The Journal of The Institute of Internet, Broadcasting and Communication (IIBC)
Vol. 23, No. 5, pp.1-8, Oct. 31, 2023. pISSN 2289-0238, eISSN 2289-0246

https://doi.org/10.7236/JIIBC.2023.23.5.1
JIIBC 2023-5-1

# K-means 클러스터링과 트랜스포머 기반의 교차 도메인 추천

# Cross-Domain Recommendation based on K-Means Clustering and Transformer

**김태훈*, 김영곤**, 박정민*****

Tae-Hoon Kim*, Young-Gon Kim**, Jeong-Min Park***

**요 약** 교차 도메인 추천은 다른 도메인에 있는 관련 사용자 정보 데이터와 아이템 데이터를 공유하는 방법입니다. 주로 사용자 중복이 많은 온라인 쇼핑몰이나 유튜브, 넷플릭스와 같은 멀티미디어 서비스 컨텐츠에서 사용됩니다. K-means 클러스터링을 통해 사용자 데이터와 평점을 기반으로 군집화를 실시하여 임베딩을 생성합니다. 이 결과를 트랜스포머 네트워크를 통해 학습한 후 사용자 만족도를 예측합니다. 그런 다음 트랜스포머 기반 추천 모델을 사용하여 사용자에게 적합한 아이템을 추천합니다. 이 연구를 통해 추천함으로써 더 적은 시간적 비용으로 초기 사용자 문제를 예측하고 사용자들의 만족도를 높일 수 있다는 결과를 실험을 통해 보여주었습니다.

**Abstract** Cross-domain recommendation is a method that shares related user information data and item data in different domains. It is mainly used in online shopping malls with many users or multimedia service contents, such as YouTube or Netflix. Through K-means clustering, embeddings are created by performing clustering based on user data and ratings. After learning the result through a transformer network, user satisfaction is predicted. Then, items suitable for the user are recommended using a transformer-based recommendation model. Through this study, it was shown through experiments that recommendations can predict cold-start problems at a lesser time cost and increase user satisfaction.

**Key Words :** cross-domain recommendation, data sharing, K-means clustering, transformer network

## Ⅰ. 서 론

### 1. 개요

최근 온라인 쇼핑몰이나 유튜브(YouTube) 또는 넷플릭스(NetFlix)와 같이 멀티미디어 서비스 컨텐츠에서 추천 시스템이 많이 적용됨에 따라 머신러닝을 이용한 추천 시스템 분야의 연구 또한 활발하게 진행되고 있다[1]. 추천 시스템의 성능을 높이기 위해서는 크게 3가지 방향성이 존재한다.

### 2. 초기 사용자 문제

첫 번째는 초기 데이터의 양이 얼마나 확보가 가능한

*정회원, 한국공학대학교 컴퓨터공학과
**정회원, 한국공학대학교 컴퓨터공학과
***정회원, 한국공학대학교 컴퓨터공학과
접수일자 2023년 9월 16일, 수정완료 2023년 9월 30일
게재확정일자 2023년 10월 6일

Received: 16 September, 2023 / Revised: 30 September, 2023 / Accepted: 6 October, 2023
*Corresponding Author: jmpark@tukorea.ac.kr
Dept of Computer Engineering Tech University of Korea, Korea

지이다. 추천할 사용자에 대한 기록이 전혀 없다면, 그 사용자에 대한 정보가 없으므로 모든 사용자에 대한 이용기록이 가장 많은 컨텐츠를 추천함으로써 범용성을 중시할 수밖에 없기 때문이다[2]. 이러한 문제를 초기 사용자 문제(Cold-start) 라고 하며, 본 논문에서는 이러한 문제를 해결하기 위해 교차 도메인 추천(Cross-domain recommendation)방식을 사용하였다.

### 3. 데이터 군집화

두 번째는 학습 데이터의 질을 높이는 방법이다. 많은 데이터의 양이 확보되어도 신경망이 제대로 학습을 할 수 없다면 추천 성능이 떨어지게 된다[3]. 따라서 본 논문에서는 K-means 클러스터링을 이용하여 사용자 기록 데이터를 군집화 함으로써 학습이 잘 되도록 유도하였다.

### 4. 다층 신경망

세 번째는 다층 신경망을 통해 학습시키기 위한 각종 파라메터들을 잘 조절하는 것이다[4]. 학습 속도를 개선하고 과적합을 방지하기 위한 배치 정규화를 통해 활성화 함수 또는 출력값을 정규화하는 작업을 한다[5]. 또한 신경망 학습 최적화를 통해 학습률 즉, 한 번 갱신할 때의 가중치의 값을 양을 조절하는 방식으로 성능을 높일 수가 있다[6].

### 5. 논문의 구성

본 논문의 구성은 2장에서 교차 도메인 추천, 그리고 트랜스포머 추천에 대해 설명한다. 3장에서는 전체적인 시스템 구조와 K-means 클러스터링과 트랜스포머와 같이 사용된 머신러닝 알고리즘의 세부 시스템 구조에 대해 설명한다. 4장에서는 제안한 추천 기법에 대해 실험하고, 추천 성능에 대한 분석과 평가를 진행한다. 5장에서는 결론과 향후 연구방향과 한계성에 대해 기술한다.

## II. 관련 연구

### 1. 교차 도메인 추천

추천 시스템에서의 초기 사용자 문제는 추천 성능에 많은 영향을 끼치기 때문에 현재까지도 많은 연구를 거치고 있는 문제점 중 하나이다. 따라서 초기 데이터 희소성 문제를 해결하기 위한 방안으로 나온 것이 교차 도메인 추천이다[7]. 교차 도메인 추천은 각각의 도메인을 독립적으로 처리하는 대신에 원본 도메인(Source domain)의 데이터를 다른 대상 도메인(Target domain)을 위해 사용하는 것을 의미한다. 교차 도메인 추천을 통해 정확도(Accuracy)과 우연성(Serendipity), 다양성(Diversity)을 향상시킴으로써 초기 사용자 문제를 개선하는 것이 목적이다[8]. 교차 도메인 추천은 기존 도메인 데이터와 다른 도메인 데이터를 서로 비교하여 겹치는 데이터가 있다면, 그 데이터를 기준으로 비어있는 잠재 공간을 채우게 된다. 이 비교하는 대상이 아이템일 경우 아이템 오버랩이라고 하고 사용자 정보일 경우 사용자 오버랩 이라고 표현하며, 둘 다 비교할 경우 풀 오버랩, 겹치는 데이터가 없을 경우 노 오버랩 이다[9].

### 2. 트랜스포머 기반 추천

본 연구에서는 K-means 클러스터링을 사용한 트랜스포머 기반의 교차 도메인 추천 방법을 제안한다. 트랜스포머 기반 추천 방식은 사용자와 아이템 간의 복잡한 관계를 학습하고, 깊은 신경망을 통해 추천 성능을 높일 수 있다. 트랜스포머 모델은 교차 도메인 추천에서 중요한 부분으로, 사용자와 아이템 간의 상호 작용을 기반으로 잠재 요인을 추출한다. 이 잠재 요인은 사용자와 아이템의 성향을 반영하여 추천 성능을 높일 수 있다. 또한, 트랜스포머 모델은 다중 헤드 어텐션을 사용하여 다양한 관점에서 정보를 처리할 수 있다. 본 연구에서는 K-means 클러스터링을 사용하여 사용자 데이터를 군집화하고, 이를 트랜스포머 모델에 입력하여 임베딩을 생성한다. 생성된 임베딩은 다층 신경망을 통해 학습되며, 이를 통해 사용자 만족도를 예측한다. 이를 통해 교차 도메인 추천을 수행하며, 추천 성능을 높일 수 있다.

### 3. K-means Clustering

K-means 클러스터링은 비지도 학습 알고리즘으로, 데이터 집합을 K개의 클러스터로 나누는 것을 목표로 합니다. 이 알고리즘은 다음과 같은 과정으로 동작한다. (1) 초기 중심점을 무작위로 K개 선택한다. (2) 각 데이터 포인트를 가장 가까운 중심점의 클러스터에 할당한다. (3) 각 클러스터의 중심을 다시 계산한다. 이를 위해 클러스터 내의 모든 데이터 포 인트의 평균을 구한다. (4) 중심점이 변하지 않을 때까지 2번과 3번의 과정을 반복한다[10]. K-means 클러스터링의 목적 함수는 다음과 같다.

$$J=\sum_{i=1}^{K}\sum_{x\in Si}\| x-\mu_i \|^2 \quad (1)$$

여기서 Si 는 중심이 Ci 인 데이터 집합 i에 속하는 모든 점이고, $\mu$i 는 i번째 클러스터의 중심점을 의미한다. 이 함수를 최소화하면 각 데이터 포인트와 해당 클러스터 중심점 사이의 거리의 제곱 합이 최소가 된다. 이를 통해 데이터 포인트들이 서로 가까운 그룹으로 묶이게 된다.

### 4. DDTCDR: Deep Dual Transfer Cross Domain Recommendation

DDTCDR 알고리즘은 교차 도메인 추천을 위한 임베딩 공유 알고리즘이다. 이 알고리즘의 핵심은 오토인코더를 사용하여 입력 데이터를 차원 축소하여 도메인을 생성하는 것입니다. 그런 다음, 소스 도메인과 타겟 도메인의 임베딩 데이터를 전이 학습을 위한 임베딩 벡터(X)를 통해 서로 데이터를 공유하면서 교차 도메인 방식으로 인공 신경망 학습을 진행한다. DDTCDR의 목적 함수는 다음과 같다.

$$L=\sum_{i=1}^{N}(y_i-\hat{y_i})^2 \quad (2)$$

여기서 $y_i$는 실제 평점이고, $\hat{y}_i$ 는 예측 평점을 의미하고, N은 학습 데이터의 개수이다. 이 목적 함수를 최소화하면 실제 평점과 예측 평점 사이의 차이가 줄어들게 된다.

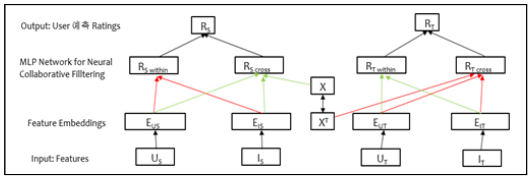


**그림 1. DDTCDR: Deep Dual Transfer Cross Domain Recommendationd의 시스템 구조**
**Fig. 1. System structure of DDTCDR: Deep Dual Transfer Cross Domain Recommendationd system**

이 알고리즘은 전이 학습을 사용하여 한 도메인에서 학습된 정보를 다른 도메인에 전이하면서 추천 성능을 향상시킵니다. 이를 통해 다양한 도메인에서의 데이터를 활용하여 추천 시스템의 성능을 향상시킬 수 있습니다.

### 5. 오토인코더 (AutoEncoder)

오토인코더(AutoEncoder)는 비지도 학습 방식으로 작동하는 신경망으로서, 주로 입력 데이터의 압축 및 복원에 사용된다. 이 알고리즘은 입력을 가능한 한 충실히 출력으로 복사하려고 하지만, 일반적으로 은닉 계층의 뉴런 수가 입력 계층보다 적어서 데이터를 압축하게 된다. 이를 통해 데이터의 중요한 특성을 추출하고, 불필요한 정보를 제거하여 차원을 축소한다. 오토인코더는 일반적으로 두 부분으로 구성된다. 첫 번째 부분은 인코더라고 하며, 입력 데이터를 저차원의 내부 표현으로 변환한다. 두 번째 부분은 디코더라고 하며, 내부 표현을 다시 원본 차원의 출력으로 변환한다. 오토인코더의 목표는 디코더의 출력이 원본 입력과 최대한 유사하도록 하는 것이다. 오토인코더의 손실 함수는 입력과 출력의 차이를 통해 계산된다. 이 손실 함수는 입력 데이터와 출력 데이터 사이의 차이를 최소화하려고 하며, 이를 통해 중요한 특성을 학습하도록 유도한다. 오토인코더는 은닉 계층의 뉴런 수가 입력 계층보다 적을 경우, 불완전한 오토인코더라고 한다. 불완전한 오토인코더는 저차원을 가지는 은닉 계층 때문에 입력을 그대로 출력으로 복사할 수 없다. 따라서 입력과 출력이 같도록 학습을 진행하는 과정에서, 중요한 특성만을 학습하게 되어 입력 데이터의 압축 표현을 얻을 수 있다.

## Ⅲ. 전체 시스템 구조

### 1. K-TCDR 추천 시스템

본 논문에서 제안하는 알고리즘은 K-TCDR (K-means Clustering and Transformer-based Cross-Domain Recommendation) 추천 시스템으로, 교차 도메인 추천에서 초기 사용자 문제를 해결하고 개인 맞춤형 추천의 정확도를 높이기 위한 목적을 가지고 있다. 기존 연구와의 차별성은 임베딩 생성 과정에서 오토인코더를 사용하는 대신, K-means 클러스터링과 트랜스포머 모델을 사용한다는 점이다. 오토인코더는 입력과 출력이 동일하므로, 새로운 데이터에 대해서도 학습 데이터와 동일하게 되어 과적합 문제가 발생할 수 있다. 이를 방지하기 위해 K-means 클러스터링을 사용하여 새로운 데이터를 적용하고, 트랜스포머 모델을 통해 교차 도메인 추천을 수행한다. 그림 2는 기존의 교차 도메인 추천 시스템 구조와 본 논문에서 제안하는 교차 도메인 추천 시스템의 차이점을 보여주는 그림이다.

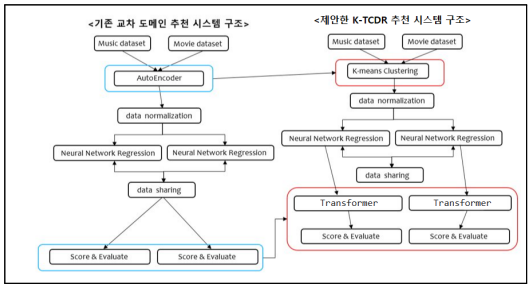


**그림 2. 기존 교차 시스템과 제안한 추천 시스템의 시스템 구조**
**Fig. 2. System structure of the existing cross system and the proposed recommendation system**

이 방법은 K-means 클러스터링을 사용하여 데이터를 여러 클러스터로 나누고, 트랜스포머 모델을 사용하여 각 클러스터 내에서 교차 도메인 추천을 수행한다. 트랜스포머 모델은 attention 메커니즘을 활용하여 다양한 도메인 간의 상호작용을 파악하고, 이를 기반으로 교차 도메인 추천을 수행한다. 이를 통해, 한 도메인에서 학습된 정보를 다른 도메인에 전이하여 추천 성능을 향상시킬 수 있다. 이러한 방식은 다양한 도메인에서의 데이터를 활용하여 추천 시스템의 성능을 향상시킬 수 있다.

## 2. K-TCDR 세부 시스템 구조

본 논문에서 제안하는 K-TCDR 시스템의 세부 구조를 설명하겠다. 먼저 타겟 도메인의 데이터와 소스 도메인의 데이터를 K-means 클러스터링을 통해 사용자 데이터를 군집화한다. 초기 중심점은 무작위로 설정되며, 사용자가 설정한 K값에 따라 군집의 수와 데이터의 응집도가 달라진다. 키워드 분석과 비교를 위해 데이터를 벡터화하는데, 이때 Bag of Words(BoW) 방식을 사용한다. BoW는 키워드의 순서는 무시하고 출현 횟수만을 고려하여 텍스트 데이터를 벡터화하는 방법이다. 숫자가 클수록 해당 키워드가 더 많이 등장한 것을 의미한다. 이렇게 벡터화된 데이터를 사용하여 식(1)을 통해 중심점과의 거리가 최소화되도록 한다. 중심점과의 거리는 0 ~ 1 사이의 값으로 나타내며, 값이 1에 가까울수록 중심점과의 거리가 멀다는 것을 의미한다. 데이터 군집화를 통해 임베딩 생성이 완료되면, 트랜스포머 기반의 인공 신경망을 사용하여 학습을 진행한다. 벡터화된 데이터와 K-means 클러스터링으로 구한 벡터값을 모두 활용하여 학습을 시킨다. 트랜스포머 모델은 attention 메커니즘을 활용하여 다양한 도메인 간의 상호작용을 파악하고, 이를 기반으로 교차 도메인 추천을 수행한다. 본 논문에서는 소스 도메인$D_S$와 타겟 도메인$D_T$의 일반적인 경향을 반영하면서 도메인 간의 차이를 고려하기 위해 사용자 잠재요인인 X를 통해 식(5)의 $D_S$와 식(6)의 $D_T$의 데이터를 공유한다. 따라서 트랜스포머 모델은 일반적인 경향의 학습을 기반으로 하고, 공유하는 임베딩인 X를 학습시킴으로써 오차를 최소화한다. 식(3)은 $D_S$의 학습을 통해 평점을 예측하는 수식이며, 식(4)는 $D_T$의 학습을 통해 평점을 예측하는 수식이다.

$$\widehat{R_S} = (1-\alpha)RS_S(EK_{US}, EK_{IS}) + \alpha RS_T(X \bullet EK_{US}, EK_{IS}) \tag{3}$$

$$\widehat{R_T} = (1-\alpha)RS_T(EK_{UT}, EK_{IT}) + \alpha RS_S(X^T \bullet EK_{UT}, EK_{IT}) \tag{4}$$

$$\epsilon(D_S) = \sum_{S(u,i)\in D_S} R_{S(u,i)} - \widehat{R_{S(u,i)}} \tag{5}$$

$$\epsilon(D_T) = \sum_{T(u,i)\in D_T} R_{T(u,i)} - \hat{R}_{T(u,i)} \tag{6}$$

이후, 신경망 학습이 완료되면 추천을 수행하여 예측된 평점을 통해 실제 평점별 사용자와 아이템 간의 유사도를 비교하여 상위 N개의 추천을 진행한다. 추천의 정확도를 평가하기 위해 MAE(Mean Absolute Error)와 nDCG(normalized Discounted Cumulative Gain)를 사용한다. MAE의 수식은 식(7)과 같으며, nDCG의 수식은 DCG/iDCG로 계산되며, DCG는 식(8), iDCG는 식(9)과 같다.

$$\frac{1}{n}\sum_{i=1}^{n}|R_i - \widehat{R_i}| \tag{7}$$

$$\sum_{i=1}^{n}\frac{rel_i}{\log_2(i+1)} \tag{8}$$

$$\sum_{i=1}^{n}\frac{rel_i^{opt}}{\log_2(i+1)} \tag{9}$$

# Ⅳ. 실험 및 결과

## 1. 실험 환경 및 데이터

데이터 세트는 last.FM과 musicbrainz의 두 개의 도메인 데이터를 사용하여 총 60K의 음악 데이터를 포함하며, 사용자의 평점과 아이템에 대한 메타 데이터가 포함되어 있다. 또한, 영화 데이터로는 IMDb에서 얻은 총 77K의 영화 데이터를 사용하였다. 음악 데이터 세트의 자세한 정보는 표 1에, 영화 데이터 세트의 정보는 표 2에 나와있다.

**표 1. 실험 데이터 셋 음악 정보**
**Table 1. Experimental dataset music information**

| Last.FM & Music brainz 60K | | |
|---|---|---|
| Rating data | Type. of users | 10000 |
| | Type. of items | 19279 |
| | No. of ratings | 60132 |
| Item meta data | Type. of album | 41586 |
| | Type. of genre | 39 |
| | Type. of publish | 20 |
| | Type. of track | 14 |
| | Type. of sales | 53281 |
| | Type. of artist | 26032 |
| | Type. of artist job | 23 |
| | Type. of art birth | 46 |
| | Type. of country | 246 |
| | Type. of city | 4757 |
| | Type. of email | 26024 |
| Training-set: 75%, Test-set: 25% | | |

**표 2. 실험 데이터 셋 영화 정보**
**Table 2. Experimental dataset movie information**

| IMDb 77K | | |
|---|---|---|
| Rating data | Type. of users | 10000 |
| | Type. of items | 74227 |
| | No. of ratings | 77624 |
| Item meta data | Type. of year | 112 |
| | Type. of genre | 1227 |
| | Type. of duration | 260 |
| | Type. of country | 4772 |
| | Type. of language | 4244 |
| | Type. of director | 31657 |
| | Type. of writer | 62241 |
| | Type. of production | 30755 |
| | Type. of actor | 77593 |
| | No. of description | 77505 |
| Training-set: 75%, Test-set: 25% | | |

## 2. 실험 매개변수 조절

K-means 클러스터링으로 임베딩을 생성할 때 클러스터의 중심점의 개수에 따라서 분류 정확성과 추천 성능에 큰 영향을 미친다. 그러나 중심점의 개수가 많아지면 성능이 더 좋아지는 것은 아니며, 반대로 너무 적으면 특징에 따른 평점 예측이 어려울 수 있다. 이를 해결하기 위해 일반적으로 널리 사용되는 값인 K = 3 ~ 10까지의 범위에서 실험을 진행하였다. 이 실험은 음악 데이터를 사용하여 영화를 추천할 경우와 영화 데이터를 사용하여 음악을 추천할 경우 모두에 적용되었다. 다음으로, 트랜스포머 기반의 추천을 이용하여 상위 N개의 추천을 할 때, 최적의 K 값을 찾기 위해 총 7개 구간의 단계를 거쳐 가장 높은 성능을 보이는 K 값을 선택하여 최소 MAE 값을 얻을 수 있도록 하였다. 추천의 정확도를 평가하는 척도로는 nDCG를 사용하였으며, 각 구간별로 가장 1에 가까운 값을 판별하였다. 제안한 알고리즘의 실험 결과와 기존 알고리즘의 실험 결과를 비교 분석하여 더 나은 성능 결과를 나타내었다는 것을 입증하였다. 표 3은 각 모듈별로 실험에 사용한 매개변수들, 그 범위, 및 설정을 정리한 표이다.

**표 3. 실험 데이터 셋 영화 정보**
**Table 3. Experimental dataset movie information**

| | | |
|---|---|---|
| K-means Clustering | Centroid | 3 ~ 10 |
| | Initialization | K-Means |
| | Metric | Euclidean |
| | Iterations | 100 |
| Neural Network Regression | Hidden node | 100 |
| | learning rate | 0.1 |
| | No. of layer | 2 |
| | Initial learning weight | 0.01 |

이 실험에서 K-means 클러스터링과 트랜스포머 기반의 교차 도메인 추천을 통합하여 사용하여, 더욱 정확하고 개인화된 추천을 사용자에게 제공하였다. 트랜스포머 모델은 attention 메커니즘을 활용하여 다양한 도메인 간의 상호작용을 파악하고, 이를 기반으로 교차 도메인 추천을 수행한다. 이를 통해, K-means 클러스터링에서 생성된 임베딩과 트랜스포머 기반의 교차 도메인 추천을 통합하여 높은 성능의 추천을 제공하였다.

## 3. 실험 결과

먼저 중심점의 개수에 따른 오차를 파악하기 위해 K 값이 3 ~ 10까지의 MAE로 실험 결과를 나타낸다. $D_S$(music)는 소스 도메인으로 음악 데이터를 사용한 경우이고, $D_T$(movie)는 타겟 도메인으로 영화 데이터를 사용한 경우다. 그리고 $D_S$(movie)는 소스 도메인이 영화 데이터인 경우이며, $D_T$(music)는 타겟 도메인이 음악 데이터인 경우이다. K가 3일 때 $D_S$(music)에서 $D_T$(movie)로 추천하는 경우와 $D_S$(movie)에서 $D_T$(music)로 추천할 때 각각 1.303331와 0.468234로 K가 3일 때의 성능이 가장 우수하다고 결론지었다.

**표 4. Mean absolute error(MAE) 실험 결과**
**Table 4. The mean absolute error(MAE) Experiment result**

| MAE | $D_s$(music) → $D_T$(movie) | $D_s$(movie) → $D_T$(music) |
|---|---|---|
| $K$ = 3 | 1.303331 | 0.468234 |
| $K$ = 4 | 1.313256 | 0.728120 |
| $K$ = 5 | 1.325624 | 0.651702 |
| $K$ = 6 | 1.342270 | 0.490275 |
| $K$ = 7 | 1.367712 | 0.617224 |
| $K$ = 8 | 1.375387 | 0.717404 |
| $K$ = 9 | 1.372834 | 0.794262 |
| $K$ = 10 | 1.387472 | 0.844376 |

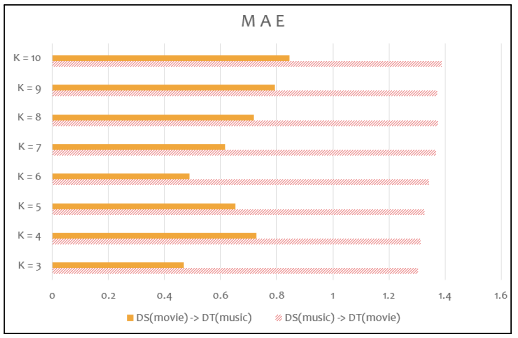


**그림 3 Mean absolute error(MAE) 결과 그래프**
**Fig. 3. The mean absolute error(MAE) result graph**

K가 6일 때 $D_S$(music)에서 $D_T$(movie)로 추천하는 경우와 $D_S$(movie)에서 $D_T$(music)로 추천할 때 각각 1.303331와 0.468234로 K가 6일 때의 성능이 가장 우수하다고 결론지었다.

**표 5. Normalized discounted cumulative gain(nDCG) 실험 결과**
**Table 5. The normalized discounted cumulative gain (nDCG) experiment result**

| nDCG | $D_S$(music) → $D_T$(movie) | $D_S$(movie) → $D_T$(music) |
|---|---|---|
| $K$ = 3 | 0.942211 | 0.968446 |
| $K$ = 4 | 0.944432 | 0.974510 |
| $K$ = 5 | 0.943183 | 0.970242 |
| $K$ = 6 | 0.948764 | 0.980360 |
| $K$ = 7 | 0.945548 | 0.963844 |
| $K$ = 8 | 0.934315 | 0.968552 |
| $K$ = 9 | 0.932536 | 0.966722 |
| $K$ = 10 | 0.933312 | 0.964772 |

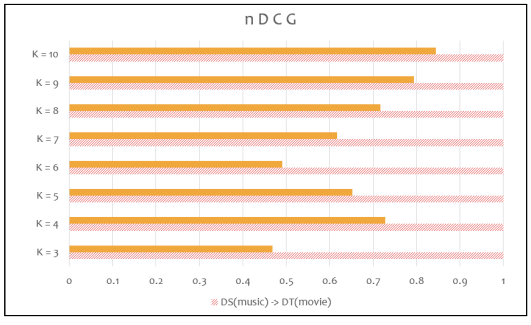


**그림 4 Normalized discounted cumulative gain(nDCG) 결과 그래프**
**Fig. 4. The normalized discounted cumulative gain (nDCG) result graph**

본 논문의 실험 결과를 이전 연구와 대조하여 분석하였다. 동일한 데이터 셋과 매개변수 설정을 기반으로 실험을 수행하였다. 이전 DDTCDR 논문에서 사용된 오토인코더 방식과 이 논문에서 제시한 알고리즘의 K-means 군집화 방식을 비교하였다. 그리고 트랜스포머 기법을 적용하여 초기 사용자 문제의 해결 능력을 측정하기 위해 실제 평점 데이터와의 MAE를 계산하였다. 또한, 추천 시스템의 정확도 평가를 위해 nDCG 결과값을 활용하였다.

**표 6. 교차 도메인에 사용된 공통 사용자 정보**
**Table 6. Public user information used for cross-domain**

| Music domain -> Movie domain | | |
|---|---|---|
| Rating data | Type. of users<br>Type. of items | 3618<br>9959 |
| Item meta data | Type. of year<br>Type. of genre<br>Type. of country | 20<br>39<br>246 |
| Movie domain -> Music domain | | |
| Rating data | Type. of users<br>Type. of items | 995<br>3836 |
| Item meta data | Type. of year<br>Type. of genre<br>Type. of country | 20<br>39<br>246 |

**표 7. 추가된 교차 도메인 신규 사용자-아이템 정보**
**Table 7. Added cross-domain with new user-item information**

| Music domain -> Movie domain | | |
|---|---|---|
| Rating data | Type. of users<br>Type. of items | 2623<br>6123 |
| Item meta data | Type. of duration<br>Type. of language<br>Type. of production | 20<br>39<br>246 |
| Movie domain -> Music domain | | |
| Rating data | Type. of users<br>Type. of items | 2623<br>6123 |
| Item meta data | Type. of publish<br>Type. of track<br>Type. of sales<br>Type. of artist job<br>Type. of artist birth<br>Type. of city | 20<br>14<br>5816<br>22<br>44<br>1530 |

첫 번째로, 표 7은 비교 실험을 진행하기 위한 공통 사용자 데이터를 보여주는 표이고, 표 8은 이 공통 사용자 데이터를 기반으로 학습된 신규 사용자-아이템 정보를 나타낸 표이다. $D_S$(music)에서 $D_T$(movie)로의 추천 성능을 살펴보면, 기존 알고리즘은 1.573235의 결과로 7% 더 높은 성능을 보였다. 그러나 $D_S$(movie)에서 $D_T$(music)로의 추천에서는 제안된 알고리즘이 0.460925의 값으로 55% 더 우수한 성능을 보였다. 이를 통해, 초기 사용자 문제에서 교차 도메인을 사용할 때, 메타 데이터의 종류가 적고 데이터 양이 많으면 기존 알고리즘이 더 좋은 성능을 보이지만, 메타 데이터의 종류가 많고 데

이터 양이 적을 때는 제안된 알고리즘이 더 높은 성능을 나타내는 것을 표 7에서 확인할 수 있다.

**표 8. 기존 알고리즘과 제안한 알고리즘의 성능 비교**
**Table 8. Comparison of performance between the existing algorithm and proposed algorithm**

| MAE | | | |
|---|---|---|---|
| $D_S$(music) → $D_T$(movie) | | $D_S$(movie) → $D_T$(music) | |
| DDTCDR | K-TCDR | DDTCDR | K-TCDR |
| 1.573235 | 1.303331 | 0.824711 | 0.468234 |
| | 17.16%↑ | | 43.22 %↑ |
| nDCG | | | |
| $D_S$(music) → $D_T$(movie) | | $D_S$(movie) → $D_T$(music) | |
| DDTCDR | K-SVDR | DDTCDR | K-SVDR |
| 0.912011 | 0.948764 | 0.939767 | 0.980360 |
| | 4.03 %↑ | | 4.32 %↑ |

## V. 결 론

본 연구는 DDTCDR 알고리즘에서 사용된 교차 도메인 추천 알고리즘의 한계를 극복하고자 시작되었다. DDTCDR 알고리즘에서는 AutoEncoder를 사용하여 임베딩을 생성하였으나, 이 방식은 입력과 출력이 동일하게 학습되기 때문에 새로운 데이터가 추가되더라도 기존의 학습 데이터와 유사해지는 과적합의 문제가 있었다. 따라서 본 연구에서는 K-means 클러스터링을 이용하여 사용자 데이터를 군집화하고, 이를 기반으로 임베딩을 생성하여 과적합 문제를 해결하였다. 또한 기존의 DDTCDR 알고리즘에서는 추천 개수의 제한이 없어서 너무 많은 아이템이 사용자에게 추천되어 추천 정확성이 떨어진다는 문제가 있었다. 이를 해결하기 위해 본 연구에서는 관련성이 높은 소수의 아이템만 추천하도록 하여 추천의 정확도를 높였다. 또한 K-means 클러스터링을 이용하여 시간 복잡도를 줄이고, 성능을 향상시키고자 하였다. 인공 신경망에서는 학습 데이터가 불규칙할 수 있으므로, 가중치를 일관되게 하기 위해 MinMax 방식의 데이터 정규화를 사용하였다. 학습에는 Stochastic Gradient Descent(SGD)를 사용하였으며, 손실 함수로는 CrossEntropy를 사용하였다. 마지막으로, 본 연구에서는 트랜스포머 기반의 추천을 이용하여 사용자에게 가장 적합한 아이템 TOP-N을 추천하였다. 추천의 정확도를 평가하기 위해 nDCG를 사용하였으며, 각 구간별로 가장 1에 가장 가까운 값을 판별하였다. 본 연구에서 제안된 알고리즘은 트랜스포머 기반의 교차 도메인 추천으로, 더욱 정확하고 개인화된 추천을 제공한다. K-means 클러스터링을 이용하여 생성된 임베딩과 트랜스포머 기반의 교차 도메인 추천을 통합하여 높은 성능의 추천을 제공하였다. 이 알고리즘은 다양한 사용자 초기 데이터를 활용하면 성능이 더 향상될 것이며, 더 나은 성능의 다층 신경망 알고리즘이 개발되더라도 기존 알고리즘을 쉽게 바꿔 적용할 수 있어 더 좋은 성능을 달성할 수 있는 잠재력이 있다.

## References

[1] E. Y. Bae and S. J. Yu, "Transitive Similarity Evaluation Model for Improving Sparsity in Collaborative Filtering", Journal of KIIT, Vol. 16, No. 12, pp. 109-114, 2018.
DOI : http://dx.doi.org/10.14801/jkiit.2018.16.12.109

[2] E. Y. Bae and S. J. Yu, "Keyword-based Recommender System Dataset Construction and Analysis", Journal of KIIT, Vol. 16, No. 6, pp. 91-99, 2018.
DOI : http://dx.doi.org/10.14801/jkiit.2018.16.6.91

[3] Tevfik Aytekin, "Clustering-based diversity improvement in top-N recommendation", Journal of Intelligent Information Systems, Vol. 42, No. 1,pp. 1-18, 2014.
DOI : https://doi.org/10.1007/s10844-013-0252-9

[4] Shual Zhang, Lina Yao, Aixin Sun, Yi Tay, "Deep Learning based Recommender System: A Survey and New Perspectives", ACM Computing Surveys, Vol. 52, No. 5, pp. 1-38, 2019.
DOI : https://doi.org/10.1145/3285029

[5] SJ Yu, "A Study of Improvement of Individual Item Diversity in Collaborative Filtering-based Recommendation", Journal of KIIT, Vol. 14, No. 8, pp. 89-94, 2016.
DOI : http://dx.doi.org/10.14801/jkiit.2016.14.8.89

[6] Mingyang Jiang, Zhifeng Zhang, "A collaborative filtering recommendation algorithm based on information theory and bi-clustering", Neural Comput & Applic 31, pp.8279-8287, 2019.
DOI : https://doi.org/10.1007/s00521-018-3959-2

[7] F Ricci, L Rokach, B Shapira, "Cross-domain recommender systems", ICDMW 2011Workshops, pp. 919-959, 2011.
https://doi.org/10.1007/978-1-4899-7637-6_27

[8] S Zhang, L Yao, B Wu, X Xu, X Zhang, L Zhu, "Unraveling Metric Vector Spaces With Factorization for Recommendation", IEEE Transactions on Industrial Informatics, Vol. 16, No. 2, pp.732-742, 2019.
https://doi.org/10.1109/TII.2019.2947112

[9] Tae-Hoon Kim, Sung Kwon Kim," SVD-based Cross-Domain Recommendation Using K-means Clustering", Journal of KIISE, Vol. 49, No.5 , pp. 360-368 May. 2022.
DOI : 10.5626/JOK.2022.49.5.360

[10] Ha-Yeon Cho, Hyeok-Min Lee, Ho-Sang Moon, Sung-Wook Shin, Sung-Taek Chung, "Improvement of Cognitive Rehabilitation Method using K-means Algorithm", The Journal of The Institute of Internet, Broadcasting and Communication (IIBC), Vol. 18, No. 6, pp.259-268, 2018.
DOI : 10.7236/JIIBC.2018.18.6.259

## 저 자 소 개

**김 태 훈(정회원)**

• Graduated with a Bachelor's degree in Computer Engineering from Tech University of Korea in 2020. Graduated with a Master's degree in Computer Engineering from Chung-Ang University in 2022. Interested in image processing, natural language processing, and recommendation systems.

**김 영 곤(정회원)**

• Young Gon Kim earned his BS in Electronic Engineering from Kyungpook University in 1983 and his MS from Yonsei University in 1985. He received his PhD from KAIST in 2000 and is now a professor at the Department of Computer Science at Tech University of Korea, focusing on Software Engineering, Information Communication Systems, and Object-Oriented Analysis and Design.

**박 정 민(정회원)**

• Jeong-Min Park earned his BS in Computer Science from Tech University of Korea in 2003, and his MS and PhD from SungKyunKwan University in 2005 and 2009, respectively. He is now a professor of Computer Science at Korea Polytechnic University, specializing in Cyber Physical Systems, Autonomic Computing, and Software Engineering.