Experimental Analysis on Cross Domain Preferences Association and rating Prediction

Zhenhua Dong
Nankai University College of Information Technical Science
Weijin Road 94, Tianjin, China
(86)13920497817
nkdongzhh@gmail.com

Qian Zhao
Funshion Online Technologies Co., Ltd
Zhichun Road Jinqiu International Building B0801 Beijin China
(86)13716282982
qiankun925@gmail.com

ABSTRACT
Cross domain recommendation and preferences association are emerging research topics. In this paper, we will study the two topics through experimental analysis methods: firstly, we use folksonamy to analyze the preferences association among different domains; secondly, we analyze the feasibility of cross domain rating prediction based on KNN model. The experimental results report the associative tag pairs of users' preferences on items across domains. In addition, we report the cross domain prediction results here.

Categories and Subject Descriptors
H.3.3. **[Information Storage and Retrieval]**: Information Search and Retrieval – *information filtering*. I.2.6. **[Artificial Intelligence]**: Learning

General Terms
Algorithms, Experimentation

Keywords
Cross domain preferences association, cross domain recommendation, tag, collaborative filtering, KNN.

1. INTRODUCTION
Recommender system is an effective way to help people to cope with the problem of information overload. Most of the currently available recommender systems predict users' interest for items in a specific domain. The cross domain learning can transfer useful knowledge from one domain to another related domain. This knowledge can be used to analyze the associations among different domains, or implement the cross domain recommendation. Cross domain recommendation has important practical significance: (1) People have various kinds of interests, the cross domain recommendation can recommend items in different domains based on the users' interests to items in a domain; (2) Cross domain recommendation can solve the semi-cold-start problem [1], and (3) increase the novelty and serendipity of recommendation [2].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
KDD'12, August 12–16, 2012, Beijing, China.
Copyright 2012 ACM 978-1-4503-1555-5/12/08...$15.00.

Our research studies two tasks in cross domain recommendations: (1) Analysis on the preferences association among different domains; (2) cross domain rating prediction. For the first task, we compute the users' preferred tags, and then mine the preference association rules based on the tags from different domains. For the second task, we adopt the cross domain KNN model to compute the user's neighbors in candidate domain, and predict the user's preferences on items in target domain based on his neighbors' preferences.

2. RELATED WORK
2.1 Cross Domain Recommendation
Cross domain recommendation transfers the knowledge from people's behaviors (such as rating, tagging) in one domain, and predict people's preference on items in another domain. Several cross-domain recommendation approaches have been proposed recently. Winoto and Tang[1] applied the KNN model to predict users' rating on the items from different domains with a small scale data set. Pan et al. proposed CST (coordinate system transfer) [3] to adapt the learned latent features of users and items from candidate domain to improve the learning of latent features of users and items in target domain. Li et al. presented CBT [4] and RMGM (rating-matrix generative model) [5] to learn the shared implicit cluster-level rating pattern, which can be used to alleviate the data sparseness. Pan et al. [6] applied TCF (Transfer by collective factorization) framework to transfer the rating knowledge from auxiliary data source in binary form to a target numerical rating matrix. Our work will apply cross domain KNN model to predict the users' preferences with more ideal data sets, and verify the effectiveness of the model.

2.2 Domain Correlation
Analysis on the correlations between user preferences is the precondition of cross domain recommendation [7]. The researchers proposed CLP (collective link prediction) and MCF (multi-domain collaborative filtering) methods to exploit the correlations among domains. Fernandez et al. [7] presented cross-domain semantic knowledge framework, and build the cross-domain semantic network. Compared to these studies, we make use of tags to bridge different domains, and analyze the users' preferences association with tags.

3. CROSS DOMAIN PREFERENCE ASSOCIATION BASED ON TAG
We present the approach of analyzing the cross domain preferences association based on tag, first introducing how to compute each user's preferred tags set, and then mining the

preference association rules based on their preferred tags in different domains.

3.1 Tag Preference

Sen et al. [8] investigated 11 different signals of a user's interest in tags, including tag apps, tag searches, tag ratings, movie clicks, movie ratings, tag quality. Only 2 of them (tag apps and movie rating) can be collected in our experiment, so we choose "Movie-rating" [8] algorithm to calculate the users' preferences on tag.

Sen et al. [8] found that the inferring preference algorithms performed better when they took into account the relevance of a tag to a movie. So firstly, we compute the relevance weighting between a tag and an item with TF-IDF:

$$w(i,t) = tf_{i,t} \cdot \log_2 \frac{N}{df_i}, \quad (1)$$

Where $tf_{i,t}$ is the number of occurrences of tag t in item i; df_i is the number of items containing t; N is the total number of items.

Secondly, we calculate a user's preference for a tag based on the user's rating for related items:

$$pref_{u,t} = \frac{\sum_{i \in I_t} w(i,t) \cdot r_{u,i}}{\sum_{i \in I_t} w(i,t)}, \quad (2)$$

Where $r_{u,i}$ is user u' rating to item i; $w(i,t)$ is the relevance weighting between tag t and item i. We ignore the items the user has not rated. When the tag preference is greater than threshold, it is the user's preferred tag.

3.2 Preferences Association

We mine the association rules between the user's preferred tags from two different domains. Following the original definition by Agrawal et al. [9], the problem of preferences association rule is defined as: Let $I = \{i_1, i_2 \ldots i_n\}$ be the users' preferred tag set in two different domains. Let $D=\{t_1, t_2 \ldots t_m\}$ be a set of transactions. Each transaction in D contains a user's preferred tag sets. We hope to mine an association X ➔ Y where X is a user's preferred tag in domain A, Y is the user's preferred tag in domain B. We select the interesting rules from all possible rules through measuring their *support*, *confidence* and *lift*:

$$Supp(X \cup Y) = \frac{Nab}{N}, \quad (3)$$

$$Conf(X \Rightarrow Y) = \frac{Supp(X \cup Y)}{Supp(X)} = \frac{Nab}{Na}, \quad (4)$$

$$Lift(X \Rightarrow Y) = \frac{Supp(X \cup Y)}{Supp(X) \cdot Supp(Y)} = \frac{Nab * N}{Na * Nb}, \quad (5)$$

Where *N* is the total number of users; *Na* is the number of users who prefer tag a in domain A; *Nb* is the number of users who prefer tag b in domain B; *Nab* is the number of users who prefers both a and b. The interesting rules satisfy the minimum thresholds on *support*, *confidence* and *lift*. We will report the preference association rules among 3 different domains in Section 5.

4. CROSS DOMAIN RECOMMENDATION BASED ON KNN

If we can acquire enough tag preference rules in two domains, the users' preferences in the two domains are related. We could make use of the user-based KNN model to predict the users' rating to items in different domains. As shown in Figure 1, there are mainly two parts in the algorithm framework: Neighbor Computation and Prediction.

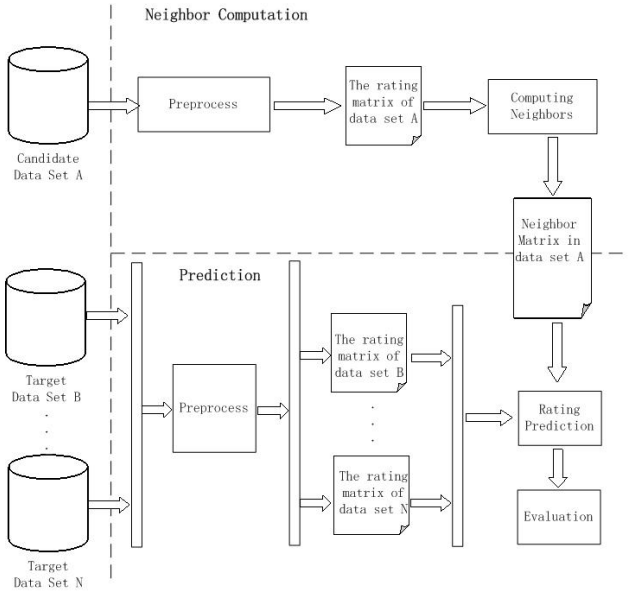

Figure 1. The framework of cross domain recommendation based on KNN.

4.1 Neighbor Computation

As shown in Figure 1, we compute the users' neighbors with the candidate data set. *Preprocess* is in charge of transferring the available rating data into the user-item *rating matrix*. Each row in the matrix is one user's rating vector, the adjusted cosine similarity between user u and v using this scheme is given by:

$$sim(u,v) = \frac{\sum_{i \in I} (R_{u,i} - \overline{R_i})(R_{v,i} - \overline{R_i})}{\sqrt{\sum_{i \in I} (R_{u,i} - \overline{R_i})^2} \sqrt{\sum_{i \in I} (R_{v,i} - \overline{R_i})^2}}, \quad (6)$$

Where $\overline{R_i}$ is the average rating of item i. If the similarity is greater than the threshold, u and v are neighbors. We compute the user's neighbors and store them in the *Neighbor Matrix*, which can be used to predict users' rating to items in other domains.

4.2 Prediction

As shown is Figure 1, the *Prediction* is in charge of cross domain rating prediction. *Preprocess* filters the incomplete data and transfers them from target domain into the user-item *rating matrix*. The *Rating Prediction* predicts the users' rating to items from target domain based on their neighbors in candidate domain with following formula:

$$R_{(u,i)} = \overline{r_u} + \frac{\sum_{v \in NN_u} (r_{(v,i)} - \overline{r_v}) \cdot s_{(u,v)}}{\sum_{v \in NN_u} s_{(u,v)}}, \quad (7)$$

Where $\overline{r_u}$ is user u's average rating to items in candidate domain (in the test dataset, we do not have u's ratings to items in target domain), $\overline{r_v}$ is user v's average rating to items in target domain, NN_u is u's neighbors set, $s_{(u,v)}$ is the similarity between u and v.

The *Evaluation* measures the quality of the cross domain rating prediction with the metrics: MAE, RMSE and Coverage [2].

5. EXPERIMENTS

5.1 Datasets

Douban[1] is a Chinese social website, which focuses on providing recommendation of cultural products, such as movie, book and music. We crawl the rating and tag information from Douban, and collect a dataset of 2,000 users and 49,000 items including movies, books and music.

Table 1 shows the items quantity, ratings quantity and average ratings in the 3 domains. The ratings scale is 1-5.

Table 1. Ratings Information

	Movie	Book	Music
Item Quantity	10447	25698	12846
Rating Quantity	75458	57323	28832
Average Rating	3.93	4.06	4.26

Table 2 shows the item quality, tag quantity and tag applications.

Table 2. Tag Information

	Movie	Book	Music
Item Quantity	9,715	15,399	9,226
Distinct Tag Quantity	10,453	14,494	8,680
Tag Applications	70,668,630	14963216	13,592,457

5.2 Results

Cross domain preferences association

Firstly, we compute the users' preferred tags with the formula (1) and (2). Table 3 shows the information of users' preferred tags.

Table 3. Users' preferred tags information

	Movie	Book	Music
{Preferred Tag, user} pair Quantity	17, 500	28, 262	11, 916
User Quantity	820	829	490

[1] www.douban.com

We acquire the interesting association rules by computing the *support*, *confidence* and *lift* based on the data in Table 3. In the experiment, we choose 2% as the *support* threshold, 33.3% as the *confidence* threshold, 1.5 as the *lift* threshold.

Finally, there are 109 interesting tag preferences association rules between book and movie, partly shown in Table 4; 70 association rules between movie and music, partly shown in Table 5; and 232 association rules between book and music, partly shown in Table 6. We translate the Chinese tags into English.

Table 4. Tag Preference Association Rules in book and movie

Book Tag	Movie Tag	Support	Confidence	Lift
Japanese comics	Japanese animation	0.0685	0.7015	1.9172
Reasoning	Thriller	0.0758	0.6753	1.5041
Female	Growth	0.0364	0.641	1.766
Japanese comics	Anime	0.0612	0.6269	1.8145
Reasoning	Japanese movie	0.07	0.6234	1.6385
Japanese literature	Japanese movie	0.1487	0.6145	1.615
Magic	Magic	0.0437	0.5769	1.5768
Network literature	Tony Leung	0.0321	0.5116	2.7207
Childhood	Childhood	0.0219	0.5	3.2981
Yi Shu	Romance	0.0204	0.4828	2.737
Art	Music	0.0758	0.4815	1.7384
Detective fiction	Horror	0.0219	0.4688	2.4736
Detective fiction	Japanese TV drama	0.0219	0.4688	4.2311
Photography	Biography	0.0306	0.4667	2.1201
Female	Erotica	0.0262	0.4615	1.8092
Humanity	History	0.0496	0.4595	1.7511
Memoirs	British file	0.0496	0.4595	1.5603
Harry Potter	Miyazaki	0.0204	0.4516	2.312

Table 5. Tag Preference Association Rules in movie and music

Movie Tag	Music Tag	Support	Confidence	Lift
Pixar	United States	0.0446	0.75	1.7663
BBC	OST	0.0403	0.6333	1.8414

Independent Film	Britpop	0.0212	0.625	2.9438
Europe	Britain	0.0425	0.625	1.887
Reasoning	OST	0.0403	0.5758	1.674
Spain	Indie	0.0403	0.5758	1.6844
Short animation	folk	0.0955	0.5488	1.5857
Reasoning	Japan	0.0382	0.5455	2.0887
Takeshi	Indie	0.0467	0.5366	1.5134
Shuji lwai	Indie	0.0722	0.5152	1.5071
Pixar	OST	0.0297	0.5	1.7977
BBC	OST	0.0297	0.4667	1.6779
Spain	Indie	0.0318	0.4545	1.5184
Childhood memories	Japan	0.0382	0.45	1.7232
Germany file	Folk	0.0403	0.4419	1.6133
Takeshi	Chinese	0.0361	0.4146	1.5139
Black	Britpop	0.0467	0.386	1.8179

Table 6. Tag Preference Association Rules in book and music

Book Tag	Movie Tag	Support	Confidence	Lift
Female	Female voice	0.0679	0.8205	1.2629
Italy	Folk	0.0361	0.7083	2.0468
Romantic	Pop	0.0616	0.7073	1.1856
Network literature	Pop	0.0616	0.6744	1.1304
Poems	Rock	0.0828	0.6724	1.0959
Romantic	Pop	0.0573	0.6585	1.1531
Photography	Rock	0.0616	0.6444	1.1118
Japanese comics	Euramerican	0.0849	0.597	1.0733
Germany	Indie	0.0722	0.5397	1.1934
Movie	OST	0.0552	0.5532	1.6084
Tsai Chih-heng	Hongkong and Taiwan	0.0234	0.4583	2.3212
Yi Shu	Cantonese	0.0276	0.4483	2.7421
Detective fiction	Japan	0.0297	0.4375	1.6753
Haruki Murakami	Indie	0.0701	0.3976	1.1631

We can find the semantics correlations from the Table 4, 5, 6 across the domains of book, movie and music. As shown in the 3 tables, 70.15% of the users who prefer the books tagged with "Japanese Comics" prefer the movies tagged with "Japanese animation" too; 75% of the users who prefer the movies tagged with "Pixar" also prefer the movies tagged with "United States"; 82.05% of the users who prefer the books tagged with "Female" has similar preferences on the music tagged with "Female voice". This kind of correlation in semantics can explain the existence of cross-domain user preferences, and implement the cross domain recommendation based on the associations. For example, we could recommend the books tagged with "Harry Potter" to the users who prefer the movies tagged with "Miyazaki".

Cross domain rating prediction

There are 3 kinds of rating dataset in the experiment: book, movie and music. We select one of them as candidate domain to generate neighbors, and select another of them as target domain to predict users' ratings. As a result, there are 6 experimental conditions for cross domain rating prediction. In order to compare the cross domain prediction with single domain prediction, we divide each single domain dataset into a training set and a test, 90% of the dataset is training set for generating neighbors, 10% of the dataset is testing set for rating prediction. Therefore, there are 3 experimental conditions for single domain rating prediction. Finally, we combine the 3 kinds of datasets together and divide the total dataset, 90% as training set and 10% as testing set. Consequently, there are totally 10 experimental conditions, all of which select the same algorithm and experimental parameters: neighbor similarity threshold is 0.5, neighbors' common items number threshold is 5.

We choose *MAE*, *RMSE* and *coverage* as the evaluation metrics:

$$MAE = \frac{\sum_{i=1}^{N} |p_i - r_i|}{N}, \tag{8}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(p_i - r_i)^2}, \tag{9}$$

$$Coverage = \frac{N}{N_{all}}, \tag{10}$$

Where p_i is a user's predictive rating to an item, r_i is the user's actual rating to the item, N is the number of all prediction. N_{all} is the number of all ratings in candidate domain dataset (testing dataset).

Table 7 shows the results of 10 conditions, including single-domain, cross-domain, and combined-domain. The first column consists of the domains where items are to be recommended, the second column shows the domains that are used to generate the neighbors. The third, fourth and fifth columns reports MAE, RMSE and Coverage accordingly. It is interested to note that: the

predictive error is not always the least when the target domain and candidate domain are same. However, if we choose the movie dataset as the candidate domain to generate the neighbors, we can get the least predictive errors, even when the target domains are book and music. We adjust the algorithm parameters, such as similarity threshold, common items number threshold, and still acquire the similar results: the neighbors generated from movie dataset can bring the least predictive errors. The predictive error (MAE=0.7735, RMSE=0.9946) of combined condition (the last row of the Table 7) is greater than most of the experimental conditions (row 1, row 4-9).

The results indicate that the accuracy of rating prediction depends on the selection of the target domain and the dataset.

Table 7. Rating Prediction Results

Target Domain	Candidate Domain	MAE	RMSE	Coverage
Movie	Movie	0.7484	0.9523	0.7669
	Book	0.8128	1.0475	0.5493
	Music	0.8303	1.0755	0.3521
Book	Book	0.7589	0.9743	0.1621
	Movie	0.7566	0.9607	0.5557
	Music	0.7625	0.985	0.1876
Music	Music	0.7449	0.9628	0.1908
	Movie	0.7222	0.9028	0.6448
	Book	0.7269	0.9255	0.3578
Combine	Combine	0.7735	0.9946	0.4855

6. Conclusion and Future Work

This paper analyses two specific questions in the cross-domain learning by experiment. Firstly, we explore the relevance of users' preferences for the items' tags among different domains by mining their association rules. Then we verify the feasibility of cross-domain rating prediction by the user-based KNN model. The experimental results show that, the correlation of users' preference for items among different domains can be expressed by the correlation of the tags on the items. Among related domains, ratings prediction with user-based KNN model is viable.

From the above, this paper in fact raises two issues to be addressed as well, which will be our future work. First, why are there such interesting associations between the tags in different domains? Second, how to improve the accuracy of cross-domain recommendation with effective models? We try to explain the former question with the Topic Model [10]. Topic Model is created to describe the generative process of the words in documents, which is applicable to any diadic data. We can simply merge the items from two different domains, and apply the Topic Model to their item-tag co-currence matrix to obtain the tags' distributions in the topic space. While the latent representation of each tag, e.g. topics, is possibly different across different domains, we could learn the subtle differences through analyzing the semantic of the tags, which also requires future experimental validation. Then the similarity between tags can be computed based on the distributions, which could explain the associations.

For the latter question, when we compute the similarity between users, the role of common items and their ratings are coarse-grained, which brings two problems here. First, although two users rated similarly on an item, they may be not interested in the same aspect of the item, which will cause a deviation. This problem also exists in the single-domain recommendation. It has been proved that we can get a better accuracy of rating prediction by building more sophisticated preferences of the users with factor models [11]. The second problem is that there is a deviation between the semantic spaces of the two domains. For example, the meaning of the same tag in book domain and movie domain may be different. A possible solution is to abstract up further to establish a common semantic space for them. By applying the Topic Model to the simply merged items from different domains to acquire the items' distributions in the topic space, and combine them with the users' ratings, we can obtain user's preference distributions in the topic space. The similarities based on these distributions are further utilized in the user-based neighborhood model. To note, simply merging items from different domains is also problematic, because the Topic Model is not able to perceive the fact that the items are cross-domain. The Topic Model could be adjusted to be aware of the items' heterogeneity. It is a challenging and exciting research direction to establish a unified semantic framework for the knowledge transformation.

7. ACKNOWLEDGMENTS

We would like to thank Cong Wang for the help in crawling the data. This work is funded by China Scholarship Council.

8. REFERENCES

[1] P. Winoto, T. Tang, If You Like the Devil Wears Prada the Book, Will You also Enjoy the Devil Wears Prada the Movie? A Study of Cross-Domain Recommendations. *New Generation Computing*, 26(3): 209-225. 2008.

[2] J.L. Herlocker, J.A. Konstan, L.G. Terveen, and J.T. Riedl, "Evaluating Collaborative Filtering Recommender Systems," *ACM Trans. Information Systems*, 22(1): 5-53, 2004

[3] W. Pan, E. W. Xiang, N. N. Liu, and Q. Yang, "Transfer learning in collaborative filtering for sparsity reduction," *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, pages 230 – 235. July 2010.

[4] B. Li, Q. Yang, and X. Xue, "Can movies and books collaborate? cross-domain collaborative filtering for sparsity reduction," *Int'l Joint Conf on Artificial Intelligence*, pages 2052 – 2057, 2009.

[5] B. Li, Q. Yang, and X. Xue, "Transfer learning for collaborative filtering via a rating-matrix generative model," *International Conference on Machine Learning*, pages 617 – 624, 2009.

[6] W. Pan, N. N. Liu, E. W. Xiang, and Q. Yang, "Transfer learning to predict missing ratings via heterogeneous user feedbacks," *Int'l Joint Conf on Artificial Intelligence*, pages 2318 – 2323, 2011.

[7] I. Fern´andez-Tob´ıas, I. Cantador, M. Kaminskas, and F. Ricci. A generic semantic-based framework for cross-domain recommendation. *In Proceedings of the 2nd International Workshop on Information Heterogeneity and Fusion in Recommender Systems*, HetRec '11, pages 25 – 32, New York, NY, USA, 2011. ACM.

[8] S. Sen , Jesse Vig , John Riedl, Tagommenders: connecting users to items through tags, *Proceedings of the 18th international conference on world wide web*, April 2009.

[9] R. Agrawal , R. Srikant, Fast Algorithms for Mining Association Rules in Large Databases, *Proceedings of the 20th International Conference on Very Large Data Bases*, pages 487-499, September, 1994.

[10] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, 50(3):993 – 1022, 2003.

[11] Y. Koren, Factorization meets the neighborhood: a multifaceted collaborative filtering model, *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, August, 2008.