918109d069fa
[5] Xavier Amatriain: Big & Personal: data and models behind Netflixrecommendations - 2013
[6] analyticsvidhya.com/blog/2015/08/beginners-guide-learn-content-based-recommender-systems/

VII. Individual Contribution

Name	SRN	Contribution
Dhvani P Shah	PES2UG19CS114	EDA(One query), Final Report, Model 1, Literature Survey
Geethika Kommineni	PES2UG19CS127	EDA(One query), Final Report, Model 1, Literature Survey
Manali Tanna	PES2UG19CS214	EDA(One query), Final Report, Model 2, Literature Survey
Milinda KN	PES2UG19CS233	EDA(One query), Final Report, Model 2, Literature Survey