Cold-Start and Data Sparsity Problems in Recommender System: A Concise Review

Nanthini M[1] and Pradeep Mohan Kumar K[2]
Dept. of Computing Technologies,
School of Computing,
SRM Institute of Science and Technology, Chennai
nm7342@srmist.edu.in, pradeepk@srmist.edu.in

Abstract. An enormous amount of data available on the e-commerce sites are ratings, ranks, reviews, opinions, complains, remarks, feedbacks, and comments about any item and it is difficult for the system to search the user interest and predict the user preference. The Recommender System (RS) came into existence, supports both customers and providers in their decision-making process. Nowadays, recommender systems are suffering from various problem such as data sparsity, cold start, scalability, synonymy, grey sheep, data imbalance, etc. One of the major problems to be considered for better recommendation is data sparsity. Cross Domain Recommendation (CDR) is one way to address data sparsity problem, cold start issue, etc. In most of the traditional system, Cross domain analysis is used to understand the feedback matrices by transferring hidden information and imposing dependencies across the domains. There is no vast comparison of existing research in CDR. This paper provides a definition of the problem, related and existing work on CDR for data sparsity and cold start, comparative survey to classify and analyze the revised work.

Keywords: Cross Domain Recommendation, Collaborative Filtering, Recommender System, Data Sparsity, Cold-Start

1 Introduction

RS is an information filtering system that tries to find the rating or preference given by the customer to the product in e-commerce websites. The most traditional recommendation methodologies are Collaborative Filtering (CF) and content-based filtering. Collaborative recommendation [1] is also known as a social recommendation, it provides recommendations based on similar people preferences, likes, dislikes, and reviews. The content-based recommendation is a cognitive filtering, it provides recommendation according to the content of an item and the user profile. Hybrid recommender systems combine the popular two approaches collaborative and content-based systems by collecting the user profiles learning the model based on the information and maintaining some useful records using an efficient information retrieval techniques and other content-based methods, and directly comparing the user profiles to find the similar users based on collaborative filtering [3] and provide recommendations. This represents that item can be suggested to the users in the form of recommendation list when items score is higher than the user's profile or are user's higher rating with a similar profile.

Monolithic hybridization is a technique that represents various levels of recommendation techniques in one algorithm while implementation. The single recommender component that combines multiple techniques by pre-analysing, processing and integrating various information sources. Hybridization is attained by making changes in algorithm to derive various kinds of input data. Fig.1 represents the types of RS in various perspective. In traditional systems, filtering techniques are providing a recommendation based on single domain information. The domain is one of the characteristics of recommender system and it is a particular thought of field, activity or user interest.

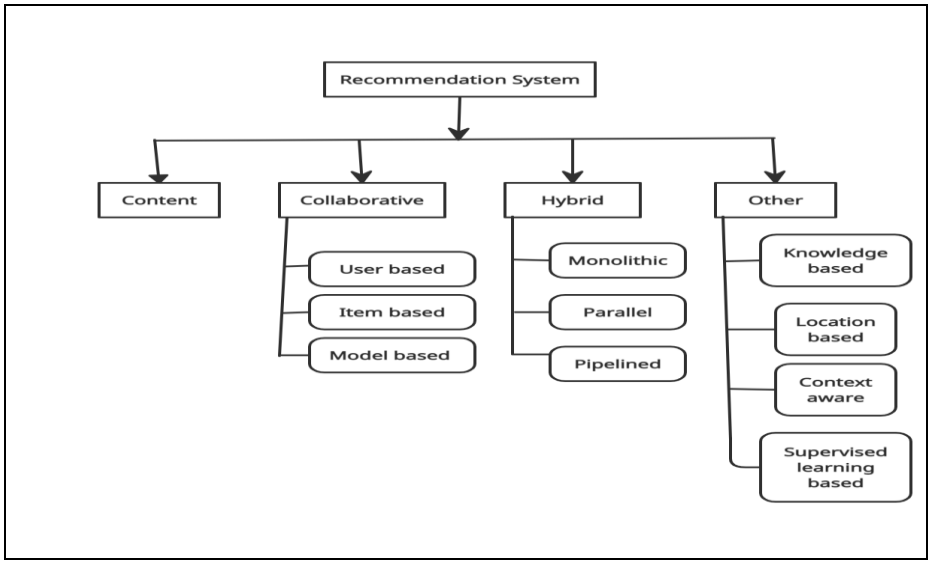

Fig.1. Types of recommender system

Considering two domains in RS, the numbers of attributes to be considered are larger and different compared to the single domain. Dimensionality Reduction (DR) [8] helps to reduce the overall dimensionality in two domains and increase the prediction accuracy of user preference in RS by discovering the hidden knowledge of the whole data set and transferring knowledge across the domains. The data available on the online websites are of various types such as continuous/discrete, numerical/categorical, structured /unstructured, etc. Continuous data is the range can take any value between its maximum and minimum value. Structured data refers to the information with the fully organized manner and can enable quick search algorithm. Unstructured data have a lack of structure which makes energy consuming task. Categorical data can take only certain values and is also known as discrete data. It can be classified as nominal or ordinal variables. One of the challenges is analyzing such variety of data and DR techniques can be applied for data analytics.

2 Related Work

This section discussed about various existing research on CDR, sparsity problem in other RS and explained the ways to address the problem faced by traditional RS.

2.1 Data Sparsity and Cold-Start in CDR

In social network, star-structured hybrid graph is used for making recommendation across the domains and the useful information is shared from data enriched domain to domain is having sparse data. Different factors such as transfer of item features considering popularity and consistency of the behavior are explored. Meng Jiang, et. al. introduced a method called Hybrid Random Walk (HRW) [10], which considers the above factors in order to identify the items can be transferable and share the knowledge between two different domains and also identify the relationships between the user and item. A CF based cross domain algorithm is developed in order to build a Linear Decomposition Model (LDM) by combining items in order to find relationship among total and local similarities of multiple domains [14]. Xu Yu, et. al. proposed an algorithm in order to compute the whole similarities in all considered domains and it is the fusion of various domain's local similarities. The weights are computed using two domains and greater than the weight in a single domain and it is reflected in the similarity measures in source domain. If the user preferences vary, it is difficult to do better recommendation with multiple domains.

A CDR is constructed based on attributes and improper classification. The traditional recommendation issue is considered as a rough unevenness categorization in the target domain which took feature vector of item and user and rating as label. Then the sparsity

problem is reduced by considering the useful vector of item and user in the target domain. Xu Yu et. al. Funk Singular Value Decomposition (SVD) is used to extract the useful features about users from two source domain for better recommendation. Later, an unevenness classification model [15] is constructed to rectify an unevenness classification problem that can efficiently solve the unequal distribution of rating. Data sparsity is one of the important problems in RS and it is dominant in newly constructed RS which is having insufficient data. CDR is considered as an efficient solution to the sparse data problem by transferring the sufficient information across the domains. While transferring information from one domain to another domain, three different cases were explored such as entities which are fully overlapped, partially overlapped and non-overlapped. Even though the entities are fully overlapped, they are having different format in each domain. The above cases reduced the overall performance of CDR in target domain. Qian Zhang et. al. proposed a Kernel Induced [16] RS based on knowledge transfer which interrelate the entities that are non-overlapped in two domains to solve the sparse data problem in RS.

Knowledge transfer from data enrich domain to domain is having sparse data play a vital role in addressing the data sparsity problem in RS. Source domain enriched with sufficient data whereas the target domain used the information from the source domain for good recommendation. Model based on Cluster-level rating is used in this work to investigate the sparsity problem without considering the overlapping entities in two domains. The existing methods have not addressed the accuracy of expected results in target domain. Ming He, et. al. proposed a model called an Adaptive Codebook Transfer Learning (ACTL) [17] to develop the codebook scale which stabilize the cost for computation and accuracy in prediction and suits the size and attributes of the source domain in CDR.

Alternate course selection process in colleges and universities is a difficult task for the students due to its unknown nature among the students. This leads to inappropriate alternate course selection and low achievements in course which compel the students to quit the course in the half way. To address the above problem and improving their selection process, Ling Huang, et. al. a cross-user-domain CF [18] is developed for the exact prediction of marks scored by each student with the help of score distribution scored by senior students. Then the top alternate courses which is having high marks scored by the senior students are recommended to the present students for their achievement. A citation-based recommendation in cross domain analysis [13] is developed for knowledge sharing. A cross-domain recommendation has evolved to help users in detecting similar knowledge across the domains and provide recommendation based on the shared information. The first technique is a simple keyword mapping. The second technique is co-citation selection.

In addition to data sparsity, data imbalance is considered as a challenging problem in CDR while transferring knowledge across the domains. This research work solves the above problem using three different learning methods such as representation, adversarial and transfer. Even though different transfer learning techniques performed well in this area, a novel RecSys-Discriminative Adversarial Network (DAN) concentrated on multiple problem such as data sparsity inside the domain and across the domain, data imbalance and knowledge transfer in the form of latent factors in CDR. The knowledge transfer is performed in an adversarial manner [19]. Four different neural architectures are developed and explored. Real time experimentation showed that the proposed technique performed well in target domain without labeled data and it is more flexible in multiple real-world scenarios and robust to cold start problem in RS.

Nowadays, online commercial businesses are more dependent on RS in order to identify the individual user's need and provide them a better recommendation. In order to provide efficient recommendation, a large amount of data with knowledge about the user and items are needed. By sharing enormous amount of data from one domain to another domain, users-items relationship was clearly explored to compliment the CF recommendation with greater accuracy in prediction. However, the efficient utilization of information across the domains is a difficult problem. Quan Do, et. al. proposed to explore latent features with similarities in multiple domains based on matrix factorization [5]. In this research work, features which are common in two domains and the features which are specific to the domains are analyzed. Both the information is very

useful in CDR. Especially the domain specific features are considered to transfer the knowledge across the domain.

CDR is an efficient system to address the sparsity problem in RS by considering knowledge from multiple domains. Wenxing Hong, et. al. proposed a Deep Neural Network based on Cross Domain technique [9] for the good recommendation. The proposed technique rectifies the prediction problem based on rating with the help of metadata information including reviews and items. The learning process happened in target domain as well as in source domains with hidden factors of users and products which increases the prediction accuracy. Mapping process and the optimization were experimented in both the domains that solves the sparsity problem effectively in target domain. CF and Matrix Factorization are failed to perform well in RS having no information about the user and the product that reduces the income of online-commercial business. It is a very challenging task to provide recommendation to the new user. Yaru Jin, et. al. proposed a novel Review Aware Recommendation [2] using cross domain analysis in order to enquire the new user recommendation in the E-commerce sites. In this work, reviews are collected using adjacency matrix and the preference vectors are take out from the domains in terms of specific and shared manner using migration model.

Cold start problem is also an important issue in RS having new users and items. RS with cold start users reduce the entire performance of the system. To address the above problem, Hanxin Wang, et. al. proposed a technique that merge e-shopping domain and the information from Ads. A CDR is developed using the deep learning algorithm, word to vector concept is employed to convert text information into latent information. Deep learning algorithm strengthen the performance of RS for sharing the knowledge from source to target domain. Text data is having enormous amount of knowledge compared to latent representation, R-metapath2Vec [6] is used to enhance the features in latent space. The existing studies on CDR is mainly focused on knowledge transfer from one domain to another domain in the same website. It is very difficult to share full information across the domain in different e-commerce sites due to some privacy concern and it leads to negative transfer problem. Hongwei Zhang, et. al. proposed a RS named as Selective Knowledge Transfer [4], that shares both hidden features of users and products from data enrich domain to sparse data domain and also address the negative transfer problem.

2.2 Collaborative Filtering Recommendation

Social recommender system [7] uses the Collaborative Filtering (CF) to make personalized recommendations. Most of the CF techniques mainly use the one-dimensional clustering methods to group users and items separately. However, the one-dimensional method usually neglects the necessary information in another dimension. Daqiang Zhang, et. al. uses the bi-clustering technique, that group both the user features and item feature concurrently in user-item matrix. Correspondingly, the bi-clustering method outperforms the one-way cluster method of overcoming the sparsity problem and reducing the high-dimensional matrices in recommender systems. The recommendation in social network gains more popularity and attaining a successful service for providing quality of service and user satisfaction. These applications enable the users to provide various implicit feedbacks and ratings in the web during their daily usage of social network. The users interacted in social networks like some items based on traditional RS filtering technique and can provide feedback about the item. With the exponential growth of online information, the sparse data gradually decrease the overall performance of RS quality and correlation factor of feedbacks. A novel Spatial Social Union (SSU) [12] is developed and it is a similarity measurement-based approach to calculate the similarity between the users that combines the interconnection between users and items in addition to location information.

3 Overview of CDR and Domain Level

This section describes the cross-domain analysis, latent variables, attributes mapping from user profiles to the latent variable and domain level attributes mapping to the latent variables.

3.1 Cross-Domain Analysis

A domain is considered as an action, or user preference. Most of the RS focused on one domain and they suggest recommendations within the domain where users gave their feedback and ranks. The combination of various domains into single domain makes the recommender engine provides recommendation list across the domain. An algorithm used in domain analysis is able to provide a recommendation of the product in the chosen domain to users who gave ratings only in the source domain. For example, if restaurant domain is referred as the source domain and the tourist spot domain is referred as the chosen domain and the knowledge is shared across restaurant domain and tourist spot domain.

3.2 Various Domain Level

- Attribute level (Chinese cuisine↔ Asian cuisine)
    - Cuisines share same kind of attributes with the different values for some attributes.
- Type-level (Restaurants↔ Tourist spot)
    - Different types, sharing some common features like location, climate.
- Item level (Shopping↔ Restaurants)
    - Entirely different set of features and attributes.
- System level (TripAdvisor↔ MakeMyTrip)
    - This level is having same kind of items and are collected and expressed in different ways and manner respectively.

3.3 Different Approaches in CDR

Knowledge aggregation is the process of combining user interests, likes and preferences from across the domains and provide recommendation in the target domain. User preferences can be clicks, ratings, logs and feedbacks are merged together for performing Knowledge linking. It is very useful to address cold-start problem.

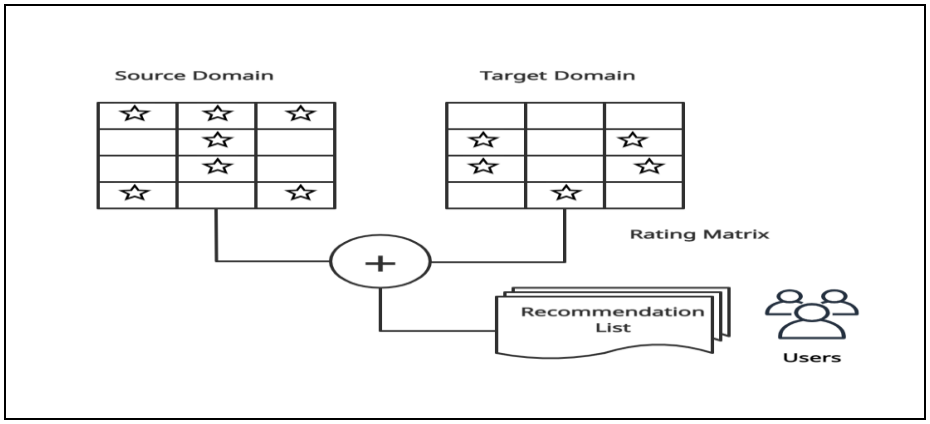

Fig. 2. Linking knowledge

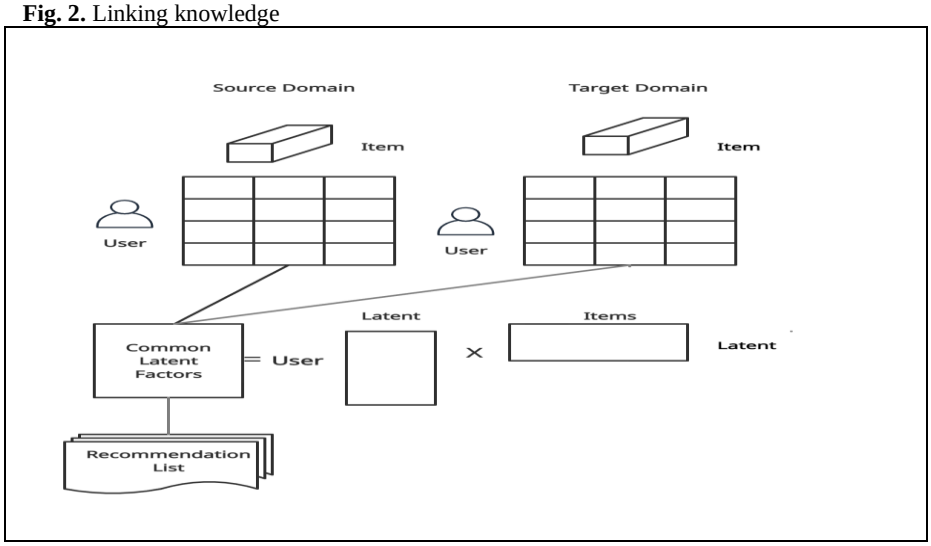

Fig. 3. Knowledge transfer

Fig. 2 represents a knowledge aggregation process using two domains. Knowledge transfer is the process of sharing knowledge using latent features and the ratings across the domains. Fig.3 shows that the common latent factors are used for knowledge transfer.

3.4 Latent Variables

Latent variables are the variables that are not present in the dataset but are inferred from the variables that are present in the original dataset (observed variables). By introducing the set of latent variables, unobserved inferences from the real-world datasets are discovered and uncovering the hidden patterns that lead to knowledge discovery. Mathematical models that aim to derive the latent variables from observed variables are called latent variable models. One advantage of using latent variable is a dimensionality reduction in big data. An enormous amount of collected variables can be merged in a model to reproduce a hidden concept, making it easier for better understanding of data. The dimensionality reduction is the process of decreasing the number of collected variables and can be classified into feature selection and feature extraction.

3.5 Attributes Mapping to Latent variables

The system uses two domain level information and user profiles for attribute mapping. The two domains considered for dimensionality reduction in RS are Restaurants and Tourist Spot [20]. Common features across the domain can be extracted using latent class analysis. Many-to-one mapping of attributes to latent variables is done to reduce the dimensionality of overall attributes in a dataset considering related domains. The user profiles are name, location, gender, history, badges collection, places visited, travel styles, contribution, price range, date of visit, ratings and reviews. Restaurant attributes are name, location, restaurant ratings, reviews by the user, cuisine and review count. Tourist spot attributes are name, ratings, tourist spot reviews, spot location and the type. Fig. 4 represents the dimensionality reduction of "d" attributes to "k" latent variables. The latent factors are classified into corresponding label for learning process. The general equation for latent variable derivation from observed attributes is given as equation (1)

$$Lat = U(\sum(f(user), f(rest), f(tour))) \quad (1)$$

Where f(user), f(rest) and f(tour) are the functions for user profiles, restaurant and tourist spot respectively which are mapping from various observed variables corresponding to the latent variables.

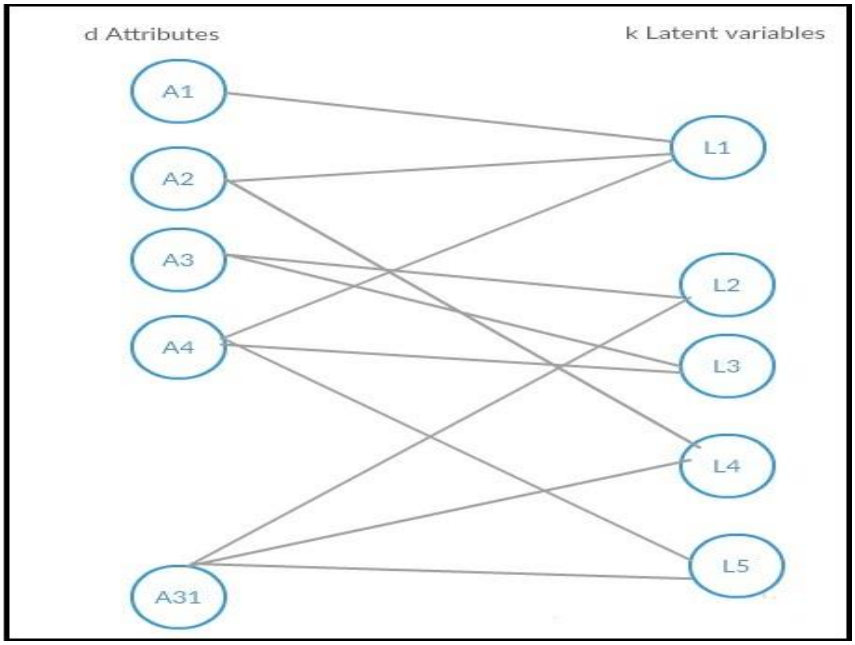

Fig. 4. Dimensionality Reduction

Table 1 represents the comprehensive survey of various techniques used in CDR for addressing data sparsity and cold start problem in CDR, open access of the journal and its scope.

Table 1. Summary of related work

Comparative Summary on Cross Domain Recommendation				
Study	**Major Attention**	**Cold Start Addressed**	**Data Sparsity Addressed**	**Open Access**
[10]	HRW with user item link in target domain	Yes	Yes	No
[11]	ROST based CDR	No	Yes	No
[14]	User based CDR with LDM	No	Yes	Yes
[15]	Funk-SVD based CDR for Multimedia Application	No	Yes	Yes
[16]	Kernel Induced RS for overlapping entities	No	Yes	No
[17]	ACTL for CDR	No	Yes	Yes
[18]	Score Prediction based Course Recommendation using cross user domain CF	No	No	Yes
[19]	RecSys-DAN based CDR	Yes	Yes	No
[5]	Matrix Tri-Factorization based CDR	No	No	No
[9]	Deep Neural Network based CDR	No	Yes	Yes
[2]	Review Aware CDR	Yes	No	Yes
[6]	DNN based CDR	Yes	No	Yes
[4]	Selective Knowledge Transfer for CDR	No	Yes	Yes
[15]	CDR for CPS	No	Yes	Yes
[13]	Co-Citation Selection based CDR	No	No	No

4 Conclusion

The most prevalent problems in RS are sparse data and new user issues. The social networks allow the users to construct relationships among the various types of items across the domains and provide a recommendation list to the users, in order to address the sparse data and new user problem. The CDR is explored to address the above problems with different types of techniques in various perspective. This work also discusses about the knowledge transfer from auxiliary domain to final domain and knowledge linking with multiple domains, latent variable analysis for cross domain technique and attribute mapping into low dimensional data for better recommendation.

References

1. Lina Yao, Quan Z. Sheng, Anne. H.H. Ngu, Jian Yu, and Aviv Segev: Unified Collaborative and Content-Based Web Service Recommendation. In: IEEE Transactions on Services Computing, vol. 8, no. 3, pp. 453-466 (2015).
2. Yaru Jin, Shoubin Dong, Yong Cai, and Jinlong Hu: RACRec: Review Aware Cross-Domain Recommendation for Fully-Cold-Start User. In: IEEE Access, vol. 8, pp. 55032-5504 (2020).

3. Go, Yang J, Park H. and Han S.: Using online media sharing behaviors implicit feedback for collaborative filtering. In: IEEE International Conference on Social Computer, pp. 439–445 (2015).
4. Hongwei Zhang, Xiangwei Kong and Yujia Zhang: Selective Knowledge Transfer for Cross-Domain Collaborative Recommendation. In: IEEE Access, vol. 9, pp. 48039-48051 (2021).
5. Quan Do, Wei Liu, Jin Fan, and Dacheng Tao: Unveiling Hidden Implicit Similarities for Cross-Domain Recommendation. IN: IEEE Transactions on Knowledge and Data Engineering, vol. 33, no. 1, pp. 302-315 (2021).
6. Hanxin Wang, Daichi Amagata, Takuya Makeawa, Takahiro Hara, Niu Hao, Kei Yonekawa, and Mori Kurokawa: A DNN-Based Cross-Domain Recommender System for Alleviating Cold-Start Problem in E-Commerce. In: IEEE Open Journal of the Industrial Electronics Society, vol. 1, pp. 194-206 (2020).
7. Daqiang Zhang, Ching-hsien Hsu, Min Chen, Quan Chen, Naixue Xiong, and Jaime lloret: Cold-Start Recommendation using Bi-Clustering and Fusion for Large-Scale Social Recommender Systems. In: IEEE Transactions on Emerging Topics in Computing, vol. 2, no. 2, pp. 239-250 (2014).
8. Dimitrios Bouzas, Nikolaos Arvanitopoulos, and Anastasios Tefas: Graph Embedded Nonparametric Mutual Information for Supervised Dimensionality Reduction. In: IEEE Transactions on Neural Networks and Learning Systems, vol. 26, no. 5, pp. 951-963 (2015).
9. Wenxing Hong, Nannan Zheng, Ziang Xiong, and Zhiqiang Hu: A Parallel Deep Neural Network Using Reviews and Item Metadata for Cross-Domain Recommendation. In: IEEE Access, vol. 8, pp. 41774-41783 (2020).
10. Meng Jiang, Peng Cui, Xumin Chen, Fei Wang: Social Recommendation with Cross-Domain Transferable Knowledge. In: IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 11, pp. 1041-4347 (2015).
11. Bin Li, Xingquan Zhu, Ruijiang Li, and Chengqi Zhang: Rating Knowledge Sharing in Cross-Domain Collaborative Filtering. In: IEEE Transactions on Cybernetics, vol. 45, no. 5, pp. 1054-1068 (2015).
12. Fei Hao, Shuai Li, Geyong Min, Hee-Cheol Kim and Stephen S. Yau: An Efficient Approach to Generating Location-Sensitive Recommendations in Ad-hoc Social Network Environments. In: IEEE Transactions on Services Computing, vol. 8, no. 3, pp. 520-533 (2015).
13. Supaporn Tantanasiriwong, and Choochart Haruechaiyasak: Cross Domain Citation Recommendation based on Co-Citation Selection. In: IEEE Transactions on Cybernetics, vol. 8, no. 6, pp. 978-987 (2014).
14. Xu Yu, Feng Jiang, Junwei Du, and Dunwei Gong: A User-Based Cross Domain Collaborative Filtering Algorithm Based on a Linear Decomposition Model. In: IEEE Access, vol. 5, pp. 27582-27589 (2017).
15. Xu Yu, Yu Fu, Lingwei Xu, and Guozhu Liu: A Cross-Domain Recommendation Algorithm for D2D Multimedia Application Systems. In: IEEE Access, vol. 6, pp. 62574-62583 (2018).
16. Qian Zhang, Jie Lu, Dianshuang Wu, and Guangquan Zhang: A Cross-Domain Recommender System w15ith Kernel-Induced Knowledge Transfer for Overlapping Entities. In: IEEE Transactions on Neural Networks and Learning Systems, vol. 30, no. 7 (2019).
17. Ming He, Jiuling Zhang, and Shaozong Zhang: ACTL: Adaptive Codebook Transfer Learning for Cross-Domain Recommendation. In: IEEE Access, vol. 7, pp. 19539-19549 (2019).
18. Ling Huang, Chang-Dong Wang, Hong-Yang Chao, Jian-Huang Lai, and Philip S. Yu: A Score Prediction Approach for Optional Course Recommendation via Cross-User-Domain Collaborative Filtering. In: IEEE Access, vol. 7, pp. 19550-19563 (2019).
19. Cheng Wang, Mathias Niepert, and Hui Li: RecSys-DAN: Discriminative Adversarial Networks for Cross-Domain Recommender Systems. In: IEEE Transactions on Neural Networks and Learning Systems, vol. 31, no. 8, pp. 2731-2740 (2020).
20. Valliyammai C., Nanthini M. and S. Ephina Thendral: Dimensionality Reduction Using Latent Variable across the Domains in Recommender System. In: International Research Journal of Electronics and Computer Engineering, vol. 2(2), pp. 33-37 (2016).