Privacy-Conscious Cross-Domain Recommendation via Domain Adaptation

HAO NIU[1] KEI YONEKAWA[1] MORI KUROKAWA[1] CHIHIRO ONO[1]
DAICHI AMAGATA[2] TAKUYA MAEKAWA[2] TAKAHIRO HARA[2]

Abstract: Item embedding techniques, which can capture the item dependency, have been popularly applied to the recommendation systems. As for the cross-domain recommendation, the embedded item vectors of different domains are expected to be aligned in a common vector space (i.e., domain adaptation), which generally requires to share some information between domains (e.g., user-item interaction data). However, it may be difficult for different domains to share such kind of user-relevant data due to the privacy policy. In this work, we propose a method to mine the item dependency between domains without sharing user-relevant data, based on which to align the item vectors of different domains. Experimental investigations confirm the effectiveness of our proposed method.

Keywords: Item embedding, Cross-domain recommendation, Item dependency, Privacy-conscious, Domain adaptation

1. Introduction

An impressive increase in the application of item embedding techniques (e.g., matrix factorization and word2vec-based item embedding methods) on recommendation systems has been observed nowadays. These techniques generate low dimensional feature vectors for items [1-2], based on which recommendation can be performed according to the items' vector similarity. The item embedding techniques developed from word2vec can effectively capture the item co-occurrence dependency based similarity, and have received much attentions recently [2].

Applying item embedding for cross-domain recommendation has also been considered [3-4]. A main idea is to align the item embedding among different domains (domain adaptation) by exploiting some shared data (e.g., user-item interaction data). However, it may be unrealistic in many cases, since different domains typically belong to different companies and sharing user-relevant data may violate privacy policy [3].

In this work, we propose a privacy-conscious method, which first mine the item co-occurrence across domains without sharing user-relevant data and then perform the domain adaptation using these mined co-occurrence items. We do the experiments on the foursquare dataset to confirm our proposed method.

2. Methodology

We use Figure 1 for the description of our method. In Figure 1, there are two domains (e.g., a restaurant domain and a shop domain). We consider the scenario that some users interact with items of both domains (common users - the users in the circle), while some users interact with only one domain (uncommon users - the users out of circle). This scenario is usual since in reality there are always some users who belong to both of the domains, even for the domains belonging to two companies (e.g, some users who use Foursquare also use Yelp).

If the userID is linked between domains, we can merge the common users' interaction data per user. Based on the merged per-user interaction data it is possible to directly generate the aligned item vectors between domains like [4]. We can also use the merged per-user interaction data to mine the item co-occurrence across domains similar to Market Basket Analysis (MBA): From the merged per-user interaction data, we can specify per user the timeslots around the time that item x_i of Domain1 is interacted with (timeslot-similar to the basket). We then select per user some timeslots that item x_i is not interacted with, to calculate the following probabilities for the item y_j of Domain2: $P(y_j|x_i)$ - the probability that y_j is interacted with in the timeslot that x_i is interacted with; $P(y_j|\bar{x}_i)$ - the probability that y_j is interacted with in the timeslot that x_i is not interacted with.

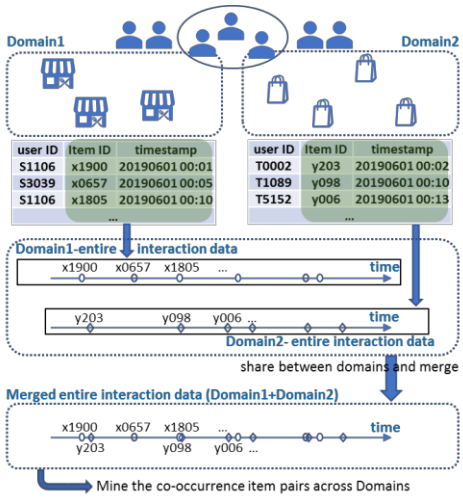

Figure 1. Description of our method.

If we get $P(y_j|x_i)>P(y_j|\bar{x}_i)$, we can treat that $<x_i,y_j>$ is the co-occurrence item pair between Domain1 and Domain2. That is because: Assuming x_i and y_j are co-occurrence items with $P(y_j|x_i)>P(y_j)$, we can easily derive that $P(y_j|x_i)>P(y_j|\bar{x}_i)$ since $P(y_j)= P(x_i)P(y_j|x_i)+P(\bar{x}_i)P(y_j|\bar{x}_i)$.

However, the analysis above is unrealistic in many cases since the user-item interaction data is user-relevant and userID linkage is also intractable for different domains (especially for the domains belonging to different companies). We thus propose the following method to mine item co-occurrence across domains which does not need to know per-user interaction data or perform the userID linkage.

1 KDDI Research, Inc., Chiyoda, Tokyo 102-8460, Japan
2 Osaka University, Suita, Osaka 565-0871, Japan

As shown in Figure 1, we extract the itemID and timestamp data of the user-item interaction data of both domains. These data in each domain generates an entire interaction data without distinguishing users, which is shared. Since the shared entire interaction data consist of only itemID and timestamp, it is not user-relevant. We then merge the entire interaction data of the two domains, and extract the co-occurrence item pairs from the merged entire interaction data, using the following proposition.

Proposition 1: If x_i and y_j meet $P(y_j|x_i)>P(y_j|\bar{x}_i)$ in the merged per-user interaction data between domains, they also meet $P_E(y_j|x_i)>P_E(y_j|\bar{x}_i)$ in the merged entire interaction data, where,

$P_E(y_j|x_i)$ - the probability that y_j is interacted with in the timeslot that x_i is interacted with in the merged entire interaction data;

$P_E(y_j|\bar{x}_i)$ - the probability that y_j is interacted with in the timeslot that x_i is not interacted with in the merged entire interaction data.

Proof of Proposition 1: Assuming that Domain1 and Domain2 have M and N users respectively, and the number of the common users of Domain1 and Domain2 is C. The way x_i is interacted with in the merged entire interaction data includes two cases: 1. x_i is interacted with by a common user, the probability of which is C/M; 2. x_i is interacted with by an uncommon user, the probability of which is (M-C)/M. We first assume that different timeslots for x_i are not overlapped and we can derive Equation(1), where, $\bar{y}_j$ means that in the timeslot y_j is not interacted with; ">" is because $P(\bar{y}_j|x_i)=1-P(y_j|x_i)<1-P(y_j|\bar{x}_i)=P(\bar{y}_j|\bar{x}_i)$. If some timeslots for x_i are overlapped, $P_E(y_j|x_i)>P_E(y_j|\bar{x}_i)$ still holds, since the overlap will further enlarge $P_E(y_j|x_i)$ due to co-occurrence. Thus, Proposition 1 is proved. ▪

$$
\begin{aligned}
&P_E(y_j|x_i)\\
&=\frac{C}{M}\left\{1-P\left(\bar{y}_j|x_i\right)P\left(\bar{y}_j|\bar{x}_i\right)^{C-1}P\left(\bar{y}_j\right)^{N-C}\right\}\\
&\quad+\frac{M-C}{M}\left\{1-P\left(\bar{y}_j|\bar{x}_i\right)^{C}P\left(\bar{y}_j\right)^{N-C}\right\}\\
&>\frac{C}{M}\left\{1-P\left(\bar{y}_j|\bar{x}_i\right)P\left(\bar{y}_j|\bar{x}_i\right)^{C-1}P\left(\bar{y}_j\right)^{N-C}\right\}\\
&\quad+\frac{M-C}{M}\left\{1-P\left(\bar{y}_j|\bar{x}_i\right)^{C}P\left(\bar{y}_j\right)^{N-C}\right\}\\
&=\left\{1-P\left(\bar{y}_j|\bar{x}_i\right)^{C}P\left(\bar{y}_j\right)^{N-C}\right\}\\
&=P_E(y_j|\bar{x}_i) \qquad (1)
\end{aligned}
$$

Therefore, from the merged entire interaction data instead of the merged per-user interaction data, we can also mine the co-occurrence item pairs across domains. For the practical applications, similar to the MBA we can further define three metrics: Support=the number of times that x_i is interacted with, Confidence=$P_E(y_j|x_i)$ and Lift= $P_E(y_j|x_i)/P_E(y_j|\bar{x}_i)$, to decide the co-occurrence item pairs (here, Lift=1 equals to $P_E(y_j|x_i)=P_E(y_j|\bar{x}_i)$. And only if all of the three metrics are larger than their thresholds, $<x_i,y_j>$ are treated as the co-occurrence item pair.

Next, we can perform the item embedding in Domain1, and transfer the feature vectors of x_i to Domain2 for guiding the item embedding of Domain2 (domain adaptation). For example, we can use the embedded vectors of x_i as the vectors of y_j for each co-occurrence item pair $<x_i,y_j>$ and freeze these vectors when performing the item embedding in Domain2.

3. Experiments and Conclusions

We evaluate our method on the Foursquare dataset of New York collected from 2012/04 to 2013/02 [5], for which we add the root categories (domains) to each interaction using the categories.csv in https://github.com/gunarto90/foursquare-category. Since users tend to visit nearby POIs, there should be some nearby POIs of different domains which are co-occurrence items. We consider 3 domains of the dataset (***1***.Food; ***2***.Outdoors and Recreation; ***3***.Shop and service), and only the POIs with no less than 5 check-in records are used in our experiments. By setting the timeslot as 10 minutes before and after the time that x_i is interacted with, and the thresholds of Support, Confidence and Lift as 10, 0.5 and 2 respectively, we mine the co-occurrence item pairs between these 3 domains. The numbers of mined pairs are shown in Table 1.

Based on the mined co-occurrence item pairs, we analyze the Recall@50 performance without and with domain adaptation (denoted by Baseline and DoAda respectively) using the data of common users between domains. The item embedding is performed using the Skip-gram model of word2vec with vector size 100. For the domain pair ***S***-***T***, we calculate the user vector as the average vector of the items that the user interacted with in ***T***, and recommend 50 items of ***S*** according to the cosine similarity between item vectors and the user vector. The Recall@50 results are also shown in Table 1, from which we can observe that the Recall@50 performance becomes better with the domain adaption based on the mined co-occurrence item pairs.

Tabel 1. Results of Recall@50[%] (average of 50 experiments)

	1-2	***2-1***	***1-3***	***3-1***	***2-3***	***3-2***
Mined pairs	13	13	10	10	4	4
Baseline	1.59	6.94	1.68	7.53	7.06	7.10
DoAda	**2.50**	**7.61**	**2.28**	**8.20**	**7.45**	**8.18**

In future, we will further study on the effect of hyperparameters (e.g., the length of timeslot and the thresholds), evaluate on the domains of different companies, and consider more effective method for the domain adaptation using the mined co-occurrence item pairs.

Reference

[1] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. Computer 8 (2009), 30–37.

[2] Hugo Caselles-Dupré, Florian Lesaint, and Jimena Royo-Letelier. 2018. Word2vec applied to recommendation: Hyperparameters matter. In Proceedings of the 12th ACM Conference on Recommender Systems. ACM, 352–356.

[3] Chen Gao, Xiangning Chen, Fuli Feng, Kai Zhao, Xiangnan He, Yong Li, and Depeng Jin. 2019. Cross-domain Recommendation Without Sharing User-relevant Data. In The World Wide Web Conference. ACM, 491–502.

[4] Yaqing Wang, Chunyan Feng, Caili Guo, Yunfei Chu, and Jenq-Neng Hwang. 2019. Solving the Sparsity Problem in Recommendations via Cross-Domain Item Embedding Based on Co-Clustering. In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining. ACM, 717–725.

[5] Dingqi Yang, Daqing Zhang, Vincent. W. Zheng, and Zhiyong Yu. 2015. Modeling User Activity Preference by Leveraging User Spatial Temporal Characteristics in LBSNs. IEEE Transactions on Systems, Man, and Cybernetics: Systems 45, 1 (2015), 129–142.

Acknowledgments This research was partially supported by JST CREST Grant Number J181401085, Japan.