An Improved Collaborative Filtering Recommendation Algorithm Combining Item Clustering and Slope One Scheme

Haipeng You, Hui Li, Yunmin Wang, and Qingzhuang Zhao

***Abstract*—With the development of electronic commerce, a lot of recommendation systems have been developed. Collaborative filtering is one of widely-used algorithm in making rating prediction for recommendation systems. However, traditional collaborative filtering suffers sparisity, scalability and cold start problems, which result in poor quality in recommendation systems. To solve these problems, we propose a recommendation algorithm combining item clustering method and weighted slope one scheme. In our algorithm we use item clustering algorithm to partition items to several clusters and apply weighted slope one scheme in each cluster to predict ratings for unknown items to target user. We make experiments on the Movielens dataset and compare our algorithm with several recommendation algorithm. The results show that our algorithm can improve the accuracy of the collaborative filtering recommendation system.**

***Index Terms*—collaborative filtering, slope one scheme, clustering algorithm, data mining**

I. Introduction

At present, with the growth of the Internet, information overload is hard to deal with. To solve this problem, recommendation system [3] has been widely used in social network applications, electronic commerce and online-video. Large recommendation system has been exist in many electronic commerce sites, such as Facebook, Amazon, YouTube, etc. In the field of recommendation system, many algorithms were proposed, such as content-based recommendation, collaborative filtering (CF) algorithms, data mining, knowledge-based recommendation and mixed recommendation. Collaborative filtering is the most mature and widely used recommendation algorithm [1]. Traditional Traditional CF methods include user-based CF and item-based CF. Compared to other recommendation algorithms, CF algorithms have a lot of merits, for example, simple to algorithm, easy to implement, good to result. But, CF algorithms have many challenges as following:

Manuscript received January 8, 2015; revised January 17, 2015. This work is supported by the National Basic Research Program of China (No. 2012CB315904), the National Natural Science Foundation of China under Granted NSFC61179028, the Basic Research of Shenzhen (No. JCYJ20130331144502026 and No. JCYJ20140417144423192), and the Natural Science Foundation of Guangdong Province under Granted NSFGD S2013020012822.

Haipeng You, Hui Li, Yunmin Wang and Qingzhuang Zhao are with the ShenZhen Engineering Lab of Converged Networks Technology, Shenzhen Graduate School, Peking University, Shenzhen, 518055, China (e-mail: kflkyhp@163.com, lih64@pkusz.edu.cn, david@hidoop.com, qingzhuang1991@163.com).

1. Sparsity: Data sparse problem is a serious issue of collaborative filtering and has been widely researched by experts. The recommendations of CF based on the similarities users or similarities items. But, many users don't have ratings to the same item, the user-item matrix will be sparse. Without enough ratings, the effectiveness of prediction to users will be greatly reduced.

2. Scalability: To ensure the real-time recommendation, we must solve the scalability problem. With the increase of users and items, the scalability problem of CF algorithm become an important factor for recommendation system. If the problem is not solved, it is hard to recommend in real time.

3. Cold start: The problem always exists in the circumstance where new items are added and new items are rated by only few users. Because of the collaborative filtering algorithms based on the similarity of users or items. A new item can't be recommended until it has been rated by users.

The slope one scheme is one of rating-based collaborative filtering algorithm, but it don't calculate the similarities between items. It adopts an easy but effective method that a simple linear regression model to predict ratings. The slope one scheme mainly consider the users rated the target item and the other items rated by the target user, and use the ratings of the users to predict the rating of the target item. However, if there are no users or few users have rated the target item, the accuracy of the algorithm will reduce. The algorithm is simple, efficient, easy to implement. But, the slope one scheme also suffers from both cold start and sparsity problems.

In this paper, we propose an improved recommendation algorithm that combines item clustering and weighted slope one scheme. The method use Density Based Spatial Clustering of Applications with Noise (DBSCAN) [5] clustering algorithm to partition the set of items into several clusters, and then we use weighted slope one scheme to get the prediction rating of target item for user based on the target user's ratings to other items of the cluster. In addition, experiments on the Movielens dataset show that our algorithm can help to increase the accuracy of recommendation system and it is also more robust to noise.

The rest of this paper is organized as follows. We provide some related work in the next section. Section III, explicitly describes the proposed algorithm. Section IV, we make an experiment to evaluate the algorithm. Finally, Section V concludes our work and presents future work.

ISBN: 978-988-19253-2-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

II. Related Work

Collaborative filtering is the most popular technique in recommendation systems. CF methods generally include model-based CF and memory-based CF. Memory-based CF always uses a similarity measure between users through ratings which grade by users to prediction [9], [11]. The similarity measure methods have Pearson's correlation, Euclidean distance, Cosine Similarity, Adjusted cosine similarity that have been studied for twenty years. The accuracy of the prediction is determined by the selected similarity measure. The drawbacks of memory-based CF include data sparisity, cold start, and scalability to new items. There are many model-based approaches to CF. PCA, SVD are based on algebra [12], [13]; Bayes methods are based on statistics [8]. On the real recommendation system, the memory-based CF is hard to real-time recommend because all compute is online. The model-based CF using offline compute prefer to memory-based CF.

The slope one scheme is one of model-based CF algorithm, and is studied for many years. In [2], the slope one scheme was first proposed, and it is simple, easy to implement and maintain, updateable on the fly. In [6], the authors apply slope one scheme to fill the vacant ratings of the user-item matrix where necessary, and use memory-based CF to produce recommendation. Zhang et al [7] introduce a novel user similarity measure based on user favorite items and apply it into slope one scheme. In the algorithm, the authors use user similarity to calculate target rating that predict by one item on all rated users, then use item similarity on summation. In [10], the authors think that users have different attitudes at different time. So the authors add a time weight on slope one scheme. The time near now, the weight high.

In this paper, we propose an improved recommendation algorithm that combines clustering and weighted slope one scheme. The DBSCAN clustering algorithm is suitable for any shape of cluster.

III. Algorithms

Figure 1 is the overview of the process of our algorithm. In the algorithm, firstly, we should have a user-item rating matrix. Secondly, users were clustered based on ratings of users. Thirdly, we use weighted slope one scheme to predict the ratings of the target users.

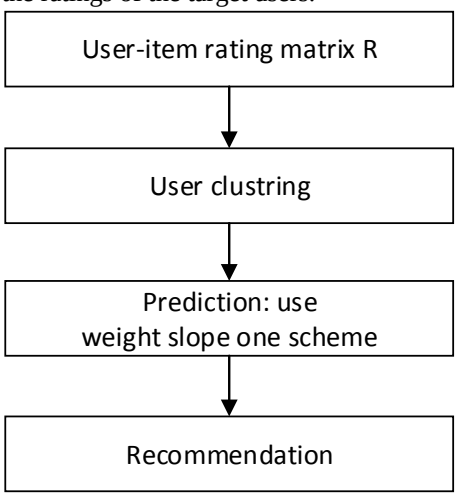

Fig. 1. The mainly process of our algorithm

A. Notation

In this paper, we use the following notations to describe algorithms. Suppose that there are m users and n items in the recommendation. The rating table is a $m \times n$ matrix which is indicated by R. In the matrix R, we use $U = \{u_1, u_2, u_3, \mathrm{L}\ u_m\}$ to indicate the set of users and $I = \{i_1, i_2, i_3, \mathrm{L}\ i_n\}$ to indicate the set of items. $r_{u,i}$ indicates the rating of user u to item i. $S(u)$ is the set of ratings of user u, $|S(u)|$ is the size of $S(u)$. $S(i, j)$ indicates the set of all user u that $S(a)$ contains both $r_{u,i}$ and $r_{u,j}$. $|S(i, j)|$ is the size of $S(i, j)$. $Num_{i,j}$ is the number of users that rate both of the two items. In the process of the slope one scheme, we need to calculate the average deviation matrix firstly. The average deviation is represented by $\{dev_{ij}\}$, and dev_{ij} indicates the average deviation of item i with respect to item j. $P(a, j)$ indicates the predication rating of user u to item j.

B. Slope One Scheme

The slope one scheme is a rating-based recommendation algorithm [2], it takes advantage of the linear relationship between items to get the deviation matrix whose values is item-item average difference. Formally the predictor is based on a simplified regression model: $f(x) = x + b$ where b is defined as the mean deviation. In many environment it is more accurate and rapid than the linear regression of $f(x) = ax + b$.

The prediction process of slope one scheme consists of two steps: firstly, calculate the average deviation dev_{ij} of the target i with other item j. Secondly, predict the rating $P(a, j)$ of the user a on the target item j.

1. Calculate the average deviation matrix $\{dev_{ij}\}$. Use the training data set and (1) to calculate the value dev_{ij} between every item i and every item.

$$dev_{ij} = \frac{\sum_{n \in S(i,j)} (r_{u,i} - r_{u,j})}{|S(i, j)|} \tag{1}$$

In (1), user u rates both item i and item j. The deviation matrix dev_{ij} can be computed once and updated quickly when new item is entered. Particularly, the function updated quickly is very important. As a real-time recommendation system, it is terrible to calculate all the data afresh. If the system can updated quickly data, the system can save many time and hardware facilities.

2. Predict the unknown rating $P(a,i)$ which means target user a to item i rating. We can use deviation matrix and the rating of target user a to item i to compute the prediction rating in (2):

$$P(a,i) = \frac{\sum_{j \in S(u)} (dev_{ij} + r_{a,j})}{|S(u)|} \tag{2}$$

Formula 2 is the basic slope one scheme. It suppose that the

ISBN: 978-988-19253-2-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

relevant items to items i play the same important role for predicting $P(a,i)$. However, different user have different ratings to items, the user that has many ratings has more influence on producing the predictions. To make up the drawback, the weighted slope one revises (2) by taking the number of ratings into consideration.

$$P(a,i)=\frac{\sum_{i\in r_u}(dev_{ij}+r_{a,j})\times Num_{i,j}}{\sum_{i\in r_u}Num_{i,j}} \quad (3)$$

C. Clustering Algorithm

There are many clustering algorithms. In most situations, people often use clustering method to pre-process the data. Clustering is a process of dividing data into different clusters and put similar data elements into same cluster. DBSCAN is a well-known algorithm for density based clustering because it can identify the groups of arbitrary shapes and deal with noisy datasets [5]. In this paper, we use the DBSCAN method. The implementation of DBSCAN is shown in the algorithm1.

Algorithm 1 DBACAN

Input: A data set containing *n* object;
the radius parameter *r*;
the neighborhood density threshold *MinPs.*
Output: A set of density-based clusters.
1. All objects are marked as unassigned.
2. Randomly select a unassigned object *p*, mark *p* assigned; if the *r*-neighborhood of *p* has at least *MinPs* objects, create a new cluster *C* and add *p* to cluster *C.*
3. Let *N* be the set of objects in the *r*-neighborhood of *p;* all points *p'* are marked as unassigned in *N*, we mark *p'* as assigned; if the *r*-neighborhood of *p'* has at least *MinPs* points, add those points in *N*; if *p'* is not a member of any cluster, add to *C.*
4. Repeated 2 and 3 until all points are assigned as "assigned".

D. Fusion Method

Always, there are many unrelated items in the dataset. The slope one scheme use all the items containing some unrelated items that may decrease the prediction accuracy. To filter the noise in slope one scheme, we firstly use clustering method to find out the related items, and use the slope one scheme on the dataset that composed by related items.

Algorithm 2 The algorithm combing clustering method and slope one scheme

Input: The training data set $m\times n$ matrix D;
the value *r*;
the value *MinPs.*
Output: The prediction value.
1. Apply the DBSCAN clustering algorithm 1 to produce several partitions. Formally, the data set D is divided into D_1, D_2, D_3, $\cdots$, D_n (the number of based on the radius parameter *r* and the neighborhood density threshold *MinPs*).
2. For each partitions, we apply the formula 2 to compute average deviation, and we get a deviation matrix.
3. After we get the deviation matrix, we use formula 3 weighted slope one scheme to compute the prediction ratings for every unknown item *i* on the cluster D_j.

IV. Experiments and Results

In this section, we describe the data sets, the evaluation metrics and the comparative experiments between our approach and the collaborative filtering and basic slope one scheme and the weighted slope one scheme. At last, we come up with the experiments results and make an analysis of the results.

A. Data Set

In the experiments we adopt Movielens 1M dataset that contains 1,000,209 ratings for 3900 movies by 6040 users. And, the ratings value ranges from 1 to 5. The sparsity level is $1-1000209\div(3900\times 6040)=0.958$, the dataset is highly sparse. We randomly selected 70% ratings as training datasets and the rest as test datasets five times. The training datasets are respectively named as train1, train2, train3, train4 and train5, and the test datasets are named as test1, test2, test3, test4 and test5.

B. Evaluation Metrics

To measure the accuracy of the algorithm, we used Mean Absolute Error (MAE) metric [4]. MAE is the most commonly used and easiest to metric. Formally, MAE is defined as the following equation:

$$MAE=\frac{\sum_{i=1}^{n}|p_i-q_i|}{n} \quad (4)$$

Where, n indicates the total number of ratings, p_i and q_i are actual and predicted ratings respectively. A lower MAE value indicates better prediction performance.

C. Experimental Results

In order to test the prediction performance of our algorithm, we compare our algorithm to the user-based CF algorithm, item-based CF algorithm, slope one scheme and weighted slope one scheme. All our experiments algorithms were coded in C++ and ran our experiments on Linux which was installed on a typical PC with Intel Core i5-3470 and 8G of RAM.

The clustering algorithm DBSCAN should input two parameter *r* and *MinPs.* The parameter *r* and *MinPs* is related with the change of the k-neighborhood distance. We give the different value to the two parameters many times. We find that *r = 12* and *MinPs = 5* is the better parameter.

We compared our algorithm with slope one scheme, weighted slope one scheme (W-Slope One), user-based collaborative filtering (U-CF) and item-based collaborative filtering (I-CF). The U-CF adopt Pearson as similarity measure and I-CF adopt Cosine as similarity measure. The results were presented in Fig. 2. The MAE is the average of five time experiments MAE results. From Fig. 2, it can be observed that our algorithm has achieved better quality of prediction than other recommendation algorithm.

ISBN: 978-988-19253-2-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

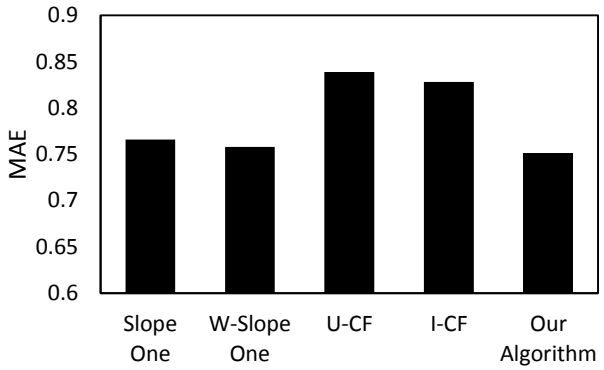

Fig. 2. The MAE of comparative experiments

V. Conclusions and Future Work

In this paper, we proposed a new approach to improve the quality of collaborative filtering recommendation systems. The algorithm combines item clustering and weighted slope one scheme. We compare our approach to other algorithms on the Movielens dataset. The results suggest our algorithm produces better result. In the future, we should to introduce other automatic method to determine the two parameters of DBSCAN. And we also can use more large data sets to evaluate the availability of the algorithm.

References

[1] D. Goldberg, D. Nichols, B. M. Oki and D. Terry, "Using collaborative filtering to weave an information tapestry", Communications of the ACM. vol. 35, no. 12, pp. 61-70, 1992.

[2] D. Lemire and A. Maclachlan, "Slope one predictors for Online Rating-based Collaborative Filtering", Society for Industrial Mathematic, California, USA, 2005, pp. 76-80.

[3] G. Adomavicius and A, Tuzlin, "Toward the next generation of recommender systems", IEEE Trans Knowledge and Data Engineering. vol. 17, no. 6, pp. 734-749, 2005.

[4] J. Herlocker, J. Konstan, L. Terveen and J. Riedl, "Evaluating collaborative filtering recommder systems", ACM Transactions on Information. vol. 22, no. 1, Jan. 2004.

[5] K. Khan, S. U. Rehman, K. Aziz, S. Fong and S. Sarasvady, "DBSCAN: Past, present and future", 2014 Fifth International Conference on the Applications of Digital Information and Web Technologies, Bangalore, India, 2014, pp. 232-238.

[6] Wang Pu and Hongwu Ye, "A Personalized Recommendation Algorithm Combining Slope One Scheme and User Based Collaborative Filtering," Proceedings of the IEEE International Conference on Industrial and Information Systems, pp. 152-154, 2009.

[7] Ziqing Zhang, Xinhuai Tang, Delai Chen, Applying User-Favorite-Item-Based similarity into Slope One Scheme for Collaborative Filtering, 2014 World Congress on Computing and Communication Technologies, pp. 5-7, 2014.

[8] J. S. Breese, D. Heckerman, and C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In fourteenth conference on uncertainty in AI. Morgan Kaufmann, pp. 43-52, July 1998.

[9] G. Mdomavicius, A. Tuzhilin: Toward the Next Generation of Recommender Systems: A Survey of the State-of-the art and possible extensions, IEEE Trans. on Knowledge and Data engine, vol. 17, no. 6, pp. 734-739, 2005.

[10] Tongqiang Jiang, Wei Lu, Improved Slope One Algorithm Based On Time Weight, Proceeding of the 2th International Conference on computer Science and Electronics Engineering, pp. 2295-2297. 2013.

[11] S.M. Weiss and N. Indurkhya. Lightweight collaborative filtering method for binary encoded data. Proceedings of the Fifth European Conference on Principles and Practice of Knowledge Discovery in Databases, pp. 484-491, 2001.

[12] B.M. Sarwar, G. Karypis, J.Konstan, and J. Riedl. Incremental svd-based algorithms for highly scaleable recommender systems. Fifth International Conference on Computer and Information Science, 2002.

[13] K. Honda, N. Sugiura, H. Ichihashi, and S. Araki. collaborative filtering using principal component analysis and fuzzy clustering. In Web Intelligence, number 2198 in Lecture Notes in Artificial Intelligence, pp. 394-402, 2001.

ISBN: 978-988-19253-2-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)