Available online at www.ijpe-online.com

International Journal of Performability Engineering

vol. 14, no. 8, August 2018, pp. 1719-1725
DOI: 10.23940/ijpe.18.08.p8.17191725

User Group-based Method for Cold-Start Recommendation

Jing He, Shuo Yuan, Yi Xiang, and Wei Zhou*

National Pilot School of Software, *Yunnan University*, *Kunming*, *650091*, *China*

Abstract

Recommendation algorithms seek to predict user ratings or preferences. Due to limited information, it is difficult to make these predictions for new users. Therefore, a dynamic cold-start recommendation algorithm would be highly helpful in such quick-changing social networks. In this paper, a novel user group-based collaborative method, called UCFRA (User group-based Cold-start Friend Recommendation Algorithm) is proposed. UCFRA integrates a graphical model and statistical population characteristics into a user group model and then combines this extended user group model with cold-start information to generate a new recommendation algorithm. Moreover, a content popularity model based on user groups and a user rating matrix is designed. In order to improve recommend precision, a user group Top-N recommendation model based on k-nearest neighbors is provided. A series of experiments involving collection of a huge data set was developed to evaluate the effectiveness of UCFRA. The experimental results showed that UCFRA is a valid algorithm.

Keywords: recommendation algorithm; cold-start systems; cluster model; content popularity model

(Submitted on April 29, 2018; Revised on June 15, 2018; Accepted on July 14, 2018)

1. Introduction

As one of the most popular applications, social networks benefit many people by continuously introducing new methods and tools. Currently, Facebook, Twitter, Line, and QQ have become a significant part of our lives. These kinds of social networks not only appeal to people looking for friendship on the internet, but also gradually satisfy many people's socialization requirements and in some cases have begun to replace traditional face-to-face social networks. Today, businesses recognized the value of connecting with others through the wide variety of networks that are springing up almost daily. These businesses can appreciate the ease with which they can keep in touch with clients, suppliers, and a host of other stakeholders. According to statistics from Hubspot [1], 92% of marketers claimed that social media marketing was important for their business, with 80% indicating their efforts in this area increased traffic to their websites.

However, full utilization of social networking [2] is not so easy. In a huge social network, it can be difficult to find common interesting relationships between people. If we only treat a social network as a low-cost management tool for a person's real interpersonal relationships, it would not go far enough. It is important to understand how people can come to know strangers easily online, which is supposed to facilitate interpersonal communications among users through their weak online relationships.

Several traditional friend recommendation algorithms have been proposed to explore people with similar interests [3-7], which is the key issue faced by social networks to meet a person's socialization requirements. However, these traditional methods are not efficient to deal with newly registered users. Therefore, cold-start algorithms are needed to solve these problems. However, previous methods still need to be promoted. To solve the cold-start problem, this paper deeply investigates the special influence caused by social popularity. Moreover, a user group-based cold-start friend recommendation algorithm (UCFRA) is proposed. This paper makes the following contributions:

* Corresponding author.
E-mail address: zwei@ynu.edu.cn

(a) A novel user group-based cold-start friend recommendation algorithm, called UCFRA, is proposed. To the best of our knowledge, this is the first work to establish a dynamic cold-start friend recommendation algorithm which integrates a graphical model and statistical population characteristics into a user group model. UCFRA is both efficient and scalable.

(b) We propose a user clustering model based on statistical population characteristics and the content popularity model based on user groups. This method solves the problem of numerical social network data and can make subsequent predictions more convenient.

(c) Development of a series of experiments involving the collection of an extremely large data set to evaluate the effectiveness of UCFRA. The experimental results showed that UCFRA is a valid algorithm.

The rest of this paper is organized as follows. Section 2 presents related work, and Section 3 gives an overview of our user group-based cold-start friend recommendation algorithm (UCFRA). The user clustering model based on statistical population characteristics is also proposed in this section. Section 4 shows the experimental results, and Section 5 presents our conclusions.

2. Related Work

The ability to find people with similar interests is a key issue in social networks. Traditional friend recommendation algorithms generally fall into one of the following three categories [8]:

(a) Content-based friend recommendations. This type of algorithm can recommend friends that share similar personal attributes, such as age, gender, occupation, and location information.

(b) Interest-based friend recommendations. This type of algorithm recommends friends based on their interest in the same hobbies or interests, but it does not care about whether the recommended person is known in real life. Therefore, this kind of algorithm usually extracts users' interests based on their activity data and then calculates interest similarities to recommend friends.

(c) Social-relationship-based friend recommendations. These algorithms use the users' existing social relationships and recommend new friends based on these social relationships, such as by recommending friends' friends.

All of the above algorithms utilize users' historical behaviors and interests to predict new friends. Thus, a large set of behavioral data is necessary for these recommendation algorithms. However, not all users have enough historical behavioral data, especially those who are newly registered. Thus, it is necessary to develop a cold-start friend-recommendation algorithm for use in circumstances where there is insufficient historical behavioral information.

Traditional cold-start friend-recommendation methods mainly fall into two categories: (1) individual recommendations based on behavioral data from other websites, and (2) non-individual recommendations based on a hot list. Depending on the conditions of the cold-start users' agreements, module one is used to obtain the users' behavioral data from other websites and then specifically analyze the similarity among users to generate a collection of users that have interests nearest to those of cold-start users' interests scored by similarity. Then, the highest scoring top N users are selected as friend recommendations to the cold start user. An example of this type of algorithm is SNetRS [9], which uses user preferences obtained from Facebook. Module two analyzes statistics to obtain a hot recommendation list according to global historical user data and recommends the hottest user as a friend to the cold-start user. Once the cold start user generates a certain amount of data, the algorithm then switches to personalized friend recommendation.

Based on the above two modules, various solutions are being investigated. Matrix factorization is one of the most commonly used methods to solve the cold-start user problem [10-12]. Social network analysis (SNA) is another common method. To develop this theory, Yang [13] first built a rating matrix for a user relationship network and then divided the users into many cliques based on SNA. When a new user joins, these cliques are used to find the nearest neighbor to recommend as a friend. Some special solutions have also been used, such as the cold-start app recommendation method proposed by Lin [14], the market-based approach proposed by Daoudet [15], and the association cluster filtering based recommendation algorithms [16-18], as well as other examples.

Although several cold-start algorithms exist, all have application reliability defects and shortcomings in large data scenarios. Some necessary information, such as demographic characteristics or the content popularity model, has not been introduced in these cold-start algorithms. In this paper, we combined the graphical model and statistical population characteristics into a user group model. Then, we integrate this extended user group model and cold-start information into a recommendation algorithm. A series of experiments involving the collection of an extremely large data set is developed to evaluate the effectiveness of UCFRA. The experimental results showed that UCFRA is a valid algorithm.

3. User Group-based Cold-Start Friend Recommendation Algorithm (UCFRA)

3.1. Graphical Model of a User Group-based Social Network

Figure 1 shows the social network of the user group. As you can observe from this picture, users are organized according to their common interests. If the users are established, it is easy to calculate their interests based on historical data. However, if they are new users (as shown in Figure 1), these traditional methods are not effective. Thus, cold-start algorithms are introduced to solve these problems.

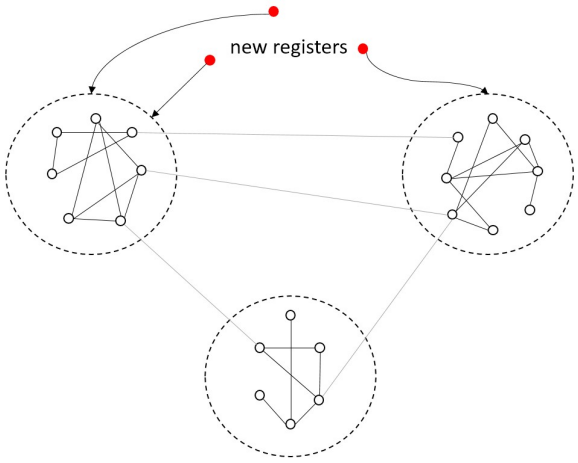

Figure 1. Graphical model of a user group-based social network

3.2. User Clustering Model based on Statistical Population Characteristics

In a social network, the demographic information of users plays an important role in understanding users' interests and information such as age, gender, and so on. People in different age groups may have different interests, and people in the same age group tend to have similar hobbies. For example, seldom do members of the post-90s generation know about hot superstars from the 60s or 70s. Likewise, members of the post-70s generation tend not to be interested in today's popular stars. Moreover, gender often also has a large influence on a person's interests. For instance, boys typically like games, sports, and military affairs, while girls generally prefer TV dramas and shopping. Therefore, we can build a user clustering model based on users' age, gender, geographical location, and other demographic information. Here, our clustering model is designed as follows:

User group sets $U=\{u_1,u_2,\cdots,u_n\}$ and demographic characteristics $F=\{f_1,f_2,\cdots,f_n\}$, $f_{i,j}$ express that user i owns demographic characteristic j. Thus, the similarity between user u and user v is

$$sim(u,v)=\frac{\sum_{f\in F} f_{u,f} f_{v,f}}{\sqrt{\sum_{f\in F} f_{u,f}^{\ 2}}\sqrt{\sum_{f\in F} f_{v,f}^{\ 2}}} \tag{1}$$

At the beginning, the algorithm randomly selects c center points. Then, the k-means clustering strategy continually iterates many times to generate s user groups with similar features. The distance between every user to the center point (c) is expressed by the Euclid Distance [19], namely

$$dis(u,r) = \sqrt{\sum_{j=1}^{n} (f_{u,j} - cen_{r,j})^2} \tag{2}$$

Here, *dis(u,r)* expresses the distance between user *u* with feature *r* to the center point, $f_{u,j}$ represents the feature attribute *j* of user *u*, and $cen_{r,j}$ is the feature attribute *j* of center point *r*.

Based on demographic characteristics, users can be clustered when a user registers an account. The main goal here is to eliminate interference due to demographic characteristics but not in the friend recommendation based on content. This technique is quite fundamental to the cold-start friend recommendation algorithm.

3.3. The Content Popularity Model based on User Groups

It is insufficient to use only demographic information to predict and analyze users. However, finding powerful prediction indicators is very important to supply and support recommendation results. The prediction indicator, which is generated from user historical behavior information, can help to improve predictions. Common behavior information features [20] include: activity index, keyword index, age and gender index, time sequence index, user behavior index, and so on. The content popularity indicator is the basic prediction index. We combined generated behavior data with a previously defined user clustering model and determined a definition for popularity, which is expressed as follows:

For each user, fol_i represents the count of records that each recommended user *i* follows from other users, rec_i is the count of records that each recommended user *i* is recommended by other users, then

$$ratio_i = \frac{fol_i}{rec_i} \tag{3}$$

For users in each user group, $fol_{i,ug}$ represents the count of records that each recommended user *i* in the user group follows from other users, $rec_{i,ug}$ represents the count of records that each recommended user *i* in the user group is recommended by other users, then

$$ratio_{i,ug} = \frac{fol_{i,ug}}{rec_{i,ug}} \tag{4}$$

$$fol\_norm_{i,ug} = \frac{fol_{i,ug}}{\sum fol_{i,ug}} \tag{5}$$

$$rec\_norm_{i,ug} = \frac{rec_{i,ug}}{\sum rec_{i,ug}} \tag{6}$$

$$diff_{i,ug} = fol\_norm_{i,ug} - rec\_norm_{i,ug} \tag{7}$$

Here, $ratio_{i,ug}$ in Equation (4), $fol\_norm_{i,ug}$ in Equation (5), and $rec\_norm_{i,ug}$ in Equation (6) all have an important influence on popularity. $ratio_{i,ug}$ and $fol\_norm_{i,ug}$ are positively correlated with popularity, while $rec\_norm_{i,ug}$ is negatively correlated with popularity.

According to the parameters mentioned above, the definition of numerical content popularity is as follows:

$$ItemPop = \frac{fol\_norm_{i,ug} \times ratio_{i,ug}}{rec\_norm_{i,ug}} = \frac{fol_{i,ug}^{\ 2} \times \sum rec_{i,ug}}{rec_{i,ug}^{\ 2} \times \sum fol_{i,ug}} \tag{8}$$

The indexes proposed above use the user's historical behavior data as implicit feedback, which can be numerically used to describe how users understand the content popularity. If we connected this model with user groups, the content popularity

model could be used to score the popularity of popular content in a user group. This method solves the problem that social network data cannot be numerical and can make subsequent predictions more convenient.

3.4. The Top-N Recommendation Model based on KNN

After using scored results to generate a recommendation list, the distance between each user in every user group and the center point is the difference. The score of content in this user group would not be subject to a weighted average [21]. At the same time, the content score standards for different users are also different. Therefore, when computing the content score of a user group, the score warp of users or content should be considered. The following describes the design of a Top-N recommendation model with acceptable deviations.

$Item = \{item_1, item_2, \cdots, item_k\}$ is defined as the content collection, $S_{g,k}$ represents the score for user g of content k, so S is a matrix of score. Each user group G includes u users, namely G belongs to U, then $UG = \{G_1, G_2, \cdots, G_n\}$ and $G_i = \{g_1, g_2, \cdots, g_u\}$. The score equation for each item in a user group can be described as follows:

$$Score(G, item_k) = \frac{\sum_{g=G} sim(g, core) \times (s_{g,k} - \overline{s_g})}{u \times (s_{g,k} - \overline{s_g})} \tag{9}$$

Here, $sim(g, core)$ represents the similarity of user g and the center point $core$ in the user group. $\overline{s_g}$ is the average score of $item_k$ in user group G. Finally, in each user group, a recommendation list will be generated based on the rank of different content score, $Score(G, item_k)$.

For a new user who owns the user feature attribute description, namely a cold-start user when he enters the systems, the Top-N recommendation model is utilized for him. Firstly, the k nearest neighbor method is used to search for the center node that is the nearest to $item_j$ in each user group. Then, according to the similarity of user a and these center nodes, along with the item score of relative user group, the content score of user a is predicted. Finally, the comprehensive score rank of the highest k prediction score of each item is formed. The prediction score for user a to $item_j$ is computed as follows:

$$predict(a, item_j) = \frac{\sum sim(a, G) \times Score(G, item_j)}{k \times Score(G, item_j)} \tag{10}$$

In Equation (10), $sim(a, G)$ is the similarity between user a and the center node of the featured user group G. $Score(G, item_j)$ represents the score of user group G with regard to $item_j$. We can then sort items j according to their prediction scores and select the top N highest scored items as the items recommended to new user a.

4. Experiments and Evaluations

To evaluate our proposed algorithm, a very large kind of micro-blog data set from Tencent Corporation, which was also used in the KDD Cup 2012, was used. The experiments showed that our algorithms were efficient.

We compared our algorithm to the two basic recommendation algorithms: the popularity rank list based recommendation algorithm (PRLRA) and the weighted popularity rank list based recommendation algorithm (WPRLRA). The performance comparison with regard to precision, recall, and F1-Score under different K parameters of these three algorithms is shown in the following figures.

Figure 2 shows the precision performance comparison under different K parameters for the three algorithms. When k is less than 80, the precision of all three algorithms is low and nearly the same. However, when k is greater than 80, the precision of our algorithm, UCFRA, is the greatest.

Figure 3 shows the performance comparison of the three algorithms with regard to recall under different K parameters. When k is less than 320, the precision of all three algorithms is low and nearly the same. However, when k is greater than 320, the precision of UCFRA and WPRLRA are nearly the same and are both better than that of PRLRA.

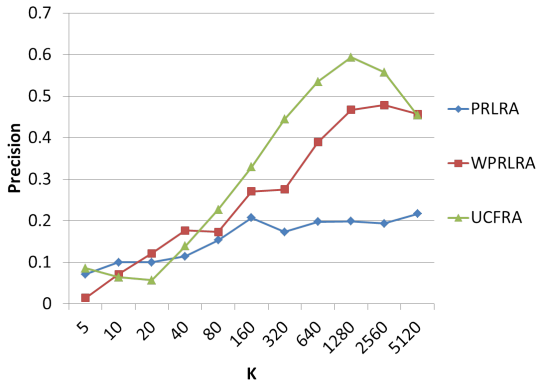

Figure 1. Precision performance comparison of three algorithms under different K parameters

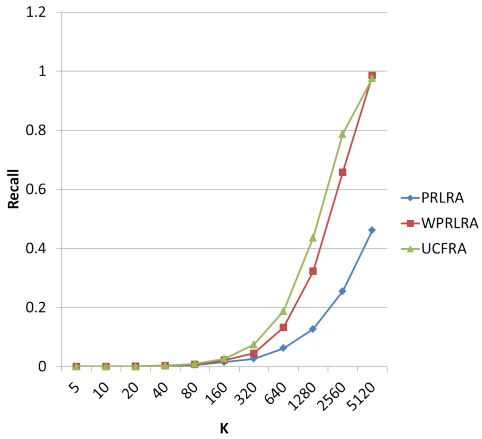

Figure 2. Recall performance comparison of three algorithms under different K parameters

Figure 4 shows the performance comparison with regard to the F_1-Score for the three algorithms under different K parameters. When k is less than 80, none of the F_1-Scores of the three algorithms are high, and they are all nearly the same. However, when k is greater than 80, the F_1-Scores of UCFRA and WPRLRA are nearly the same and are both better than that of PRLRA.

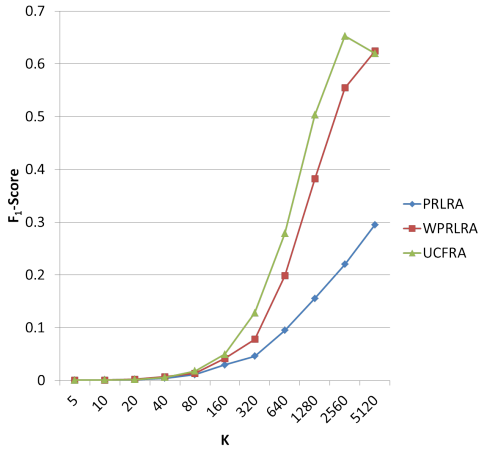

Figure 3. F_1-Score performance comparison of three algorithms under different K parameters

5. Conclusions

This paper introduced the background and significance of the social network recommendation problem, focusing on cold-start recommendations for new users. Then, we proposed a user cluster model based on population statistical characteristics, a

content popularity model based on user groups, and a Top-N recommendation model based on KNN. After combing the above-mentioned models, the author proposed a cold-start friend recommendation algorithm based on user groups. Then, we used Tencent's micro-blog data collection as the experiment data set to compare and analyze the algorithm in detail. The experimental results showed that the cold-start friend recommendation algorithm based on user groups proposed in this paper is superior to common recommendation algorithms based on a popularity ranking list and a weighted recommendation algorithm based on a popularity ranking list.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (61762089, 61663047) and the Open Foundation of Key Laboratory in Software Engineering of Yunnan Province (2017SE206).

References

1. M. An, "The Future of Content Marketing: How People are Changing the Way They Read, Interact, and Engage with Content," (https://research.hubspot.com/the-future-of-content-marketing, accessed June 25, 2016)
2. T. A. Pempek, Y. A. Yermolayeva, and S. L. Calvert, "College Students' Social Networking Experiences on Facebook," *Journal of Applied Developmental Psychology*, Vol. 30, No. 3, pp. 227-238, 2009
3. M. Moricz, Y. Dosbayev, and M. Berlyant, "PYMK: Friend Recommendation at Myspace," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp.999-1002, New York, 2010
4. M. Ye, P. Yin, and W. C. Lee, "Exploiting Geographical Influence for Collaborative Point-of-Interest Recommendation," in *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 325-334, 2011
5. S. Atisha and R. Vineet, "A Survey on Recommender Systems based on Collaborative Filtering Technique," *International Journal of Innovations in Engineering and Technology*, Vol. 2, No. 2, pp. 8-14, 2013
6. Z. Yu, C. Wang, and J. Bu, "Friend Recommendation with Content Spread Enhancement in Social Networks," *Information Sciences*, No. 309, pp.102-118, 2015
7. S. Huang, J. Zhang, and L. Wang. "Social Friend Recommendation based on Multiple Network Correlation," *IEEE transactions on multimedia*, Vol. 18, No. 2, pp. 287-299, 2016
8. C. He, D. Parra, and K. Verbert, "Interactive Recommender Systems: A Survey of the State of the Art and Future Research Challenges and Opportunities," *Expert Systems with Applications*, No. 56, pp. 9-27, 2016
9. P. Jyoti, J. Maitri, K. Abbas, and T. Malhar, "Recommendation System using Social Networking," *International Journal of Computer Science*, *Engineering and Information Technology*, Vol. 2, No. 5, pp. 45-54, 2012
10. M. Saveski and A. Mantrach, "Item Cold-Start Recommendations: Learning Local Collective Embedding," in *Proceedings of the 8th ACM Conference on Recommender Systems*, pp. 89-96, 2014
11. U. Ocepek, J. Rugelj, and Z. Bosnić, "Improving Matrix Factorization Recommendations for Examples in Cold Start," *Expert Systems with Applications*, Vol. 42, No. 19, pp. 6784-6794, 2015
12. F. Peng, X. Lu, and C. Ma, "Multi-Level Preference Regression for Cold-Start Recommendations," *International Journal of Machine Learning and Cybernetics*, pp. 1-14, 2017
13. Y. J. Yang, H. Z. Zhang, and X. F. Wang, "On Alleviation of New User Problem in Collaborative Filtering using SNA Theory," *International Journal of u- and e- Service*, *Science and Technology*, Vol. 6, No. 6, pp. 121-132, 2013
14. J. Lin, K. Sugiyama, M.Y. Kan, and T. S. Chua, "Addressing Cold-Start in App Recommendation: Latent User Models Constructed from Twitter Followers," in *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 283-292, 2013
15. M. Daoud, S. K. Naqvi, and T. Siddiqi, "An Item-Oriented Algorithm on Cold-start Problem in Recommendation System," *International Journal of Computer Applications*, Vol. 116, No. 11, pp. 19-24, 2015
16. C. Huang and J. Yin, "Effective Association Clusters Filtering to Cold-Start Recommendations," in *Proceedings of the 7th International Conference on Fuzzy Systems and Knowledge Discovery* (*FSKD*), pp. 2461-2464, 2010
17. M. K. Najafabadi, M. N. Mahrin, and S. Chuprat, "Improving the Accuracy of Collaborative Filtering Recommendations using Clustering and Association Rules Mining on Implicit Data," *Computers in Human Behavior*, No. 67, pp. 113-128, 2017
18. S. Gupta and S. Goel, "Handling User Cold Start Problem in Recommender Systems using Fuzzy Clustering," *Information and Communication Technology for Sustainable Development*, Springer, Singapore, pp.143-151, 2018
19. D. G. Ferrari and L. N. De Castro, "Clustering Algorithm Selection by Meta-Learning Systems: A New Distance-based Problem Characterization and Ranking Combination Methods," *Information Sciences*, No. 301, pp. 181-194, 2015
20. J. Wei, J. He, and K. Chen, "Collaborative Filtering and Deep Learning based Recommendation System for Cold Start Items," *Expert Systems with Applications*, No. 69, pp. 29-39, 2017
21. G. Roumelis, M. Vassilakopoulos, and A. Corral, "Efficient Query Processing on Large Spatial Databases: A Performance Study," *Journal of Systems and Software*, No. 132, pp. 165-185, 2017