Reaffirming the Critical Role of Transformative Research and Knowledge Production in the Age of Post-Truth

5th DLSU Senior High School Research Congress

An Exploratory Study of OTT Platform Movie Recommendation using Cosine Similarity

Ned Jeonyl P. Arban[1], Paul Clarence B. Arce[1], Ralph Kobe R. Bernabe[1], Jin Woo C. Kim[1*], Justine Patrique A. Tillermo[1], and Katrina Ysabel C. Solomon[2]

[1] *De La Salle University Integrated School*

[2] *Advanced Research Institute for Informatics, Computing and Networking (AdRIC), De La Salle University*

**corresponding author: jin_woo_c_kim@dlsu.edu.ph*

Abstract: "Over the Top" platforms, or OTT platforms, are where movies and TV shows can be watched. The main focus of the research is the recommendation system of OTT platforms, studying its mechanisms. The researchers also aim to identify relevant features that are most useful to a recommendation system. The researchers conducted data preprocessing such as the one hot encoding method. Cosine similarity was employed as the foundational algorithm for the recommendation system. Upon generating several recommendations using different sets of features, the most relevant ones were determined through a survey. By utilizing the cosine similarity algorithm, the research aims to improve the OTT platform recommendation system. This study also seeks to gather data sets using standard pre-processing methods and identify the features that will result in the best recommendations when using the cosine similarity algorithm. The researchers compared different data sets and similarity scores based on various features. The researchers found that the data set with all gathered features had the highest level of similarity and is likely to be used in the recommendation system.

Keywords: OTT platforms; recommendation systems; data preprocessing; cosine similarity; feature selection

1. INTRODUCTION

OTT platforms, also known as "Over the Top" platforms, are used to stream movies or TV shows. The most common and popular examples of OTT platforms are Netflix, Disney+, Amazon Prime, and Hulu. These platforms all share the same system that allows them to send users the right recommendations of movies and TV shows according to their unique tastes, which is called a recommendation system.

Recommendation systems are significant for OTT platforms because they give users a tool to easily browse a movie library (Vijayanagar, 2021). For a recommendation system to be efficient, there are several important factors that OTT platform developers need to take into account (Jane, 2021). According to the related literature reviewed by the researchers, a significant number of data is required to identify the correct patterns and recommendations.

A recommendation system could be made by obtaining a dataset, importing the libraries, identifying features, combining relevant features into a single feature, extracting features, using the cosine similarity algorithm, generating a similar movie matrix, sorting the similar movie matrix in order, and printing the similar movies (Mahnoor, 2020). One of the biggest and most important things in a recommendation system is the use of cosine similarity. It is used to classify various data in order to determine the similarities of the data objects. It was also stated that cosine similarity is used in one of the most popular OTT platforms of today which is Netflix (Saleena, 2021).

Although the recommendation system has become very popular and is used in many OTT platforms, there are still some limitations found in the system. One such issue is misunderstanding the preferences of the users. Another issue is the system recommending the wrong content to users. It is essential in an OTT platform to understand the users' preferences since this will be the data used to be able to give the audience the correct recommendations of content. The researchers believe that implementing a recommendation system wherein all user preference features would be taken into account, would help its users find the most suitable movie or TV show they would like to watch next.

The researchers aim to have an exploratory study about the recommendation systems The researchers aim to build data sets with selected features an OTT platform user takes into account when choosing recommended movies, such as genre, cast, and production companies, and use the made set of features as the basis of how the recommendation would run.

2. METHODOLOGY

2.1. Data Preprocessing

The chosen data set of the researchers contained special cases where if it is directly fed to the algorithm, the code would not read the data and result in an error. This is why the researchers conducted data preprocessing to ensure that the data would be suitable for the algorithm.

One hot encoding is a method for representing binary vectors of categorical variables. Each category is given a binary value, with only one bit set to activated (1) and all other bits set to deactivated (0). Machine learning algorithms that require numerical input frequently use this encoding to represent categorical data in a format appropriate for analysis and modeling. While allowing for mathematical operations on the data, one hot encoding helps maintain the categorical information. The researchers used the one-hot encoding to the features such as the movie's cast, production company, and director where the three mentioned features were all in special cases, making the algorithm not work.

Through the process, the special cases of the categorical features would be changed to numerical by using the "0,1" language where "0" is false and "1" is true. By enabling the said language, features originally in labels would be read by the algorithm making it possible to run the code.

2.2. Cosine Similarity Algorithm

The researchers utilized the cosine similarity algorithm in Python to rank movies in the data set and create a recommendation system. Python will serve as the application for activating user machine learning and will form the foundation of the project's recommendation system.

Cosine similarity is a mathematical measure determining how similar two vectors are. It calculates similarity based on the angle between the vectors while disregarding their magnitudes. When the cosine of the angle is calculated, the resulting value varies from -1 to 1, with 1 representing identical vectors, 0 representing orthogonal or unrelated vectors, and -1 representing vectors pointing in opposing directions. Cosine similarity is commonly utilized in practical applications such as information retrieval, text analysis, and recommendation systems. It aids in rankings of documents based on their resemblance to a query in information retrieval, allowing for the efficient retrieval of relevant information. Cosine similarity is useful in text analysis for applications such as document clustering, which discovers groups of documents with similar content. It facilitates the identification of objects or products that are comparable to a user's preferences in recommendation systems, allowing for customized recommendations. The researchers implemented the cosine similarity which will be the center of the recommendation system. When searching for a movie, the cosine similarity will be the one to find the topmost similar movies to the one queried by the user.

2.3. Feature Selection

As the data set is preprocessed and the algorithm is implemented, the researchers, as stated in the introduction, used various sets of features and attempted to make the set of features describing the content users would most likely consider in choosing to watch next.

2.3.1. Content-based Feature Set

According to an article by Neerja Doshi (2018), She has done similar research with the researchers using a set of features called Content-based sets. However, it was mentioned that vote scores could also be included due to the fact the features affect the movie's content. This is why the researchers would be making two data sets wherein the first set of features would contain the director, cast, production company, run time, and genre, and the second set consists of all the features in the previous set with the addition of vote scores and vote average.

2.3.2. User Preference-based Feature Set

According to an article by Mahnoor Javed (2020), a set of features called the user preference-based set was used, which is compiled solely of what a user thinks before choosing a film to watch. "A user who likes a horror movie will most probably like another horror movie. Some users

may like seeing their favorite actors in the cast of the movie. Others may love movies directed by a particular person. Combining all of these aspects, our shortlisted features are sufficient to train our recommendation algorithm." (Javed, 2020). The user preference set contains the features cast, genre, and director.

2.4. Survey

After implementing the algorithm to generate different sets of recommendations based on the several feature sets aforementioned, the researchers conducted a survey to determine which feature set is the most preferred of the participants. This is why the survey result will be the new User Preference-Based Set, for it is considered as using human participants, making it highly viable to use the said label.

3. RESULTS AND DISCUSSION

3.1. Data Set Description

The raw data set is frequently referred to as the initial data set before preprocessing. The original dataset contains 23 features, including the following: the movie's budget, genre, homepage information, original language, original title, overview, popularity, production company, country of origin, release date, gross box office receipts, runtime, spoken languages, current status, tagline, title, vote average, vote total, and the names of its cast, crew, and its directors. Budget and revenue are examples of data types that are already numerical. However, there are more data types that are not numerical but are used in certain circumstances. To put it simply, these exceptional circumstances are the data type for terms like "cast" and "director," and they need to undergo numerical preprocessing. Yussr Elsayed Ibrahim created the dataset which was taken from Kaggle. In this online database, shareholder data may be maintained without limitation.

3.2. Data Preprocessing

The researchers preprocessed the data to be coded and utilized for cosine similarity, The researchers included additional aspects that were initially a part of the earlier data set and removed specific database components that were not in some way necessary to the coding. Considering whether the dataset is suitable for cosine similarity requires considering the many columns and other aspects. Before beginning the actual coding, the researchers had to preprocess some areas since they had polynomial properties in each column.

For the cosine similarity to function correctly, all of the data's components must be numerical for the data to be read, even though the researchers made some of the data numerical. One hot encoding transposed all the words in the horizontal columns and replaced the vertical columns with 0 and 1, where 0 denotes a no response and 1 with a response. The sample output of the one hot encoding process is seen in Figure 1.

Figure 1

Result of One Hot Encoding

	genre_Fantasy	...	director_David Hayter	director_Lance Hool	director_Tran Anh Hung	director_Mike Marvin	director_Britt Allcroft
	1	...	0	0	0	0	0
	1	...	0	0	0	0	0
	0	...	0	0	0	0	0
	0	...	0	0	0	0	0
	0	...	0	0	0	0	0

3.3. Cosine Similarity

The researchers used cosine similarity to create the recommendation system immediately after all the preprocessing was completed. A search bar was built so users could look up movies and compare them to other recommendations the group had generated with different feature sets. The top 10 most comparable movies from the movie being searched for are displayed using cosine similarity. The researchers looked into the sets of features to employ since they considered the ideal features to use while creating a recommendation system. The dataset's complete feature set is the initial collection of features to be used. Each item in the data set is assigned a similarity score by the algorithm in comparison with the selected movie title. The score ranges from 0-1, with 1 being the closest in similarity. The program then ranks then filters the top 10 recommended titles starting from the one with the highest similarity score.

As seen in Table 1, the feature set where all features are used, shows a similarity score of almost a perfect score. The results show that almost all the features were in a

match with the chosen movie with the movies in the movies in the dataset.

Table 1

Recommended Movies using All Features

Original Title	**Similarity Level**
Psycho	1.000000
Saw II	1.000000
The Godfather	1.000000
The Last Exorcism	0.999999
Tom Jones	0.999999
Paranormal Activity 3	0.999999
Roger & Me	0.999999
Fantasia	0.999998
Four Weddings and a Funeral	0.999998
One Flew Over the Cuckoo's Nest	0.999998

3.4. Establishing New Data Sets

After making new data sets with different sets of features, the researchers implemented the cosine similarity on these data sets and used the results of the cosine similarity and use it in asking the respondents whether which set of recommendations was helpful in choosing the next movie to watch.

3.4.1. Content-Based Set

It could be seen in Table 2 that the similarity score is almost near the perfect score. It is notable since the feature set where all features were used had a higher similarity score than the Content-based feature. Basing on the similarity score, the researchers determined that the set of features where all features are used could be a better set for the recommendation system.

Table 2

Recommended Movies using Content-based Set

Original Title	**Similarity Level**
A Nightmare on Elm Street 2: Freddy's Revenge	0.999830
Willard	0.999798
Final Destination	0.999794
Magnolia	0.999793
Freddy vs. Jason	0.999792
Lost Souls	0.999792
Blade	0.999791
Boogie Nights	0.999778
The Conjuring	0.999777
Jason X	0.999777

Looking at the results of the content-based feature set with the vote scores seen in Table 3, it could be said that the results show that the similarity scores of the results are near the perfect score. Comparing the similarity score of the content-based feature set with the vote scores of just the average content-based feature set, the feature set with the addition of the vote scores shows a higher similarity rate. Although the third feature set was higher than the content-based set, it could not compare to the set of features where all features were used, whereas the feature set where all features were used had a higher similarity score than that of the other feature sets.

Table 3

Recommended Movies using Content-based Set with the Vote Scores Feature

Original Title	Similarity Level
Alice Through the Looking Glass	0.999998
Die Hard 2	0.999998
Angels & Demons	0.999998
The Girl with the Dragon Tattoo	0.999998
Speed	0.999998
The Intern	0.999998
Top Gun	0.999998
Paranormal Activity	0.999998
The Witch	0.999998
Rocky	0.999998

3.4.2. User Preference-Based Set

The user-preference based set shown in Table 4 indicates that the similarity score is low in number but could also be hypothesized that the said dataset is the set of features that the majority of the people would prefer rather than the other type of datasets, and to conclude the fact, the researchers will be holding a survey to choose which set of features would be most preferred by people. Although comparing the features all in all, it could be concluded that the set of features where all features are used is the best feature for the recommendation system judging by the similarity scores.

Table 4

Recommended Movies using User Preference-Based Set

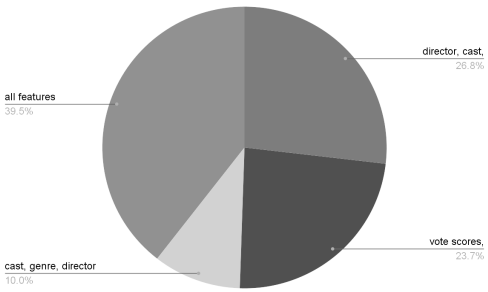
Original Title	Similarity Level
Ouija	0.408248
Maniac	0.408248
The Helpers	0.408248
The Loved Ones	0.408248
Antibirth	0.408248
The Wicked Within	0.408248
Dracula: Pages from a Virgin's Diary	0.408248
Mama	0.408248
Hellraiser	0.408248
Stitches	0.408248

3.5. Survey

The researcher conducted a survey to find out which features are most suitable for a recommendation system, and which helped the most in choosing the next movie to watch. The researchers have collected 39 participants to partake in the research. The results are as follows:

Figure 2

Survey Results



As seen in Figure 2, 39.5 percent of the respondents have chosen the set of features where all features were used in the recommendation selection process. While it can be concluded that users limit themselves to a small number of features when considering which movies to watch next, the survey indicates that the more features are used, the more information is considered when generating a recommendation, making it more detailed and may provide a more appropriate set of recommendations. The underlying information contained in features not usually considered consciously by humans may actually prove to be helpful for recommendation systems.

4. CONCLUSIONS

The researchers first compared the set of datasets to the similarity scores in the set of features. As a result, it could be seen that the set where all features were used is the set of the dataset with the highest similarity level, which could be said that the set with all features is the most likely to be used when dealing with a recommendation system. The survey also shows that the set with all features obtained the highest vote of 39.5 percent from the participants who were asked to select the sets of recommendations they would most likely use. In the two aspects of the numerical spectrum and also human preference spectrum, the set using all features stands out making it the set of features that would most likely be helpful for the users of OTT platforms to choose the next movie to watch after a certain one. As a recommendation, by using the made recommendation system, the researchers recommend building a streaming website where the recommendation system can function on the "For you" page, and also the search engine. After creating a website, it would be best if another algorithm will be implemented taking into account all the searches of the website's user and building a new data set of its own using machine learning, which could be more accurate than the set of features the researchers have made in the study.

5. ACKNOWLEDGMENTS

Sincere gratitude to pastor Jang Myung Joon, who has been the resource person of the group for giving the original dataset to complete the study and has also been the mentor for the researchers who showed directions and also explained fundamentals and mechanisms of the algorithms explored.

6. REFERENCES

Kwangil Park. "A Methodology Combining Cosine Similarity with Classifier for Text Classification." Taylor & Francis, https://www.tandfonline.com/doi/full/10.1080/.

"Cosine Similarity." GeeksforGeeks, 6 Oct. 2020,https://www.tandfonline.com/doi/full/10.1080/0883951\

Dr. Vaidya, A., Vaidya, A., & Patil, A. (2022). OTT Platforms Usage Among Youth During Lockdown. IJRAR, 9(1), 1-17. ResearchGate.net. Retrieved: https://www.researchgate.net/profile/Alpana-Vaidya/pub

Garima Gupta, Komal Singharia (2021, Feb 1). Consumption of OTT Media Streaming in COVID-19 Lockdown: Insights from PLS Analysis. Sage Journals https://journals.sagepub.com/doi/full/10.1177/0972262921989118

Gal, Noa. "6 Ott Challenges & Pitfalls - and How to Avoid Them." TV Platforms & Content Protection, Viaccess-Orca Israel LTD, 20 Apr. 2022, https://www.viaccess-orca.com/blog/ott-challenges.

Jane, C. (2021, September 28). Why do you need a Recommendation Engine for your OTT Platform? https://dev.to/charlot/why-do-you-need-a-recommendation-engine-fo

Javed, M. (2020, November 4). Using Cosine Similarity to Build a Movie Recommendation System. Towards Datascience. https://towardsdatascience.com/using-cosine-similarity-to-build-a-

John, S. (2022, January 6). Netflix Movies and TV Shows recommender using cosine similarity. Medium. https://www.researchgate.net/publication/357701582_Iss

Jhala B., & Patadiya V. (2021, September). A Study On Consumer Behavior Towards OTT Platforms In India During Covid Era. https://www.researchgate.net/publication

Jose Immanuvel, J., Sheelavathi, A., Priyadharshan, M., Vignesh, S., & Elangko, K. (2022, June 17). Movie Recommendation System. Ijraset. https://medium.com/web-mining-is688-spring-2021/netflix-movies

Roy, D., Dutta, M. (2022). A systematic review and research perspective on recommender systems. J Big Data 9, 59. https://doi.org/10.1186/s40537-022-00592-5

Shetty, B. (2019, July 24). An in-depth guide on how Recommender System works. BuiltIn. https://builtin.com/data-science/recommend