Integrated contents recommendation platform across different domains

Jihyeong Lee, 2017314320
Chanhoo Keum, 2017314910
Inseo Nam, 2019312643
Seoyoon Hong, 2018310346
Jinhwa Hong, 2017310820

Sungkyunkwan University Capstone Design Project

Abstract. Many contents such as movies, dramas, and webtoons are scattered in various domains. We have the inconvenience of having to move the domain to see each content we want, and there is no recommendation system among them, so we don't know which content suits our interests. In this project, various contents will be included in one service, and a recommendation system will be implemented based on these contents with contents based[1] and collaborative filtering model[2].

Keywords: Cross Domain Recommendation · BERT · Item2Vec · Matrix Factorization · Contents-based Recommendation· Collaborative Filtering

1 Introduction

These days, lots of platforms including OTT services such as Netflix, Watchapedia, Amazon prime uses state of art AI recommendation systems. And there are also such services as Watchapedia and IMDB which provide targeted recommendations based on ratings from the users and similar contents for each content.

However, there are some limitations from user's view. With OTT platform that provides actual contents, user can be provided personalized recommended contents from various domains(e.g. movie, drama, documentary) but the recommendations are limited within the platform, which means user cannot get recommendations of other contents that does not exist in the platform one is using. And in case of rating services, they do not provide cross-domain recommendation and a user can get recommendations only after when one gives the rating for several contents for each domain, e.g. books, movies, drama. They

[1] recommendation model using item features to recommend other items similar to what the user likes, based on their previous actions or explicit feedback.

[2] recommendation model using similarities between users and items simultaneously to provide recommendations.

cannot get personalized recommendation in domains they have left their ratings and not in the other domains.

Therefore, we are planning to develop a service that provides cross-domain recommendations with fewer ratings which is not restricted to any content-providing platforms. We are expecting that a user can get list of similar contents of one content and targeted recommendations not only from one domain but also from other domains, even if he or she have left ratings in only one domain. This service also could be a good tool for content providers to understand the needs of users given the current market trend where developing cross-domain contents become active more and more.

2 Motivation & Objective

2.1 Motivation

There are two main problems of collaborative filtering(CF) recommenders, which are the cold-start and the data sparsity.[1] The cold-start problem is the recommender system cannot generate accurate recommendations for users or items about which it has not yet gathered sufficient information. The start-up of the recommender with almost no user and lack of user interaction may suffer for this problem. The data sparsity refers to the difficulty in finding sufficient reliable similar users since the users only rated a small portion of items, which leads to the sparsity of the ratings in the user-items matrix.

Despite of existing recommendtaion services from big-tech companies, these two problems still remain, and there are a lot of studies ongoing to address this situation. Cross-domain recommendation arise to address the sparsity problem so that recommendations can be made through sparse rating dataset across several domains.

We expect to achieve moderate performance using basic techniques used to implement recommendation system, and therefore provide practical cross-domain recommendation service with lighter model.

2.2 Objective

We aims to provide recommendation system service for three domains-movie, drama and webtoon-that utilizes two different approaches: content-based modeling and collaborative filtering(CF). The purpose of this service is to provide personalized recommendations to our users based on their interests and preferences.

With the content-based modeling approach, we will analyze the features and characteristics of the items that the user has interacted with in the past and recommend similar items that match their interests. This approach is based on the assumption that users who have liked similar items in the past will like similar items in the future.

On the other hand, the collaborative filtering approach will recommend items to the user based on the preferences of similar users. This approach is based on the idea that users who have similar tastes and preferences will likely enjoy the same items.

Our service aims to provide independent recommendation and evaluation services separate from OTT platforms, as well as offer appropriate recommendations to users who may not leave a lot of ratings. We strive to deliver personalized recommendations to our users based on their interests and preferences, as well as offer a more satisfactory experience on our platform. Even users who may not have left many ratings will still receive relevant and personalized recommendations through our system.

3 Background & Related Work

3.1 Background

Over the past few years, dramas such as Squid Game and Strange Lawyer Woo Young-woo have been popular not only in Korea but also abroad. And these contents were mainly consumed by people based on OTT. According to a newspaper, about 70% of Koreans currently use OTT. As many users exist, it will be necessary to provide various services. The contents provided by OTT are also different, mainly movies and dramas. And there is another representative content consumed by young people. It's a webtoon. However, due to the nature of OTT, it is difficult to support webtoon content because it provides video. Therefore, based on this point, we are going to provide a so-called Cross Domain Recommendation (CDR) service that recommends content suitable for users by bringing together many contents, including various movies, dramas, and webtoons scattered in various domains.

3.2 Related Work

koBERT is a Korean version of BERT, a natural language processing model in Korean. BERT is a natural language processing model developed by Google that has the ability to understand the contextual meanings of words in sentences by learning large amounts of text data. koBERT was created by fine-tuning these BERT models to match Korean data. Representatively, there is a koBERT model provided by SKT Brain. This significantly improves natural language processing performance in Korean text data, and shows good performance on various natural language processing tasks. koBERT can be used in various natural language processing tasks such as classifying Korean sentences, analyzing emotions, and answering questions. Also, using this, applications such as Korean sentence generation can be developed.

Ensemble method is a technique that combines multiple models to create a single predictive model. An ensemble is a collection of systems equivalent to a system in statistical mechanics. This is one of the ways to achieve higher predictive performance by leveraging the collective intelligence that can be produced with multiple single models, rather than expecting the results of performance in a single model. The Ensemble method is divided into two main types. The first type is a method of learning various models and deriving final predictions by weighted average of their predictions. This method is called the Voting Ensemble. The second type is how to generate new learning data using the predictive results of each model, and then use this data to re-learn the model. This method is called a bagging ensemble. We aim to achieve high performance in collaborative filtering with ensemble method. Below are the candidate models that we are going to take experiment with.

itemKNN or Item-based Collaborative Filtering is one of the most popular algorithms in recommended systems. Based on historical user-item evaluation data, the algorithm aims to predict which items users may prefer. ItemKNN is a type of Item-to-Item Collaborative Filtering that measures similarity between items and predicts user-item evaluation data based on it. Specifically, after generating a feature vector for each item, we calculate the cosine similarity between the items to find similar items. And using this, predictions are made using the evaluation scores of similar items among items that the user has not yet evaluated.

userKNN or User-based Collaborative Filtering is one of the most popular algorithms in recommended systems. Based on historical user-item evaluation data, the algorithm aims to find users with similar tastes and use their item evaluation data to predict which items users would prefer. UserKNN is a type of User-to-User Collaborative Filtering that measures similarity between users and predicts user-item evaluation data based on it. Specifically, after creating a feature vector for each user, we calculate the cosine similarity between users to find users with similar tastes. And using this, predictions are made using item scores evaluated by other users with similar tastes to the user you want to predict.

BPR (Bayesian Personalized Ranking)[3] is one of the most popular collaborative filtering algorithms in recommended systems. The algorithm uses the Bayesian probability model-based Personalized Ranking method to predict your preferred items. BPR is one of the Matrix Factorization algorithms, which represents user-item evaluation data as matrices and decomposes them into latent factors. Using this latent factor, BPR uses a Personalized Ranking method to predict which items users prefer. The Personalized Ranking method aims to rank your favorite items in order. BPR uses a very important concept, Pairwise Ranking, when learning this Personalized Ranking. This is a way to learn which one the model prefers for two items. For example, if the user rated Movie A as preferred to Movie B, the model learns these Pairwise Ranking so that A ranks higher than B. Since BPR is based on the Bayesian probability model, the algorithm considers a prior distribution of user-item evaluation data. This allows the model to make more accurate recommendations using prior information as it learns.

LightGCN is one of the algorithms used in Collaborative Filtering (CF)-based recommendation systems[4]. The basic idea of GCN is to learning representation for nodes by smoothing features over the graph. It is based on Graph Convolutional Network (GCN) and learns by graphing the interactions between users and items. LightGCN generates graphs that represent users and items as nodes and the interaction between users and items as edges. This graph shows the relationship between a user and an item, and uses it to learn the potential factors of each user and item. LightGCN uses GCN to learn latent factors by considering the interaction between users and items through this graph.

FM (Facatorization Machine) is a model based on matrix factorization and is one of the models used in various fields, including recommended systems[5]. FM deals with very intense feature vector X. To overcome the limitations of linear models that fail to reflect interactions between properties, the proposed model can perform better than linear models considering interactions between properties. To this end, FM learns the vector representations of each characteristic and the weights for the interactions between them. Various features can be concatenated and used as a feature vector, and any implicit attribute can be added to a feature vector in the form of a real number.

4 Problem Statement & Proposed Solution

We need to discover content-based similarities for three different domains. This requires analyzing descriptions (such as genres, titles, and plot summaries) of individual works. Additionally, we aim to recommend works from other domains to users who have left ratings for one domain. To achieve this, we require an approach that can identify the similarity between users and the items they have rated across different domains.

5 Planning in detail

5.1 Data

We plan to extract movie, drama, and webtoon data from each of the three domains based on web crawling. Currently, about 2,500 data are extracted from Watchapedia. More movie data is planned to be extracted from IMDb. Dramas and webtoons also plan to extract data through evaluation sites covering both domestic and overseas or like Naver, Google websites and use them for learning.

5.2 Model

We plan to convert the title and genre of each content into an embedded vector using koBERT. After the conversion, top-k items with high similarity are recommended by calculating the similarity to the content that the user has watched or likes.

Collaborative filtering is looking forward to finally learning user embedding and encoder. If learning is well conducted, it seems that it is possible to recommend an unsupervised learning-based system by forming a cluster between similar users or a cluster between similar contents.The parameter of the encoder has a method of using it independently of the method of sharing. We plan to verify the accuracy through experiments. In the case of CNN and SasRec, we expect that our model will also be able to perform more through this method because the performance has been improved by sharing parameters.

We are planning to try ensemble method with models such as itemKnn, userKnn, and BPR first, and also graph-based LightGCN and etc. Traditionally, the voting method is widely used, but it is planned by learning the weight and doing a weighted sum.

In the case of FM(Factorization Machine), it is an advantageous model to consider various features. Various features are attached next to each other in parallel and used for learning. Due to these characteristics, there is an advantage of being able to utilize various domain information at the same time, and I think this is a model that fits well with our project.

5.3 Service

We aim to provide detailed information on each work and a list of similar works, as well as personalized recommendation lists for three different domains once the user has left ratings for a few works. Through this, users will be able to receive recommendations for works that match their tastes across various domains, in addition to content-based recommendations.

5.4 Development plan

We are planning to do crawling, preprocessing the data and modeling for about 5 weeks. For the next two weeks, we will build a server to store this data to recommend contents to user effectively, and we will create a website to provide the service for the rest of the year. So, we will proceed with the project for about 10 weeks and finally provide the recommended system service.

Table 1. Development Time Table

	week5	week6	week7	week8	week9	week10	week11	week12	week13
Data Crawling	■	■							
Data Preprocessing		■	■						
Modeling			■	■	■				
Back-end					■	■	■		
Front-end							■	■	■

References

1. Berkovsky, S., Kuflik, T., Ricci, F.: Cross-domain mediation in collaborative filtering. User Modeling 2007, 355–359. https://doi.org/10.1007/978-3-540-73078-1_44
2. Zhu, F., Wang, Y., Chen, C., Zhou, J., Li, L., Liu, G.: Cross-domain recommendation: Challenges, progress, and prospects. Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence. (2021)
3. Rendle, S., Freudenthaler, C., Gantner, Z., Schmidt-Thieme, L.: BPR: Bayesian personalized ranking from implicit feedback. arXiv.org. https://arxiv.org/abs/1205.2618 (2012)
4. He, X., Deng, K., Wang, X., Li, Y., Zhang, Y., Wang, M.: LIGHTGCN: Simplifying and powering graph convolution network for recommendation. arXiv.org. https://arxiv.org/abs/2002.02126 (2020)
5. Rendle, S.: Factorization machines. 2010 IEEE International Conference on Data Mining. https://doi.org/10.1109/icdm.2010.127 (2010)