CGMF: Coupled Group-based Matrix Factorization for Recommender System

Fangfang Li[1,2], Guandong Xu[2], Longbing Cao[2], Xiaozhong Fan[1], and Zhendong Niu [1]

[1]School of Computer Science and Technology, Beijing Institute of Technology, China
Fangfang.Li@student.uts.edu.au;{fxz,zniu}@bit.edu.cn
[2]Advanced Analytics Institute, University of Technology, Sydney, Australia
{Guandong.Xu,Longbing.Cao}@uts.edu.au

Abstract. With the advent of social influence, social recommender systems have become an active research topic for making recommendations based on the ratings of the users that have close social relations with the given user. The underlying assumption is that a user's taste is similar to his/her friends' in social networking. In fact, users enjoy different groups of items with different preferences. A user may be treated as trustful by his/her friends more on some specific rather than all groups. Unfortunately, most of the extant social recommender systems are not able to differentiate user's social influence in different groups, resulting in the unsatisfactory recommendation results. Moreover, most extant systems mainly rely on social relations, but overlook the influence of relations between items. In this paper, we propose an innovative coupled group-based matrix factorization model for recommender system by leveraging the user and item groups learned by topic modeling and incorporating couplings between users and items and within users and items. Experiments conducted on publicly available data sets demonstrate the effectiveness of our approach.

1 Introduction

With the advent of online social networks, more and more social information is incorporated to RS, and social RS is becoming an active area in RS [7] [3]. The main motivation behind social RS is to leverage the auxiliary friend relations of users to tackle the common challenges in RS, e.g., cold-start and sparsity. For example, for a new user to RS, it is usually difficult to find the like-minded users due to the lack of the new user's ratings. However, through the social information known from social networking, this difficulty could be partially overcome. The underlying assumption of the social recommendation approach is that a user's taste is influenced by his/her friends in social networking. Accordingly, assigning more weights to items that the friends are interested in will potentially improve the satisfaction of recommendations. However, the extant social RS treats user's friends equally, but ignores the fact that user's social interests are intrinsically multifaceted.

Everyone has specific preference in particular groups. This indicates that a user may trust different subsets of friends in different groups. More specially, a user may have friends working in different domains, and join in activities across different domains. This is evidenced by that the extant social networks such as Google+, Facebook, Twitter already have such mechanisms to divide users into groups for sharing different information with different groups. Undoubtedly, in social RS, utilizing such social group information will be able to provide better personalized services for users. But most of extant social Web applications such as Tweeter, Sina Weibo, Delicious etc. do not provide reliable mechanisms to allow users to differentiate social connections from individual groups. Some recent researches integrate the distinguished group information into recommendation algorithms, e.g. [13] leverages the social trust circles from item-category information for social recommendation. Despite of the superior results demonstrated from the given multi-category rating data sets, this approach has a major limitation that it relies on the explicit item category information to form user circles, upon which social recommendation is made. However, such information is not always available in existing social networks e.g., Facebook or Twitter might not have such explicit category information, resulting in difficulties in applying the proposed algorithm. In this work, we attempt to address this unknown category information problem based on hidden topic modeling.

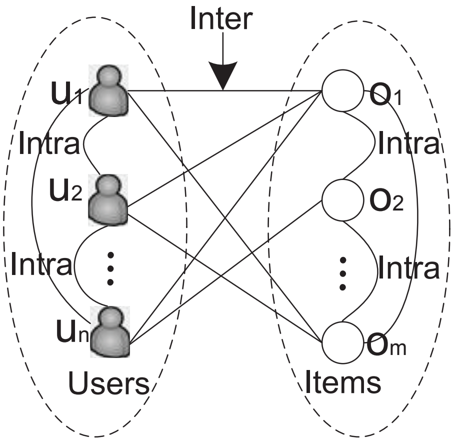


Fig. 1. Coupling Relations in Recommender Systems

The extant social RS mainly focused on capturing social friendships within users and mutual relations between users and items. However, just considering social friendships and mutual relations are not enough for recommendation. Actually, items are often coupled together, if item o_i is closely relevant to item o_j, the preference of user u on item o_i would be influenced by item o_j. The effectiveness of recommendation would be probably increased through analyzing these relations of items. In fact, the complex relations in social RS can be abstracted into two classes. One is inter-couplings such as user preferences or ratings on items, the other one is intra-coupling including user intra-coupling and item intra-coupling [1], which are shown in Fig. 1. These coupling relations should be considered simultaneously and evenly in learning the recommendation model. The inter-couplings between users and items have been well studied [2] and utilized in the extant RS for enhancing the result of recommendation. However, the user intra-coupling relations are not sufficiently exploited and item intra-coupling relations are often ignored in social RS. In addition, few solutions have been proposed to integrate user-intra, item-intra and user-item inter-coupling relations into a unified manner. A complete considerations of such couplings can provide a practical mean for enhancing the effectiveness of social RS and solving the cold start and data sparsity problems. In this work, we incorporate the coupling relations and group information into the matrix factorization model.

The contributions of the paper are concluded as follows:

- We propose a Coupled Group-based Matrix Factorization model (CGMF) which incorporates couplings between/within users and items such as user intra-coupling, item intra-coupling and user-item inter-coupling.
- We apply topic modeling on item descriptions to automatically extract hidden topics of items and derive the user groups, then integrate group information into matrix factorization as an additional constraint in learning recommendation models.
- We conduct experiments to verify our algorithms and recommendation models.

The rest of the paper is organized as follows. Section 2 presents the related work. In Section 3, we analyze the coupled interactions between users and items and within users and items. Then we introduce the group formation algorithm in Section 4. After that, the CGMF model is proposed integrating user and item groups. Experimental results and analysis are presented in Section 6 followed by the conclusion.

2 Related Work

Collaborative filtering (CF)[11] is one of the most successful approaches taking advantage of user rating history data to predict users' interests. Research efforts have been invested to make use of complimentary information in order to address the cold-start and sparsity problem. Slope One is a family of algorithms used for collaborative filtering, introduced in [6]. Arguably, it is the simplest form of non-trivial item-based collaborative filtering based on ratings of another item.

However, CF algorithms do not consider user intra-coupling and item intra-coupling existed in RS and the users and items are assumed to be independent and identically distributed.

Matrix factorization [4] [5] is a latent factor model which is generally effective at estimating overall structure that relates simultaneously to most of or all items. The basic matrix factorization approach for RS is based on an assumption that users are independent and identically distributed. This approach ignores the social activities between users, which is not consistent with the reality that we normally ask friends for recommendations. With the advent of social network, many researchers have started to analyze social recommender systems and various models integrating social networks such as Social Recommendation (SoRec) [8], Social Trust Ensemble (STE)[7], Recommender Systems with Social Regularization [9], etc. have been proposed. Social Matrix Factorization approaches actually consider the social activities of users, but social relations are mixed together and treated equally. As a result, it is impossible to differentiate social recommendations from different friends in terms of their preferred areas. Apart from this, item intra-couplings are also ignored.

3 Coupled-Interaction Analysis

Coupled-Interaction includes intra-couplings within users and items and inter-couplings between users and items.

3.1 Inter-coupled Interaction

Users often directly interact with items, for example, some users will give their rating after they watched a movie, which is the most intuitionistic interaction between users and items. Reflecting the relations between users and items, inter-coupling between user u and item o_i can be directed computed by $\delta^{Ie}_{u,i} = P_u Q_i^T$, where each item o_i is associated with a vector $Q_i \in \mathbb{R}^d$, and each user u is associated with a vector $P_u \in \mathbb{R}^d$.

3.2 Intra-coupled Interaction

Besides inter-coupling, RS also have massive intra-couplings which contain user-intra-coupling and item-intra-coupling. Intra-coupling within users and items can be modeled by Eqn. 1

$$\delta^{Ia}_{u,i} = \sum_{v \in F(u)} S_{u,v} P_v {Q_i}^T + \sum_{j \in N(i)} W_{i,j} P_u {Q_j}^T \qquad (1)$$

where the first part is user-intra-coupling and the second part is item-intra-coupling. $S_{u,v}$ is the friendship relation of users u and v, and W_{ij} is the relevance of items o_i and o_j.

Eqn. 1 not only says that user profile P_u should be similar to his friends' profile P_v, but also says if user u is interested in item o_i, he/she will also interest in item o_j which is closely relevant to item o_i.

After coupled interactions are considered, MF prediction model is modeled as follows:

$$\hat{R}_{u,i} = r_m + \delta_{u,i}^{Ie} + \delta_{u,i}^{Ia} \tag{2}$$

which $\delta_{u,i}^{Ia}$ represents the intra-couplings within users and items, $\delta_{u,i}^{Ie}$ represents the inter-coupling between users and items.

4 Group Formation

Topic modeling is a proper way to automatically divide all items into different topics which can be considered as groups. Theoretically, LDA is a probabilistic generative model for a text corpus. The basic idea of LDA is based on the hypothesis that a person has certain topics in mind when writing an article. To address a topic, the author needs to pick up a word with a certain probability from a bag of words reflecting that topic. In this manner an item is represented as random mixtures over latent topics and each topic is characterized by a set of related words with a probability distribution. In the context of social networks, the obtained topics represent the commonly shared perception of the items by collaborative users, and the words of the specic topic constitute a common vocabulary contributed to the topic. In a summary, topic modeling can be used to partition different items into different groups especially when items have corresponding text description information.

Through the LDA model, we can capture the hidden topics and item assignments to these topics. Once we get the topic probability distribution of the items, we can easily analyze the user's affiliation on such topics according to the following Eqn. 3.

$$p_r\left(u|g_k\right) = \sum_{o_i} p_r\left(u|o_i\right) p_r(o_i|g_k) \tag{3}$$

with an item $o_i = \{w_{i,n}, n = 1, \ldots, N_i\}$ is generated by picking a distribution over the topics from a Dirichlet distribution, $p_r\left(u|g_k\right)$ and $p_r\left(u|o_i\right)$ are the probality that the user u belongs to group $g_k (1 \leq k \leq K)$ and the interests on item o_i. The whole process of item group formation and user probability distribution on these groups is described as following algorithm 1.

5 Coupled Group-based MF Model

After considering inter-couplings and intra-couplings, we aim to integrate them with group information in a unified model, namely Coupled Group-based MF model as follows.

$$\hat{R}_{u,i}^{(g)} = r_m + {\delta_{u,i}^{Ie}}^{(g)} + {\delta_{u,i}^{Ia}}^{(g)} \tag{4}$$

Algorithm 1: Group Formation Algorithm

```
Input: Items set ${o_1, o_2, ..., o_m}$
Output: Item groups ${g_1, ..., g_K}$ and the probability distribution matrix of
        users belongs to these groups
Classify items set to different topic groups ${g_1, ..., g_K}$ in terms of significant
text-related information by LDA topic modeling;
Compute user distribution on the classified topics by Eqn. 3;
Assign the probabilistic weight to users which indicates how much the users
belong to the groups.
```

Different from MF, the prediction task of matrix $\hat{R}$ is transferred to compute the mapping of users and items to factor matrices P, Q, coupling relations and group information. Once this mapping is completed, RS can easily predict the rating a user will give to any item in a specific group by using Eqn. 4. The computation of the mapping can be optimized by minimizing the regularized squared error on the set of observed ratings. The objective function is given as Eqn. 5.

$$L^{(g)} = \frac{1}{2} \sum_{(u,i) \in K} \left(R_{ui}^{(g)} - \hat{R}_{u,i}^{(g)} \right)^2 + \\ \frac{\lambda}{2} \left(\|Q_i^{(g)}\|^2 + \|P_u^{(g)}\|^2 + \sum_{v \in F(u)^{(g)}} \|S_{u,v}^{(g)}\|^2 + \sum_{j \in N(i)^{(g)}} \|W_{i,j}{}^{(g)}\|^2 \right) \tag{5}$$

The training process starts at randomly initiate values of $P^{(g)}$ and $Q^{(g)}$. Then it iterates to update $P^{(g)}$, $Q^{(g)}$, $S_{u,v}^{(g)}$ and $W_{ij}^{(g)}$ by the gradient decent approach on the objective function $L^{(g)}$ until convergence. After $P^{(g)}$ and $Q^{(g)}$ are learned from the training process, we can predict the ratings for user-item pairs (u, o_i) by Eqn. 4.

6 Experiments and Results

In this section, we evaluate our proposed model and compare it to the existing approaches respectively using Movielens, LastFm and DBLP citation database [12].

6.1 Data Set

MovieLens data set has been widely explored in collaborative filtering research in last decade. MovieLens 10M data set consists of 72,000 users, 10,000 movies and 10 million ratings data. MovieLens is a classic data set for evaluating recommendation models, however, this data set does not contain the friendship of users i.e.

the user intra-coupling. Therefore, the following experiments on MovieLens can not show the sensitivity of user intra-couplings. But the data set actually has a special genre feature which is applied for grouping all the movies, so the results on MovieLens show the influence of inter-couplings and item intra-couplings.

Different from MovieLens, LastFm data set contains social networking, tagging, and music artist listening information involving 1892 users, 17632 artists, 12717 bi-directional user friend relations, 92834 user-listened artist relations, and 11946 tag assignments. However, the data set does not have the rating data of users on artists. We know that the listening count indirectly reflects the preference of users on the artist, therefore, we normalize the listening count for the users on artists to [0,5] to indicate the implicit preferences. The data set has tagging information which is used for group formation by topic modeling. Therefore, the results on LastFm show the sensitivity of couplings between and within users and items, and group information after this adjusted setting.

The DBLP citation database contains 1,572,277 papers and 2,084,019 citations. Each paper is associated with title, abstract, authors, year, venue, citation number and references. The data set consists of various coupling relations such as co-author and citation relations. For the DBLP data set, in the experiments, authors and papers are separately thought as users and items. Co-author relations are taken as friendships between users since a user must have friendship with the co-authors of his/her publications. And the item intra-couplings are captured by the citation network and the text similarity based on the paper's title and abstract. The "write" and "write-by" relations are converted to "0-1" ratings representing the preference of the user to the paper.

Overall, the following experiments on MovieLens and LastFm data sets are separately used for movie and artist recommendation, while DBLP data set is explored for paper recommendation for testing our CGMF model.

6.2 Experimental Settings

The 5-fold cross validation is performed in our experiments. In each fold, we have 80% of data as the training set and the remaining 20% as testing set. Here we use Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) as evaluation metrics.

RMSE and MAE are defined as follows:

$$RMSE = \sqrt{\frac{\sum_{(u,i)|R_{test}} (r_{u,i} - \hat{r}_{u,i})^2}{|R_{test}|}} \tag{6}$$

$$MAE = \frac{\sum_{(u,i)|R_{test}} |r_{u,i} - \hat{r}_{u,i}|}{|R_{test}|} \tag{7}$$

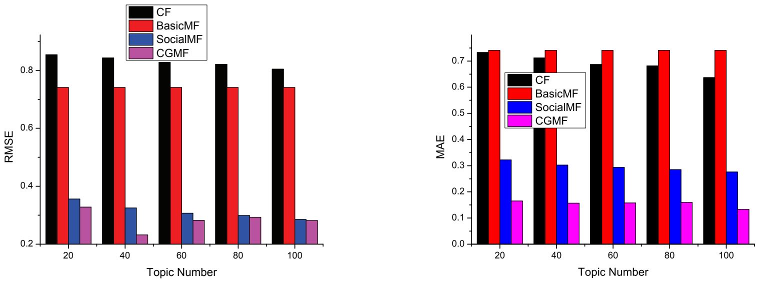
where R_{test} is the set of all pairs (u, o_i) in the test set.

To evaluate the performance of our proposed CGMF we consider three baseline approaches:

- CF: This is the well-known item based collaborative filtering method called Slope One.
- BasicMF: This method is a probabilistic matrix factorization approach in [10] which does not take the social network into account.
- SocialMF: This is the model which just considers the social friendships but ignores the item intra-couplings and group information.

6.3 Experimental Results and Discussions

Effectiveness of Couplings and Groups DBLP and LastFm data sets have ample couplings within users and items and between users and items, and text related information used for forming groups, so the experimental results on the two data sets can demonstrate the impacts of couplings and group information. We depict the effectiveness comparisons with respect to each method on DBLP and LastFm data sets in figures 2 and 3. From Fig. 2, we can clearly see that, our proposed CGMF method outperforms the counterparts in terms of RMSE and MAE. Compared to SocialMF, CGMF achieves up to 4.3% improvement on RMSE and 14.7% on MAE, while immense improvements compared to CF and Basic MF. On LastFm data sets, Fig. 3 evidences that CGMF performs much better than the benchmark methods. We can see that CGMF can reach a prominent improvements compared to CF and Basic MF approaches, which is resulted from considering complete coupling relations.

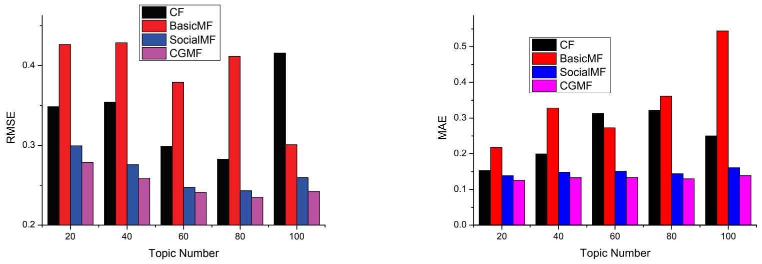


(a) RMSE

(b) MAE

Fig. 2. RMSE and MAE Comparison on DBLP with Different Number of Topics

Effectiveness of Item Intra-couplings and Groups Because the users of MovieLens data set do not have friendships which mean the user intra-couplings can not be captured, but MovieLens actually has natural genre feature which is used to form groups, the experimental results on Fig. 4 can show the performance of item intra-couplings and group information. The results indicated that our proposed CGMF can reach an average improvements of 3.5% and 3.1% on RMSE,

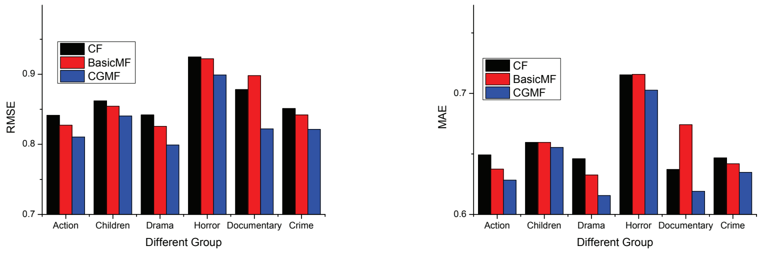


(a) RMSE

(b) MAE

Fig. 3. RMSE and MAE Comparison on LastFm with Different Number of Topics

and 1.6% and 1.8% on MAE compared to CF and Basic MF approaches. The biggest improvements are 4.3% on RMSE, and 3.0% on MAE compared to CF in the Dramma group. Compared to Basic MF, our proposed CGMF can improve by 7.6% on RMSE and 5.5% on MAE in the Documentary group.



(a) RMSE

(b) MAE

Fig. 4. RMSE and MAE Comparison on MovieLens with Different Group

Adaptiveness of Different Group Formation Methods The experimental results on DBLP and LastFm data sets indicate that topic modeling is an effective group formation method, while results on Movielens show that category information which is used for group formation can also contribute to improving the effectiveness. That is to say, no matter which group formation methods (topic modeling or category information) are chosen, our proposed CGMF can be applied. The very significance of CGMF is topic modeling can be chosen for grouping when the data set does not have category information.

Overall, the RMSE and MAE figures on all the three data sets show the significant improvements of CGMF compared to benchmark methods. Therefore, we can conclude that by taking the couplings and group information into consideration, our approach can reach a better recommendation.

7 Conclusion

This paper proposed a coupled group-based matrix factorization model for recommender system, which incorporates coupling relations between and within users and items. CGMF first extracts items hidden groups via a Latent Dirichlet Allocation (LDA) model and derives user groups. Then coupling relations and group information are incorporated into CGMF. The experiments conducted on the real data sets demonstrated the superiority of the approach against the state-of-the-art methods.

8 Acknowledgements

This work is sponsored in part by Australian Research Council Discovery Grants (DP1096218 and DP130102691) and ARC Linkage Grant (LP100200774).

References

1. Longbing Cao, Yuming Ou, and Philip S. Yu. Coupled behavior analysis with applications. *IEEE Trans. Knowl. Data Eng.*, 24(8):1378–1392, 2012.
2. Wei Feng and Jianyong Wang. Incorporating heterogeneous information for personalized tag recommendation in social tagging systems. In *KDD*, pages 1276–1284, 2012.
3. Mohsen Jamali and Martin Ester. A matrix factorization technique with trust propagation for recommendation in social networks. In *RecSys*, pages 135–142, 2010.
4. Yehuda Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *KDD*, pages 426–434, 2008.
5. Yehuda Koren, Robert M. Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *IEEE Computer*, 42(8):30–37, 2009.
6. Daniel Lemire and Anna Maclachlan. Slope one predictors for online rating-based collaborative filtering. In *Proceedings of SIAM Data Mining (SDM'05)*, 2005.
7. Hao Ma, Irwin King, and Michael R. Lyu. Learning to recommend with social trust ensemble. In *SIGIR*, pages 203–210, 2009.
8. Hao Ma, Haixuan Yang, Michael R. Lyu, and Irwin King. Sorec: social recommendation using probabilistic matrix factorization. In *CIKM*, pages 931–940, 2008.
9. Hao Ma, Dengyong Zhou, Chao Liu, Michael R. Lyu, and Irwin King. Recommender systems with social regularization. In *WSDM*, pages 287–296, 2011.
10. Ruslan Salakhutdinov and Andriy Mnih. Probabilistic matrix factorization. In *Advances in Neural Information Processing Systems*, volume 20, 2008.
11. Xiaoyuan Su and Taghi M. Khoshgoftaar. A survey of collaborative filtering techniques. *Adv. in Artif. Intell.*, 2009:4:2–4:2, January 2009.
12. Jie Tang, Duo Zhang, and Limin Yao. Social network extraction of academic researchers. In *ICDM*, pages 292–301, 2007.
13. Xiwang Yang, Harald Steck, and Yong Liu. Circle-based recommendation in online social networks. In *KDD*, pages 1267–1275, 2012.