Journal of Physics: Conference Series

Improving the Cold Start Problem in Music Recommender Systems

Improving the Cold Start Problem in Music Recommender Systems

KeYin Cao[1a], Yu Liu[1,2*] and HuaXin Zhang[1]

[1]College of Information Science and Engineering, Guilin University of Technology, Guilin, Guangxi, 541006, China

[2]Guangxi key Laboratory of Embedded Technology and Intelligent System

[a]colin_95@glut.edu.cn [*]Corresponding author's e-mail: Lewis_5709@163.com

Abstract. The music recommender system can help us extract valuable music from a huge amount of raw information, due to the increasingly severe data sparsity and cold start problems in music recommendation system, the recommendation results will be inaccurate. But the traditional algorithm cannot effectively solve these problems; the existing improved algorithm still requires specific parameters in advance due to its poor stability. In this paper, a proposition for a cold start based on community detection algorithm is proposed. By projecting the bipartite network, calculating the similarity between the User and the Item, the Louvain algorithm is used to perform community detection on the projected One-mode network, so that the new record can be updated to the original community, and then the user group is recommended for music. With higher accuracy, solid stability, and shorter running time (when compared with other ease cold start algorithms), the algorithm in the larger community can be safely applied to the music recommendation system or prediction field.

1. Introduction

Music recommendation systems have been more frequently applied in online music streaming services (e.g., Spotify, 163Music, Apple Music). The recommendation system is a tool designed to simplify user decision-making. It automatically recommends songs by mining user behavior data to help users browse a large amount of music [1]. They are mainly divided into two categories: content-based filtering and collaborative filtering. With the emergence of a large amount of information, the cold start challenge brought by the problem of data sparseness or insufficient new records is getting worse. Insufficient user historical behavior records, or lack of access records for songs, the recommendation results will be inaccurate, and even affect the opportunity to recommend other songs and resulting in negative feedback, which called cold start problem.

To solve the cold start problem in the recommendation system, many methods have been proposed. Generally, existing collaborative filtering methods are classified into neighbor-based collaborative filtering and model-based collaborative filtering, in which the neighbor-based method uses a common experience group to generate recommendations that may be of interest to target users. Since the calculation of similarity (between users or projects) is a key component of a neighbor-based approach, many similarity measures have been designed. These include the Pearson correlation coefficient (PCC), cosine similarity [2], and some other heuristic similarity measures [3], which apply domain-specific meanings of rating data. However, most of them do not consider the topological characteristics of the bipartite network of the recommendation system. They cannot provide accurate music recommendations for users in an incremental way. Model-based methods include matrix factorization (MF) [4], K-Means Clustering based on Bayesian [5], and random walk models [6] [7]. Among them, the SVD singular value decomposition model has become the most widely used method in recommendation systems. These models are added with features such as demographic characteristics and social information. However, due to privacy policies, this information is difficult to collect. There is also a content-based method, which can usually use the automatically extracted music content characteristics to process recommendations for new songs [8] [9]. This paper mainly discusses collaborative filtering recommendation methods.

In this paper study, community detection technique is used in cold-start recommendation algorithm to solve the cold start problem in RS. In this algorithm, the interactions between users and music are regarded as a bipartite network, calculates Pearson correlation coefficient projection the bipartite network into two single-mode networks, and uses the Louvain algorithm to obtain the initial user and song community groups. The algorithm can gradually update the community based on the newly arrived data and obtain higher accuracy in the case of a cold start. After verification in multiple data sets, the cold start algorithm based on community detection can effectively improve the cold start problem in the music recommendation system, and the algorithm has high stability.

2. Related Work

The Cold start problem has attracted wide attention in the recommendation field. The researchers found that the cold start problem is an extreme stage of the data-sparse problem; they proposed a series of solutions to the data-sparse and cold start problems. The matrix factorization model projects users and items into the matrix, which explicitly represents the user's interest data. Singular Value Decomposition (SVD) is one of the widely used matrix factorization techniques. S. Funk applied SVD to the recommendation system, Koren proposed the SVD ++ algorithm based on user bias, item bias, and implicit feedback information from users, Luo introduced a co-decomposition SVD model to enhance a single data source and alleviate the problem of overfitting in matrix factorization [10]. These methods can effectively reduce the sparseness of the data and improve the accuracy of the recommendation. However, the amount of data reaches more than one million, these method is faced with the requirement that the matrix must be dense, the missing values need to be supplemented, and it can not effectively deal with online and dynamic problems.

The social information of users is useful in the development of socialized interest, which can form special interest communities. Some studies have found that user derived from communities can enhance the efficiency of recommender systems. Lei Shi proposed a matrix factorization method based on local representation, which can use a decision tree to create user groups of the same interest, and classify user groups into users by two rounds to achieve group recommendation [11]. Ke Han C uses k-SNAP's graph summary method and content similarity algorithm to perform two-stage clustering on users [12]. Chao Z proposed a clustering method that calculates Pearson-Jaccard correlation coefficients from the user and project directions, reducing the impact of data sparsity from both directions [13].

3. Method

In this section, the input information is used to construct a bipartite network of user behavior and music items, and the user data set and music data set are represented as $\mathrm{U} = \{\mathrm{u}_1, \mathrm{u}_2, \mathrm{u}_3 \cdots \mathrm{u}_\mathrm{n}\}$, $\mathrm{I} = \{\mathrm{i}_1, \mathrm{i}_2, \mathrm{i}_3 \cdots \mathrm{i}_\mathrm{m}\}$, R is the weight of the edge representing the normalized score. Fig 1 shows Bipartite network.

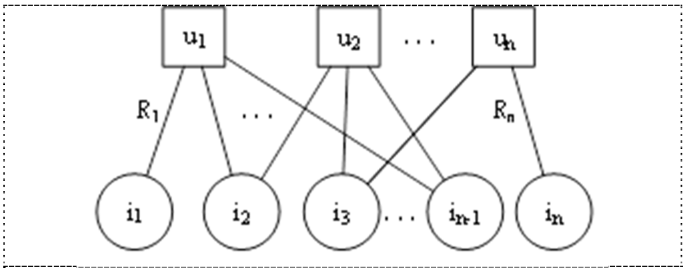

Figure 1. Bipartite network.

3.1. Bipartite Network Community Detection

For the purpose of initialize the community of the bipartite network, perform projection on users and songs one-mode network to reduce the time taken community detection. In other words, the Pearson correlation coefficient is used to calculate the user preference similarity matrix W_u, and the minimum similarity threshold is set η to highlight the similarity. Similarly, the song similarity matrix W_i can also be calculated. Pearson correlation coefficient can be written as

$$\rho_{i,j} = \frac{\sum_{i \in V_{u1} \cap V_{u2}} \left(r_{u1,i} - \overline{r_{u1}} \right) \left(r_{u2,i} - \overline{r_{u2}} \right)}{\sqrt{\sum_{i \in V_{u1} \cap V_{u2}} \left(r_{u1} - \overline{r_{u1}} \right)^2 \sum_{i \in V_{u1} \cap V_{u2}} \left(r_{u2} - \overline{r_{u2}} \right)^2}} \tag{1}$$

Where V_{u1} and V_{u2} represent the common evaluation set of users u_1 and u_2, $r_{u1,i}$ represents the ratings of user u_1 on music i, and $\overline{r_{u1}}$ represents the average ratings of user u_1 on all music.

3.2. Update Bipartite Network

To ensure accuracy in the new record recommendation process, we implemented an update method to solve the problem of scalability and new record in the music recommendation system.

When the new record enters the recommendation system, the similarity of related users in the projection similarity network is recalculated first, and the user similarity network is updated according to the similarity matrix. Next, for the old users, the Louvain algorithm is used to find the most suitable community for users, and the user's community may change, representing the migration of user preferences. For new users, update users directly to the lowest average community. In the same way, implement similar strategies to update communities in similar networks of songs. The rest of the algorithm for updating the newly recorded community will be discussed in detail in Section 3.3.

3.3. Personalized Recommendation

In order to implement cold-start music recommendation, user preferences are first defined. This article is based on user preferences for music, rather than using explicit rating information to find similar users and music. Rong Jin proposed a normalized model [16], with preference defined as:

$$p\left(R\, is\ preferred\right) = p\left(Rating \le R\right) - p\left(Rating = R\right)/2 \tag{2}$$

Where Rating is the rating level, R represents all the ratings of the user, and $p\left(Rating \le R\right)$ p(Rating ≤ R) means that if there is an item rating level below R, the user prefers songs with a rating of R. By subtracting the second item, if many objects are rated as R, then the preference priority of that level can be ignored. Refer to the latent factor (LFM) algorithm, calculate the latent factor for each community, its preference matrix $\tilde{R}$ is defined as:

$$\tilde{R}_{u,i} = \left(p_u + s_{uv} \right)' \left(q_i + g_{ij} \right) \tag{3}$$

p_u and q_i are potential preference factors related to users and songs. s_{uv} and g_{ij} have the same value for all nodes in the same community. They are k-dimensional clustering factors related to users and songs. The second step is to perform embedded the preference potential factor into the alternative least squares algorithm (ALS) by the inverse fitting algorithm, which can avoid large-scale matrix operation and improve efficiency.

$$F=\sum_{\begin{pmatrix} i\in I,u\in U \\ v\in V,j\in J \end{pmatrix}}\left[\left(R-\tilde{R}_{u,i}\right)^2+\lambda\left(\left\|p_u\right\|^2+\left\|q_i\right\|^2+\left\|s_v\right\|^2+\left\|g_j\right\|^2\right)\right] \tag{4}$$

Where λ is a smoothing parameter, we using the least-squares method to fix some values in turn, the optimal solution of unknown variables in the F can be obtained. In the whole recommendation process, firstly, the bipartite network is projected into the one-mode network of users and songs according to the Pearson correlation coefficient. Community detection of two one-mode networks using Louvain algorithm. Second, the community was updated after the new records arrived, and in the process only a small number of users and song items changed the community. Finally, the preference information is embedded in the ALS algorithm to update the community clustering factor, avoiding retraining the entire model.

4. Experiments and Discussion

In this section, we will demonstrate the performance of our algorithm through a series of experiments in the field of overlapping communities. We compare our algorithm with the SVD algorithm, and an improved Bipartite cluster method[17] , to prove the validity of our method. The experimental environment is Intel i7 8-core CPU, 8G RAM, CentOS 6.5 operating system PC. In order to eliminate the effects of randomness, all experiments were performed under the same conditions, and the experiment was repeated 10 times and averaged.

4.1. Dataset and Evaluation metrict

The dataset uses three music datasets to evaluate all models such as Million Song, Yahoo Music. Finally, in order to verify the performance of the algorithm, this paper uses the movie rating dataset MovieLens; This paper uses two popular metrics, the Mean Absolute Error (MAE) and Precision to measure the proximity of prediction grades to true ratings.

$$MAE=\frac{\sum_{i-1}^{N}\left|r_{u,i}-\hat{r}_{u,i}\right|}{N} \tag{5}$$

$$Precision=\frac{\sum_{i=1}^{N}P_i}{N} \tag{6}$$

4.2. Results

The smoothing parameter λ is selected from the set {0.1, 0.5 , 2, 4, 8, 12}, and the threshold value η of the network projection is set from the set { 0.3, 0.4, 0.5, 0.6, 0.7, 0.8}. For smaller groups in the projection network, group them into a special community. The first 80% of the data is used for training, and the last 20% is used to verify the accuracy of the recommendation.

As shown in Table1, the user prediction evaluation and real evaluation data are calculated in three relevant data sets. The community-based cold start algorithm is the best in the MovieLens data set with more evaluation data, and the Bipartite-cluster algorithm times. In other words, the SVD algorithm is the worst. The performance of the cold-start based on community detection algorithm and the Bipartite-cluster algorithm in the Million Song and Yahoo Music datasets with a small number of evaluations is not much different, but they are stronger than the SVD performance. Therefore, it can be concluded that the method of using community detection technology can improve the cold start problem in a music recommendation system and provide users with music recommendations with fairly good accuracy.

Table 1. MAE results

Dataset	Methods	New user	New item	Non-cold start
	SVD	1.248	0.792	0.965
Million Song	Bipartite-cluster	0.947	0.841	0.746
	Community cold-start	0.916	0.832	0.733
	SVD	1.319	1.273	1.218
Yahoo Music	Bipartite-cluster	1.291	1.339	1.122
	Community cold-start	0.964	1.091	0.955
	SVD	1.112	0.906	0.731
MovieLens	Bipartite-cluster	0.799	0.774	0.715
	Community cold-start	0.782	0.819	0.726

In order to verify the performance of the recommended songs, this paper also tests the Precision and running time on the Yahoo Music dataset. Fig 2 shows the accuracy of the three algorithms in different recommended items. The number of recommended songs is set to [5, 10, 15, 20].

As the number of recommendation music increases, the accuracy of the three algorithms decreases. However, the accuracy of cold-start algorithm based on community detection remained relatively stable and slightly higher than the SVD and Bipartite-cluster algorithms. The three algorithms make song recommendations on the six batches of new records on the Yahoo Music dataset. Bipartite cluster algorithm and cold start algorithm based on Community Detection remove the training time. The running time of the algorithm is shown in the Fig. 3.

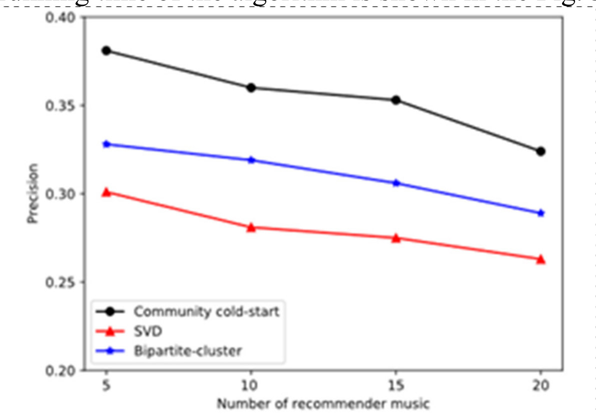

Figure 2. Precision results.

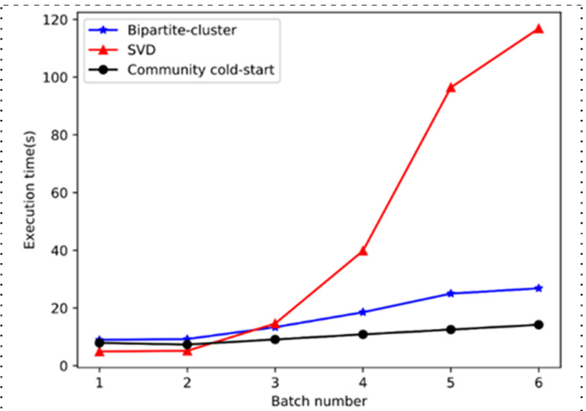

Figure 3. Execution times.

These results provide substantial evidence for the conclusions: The MAE scores of the three algorithms are similar for evaluating data-intensive datasets, and the cold-start algorithms based on community detection when sparse data and encountering new users/items have better performance than others. The method is more accurate and has a shorter run time. To sum up, the cold start algorithm based on community detection was found effective when applied to cold start problem in music recommendation systems.

5. Conclusion

In this paper, a cold start algorithm based on community detection was proposed, which projected the input user and music bipartite network to reduce the subsequent complexity. The experimental results on three real user evaluation data sets show that the cold detection algorithm based on community detection has better recommendation accuracy and stable performance. In future work, the algorithm will be further optimized to meet the needs of users like different classification music; also, the program will be tested on more data sets to find the structure where user preferences overlap. The next step is to implement a complete music recommendation system and display the results of music recommendations on a Web page.

Acknowledgments
This research work was supported by Guangxi key Laboratory Fund of Embedded Technology and Intelligent System and The National Natural Science Foundation of China under research project 41264005.

References
[1] Sandeep R. Collaborative Filtering Techniques in Recommendation Systems[J]. Data, Engineering and Applications,2019, pp.11-21.
[2] Su X, Khoshgoftaar T M. A Survey of Collaborative Filtering Techniques[J]. Advances in Artificial Intelligence, 2009, 2009:1-19.
[3] Ahn H J. A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem[J]. Information Sciences, 2008, 178(1):37-51.
[4] Koren Y, Bell R, Volinsky C. Matrix Factorization Techniques for Recommender Systems[J]. Computer, 2009, 42(8):30-37.
[5] Park, H.B, Y.S., Moon, S.C., Ko, H.H. Clustering Method of Weighted Preference Using K-means Algorithm and Bayesian Network for Recommender System[J]. Journal of Information Technology Applications & Management Special Section, 2013. 219 – 230.
[6] Fouss F, Pirotte A, Renders J M, et al. Random-Walk Computation of Similarities between Nodes of a Graph with Application to Collaborative Recommendation[J]. IEEE Transactions on Knowledge and Data Engineering, 2007, 19: p.355-369.
[7] Li X, Su X, Wang M. Social network-based recommendation: a graph random walk kernel approach.[C]// Acm/ieee-cs Joint Conference on Digital Libraries. ACM, 2012.
[8] Seyerlehner K, Widmer G, Knees P. A Comparison of Human, Automatic and Collaborative Music Genre Classification and User Centric Evaluation of Genre Classification Systems[C]// Adaptive Multimedia Retrieval Context, Exploration, & Fusion-international Workshop, Amr, Linz, Austria, August, Revised Selected Papers. Springer-Verlag, 2010.
[9] Vall A, Eghbal-Zadeh H, Dorfer M, et al. Music Playlist Continuation by Learning from Hand-Curated Examples and Song Features: Alleviating the Cold-Start Problem for Rare and Out-of-Set Songs[J]. 2017.
[10] Koren Y. Factorization meets the neighborhood: A multifaceted collaborative filtering model. In: Proc. of the 8th Int'l ACM SIGKDD Conf. on Knowledge Discovery and Data Mining. ACM Press, 2008. 426-434.
[11] Shi L, Zhao W X, Shen Y D. Local Representative-Based Matrix Factorization for Cold-Start Recommendation[J]. Acm Transactions on Information Systems, 2017, 36(2):1-28.
[12] Ke-Han C. User Clustering Based Social Network Recommendation[J]. Chinese Journal of Computers, 2013.
[13] Chao Z, Ying-Hua S, Hua-Feng X, et al. Collaborative Filtering Recommendation Algorithm Based on Bidirectional Clustering on Users and Items[J]. journal of qingdao university (natural science edition), 2018.
[14] Xu B, Bu J, Chen C, et al. An exploration of improving collaborative recommender systems via user-item subgroups[C]// 2012.
[15] Zhang Y, Zhang M, Liu Y, et al. Improve collaborative filtering through bordered block diagonal form matrices[J]. 2013.
[16] Ratings P A, Jin R, Si L, et al. Collaborative Filtering with Decoupled Models for[C]// Twelfth International Conference on Information & Knowledge Management. ACM, 2003
[17] Qin J, Ren K, Fang Y, et al. Sequential Recommendation with Dual Side Neighbor-based Collaborative Relation Modeling[J].2019.