사용자 정보 및 장르별 평균 평가를 이용한 내용 기반 영화 추천 시스템

일홈존 [1], 박두순 [1], 김대영 [1*]
[1] 순천향대학교 컴퓨터소프트웨어공학과
i_sadriddinov@mail.ru, parkds@sch.ac.kr, dyoung.kim_at_sch.ac.kr

Content-based Movie Recommendation system based on demographic information and average ratings of genres.

Sadriddinov Ilkhomjon Rovshan Ugli[1], Doo-Soon Park[1]
Dae-Young Kim[1*]
[1]Dept. of Computer Software Engineering, Soonchunhyang University

Abstract

Over the last decades, information has increased exponentially due to SNS(Social Network Service), IoT devices, World Wide Web, and many others. Therefore, it was monumentally hard to offer a good service or set of recommendations to consumers. To surmount this obstacle numerous research has been conducted in the Data Mining field. Different and new recommendation models have emerged. In this paper, we proposed a Content-based movie recommendation system using demographic information of users and the average rating for genres. We used MovieLens Dataset to proceed with our experiment.

1. Introduction

Recommendation systems play a pivotal role in our daily usage of various platforms. Ordering a hotel, selecting one's next music, choosing a book to read, everywhere we can see a recommendation system. This paper elucidates one such recommendation model. More precisely, it is a content-based movie recommendation system that offers a list of movies to users based on their demographic information and average ratings for each genre[1].

The paper contains the following sections: Section 2. explores the related works; Section 3. Explains the experiment procedures proceeded for this paper; Section 4. is about the conclusion and future work.

2. Related work

Recommender systems(RS) are a very powerful technique that can serve a lot of customers and non-costumers. There are plenty of areas where one can witness applications of RS[2]. There are different kinds of RS: Content-based filtering, Collaborative filtering, and Hybrid Recommendation system.

*corresponding author: 김대영
-Acknowledgements: This research was supported by the National Research Foundation of Korea (No. NRF-2022R1A2C1005921) and BK21 FOUR (Fostering Outstanding Universities for Research) (No.5199990914048) and the MSIT(Ministry of Science, ICT), Korea, under the National Program for Excellence in SW, supervised by the IITP(Institute of Information & communications Technology Planning & Evaluation) in 2021"(2021-0-01399)

2.1 Content-based filtering

This type of RS recommends an item based on mainly two things: one is a user profile, and another one is an item property. For example, if a user has similar demographic information to another user, then the system concludes that it can recommend the same list of items[3].

2.2 Collaborative filtering

In this method, the system tries to learn the actions of users and generates the model accordingly. User interactions are considered an important aspect of recommendations. The problem with this technique is when the dataset lacks user interaction – data sparsity. In such a case, the model's accuracy reduces sharply[4].

2.3 Hybrid Recommendation system

The hybrid technique is by its name the combination of two models discussed above. It takes advantage of both models and creates a more accurate and effective RS[5]. This method is widely used on many recommendation platforms.

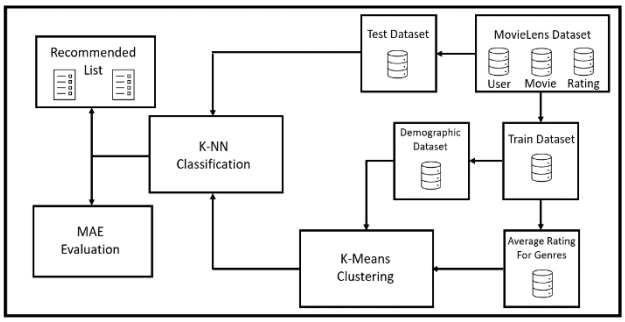
3. Experiment

The Environment for our experiment can be seen in the following table:

<Table 1> Experimental Environment

Division	Detailed contents
CPU	Intel(R) Core(TM) i5-7500 CPU @ 3.40GHz
RAM	8GB
HDD	128GB
OS	Windows 10 Pro 64bit
DEV.TOOL	R version 4.0.3 (2020-10-10) RStudio Version 2022.07.1

The following flow chart elucidates the whole process of the experiment:



(Figure 1) The architecture of the Recommendation System.

For our experiment, we used a MovieLens Dataset[6]. There are three separate datasets, each contains different variables related to either a movie or a user.

We split our User dataset into two sub-datasets: Train and Test. For each dataset, the following procedures are implemented.

a) For each user, we calculated an average rating based on the 18 different genres. To do this we first, found the list of movies watched by a specific user. Then we classified all movies according to different genres. Within each genre group, we calculated the average rating.

b) We merged the demographic dataset and average ratings for the genres dataset based on user IDs.

As a result, we constructed a new dataset that contains 22 columns: user's ID, user's age, user's gender, user's occupation, and 18 different genres.

Using the new dataset a k-means clustering algorithm is applied. This method took all pairs of users and calculated a euclidean distance.

As a k value, we tested from 2 to 10. As a result, 4 was relatively better for the k value.

Once the clustering step was finished we moved to the classification part. As a classification algorithm, we used k-nn(k nearest neighbor). This method classified users from the Test dataset into 4 different clusters using the euclidean method. The users' attributes were compared with the centroids of each cluster. The closest cluster's number is assigned to the user.

After clustering and classification, we evaluated our model. As an evaluation metric, we

used MAE(Mean Absolute Error). This metric can be calculated using the following equation 1:

$$MAE = \frac{\sum_{i=1}^{n}|y_i - x_i|}{n} \quad (1)$$

Where n is the total number of values, y_i is the predicted rating, and x_i is the real rating.

We have selected 10 users from the Test dataset as a random sample. Then we predicted ratings for 10 movies for each random user. After that, we applied the MAE method to check the accuracy of our model. As the parameters, we gave 100 real ratings and 100 predicted ratings. We got the 0.86 from our MAE calculation.

4. Conclusion and Future work

In this paper, we prosed a content-based movie recommendation system using 3 demographic attributes of users and the average ratings of each user for 18 different genres. For clustering, we used a k-means algorithm that used a euclidean distance. On the other hand, the knn technique was used to classify the new users into clusters. Finally, we calculated MAE for predicted ratings. The accuracy of our model was 0.86 based on MAE. In the future, we want to improve the clustering method. Since the dataset we used contains mixed data types: qualitative and quantitative. So it is important to consider this aspect while applying clustering methods.

Reference

[1] Al-Safi, J.; Kaleli, C. Item Genre-Based Users Similarity Measure for Recommender Systems. *Appl. Sci.* **2021**, *11*, 6108.

[2] Vilakone, P., Park, DS., Xinchang, K. *et al.* An Efficient movie recommendation algorithm based on improved *k*-clique. *Hum. Cent. Comput. Inf. Sci.* **8**, 38 (2018).

[3] Doo-Soon Park. Improved Movie Recommendation System based-on Personal Propensity and Collaborative Filtering. KIPS transactions on computer and communication systems 2 (11), 475-482.

[4] Phonexay Vilakone, Khamphaphone Xinchang, and Doo-Soon Park. Personalized Movie Recommendation System Combining Data Mining with the k-Clique Method. J Inf Process Syst, Vol.15, No.5, pp.1141~1155, October 2019.

[5] Phonexay Vilakone, Khamphaphone Xinchang, and Doo-Soon Park. Movie Recommendation System Based on Users' Personal Information and Movies Rated Using the Method of k-Clique and Normalized Discounted Cumulative Gain. J Inf Process Syst, Vol.16, No.2, pp.494~507, April 2020.

[6] F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. ACM Trans. Interact. Intell. Syst. 5, 4, Article 19 (January 2016)