Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.12785/amis/090332

Auxiliary Domain Selection in Cross-Domain Collaborative Filtering

Chang Yi, Ming-Sheng Shang and Qian-Ming Zhang*

Web Sciences Center, School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, People's Republic of China

Received: 2 Aug. 2014, Revised: 3 Nov. 2014, Accepted: 4 Nov. 2014
Published online: 1 May 2015

Abstract: The problem of data sparsity largely limits the accuracy of recommender systems in collaborative filtering model. To alleviate the problem, cross-domain collaborative filtering was proposed by harnessing the information from the auxiliary domains. Previous works mainly focused on improving the model of utilizing the auxiliary information yet little on the selection of auxiliary domains, although it is observed that the result of recommendation depends on the characteristics of auxiliary dataset. In this paper, we study the validity of cross-domain collaborative filtering by movie recommendation via different auxiliary domains of different movie genres. Through extensive experiments we find that the number of overlapping users between target domain and auxiliary domain is an indicator of choosing beneficial domains, while the low Kullback-Leibler divergence between non-overlapping user ratings, rather than the overlapping user ratings, is much more significant. The results are helpful in selection of auxiliary domains in cross-domain collaborative filtering.

Keywords: collaborative filtering, transfer learning, cross-domain, recommender system, KL divergence

1 Introduction

People in online systems usually feel lost when they need to select a suitable production due to the information overload problem. One solution to this preoblem is recommender system. Recommender system is a kind of information filtering tool that seeks to predict the preference of a user to an item. Due to its importance, in recent years recommender system is widely investigated by many experts in a vatiety of applications, and many types of recommendation algorithms have been proposed [1,2]. Therein, Collaborative Filtering (CF) model [3,4] is most successfully applied. The CF family recommendation algorithms normally require the support from a large amount of information on the behaviors history of users, and the accuracy of prediction is largely depending on the density of the given rating datasets. In online systems, however, the data is usually very sparse since a large portion of users only rate a very limited number of items. Sparsity problem has become a major bottleneck of recommender systems [5]. To alleviate this problem, numerous solutions have been proposed, most of which are by introducing addition information, for example, cross-domain collaborative filtering by harnessing the information from the auxiliary domains [6], and CF method with social tags [7] and so on.

Cross-domain collaborative filtering (CDCF) is a powerful tool to solve the data sparsity problem, which learns useful knowledge from other domains [6]. The idea behind the method is reasonable, for example, a user has few activities or is even new in the movie recommender system (target domain), but his interest can be well expressed in a related domain (auxiliary domains), such as the book store. So, by using CDCF method we can recommend movies to this user with the help of the histories of his reading perference. Transfer learning is a typical and effective method in CDCF. In [8], by considering the auxiliary knowledge as a CODEBOOK, Li proposed an algorithm based on a shared rating pattern for solving adaptive transfer learning problems. Latter, the idea was extended to a probabilistic model named Rating-Matrix Generative Model [9] to solve collective transfer learning (multi-task learning) problems in CF. Followed this idea, lots of methods are proposed to built the bridge between various of domains including cross-domain collaborative filtering over time [10], cross heterogeneous user feedbacks [11,12] and Cross-domain

* Corresponding author e-mail: msshang@uestc.edu.cn

topic learning [13]. All these methods achieved success on accurate recommendation. However, it is noticed that the prediction accuracy is not always good for every auxiliary domain, and little attention has been paid on this topic. Therefore we are inspired to study how to choose an effective auxiliary domain for a target domain. We expect to find some indices that can help to select proper auxiliary domain in advance.

In this paper, we use the benchmark dataset used in recommender system research, i.e., the Movielens dataset collected by the Grouplens group, to study the efficiency of Cross-domain collaborative filtering on different auxiliary domain, the 18 genres movies. And we adopt the Rating-Matrix Generative Model [9] method based on CDCF model due to its virtue in collecting multi-domains. Through extensive experiments we find that if the auxiliary domain is highly correlated with the target domain on rating pattern, indicated by low Kullback-Leibler (KL) divergence [14,15], the recommendation accuracy will be more likely to be improved. Specifically, this correlation is not significant unless we consider the non-overlapping users between the target domain and auxiliary domain. Compared with the consideration of overlapping users, the non-overlapping users should be paid more attention. The results are very helpful to select a contributing auxiliary domain in cross-domain collaborative filtering systems.

The rest of the paper is organized as followed: materials and methods are introduced in Section 2; in Section 3 we describe the experiments and analyze the results; Section 4 presents the conclusion and outlook.

2 Materials and Method

2.1 Dataset

The Movielens dataset collected by the Grouplens group in Minnesota University is a very good dataset for our study. Since it is the benchmark dataset widely used in recommender system research, it is public available, and for this paper's focus, it provides movie genres so that can be divided into many auxiliary domains conveniently. We use the 1M dataset which including 1 million ratings rated by 6040 anonymous users to 3952 movies. The ratings range from 1 to 5 where 5 means the user likes the movie very much and vice visa. The movies can be classified into 18 different genres according to the tags (a genre is defined to correspond to a domain in this paper). According to these movie genres, users are also divided into 18 groups. Some basic information are shown in Table 1, where ID is valid and consistent through the whole paper. N_{user} and N_{movie} are the numbers of users and movies respectively; $\langle K_{user} \rangle$ denotes the average number of movies rated by each user, calculated by $|E|/N_{user}$, where $|E|$ is the number of ratings in a given genre; Similarly $\langle K_{movie} \rangle$ denotes that how many ratings are rated to each movie, written as $|E|/N_{movie}$. $\langle r \rangle$ is the average rating values and the σ_r is the corresponding standard deviation. The domains are listed in the increasing order based on N_{user}.

It should be pointed that, these groups are allowed to be overlapped because users usually have more than one kind of taste. And as can be seen in Table 1, the largest user group contains 90% of total users.

To evaluate the accuracy of recommender system in this paper, the initial dataset is divided into two parts: the training set E^T and the probe sets E^P. The training set is treated as known information while the probe set is used for testing and no information in this set is allowed to be used for the recommendation. The test set is randomly sampled and always contains 20% ratings of the whole ratings for a domain, namely $|E^T| : |E^P| = 8 : 2$. Since there are 18 genres dataset, when considering a target domain, the other 17 datasets are used as auxiliary domain respectively.

2.2 Method

There are many cross-domain collaborative filtering models proposed to handle different application scenarios, among which the Rating-Matrix Generative Model (RMGM) is one of the effective cross-domain methods based on probability model [9]. This approach lends itself well to an adequate modeling of collaborative auxiliary effects. It assumes that there is a latent user-item rating pattern in the related domains as a "bridge" to estimate the missing ratings. The ratings rated by users to items can be clearly described by a rating matrix. Suppose there are m users and n items, we can build a $m \times n$ matrix M, where the entry M_{ui} means the rating of user u to item i. And each matrix can be clustered according to the user (item) similarity. Then the common knowledge is abstracted as an overlapping implicit cluster-level rating matrix from both matrices. The entries are the average ratings of the corresponding user-item co-clusters. Then the predicted rating can be obtained by a rating function in terms of the combination of the latent user-cluster and item-cluster variables.

A simple example of generating cluster-level rating matrix is shown in Figure 1. The original matrix A and B in subfigure (a) and (b) are the related rating matrix. Notice that, the label "a" in Matrix A is different from that in Matrix B and so as to b, c, d, e and f. Subfigure (c) and (d) correspond to Matrix A and B respectively by permuting the rows (users) and columns (items) with the rule that the same ratings need to be grouped into clusters (the red boxes) as far as possible. At last, the corresponding cluster level rating matrix can be built in subfigure (e): entries in blue box correspond to subfigure (c) and entries in red box correspond to subfigure (d).

Suppose we are given several rating matrices in related domains, and there are K user clusters $\{C_u^1,...,C_u^k\}$, and L item clusters $\{C_v^1,...,C_v^l\}$ in the shared cluster-level rating

Table 1: Basic information of every movie genre.

ID	Genre	N_{user}	N_{item}	$\langle k_{user} \rangle$	$\langle k_{item} \rangle$	$\langle r \rangle$	σ_r
1	Documentary	2243	110	3.53	15.98	3.93	1.0672
2	Western	4100	67	5.04	73.61	3.64	1.2096
3	Film-Noir	4150	44	4.40	173.91	4.08	0.8698
4	Musical	4754	113	8.74	166.13	3.67	1.2123
5	Animation	4808	105	9	37.23	3.68	1.1709
6	Fantasy	4850	68	7.48	180.6	3.45	1.2841
7	Mystery	5133	104	7.83	365.25	3.67	1.1810
8	Children's	5283	250	13.66	48.35	3.42	1.3476
9	Horror	5300	339	14.41	1123.32	3.22	1.5019
10	Crime	5662	201	14.05	1807.75	3.71	1.1615
11	War	5769	141	11.88	202.14	3.89	1.1348
12	Adventure	5894	281	22.73	1185.42	3.48	1.2757
13	Sci-Fi	5911	274	26.61	1512.44	3.47	1.3392
14	Romance	5961	459	l24.75	321.4	3.61	1.1380
15	Thriller	5989	485	31.67	692.26	3.57	1.2247
16	Action	6012	495	42.82	530.84	3.49	1.2848
17	Crime	6031	1163	59.12	2528.94	3.52	1.2560
18	Drama	6037	1493	58.73	5291.48	3.77	1.0937

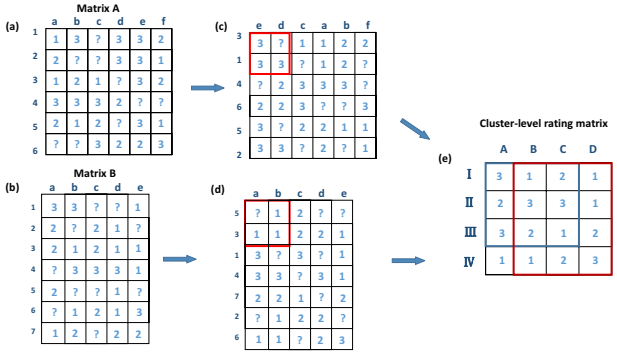

Fig. 1: A simple example of implementation of RMGM.

patterns (illustrated in Figure 1 (e)). Then the cluster-level rating matrix can be expressed as the probability of user u in C_u^k and item v in C_v^l simultaneously, represented by $P(C_u^k, C_v^l|u,v)$ (the random variable u and v are assumed to be independent from each other). The missing rating of user u to item v can be obtained by $r \cdot P(C_u^k, C_v^l|u,v)$, where r is the average rating in user-item co-cluster (C_u^k, C_v^l).

However, each user-item co-cluster (C_u^k, C_v^l) can also have various ratings with different probabilities. Thus the rating function can be defined as

$$\begin{aligned} f_R(u,v) &= \sum_r P(r|u,v) \\ &= \sum_r \sum_{k,l} P(r|C_u^k, C_v^l) P(C_u^k, C_v^l|u,v) \\ &= \sum_r \sum_{k,l} P(r|C_u^k, C_v^l) P(C_u^k|u) P(C_v^l|v) \end{aligned} \quad (1)$$

The probabilities for each u and v can be iteratively computed until the probabilities converge to a stable state based on the expectation-maximization algorithm. Then the predicted user-item joint ratings can be obtained by Eq. (1). For example, suppose we want to estimate the missing rating 3d (3rd row and d column) in the Matrix A in Figure 1. The probability of this rating belongs to the user cluster I and item cluster A is 0.9, therein the average rating of users is 3. With the rest 0.1 probability the user and item are clustered in user-item group IC whose average rating is 2. Then the estimation of the missing rating 3d is 2.9 according to the belonging probability and average ratings of user-item group.

2.3 Metric

There are many metrics to evaluate the performance of recommender systems [16]. The commonly used metrics are the Mean Absolute Error (MAE) and Root Mean Squared Error. The information retrieval metrics such as precision and recall are also useful. Recently, the diversity, novelty and coverage are also considered as important aspects in evaluation. In the RMGM model, we are to predict the rating sores in the probe set, so we use MAE to evaluate the accuracy of the prediction produced by applying different auxiliary domains, defined as

$$MAE = \sum_{r_{ua} \in r^P} \frac{|r_{ua} - \widehat{r_{ua}}|}{|r^P|}, \quad (2)$$

where u and a represent the selected user and item respectively. r_{ua} is the predicted rating, $\widehat{r_{ua}}$ is the actual rating in the probe set, and $|r^P|$ is the number of ratings in the probe set. Obviously, the lower value of MAE indicates a more accurate prediction.

3 Results and Analysis

3.1 Prediction Accuracy on Auxiliary Domains

In our experiment, we consider every domain (movie genre) as the auxiliary information when given a target domain. Denote that, the pair of target domain and auxiliary domain (x,y) is different from (y,x). So there are 18×18 different pairs in total, and MAE will be calculated for every pair of domains, named $MAE(x,y)$ where y is the ID of the target domain and x corresponds to the auxiliary domain. For an intuitive comparison, we use the MAE difference $MAE(x,y) - MAE(y,y)$, written as $M(x,y)$ replacing $MAE(x,y)$. Because of the definition of MAE, the negative value of $M(x,y)$ indicates that the auxiliary domain x is effective for the target domain.

Figure 2 represents the cumulative distribution of $M(x,y)$ for all the pairs of target domain y and auxiliary domain x. For each pair of y and x, the MAE is calculated by 10 independent runs. Clearly auxiliary domains play positive roles in only 50% of the cases, which means that randomly select a domain is usually ineffective. Therefore, how to choose a suitable auxiliary domain is critical to the performance of cross-domain collaborative filtering models. In the next section, we are going to find some correlation betweent the character of auxiliary domain and the predicted accuracy.

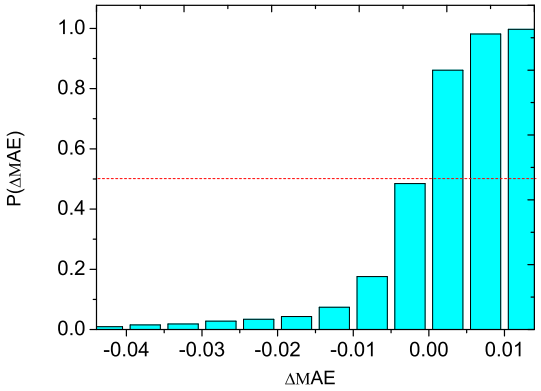

Fig. 2: The cumulative distribution of $M(x,y)$.

3.2 User Confidence Coefficient Analysis

The overlapping of user groups is firstly taken into consideration as it is natural to expect that an auxiliary domain sharing more common users with the target domain may bring more available information. As Table 1 shows, Documentary Genre attracts much fewer people than other genres. It means all the user groups are highly overlapped. To quantify the overlapping degree, we calculate the confidence coefficient between two selected domains, e.g. S and T, by

$$UCC(S,T) = \frac{|\,U_S \cap U_T\,|}{|\,U_T\,|}, \tag{3}$$

where U_S denotes the group of users who rate the movies in S, and $|\,U_S\,|$ denotes the number of elements in U_S. Clearly, higher value of $UCC(S,T)$ means more users in U_T are covered by the users in U_S. The confidence coefficients of users between every pair of target domain and auxiliary domain are shown in Figure 3, where the little square (x,y) corresponds to the value of $UCC(x,y)$, where x is the index of column (the ID of auxiliary domain), and y is the ID of target Domain. The ID of these domains is in accordance with the ID in Table 1. As the target domain and auxiliary domain are different, Figure 3 is asymmetric. The $UCC(S,T)$ between different genres rang from 0.3 to nearly 1. And the larger the user group forms, the higher the $UCC(S,T)$ is.

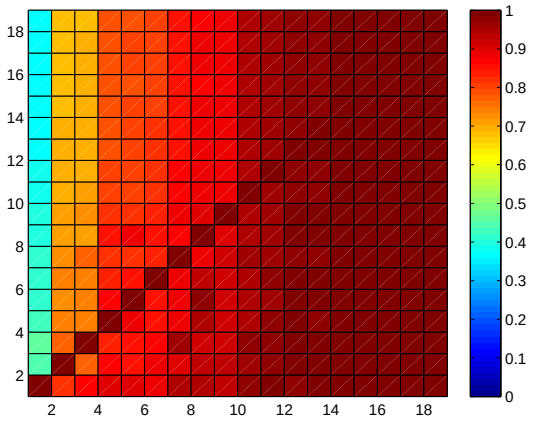

Fig. 3: Confidence coefficients of users between every pair of domains.

To examine the correlations, we introduce the Pearson Correlation Coefficient (PCC) [16], defined as

$$PCC(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}, \tag{4}$$

where X and Y are two vectors, and $\overline{X}$ is the mean value of elements in X. In this case, the two vectors are $UCC(:,y)$ and $M(:,y)$, respectively, where ":" means the

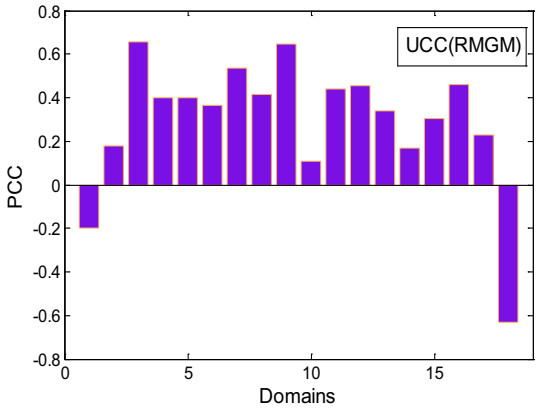

Fig. 4: The correlations between the MAE difference and the user confidence coefficient.

set $\{1,2,3,\cdots,18\}$ but except y which is the ID of the target domain.

Figure 4 shows the Pearson Correlation Coefficient between $UCC(:,y)$ and $M(:,y)$ for every domain considered as the target domain, where the ID of these domains is in accordance with the ID in Table 1. Clearly we can see that most of the correlation coefficients are positive. This phenomenon is a little surprising that the more users in the target domain can be covered by the auxiliary domain, the less accurate predictions we will get. From Figure 4, we can also see that the user confidence coefficient is not stable, as $UCC(:,y)$ is negatively related to $M(:,y)$ when y is set to 1 and 18 in Figure 4. Thus we can say that the overlapping of users between the auxiliary domains and the target domain is only a rough indicator.

3.3 Kullback-Leibler Divergence Analysis

In the RMGM model we need to group the users and movies into clusters according to the ratings. So the similarities between users' ratings are also valuable to take into consideration. The rating pattern in each domain can be viewed as a distribution of rating r ($r \in [1,5]$). And Kullback-Leibler (KL) divergence [14,15,17] is very suitable to measure the rating difference. KL divergence is a non-symmetric measure of the difference between two probability distributions P and Q. Specifically, the KL divergence of Q from P is a measure of the information lost when Q is used to approximate P. It is defined to be

$$KL(P,Q) = \sum_i P(i) ln \frac{P(i)}{Q(i)}, \tag{5}$$

which is the expectation of the logarithmic difference between the probabilities vectors P and Q. If and only if Q is consistent with P, $KL(P,Q) = 0$. Here P and Q represent the probabilities vectors of ratings. Because ratings are ranging from 1 to 5, P and Q are set to be 5-dimensional vectors of the auxiliary domain and the target domain respectively. To keep consistency, Eq. (5) are rewritten as

$$KL(X,Y) = \sum_i X(i) ln \frac{X(i)}{Y(i)}, \tag{6}$$

where X and Y are the probability vectors of the auxiliary domain x and the target domain y respectively.

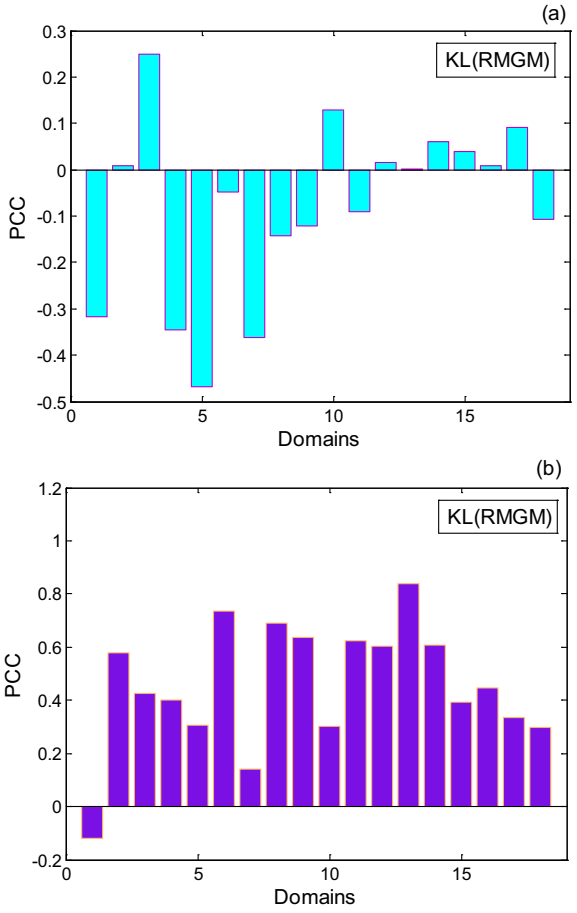

Fig. 5: Correlations between the MAE difference $M(:,y)$ and $KL(:,y)$ for RMGM model.

Figure 5 presents the correlations between the MAE difference $M(:,y)$ and $KL(:,y)$ for RMGM model, where $KL(:,y)$ in Subfigure (a) is calculated including the whole users in both domains, while in subfigure (b) it is calculated only according to the non-overlapping users. The ID of these domains is in accordance with the ID in Table 1. We still examine the PCC between this factor $KL(:,y)$ and $M(:,y)$ for every target domain, shown in

Figure 5 (a). Observing this irregular phenomenon, we think this metric is used in an improper way as the differences between rating pattern are mainly caused by the non-overlapping users. As we expect, for a pair of domains, the average KL divergence between the overlapping users is lower than that between different users. Those are 0.2008 and 0.3169 respectively. So we further examined the KL divergence of ratings by removing out the ratings rated by the overlapping users. Significantly, all the correlation coefficients are positive only except the first one as shown in Figure 5 (b). We can also find nearly all the correlations are highly positive of which the values are above the empirical value 0.2. That is to say, if the rating pattern of the non-overlapping users in an auxiliary domain is more similar to the pattern of non-overlapping users in the target domain, this auxiliary domain will be more effective than other domains.

4 Discussion and Conclusion

The main contribution of this paper is to clarify that different movie genres play different roles as auxiliary domain in CDCF model, and we tried to find an effective indicator to select a contributing auxiliary domain. The two indices, i.e., the overlapping rate, and the Kullback-Leibler divergence are found as the character of the auxiliary domain. The overlapping of user groups is a nice one but not very exact. The rating pattern of the whole users in a domain is not exact either. But if we eliminate the effect of the overlapping users, the rating pattern of the rest users is much correlated the effectiveness of the auxiliary domain.

However, there still a lot of on going works. First, the overlapping of user groups is negatively related to the effectiveness of the auxiliary domain. It is difficult to explain and only a conjecture is given in this paper. It is very probable that this phenomenon is correlated to the next one — the rating pattern of the non-overlapping users is a nice indicator positively correlated to the effectiveness. The results in the papers may shed some light on this problem of auxiliary domain selection in applying cross-domain collaborative filtering models, we expect that to further understanding the dataset may give more profound results..

Acknowledgements

This work is supported by the National Natural Science Foundation of China (61370150 and 91324002). QMZ acknowledges the program of outstanding PhD candidate in academic research by UESTC (No.YBXSZC 20131034).

References

[1] D.H Park, H.K Kim, I.Y Choi, J.K Kim, Expert Systems with Applications **39(11)**, 10059-10072 (2012).
[2] L. Lü, M. Medo, C.H. Yeung, Y.C. Zhang, Z.K. Zhang, T. Zhou, Physics Reports **519**, 1-49 (2012).
[3] F. Cacheda, V. Carneiro, D. Fernndez, V. Formoso, ACM Transactions on the Web,**5**, 2 (2011)
[4] G. Adomavicius, A. Tuzhilin, IEEE Trans. Knowl. Data Eng **17**, 734 (2005).
[5] M.D. Ekstrand, J.T. Riedl, J.A. Konstan, Foundations and Trends in Human-Computer Interaction **4**, 175-243 (2011).
[6] B. Li, Tools with Artificial Intelligence (ICTAI), 23rd IEEE International Conference on IEEE, 1085-1086 (2011).
[7] M.S. Shang, Z,K. Zhang, T. Zhou, Y.C. Zhang. Physica A **389**, 1259-1264 (2010)
[8] B. Li, Q. Yang, X. Xue, Proceedings of the 21st international joint conference on Artificial intelligence, California, USA, 2052-2057 (2009).
[9] B. Li, Q. Yang, X. Xue, Proceedings of the 26th Annual International Conference on Machine Learning (ACM), Canada, Montreal, 617-624 (2009).
[10] B. Li, X.Q. Zhu, R.J. Li, C.Q. Zhang, X.Y. Xue, X.D. Wu, Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Volume Three, AAAI Press, Spain Barcelona, Catalonia, 2293-2298 (2011).
[11] W. Pan, E.W. Xiang, N.N. Liu, Q. Yang, Proceedings of 24th AAAI Conference on Artificial Intelligence, AAAI **10**, 230-235 (2010).
[12] W. Pan, N.N. Liu, E.W. Xiang, Q. Yang, Proceedings of the Twenty-Second international joint conference on Artificial Intelligence (IJCAI'11), Barcelona, Catalonia, Spain, 2318-2323 (2011).
[13] J. Tang, S. Wu, J.M. Sun, H. Su, Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 1285-1293 (2012).
[14] S. Kullback, R.A. Leibler, The Annals of Mathematical Statistics **22**, 79-86 (1951).
[15] J.H. Lin, Information Theory, IEEE Transactions on **37**, 145-151 (1991).
[16] J.L. Herlocker, J.A. Konstan, L.G. Terveen, J.T Riedl, ACM Transactions on Information Systems **22**, 5-53 (2004).
[17] B.D. Shai, J. Blitzer, K. Crammer, F. Pereira, Advances in neural information processing systems **19**, 137-144 (2007).

Chang Yi is a master student of Computer Science in University of Electronic Science and Technology of China. Her research interests are mainly on complex networks and recommender system.

Ming-Sheng Shang received his Ph.D Degree in Computer Science from University of Electronic Science and Technology of China (UESTC). Now he is a full professor of UESTC. His research interests include data mining, complex networks, cloud computing and their applications.

Qian-Ming Zhang is a doctoral student in University of Electronic Science and Technology of China. He focuses on studies on complex networks, including link prediction and recommender system.