소셜 네트워크 분석 및 정규화된 할인 누적 이익을 이용한 영화 추천 시스템

비라콘 폰싸이*, 신장 캄파폰*, 이한나** 박두순***
*순천향대학교 컴퓨터학과
,*순천향대학교 컴퓨터소프트웨어공학과
e-mail : xayus@yahoo.com

Movie Recommendation System using Social Network Analysis and Normalized Discounted Cumulative Gain

Phonexay Vilakone*, Khamphaphone Xinchang*, Hanna Lee**, Doo-Soon Park***,
*Dept. of Computer Science and Engineering, Soonchunhyang University
,*Dept. of Computer Software Engineering, Soonchunhyang University

Abstract

There are many recommendation systems offer an effort to get better preciseness the information to the users. In order to further improve more accuracy, the social network analysis method which is used to analyze data to community detection in social networks was introduced in the recommendation system and the result shows this method is improving more accuracy. In this paper, we propose a movie recommendation system using social network analysis and normalized discounted cumulative gain with the best accuracy. To estimate the performance, the collaborative filtering using the *k* nearest neighbor method, the social network analysis with collaborative filtering method and the proposed method are used to evaluate the MovieLens data. The performance outputs show that the proposed method get better the accuracy of the movie recommendation system than any other methods used in this experiment.

1. Introduction

Today, the researchers pay attention to the way community's detection in huge networks and one of a method that the researcher is interested that is the social network analysis method which use to detect the community in the social network [1]. Some of the researchers introduced this method in the movie recommendation system to get a better accuracy of the recommended movies to the users. Beside to the method of the community's detection in the social network, the many of the researchers also pay attention to the method of ranking measure like normalized discounted cumulative gain which is a family of ranking measures widely used in practice and the result of this method is very efficient. The purpose of this paper is to get an upwards proficient method than the social network analysis method. The proposed movie recommendation system using social network analysis and normalized discounted cumulative gain method. We used the centrality of social network analysis i.e. betweenness centrality method.

The idea is clustering by using community detect based on edge betweenness centrality method for the user, then finding the group for new users based on their personal characteristics such as gender, age, and occupation with the help of cosine similarity measure method. After that, the system will recommend movies in the group with similarity to target users. The purpose of this paper is to develop techniques that can recommend the most appropriate movies to new users based on their personal characteristics.

The remaining of the paper is presented as follows: In segment 2, related works in the area are presented. In segment 3, the proposed method is to explain further. In segment 4, the detail of experimental analysis is presented, the results of the experiments are presented and the comparison to evaluate of the performance is presented by the proposed method with the collaborative filtering using a *k*-nearest neighbor and the social network analysis with collaborative filtering. Finally, in segment 5, the conclusions are conducted and future works are presented.

2. Related Work

Before introducing the proposed method in the next segment. The meaning of this method, along with some theoretical which the method used in this paper are presented below.

***Corresponding Author: Doo-Soon Park
※ 본 연구는 과학기술정보통신부 및 정보통신기술진흥센터의 대학ICT연구센터육성지원사업의 연구결과로 수행되었음(IITP-2017-2014-0-00720-002)

2.1 Social Network Analysis Method

A social network is a process of investigating social structures using networks and graph theory. It characterizes networks structure in terms of nodes like actor individual, people, or thing in the network. Centrality is a representative indicator used in social network analysis. There is 3 kind of centrality including degrees' centrality, betweenness centrality, and closeness centrality. Betweenness centrality is the measure of the center of the graph based on the shortest path. The betweenness centrality for each vertex is the number of the shortest paths that pass through that vertex. Girven et al [2], presented community detect in social network and biological network based on edge betweenness centrality to avoid the shortcomings of the hierarchical clustering method, they find which edge in the network are most betweenness another pair of vertices by used betweenness centrality to edges and determine the edge betweenness of edge like the shortest path number between pair of vertices that runs long it. The algorithm used for identifying communities is simply stated as follow:

- Calculate the betweenness centrality for all edges in the network,
- Remove the edge with the highest betweenness,
- Recalculate betweennesses centrality for all edges affected by the removal,
- Repeat from step 2 until no edges remain.

2.2 Normalized Discounted Cumulative Gain(NDCG)

The normalized discounted cumulative gain which is one of the most popular evaluation measures in Web search engine [3, 4] algorithms or related applications. Using a graded relevance scale of documents in a search-engine result set, NDCG measures the usefulness, or *gain*, of a document based on its position in the result list. NDCG has two advantages compared to many other measures. First, NDCG allows each retrieved document has graded relevance while most traditional ranking measures only allow binary relevance. That is, each document is viewed as either relevant or not relevant by previous ranking measures, while there can be degrees of relevancy for documents in NDCG. Second, NDCG involves a discount function over the rank while many other measures uniformly weight all positions. This feature is particularly important for search engines as users care top ranked documents much more than others. For a query, the *normalized discounted cumulative gain*, or NDCG, is computed as:

$$NDCG = DCG / IDCG \quad (1)$$

The traditional formula of DCG accumulated at a particular rank position p is defined as:

$$DCG = \Sigma^{p}_{i=1} [rel_i / log_2 (i+1)] \quad (2)$$

Where IDCG is ideal discounted cumulative gain,

$$IDCG = \Sigma^{|rel|}_{i=2} [2\hat{}rel_{i-1} / log_2(i+1)] \quad (3)$$

And |REL| represents the list of relevant documents (ordered by their relevance) in the corpus up to position *p*.

The NDCG values for all queries can be averaged to obtain a measure of the average performance of a search engine's ranking algorithm. Note that in a perfect ranking algorithm, the DCG will be the same as the IDCG producing an NDCG of 1.0. All NDCG calculations are then relative values on the interval 0.0 to 1.0 and so are cross-query comparable.

3. Movie Recommendation System using Social Network Analysis and Normalized Discounted Cumulative Gain

The process of gathering data and work processes for the guidance system is presented in Fig. 2. Fig. 2 presents the movie guide system using social network analysis and normalized discounted cumulative gain.

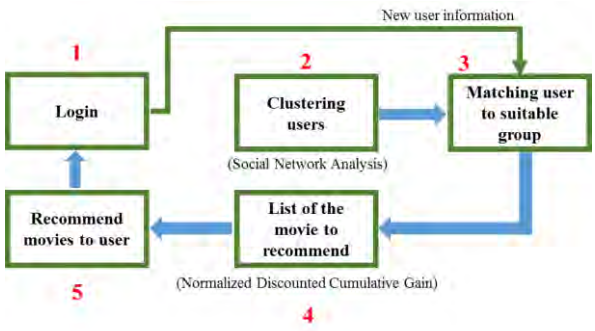

Fig. 1. Flow chart of proposed method.

The illustration of system movie recommendation system using social network analysis and normalized discounted cumulative gain is shown in Fig. 1.

Number 1 in Fig. 1, when new users want to join the system they need to log in.

Number 2 in Fig. 1, the personal information of the user is used to make relationship matrix between users, this matrix refer to user connection each other in the network. After that, we cluster the user into several groups by using detect community based on edge betweenness.

Number 3 in Fig. 1, after the clusters of the users, are found, then we want to find the most similar cluster with respect to the new user, in this process, cosine similarity measure algorithm helps to find a group similar to the new user by computing distance between new user and users in each group from their personal information. Then the largest group with have number ones will be chosen as the most similar group for the relevant user.

Number 4 in Fig. 1, the movies that watched by the users in the group will calculate the score according to the score of the movie.

Number 5 in Fig. 1, the system will recommend the most movies that watched by users in the group with a high ranking to the new user.

4. Experimental Analysis

4.1 Datasets

To evaluate the proposed method, the MovieLens [5] dataset was used for the experimentation. This dataset is separated into two parts. First part is experimental data in which there are 800 users and we used this data for the training the purposed method. The second part is test data in which there are 143 users and we used this data for testing the purposed method. In the Movielens dataset, there are 100,000 ratings from 1,684 movies of 943 users, and the necessary information for users is including age, gender, and occupation.

4.2 Analysis Result

After developing the proposed movie recommendation system using the social network analysis and normalized discounted cumulative gain, the number of movies rated by new users in movies recommended by the system was predicted. This paper uses a widely used evaluation metric for benchmarking the output of the proposed method. The mean absolute percentage error (MAPE) is the method of predicting the accuracy of the predictive method in the statistics given by the formula. [6-8]:

$$\text{MAPE} = ((100\%) / n) \Sigma^{n}_{t} | (A_t - F_t) / A_t \qquad (4)$$

Where A_t is the actual value and F_t is the forecast value.

We compared the MAPE with the collaborative filtering using the k nearest neighbor method, the social network analysis with collaborative filtering method and the social network analysis with k nearest neighbor method to estimate the effectiveness of the proposed approach. If the number of results is low, it means that our methods are salutary. For the MAPE calculation, the Eq. (4) is used.

First, we measure the MAPE for the proposed method. The MAPE value is 16.50% (see below).

$$\text{MAPE} = ((100\%) / n) \Sigma^{n}_{t} | (A_t - F_t) / A_t = 16.50\% \qquad (5)$$

Second, we calculated MAPE for social network analysis with collaborative filtering method. The value of MAPE was 20.20%. Finally, we calculated MAPE for collaborative filtering with the *k* nearest neighbor method. The value of MAPE was 21.9% as shown in Fig. 3 below.

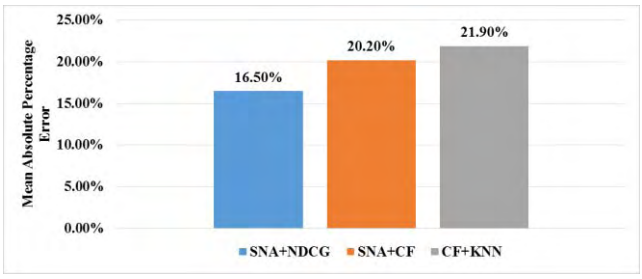

Fig. 2. Comparison of Existing Method

After comparing these methods, the MAPE results calculated by the movie recommendation system using the collaborative filtering using the *k* nearest neighbor method, the social network analysis with collaborative filtering method, and the proposed approach are more accurate and effective, confirm that the performance of three methods at our best.

5. Conclusions

In this paper, we have proposed an alternative approach for recommendation system using the social network analysis and normalized discounted cumulative gain method to increase more accuracy in the movie recommendation system. We clustering the community or group to the user based on edge betweenness centrality and cosine similarity measure helps to find users similar to the target user by using their personal information and recommended movies to the user with the help of normalized discounted cumulative gain method. The method that we proposed in this paper is very effective for a movie recommendation system. In addition, the method presented in this paper showed the best performance, followed by the social network analysis and CF, and followed CF with *k*NN.

Future studies, we will modify the algorithm of social network analysis to decrease the time of experimental and apply this algorithm to the huge dataset.

References

[1] Vilakone P, Xinchang K, Park D S and Hao F. “An efficient movie recommendation algorithm based on improved *k*-clique”, Human-centric Computing and Information Sciences, 8:38, pp. 1-15, 2018

[2] Girvan M and Newman M E J. “Community structure in social and biological networks”, PNAS June 11, 2002. 99 (12) 7821-7826.

[3] J¨arvelin K and Kek¨al¨ainen J. “IR evaluation methods for retrieving highly relevant documents”, In Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, pp 41–48(2000).

[4] J¨arvelin k and Kek¨al¨ainen J. “Cumulated gain-based evaluation of IR techniques”, ACM Transactions on Information Systems (TOIS), 20(4): pp 422–446(2002).

[5] Harper F M and Joseph A K. “The MovieLens Datasets: History and Context”, ACM Transactions on Interactive Intelligent Systems (TiiS) 5, 4, Article 19, December, pp 19(2015).

[6] Tofallis. “A Better Measure of Relative Prediction Accuracy for Model Selection and Model Estimation”, Journal of the Operational Research Society, 66(8), pp 1352-1362(2015).

[7] Hyndman R J and Anne B K. “another look at measures of forecast accuracy”, International journal of forecasting 22.4, pp 679-688 (2006).

[8] Kim S and Kim H Y. “A new metric of absolute percentage error for intermittent demand forecasts”, International Journal of Forecasting, volume 32 issue 3, pp 669-679(2016).