Journal of KIIT. Vol. 20, No. 2, pp. 47-52, Feb. 28, 2022. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2022.20.2.47

Adaptive Cross-Domain Recommendation Model based on Association Analysis

Eun-Young Bae*, Seok-Jong Yu**

Abstract

A recommender system is an information filtering system that predicts a user's preferences or ratings for an item. The initial user and data sparsity problems have been the focus of several studies. One of the ways suggested to address these problems is cross-domain recommendation, which utilizes the knowledge learned from other domains for making recommendations for a target domain lacking evaluation information. Most previous studies on cross-domain recommendation do not reflect the characteristics between the target and source domains, the quality of recommendation may vary depending on the source domain used. Moreover, a scalability issue may arise because the data processing effort increases when multiple domains are aggregated and used for making recommendations. In this study, we investigated the impact of using a source domain with a high association degree and a target domain together on the accuracy of recommendation using Amazon dataset. We analyzed and compared the accuracies of proposed recommendation methods with reference to the conventional methods.

요 약

추천 시스템이란 항목에 대한 사용자의 선호도나 평점을 예측하는 정보 필터링 시스템이다. 추천 시스템에서 초기 사용자 문제와 데이터 희박도는 가장 많은 연구가 이루어지고 있는 분야이며, 이 문제에 대한 해결 방안으로 교차 도메인 추천 방법이 있다. 교차 도메인 추천은 평가 정보가 부족한 대상 도메인에서 추천하기 위하여 또 다른 도메인에서 학습한 지식을 활용한다. 기존 교차 도메인 추천 연구들은 대상 도메인과 원본 도메인 간 특성을 반영하지 못하기 때문에 어떤 원본 도메인을 사용하느냐에 따라 추천의 질에 차이가 생길 수 있으며, 여러 도메인을 합병하여 추천에 사용하는 경우 데이터 처리량 증가에 의한 확장성 문제가 발생할 수 있다. 본 연구에서는 아마존 온라인 쇼핑몰 데이터셋을 대상으로 여러 항목 범주에 걸쳐 사용자가 중복된 환경에서 대상 도메인과 연관도가 높은 도메인의 활용이 추천 정확도에 미치는 영향을 분석하였다.

Keywords
recommender system, cross-domain recommendation, association rule, data mining, information filtering

Ⅰ. Introduction

Recommender systems are information filtering systems that predict a user's preferences or ratings for an item. Recommender systems are becoming increasingly advanced to support large-scale services in many sectors such as Amazon's product recommendations, Netflix's movie recommendations,

* Ph.D, Computer Science, Sookmyung Women's Univ.
- ORCID: https://orcid.org/0000-0002-7253-0778
** Professor, Computer Science, Sookmyung Women's Univ.
- ORCID: https://orcid.org/0000-0002-1631-4034
· Received: Nov. 10, 2021, Revised: Dec. 07, 2021, Accepted: Dec. 10, 2021
· Corresponding Author: Seok-Jong Yu
Dept. of Computer Science, Sookmyung Women's University, Korea
Tel.: +82-2-710-9831, Email: sjyu@sookmyung.ac.kr

Facebook's friend recommendations, Google's advertisement recommendations, and YouTube's content recommendations. Two thirds of movie rental earnings by Netflix, 35% of sales by Amazon, and no less than 38% of Google News displayed are based on recommendations [1].

Among the various tasks for the recommender system to tackle, a user cold-start problem is one of the most actively researched areas [2]. To make an appropriate recommendation for the user, the user preference information or profile is required. However, if the user's preferences or meta information is not present in the target domain, it is difficult to make a satisfactory recommendation to a user. Several studies have focused on addressing this initial user problem. One of the solutions to address this problem is cross-domain recommendations [3][4].

Cross-domain recommendations use information from source domain rather than target domain for recommending an item. Based on the overlap conditions between two domains in respect of the user and item, we can have the following scenarios: a user, an item, both user and item, or neither overlap between domains [5][6].

In this study, for an environment with overlapping users such as the Amazon online shopping mall, we intend to investigate which source domain can improve the accuracy of recommendation when the recommendation system does not have the initial user preference information. When multiple domains with preference information are aggregated and used, a lot of data must be processed, which increases the computation load.

In this study, we do not select domains passively. By adaptively selecting and using domains with high association degree with the recommended domain, we expect that the accuracy of recommendation can be improved for the initial user. We intend to confirm this through experiments.

II. Related Works

2.1 Cross-domain recommendation

With the proliferation of e-commerce sites and online social media, user preference data reflecting diverse tastes and interests are used, and profiles across multiple systems are maintained. Instead of processing each domain independently, it is possible to create a more comprehensive user model and better recommended items by utilizing all available user data from multiple domains, such as movies, books, and music. This is called cross-domain recommendation. It can help alleviate the cold-start problem or sparsity problem, and generate customized cross-selling or bundle recommendation for items across multiple domains, and enhance accuracy, serendipity, and diversity [7].

2.2 Association rule mining

Association rule mining aims to create associations between items by using the history of consumer purchasing items to find associated items. Based on purchase history data, we may conclude that “Customers who purchase item A are more likely to purchase item B too.” It is also referred to as market basket analysis or affinity analysis. Apriori and frequent pattern (FP) tree-growth algorithms have been used for this purpose [8].

Apriori algorithm analyses the association among data bits based on the frequency of occurrence. It uses an iterative approach, which is performed in units of levels to be used for a set of k-th item to find a set of the (k+1)th item. The FP-growth algorithm is designed to extract a frequent item set using the FP-tree without generating a candidate frequent item set. It is able to extract a frequent item set faster when compared with Apriori algorithm [9].

Because association rules are expressed as

conditions and reactions, there are advantages in that the results are easy to understand, data can be used by itself without conversion, and calculation is significantly simple. However, when the number of items increases, the calculation effort required for analysis increases exponentially and the association rules are found with significantly segmented items. Therefore, it can result in a meaningless analysis, and less-frequently bought items might be excluded from the rules because of the small number of transactions.

Provided that X and Y are item sets that include one or more items, the association rule is expressed as X->Y. The item set on the left is the antecedent clause (Antecedent, lhs), and the item set on the right is the consequent clause (Consequent, rhs). X->Y is interpreted as "If an item group X is purchased, an item group Y is also likely to be purchased."

The following three major indicators are used in the association rule analysis.

Equation (1) is referred to as Support, which is the measure of the importance of the rule and indicates the value of the entire transaction size. Support is calculated as the probability of a transaction that includes both X and Y in their entirety. Equation (2) is referred to as Confidence, which is the measure of the reliability of the rule and indicates that the purchase of item X translates into how much of item Y is purchased. Equation (3) is referred to as Lift, which is the ratio of the cases where the transaction includes item Y to the cases where item Y is randomly purchased when item X is purchased.

$$Support\,(X \rightarrow Y) = P(X \cap Y) \tag{1}$$

$$Confidence\,(X \rightarrow Y) = P(Y|X) = \frac{P(X \cap Y)}{P(X)} \tag{2}$$

$$Lift\,(X \rightarrow Y) = \frac{P(Y|X)}{P(Y)} = \frac{P(X \cap Y)}{P(X)P(Y)} \tag{3}$$

III. Cross-domain recommendation model using adaptive domain selection based on association degree

We propose an adaptive cross-domain recommendation model (ACDRM) that is designed to improve the initial user problem and increase the accuracy of recommendation by selecting a domain that is highly associated with the target domain adaptively and by making cross-domain recommendations.

3.1 Problem definition

Most of the previous studies on cross-domain recommendations use passively predefined domains for making recommendations. However, this case does not reflect the characteristics between domains well. Therefore, the quality-of-recommendation results may differ depending on the domain that is used. In addition, when multiple domains are used as the source domain, the computation load increases because the data size increases.

In this study, we verify whether we can come to similar conclusions despite using different datasets and models than those used in the previous studies. We produce recommendation results using (a) a domain group with high association degree and a domain with high association degree and (b) an arbitrary domain group and a domain as the source domain. Furthermore, we evaluate the accuracies of these recommendations and compare them with those of the conventional methods.

3.2 Association degree

The association degree was determined using the inter-domain cross-domain association rule rate (ICARR) index. It is calculated as the rate of the cross-domain association rules among the generated association rules. It produces association rules between

items by aggregating two domains. The formula is given by Equation (4).

$$ICARR = \frac{CNT(ICAR)}{CNT(IAR)} \quad (4)$$

where *CNT(IAR)* and *CNT(ICAR)* refer to the number of association and cross-association rules between items, respectively.

3.3 Generation of cross-association rules

First, a domain tag is assigned to all items in a domain. Next items bought together by a buyer in a single transaction are indexed. Finally, Apriori algorithm [9] is used to process the data to generate association rules. The objective of Apriori algorithm is to disclose the association relationships among data based on the frequency of their occurrence. To generate the apriori association rules, transaction data should exist and minimum support must be set. The rules are generated in two steps.

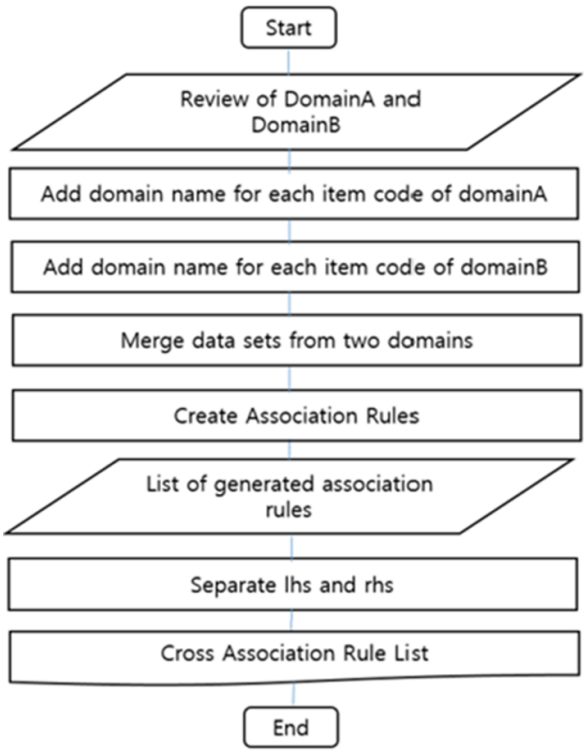

Fig. 1. Cross-association rules extraction algorithm

The first step is to find a set of frequent items from a set of candidate items. To this end, we find a set of items that have a support level above a predetermined minimum value. In the next step, association rules are generated using the sets of frequent items found.

Fig. 1 presents the flowchart of this algorithm. To obtain the cross relationship between domains in association rules, the rules are generated after appending the domain name for each item number of all items in each domain. For the generated set of association rules, the cross-domain rules between antecedent and consequent clauses are extracted. Of all the association rules, the rate of the cross-domain association rules (ICARR) is calculated for determining the association degree.

IV. Experiments and evaluation

As the experimental data, we used Amazon's product dataset preprocessed by McAuley [10]. This dataset does not have overlapping items between domains, although it has overlapping users. This dataset contains reviews of items such as ratings and metadata about items purchased between May 1996 and July 2014. This dataset has 24 domains such as video games, toys, books, electronics, movies and TVs, CDs and etc. In this study, we performed experiments for 10 domains. Table 1 lists the name of algorithms used in the experiments

Table 1. Experimental algorithms

Algorithm	Abbr.
Singular value decomposition	SVD
Cosine similarity item-based collaborative filtering	CI
Pearson similarity item-based collaborative filtering	PI

As a result of calculating the association degree with other domains based on the VG domain(Video

Game), the ICARR value for VG-Toy was 70% and for VG-Video was 35.9%. Fig. 2 shows the RMSE comparisons of recommendations using various domain scenarios: (a) single domain with high association degree (SSH), (b) single domain with low association degree was used (SSL), (c) two domains with high association degree were aggregated and used (MSH), (d) one random domain was used (SSA), (e) two random domains were aggregated (MS), and (f) three random domains and four random domains were aggregated (MS).

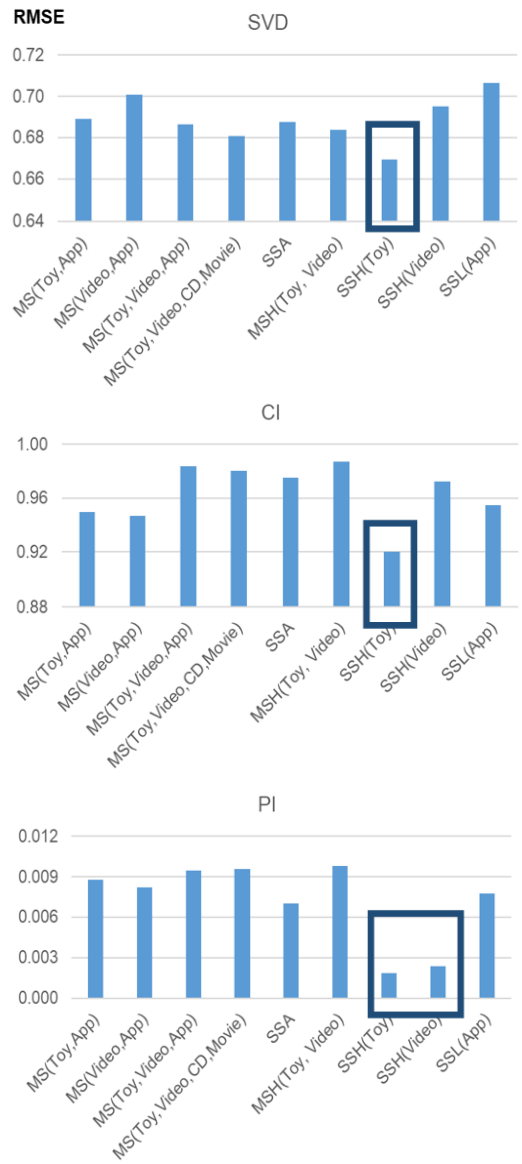

Fig. 2. RMSE comparison using multiple domains (Target: VG)

For “Toy” with the highest association degree, the RMSE(Root mean squared error) value was the lowest for all algorithms. For “Video” with high association degree, the recommendations exhibited high accuracy when using the remaining algorithms other than SVD and CI, which resulted in relatively high error values.

MAE(Mean squared error) measurement results also supported the RMSE comparison results. In a nutshell, the experimental results confirmed that the appropriate selection of single domain with high association degree can improve the accuracy of recommendation when compared with the aggregation of multiple domains.

Ⅴ. Conclusion and future research

We performed experiments on cross-recommendation by adaptively selecting source domains based on a target domain. The experimental results confirmed that the cross-recommendation accuracy using single domain with high association degree was better than the accuracy of recommendation using a domain with low association degree. However, there were some limitations. First, adaptive source domain selection requires datasets from multiple domains with overlapping users. Second, we generated association rules after equally specifying the minimum support and confidence for all domain pairs for obtaining association rules between items and categories. The results could be affected by support and confidence values, and this part can be covered in the future studies.

References

[1] IEEE Explore, https://jessesw.com/Rec-System/ https://ieeexplore.ieee.org/Xplore/home.jsp

[2] X. Su and T. M. Khoshgoftaar, "A survey of collaborative filtering techniques", Advances in Artificial Intelligence, Article No. 4, pp. 2, Jan. 2009. https://doi.org/ 10.1155/2009/421425.

[3] A. Bansal, S. Kumar, R. Yadav, and N. Dhage, "A review on cross domain recommendation", International Conference on Electronics, Communication and Aerospace Technology, Coimbatore, India, pp. 617-620, Apr. 2017. https://doi.org/10.1109/ICECA.2017.8212739.

[4] P. Cremonesi, A. Tripodi, and R. Turrin, "Cross-domain recommender systems", IEEE 11th International Conference on Data Mining Workshops, Vancouver, BC, Canada, pp. 496-503, Dec. 2011. https://doi.org/10.1109/ICDMW.2011.57.

[5] B. Li, "Cross-Domain Collaborative Filtering: A Brief Survey", 23rd IEEE International Conference on Tools with Artificial Intelligence, Boca Raton, FL, USA, pp. 1085-1086, Nov. 2011. https://doi.org/10.1109/ICTAI.2011.184.

[6] O. S. Revankar and Y. Haribhakta, "Survey on Collaborative Filtering Technique in Recommendation System", International Journal of Application or Innovation in Engineering & Management, Vol. 4, No. 3. pp. 85-91, Mar. 2015.

[7] I. Cantador, I. Fernandez-Tobıas, S. Berkovsky, and P. Cremonesi, "Cross-domain recommender systems. In Recommender Systems Handbook", Springer, 2015.

[8] Y. Ye and C. C. Chiang, "A Parallel Apriori Algorithm for Frequent Itemsets Mining", 4th International Conference on Software Engineering Research, Management and Applications, Seattle, WA, USA, pp. 87-93, Aug. 2006. https://doi.org/10.1109/SERA.2006.6.

[9] E. Y. Bae and S. J. Yu, "A Study on the Cross Domain Recommendation System Using Adaptive Source Domain Selection", Journal of KIIT, Vol. 17, No. 10, pp. 9-16, Oct. 2019. http://dx.doi.org/10.14801/jkiit.2019.17.10.9.

[10] Amazon Product Dataset, http://jmcauley.ucsd.edu/data/amazon/links.html, UCSD. [accessed: May 19, 2019]

Authors

Eun-Young Bae

1993 : B.S degree in Statistics, Sookmyung Women's University
2003 : M.E degree in Information Processing, Sogang University
2020 : Ph.D degree in Computer Science, Sookmyung Women's University
Research Interests : Recommender Systems, Collaborative Filtering, Data Visualization

Seok-Jong Yu

1994 : B.S degree in Computer Science, Yonsei University
1996 : M.S degree in Computer Science, Yonsei University
2001 : Ph.D degree in Computer Science, Yonsei University
2005. 3~Present : Professor, Computer Science, Sookmyung Women's University
Research interests : Recommender Systems, Collaborative Filtering, Data Visualization