협업 필터링을 사용한 유사도 기법 및 커뮤니티 검출 알고리즘 비교

일홈존 [1], 홍민표 [1], 박두순 [1*]
[1] 순천향대학교 컴퓨터소프트웨어공학과
i_sadriddinov@mail.ru, hmp4321@naver.com, parkds@sch.ac.kr

Comparison of similarity measures and community detection algorithms using collaboration filtering

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

Abstract

The glut of information aggravated the process of data analysis and other procedures including data mining. Many algorithms were devised in Big Data and Data Mining to solve such an intricate problem. In this paper, we conducted research about the comparison of several similarity measures and community detection algorithms in collaborative filtering for movie recommendation systems. Movielense data set was used to do an empirical experiment. We applied three different similarity measures: Cosine, Euclidean, and Pearson. Moreover, betweenness and eigenvector centrality were used to detect communities from the network. As a result, we elucidated which algorithm is more suitable than its counterpart in terms of recommendation accuracy.

1. Introduction

The flood of information caused many inconveniences for both sides: companies and customers. Customers were not able to use the services more comfortably since these platforms were not as efficient as expected in terms of accuracy and speed. On the other hand, companies failed to attract more clients due to their incompetent system. To resolve this issue longitudinal research experiments have been conducted in this domain. For example, Netflix[1] announced a competition to design an algorithm that would enhance the precision of the recommended movies on this platform, in 2006. This event attracted more companies to consider seriously the accuracy of their recommendation engines. To date, thousands of research papers were published per year. In the course of the experiments, researchers discovered and devised novel algorithms for recommendation systems. Even methods from other spheres were also implemented to improve the

*corresponding author: Doo-Soon Park
-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)

efficiency of this system. For instance, approaches from Social Network Analysis have been combined to give more precise recommendations by detecting the relationships among users.

We proposed a movie recommendation system using collaborative filtering by implementing three different similarity measures and two community detection methods. The main purpose of this empirical research is to compare the efficiency of different algorithms in our data set. The data set was derived from the GroupLens research group at the University of Minnesota, USA[2].

The further structure of this paper is as follows. In the second section, we provided background knowledge about related work. The third section explained our experiment in greater detail. Finally, in the fourth section, we discussed future work and a conclusion.

2. Related Work

This section gave a brief explanation of the related research, such as recommendation system, Similarity measures, and Social Network Analysis.

2.1 Recommendation System

A recommender system is an approach that filters out unnecessary information from the big data set available to provide relevant and proper suggestions to a user[3]. There are several types of recommendation algorithms. For instance, content-based, collaborative filtering, and hybrid recommendation algorithm.

Content-based recommendation method mainly focuses on the features of the item and the preferences or personal information of a user[4].

The collaborative filtering approach is simply from its name and gives recommendations according to the collaboration of users that are similar to each other in terms of preference towards items.

Hybrid recommender systems are the combination of two recommendation algorithms mentioned above. The main advantage of this method is that it can efficiently apply the powerful strength of other algorithms by discarding the drawbacks.

2.2 Similarity measures

There are various kinds of similarity measures devised. In this section, we mainly concentrate on the three most widely used algorithms: Cosine, Euclidean, and Pearson.

Cosine similarity is a representation of how two given sequences of numbers are similar[5]. This method is widely applied to various domains including documents. Equation 1 illustrates the formula of Cosine similarity.

$$\boldsymbol{Cos\ \alpha} = \frac{\boldsymbol{A} \times \boldsymbol{B}}{|\boldsymbol{A}| \times |\boldsymbol{B}|} = \frac{\sum_{i=1}^{n} \boldsymbol{A_i} \times \boldsymbol{B_i}}{\sqrt{\sum_{i=1}^{n} (\boldsymbol{A_i})^2} \times \sqrt{\sum_{i=1}^{n} (\boldsymbol{B_i})^2}} \quad (1)$$

Another similarity measure used in our paper is Euclidean. It is simply the distance between two points[6]. The Euclidean metric is the most pervasive method among its counterparts. And the following Equation 2 represents its formula.

$$\boldsymbol{d(x,y)} = \sqrt{\sum_{k=1}^{n} (\boldsymbol{x_k} - \boldsymbol{y_k})^2} \quad (2)$$

On the other hand, Pearson Correlation measures the ratio between the covariance and the standard deviation of two numeric vectors[7].

$$\boldsymbol{r_{xy}} = \frac{\boldsymbol{n \sum x_i y_i - \sum x_i \sum y_i}}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} \quad (3)$$

2.3 Social Network Analysis

The SNA is a field that analyzes the social network using graph properties from Graph Theory. The structure of a network or graph is mainly constructed by nodes and edges. Based on the implementation domain nodes and edges might be represented differently, for example, nodes can be a specific location, a person in a social network, an actor, and items; edges, on the other hand, illustrate the distance between locations, the relationship of people, and others[8].

The implementation of social network analysis algorithms may effectively solve many real-life problems. For instance, in Social Network Service(SNS)s it is crucially important to find a group of people who share the same interest and preferences to recommend commercial advertisements. The group of people here is called a subgraph or community in social network analysis and this procedure is known as community detection[9]. There are a lot of community detection algorithms have been designed till now. The most common ones

are Betweenness and Eigenvector centrality algorithms.

Betweenness centrality is a method of involvement of nodes along the shortest path within a network[10]. In simple terms, for a specific node $v_i \in G$, the betweenness centrality of a node v_i is:

$$\textbf{betweenness centrality}(v_i) = \sum_{i,k} \frac{\sigma_{v_i,\sigma_{v_k}}(v_i)}{\sigma_{v_i,\sigma_{v_k}}} \quad (4)$$

The significance of this measure is to find the node with high importance in the whole network.

Another centrality measure that was used in our experiment is Eigenvector Centrality. This measure also estimates the influence of the node in a given network. Each node with greater eigenvector scores has more connections than those with low scores[11]. It is very similar to another centrality measure known as degree centrality. The sharp difference between them might be seen in the degree of nodes.

3. Experiment

Fig. 1. displays the flow of the proposed recommendation system.

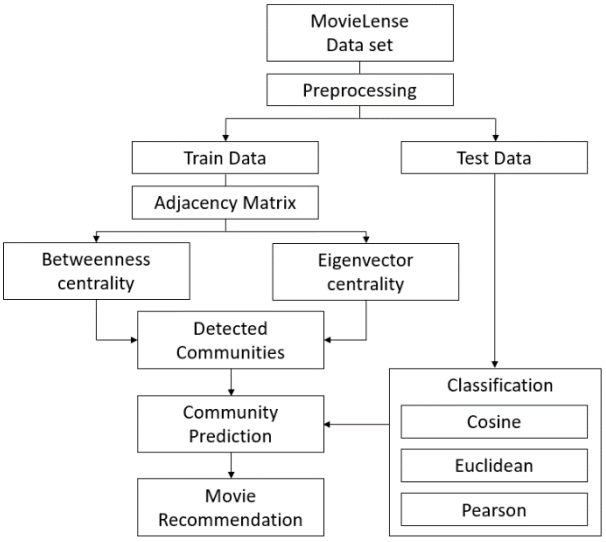

Figure 1. Architecture of the recommendation.

First of all, we downloaded the MovieLense dataset. We used two sub-datasets: user.csv and data.csv. The first file contains 943 rows and 5 columns, such as user ID, age, gender, occupation, and zip code. Another dataset consists of 100,000 rows and 4 columns like user ID, movie ID, rating, and timestamp[12].

Once downloading step was finished, we started preprocessing of datasets. We grouped age according to United Nations - Population Division(2019 Revision)[13]. And before making an adjacency matrix, we converted factor variables to numeric ones. Then we split our dataset into two parts with a ratio of 80:20. 80 per cent of the whole dataset was assigned to the train dataset, and the rest of it was for the test dataset.

The process of creating an adjacency matrix is as follows. If user A contains the same demographic vector such as age, gender, and occupation as user B, then we assign 1 to $Adj_{A\ B}$. Here, the Adj is an adjacency matrix and A and B are the row and column for A and B respectively.

In order to detect communities from the network, we converted our adjacency matrix to Igraph network object. Igraph is a library of R where many functions related to Social Network Analysis exist. After conversion, we passed our network to a function that does all calculations and provides communities and their members as a result. As the first algorithm, we used the Betweenness Centrality method to detect communities. In our experiment, we detected 41 communities in total.

After finding communities we discarded those that contain only one or two members inside a community. After completing this step, we moved to the recommendation stage by finding the top-10 movies inside each cluster. We used the average rating of each movie viewed by all members of a community. To classify all users from the test dataset we applied similarity algorithms. For each user from the test dataset, we calculated the distance between the user and the centre of each cluster(community). To measure this distance three similarity measures were implemented: Cosine, Euclidean, and Pearson. We assigned the user to the closest cluster in distance and recommended the top-10 movies. The same process was repeated for Eigenvector Centrality.

The evaluation was performed for both algorithms and all three similarity measures. To check the accuracy of movie recommendations we applied the MAE(Mean Absolute Error) metric since this method is one of the most commonly used evaluations. MAE returns values that are more interpretable as it is simply the average of all errors. We predicted the rating for a movie using Equation 5.

$$R_{U,i} = (\sum_{x}^{n} R_{x,i})/n \quad (5)$$

Here, $R_{U,i}$ is a rating for a movie i given by user U. $R_{x,i}$ is also a rating for the same movie but given by the user x. The main purpose of this equation is to predict an average rating for a specific movie using the most similar users rating for a target user.

After the prediction stage is finished, we calculated the MAE for our system using Equation 6.

$$MAE = \frac{\sum_i^N |P_{ui} - r_{ui}|}{N} \quad (6)$$

Here, P represents the predicted ratings for movies, and r is the real ratings. Moreover, N is the number of movies that were involved in this calculation.

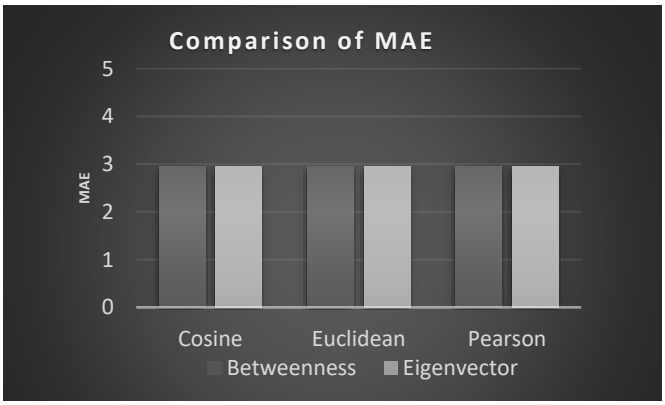

Figure 2. Comparison of accuracy by algorithms.

The chart above shows the MAE of the movie recommendation system for different algorithms. For all different algorithms, the MAE was the same 2.957.

4. Conclusion

The purpose of our approach was to compare two centrality algorithms and three different similarity measures. We applied a collaborative filtering recommendation system to the MovieLense data set. Finally, we calculated an evaluation by using the MAE metric. For all different algorithms, we got the same performance. We can conclude that both community detection methods produced the same result. Moreover, all three similarity algorithms also generated almost identical outputs when they were compared with each other.

Reference

[1] https://www.netflix.com/

[2] B. N. Miller, I. Albert, S. K. Lam, J. A. Konstan, and J. Riedl, “MovieLens unplugged: experiences with an occasionally connected recommender system,” in Proceedings of the 8th International Conference on Intelligent User Interfaces, Miami, FL, pp. 263-266, 2003.

[3] Phonexay Vilakone, Khamphaphone Xinchang, Doo-Soon Park, "Personalized Movie Recommendation System Combining Data Mining with the k-Clique Method", Journal of Information Processing Systems, Volume 15, No 5 (2019), pp. 1141 – 1155. 2019.

[4] C. C. Aggarwal, Recommender Systems. Cham: Springer International Publishing, 2016.

[5] A. R. Lahitani, A. E. Permanasari and N. A. Setiawan, "Cosine similarity to determine similarity measure: Study case in online essay assessment," 2016 4th International Conference on Cyber and IT Service Management, pp. 1-6, 2016.

[6] A. Suebsing and N. Hiransakolwong, "Feature Selection Using Euclidean Distance and Cosine Similarity for Intrusion Detection Model," 2009 First Asian Conference on Intelligent Information and Database Systems, pp. 86-91, 2009.

[7] M. C. ABOUNAIMA, F. Z. E. MAZOURI, L. LAMRINI, N. NFISSI, N. E. MAKHFI and M. OUZARF, "The Pearson Correlation Coefficient Applied to Compare Multi-Criteria Methods: Case the Ranking Problematic," 2020 1st International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET), pp. 1-6, 2020.

[8] Khamphaphone Xinchang, Phonexay Vilakone, Doo-Soon Park, “Movie Recommendation Algorithm Using Social Network Analysis to Alleviate Cold-Start Problem”, Journal of Information Processing Systems, Volume 15, No 5(2019), pp 616-631. 2019

[9] R. Kanawati, "Community detection in social networks: The power of ensemble methods," 2014 International Conference on Data Science and Advanced Analytics (DSAA), pp. 46-52, 2014.

[10] G. Ausiello, D. Firmani and L. Laura, "The (betweenness) centrality of critical nodes and network cores," 2013 9th International Wireless Communications and Mobile Computing Conference (IWCMC), pp. 90-95, 2013.

[11] A. Bihari and M. K. Pandia, "Eigenvector centrality and its application in research professionals' relationship network," 2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE), pp. 510-514, 2015.

[12] https://grouplens.org/

[13] https://ourworldindata.org/grapher/population-by-broad-age-group?country=~OWID_WRL