Check for updates

Content-based Filtering for Improving Movie Recommender System

Xinhua Tian

New College, University of Toronto, Toronto, Canada.
xinhua.tian@utoronto.ca

Abstract. People constantly receive personalized information recommendations, and movie recommendation is one of the most recognized applications. Effective algorithms support the analysis of users' behavior, which helps to improve the rating system. Content-based filtering (CBF) is a major technique in recommender systems that operates on the premise of leveraging the relationship between user preferences and item characteristics to predict items. This paper provides a detailed look about the challenges that this method presents, emphasizing concerns with new users, inherent method limitations, issues with feature sparsity, the challenge of feature extraction, and the potential risk of over-specialization in suggestions. In synthesizing these challenges and innovations, this study highlights the potential of content-based filtering, marking its key role in the ongoing pursuit of personalized content delivery, while suggesting methods for improvement.

Keywords: Recommendation system, content-based filtering, recommended movie, machine learning, recommendation based on similarity and popularity.

1 Introduction

With the advancement of science and technology, more and more electronic devices have been invented and entered human life. The development of electronic information technology has led to the appearance of software. Mobile phones and computers can install many applications that facilitate life based on microchips. Therefore, it provides a vital choice for human to perform entertainment activities at home, such as watching movies. The convenience of watching movies has led to the rapid development of software, and what follows is how movie system provides better recommendation for users. Then, the accuracy of personalized recommendation has come up with a problem. To increase the accuracy, researchers conceived different techniques to find users' preference, in case to provide users' superior experience on recommending movies. For example, researchers and application developer know that users' action through each application could provide useful information for the algorithm. They started to ask users' social media accounts to gather more effective information, which helped the growth of machine learning. This is a way that researchers and programmers used for better analyze their users to provide movies they may preferred. Also, Jayalakshmi et al. has discussed several methods using for

© The Author(s) 2024
B. H. Ahmad (ed.), *Proceedings of the 2023 International Conference on Data Science, Advanced Algorithm and Intelligent Computing (DAI 2023)*, Advances in Intelligent Systems Research 180,
https://doi.org/10.2991/978-94-6463-370-2_61

recommender system for recommending movies in their article, such as K-means clustering, collaborative filtering (CF), content-based filtering (CBF) [1].

This study uses content-based filtering to analyze the data from various angles. Anaconda navigator provides the environment to clean and analyze the dataset into above ways. Dataset from GitHub had thousands of ratings for thousands of movies. By getting IMDb top movies using their rates to understand the structure of this dataset. Then, this study creates different columns in the data frame, and the study uses many ways to approach the dataset, such as getting the top romance movies. By using all these variables, users rated movies to help build the model of preferences and attributes. After using several data to train the model, given the example as the movie "The Dark Knight" and "Mean Girls", the model has provided top 10 similar movies based on the attributes it collected. After improving the model, the algorithm can provide the similarity between the user and other watchers for recommended movies.

2 Related Work

There are lots of different research discussing about recommender system for movies based on content-based filtering. Some research focused on the hybrid recommender system for relevant movies, and other research talked about movie recommendation based on users' network. One study from Ayyaz et al. discussed bout the hybrid recommender system using a fuzzy based technique to predict movies for users [2], and another study from Walek and Fojtik proposed a new algorithm called Predictory which connected two different filtering method with fuzzy expert system [3]. Different from above, Tahmasebi et al. talked about a deep autoencoder in their study, which utilized to provide movies for twitter users based on their social influence [4]. Similarly, Son and Kim used multiattribute network to propose a new method using content-based filtering, and they re-modeled the MovieLens data to measure the content difference [5].

Previous study research also completes the movie recommender system towards similarity and popularity, and several research even involved a hybrid method to overcome the drawbacks of single filtering. For example, Pera and Ng introduced a new recommender called GroupReM, which help group member choose their items by finding their similarity [6]. Furthermore, the hybrid method using in the article of Bahl et al. minimum the error may occurred by content-based filtering model, and they utilized the singular value decomposition (SVD) to enhance the efficiency of algorithm [7]. Similarly, Roy et al. presents a genre-based hybrid filtering for the new movies, and the nonlinear similarities has been utilized for predict the preference of users [8]. Although these studies proposed lots of different effective method to improve the recommender system using in movies, the attributes getting from retrieved content and users' rating are uninterruptedly the essential considered factors.

Content-based filtering uses the key words from previous products with higher rating from the user and compared with key words of other products to recommend [9]. This modelling function provides a more personalized recommendations based on

the individual's preferences, which also help to create the user portrait. In addition, personalizing suggestions is often a challenge for new users, especially when the system lacks detailed information on user behavior or preferences. However, unlike collaborative filtering requesting user interaction for appropriate recommendations, content-based filtering can directly make recommendations based on specific attributes or characteristics of users. This implies that customized recommendations may be given for new users based on their basic data, such as age, gender, location, or tags that they may first offer. This solved the cold start problem [10]. Therefore, content-based filtering provides a more flexible and faster way to provide personalized recommendations for new users without relying on a large amount of user behavior data. Different from previous studies, this research focus on the content-based filtering on movie recommendation and provides an improved algorithm after training the dataset.

3 Method

In this research, the model is generating using content-based filtering. Content-based filtering is one of the filtering systems from recommender system, which could handle a large collection from users' preference. To describe the accuracy of diversity, it is not possible to use collaborative filtering to provide users with suitable and qualified items. User's preference can be various. Two types of information are needed to complete the collaborative filtering model, users, and their ratings. The lack of information makes it difficult to deal with the "cold start" of a user. For example, if a user registers a website, he/she is marked as "new user", which indicates there are no ratings for this user. This demonstrated it is difficult to make recommendations to him/her without the support of additional information. However, content-based filtering solved this problem by creating "labels" for users. The information shows on the user's profile while they register is regarded as features, such as their ages, locations. Therefore, the core concept for content-based filtering is recommending new items to the user based on the highly rated item from the user in the past. If a user liked something in the past, the algorithm retrieves the feature of this item. The system will recommend the user item with the similar feature in the future. The application of this recommender system in this research has significant consequences since both the crew and the product team of the movie can be used as qualitative "labels" [8].

Content-based filtering are based on information retrieval and information filtering. The first initiation of this filtering to process is to create a detailed user's profile. The algorithm can retrieve user's attributes, and each attribute finding in user's profile is labeled as a feature for the filtering algorithm to approaching later. This step is called profile construction with attributes made. Products using in content-based filtering are involved with movies, books, and music, these provide the algorithm genres, the crew, writers, directors etc. This research will discuss about the movies. Mathematically, assuming a movie M have number n's feature, the feature set could be representing as an N-dimension vector $I_i = \{x_1, x_2, \ldots x_n\}$. Then, the profile with

user's preference is needed, which utilized for future development and calculating the weighted arithmetic mean of features the user has interacted with in the past. In this way, assume we have a user U, the profile of this user is $\mathrm{U_u} = [y_1, y_2, ..., y_n]$, and the interaction weights of users and products are w_i. Then, score, the value of y, should be $U_u = \frac{1}{N} \sum_{i=1}^{N} w_i \cdot I_i$. In addition, score of the user is calculated by the user's profile and the dimension vector. The score is the prediction for each item, and cosine similarity is the most common way to approach the value. The score is $score(u,i) = cos(\theta) = \frac{U_u \cdot I_i}{||U_u|| \cdot ||I_i||} = \frac{\sum_{j=1}^{n} y_j \cdot x_j}{\sqrt{\sum_{j=1}^{n} y_j^2} \cdot \sqrt{\sum_{j=1}^{n} x_j^2}}$. According to this score, the algorithm will calculate the similar items, and it will recommend number k's highest rating movies based on user's preference. The number k depends on the input value. This step is the main idea of content-based filtering called generating recommendation.

4 Experiment

IMDb, shorten from Internet Movie Database, is an online database that contains worldwide data about entertainment media, including movies, TV shows, podcasts, video games etc. This platform provides the information about the cast, production crew, and plot summaries. It also gives ratings and reviews from audiences or fans for users, which used to present the suggested movies or TV shows what users probably be interested in based on the algorithm. In addition to be specific, this research only discusses about recommender system of movies. From the IMDb website, there are divided sections as "featured today", "what to watch (from your watchlist)", "Top 10 on IMDb this week", "Fan favorites", "More to watch" offering various movies for users on the home page. Given IMDb's significant variety of offerings in movies and its personalized recommender system, understanding how these recommendations are generated is fundamental. For better analyzing and describing the mechanism behind the recommender system in IMDb, an accurate dataset is required to produce a theoretical model. The dataset using in this research is called Movie Recommender founded on GitHub. This dataset is a CSV file called movies, which includes movies name, title, genre, and so on. Besides, this research utilizes content-based filtering to generate appropriate model based on the variables created before.

In this experiment, analysis of this dataset is the first step. To clean the dataset, generating the top movie charts based on the IMDb ratings provides the meaning of votes, which uses for counting the number of users' supportive through each movie they rated. This research use 95% quantile as the cutoff, and the percentage of each movie's votes assists in determining the qualify movies for this research. After calculating, the movie was qualified for this research at least have 434 votes on IMDb, which means there are 2274 movies qualified to use in this research. Also, this dataset has 45463 genres.

Then, the recommender system starts to generate. Table 1 is the top 10 recommendation for the movie The Godfather. The first column represents the title number in this dataset, and the second column are the name of movie. Table 2 is the

top 10 recommendation for the movie The Dark Knight. The first recommendation is The Dark knight Rises, which supposes to be the installment of The Dark Knight. Also, the table shows that other recommendations are related to the Batman, but the recommendations supposed to be more personalized and diversified. This requires the improvements of screening the keywords, and the step of pre-processing the keywords is to calculate the frequent counts each keyword appears in the dataset. See table 3, the results of top 10 movies have changed to be more accurate and diversified compared with table 2.

Table 1. Recommendations for The Godfather

Data No.	Title
973	The Godfather: Part II
8387	The Family
3509	Made
4196	Johnny Dangerously
29	Shanghai Triad
5667	Fury
2412	American Movie
1582	The Godfather: Part III
4221	8 Women
2159	Summer of Sam

Table 2. Recommendations for The Dark Knight

Data No.	Title
7931	The Dark Knight Rises
132	Batman Forever
1113	Batman Returns
8227	Batman: The Dark Knight Returns, Part 2
7565	Batman: Under the Red Hood

524	Batman
7901	Batman: Year One
2579	Batman: Mask of the Phantasm
2696	JFK
8165	Batman: The Dark Knight Returns, Part 1

Table 3. Recommendations for The Dark Knight After Pre-Processing the Keywords

Data No.	Title
8031	The Dark Knight Rises
6218	Batman Begins
6623	The Prestige
2085	Following
7648	Inception
4145	Insomnia
3381	Memento
8613	Interstellar
7659	Batman: Under the Red Hood
1134	Batman Returns

To get more information about the recommendations, the algorithm has been improved. The improved algorithm collaborates in tracking down the counts of votes, the average of the votes, the year it released, and the weighted rating. The weighted rating (WR) represents both the average rating of a movie and the votes. The equation

of WR is $WR = \frac{v}{v+m}R + \frac{m}{m+v}C$, where v is the votes of a movies, m is minimum votes qualified which is 434 in this dataset. And R is the average rating, C is the average votes. From table 5 and table 6, the results of recommender system using content-based filtering are comprehensive, which presents a specific ranked movie list for the user. Compared with the table 3 and table 4, table 5 and 6 provides detailed data to support the reason chosen these top 10 movies. As a result, the improved recommender system based on content-based filtering provides an optimization strategy, and it provides users with highly personalized content and is very effective.

Table 4. Recommendations for Mean Girls After Pre-Processing the Keywords

Data No.	Title
3319	Head Over Heels
4763	Freaky Friday
1329	The House of Yes
6277	Just Like Heaven
7905	Mr. Popper's Penguins
7332	Ghosts of Girlfriends Past
6959	The Spiderwick Chronicles
8883	The DUFF
6698	It's a Boy Girl Thing
7377	I Love You, Beth Cooper

Table 5 Improved Recommendations for The Dark Knight

	Title	Vote_count	Vote_average	Year	WR
7648	Inception	14075	8	2010	7.917588
8613	Interstellar	11187	8	2014	7.897107
6623	The Prestige	4510	8	2006	7.758148
3381	Memento	4168	8	2000	7.740175
8031	The Dark Knight Rises	9263	7	2012	6.921448

6218	Batman Begins	7511	7	2005	6.904127
1134	Batman Returns	1706	6	1992	5.846862
132	Batman Forever	1529	5	1995	5.054144
9024	Batman v Superman: Dawn of Justice	7189	5	2016	5.013943
1260	Batman & Robin	1447	4	1997	4.287233

Table 6. Improved Recommendations for Mean Girls

	Title	Vote_count	Vote_average	Year	WR
1547	The Breakfast Club	2189	7	1985	6.709602
390	Dazed and Confused	588	7	1993	6.254682
8883	The DUFF	1372	6	2015	5.818541
3712	The Princess Diaries	1063	6	2001	5.781086
4763	Freaky Friday	919	6	2003	5.757786
6277	Just Like Heaven	595	6	2005	5.681521
6959	The Spiderwick Chronicles	593	6	2008	5.680901
7494	American Pie Presents: The Book of Love	454	5	2009	5.11969
7332	Ghosts of Girlfriends Past	716	5	2009	5.092422
7905	Mr. Popper's Penguins	775	5	2011	5.087912

Therefore, this study focused on understanding IMDb's personalized movie recommendation system. To ensure accuracy, the first step used data cleaning to get

the top movies. Using a weighted rating formula, which considers vote counts, average votes, release year, and movie's rating, this study developed a recommendation mechanism. The efficacy of our improved method was determined by comparative analysis utilizing multiple tables.

5 Conclusion

In this study, it presents the result of the research to develop an improved recommender system based on content-based filtering. By analyzing and calculating the dataset, the recommend system could provide a specific recommended movie list containing with other's opinion (votes) and scientific algorithm (weighted rating). The recommendation system has evolved significantly, the introduction of similarity and popularity supports to change the user experiences. As evidenced in prior studies, this process of improved recommender system not only enhance the quality of recommended movies, but also overcomes the limitation of collaborative filtering. Content-based filtering extracts and analyzes the keywords from movies to promote deeper recommendations with individual users. Although the table of result presents the personalized movies for the user, this algorithm still has some drawbacks. Because it depends too much on the user's prior browsing activity, it cannot suggest information in new categories that the user has not yet encountered. To reduce restrictions on content variety, further optimization techniques are needed. Furthermore, instead of using single filtering method, combining various filtering methods, the advantage of them could be used to remedy defects.

6 Reference

1. Jayalakshmi, S., Ganesh, N., Čep, R., and Senthil Murugan, J.: Movie Recommender Systems: Concepts, Methods, Challenges, and Future Directions. Sensors (Basel, Switzerland) 22(13), 4904 (2022)
2. Ayyaz, S., Qamar, U., Nawaz, R.: HCF-CRS: A Hybrid Content based Fuzzy Conformal Recommender System for providing recommendations with confidence. PLoS ONE 13(10), e0204849–e0204849 (2018).
3. Walek, B., Fojtik, V.: A hybrid recommender system for recommending relevant movies using an expert system. Expert Systems with Applications 158, 113452 (2020).
4. Tahmasebi, H., Ravanmehr, R., Mohamadrezaei, R.: Social movie recommender system based on deep autoencoder network using Twitter data. Neural Computing & Applications 33(5), 1607–1623 (2021).
5. Son, J., Kim, S. B.: Content-based filtering for recommendation systems using multiattribute networks. Expert Systems with Applications 89, 404–412 (2017).
6. Pera, Maria S., Ng, Yiu-Kai: A group recommender for movies based on content similarity and popularity. Information Processing & Management 49(3), 673–687 (2013).
7. Bahl, D., Kain, V., Sharma, A., Sharma, M.: A novel hybrid approach towards movie recommender systems. Journal of Statistics & Management Systems 23(6), 1049–1058 (2020).

8. Roy, A., Ludwig, S. A.: Genre based hybrid filtering for movie recommendation engine. Journal of Intelligent Information Systems 56(3), 485–507 (2021).
9. Jannach, D.: Recommender systems: an introduction. Cambridge University Press, New York (2011).
10. Deldjoo, Y. et al.: Movie genome: alleviating new item cold start in movie recommendation. User Modeling and User-Adapted Interaction 29(2), 291–343 (2019).

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.