Exploiting Cross-Domain Sequence Data for Cross-Domain Recommendation

Hao NIU†, Kei YONEKAWA†, Mori KUROKAWA†, Chihiro ONO†, Daichi AMAGATA††, Takuya MAEKAWA††, and Takahiro HARA††

† KDDI Research, Inc.,
Garden Air Tower, 3-10-10, Iidabashi, Chiyoda-ku, Tokyo, 102-8460, Japan
†† Osaka University,
1-5 Yamadaoka, Suita, Osaka 565-0871, Japan
E-mail: †{ha-niu,ke-yonekawa,mo-kurokawa,ono}@kddi-research.jp,
††{amagata.daichi,maekawa,hara}@ist.osaka-u.ac.jp

Abstract The recommendation techniques based on the sequence data of user actions attract much research attentions recently, which can mine the short-term item dependency for recommending users their interested items. Till now, these techniques are mainly analyzed in the single domain. The application of them for cross-domain recommendation is still limited. In this work, we first integrate the sequence data of a part of common users between domains, and then exploit the integrated cross-domain sequence data for the cross-domain recommendation. There are now many domains (e.g., companies) partnering with each other to improve user experience and expand themselves, some of which have realized common userID or userID linkage. Therefore, the common user information can be obtained between domains. In addition, it is also possible to share the data of a part of common users between domains with users' consent. We thus can integrate the sequence data of these users in order of time, based on which it is able to directly mine the short-term cross-domain item dependency. Then, we can recommend the items of one domain to the users of the other domain based on the mined short-term cross-domain item dependency. We evaluate the above approach on the Retailrocket data and further deploy it to our practical use scenario, both of which confirm its effectiveness.

Key words sequence, item dependency, cross-domain recommendation, data integration, item embedding

1 Introduction

Many of traditional recommender systems apply the matrix factorization or matrix completion techniques to the user-item interaction matrix [1,2]. The user-item interaction matrix (e.g., user-item rating matrix) aggregates all the interactions between each user and item, from which the long-term item dependency of the entire period considered can be captured. However, it is difficult to mine the short-term item dependency from user-item interaction matrix, since it ignores the time/order information involved in the interaction data.

In many practical use scenarios, the short-term item dependency is more important. For example, users of E-commerce usually change their preferences over time, and the item dependency is assumed to exist only in a short period of time (e.g., one or a few days). The real interaction data of most companies records not only user-item interactions but also the interaction time, based on which the interaction data can be processed to be a large number of user interaction sequences. Using the processed interaction sequences, it is able to mine both the long-term and short-term item dependency for making recommendations. There have been a plenty of research works on recommendation techniques using user sequence data, such as sequential pattern mining, item embedding and neural network techniques listed in [2–4].

On the other hand, more and more companies tend to partner with each other for serving users better and advance the development of themselves, which increases the demand for cross-domain recommendation (CDR). Some of these companies realizes partnership by means of common userID [5] or userID linkage. For example, the common userID is widely used by the companies belonging to a same group; the userID linkage can be realized by allowing login with other companies' userID. It is also possible to realize inexact userID

linkage based on IP, CookieID, AAID/IDFA or other data. Therefore, we can make an effective CDR by leveraging the data of these common users.

Till now, the CDR techniques based on user sequence data are still limited, among which the techniques leveraging the data of common users become more rare. Two approaches applying item embedding can be found in [5,6]. Authors of [5] first generate the feature vectors of the users and items for different domains respectively by performing word2vec model on the user sequence data. Then, they use Canonical Correlation Analysis (CCA) to align the vectors of different domains to a common vector space for cross-selling between domains. Different from [5], authors of [6] use both the single-domain and cross-domain sessions of the common users to obtain aligned feature vectors between domains within a unified framework, for solving the spare problem of one domain with the help of other domains (a session can be regarded as a short-term sequence).

However, [6] does not exploit the cross-domain sequence data (CDSD) directly for CDR, but first mines cross-domain item co-cluster correlation from the CDSD and then uses the mined co-cluster correlation to align the item embedding of different domains, which results in a complicated framework. *In fact, after integrating the sequence data of the common users between domains, we can treat the integrated CDSD as the sequence data of one single domain, and directly apply the existing single-domain sequence-aware recommendation techniques on the CDSD to realize CDR* [7]. That is because we can capture both the long-term and the short-term cross-domain item dependency from the CDSD, even using the existing single-domain sequence-aware recommendation techniques. We name this approach as CDSD4CDR in this work.

To confirm the effectiveness of CDSD4CDR, we evaluated its performance using Retailrocket data and also deployed it to our practical use scenario. The single-domain sequence-aware recommendation technique that we apply to the CDSD4CDR is the word2vec based item embedding, since word2vec model can effectively capture the shore-term item co-occurrence. We also performed no CDSD based approaches for CDR. Specifically, the no CDSD-based approaches include cross-domain collaborative filtering using matrix factorization which does not exploit the sequence data, and the CCA approach of [5] which exploits the sequence data of each domain separately. The results illustrate that CDSD4CDR significantly outperforms no CDSD based approaches for both the Retailrocket data and our practical use scenario.

2 Related Works

2 1 Sequence-Aware Recommendations

Compared to the traditional recommendation approaches which focus on the long-term item dependency, the sequence/session-aware recommendation techniques can capture not only the long-term item dependency but also the short-term item dependency. The existing sequence/session-aware recommendation techniques are well-summarized in [2–4]. For example, the Markov Models uses the Markov chain to model the short-term item dependency [8]; the item embedding techniques originated from word2vec can also captures the short-term item dependency (co-occurrence) [9]; the Neural Models use the neural network to learn both the long-term and short-term dependency [10].

2 2 Cross-Domain Recommendations

Most general approach for CDR is collective matrix factorization [11] based one, which jointly factorizes the user-item interaction matrices with the assumption of user or item overlap between domains. Some techniques are also proposed which do not require that the users or items in the two domains are overlap. For example, [12] deals with the problem of data sparsity in a target domain by transferring cluster-level user-item rating patterns from other domains. Owing to the flexibility of the back-propagation scheme and the effectiveness on real world dataset, recent years have seen an explosion of neural network based CDR models [13, 14]. There are also some but not too much works exploiting the sequence data for CDR like [5, 15]. In addition, [6] is the work exploiting the CDSD for CDR.

3 Methodology

In this section, we will describe how to integrate the cross-domain sequence data and how to realize CDSD4CDR using the existing single-domain sequence-aware recommendation techniques as is described in [7]. Without loss of generality, we use two practical heterogeneous domains, an E-commerce site and an ad-platform, as shown in Figures 1 for the description. In the E-commerce site, users' product viewing and purchase actions are recorded; on the ad-platform, users' web page viewing actions are collected when they view the web pages with the ad tags of the ad-platform. The userID sets of the E-commerce site and the ad-platform are denoted respectively by **E**={E-IDm, m=1,2,...,M} and **A**={A-IDn, n=1,2,...,N}. We perform the userID linkage by inserting ad tags of the ad-platform into the web pages of the E-commerce site. If a user logs into the E-commerce site with userID (E-IDi), this userID is linked to the cookieID (A-IDj) generated by the ad-platform. It is regarded that the two IDs belong to the same user (i.e., the user is listed as a common user). In

addition, we can share the action data of some listed common users between the two domains with their consent.

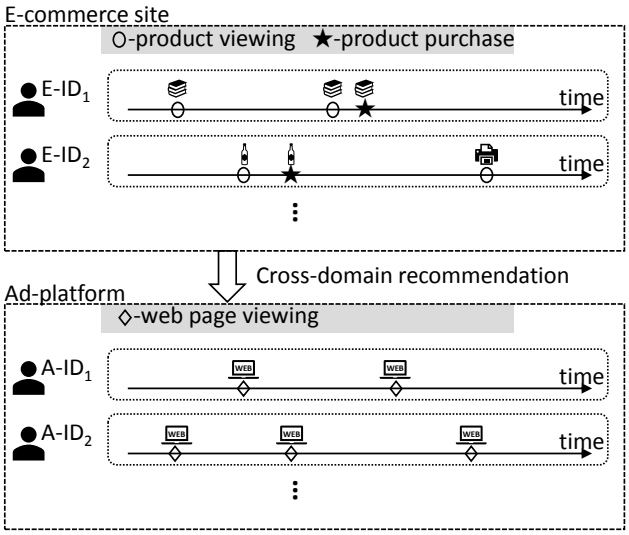

Figure 1 Cross domain recommendation for two heterogeneous domains.

We regard the products in the E-commerce site and web pages on the ad-platform as items, and our purpose is to recommend the products of the E-commerce site to the users of the ad-platform only based on their actions on the ad-platform. Traditionally, we can concatenate the user-item interaction matrices of the two domains according to the common user information, and then use the matrix factorization techniques (e.g., Singular Value Decomposition - SVD or Non-negative Matrix Factorization - NMF) to realize the CDR. Also, the CCA approach of [5] realizes CDR by exploiting the user sequence data of each domain separately. Specifically, the CCA approach includes four steps:

- Generating the items' and users' vectors based on the sequence data for each domain respectively;
- Deriving the transform matrix for each domain by using the CCA algorithm and the common user information;
- Transforming/Aligning the items' and the users' vectors of each domain to the common latent space;
- Cross-domain recommendation according to the cosine similarities of the users' and the items' vectors.

On the other hand, we can integrate the sequence data between domains per common user to extract and leverage the short-term cross-domain item dependency for the CDR (CDSD4CDR). The sequence data of each common user is integrated based on another important common information between domains - *time*. For example, if E-ID1 and A-ID2 in Figures 1 are linked to be the same user, their sequence data is integrated as Figures 2, where PX and WX indicate the itemIDs of the two domains. After the integration for all the available common users, a dataset can be generated

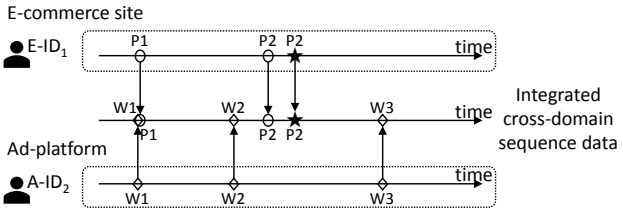

Figure 2 Cross-domain sequence data of one common user.

in which each element is a cross-domain sequence including one common user's actions of both domains. Therefore, the cross-domain item dependency can be studied by applying the existing single-domain sequence-aware recommendation techniques to this dataset (CDSD), e.g., sequential pattern mining, item embedding, neural network and other techniques listed in [2–4].

The cross-domain item dependency between the E-commerce site and the ad-platform may be the cross-domain item co-occurrence, since some users may refer related web pages when viewing or purchasing products in the E-commerce site. Therefore, we realize CDSD4CDR using item embedding in this work, which is realized by three steps:

- Generating the CDSD by integrating the sequence data of the available common users between domains;
- Learning the aligned items' and users' vectors of both domains directly by applying the item embedding techniques on the CDSD;
- Cross-domain recommendation according to the cosine similarities of the users' and the items' vectors.

4 Experiments

To evaluate the effectiveness of CDSD4CDR, we compare it with matrix factorization based CDR, and the CCA approach based CDR [5]. Aiming at a simple deployment on the practical systems and a fair comparison with [5], we apply the skip-gram model of word2vec to the CDSD to directly learn the aligned item embedding of different domains. As for the matrix factorization, we use two popular techniques: SVD and NMF. The comparisons are first performed on Retailrocket data (https://www.kaggle.com/retailrocket/ecommerce-dataset/home), and then we deploy the above approaches to our practical use scenario.

4.1 Retailrocket Data

The Retailrocket data is collected from 20150503 to 20150918, which includes three kinds of user actions: view, adding to cart and purchase of items. There are totally 25 root categories, and the root category of each item can be obtained based on the item properties. We generate two

Table 1 Statistics of the divided two domains

Domains	*Interactions*	*Users*	*Items*
Domain1	1291374	668816	104278
Domain2	1209124	601517	80966
Common users: 34304			

Table 2 Performance of CDR between the divided two domains of the Retailrocket data

Approach	[Domain1, Domain2]		[Domain2, Domain1]	
	Recall@50	MRR@50	Recall@50	MRR@50
CCA	0.062	0.009	0.052	0.008
SVD	0.080	0.011	0.079	0.015
NMF	0.099	0.013	0.106	0.014
CDSD4CDR	**0.187**	**0.031**	**0.174**	**0.045**

domains by dividing the 25 root categories into two sets randomly. The root category sets of each domain are shown as follows: Domain1 - [1482, 859, 1057, 1532, 653, 1394, 1182, 250, 1579, 1692, 791, 755, 1600]; Domain2 - [679, 231, 140, 803, 378, 431, 1698, 1224, 1490, 1452, 395, 659]. The items of each domain are the items whose root categories belong to the root category sets of this domain. To avoid the item overlap between domains, for the items with multiple categoryids we randomly use one categoryid. The statistics of the two domains are shown in Table 1.

Next, we perform the CDR between the two domains. In our experiments, 80% of the whole common users are used for the training (training-users), and the common users left are used for the test (test-users). Specifically, we integrate the sequence data of training-users between domains to CDSD, and then learn the items' vectors of both domains simultaneously by applying the skip-gram model on CDSD. Then, we generate the users' vectors in each domain as the average vectors of the items that they interacted with in the same domain. The recommendation is performed by recommending the items of one domain to the test-users in the other domain, by calculating the cosine similarities between the item vectors and the user vectors of the other domain.

For the CDSD4CDR and the CCA approach, we use the skip-gram model with hierarchical softmax and window size 10. The vector size of all vectors is set to be 100. We did 5 experiments for the CDR between Domain1 and Domain2. The recall@50 and Mean Reciprocal Rank (MRR)@50 performance is analyzed by comparing each user's recommendation set with the set of the items interacted with by this user. The average performance is shown in Table 2, where [X, Y] means recommending the items of domain X to the users of domain Y. We can observe that CDSD4CDR significantly outperforms the matrix factorization (SVD and NMF) and the CCA approach, because CDSD4CDR can effectively capture the short-term cross-domain item dependency of the E-commerce data.

Table 3 Statistics of the extracted data

Domains	*Interactions*	*Users*	*Items*
E-commerce	3M	0.13M	0.02M (products)
Ad-platform	48M	0.11M	0.15M (web pages)
Common users: 0.08M			M: millions

Table 4 Performance of CDR in the practical use scenario

Approach	(1)		(2)	
	Recall@50	MRR@50	Recall@50	MRR@50
CCA	0.014	0.002	0.014	0.002
SVD	0.093	0.024	0.083	0.016
NMF	0.099	0.022	0.083	0.014
CDSD4CDR	**0.126**	**0.044**	**0.114**	**0.032**

Since the long-term item dependency is not significant in the E-commerce data compared to the short-term item dependency, the matrix factorization techniques perform worse than CDSD4CDR. In addition, the CCA approach performs worst, because it is based on the single-domain sequence data and difficult to extract the cross-domain item dependency.

4 2 Practical Use Scenario

We deploy the above approaches on the practical use scenario described in Section 3, which consists of two heterogeneous domains, an E-commerce site and an ad-platform. By inserting ad tags of the ad-platform into the web pages of the E-commerce site, we linked about 0.18 million IDs (users) between two domains. We exact these common users' interaction data of both domains during the first four weeks of 12/2017, to illustrate the CDR performance of the above approaches. The statistics of the extracted data are given in Table 3, and the performance evaluation is similar to that of Retailrocket data. Two kinds of CDR performance averaged over 5 experiments are illustrated in Table 4: (1) Generating the test-users' vectors per week and evaluating the recommendation performance using products interacted with by these users in the same week; (2) Generating the test-users' vectors per week and evaluating the recommendation performance using products interacted with by these users in the next week. (2) corresponds to the practical application of CDR. From Table 4, we can observe that CDSD4CDR also outperforms other approaches greatly. The results confirm again that the short-term item dependency is much more significant than the long-term item dependency in the E-commerce data.

5 Conclusions

In this work, we evaluate the performance of exploiting the integrated cross-domain sequence data for the cross-

domain recommendation (CDSD4CDR). By comparing to the no CDSD based CDR approaches on the Retailrocket data and in a practical use scenario, the effectiveness of CDSD4CDR is confirmed. Although in this work we used the E-commerce and ad-platform data, CDSD4CDR also works for other kinds of data if the short-term item dependency is significant than the long-term item dependency, e.g., the news recommendation data. In future, we will analyze the performance of CDSD4CDR on the different kinds of data. We will also apply different single-domain sequence-aware recommendation techniques to the CDSD4CDR and design new techniques dedicated to the CDSD4CDR like [6], and evaluate their performance.

Acknowledgments

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

References

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

[2] Massimo Quadrana, Paolo Cremonesi, and Dietmar Jannach. Sequence-aware recommender systems. *ACM Computing Surveys (CSUR)*, 51(4):66, 2018.

[3] Dietmar Jannach. Keynote: Session-based recommendation–challenges and recent advances. In *Joint German/Austrian Conference on Artificial Intelligence (Künstliche Intelligenz)*, pages 3–7. Springer, 2018.

[4] Shoujin Wang, Longbing Cao, and Yan Wang. A survey on session-based recommender systems. *arXiv preprint arXiv:1902.04864*, 2019.

[5] Masahiro Kazama and István Varga. Multi cross domain recommendation using item embedding and canonical correlation analysis. In *RecSys Posters*, 2017.

[6] Yaqing Wang, Chunyan Feng, Caili Guo, Yunfei Chu, and Jenq-Neng Hwang. 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*, pages 717–725. ACM, 2019.

[7] Hao Niu, Kei Yonekawa, Mori Kurokawa, and Arei Kobayashi. Transfer learning among time series data. *IEICE Technical Report*, 118(284):425–428, 2018.

[8] Mehdi Hosseinzadeh Aghdam, Negar Hariri, Bamshad Mobasher, and Robin D Burke. Adapting recommendations to contextual changes using hierarchical hidden markov models. *RecSys*, 15:241–244, 2015.

[9] Mihajlo Grbovic, Vladan Radosavljevic, Nemanja Djuric, Narayan Bhamidipati, Jaikit Savla, Varun Bhagwan, and Doug Sharp. E-commerce in your inbox: Product recommendations at scale. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1809–1818. ACM, 2015.

[10] Massimo Quadrana, Alexandros Karatzoglou, Balázs Hidasi, and Paolo Cremonesi. Personalizing session-based recommendations with hierarchical recurrent neural networks. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*, pages 130–137. ACM, 2017.

[11] Ajit P Singh and Geoffrey J Gordon. Relational learning via collective matrix factorization. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 650–658. ACM, 2008.

[12] Bin Li, Qiang Yang, and Xiangyang Xue. Can movies and books collaborate? cross-domain collaborative filtering for sparsity reduction. In *Twenty-First International Joint Conference on Artificial Intelligence*, 2009.

[13] Guangneng Hu, Yu Zhang, and Qiang Yang. Conet: Collaborative cross networks for cross-domain recommendation. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 667–676. ACM, 2018.

[14] Ali Mamdouh Elkahky, Yang Song, and Xiaodong He. A multi-view deep learning approach for cross domain user modeling in recommendation systems. In *Proceedings of the 24th International Conference on World Wide Web*, pages 278–288. International World Wide Web Conferences Steering Committee, 2015.

[15] Dilruk Perera and Roger Zimmermann. Lstm networks for online cross-network recommendations. In *IJCAI*, pages 3825–3833, 2018.