Exploring Social Network Information for Solving Cold Start in Product Recommendation

Chaozhuo Li(✉), Fang Wang, Yang Yang, Zhoujun Li, and Xiaoming Zhang

State Key Laboratory of Software Development Environment,
Beihang University, Beijing, China
{lichaozhuo,fangwang,lizj}@buaa.edu.cn

Abstract. Cold start problem is a key challenge in recommendation system as new users are always present. Most of existing approaches address this problem by leveraging meta data to estimate the tastes of new user. Recently, social network has been becoming an integral part of daily life. Usually, social network information reflect users preferences to some extent, combining this kind of data would contribute to address the cold start problem. Existing approaches of this kind are either leverage relationships between users or utilize meta data such as demographic information. The huge textual information in social network has been neglected. In this paper, we propose a novel recommendation framework, in which the textual data in social network are used to improve the recommendation accuracy for new users. In particularly, both of new user's interests and items are modeled by mining the textual data in social network. Experimental results demonstrate that our approach is superior to other baseline methods in both precision and diversity.

Keywords: Recommendation system · Social network · Cold start

1 Introduction

In the recent decade, recommendation system has become an integral part of peoples lives, as a means to help users in information overload scenarios by proactively finding items or services on their behalf. Recommendation technology has been successfully applied in many e-commerce sites, such as Amazon.com, WalMart.com and Netflix.com. Existing work usually relies on user's historical item ratings, especially for the Collaborative Filtering (CF) based recommendation systems [12,15,16]. When it comes to new users without rating records, however, the performance of these strategies falls a great deal, which is known as the cold start problem and is one of the most challenging problems in recommendation systems.

Because of the prevalence of social network, many e-commerce systems allow users to provide their social network id (e.g., Twitter and Weibo). A large portion of costumers are pleasure to offer their social network URLs. For example, Douban website is a famous online bookstore in China, in which more than 23 percents of users has published social network URLs in their profiles. We argue

J. Wang et al. (Eds.): WISE 2015, Part II, LNCS 9419, pp. 276–283, 2015.
DOI: 10.1007/978-3-319-26187-4_24

that user's social network information can help to capture user's interests. For example, a microblog *"Captain Jack Sparrow in Pirates of the Caribbean"* implies the microbloger may be interested in fantasy-adventure films or books. Besides, many social network websites allow users to define their own tag of interests. This informations are also meaningful for product recommending. For example, social network users labeled with tag *"Comic"* are most likely to purchase comic related commodities.

In this paper, we design an efficient product recommendation system for handling the cold start problem, by leveraging users social network information. We run it on book domain as an example. The proposed techniques can also be adapted to recommending other kinds of products. Firstly, books with similar topics are aggregated into the same cluster which is used instead of specific books to represent user's reading interests. We propose an innovative approach to model the user's reading interests on book clusters by mining the textual data in social network. Specifically, a probability model used to mine user interest from user's tags and a classifier-based microblog mining method are proposed. We also construct item models based on book clusters using a matrix factorization model. Finally, the recommendations to new user can be made based on users interest model and the learned item models.

2 Recommendation System Using Social Network Textual Data

2.1 Problem Definition

In this paper, we aim to recommend bo oks to new users using their social network information. There are three major processes as below.

User interest modeling process can be represented as the constructing of a vector V: $V \in R^{1\times p}$ for a single user, p is the count of book clusters. V_i indicates user's preference on book cluster i.

Item modeling process is represented as the constructing of a matrix I: $I \in R^{p\times n}$, n is the count of books. The item modeling process can be abstract into a mathematical expression: $S \approx UI$, where matrix $U : U \in R^{m\times p}$ stands for the interest models of m training users, S is the rating matrix. Given matrices U and S, we provide a matrix factorization model to build I by solving the following optimization problem:

$$\text{minimize} \quad \frac{1}{2}\|S - UI\|_F^2 + \frac{\lambda}{2}\|I\|_F{}^2 \tag{1}$$

After the construction of I, the final problem is how to make recommendations to new users with their interest models and I.

2.2 Overview of Our System

Figure 1 shows the detailed structure of our system. The major components are user interest modeling, item modeling and recommendation-making. In order to explain more clearly, framework is divided into two processes: training process and predicting process.

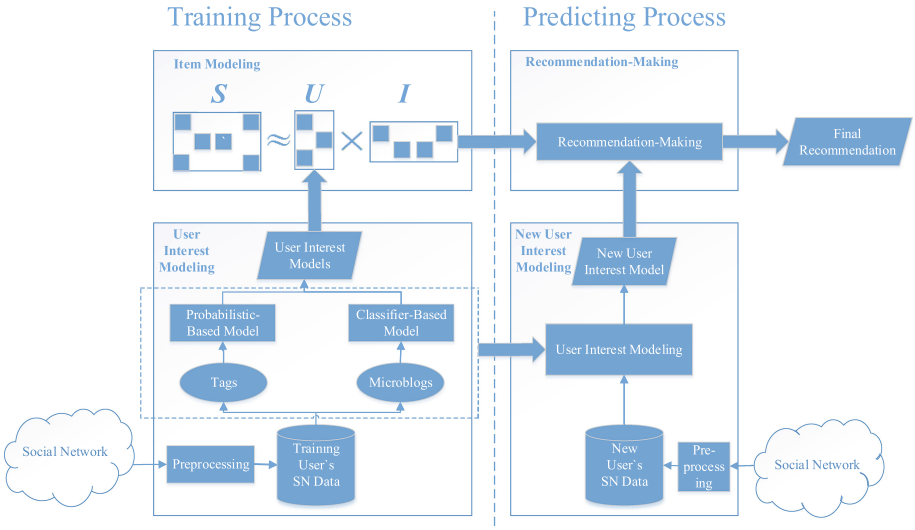

Fig. 1. Framework of our recommendation system contains three major parts: user interest modeling, item modeling and making recommendation. In order to explain more clearly, it is divided into two processes: training process and predicting process.

2.3 Crawling and Preprocessing

Data we used in this paper are crawled from two websites: Douban and Sina Weibo. Douban Website (http://book.douban.com/) is similar to Netflix, where users can label books with tags which describe book's contents by several terms, the users also can mark whether they have read a specific book. Thus in rating matrix S, if user u has read book i, $S_{u,i} = 1$, otherwise $S_{u,i} = 0$. Sina Weibo is the biggest social network platform in China, where users can label themselves with tags and publish microblogs. In addition, Douban users present their social network information in profiles. After matching Weibo users and Douban users by Weibo URLs provided in Douban users profiles, social network users reading list can be obtained.

2.4 User Interest Modeling

User interest modeling aims to extract user's interests from social network text. We propose "*book clusters*" as the intermediary between users and books. Each book cluster contains a group of similar books. The user interest modeling aims to mapping user's social network data to his tastes on book clusters.

User Interest Modeling Using Tags. In this section, we propose a probabilistic-based model using chi-square goodness score of fit test. Here gives some definitions of variables: T is the set of books read by users whose tags of interests contains tag t, C is the set of books in book cluster c. $C \cap T$ is the set of books which have been

read by users who labeled with tag t and also belongs to book cluster c. $*$ is the set of all books. All sets but $*$ allow repeated records.

The global distribution of book clusters: $P(C|*) = \frac{Count(C)}{Count(*)}$. The conditional distribution of book clusters given readers with specific tag of interests: $P(C|T) = \frac{Count(C \cap T)}{Count(T)}$.

$S(T, C)$ is the degree of discrepancy between book cluster's distribution given all users and distribution given users labeled by specific tag.

$$S(T, C) = \frac{(P(C|T) - P(C|*))^2}{P(C|*)} \times Count(C \cap T) \tag{2}$$

To filter useless tags, we sort all scores in descending order and select top K records as representative tags and their correlation coefficients with book clusters. Considering the fact that users may have several different tags, we select the largest $S(T, C)$ for each tag t of the user as his interests score on book cluster c.

$$P_t'(U, C) = \underset{all\ T\ of\ U}{Max}(S(T, C)) \tag{3}$$

For each user, we can get a numeric vector $V : V \in R^{1 \times p}$ contains his preference on all book clusters. Then we normalize the vector using "$Min-Max\ Normalization$" strategy. Finally, we achieve user's interest scores on book clusters as user's interest model.

User Interest Modeling Using Microblogs. We design a classifier-based model to extract interest model from user's microblogs. For each book cluster, we train a binary classification model to decide whether the input microblog is related to the book cluster. The count of microblogs related to a book cluster are regarded as the user's interest score on this book cluster.

In training process, for each book cluster, using each book's tags, a SVM classifier is trained for each book cluster by one-vs-all multi-class classification. Books contained in a specific book cluster are represented in "Bag of Words". In predicting process, we take every single microblog as input and utilize the trained models to decide whether the microblog is related to the specific cluster. After processing all microblogs, we can get a vector $V \in R^{1*p}$ contains counts of microblogs related to every book cluster. Greater count values imply the user is more interested in this book cluster. After the above processes, we get user's interest scores on book clusters from microblogs as user's interests model.

Combination. Here we achieve two matrices representing user's interests models: U_{tag} and $U_{microblog}$. U_{tag} is predicted from user's tags, $U_{microblog}$ is predicted from user's microblogs. We choose a simple but effective method to combine these two matrices.

$$U = \lambda U_{tag} + (1 - \lambda) U_{microblog} \tag{4}$$

2.5 Item Modeling

We propose a novel way to combine book popularity and affiliation into item model. Item models are constructed from user's reading history and user's interest models. User's reading history can reflect book's popularity, combining with user's interests on book clusters, we can achieve book's relevance with book clusters.

Matrix S represents training users' rating scores, matrix U represents training users' interest models constructed according to Eq. (11). We propose to compute matrix $I : I \in R^{p \times n}$ given U and S by solving the optimization problem (1). Similar to solving optimization problem in [13], the optimal solution is given by

$$I = (U^T U + \lambda E)^{-1} U^T P \tag{5}$$

where E is a $k \times k$ identity matrix. λ is a small positive numeric avoiding the chance that determinant of $U^T U$ equals to 0. Each column in matrix I represents a model of book. Larger $I_{i,j}$ indicates the relevance between book cluster i and book j is tighter, users who interest in book cluster i are likely to enjoy book j.

2.6 Making Recommendation for New User

In this section, we show how our framework help solving cold start problem. Assume new users social network information (tags, microblogs) are available, we can construct his interest model $V_{cluster} : V_{cluster} \in R^{1 \times p}$, p is the count of book clusters. Combining new user's interest model and items' models, we can achieve his tastes on different books: $V_{book} = V_{cluster} \times I$. After sorting elements in V_{book} in descending order, we select top K books as the final recommendation list.

3 Experiments

3.1 Data Sets

We crawled user's social network information from Sina Weibo (http://weibo.com) and got users reading history from Douban Website (http://www.douban.com). We also crawled book's information from Douban Book Website (http://book.douban.com). Finally we get 10242 active users and their related data. 2000 most popular books are picked as recommend targets. We select top K ranked books as final recommendation list. 10-fold cross validation method is used to assess the results.

3.2 Baseline Methods

Some frequently strategies used in cold start recommendation are shown below. **Random Strategy**: Recommended books are randomly selected from all books [12]. This is the simplest method to solve cold start problem. **Most Popular Strategy**: This is a naive method to select most popular books as recommendation list which is same to all new users [8,13]. **Tags Only Strategy**: User interest models

are extracted only from user's social network tags. **Microblogs Only Strategy-trategy**: User interest models are extracted only from user's microblogs. **Nearest Neighborhood Strategy**:Ten nearest users are selected as new users neighbors according to their interest models and neighbor's most favorite books are picked as final recommendation list [6].

3.3 Evaluation Protocol

We adopt several binary relevance based information retrieval performance metrics. **Pre @ K**: In the K recommended books, this is the ration of books user has read. For all test users, we take the average of Pre @ K scores to evaluate different strategies. **MRR**: Mean Reciprocal Rank is a popular metric strategy used in information retrieval field, the reciprocal value of the mean reciprocal

Table 1. Experimental results

Strategy	K	PRE @ 10	MRR	MAP	DS
Random	3	0.01	0.018	0.018	7
Most popular	3	0.200	0.423	0.341	4
Tags Only	3	0.263	0.423	0.378	6.85
Microblogs Only	3	0.152	0.221	0.316	5.94
Nearest Neighborhood	3	0.156	0.196	0.211	4.33
Our Strategy	3	**0.266**	**0.434**	**0.399**	**8.53**
Random	5	0.015	0.036	0.036	12
Most popular	5	0.201	0.416	0.362	12
Tags Only	5	0.237	0.419	0.369	9.81
Microblogs Only	5	0.15	0.214	0.304	7.96
Nearest Neighborhood	5	0.163	0.204	0.214	6.8
Our Strategy	5	**0.249**	**0.447**	**0.426**	**12.61**
Random	10	0.022	0.056	0.025	**20.63**
Most popular	10	0.184	0.414	0.361	18
Tags Only	10	0.21	0.429	0.376	18.247
Microblogs Only	10	0.146	0.219	0.337	11.10
Nearest Neighborhood	10	0.163	0.217	0.242	17.12
Our Strategy	10	**0.229**	**0.459**	**0.405**	20.42
Random	20	0.019	0.071	0.015	**34.26**
Most popular	20	0.147	0.413	0.351	33
Tags Only	20	**0.197**	0.434	0.358	28.52
Microblogs Only	20	0.126	0.246	0.316	14.50
Nearest Neighborhood	20	0.148	0.208	0.227	22.75
Our Strategy	20	0.193	**0.456**	**0.397**	31.19

rank corresponds to the harmonic mean of the ranks. **MAP**: Mean Average Precision is a popular metric method used in IR domain. **Diversity**: Diversity is a very important quality in recommendation systems. A recommendation system with higher diversity can satisfy users' special tastes better. In fact there is no standard formula to evaluate the diversity score. Thus we simply choose the count of distinct book clusters in final recommendation list as diversity.

3.4 Experimental Results

The complete results are shown in Table 1. Given different size (K) of recommendation list, we compare performance of our framework with baseline methods. According to result, using users interest model extracted from his social network information, our strategy achieve a better performance.

4 Related Works

Cold start recommendation has received a lot of attentions in recent decades. In general, there are two main categories of research approach to solve the cold start problem. The first category is to make cold start recommendation without using external information. To improve precision of recommendation with few rating scores, Zhang [4] aimed to predict unrated items from like-minded user clusters and similar item clusters. The second category utilizes external information to recommend items for new users. Schein proposed an aspect model with latent variable method which combines both collaborative and content information in model fitting [5]. A predictive feature-based regression model that leverage information of users and items, was proposed by Park [8].

5 Conclusion

In this paper, we proposed a novel approach for solving cold start problem in production recommendation by leveraging social network textual information. Book clusters are viewed as the intermediaries between users and books. User interest models representing user's interests on book clusters are extracted from user's textual data (tags, microblogs). Item models representing correlations between book and book clusters are built by a well-designed matrix factorization method. Recommendations for new user without rating history are made based on his interest model and the trained item models. Experimental results show our framework is able to attain both high precision and diversity.

Acknowledgments. This work is supported in part by the National Natural Science Foundation of China (Grant Nos. 61170189, 61370126, 61202239), National High Technology Research and Development Program of China under grant (No. 2015AA016004), the Fund of the State Key Laboratory of Software Development Environment (No. SKLSDE-2015ZX-16), and Microsoft Research Asia Fund (No. FY14-RES-OPP-105).

References

1. Ahn, H.J.: A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem. Inf. Sci. **178**(1), 37–51 (2008)
2. Chen, C.C., Wan, Y.H., Chung, M.C., Sun, Y.C.: An effective recommendation method for cold start new users using trust and distrust networks. Inf. Sci. **224**, 19–36 (2013)
3. Victor, P., Cornelis, C., De Cock, M., Teredesai, A.M.: Key figure impact in trust-enhanced recommender systems. AI Commun. **21**(2), 127–143 (2008)
4. Zhang, D.Q., Hsu, C.H., Chen, M., Chen, Q., Xiong, N., Lloret, J.: Cold-start recommendation using bi-clustering and fusion for large-scale social recommender systems. IEEE Trans. Emerg. Top.Comput. **2**(2), 239–250 (2014)
5. Schein, A.I., Popescul, A., Ungar, L., Ungar, H., Pennock, D.M.: Methods and metrics for cold-start recommendations. In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 253–260, (2002)
6. Bobadilla, J., Ortega, F., Hernando, A., Bernal, J.: A collaborative filtering approach to mitigate the new user cold start problem. Knowl.-Based Syst. **26**, 225–238 (2012)
7. Phelan, O., McCarthy, K., Bennett, M., Smyth, B.: Terms of a feather: content-based news recommendation and discovery using twitter. In: Clough, P., Foley, C., Gurrin, C., Jones, G.J.F., Kraaij, W., Lee, H., Mudoch, V. (eds.) ECIR 2011. LNCS, vol. 6611, pp. 448–459. Springer, Heidelberg (2011)
8. Park, S., Chu, W.: Pairwise preference regression for cold-start recommendation. In: Proceedings of the third ACM Conference on Recommender Systems, pp. 21–28 (2009)
9. Zhou, K., Yang, S.H., Zha, H.Y.: Functional matrix factorizations for cold-start recommendation. In: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, pp. 315–324 (2011)
10. Claypool, M., Gokhale, A., Miranda, T., Murnikov, P., Netes, D., Sartin, M.: Combining content-based and collaborative filters in an online newspaper. In: Proceedings of ACM SIGIR Workshop on Recommender Systems, vol. 60 (1999)
11. Lin, J., Sugiyama, K., Kan, M., Chua, T.: Addressing cold-start in app recommendation: Latent user models constructed from twitter followers. In: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 283–292 (2013)
12. Su, X.Y., Khoshgoftaar, T.: A survey of collaborative filtering techniques. Adv. Artif. Intell. **4**, 2009 (2009)
13. Liu, N., Meng, X.R., Liu, C., Yang, Q.: Wisdom of the better few: cold start recommendation via representative based rating elicitation. In: Proceedings of the fifth ACM Conference on Recommender Systems, pp. 37–44 (2011)
14. Zhang, M., Tang, J., Zhang, X.C., Xue, X.Y.: Addressing cold start in recommender systems: a semi-supervised co-training algorithm. In: Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 73–82 (2014)
15. Kim, B.M., Li, Q.: Probabilistic model estimation for collaborative filtering based on items attributes. In: Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence, pp. 185–191 (2004)
16. Yu, K., Schwaighofer, A., Tresp, V.: Probabilistic memory-based collaborative filtering. IEEE Trans. Knowl. Data Eng. **16**(1), 56–69 (2004)