Netflix Recommendation System based on TF-IDF and Cosine Similarity Algorithms

Mohamed Chiny[1][a], Marouane Chihab[1], Omar Bencharef[2][b] and Younes Chihab[1][c]
[1]*Laboratory of Computer Sciences, Ibn Tofail University, Kenitra, Morocco*
[2]*Department of Computer Sciences, Cadi Ayyad University, Marrakesh, Morocco*

Keywords: Netflix, EDA, Recommendation, NLP, TF-IDF, Cosine similarity

Abstract: Netflix is the most popular on-demand broadcast platform available today. Its service is available in 190 countries and includes movies and TV shows. In our study, we conducted an exploratory analysis of data obtained from Flixable, which is a search engine that lists the content available on Netflix. A dataset of 7,787 unique records was analyzed to highlight essential information about the content available on this platform. We also implemented a recommendation system using the TF-IDF and Cosine similarity algorithms, which are models widely used in Natural Language Processing (NLP). The exploratory analysis has revealed interesting data on the current trends of the content delivered on Netflix. Despite the limitation of the recommendation system in its current state, it looks promising when additional features are taken into consideration.

1 INTRODUCTION

Streaming services have become increasingly important and often promising gateways to unlimited cultural and entertainment content, as is the case with Netflix (Colbjornsen, 2020).

Netflix is a company founded in 1997. It is specialized in the distribution and exhibition of film and television works through a dedicated platform. This platform offers an online subscription rental service for movies and television series. It allows subscribers to access programs for a fixed monthly fee. It lists the programs that the customer wishes to see in their "queue" as a recommendation (Bennett, 2007).

Netflix has become one of the world's leading entertainment services with 204 million paid subscriptions in more than 190 countries, making television series, documentaries and feature films available in a wide variety of genres and languages. However, 60% of the movies rented by Netflix are selected based on personalized recommendations. Indeed, a recommendation system could not only generate more direct revenue but also additional revenue by introducing users to new categories (Dias, 2008). Therefore, a referral system can have a significant impact on a company's revenues (Chiny, 2021; Dawei, 2020).

As the information industry and the Internet develop rapidly, the use of Big Data is entering people's minds and attracting attention. It gives rise to the recommendation system that allows the desired information to be quickly extracted from excessive details. In the recommendation system, the user-based collaborative filtering algorithm has become a search hotspot (Tie-min, 2020; Zhang, 2019). Therefore, it is evident that Netflix prioritizes one type of content more than others, based on user recommendations and on the viewing rate of on-demand broadcasts.

In our study, we conducted an exploratory study of data regarding television and film programs available on Netflix. The data set is collected from Flixable (Flixable, 2021), a third party search engine of Netflix. Indeed, in 2018, they published an interesting report (Netflix, 2018) which shows that the number of movies on the streaming service has decreased by more than 2000 titles since 2010, while

[a] https://orcid.org/0000-0001-6293-7606
[b] https://orcid.org/0000-0002-9458-0389
[c] https://orcid.org/0000-0003-0031-7609

Chiny, M., Chihab, M., Bencharef, O. and Chihab, Y.
Netflix Recommendation System based on TF-IDF and Cosine Similarity Algorithms.
DOI: 10.5220/0010727500003101
In *Proceedings of the 2nd International Conference on Big Data, Modelling and Machine Learning (BML 2021)*, pages 15-20
ISBN: 978-989-758-559-3

the number of TV shows on Netflix has almost tripled. It would be interesting to explore all the other information that can be obtained from the same set of data. On the other hand, we have implemented a Netflix content recommendation system based on similarity analysis using Term Frequency - Inverse Document Frequency and Cosine Similarity algorithms. Initially, the explored textual corpus consists only of the title and the description of the work. However, we plan to improve our system by considering other determining features, in this case, the demographics of the users.

The particularity of our system is its simplicity and the reduced volume of data needed for training, which means that it can be implemented and used in a short time.

In the rest of this paper, we will deal with Related works, the methodology adopted for the study, the exploration of the results and then the discussion and conclusion.

2 RELATED WORKS

Recommender systems are trained to suggest fast and relevant results to users. In the existing literature, many works have tried to produce efficient recommendation engines using Term Frequency - Inverse Document Frequency and Cosine Similarity. This is the case of the hybrid recommendation system presented to the user via a web interface. This system uses a small data model (Muthurasu et al., 2019) or the image-based recommendation system to guide users to movies and TV shows on streaming platforms (Mehta et al., 2021). Another recommendation system that improves user experience has been proposed by adopting a Machine Learning-based approach and Cosine Similarity. Popularity, genre as well as content are considered for the design of this tool (Singh et al., 2020).

3 METHODOLOGY

Our approach is to understand Netflix's current trend in the type of offered programs by conducting an exploratory data analysis (EDA). Second, we set up a system of emission recommendations based on two techniques derived from NLP, which are TF-IDF and Cosine Similarity.

3.1 Data Collection

The data for this study were obtained from Flixable (Flixable, 2021), which is a search engine that lists content available on Netflix. The dataset contains records in 12 columns, including the title of the show, the category to which it belongs (film or TV show), the name of the director and main actors, the name of the countries where the show is available for viewing on Netflix, the rating on Netflix, the duration of the show and its description.

3.2 Data Preprocessing

Before proceeding with the exploratory analysis of the data, the data were filtered to exclude missing or inconsistent information. This step resulted in 7787 unique records ready for use. Concerning the second part, which consists of implementing a recommendation system based on TF-IDF and Cosine Similarity, we proceeded with tokenization and then excluded stop words and undesirable characters such as punctuation marks.

3.3 Exploratory Data Analysis (EDA)

Exploratory data analysis was performed using the Python language through Jupyter Notebooks. A set of software libraries specialized in EDA were used, such as NumPy, Pandas, Seaborn and Matplotlib. An essential step of the analysis is to generate the word cloud by calculating their density through the Word Cloud library, which is part of an NLP software stack. For the recommendation part, we used the TF-IDF and Cosine Similarity algorithms that we applied to the titles and the descriptions of the programs available on Netflix to analyze their similarities.

3.3.1 Term Frequency – Inverse Document Frequency (TF-IDF)

The TF-IDF algorithm is used to evaluate the importance of words in a textual corpus. The importance is proportional to the number of times the words appear in the document and inversely proportional to the frequency of words appearing in the corpus (Kang, 2016; Guo, 2016; Shengqi, 2020).

TF represents the frequency of words, indicating the number of times they appear in a corpus (func 1). This consists of calculating the number of word appearances out of the total number of words present in the corpus.

$$tf_i = \frac{n_i}{\sum_k n_k} \quad (1) \qquad idf_i = \log \frac{|D|}{|\{d_j : t_i \in d_j\}|} \quad (2)$$

IDF is the measure of the importance of the term in the corpus as a whole. It consists in calculating the logarithm of the inverse of the proportion of documents in the corpus that contain the term (func 2). This consists of calculating the total number of documents contained in the corpus over the number of documents where the word is present. The logarithm of this result constitutes the value of the IDF. The weight of TF-IDF is calculated by multiplying the two measures ($tf_i * idf_i$), so the greater the weight, the more significant the word concerned in the corpus.

3.3.2 Cosine Similarity

The cosine similarity algorithm measures the similarity by using the cosine angle between two vectors A and B, which can represent words, sentences, paragraphs or even entire documents (func 3). The closer the value of the cosine to 1, the smaller the angle between the two vectors. If we apply this algorithm to textual corpora, then it can evaluate the level of similarity between documents (Yunxiang, 2020).

$$\cos\theta = \frac{A.B}{\|A\| \|B\|} \quad (3)$$

4 RESULTS

4.1 Exploratory Data Analysis

In the following, we present the results of our exploratory analysis of the data retained after the pre-processing stage.

4.1.1 Ranking of Countries According to the Number of Emissions Available

The German portal Statista published an interesting report in July 2018 about the number of films and series available on Netflix per country (Statista, 2018). The top 4 countries in this ranking are, in order, USA, Canada, UK and India. However, our results show a shift in position between India and Canada (Figure 1). Indeed, India has moved up to second place and Canada to fourth. This change in ranking has affected many countries such as Australia, which gave way to Japan and France, which took the place of Brazil...

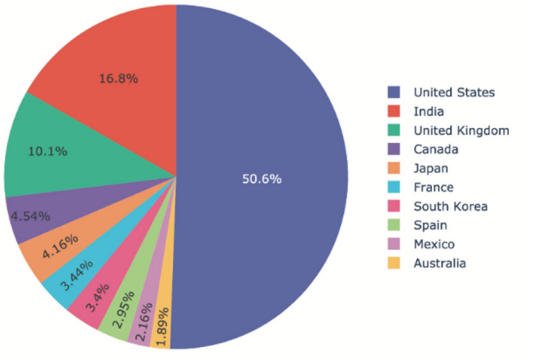

Figure 1: Top 10 countries by number of issues available on Netflix

Apart from the rankings, it can be seen that the USA monopolizes about half of the programs available on Netflix, followed by India with 16.8% and the United Kingdom with 10.1%. Therefore, more than three-quarters of all Netflix broadcasts are available in three countries, and Netflix is available in streaming in more than 190 countries (Help Center Netflix, 2021).

Africa is also missing from the Top 10 countries where Netflix content is available. Indeed, Egypt comes in 12th position with 1.6% of content, followed by Nigeria in the 20th position with a rate of 1.08%.

4.1.2 Ranking of Program Categories Proposed by Netflix

The Netflix online video platform offers content in a variety of genres. However, in our study, we were interested in the categories offered, which can be segmented into two main categories: movies and TV shows.

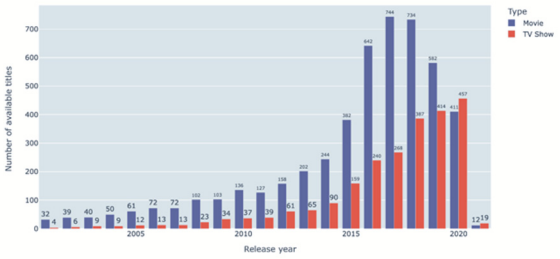

Figure 2: Change in the number of issues offered by category

We found that Movies account for 69% of the content streamed on Netflix versus 31% on TV Show. It is clear that Movies are the most popular and most available content among the platform's subscribers.

Nevertheless, if we analyze the trend of content being put online over the years in figure 2, we can see that the rate of evolution of TV shows continues to increase, while from 2018, the number of movies decreases, and for the first time, in 2020, the number of TV shows broadcast exceeds the number of movies. This observation also applies to the beginning of 2021, with 19 new TV shows broadcast compared to 12 movies.

This transition, which was identified in 2020, raises questions to understand its causes. Indeed, one of the possible causes may be due to the Covid 19 pandemic, which forced billions of people worldwide to confine themselves to their homes for varying periods depending on the country. These conditions could probably influence the habits of users concerning the monitoring of their television programs.

Although we can see that the gap between the number of movies and TV shows available on Netflix has been decreasing more and moreover the last four years, and it could well be that the transition that marked the year 2020 is only a natural continuation of the global pace that started a few years ago.

4.1.3 The Global Context of the Works Offered on Netflix

A word cloud is a kind of semantic digest of a textual corpus. It is the visual representation of the most frequent words in a document. The more visible the word is, the more frequent it is.

The word cloud is generally one of the first visualization steps that allow building a global idea about the subject matter of a textual corpus. It is a step before implementing more advanced NLP algorithms, which are likely to convey more precise information about the semantics of the document.

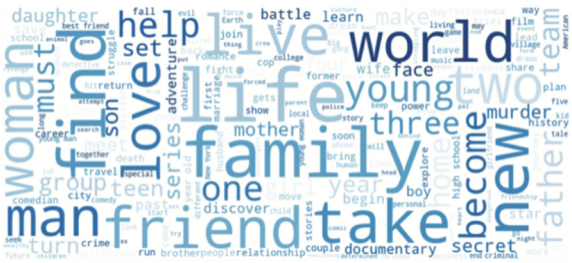

Figure 3: Cloud of most frequent words in the program description

Figure 3 shows the most commonly used words in the description of the works distributed by Netflix. Words such as "Family", "World", "Life", and "friend" are the most common, suggesting that most online programs are dealing with social issues.

However, the frequency of a word is not a good indicator of its relevance to the context. Indeed, the principle of the frequency of a word is taken up by the Bag-of-word model, which is a simple representation of documents in the form of vectors calculated according to the frequency of a word in the corpus. In the NLP domain, this model is not sufficient to have a more solid idea about the context of the document. This is why techniques such as Word Embedding (Ensaf, 2016), Word2vec (Mikolov, 2013) or TF-IDF are more used in this sense.

4.2 Recommendation System 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. However, Netflix states that its recommendation algorithm does not consider user demographics such as age and gender (Help Center Netflix, 2021).

We found it interesting to implement a recommendation system that considers other metrics such as the contextual similarity of the titles and descriptions of the proposed works. First, we applied the algorithm to all the programs available on the platform regardless of their actual availability in the geographical area where the client is located.

Table 1 illustrates the recommendation of 10 issues whose titles and descriptions are deemed to have similarities with the "NCIS" series. These are the implementation of the TF-IDF and Cosine Similarity algorithms.

This recommendation system uses as a document only the title and the description of the proposed issue on Netflix. In other words, the application of the IF-IDF and Cosine Similarity algorithms on such a small corpus will certainly not give relevant results.

Table 1: List of recommended emissions following the selection of NCIS

Title	Calculated Score
La Piloto	0.219489577073954
Twins Mission	0.185445602841902
Curon	0.165079412202837
Mugamoodi	0.161729302850993
Hashoter Hatov	0.160847112954766
Nagi-Asu: A Lull in the Sea	0.154977034987895
The Indian Detective	0.154310875273481
The Unremarkable Juanquini	0.142036012436844
Michael	0.141768479700181
Happyish	0.140348670951172

However, this approach is a starting point towards a more sophisticated recommendation system that considers other features such as the show's duration, the Netflix score, the leading players, etc. In addition, the fact that Netflix does not take into account the demographics of the subscribers led us to take this feature into account as well to be able to increase the probability of acceptance of the recommendations by the client.

5 DISCUSSION AND CONCLUSION

Netflix is the most popular on-demand broadcast platform at the moment. It broadcasts thousands of programs to subscribers in 190 countries around the world. Since its inception, movies have been the most dominant content in online programming. However, the year 2020 marks the domination of TV shows for the first time, which is also the case for the beginning of the year 2021.

In our study, we conducted an exploratory analysis of Netflix data. This analysis highlighted information such as the countries where Netflix content is most available, including the USA, India and the United Kingdom, the distribution of programs broadcast by category (69% movies and 31% TV shows), and the semantic digest of the words used in the descriptions of the works broadcast. On the other hand, and due to the increase in the number of TV shows offered at the expense of movies in 2020, we thought it would be interesting to conduct a more in-depth study on this subject to understand the causes. Indeed, among the probable causes, we could cite the COVID-19 pandemic, which has forced billions of people worldwide to confine themselves to their homes, which has probably impacted their TV viewing habits.

The second part of this work consists of implementing a system of program recommendation by applying the TF-IDF and Cosine Similarity algorithms on the titles and the descriptions of the works. However, the relevance of the results of this system can be criticized because of the limited number of occurrences present in the used corpus. Nevertheless, it can be an excellent start to understanding other features that could refine the recommendations, such as the show's duration, the score attributed on Netflix, the actors highlighted in the program, etc. In addition, we thought it would be interesting to include the demographic data of the subscribers since Netflix does not consider this factor when proposing recommendations.

Nevertheless, our work is characterized by the simplicity and the low volume of training data required to implement the recommendation system. This gives it the advantage of being easily implemented and used.

REFERENCES

Bennett, J., Lanning, S., 2007. The Netflix prize. In Proceedings of KDD cup and workshop, Vol. 2007, p. 35). New York, NY, USA.

Chiny, M., Bencharef, O., Hadi, M.-Y., Chihab, Y., 2021. A Client-Centric Evaluation System to Evaluate Guest's Satisfaction on Airbnb Using Machine Learning and NLP. Applied Computational Intelligence and Soft Computing.

Colbjornsen, T., Talleras, K., Ofsti, M., 2020. Contingent availability: a case-based approach to understanding availability in streaming services and cultural policy implications. INTERNATIONAL JOURNAL OF CULTURAL POLICY.

Dawei, W., Yuehwern. Y., Ventresca, M., 2020. Improving neighbor-based collaborative filtering by using a hybrid similarity measurement. Expert Systems with Applications.

Dias, M. B., Locher, D., Li, M., El-Deredy, W., Lisboa, P.J., 2008. The value of personalised recommender systems to e-business: a case study. In Proceedings of the 2008 ACM conference on Recommender systems (pp. 291–294). ACM.

Ensaf, H.-M., Mohammed, E.-M., Mohamed, H.-H. H., 2020. An Enhanced Sentiment Analysis Framework Based on Pre-Trained Word Embedding. International Journal of Computational Intelligence and Applications.

Flixable, 2021. Full List of Movies and TV Shows on Netflix. https://flixable.com.

Guo. A. and Yang, T., 2016. Research and improvement of feature words weight based on TF-IDF algorithm. In 2016 IEEE Information Technology, Networking,

Electronic and Automation Control Conference, pp. 415–419, Chongqing, China.

Help Center Netflix, 2021. Dans quels pays le service Netflix est-il disponible?. https://help.netflix.com.

Kang, G., Tang, M., Liu, J., Liu, X., and Cao, B., 2016. Diversifying web service recommendation results via exploring service usage history. IEEE Transactions on Services Computing, vol. 9.

Mehta, R., Gupta, S., 2021. Movie Recommendation Systems using Sentiment Analysis and Cosine Similarity. International Journal for Modern Trends in Science and Technology, 7(01): 16-22.

Mikolov, T., Yih, W.-t., Zweig, G., 2013. Linguistic regularities in continuous space word representations. In Proceedings of the 2013 conference of the north American chapter of the association for computational linguistics: human language technologies.

Muthurasu, M., Rengaraj, N., Mohan, K. C., 2019. Movie Recommendation System using Term Frequency-Inverse Document Frequency and Cosine Similarity Method. International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-7 Issue-6S3.

Netflix Museum, 2018. https://flixable.com/netflix-museum.

Shengq, W., Huaizhen. K., Chao, L., Wanli, H., Lianyong, Q., Hao, W., 2020. Service Recommendation with High Accuracy and Diversity. Wireless Communications and Mobile Computing.

Singh, R. H., Maurya, S., Tripathi, T., Narula, T., Srivastav, G., 2020. Movie Recommendation System using Cosine Similarity and KNN. International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Vol. 9(5).

Statista, 2018. Les catalogues Netflix à travers le monde. https://fr.statista.com/infographie/11504/les-catalogues-netflix-a-travers-le-monde.

Tie-min, M., Xue, W., Fu-cai, Z. , Shuang, W., 2020. Research on diversity and accuracy of the recommendation system based on multi-objective optimization. Neural Computing and Applications.

Yunxiang, L., Qi, X., Zhang, T., 2020. Research on Text Classification Method based on PTF-IDF and Cosine Similarity. Journal of Information and Communication Engineering: Volume 6 pp. 335-338 (Issue 1).

Zhang, H., Sun, Y., Zhao, M., Chow, T. W. S., Wu, Q. M. J., 2019. Bridging User Interest to Item Content for Recommender Systems: An Optimization Model. IEEE Transactions on Cybernetics, 1–13. doi:10.1109/tcyb.2019.2900159.