CONSTRUCTING CORE USER SET FOR RECOMMENDATION BASED ON RELEVANCE OF INDIVIDUAL USERS

Gaofeng Cao, Li Kuang
School of software, Central South University
Changsha, China
caogaofeng@csu.edu.cn . kuangli@csu.edu.cn

Abstract

With the rapid explosion of information, more and more people find it difficult to get the information that they really require and are interested in quickly and effectively. Recommender systems have been developed to effectively help users to make decisions to a certain degree. However, existing researches on recommender system mainly focus on improving the accuracy of recommendation by designing new algorithms. In this paper, we propose a different idea to improve the efficiency, accuracy and novelty of recommendation by extracting core users who carry most of the information for recommendation. Specifically speaking, we propose new approaches to identifying core users based on interest similarity between users and users' own attributes. First, the interest similarity between all user pairs are calculated and sorted. The users' attributes are quantified and integrated. And then three strategies are used to select core users, which are frequency-based, rank-based, and fusion-sorting-based. In the experiment part, we compare our proposed methods with other existing methods from various perspectives, including accuracy, novelty, and long-tail distribution. Experiments show the effectiveness of our core user extraction method and prove that 20% of core users enable recommender systems to achieve more than 90% of the accuracy of the top-N recommendation.

Keywords: Core Users Extraction; Interest Similarity; User' Attribute; Recommender System;

1 INTRODUCTION

With the development of internet technology, recommender system is proposed and applied in many online platforms as an effective way to solve the problem of "information overload". It can provide more personalized services by predicting potential interests according to users' historical choices. Recommender systems have already been applied widely.

Recommender systems typically produce a list of recommendations in one of three ways: content-based (Peng, 2010), collaborative filtering (Chen, 2012; Li, 2009) and context-aware approaches (Wang, 2012; Yap, 2006). Different from the previous studies and approaches, the perspective of physics have also found applications in designing recommendation algorithms (Zhou, 2007; Zhang, 2007), and the most representative ones are mass diffusion and heat conduction.

Some latest study has proved that there exists a group of core users who carry most of the information for recommendation. With them, the recommender system can already generate satisfactory recommendation. It is meaningful to identify core users in recommendation since it can improve the efficiency as well as the robustness of recommender system by excluding irrelevant and unreliable users. However, the proposed approach to extracting core users are mostly based on out-degrees of nodes in graph theory, while the semantic meaning between nodes is not considered. Moreover, there exists data sparsity problem for

the proposed similarity-based approach, since the similarity of two users can be calculated if and only if they have common items.

In order to solve the problems, in this paper, we propose new approaches to identifying core users based on interest similarity between users, as well as two important attributes of individuals including one's popularity degree and resource. The identification of core users is realized through four steps: First, the interest similarity between all user pairs are calculated and sorted from high to low. In particular, for interest similarity, we propose to cluster items based on K-means and then calculate the interest similarity of two users based on similar items. Therefore, we can get the nearest neighbor list of each user with highest interest similarity. Second, the popularity degree and resource of each individual user is calculated. Third, based on the calculated neighbor lists and values, we propose three strategies to select core users. Fourth, we validate the effect of the extracted core users by various methods in recommendation.

The contributions of this paper include:

The paper first propose two strategies to extract core users based on interest similarity between users. By considering the semantic meaning between items, we improve the calculation of user similarity. Based on the nearest neighbor lists of each user with highest interest similarity, core users are selected according to the frequency of occurrence and the weight of rank in the lists.

The paper then propose a strategy to extract core users based on two important attributes of individuals, including one's popularity degree and resource. Popularity degree is combined with the resources to determine one's score, and core users are selected from the ones with highest scores.

In experiment, we compare the three methods as well as another two existing ones and make a comprehensive analysis of the results from the aspects of accuracy of recommendation, novelty and long tail distribution.

2 APPROACH TO EXTRACTING CORE USERS IN RECOMMENDATION

A recommender system consists of a set of users $\{u_1, u_2, \dots, u_m\}$ and items $\{o_1, o_2, \dots, o_n\}$, and can be represented by a bipartite network where users are connected to item they have read, purchased, collected, rated, commented and etc.

2.1 SIMILARITY DEGREE BETWEEN TWO USERS

When calculating user similarity based on cosine or Pearson formula through common item ratings, the effect of similar item is not considered. If two users do not have common ratings, the similarity degree of them cannot be calculated. It is obviously unreasonable.

When the rating data is sparse, there will be few ratings on the same items for two users. Obviously it is not accurate to decide whether the two users are similar in this case. In order to solve the problem, we combine item similarity and the traditional Pearson user similarity formula to improve the accuracy of user similarity.

First, we cluster the items using k-means algorithm. We use semantic similarity based on ontology (Sánchez, 2012) to calculate the similarity of two items.

We can then integrate the item similarity into the Pearson formula, as shown in formula (1):

$$\frac{\sum_{i=1}^n (r_{ui} - \bar{r}_u)(r_{vi} - \bar{r}_v) \sqrt{\frac{1}{n} \sum_{i=1}^n (r_{ui} - \bar{r}_u)^2} \sqrt{\frac{1}{n} \sum_{i=1}^n (r_{vi} - \bar{r}_v)^2}}{\sqrt{\sum_{i=1}^n (r_{ui} - \bar{r}_u)^2} \sqrt{\sum_{i=1}^n (r_{vi} - \bar{r}_v)^2}}$$

where r_{ui} is u_i 's rating to item i , $\bar{r}_u$ is u_i 's rating to item i , r_{vi} is v_i 's rating to item i , $\bar{r}_v$ is v_i 's rating to item i , $\sqrt{\frac{1}{n} \sum_{i=1}^n (r_{ui} - \bar{r}_u)^2}$ indicates the item similarity of u_i and v_i , and $\sqrt{\frac{1}{n} \sum_{i=1}^n (r_{vi} - \bar{r}_v)^2}$ is the average ratings of u_i and v_i are similar items, n represents the number of similar item pairs.

2.2 TWO ATTRIBUTES OF AN INDIVIDUAL USER

In this part, we define two attributes to measure the quantity of information carried by the user, which are popularity degree and resource.

2.2.1 Popularity degree of a user

The popularity degree of user is the total number of items which have been seen and rated by a user. Any user who has rated more items has greater chance to have common or similarity items with other users and so might be useful for recommendation of the uncommon items to them.

2.2.2 Resource of a user

Since recommender systems can be naturally represented by a user-item bipartite network, some classic network-based propagation processes such as mass diffusion, are applied to recommender systems, and it has been proved that the diversity-accuracy dilemma and recommendation accuracy for small-degree items can be effectively solved (Zhang, 2012).

We assume that a unit resource is assigned to each user, who distributes her/his resource to all neighboring items, and then each item redistributes the received resource to all connected users. Accordingly, the resource that user u_i has received from user u_j is

$$\frac{r_{uj}}{d_j} \times \frac{1}{d_i} \quad (2)$$

where d_j is the degree of u_j , namely the number of items that connected with u_j , d_i is the degree of u_i , namely the number of users who have rated u_i .

In this paper, we set the initial resource for each user is 1. All users reallocate resources according to formula (2) to get the final results as users' resources. For the sake of simplicity, Fig. 1 is a case of resource redistribution by one user.

Figure. 1 A user's resource redistribution

2.3 STRATEGIES FOR SELECTING CORE USERS

In this paper, we propose three strategies (Zeng, 2013; Cao, 2016) to select core users.

The first one is to select the users who appear the most times in all other users' nearest neighbor list, and we named it as frequency-based strategy. The detailed process is: First, we compute the top- N (e.g. $N=15, 20, 25$) neighbors of each user with highest similar interests based on formula (1). Second, for each user, count how many times he has appeared in other users' top- N lists. Third, the users who appear most frequently are selected as core users.

The second one is to select the users who have the highest weight of location in all the other users' nearest

neighbor lists and we name it as rank-based strategy. The strategy takes into account the ranks of each user in other users' top- N neighbor lists. Suppose user u_i belongs to user u_j 's top- N neighbor list and his position is the p_{th} , then the score of u_i is $1/p$. If user u_i also appears in other users' top- N neighbor lists, we sum the scores as his final weight:

$$w_i = \sum_{j \in \mathcal{N}(u_i)} \frac{1}{p_{ij}} \quad (3)$$

, where $\mathcal{N}(u_i)$ is the top- N neighbor list of user u_i whose neighbor lists contain u_i . p_{ij} is u_i 's position in u_j 's top- N neighbor list. Finally, the users with the largest weights will be selected as core users.

The third one is to combine user's popularity degree with users' resource to select core users, and we name it as resource-degree-sort-based method. First, all users are sorted in a descending order according to their popularity degrees, and we can get a List A_u . Second, all users are sorted in a descending order according to their resources, and we get a list named List B_u . But the order of a user in the two lists may be different, so we combine the two lists to rearrange users' order. The combination of the two lists is calculated as formula (4) shows:

$$C_u = \frac{A_u + B_u}{2} \quad (4)$$

Where N is the number of all users, p_{ij} indicates the position of a user in List A_u and List B_u , respectively. Finally, we can get a new List C_u based on the value of C_u . Then we select top- N users in List C_u as core users.

2.4 THE RECOMMENDATION ALGORITHM

We then validate the effect of the extracted core users in recommendation. We make use of network-based inference method (NBI for shot) (Zeng, 2011) as recommendation algorithm.

The MD method can be described in a more intuitive way: The initial resources placed on objects are first evenly divided among neighboring users and then evenly divided among those users' neighboring objects. In a real network, there can be a lot of neighboring users who have common objects with the target user. We argue here that only a few of the most relevant neighboring users should be taken into account in the diffusion.

For NBI based on core users, there will be a little difference. Given the set of core users C and the target user i , only the users in C will receive the resources from i , collected items in NBI. Other users will not receive resources even though they have same and similar item with i . Then the users who have received resources redistribute the resources back to the item side.

3 EXPERIMENT

In this section, we introduce the dataset used in the experiment, and analyze the experimental results.

3.1 DATA DESCRIPTION

We use the MovieLens dataset in experiments, which is a popular dataset used by researchers and developers in the field of recommendation. The dataset contains ratings from 943 users on 1,682 movies. The rating records are integers between 1 and 5.

In order to verify the accuracy of the recommendation algorithm, the data set is randomly divided into training set and test set according to the proportion of 8:2.

3.2 EVALUATION METRIC

In our experiments, we use the following evaluation indicators to analysis and comparison of our proposed three methods and the two methods in literature.

3.2.1 RECALL

For each user i , R_i is defined as the ratio of recall.

$$R_i = \frac{|\mathcal{R}_i \cap \mathcal{T}_i|}{|\mathcal{T}_i|} \quad (5)$$

where $|\mathcal{R}_i|$ is the number of movie that associate with user i in the test set, $|\mathcal{T}_i|$ is the number of the items in the test set which appear in the recommendation list with recommendation length L for target user i . A higher recall value indicates a higher accuracy of the recommendation algorithm.

3.2.2 NOVELTY

Novelty (Zhang, 2010) is the ability to recommend non-popular items to the user. In this paper we use the average degree of the recommended items to measure the novelty of the recommender system. The novelty index is calculated as formula (6) shows:

$$N = \frac{1}{L} \sum_{i=1}^L \frac{1}{d_i} \quad (6)$$

where d_i is the degree of an item, M is the number of users, L is the length of recommend list, $\mathcal{R}_i$ is the recommend list.

3.2.3 FREQUENCY OF THE RATED ITEMS

In our study, we use the frequency of the rated items (Morid, 2011) to measure the recommended ability of users, computed by equation (7):

$$F_i = \frac{1}{|\mathcal{R}_i|} \sum_{j \in \mathcal{R}_i} \frac{1}{r_j} \quad (7)$$

where r_j is the number of users who have rated item j , $|\mathcal{R}_i|$ is the number of items rated by user i , i.e., the degree of user i . This factor shows whether a user is interested in popular and high frequently rated items or non-popular and low frequently rated items.

3.3 EXPERIMENTAL RESULTS AND ANALYSIS

In this part, we compare our proposed three methods to extract core users, which are Similarity-Frequency-based, Similarity-Rank-based, and resource-degree-sort-based

methods. In addition, we compare our approaches to the similarity-based approach with the best performance mentioned in (Zeng, 2014) and best performance mentioned in (Zhang, 2016), where cosine distance based on common item ratings is used to calculate users' similarity. In order to differentiate our proposed users' similarity approach and theirs, we name theirs as Similarity-1 and ours as Similarity-2. In Figure 2 and Figure 3, the solid line shows the methods proposed in this paper, and the dotted line represents the methods proposed by others.

The results are presented in Figure 2 where P denotes the proportion of core users in all users. In the experiment, we set the nearest neighbors is top-10 and the recommendation list length is 15.

3.3.1 ANALYSIS ON THE RECALL OF EXTRACTING CORE USERS

Figure 2 The ratio of core users and the recommendation accuracy

3.3.2 ANALYSIS ON THE NOVELTY OF EXTRACTING CORE USERS

Figure 3 The ratio of core users and the novelty

3.3.3 ANALYSIS ON THE RECOMMENDED ABILITY OF EXTRACTING CORE USERS

Figure 4 Distribution Scheme of core user according the their Degree and AvgFreq Values

From the above experimental results, the accuracy is higher when we choose the Similarity-Rank-based method to identify core users. Furthermore, taking the Similarity-Rank-based method for MovieLens dataset, through statistical calculation, 90.39% (0.239/0.264) of the accuracy can be achieved when we only use 20% of core users ($P = 0.2$) and the recommendation accuracy tends to increase slowly with the increase of P . Meanwhile, the novelty is the best for Similarity 2-Rank-based and Similarity 2-Frequency-based methods. It is sufficient to explain the interest similarity plays a more important role than other factors in the process of recommendation. The result fits well with the idea of collaborative filtering, in which the interest similarity between users plays an important role in measuring user's relations.

From the long tail distribution, the core users selected by Resource-degree-sort-based method is relatively concentrated in the region where the user degree is greater than 80, i.e. the head of the distribution. However, the core users selected by other methods are relatively scattered.

Among the 5 methods, the distribution of user degree for core users selected by Similarity 2-Rank-based method is the most reasonable, since the proportions for the three

intervals are approximate to 50%, 30% and 20% respectively.

4 CONCLUSION AND FUTURE WORK

In this paper, we propose new approaches to identifying core users based on interest similarity between two users and users' own attributes. First, the interest similarity between all user pairs are calculated and sorted. The user attributes are quantified and integrated. Through the comparison of user similarity, popularity and resource, we find that user similarity plays a very important role in the recommendation system. And then three strategies are used to select core users, which are frequency-based, rank-based, and fusion-sorting-based. Using core users who carry more information not only produces satisfactory recommendations but also helps us to improve the efficiency of recommendation, and it also improves the robustness of the recommendation algorithm by eliminating irrelevant and unreliable users. The experimental results show that our proposed method can greatly reduce the computational complexity while maintaining the accuracy and diversity of the recommendation.

Future work include: (1) find more approaches to defining core users from different aspects, for example, fusing similar relationships and trust relationships, (2) find a relatively stable approach to generating core users, so as to reduce the frequency of updates, (3) Study the use of core users to solve the problem of user interest transfer and cold start.

ACKNOWLEDGMENT

The research is supported by "Natural Science Foundation of Hunan Province" (No.2016JJ3154), "National Natural Science Foundation of China" (No.61202095), "Scientific Research Project for Professors in Central South University, China" (No.904010001), and "Innovation Project for Graduate Students in Central South University" (No.502210017).

REFERENCES

- Chen, Y. W., Xia, X., Shi, Y. G. (2012). A collaborative filtering recommendation algorithm based on contents' genome[C]// Iet International Conference on Information Science and Control Engineering. 2012, pp.618-621.
- Cao, G. F., Kuang, L. (2016). Identifying Core Users Based on Trust Relationships and Interest Similarity in Recommender System. In Web Services (ICWS), 2016, pp. 284-291, doi:10.1109/ICWS.2016.44.
- Li, N., Li, C. (2009). Zero-Sum Reward and Punishment Collaborative Filtering Recommendation Algorithm[C]// Ieee/wic/acm International Joint Conferences on Web Intelligence and Intelligent Agent Technologies, 2009. Wi-Iat. IET, 2009, pp.548-551.
- Morid, M. A., Shajari, M., Golpayegani, A. H. (2011). Who are the most influential users in a recommender system?. International Conference on Electronic Commerce, pp.1-5.

Peng, T., Wang, W., Gong, X Y. (2010). A graph indexing approach for content-based recommendation system[C]//Multimedia and Information Technology (MMIT), 2010 Second International Conference on. IEEE, 2010, pp. 93-97.

Sánchez, D., Batet, M., Isern, D., Valls, A. (2012). Ontology-based semantic similarity: A new feature-based approach. Expert Systems with Applications, 39(9), pp.7718-7728.

Wang, L. C., Meng, X. W., Zhang, Y. J. (2012). Context-Aware Recommender Systems[J]. Journal of Software, 2012, 32(3), pp.217-253.

Yap, G. E., Tan, A. H., & Pang, H. H. (2006). Discovering Causal Dependencies in Mobile Context-Aware Recommenders. International Conference on Mobile Data Management (pp.4). IEEE Computer Society, doi:10.1109/MDM.2006.72

Zhou, T., Ren, J., Medo, M., Zhang, Y. C. (2007). Bipartite network projection and personal recommendation. Physical Review E Statistical Nonlinear & Soft Matter Physics, 76(2), pp.70-80.

Zhang, Y. C., Blattner, M., Yu, Y. K. (2007). Heat conduction process on community networks as a recommendation model. Physical Review Letters, 99(15), pp.12505-12508.

Zhang, F., Zeng, A. (2012). Improving information filtering via network manipulation. EPL (Europhysics Letters), 100(5), 58005, doi: 10.1209/0295-5075/100/58005

Zeng, W., Zeng, A., Liu, H., Shang, M. S., Zhou, T. (2014). Uncovering the information core in recommender systems. Scientific reports, Feb 2014, pp. 6140-6152, doi:10.1038/srepZeng,

Zeng, W., Zhu, Y. X., Lü L., Zhou, T. (2011). Negative ratings play a positive role in information filtering. Physica A Statistical Mechanics & Its Applications, 390(23-24), Nov 2011, pp. 4486-4493, doi:10.1016/j.physa.2011.07.005.

Zhang, Z. K., Liu, C., Zhang, Y. C., Zhou, T. (2010). Solving the cold-start problem in recommender systems with social tags. Epl, 92(2), pp.28002-28007.

Zhang, L., Yu, L., Cao, S. (2016). Constructing the Core User Set for Collaborative Recommendation Based on Samples Selection Idea[J]. International Journal of u-and e-Service, Science and Technology, 2016, 9(3), pp. 27-34.

Authors

Gaofeng Cao received his bachelor degree in Computer Science and Technology from Taiyuan University of Technology in 2015. He is currently a graduate student in the School of Software, Central South University. His research interests include

Recommender system, Machine Learning.

Li Kuang received her B.S. and Ph.D. degrees in computer science from Zhejiang University, Hangzhou, China, in 2004 and 2009 respectively. She is currently an associate professor at the school of software, Central South

University. Her research interests include Service Computing, Mobile Computing and Social Computing.