Rating-Based Collaborative Filtering Using Spectral Clustering Algorithm

Yongjie Yan, Hui Xie* and Li Ma

School of Mathematics and Computer Science, Jiangxi Science and Technology Normal University Nanchang 330038, China

*Corresponding author: huixie@aliyun.com

Abstract. Clustering analysis has been an important area of machine learning and data mining research, it can help us know the connection between things more clearly. In recent years, the research of spectral clustering algorithm has been a new and efficient clustering analysis algorithm. In this paper, the sparsity and the real-time problem of traditional recommendation algorithms, a new recommendation algorithm based on spectral clustering is proposed. The spectral clustering process can improve the efficiency of spectral clustering algorithm. Spectral clustering can be performed offline, which will accelerate the speed of online recommendation. The experimental results on Movie lens show that the new algorithm improves recommendation quality in MAE and coverage.

1. Introduction

As an effective way to solve the problem of “information overload” in the Internet, recommendation system is widely used in the field of e-commerce. On the one hand, it can help users find products of interest, improve their online shopping experience, on the other hand, it can improve the sales of e-commerce websites. However, due to the rapid growth of the number of users and commodities in the e-commerce system, the traditional recommendation system is faced with low prediction accuracy, sparsity, real-time and other problems, and the recommendation quality is difficult to guarantee.

With the rise of E-commerce, how to use social network to improve the performance of traditional recommendation system has become a research hotspot in recent years. Personalized recommendation based on social network is also called social recommendation. It is based on the information of users in social network to generate recommendations for users. Literature [1] compares online recommendation with friend based recommendation, and the results show that friend based recommendation is better than online recommendation. The research of literature [2] on the influence of social background on recommendation confirms that people prefer recommendation from acquaintances. The above research provides an important theoretical basis for social recommendation. On this basis, researchers [3-8] have studied the social recommendation algorithm, combining the social network with the traditional recommendation algorithm, improving the prediction accuracy of the recommendation system. However, the massive user information in social network seriously affects the performance of recommendation system, and social recommendation still faces the problem of sparsity and real-time. For this reason, Zhang [9] applied clustering method to social...
recommendation system, and achieved good results. However, the traditional clustering method is sensitive to the spatial distribution of data sets, and the clustering effect is not ideal.

Recommendation technology analyzes the interest model from the user's history information. Then, based on the user's interest model, the recommendation system can choose the appropriate items to recommend to the user. This initiative behavior that recommendation system selects items from huge amounts of data for the user, can greatly improve the users' experience, so the study of recommendation technology has a certain practical significance.

Spectral clustering algorithm is a clustering algorithm based on graph theory [10, 11]. Compared with the traditional clustering algorithm, it is not only easy to realize, but also can recognize the sample space of any shape, and can converge to the global optimal solution. Spectral clustering algorithm is often used in image segmentation, but it is relatively less used in recommendation system.

Collaborative filtering recommendation technology is one of the most successful technologies in the recommendation system. It began to research and promote the prosperity of the whole recommendation system research in the 1990s. A large number of papers and research belong to this category. Collaborative filtering is a widely used and successful recommendation paradigm that model users' collaborative behaviors reflected in previous transactions for recommendation.

2. Recommendation System Based On Spectral Clustering Algorithm

2.1. Classic Collaborative Filtering

Collaborative filtering algorithm is one of the most widely used and effective algorithms in recommendation system. Most of the existing social recommendation algorithms are based on collaborative filtering technology. The basic idea of traditional collaborative filtering algorithm is to find a set of users similar to the target user, and use the characteristics of similar user set to predict the characteristics of the target user. The input data is the user rating matrix, which is recorded as \( r(m, n) \), as shown in Table 1, where \( r_{i,j} \) represents the rating of user \( i \) on project \( j \).

| User   | Item 1 | ... | Item j | ... | Item n |
|--------|--------|-----|--------|-----|--------|
| User1  | \( R_{1,1} \) | ... | \( R_{1,j} \) | ... | \( R_{1,n} \) |
| ...    | ...    | ... | ...    | ... | ...    |
| User i | \( R_{i,1} \) | ... | \( R_{i,j} \) | ... | \( R_{i,n} \) |
| ...    | ...    | ... | ...    | ... | ...    |
| User m | \( R_{m,1} \) | ... | \( R_{m,j} \) | ... | \( R_{m,n} \) |

Table 1. User rating matrix

A classic Collaborative Filtering (CF) is described in detail in [12]. In the classic CF model, the similarity between two users is calculated using the Pearson correlation over the ratings of their common items. The formula for the Pearson correlation is the following, stated in [12]:

\[
userSim(u, v) = \frac{\sum_{i \in CR_u,n} (r_{ui} - \bar{r}_u)(r_{vi} - \bar{r}_v)}{\sqrt{\sum_{i \in CR_u,n} (r_{ui} - \bar{r}_u)^2} \sqrt{\sum_{i \in CR_u,n} (r_{vi} - \bar{r}_v)^2}} \tag{1}
\]

In the formula, \( r \) stands for rating, \( u \) denotes the center user and \( n \) a neighbor. \( CR_{u,n} \) denotes the set of co-rated items between \( u \) and \( n \). After performing this calculation, we select the top ten most similar users. Next, we rank the articles of these users to recommend to the center user, using the formula of predicted rating for user \( u \) with average adjusts described in [12]:

2
In general, the CF method makes the product of a previously successful search available to other users. A user with low personal experience/knowledge can benefit from the prior experience of others. However, there are a few possible ways the collaborative filtering results can be suggested to the user (e.g., including the user’s own search results vs just the recommended results, no user’s own results), and there are different types of users, with knowledge at different levels and in different subject domains. Which way or combination of ways is effective for what type of user and for which subject domain(s) has not been investigated.

2.2. Spectral Clustering Algorithm (SCA) Model

Spectral Clustering Algorithm (SCA) is a crucial algorithm for learning on large data sources. We will analyse Ng et al [13] spectral clustering paper from NIPS 2001, which proposes a concise algorithm and provides a theoretical backing for its choices. This paper reduces the problem to simple k-means clustering in the eigenvector space of the Palladian matrix of the data, provides theoretical bounds relating the clusters in this space to clusters in the data space, and shows initial experimental results using this algorithm.

Spectral clustering aims to solve the problem of grouping large amounts of high dimensional data into a small number of clusters. In the absence of labels for data, clustering can be extremely useful for data analysis, visualization, and unsupervised learning. We desire a clustering algorithm that takes as input the following:

**Ideal input:**

\[ X : \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^{n \times k} \]

\( s(x_i, x_j) \): A real valued similarity.

**Ideal output:**

\( C : \{C_i, C_{i_2}, \ldots, C_{i_k}\}; i \in \{1, \ldots, k\} \)

\( k \): Number of clusters

The Ng, Jordan, Weiss (NJW) algorithm [13] presents a spectral approach to clustering. It takes as input \( X \) as defined in the ideal input, and the following:

\( k \): The number of clusters expected

\( \sigma \): A scaling parameter for the similarity function.

The similarity metric is pre-defined as follows:

\[ s(x_i, x_j) = \exp\left\{ \frac{\|x_i - x_j\|^2}{2\sigma^2} \right\} \]  

The algorithm is as follows:

1. Construct the matrix \( A_y \) such that

\[ A_y = \begin{cases} s(x_i, x_j) & \text{if } i \neq j \\ 0 & \text{otherwise} \end{cases} \]
2. Let $D$ be $\text{diag}(A)$. Let $A = D^{-1/2}AD^{-1/2}$. (Normalize $A$.)

3. Find the $k$ largest eigenvectors of $A$ : $v_1, \ldots, v_k$, and form the matrix $V = [v_1, \ldots, v_k]$.

4. Form the matrix $Y$ by normalizing the rows of $X$ to have unit length:
   
   $Y_{ij} = X_{ij} / \sum_j X_{ij}$.

5. Treat $Y$ as a data matrix, so that $Y = [y_1, \ldots, y_n]^T$, and cluster these rows of $Y$ to obtain the cluster assignments for $y_i : C_k', C'_1, \ldots, C'_k$.

6. Let $C_k = C'_k$; that is, assign $x_i, x_j$ to the same cluster, iff $y_i, y_j$ were assigned to the same cluster.

3. Dataset And Metrics

3.1. Dataset

In the experiment, we used the data of Movielens. Movielens is a web-based research recommendation system, which was released in the autumn of 1977. The Movie Lens academic dataset provides ratings of the 6,040 users to 3,187 movies (figure 1). The Whole Training Set (WTS) available to students includes 801,051 (4.15%) ratings, user meta-data (gender, age, and occupation), and movie meta-data (name and genres). Ratings are integers between 1 to 5.

![Figure 1. The structure of Movie Lens dataset.](image)

We model users and movies as nodes, with ratings as directed edges. The distributions of some statistics about this dataset are shown in figure 2-4. The ratings tend to more positive than negative (figure 2).
Figure 2. Distribution of movie average rating.

An important fact about this dataset is that the variance of average rating per movie (Figure 3), which suggests more information is hidden under item-item pairs.

Figure 3. Distribution of rating counts per movie.

3.2. Metrics

The accuracy of recommendation system is the most basic index. The ultimate goal of the improved collaborative recommendation algorithm is to improve the accuracy of the results of this paper, so we mainly consider the accuracy of the algorithm. In order to evaluate the recommendation accuracy of the improved recommendation algorithm, root mean square error (RMSE) and mean absolute error (MAE) are used to measure the effect of the recommendation system. Mae and RMSE are measures of the deviation between the recommended value and the real value specified by the user. RMSE and Mae values can be obtained by calculating the score deviation between the actual score and the predicted score between users. The lower the RMSE and Mae values are, the higher the accuracy of the proposed algorithm is. Formally,
Here, $p_i$ is the predicted score to target user on the project $i_j$, $N$ is the count of projects predicted and $r_i$ is the real score to target user on the project $i_j$.

### 4. Experimental Result

In the movie scoring prediction problem, each training sample contains three types of information: user ID, movie ID and scoring data. We need to design a model starting from user ID and movie ID. The correlation between them is found in the training data, and the score data is predicted. To verify the superiority of the improved collaborative recommendation algorithm, we make the following experimental design that compared the traditional collaborative filtering recommendation algorithm. Slope One algorithm is a project-based collaborative filtering recommendation algorithm proposed by Lemire et al. In 2005 [14]. In practice, this algorithm has the advantages of simple calculation, easy implementation and maintenance, providing recommendation services for new users, efficient query corresponding, reasonable accuracy, etc. It generates prediction by calculating the deviation between users’ projects. Trust Walker, a recommendation algorithm based on random walk model, is proposed in literature [15].

We implemented the spectral clustering algorithm to obtain a low-rank representation adapted to Movielens sparse data. Our implementation could benefit from easy parallelization. To select the best value of the parameter $\sigma$ described in equation (3) as the target reduced rank, we run the algorithm with each candidate value over 5 random train / test splits of our dataset. The spectral clustering algorithm can fit the training set arbitrarily well, but we were not able to reproduce the good generalization results achieved on the Movielens dataset. For most values of the hyper parameters, we obtained a test RMSE of 2.1 and more. However, we re-used this algorithm in subsequent methods as an efficient dimensionality-reduction tool. Figure 4 gives a Results of RMSE value with different k number of clusters.

Figure 4. Results of RMSE value with different k number of clusters.
As can be seen from figure 5, the RMSE of Spectral Clustering Algorithm (SCA) model recommendation results is lower than the other two algorithms.

![Figure 5](image-url)  
**Figure 5.** Results of RMSE value on Movie Lens data set.

5. Conclusion
The user spectrum clustering can be done offline, which greatly reduces the execution time of the online algorithm and improves the real-time performance of the algorithm. The experimental results on the Movie Lens dataset show that the algorithm proposed in this paper has greatly improved MAE and RMSE, and greatly improved the performance of the recommendation system. However, after the offline clustering stage, the user category and the number of clusters are determined and are not easy to change. In practical applications, the number of users, user behavior and user relationship are constantly changing. Therefore, how to adjust the user category and the optimal number of clusters according to these changes is worthy of further study.

Acknowledgments
This work was sponsored by the National Natural Science Foundation of China (71561013), the Fund of Humanities and Social Sciences in Universities of Jiangxi Province (JC17221, JD18083), Science and technology research project of Jiangxi Provincial Department of Education (No.170678).

References
[1] Sinha R, Swearingen K. Comparing Recommendations made by online systems and friends. Processings of DELOS Workshop: Personalisation and Recommender Systems in Digital Libraries, 2001.
[2] Bonhard P, Sasse M A. 'konwing me, knowing you'-using profiles and social networking to improve recommender system. BT Technology Journal, 2006, 24 (3): 84 - 98.
[3] Lam C. SNACK: Incorporating Social Network Information in Automated Collaborative Filtering. The fifth ACM Conference on Electronic Commerce, 2004: 254 - 255.
[4] Konstas L, Stathopoulos V, Jose J M. On Social Networks and Collaborative Recommendation. Proceedings of the 32nd international ACM SIGIR conference on research and development in information retrieval, 2009: 195 - 202.
[5] Symeonidis P, Tiakas E, Manolopoulos Y. Product recommendation and rating prediction based on multi-modal social networks. Proceedings of the 5th ACM Conference on Recommender Systems, New York, 2011: 61 - 68.
[6] Zhang Juan, Hu Xuegang, Zhang Yuhong, et al. An efficient Li Bin. Cross-domain Collaborative Filtering: A Brief Survey. Proceedings of the 23rd International Conference on
Tools with Artificial Intelligence. [S. l. ]:IEEE Press,2011:1085 - 1086.

[7] Ning X, Karypis G. Multi-task Learning for Recommendation System. Proceedings of the 2nd Asian Conference on Machine Learning. Tokyo,Japan,2010:269 - 284.

[8] Tsai C F, Hung C. Cluster ensembles in collaborative filtering recommendation. Applied Soft Computing, 2012, 12 (4): 1417 - 1425.

[9] Zhang Juan, Hu Xuegang, Zhang Yuhong, et al. An efficient ensemble method for classifying skewed data streams. Proceedings of the 7th International Conference on Intelligent Computing: Bio-inspired Computing and Applications. 2011: 144 - 151.

[10] Lyzinski V, Sussman D L, Fishkind D E, et al. Spectral clustering for divide-and-conquer graph matching. Parallel Computing, 2015, 47 (3): 70 - 87.

[11] Cribben I, Yu Y. Estimating whole-brain dynamics by using spectral clustering. Journal of the Royal Statistical Society Series C, 2017, 66 (3): 607 - 627.

[12] Schafer, J., Frankowski, D., Herlocker, J. and Sen, S. Collaborative Filtering Recommendation Systems. The Adaptive Web. 2007, 291 - 324.

[13] Ng A Y, Jordan M I, Weiss Y. On Spectral Clustering: Analysis and an Algorithm. Adv. Neural Inf. Proc. Syst. 14. MIT Press, 2001.

[14] Lemire D, Maclachlan A. Slope One Predictors for Online Rating-Based Collaborative Filtering. In: Proceedings of the Fifth SIAM International Conference on Data Mining, 2005, 471 - 480.

[15] Jamali M, Ester M. TrustWalker: a random walk model for combining trust-based and item-based recommendation. In: Pro of KDD’09, New York: ACM Press, 2009, 397 - 405.