EasyChair Preprint

№ 10290

Recommendation Engine for Netflix

S Sureshkumar, Harsha Vinjamuri, Vinay Gollgopu,
Suryaprakash Gurakala and Tharun Padigunta

May 30, 2023

RECOMMENDATION ENGINE FOR NETFLIX

SureshKumar1* Vinjamuri Harsha1 Gollagopu Vinay2 Gurakala Suryaprakash2 Padigunta Tharun2
Department of Computer Science and Engineering
Kalasalingam Academy of Research & Education
KrishsnanKovil, TamilNadu

Abstract: *A Netflix recommendation engine is a machine learning-based system that recommends movies and TV programs based on a user's prior viewing history and preferences. This system helps users obtain content they might not have found otherwise, while also improving the user experience and increasing platform engagement. The recommendation engine uses a combination of collaborative filtering and content-based filtering methods to generate personalized recommendations. Collaborative filtering involves analyzing a user's behavior and identifying patterns of similarity between users, whereas content-based filtering involves analyzing the content's characteristics, such as genre, actors, and storyline, to recommend comparable content.*

Keywords: *Netflix , content-based filtering, machine learning, recommendation engine, Collaborative filtering.*

1 INTRODUCTION

Netflix is the world's biggest on-demand internet streaming media and online DVD movie renting service supplier. Marc and Reed founded the company on August 29, 1997, in Los Gatos, California. It has 69 million users in over 60 nations who watch over 100 million hours of TV programs and movies per day. Netflix is the world's top online entertainment provider, offering TV shows, movies, and feature films in a broad range of categories and languages. I was intrigued to evaluate the material published on the Netflix platform, which led me to create these simple, interactive, and exciting visualizations and discover comparable groups of people.

We did an exploratory analysis of data from Netflix movies and TV shows. The data collection is derived from Flixable, a Netflix third-party search engine. It would be interesting to look into all of the other information that can be obtained from the same collection of data. Initially, the examined written material includes only of the title and summary of the work. However, we intend to enhance our method by taking into account other determining factors, in this instance, the demographics of the users. Our system's unique feature is its ease and the small amount of data required for training, which means it can be applied and used quickly.

The recommendation engine analyzes user data such as watching history, ratings, and search browses to identify patterns and forecast which content the user is likely to love and enjoy. To generate recommendations, this algorithm uses a combination of collaborative filtering and content-based filtering techniques.
Collaborative filtering is the process of analyzing user behavior and preferences in order to find comparable users with similar hobbies and watching patterns. After identifying comparable users, the algorithm suggests content that the similar users have viewed and liked but that the user has not yet seen. In contrast, content-based filtering entails evaluating the traits of the content itself in order to suggest comparable content to users. The remainder of this paper will cover related works, the methods used for the study, the analysis of the findings, and finally the debate and conclusion.

2 METHODOLOGY

The goal is to perform an exploratory data analysis to comprehend Netflix's current trend in the types of programs provided. (EDA). Second, we established an emission suggestion system based on two NLP-derived methods, TF-IDF and Cosine Similarity.

In natural language processing (NLP) and data pre-processing We deleted punctuation and stop words because they are unnecessary parts, and then we counted vocabulary items using countvectorizer, which creates a lexicon of the most frequently used and essential terms. Then we use stemming to normalize the text before transforming it into a meaningful depiction of numbers that can be used to train a machine learning method for prediction.

Tf-IDF vectorization is used in this case. We obtained seref,orai,lukasz from this, and these terms are top in vocabulary present in description, and top vocabulary present in listed_in are tv,thriller,teen,talk, and so on.We obtained the finest cluster arrangements by using various clustering algorithms such as Kmeans , hierarchical clustering, Agglomerative clustering, and DBSCAN on data. We find that the optimum amount of clusters is three. Then, after developing a suggestion system based on description and genre, we obtained the best pictures based on our recent interests.

2.1 DATA COLLECTION

We gather user information such as browsing history, interests, and profiles. and also gathered details about Netflix such as type, title, director, cast, nation, and so on. Flixable, a search engine that lists content accessible on Netflix, provided the data for this research. The dataset contains records in 12 columns, including the title of the show, the category to which it belongs (film or TV show), the names of the director and main actors, the countries where the show is available for viewing on Netflix, the Netflix rating, the duration of the show, and its description.

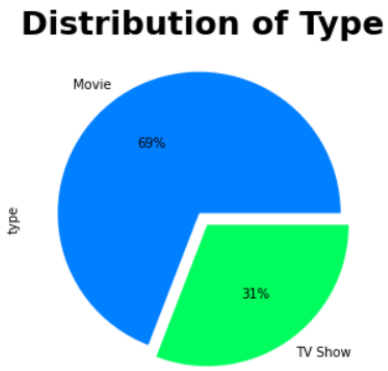

Fig 2.1 Distribution of data type

2.2 DATA PRE-PROCESSING

In text analysis (NLP) and data pre-processing, I removed punctuation and stop words because they are useless parts, and then we measure vocabulary items with count vectorizer(), which creates a dictionary of the most frequently used essential terms.

We use stemming to standardize the content here. We turn text into a meaningful depiction of numbers that is used to train a machine learning method for prediction. We use Tf-IDF vectorization because it includes information on both the more essential and less important terms. Seref, Orai, and Lukasz Top vocabulary present in description and top vocabulary present in listed_in are tv, thriller, teen, talk, and so on.

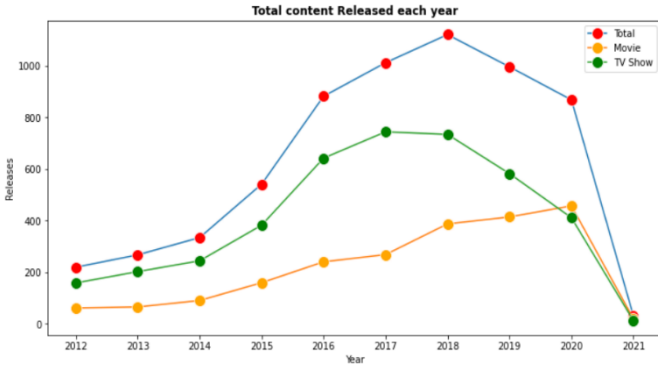

Fig 2.2 data processing

2.3 COLLABRATIVE FILTERING

Analyzing historical data on user actions and using it to forecast what they might like based on their similarity to other users. The fact that CFs only examine interactions between people rather than items and users is a significant benefit of this method. As a result, they do not need to understand the real context of the data in order to make suggestions. That is, they can make a forecast without "understanding" a film, a buddy, or a piece of music, for example. As a result, regardless of the data's content, this method can be used widely. However, if users do not evaluate an adequate number of products, these methods may underperform. Collaborative Filtering may also have scaling or sparsity problems. Collaborative

2.4 CONTENT – BASED FILTERING

This component uses the information about the content, such as genre, actor, and director, to recommend content that is similar to what the user has watched in the past. based on the user's prior activities or explicit feedback, recommends other products comparable to what they like. To make suggestions, it looks for parallels in goods, services, or content characteristics, as well as information gathered about the individual. This enables scaling to a big number of users easier. The model can capture a user's particular preferences and suggest niche products that few other users are interested in.

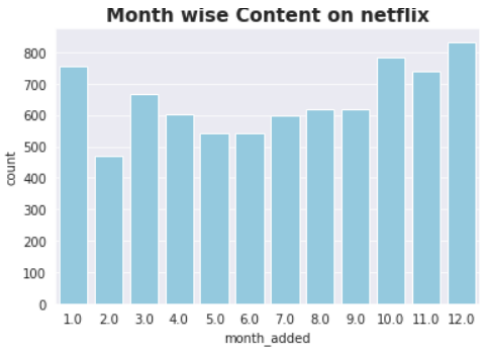

Fig 2.4 month wise content

3 RESULTS

3.1 SILHOUETTE SCORE

The Silhouette Score is computed for each sample using the mean intra-cluster distance (a) and the mean nearest-cluster distance (b). A sample's Silhouette Score is (b - a) / max(a, b). To be more specific, b is the distance between a sample and the closest cluster in which the sample does not belong.

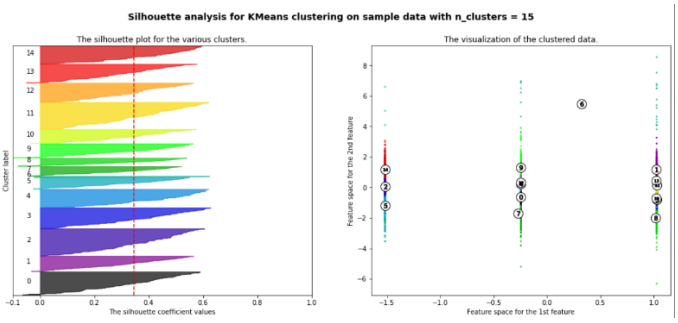

Fig 3.1 silhoutte score

3.2 ELBOW METHOD

It is a heuristic for calculating the number of groups in a data collection. The technique entails showing the explained variance as a function of cluster count and selecting the curve's elbow as the number of clusters to use.

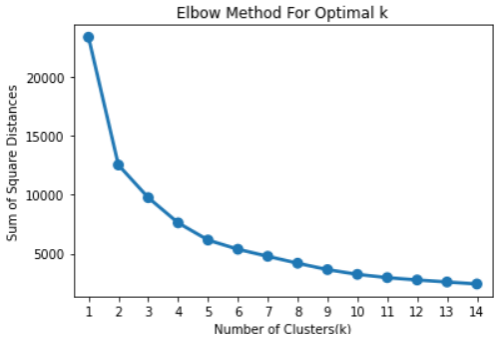

Fig 3.2 elbow method

3.3 DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is the foundation method for density-based clustering. It can find clusters of various forms and sizes in a significant quantity of data that contains noise and outliers.

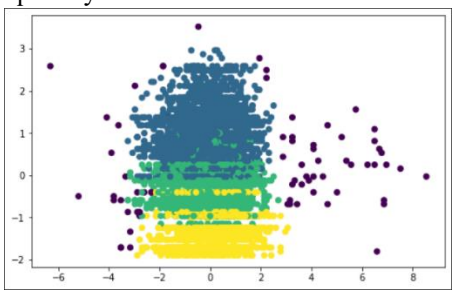

Fig 3.3 dbscan

3.4 AGGLOMERATIVE CLUSTERING

A form of hierarchical clustering method is agglomerative clustering. It is an autonomous machine learning method that separates the population into several clusters, with data points in the same cluster being more similar and data points in various clusters being distinct.

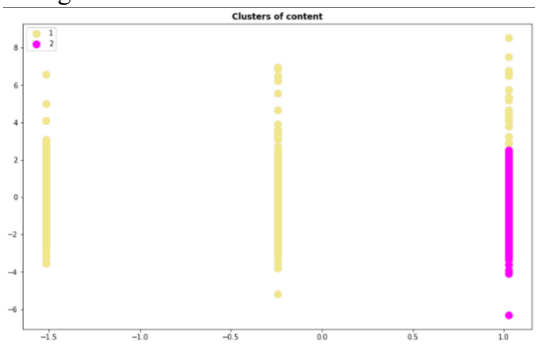

Fig 3.4 Representation of agglomerative clustering

3.5 DENDOGRAM

A dendrogram is a picture that depicts a tree. This diagrammatic depiction is commonly used in a variety of settings, including hierarchical clustering, where it depicts the order of the groups generated by the related analyses.A dendrogram can be used to depict the connections between any type of entity as long as their similarity to each other can be measured. Lexomic analysis compares the distribution of various terms across entire texts or parts of texts.

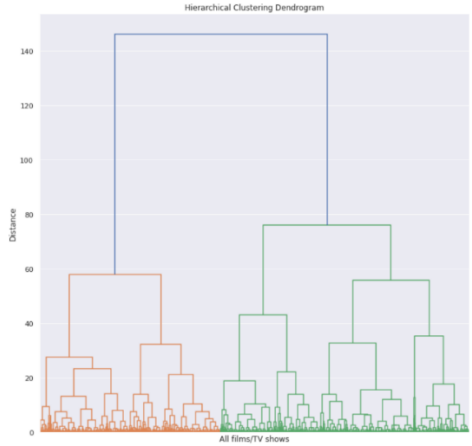

Fig 3.5 dendogram

4 RECOMMENDATION METHOD BASED ON TF-IDF AND COSINE SIMILARITY

Netflix offers, from the platform's home page, the classification of its movies and TV shows by genre. In fact, Netflix values content according to the movies viewed by users. However, the Netflix service has a system of personalized recommendations to propose to the subscriber the programs likely to interest him. Netflix's recommendation system is based on several criteria such as the customer's interaction with the service, the choice of other users whose tastes are deemed similar to those of the customer in question, the metadata specific to the programs, the time of day the user connects to the platform, the duration of viewing time, etc

5 CONCLUSION

We received the best pictures based on our recent interests after developing a recommendation system based on description and genre. For example, the first two hits for "Indiana Jones and the Last Crusade" are strongly suggested, and they are also Indiana Jones series movies.It will undoubtedly improve watch time, content retention, and have a positive effect on Netflix's company by increasing profitability and business possibilities.

6 REFERENCES

[1] An Integrated PCA-DAEGCN Model for Movie Recommendation in the Social Internet of Things.

[2] Movie Recommendation System using bag of words and scikit-learn.

[3] Netflix Recommendation System based on TF-IDF and Cosine Similarity Algorithms.

[4] Building a content-based recommendation engine model using Adamic Adar Measure; A Netflix case study.

[5] Hybrid Recommendation Algorithm for Personalization of Customer Experience.

[6] Recommendation System for Netflix.