

# 셋톱박스 오디언스 타겟팅을 위한 세션 기반 개인화 추천 시스템 개발\*

차지수

연세대학교 산업공학과  
(jisoo.cha@yonsei.ac.kr)

양재원

연세대학교 산업공학과  
(jwyang1993@yonsei.ac.kr)

장서호

연세대학교 디지털메달리틱스융합학과  
(hoya915@yonsei.ac.kr)

김우주

연세대학교 산업공학과  
(wkim@yonsei.ac.kr)

정구섭

연세대학교 산업공학과  
(supkoon@yonsei.ac.kr)

백상덕

연세대학교 산업공학과  
(eagleagle@yonsei.ac.kr)

박태준

한국방송광고진흥공사  
(tj1000@kobaco.co.kr)

김우영

연세대학교 산업공학과  
(timothy@yonsei.ac.kr)

이원준

연세대학교 산업공학과  
(streo222@yonsei.ac.kr)

정찬우

한국방송광고진흥공사  
(cwjeong20@kobaco.co.kr)

셋톱박스 오디언스(TV 시청자) 타겟팅의 핵심은 오디언스의 시청패턴을 분석하여 광고의 효과성이 높을 것으로 예상되는 오디언스에게 맞춤형 광고를 내보내는 것이다. 세션 기반 추천 시스템은 인터넷 광고 추천, 유저 검색 기록 기반 추천 등에 많이 이용되고 있지만, TV 광고의 측면에서 셋톱박스 데이터 수집의 어려움을 이유로 연구하기에 어려움이 있었다. 또한 오디언스 개인의 식별정보가 있는 데이터에서, 오디언스의 선호가 반영되는 시청 패턴을 모델링하는 데 한계가 있었다. 따라서 본 연구에서는 한국방송광고진흥공사(KOBACO)와 방송3사(SKB, KT, LGU+)와의 협업을 통해 익명화된 오디언스 4,847명의 6개월간 시청 데이터를 확보하여 연구를 진행하였으며, 유저-세션-아이템의 계층적 구조를 가지는 개인화 세션 기반 추천 시스템을 개발하여 성능 검증을 진행하였다. 그 결과, 셋톱박스 오디언스 데이터셋과 그 외 검증을 위한 2개의 데이터셋에서 제안된 모델이 비교 대상 모델보다 높은 성능을 보이는 것을 확인하였다.

**주제어** : 세션 기반 추천 시스템, 오디언스 타겟팅, 그래프 뉴럴 네트워크, 추천 시스템, 표현 학습

논문접수일 : 2023년 5월 17일

논문수정일 : 2023년 6월 22일

게재확정일 : 2023년 6월 25일

원고유형 : 학술대회 Fast Track

교신저자 : 김우주

## 1. 서론

정보통신기술이 발달하면서, 고객들의 관심사를 파악하여 고객 맞춤형 광고를 추천하는 시스템이 날이 갈수록 발전하고 있다. 전자상거래 및

인터넷 광고에서 이러한 추천 시스템이 활발하게 사용되고 있다(나혜연, 남기환, 2020; 홍태호 등, 2022; 장동수 등, 2023). 이 중 유저의 취향, 혹은 선호를 파악하여 아이템을 추천하는 시스템을 개인화 추천 시스템이라고 부르는데, 방송

\* 본 논문은 국토교통부의 스마트시티 혁신인재육성사업 및 중소기업기술정보진흥원의 지원을 받았습니다

업계에서도 TV 셋톱박스 데이터를 바탕으로 유저를 분류하여 개인 맞춤형 광고를 방영한다. 이를 셋톱박스 오디언스<sup>1)</sup> 세그멘테이션이라고 부른다. 한국에서 TV 셋톱박스는 SK BroadBand 기준 910만명 이상<sup>2)</sup>으로 많은 가입자를 보유하고 있다. 이런 셋톱박스 데이터를 분석하여 오디언스의 구매 가능성이 높은 맞춤형 광고를 송출하는 것은 방송업계와 광고업계의 주요 과제이다. 현재까지 방송업계와 광고업계는 셋톱박스 오디언스를 여러 가지 분석 방법론을 사용하여 분석하고 분류하여 타겟팅하고 있다. 그러나 셋톱박스의 수가 늘고, 채널의 수가 늘어남에 따라 오디언스의 선택의 폭이 넓어지면서 보다 더 정밀한 타겟팅을 필요로 하게 되었다.

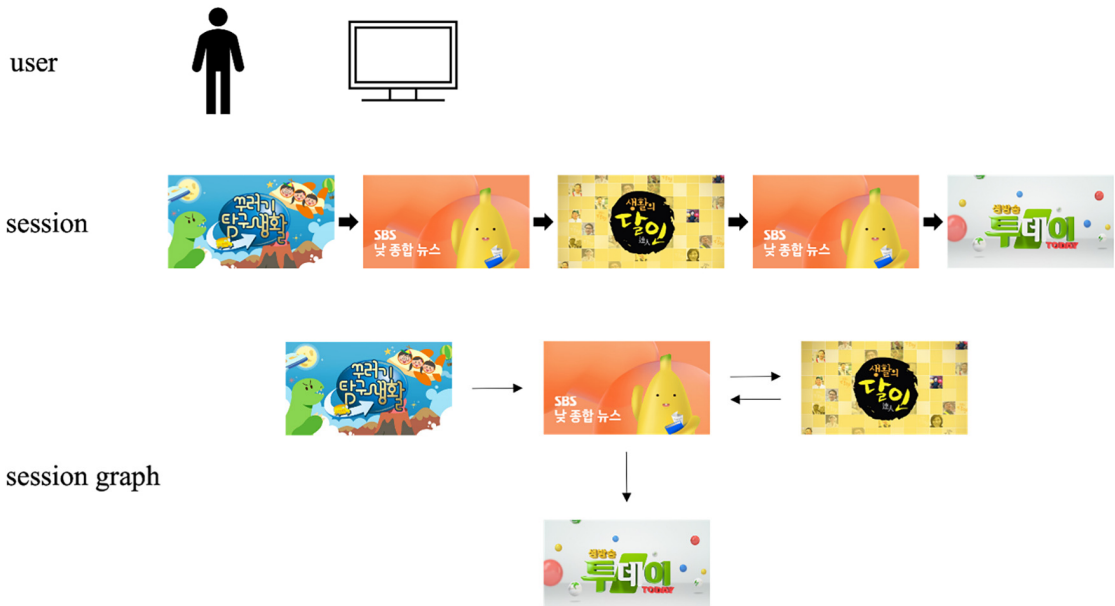
오디언스 타겟팅을 위한 기법으로 본 연구에서는 세션 기반 추천 시스템을 제안한다. 전통적인 추천 시스템은 Matrix Factorization(Mnih, Salakhutdinov, 2007), Markov chain(Shani, Braffman, and Heckerman, 2002), K-Nearest Neighbors(Sarwar et al., 2001) 등의 머신러닝 기법이 활용되었으나, 추천 정확도가 높지 않고, 특히 유저 여러 명의 세션 길이가 긴 세션에 대해 적용하기 힘들다는 면이 있었다. 오디언스의 시청 데이터는 긴 시간 축적되어 기록되기 때문에, 전통적 추천 시스템을 사용하면 학습 시간 문제로 적용이 힘들다. 이 문제의식에서 시작된 세션 기반 추천 시스템은 추천 시스템의 한 분야이다. 세션은 보통 특정 시간 구간에서 유저와 아이템의 상호작용을 의미한다. 기존 추천 시스템에서는 아이템의 시간 순서를 고려하지 않았지만, 세션 기반 추천 시스템은 유저가 상호작용한 아이템의 순서를 고려하여 학습을 진행한다. 셋톱박스 오디언스를 예로 들면, 개별

오디언스가 일정 기간 시청한 일련의 TV 프로그램 시퀀스를 한 세션으로 볼 수 있다. 전체 구간으로 본다면 오디언스가 셋톱박스에 가입한 시점으로부터 현재까지의 오디언스 활동을 하나의 세션으로 볼 수 있으나, 세션마다 반영된 오디언스의 행동패턴을 고려하면 한 오디언스는 세션을 여러 개 가질 수 있다. 한 세션을 TV를 켜고 끄는 것으로 나눌 수도 있고, 혹은 하루 단위의 프로그램 시청을 한 세션이라 볼 수도 있다.

세션 기반 추천 시스템은 초기에는 Markov Chain 등의 머신러닝 기법이 사용되었으나, 계산량이 많고 모델링할 유저의 수가 늘어나면 계산 시간이 크게 증가한다는 문제점이 있었다. 따라서 최근에는 딥 러닝 알고리즘이 세션 기반 추천 시스템에 제안되었다. 딥 러닝 알고리즘은 Multi Layer Perceptron과 비선형 활성화 함수(activation function)를 사용하여 복잡한 함수를 근사시킬 수 있다는 universal approximation theorem(Hornik, Kurt, 1989)이 이론적으로 뒷받침되어 현재 추천 시스템에서도 많이 사용되고 있다. 세션 기반 추천 시스템에 일반적으로 사용되는 딥 러닝 알고리즘은 시간 순서를 이용한 순환 신경망(Recurrent Neural Network)으로, Hidasi et al. (2016)가 이를 추천 시스템에 최초로 소개하였다. 순환 신경망은 시간 순서가 있는 데이터에 강력한 성능을 보이므로 세션 기반 추천 시스템에 널리 사용되어 왔다. 그러나 순환 신경망은 아이템 간의 복잡한 상호작용을 고려하지 못한다는 문제점이 있다. 예를 들어 TV를 시청할 때 시청한 프로그램의 시퀀스가 한 세션이 되는데, 프로그램 간의 관계는 서로 독립적이지 않다. 같은 작가, 혹은 같은 장르를 공유하는 프로그램들은 오디언스가 TV

1) 셋톱박스 시청자를 의미한다.

2) [https://skbroadband.com/footer/page.do?menu\\_id=F06000000](https://skbroadband.com/footer/page.do?menu_id=F06000000)



〈그림 1〉 오디언스 시청기록을 그래프로 구성하는 방법<sup>3)</sup>

시청을 할 때 시간대가 비슷하다면 같이 시청될 것이라고 볼 수 있으며, 모델링을 할 때도 이러한 점을 고려해야 한다.

이런 관점에서 Graph 구조를 차용한 세션 기반 추천 시스템이 제안되었다. 그림 1을 보면, 세션 내 유저의 아이템과의 상호작용은 그래프로도 나타낼 수 있다. 그래프를 사용한 딥 러닝 모델을 활용하면 유저, 아이템뿐만 아니라, 아이템 간의 상호작용이 모델링에 반영될 수 있다는 강점이 존재한다. 그래프를 사용한 세션 기반 추천 시스템이 처음 소개된 논문으로는 SR-GNN(Wu et al., 2019)이 있다. 그러나 이 모델에도 한계점은 존재한다. 그래프를 사용한 세션 기반 추천 시스템은 인풋으로 들어오는 각 세션을 식별한다. 다시 말해 유저가 본 세션에서 유저 정보를 제외하고 세션 정보만을 이용한다. 식별화한

세션으로 세션 기반 모델링을 하면 세션의 계층적 상위구조인 유저의 정보를 활용하기 힘들므로 세션 기반 추천 시스템에서 유저의 정보를 결합시키는 부분을 연구할 필요가 있다.

본 연구의 문제의식은 다음과 같다. 첫째, TV 오디언스 시청기록의 데이터 수집이 현실적인 여건상 어려우므로, 현재까지 오디언스 시청기록을 딥러닝 알고리즘으로 모델링하려는 시도가 없었고 이를 효과적으로 모델링하기 위한 방안 또한 아직 제안되지 않았다. 둘째, 현재까지의 세션 기반 추천 알고리즘에서는 많은 경우 각 세션을 익명적으로 보며 유저 정보를 모델링에 개입시키지 않아서, 유저 정보가 가용한 데이터에서도 이를 무시하고 모델링하는 방법이 사용되었다. 둘째, 그래프 기반 추천 기반 알고리즘에서는 아이템 간의 상호작용은 고려하지만, 개별

3) 지상파 방송사 SBS의 프로그램 썸네일을 사용하였다. [https://www.sbs.co.kr/tv?div=gnb\\_pc](https://www.sbs.co.kr/tv?div=gnb_pc)

유저와의 상호작용은 고려하지 않았다. 다시 말해서, 어떤 유저가 한 아이템에 가지는 관심의 정도가 다른 유저와 모두 같은 문제가 있다.

본 연구에서는 한국방송광고진흥공사(KOBACO)와 방송3사(SK, KT, LGU+)와 협력하여 익명 처리를 거친 고객 4847명의 셋톱박스 데이터를 확보하여 셋톱박스 시청이력 모델링을 진행하고 고도화하였다. 또한 TV 시청에는 오디언스의 개인적 취향이 많이 반영된다고 가정하고 데이터에 맞는 개인화 추천 시스템을 모델링한 연구모델에 대해 설명할 것이다. 또한 KOBACO 데이터와 기타 벤치마크 데이터로 모델의 성능을 검증할 것이다. 이 일련의 과정의 결과를 통해 셋톱박스 오디언스를 활용한 광고 타겟팅 과제에서 개인화 추천 시스템이 효과적인 것임을 제안하고, 연구의 향후 과제와 한계점에 대해 설명할 것이다.

## 2. 이론적 배경

### 2.1. 세션 기반 추천 시스템(Session-Based Recommendation System)

추천 시스템은 유저의 관심도가 가장 높은 아이템을 유저에게 추천해 주는 일련의 알고리즘을 말한다. 대표적인 추천 시스템 방법론으로는 Matrix Factorization(Koren et al., 2009)이 있는데, 이는 유저-아이템 매트릭스를 행렬분해하여 유저와 아이템의 잠재요소로 분해하는 방법론으로, 유저와 아이템의 단편적 관계를 모델링하는데 주로 사용되었다. 하지만 Matrix Factorization은 유저와 아이템의 상호작용 유무만 모델링 되는 관계로, 시시각각 변화하는 유저의 의도를 담아

내는 데 한계가 존재한다. 이를 위해 세션 기반 추천 시스템이 도입되었는데, 이는 개별 유저의 아이템과의 상호작용의 시퀀스를 모델링한다. 일반적으로 세션은 특정 시간 구간에서 유저와 아이템의 상호작용을 의미한다. 유저의 활동기록은 시간정보의 시퀀스로 기록되어 있는 경우가 많으며, 유저의 의도를 보다 세밀하게 파악하기 위해서는 시퀀스를 여러 개의 세션으로 나눌 필요가 있다. 이를테면 TV를 볼 때, 아침에는 하루 동안 일어난 일에 대한 정보를 얻기 위해 TV를 시청하며, 저녁에는 홈쇼핑에서 물건을 사기 위해 TV를 시청할 수 있다. 이렇게 하나의 시퀀스를 여러 개의 세션으로 나누면 세션마다 유저의 의도를 개별적으로 볼 수 있게 된다.

세션 기반 추천 시스템은 유저-아이템 상호작용의 시퀀스를 모델링하는 방법이다. 대표적인 방법으로는 Markov Chain을 이용한 알고리즘(Shani, Brafman, and Heckerman, 2002)이 있다. Markov Chain을 이용한 히든 마르코프 모델은 유저의 의도를 은닉상태(hidden state)로 모델링하여 은닉 상태 간의 전이확률(transition probability)과 은닉 상태에서의 방출확률(emission probability)을 추정하는 알고리즘인데, 이 또한 추천 시스템에서 활용된 바 있다(Sahoo et al., 2012). 그러나 이런 전통적인 방법은 하나의 세션을 모델링하는 데 적합하며, 여러 개의 세션을 모델링하여 추천 알고리즘에 활용하려고 할 때는 세션 수가 증가함에 따라 적합(fitting) 시간이 많이 걸리는 단점이 존재한다. 또한 현재 상태가 바로 직전 상태만의존적이라는 Markov Chain의 기본 가정은, 시시각각 유저의 의도와 선호가 반영되는 세션 데이터에 적절하지 않을 수 있다.

## 2.2. 딥러닝을 사용한 세션 기반 추천 시스템

딥러닝을 사용하면 개별 세션의 로그우도(log-likelihood)를 계산하는 Markov Chain 방법과 달리 인풋 도메인에서 타겟으로 가는 일반화된 함수를 학습하기 때문에, 보다 많은 데이터를 빠른 시간 내에 학습 가능하다는 장점이 존재한다. 또한 상태간 강한 독립성 가정을 하지 않고도 연산이 가능하다는 장점이 있다. NARM(Li et al., 2017)은 GRU(Chung et al., 2014)와 어텐션 메커니즘을 활용하여 세션에 대한 hidden representation을 도출하는 인코더-디코더 구조의 세션 기반 추천 시스템이다. STAMP(Liu et al., 2018)는 어텐션 메커니즘과 다층 퍼셉트론(multi-layer perceptron)을 사용하여 유저의 의도를 담아내려고 하였다. 딥러닝을 사용한 세션 기반 추천 시스템은 전통적인 모델에 비해 세션 기반 추천 시스템에서 만족할 만한 성능을 보였지만, 여전히 유저-아이템의 상호작용을 모델링할 뿐 아이템과 아이템 간의 상호작용에 대해서는 모델링이 어려운 한계가 있었다.

## 2.3. 그래프를 적용한 세션 기반 추천 시스템

그래프(graph)는 노드(node)와 엣지(edge)로 구성되어 있는 자료구조이다. 노드는 하나의 데이터 포인트를 의미하고, 엣지는 노드 간의 상관관계를 의미한다. 이 때 엣지는 노드 간의 상관관계의 정도를 가중치(weight)로 가질 수 있고, 엣지가 두 개의 노드를 연결할 경우 방향성 또한 가져서 노드 간의 조건부 상관관계를 뜻할 수 있다. 그래프의 연결 관계를 표현하기 위한 여러 가지 자료구조가 있지만 인접행렬(adjacency matrix)은 이를 사용한 행렬곱으로 노드 간의 연결관계를 연산할 수 있기 때문에 널리 사용된다. 그래프

뉴럴 네트워크(Graph Neural Network, GNN)은 그래프 자료구조에 뉴럴 네트워크를 적용하는 방법론이다. 그래프 뉴럴 네트워크는 한 노드를 기준으로 주변 노드의 정보를 종합(aggregate)해서 노드의 정보를 업데이트하는 메시지 전파(message passing) 알고리즘이다. 이는 그래프 노드에 대해 표현 학습을 수행하는 방법으로 널리 사용되었다. GCN(Kipf & Welling, 2016)은 체비셰프 다항식(Chebyshev Polynomial)을 사용하여 그래프 스펙트럼에 그래프 필터를 적용하였고, GraphSAGE(Hamilton et al., 2017)는 그래프에 대한 inductive representation learning을 위해 보다 효과적인 aggregation 방법을 제안하였다. 또한 GAT(Velickovic et al., 2017)는 aggregation function을 어텐션 네트워크로 정의하여 노드 분류(node classification) 등의 downstream task에서 좋은 성능을 보였다.

아이템 간의 상호작용을 모델링하기 위해 딥러닝을 적용한 추천 시스템에 그래프 뉴럴 네트워크가 제안되었다. Wu et al.(2019)은 아이템을 노드, 아이템 간의 시간순서를 엣지로 해서 세션 그래프를 만들어 GNN으로 유저의 intention을 모델링하였다. 여기에서는 그래프가 시간 순서로 구축되었으므로 그래프에 순환 신경망의 개념을 적용한 Gated Graph Neural Network를 사용해 아이템의 임베딩과 세션을 학습하였다. 그래프를 적용한 세션 기반 추천 시스템은 많은 추천 벤치마크에서 좋은 성능을 보였으나, 세션에 대한 유저 정보를 익명화하여 유저 정보가 존재하는 데이터가 주어질 때 유저 정보의 손실을 야기하였다. 그래프를 사용한 세션 기반 추천 시스템에 유저 정보를 추가하는 작업은 현재 활발히 진행되고 있다.

### 3. 연구모델

본 연구에서는 오디언스 광고 타겟팅을 위한 개인화 추천 시스템 학습 방법을 다음과 같이 제안한다. 먼저 모델 학습에 있어 우리가 해결하고자 하는 태스크를 정의하고, 그 후 제안된 모델을 설명할 것이다.

#### 3.1. 모델 개요

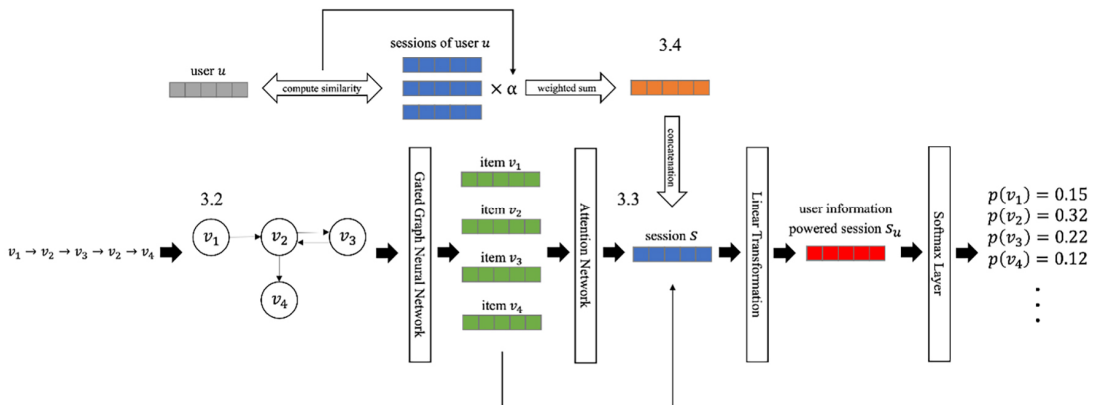
개인화 추천 모델 데이터를 계층적으로 구성하면 상위 계층부터 유저, 세션, 아이템으로 각각 구성할 수 있다. 유저는 복수의 세션을 가질 수 있고 세션은 복수의 아이템을 포함한다. 아이템과 유저는 고유한 개수를 가지고 있으나, 세션은 아이템의 시간 순서 배열이기 때문에 동일한 세션은 존재하지 않는다고 가정한다.

$V = \{v_1, v_2, \dots, v_m\}$ 을  $m$ 개의 item set으로 하고, 길이  $n$ 인 세션  $s_u = [v_{t_1}, v_{t_2}, \dots, v_{t_n}]$ 는 유저  $u$ 가 상호작용한 아이템을 시간 순서로 배열한 시퀀스라고 둔다. 이때 item set은 중복이 허용되지 않고, 세션 내의 아이템은 중복이 허용됨에

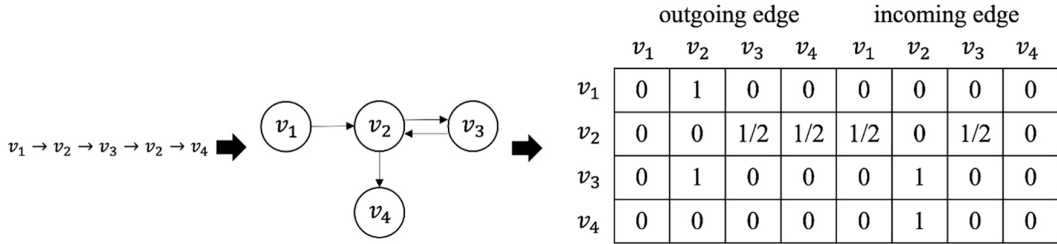
유의한다. 이때 세션  $s_u$ 를 가지고 그래프  $G$ 를 그릴 수 있는데,  $V_s$ 를 세션  $s_u$ 에 존재하는 중복이 허용되지 않는 item set이라고 두고 시간 순서대로 방향성을 가진 엣지를 연결하면 그래프  $G = (V_s, E_s)$ 를 얻을 수 있다. 우리가 해결하고자 하는 문제는 세션  $s_u = [v_{t_1}, v_{t_2}, \dots, v_{t_{n-1}}]$ 이 주어졌을 때, 바로 다음 아이템  $v_{t_n}$ 을 예측하는 것이다. 본 논문에서는 개인화된 추천을 위해 해당 문제에 유저에 대한 임베딩을 개입시켜 성능 향상을 도모하였다.

#### 3.2. 그래프 구조와 인접행렬(Adjacency Matrix)

세션  $s_u = [v_{t_1}, v_{t_2}, \dots, v_{t_{n-1}}]$ 가 주어졌을 때 세션 그래프는 그림 2의 3.2 부분의 그래프와 같이 그릴 수 있다. 그래프의 각 노드는 유저가 상호작용한 item set  $V_s$ 의 각 원소를 나타내고, 엣지는 시간 순서 기준 이전 노드에서 다음 노드로 방향성 있게 연결되는 그래프 구조이다. 또한 주어진 세션에 따라 엣지가 여러 번 연결될 수 있지만, 중복되는 엣지는 무시한다. 추후 연구에서 엣지의 연결 횟수에 따라 가중치를 주는 방법이



〈그림 2〉 제안된 모델



〈그림 3〉 인접행렬 구축 예시

사용될 수 있고, 세션의 각 아이템에서 머무는 시간(dwelling time)과 아이템 간의 시간 간격(interval time)을 엣지 속성으로 사용하는 것을 고려해 볼 수 있을 것이다.

본 논문에서 제안된 모델의 기본모델(backbone model)에서는 SR-GNN(Wu et al., 2019)에서 사용한 인접행렬의 구축 방법과 GGNN(Gated Graph Neural Network)을 사용하였다. 인접행렬의 구축 예시는 그림 3과 같다. 세션  $s = [v_{t_1}, v_{t_2}, \dots, v_{t_{n-1}}]$ 의 고유한 아이템의 개수를  $|V_s|$ 라고 할 때, 해당 세션 그래프의 인접행렬  $A$ 는  $A \in R^{|V_s| \times 2|V_s|}$ 의 형태로 구축된다. 다시 말해서 인접행렬은 세션의 고유한 아이템 개수만큼의 행, 그 두배의 개수의 열을 가지도록 구축된다. 행에서  $i$ 번째 해당하는 아이템을  $v_i$ 라고 하고, 열에서  $j$ 번째에 해당하는 아이템을  $v_j$ 라고 하면,  $A_{i,j}$ 는 인접행렬  $A$ 에서 아이템  $v_i$ 와  $v_j$ 의 연결관계이다. 이때  $0 < j \leq |V_s|$ 일 때는  $v_i$ 에서  $v_j$ 로 가는 outgoing edge가,  $|V_s| < j < 2|V_s|$ 일 때는  $v_i$  기준으로  $v_j$ 에서 들어오는 incoming edge가 각각  $1/\sum_{j=1}^{|V_s|} A_{i,j}$ ,  $1/\sum_{j=|V_s|+1}^{2|V_s|} A_{i,j}$ 으로 정의된다. 이 인접행렬  $A$ 는 outgoing edge와 incoming edge에 대한 정규화된 인접행렬을 각각 구축한 다음 열 방향으로 결합(concatenation)한 것과 동일하다.

### 3.3. 세션 벡터(Session Vector)

본 논문에서는 아이템 임베딩과 유저 임베딩, 두 종류의 임베딩(embedding)을 학습하였다. 임베딩의 은닉 사이즈(hidden size)를  $h$ 라고 할 때, 아이템 임베딩은 각 아이템을 표현하는 학습 가능한 임베딩  $v \in R^h$ 이고, 유저 임베딩은 유저를 표현하는 학습 가능한 임베딩  $u \in R^h$ 로 구성된다. 두 가지 임베딩 모두 제안된 모델에서 오차 역전파(backpropagation)가 수행되면서 학습된다. 기본모델에서는, 각 아이템에 대한 lookup table을 통해 각각의 노드가 각 아이템의 임베딩이 되도록 그래프를 구성한다. 이 아이템 임베딩은 GGNN을 통해 주변 아이템의 임베딩을 받아서 그래프 필터를 거쳐 업데이트 된다.

업데이트 이후에는 각 아이템에 대한 임베딩이 새롭게 구성되고, 본 논문에서는 두 종류의 세션 임베딩을 결합하여 최종 세션 벡터를 구성하는 방법인 하이브리드 세션 임베딩을 사용하였다. 두 종류의 세션 임베딩은 local embedding과 global embedding이다. 먼저 local embedding은 마르코프 과정(Markov Process)과 유사하게, 유저가 상호작용할 다음 아이템은 바로 직전 아이템과만 상관이 있을 것이라는 가정 하에 인풋 세션의 마지막 아이템의 임베딩을 유저의 국소적 관심 부분인 local embedding으로 본다. Global

embedding은 세션 전체에 대해 유저의 관심이 퍼져 있을 것이라 가정하고 마지막 아이템의 임베딩과, 다른 모든 아이템의 임베딩과의 유사성을 내적(dot product)으로 비교하여 attention score를 만든 뒤, 세션의 각 아이템의 임베딩에 attention score를 가중치로 하여 가중합(weighted sum)한다. 그 뒤 local embedding, global embedding 두 임베딩을 결합(concatenation) 하고 Linear Transformation Layer를 통과시켜서 세션  $s$ 에 대한 global 정보, local 정보를 결합한 세션 벡터를 만든다. 이 세션 벡터는 유저의 local intention과 global intention을 모두 반영한다고 할 수 있겠다. 세션 벡터는 그림 2의 3.3 부분에서 확인할 수 있다.

### 3.4. 유저 임베딩과의 결합

본 논문에서는 기반모델(backbone model)으로 도출된 임베딩에 유저 정보를 추가하여 모델을 설계하였다. 현재까지의 과정으로 나온 세션 임베딩을 해당 세션의 유저 임베딩과 내적하면 유사도를 비교할 수 있는데, 같은 유저의 모든 세션에 대해 유저 임베딩과 서로 유사도를 비교해서 해당 유저의 세션 개수만큼의 길이의 attention score 벡터를 얻는다. 이때 attention score와 해당 유저에 대한 모든 세션 임베딩을 가중합하여 해당 유저의 정보가 포함된 유저 벡터를 얻을 수 있다. 이를 원래 모델 인풋의 세션 벡터와 결합하여 학습가능한 linear transformation layer를 통과시키면 유저 정보가 부가된 세션 벡터(user-powered session vector)를 최종적으로 얻어 학습에 사용할 수 있다. 유저 임베딩과의 결합 부분은 그림 2의 3.4 부분에서 확인 가능하다.

### 3.5. 목적함수(Objective Function)

학습 데이터 내의 총 아이템 집합의 크기를  $|V|$ 라고 할때, 타겟 아이템과의 유사성을 산출하기 위해 최종적으로 나온 세션 벡터를 모든 아이템 집합의 정보가 있는 행렬  $M \in R^{|V| \times h}$ 와 곱한다. 그러면 세션 벡터의 모든 아이템과의 유사성 벡터  $\alpha \in R^{|V|}$ 를 얻을 수 있다.  $\alpha$ 의 각 요소(element)를 정규화하여 확률로 만들기 위해 소프트맥스 계층(softmax layer)를 통과시키고 이산형 데이터에서 주로 사용되는 손실함수 크로스 엔트로피(cross entropy loss)를 사용해 예측된 값과 실제값의 차이를 계산한다. 마지막으로 역전파(backpropagation)를 사용하여 학습 가능한 파라미터들을 업데이트한다.

### 3.6. 추천(Make Recommendations)

학습이 완료된 후 아이템을 추천할 때는 소프트맥스 계층을 거친 각 아이템의 확률을 내림차순으로 정렬하여 추천한다. 만약 5개의 아이템을 추천한다고 하면 높은 순서대로 5개의 아이템이 최종적으로 결과로 나오게 된다.

## 4. 실험 내용

먼저 4.1. 장에서는 실험에 사용된 데이터셋을 간단하게 소개하고, 4.2. 장에서는 실험에 사용된 평가지표를 소개할 것이다. 4.3. 장에서는 비교 대상이 되는 모델을 소개하고, 4.4. 장에서는 학습에 사용된 파라미터를 소개한다. 마지막으로 4.5. 장에서는 실험 결과 설명 및 성능 비교를 수행할 것이다.



#### 4.1. 데이터셋 소개

연구에서 주목한 셋톱박스 추천을 위한 데이터인 KOBACO 데이터셋 이외에, 제안된 모델의 추천 시스템으로서의 성능을 검증하기 위해 유저 정보가 존재하는 세션 데이터셋을 두 가지 추가하여 실험을 진행하였다. 모든 데이터는 유저별로 세션을 시간 순으로 정렬하여 유저별 세션의 개수 기준으로 앞쪽 80%를 학습 데이터로, 나머지 20%를 검증 데이터로 사용하였으며, 세 데이터의 세부사항은 소제목 4.1.1., 4.1.2., 4.1.3.에서 각각 소개할 것이다

##### 4.1.1. KOBACO 데이터

본 논문에서는 한국방송광고진흥공사(KOBACO)와 협업하여 방송3사(SKB, KT, LGU+)의 셋톱박스 세션 데이터를 확보하여 실험을 진행하였다. 협약에 따라 전체 데이터셋은 공개하지 않는다. 셋톱박스는 보통 한 가족이 사용하는 경우가 많지만 개인화 추천을 위해 1인 가구의 셋톱박스만 수집되었다. 셋톱박스 원 데이터(raw data)는 익명화된 셋톱박스 아이디, 셋톱박스를 켜고 끈 시간, 채널정보밖에 없기 때문에 채널 편성표 데이터를 확보하여 프로그램으로의 매핑을 진행하였다. 매핑 이후에는 각 유저의 6달간의 프로그램 시퀀스를 얻을 수 있었다. 이 시퀀스에 대해 본 연구에서는 세션을 하루 단위로 나누었다. 지상파 방송에서 편성은 보통 오전 6시 기준으로 분리되므로 오전 6시부터 다음날 오전 5시 59분 59초까지를 하나의 세션으로 보았다. 또한 학습의 안정성을 위해 너무 긴 시간 시청된 프로그램과 너무 짧은 시간 시청된 프로그램을 제외하였다. 또한 전체 데이터에서 5번 이하로 등장한 프로그램을 제외하고, 길이 1인 세션을 제외하였다.

4847명의 전체 셋톱박스 유저 중 300명의 유저를 무작위 추출(random sampling)하여 실험에 사용하였다. 실험에 사용된 데이터셋의 설명은 표 1과 같다.

##### 4.1.2. Cosmetics 데이터

Cosmetics 데이터셋은 Michael (2020)이 배포한 medium cosmetics online store에서 수집된 5개월간의 유저의 쇼핑 데이터이다. 전체 데이터셋은 사용 가능하게 공개되어 있다. 전체 유저 중 학습에 사용되기에 충분한 유저 시퀀스를 가진 상위 100명의 유저만 뽑아 학습을 진행하였다. 세션이 미리 나누어져 있는 데이터셋이라 세션을 나누지 않았고, 길이가 3 미만인 세션을 제외하였다. 실험에 사용된 데이터셋의 설명은 표 1과 같다.

##### 4.1.3. AppUsage 데이터

AppUsage 데이터셋은 Efficient Computing Lab 홈페이지에서 공개적으로 열람 가능한 유저의 스마트폰 앱 사용 데이터이다(Shepard et al., 2011). 세션이 미리 나누어져 있지 않아 아이템의 시간 간의 간격이 10분 이상인 경우 세션을 분리하였다. 그 결과로 나온 세션 중 길이 3 미만인 세션을 제외하였고, 유저의 의도가 반영되지 않는 시스템 기본 앱을 제거하였다. 실험에 사용된 데이터셋의 설명은 표 1과 같다.

〈표 1〉 실험에 사용된 데이터셋 개요

| Dataset                | KOBACO  | Cosmetics | AppUsage |
|------------------------|---------|-----------|----------|
| Number of users        | 300     | 100       | 34       |
| Number of items        | 9,441   | 25,883    | 2,288    |
| Number of sessions     | 35,221  | 11,668    | 76,335   |
| Number of interactions | 741,001 | 298,685   | 647,797  |
| Mean session length    | 21.039  | 25.599    | 8.486    |

## 4.2. 평가지표

학습이 완료된 후 추천할 때는 소프트맥스 계층을 거친 각 아이템의 확률을 내림차순으로 정렬하여 추천한다. 이 추천된 결과를 가지고 제안된 모델의 성능을 검증한다. 성능 검증에 사용되는 평가지표로는 추천 시스템에서 흔히 사용되는 Hit rate, MRR(Mean Reciprocal Rank)을 사용할 것이다. 먼저 Hit Rate는 추천된 아이템 중 정답인 아이템이 존재할 비율이다. 또 추천된 아이템의 개수를 직접 정할 수 있다. 만약 개수를 5로 정할 경우 HR@5로 표현한다. 본 논문에서는 1,5,10,20개의 추천된 아이템에 대한 Hit Rate를 평가지표로 삼았다. 숫자가 작을수록 추천이 더 정확해야 하므로 더 어려운 추천 태스크라고 할 수 있다. MRR은 mean reciprocal rank의 약자로, 단순 추천된 아이템 중 정답이 있는지의 지표인 Hit Rate와 달리 추천된 아이템 중 정답이 존재

한다면, 추천된 아이템에서의 정답의 순위의 역수를 계산한다. 만약 정답 아이템이 추천된 아이템 중 1위로 존재할 때는 1, 3위로 존재한다면 1/3이 된다. 만약 추천된 아이템 중 정답 아이템이 없다면 0으로 둔다. MRR 또한 Hit Rate와 같이 몇 개의 추천된 아이템에 대해 지표를 계산할지 정할 수 있다. 본 논문에서는 위의 Hit Rate과 동일한 1,5,10,20을 평가지표로 삼았다.

## 4.3. 비교 대상 모델 소개

연구에서 제안된 모델의 기반모델(backbone model)로 사용된 SR-GNN(Wu et al., 2019)과 비교해서 성능 향상이 얼마나 되었는지를 측정하였다. 이 모델은 세션 기반 추천 태스크에서 준수한 성능을 보여주고 있지만, 유저 정보가 없이 단순 세션 정보만으로 추천을 수행하여 유저 정보를 사용하지 않는다는 단점이 있었다. 따라서 본 연구

〈표 2〉 실험 결과

| Dataset   | Model     | Hit Rate             |              |              |              |
|-----------|-----------|----------------------|--------------|--------------|--------------|
|           |           | Hit@1                | Hit@5        | Hit@10       | Hit@20       |
| KOBACO    | SR-GNN    | 17.22                | 34.25        | 41.69        | 49.55        |
|           | Our Model | <b>17.29</b>         | <b>34.36</b> | <b>41.84</b> | <b>49.79</b> |
| Cosmetics | SR-GNN    | 15.35                | 25.12        | 29.15        | 32.66        |
|           | Our Model | <b>17.72</b>         | <b>28.31</b> | <b>32.66</b> | <b>36.57</b> |
| AppUsage  | SR-GNN    | 44.37                | 81.03        | 88.59        | 92.79        |
|           | Our Model | <b>44.45</b>         | 81.03        | <b>88.69</b> | <b>92.95</b> |
| Dataset   | Model     | Mean Reciprocal Rank |              |              |              |
|           |           | Mrr@1                | Mrr@5        | Mrr@10       | Mrr@20       |
| KOBACO    | SR-GNN    | 17.22                | 23.50        | 24.50        | 25.04        |
|           | Our Model | <b>17.29</b>         | <b>23.57</b> | <b>24.58</b> | <b>25.12</b> |
| Cosmetics | SR-GNN    | 15.35                | 19.00        | 19.53        | 19.78        |
|           | Our Model | <b>17.72</b>         | <b>21.69</b> | <b>22.28</b> | <b>22.55</b> |
| AppUsage  | SR-GNN    | 44.37                | <b>58.46</b> | <b>59.50</b> | 59.80        |
|           | Our Model | <b>44.45</b>         | 58.44        | 59.49        | 59.80        |

의 실험에서는 제안된 모델에서의 유저 정보의 포함 유무가 성능에 얼마나 영향을 주었는지, 유저 정보의 결합 방식이 적절한지 이전 소제목에서 설명한 평가지표를 통해 살펴볼 것이다.

#### 4.4. 파라미터 세팅

이 장에서는 실험에 사용된 세팅을 설명한다. 먼저 아이템 임베딩과 유저 임베딩은 동일하게 128을 사용하였다. 배치 사이즈 256으로 20 epoch의 실험을 수행하였으며, 초기 learning rate는  $5e-4$ 를 사용하였고, mini-batch Adam optimizer를 사용하였다. Step learning rate scheduler를 사용하여 3번째 epoch 마다 0.1씩 learning rate를 감소시켰다. 각 뉴럴 네트워크의 파라미터는 랜덤 uniform 초기화하였다. 또한 학습 종료 시까지 epoch마다 학습 후 검증을 수행하고 검증 성능 중 가장 높은 성능을 해당 모델의 최종 검증 성능으로 보았다.

#### 4.5. 실험 결과 및 성능 비교

표 2에 각 모델의 3개의 검증 데이터셋에 대한 성능을 비교하였다. 비교 대상 모델과 성능을 비교해서 성능이 더 좋은 부분을 굵은 글자로 처리하였다. 먼저 KOBACO 데이터셋에서는 비교 대상에 비해 제안된 모델의 성능지표가 모두 우수하였다. 셋톱박스 오디언스의 시청 패턴에는 오디언스의 선호가 반영되어 있고, 제안된 모델이 오디언스의 임베딩을 잘 모델링했다고 볼 수 있다. 둘째로 Cosmetics 데이터셋에서도 제안된 모델은 우수한 성능을 보였다. 특히 가장 어려운 태스크인 Hit@1, MRR@1에서 유의미한 차이가 보인다. 제안된 모델은 비교 대상 모델에 비해 정확한 추천을 하는 비율이 높다는 것을 알 수

있다. 마지막으로 AppUsage 데이터에서는 전반적으로 비교 대상 모델에 비해 특별하게 우수한 성능을 보이지는 않았으나, 가장 어려운 태스크인 Hit@1, MRR@1에서 추천 정확도가 보다 높은 것을 알 수 있었다.

### 5. 결론

본 연구에서는 셋톱박스 오디언스 세그멘테이션을 위해 세션 정보에 유저 정보를 결합한 개인화 추천 시스템을 제안하였다. 또한 유저의 개인적인 선호나 의도가 시청 이력에 반영될 것이라는 가정 하에 세션 데이터에 유저 정보를 결합하여 학습하는 모델을 개발한 결과, 연구의 주 데이터인 KOBACO에서 성능 향상이 있었음은 물론 개인화 추천 시스템에 널리 사용되는 Cosmetics, AppUsage 데이터로 검증했을 때에도 괄목할 만한 성능을 보였다.

본 연구의 실용적, 학술적 의의는 두 가지로 나눌 수 있다. 첫째, 본 연구는 우리가 알기로 실제 셋톱박스 데이터를 통해 셋톱박스 오디언스의 프로그램 세션에 추천 시스템을 적용한 최초의 연구이다. KOBACO 측과 협업하여 확보한 셋톱박스 데이터는 방송3사의 데이터를 모두 포괄하며, 셋톱박스 오디언스에 개인정보 활용 동의를 얻은 데이터로써 실용적, 학술적 가치가 충분한 데이터이다. 이 연구는 각 방송사에서 셋톱박스 오디언스 타겟팅에 사용되는 인공지능(Artificial Intelligence, AI) 모델을 고도화할 수 있는 발판이 될 수 있을 것이며, 유저-세션-아이템의 계층적 구조를 사용하는 다른 산업계에도 충분한 실용적 가치를 제시한다. 둘째, 본 연구는 세션 데이터와 유저 데이터의 유사성을 이용한 세션 기반

개인화 추천 시스템을 제시하였다. 기존 개인화 추천 시스템은 사용자가 여러 가지 세션을 가진 모델이 아니라, 유저의 시퀀스 하나를 이용한 모델이 주로 개발되었다. 하지만 유저 당 하나의 시퀀스가 있을 때에는 유저의 선호와 의도가 복잡하게 섞여 있을 수 있기 때문에 정확한 추천을 하기 힘들어질 수 있다. 반면 본 연구에서는 시퀀스를 세션으로 나눈 후 유저의 의도를 모델링하였고, 어텐션 구조를 통해 세션 벡터에 유저의 정보를 결합한 결과 비교 대상인 기반모델보다 유의미한 성능 향상을 보였다.

그러나 본 연구에서도 몇 가지 한계점들이 존재한다. 첫째로 본 연구에서는 KOBACO 데이터셋의 경우 원 데이터에서 임의로 세션을 하루 단위로 나누었다. 하지만 오디언스의 시청의도가 항상 하루 단위로 나누어지지 않을 수 있고, 우리가 단순히 연구의 편의를 위해 편성표 기준으로 나누었기 때문에 세션을 나누는 데에서 엄밀한 비교실험이 수행되지 못하였다. 추후 본 데이터로 세션을 나누는 기준을 달리하여 추가적인 연구가 이루어질 수 있다. 둘째로 모델에서 유저 임베딩이 적절히 학습되었는지는 알 수 없다. 오디언스 세그멘테이션을 위해서는 학습된 유저 임베딩을 가지고 분류를 수행해야 하지만, 우리가 가지고 있는 데이터의 셋톱박스 오디언스에 대해서 광고 효과가 측정된 적은 없다. 따라서 해당 추천 태스크의 유저 임베딩이 실제 광고 타겟팅을 위한 downstream task에 적절한 성능을 보이는지에 대해서는 추가적인 연구가 수행되어야 할 것이다. 셋째로 모델의 성능 면에서 성능 향상이 이루어질 수 있는 변수가 몇 가지 있으나, 본 연구에서는 사용되지 않았다. KOBACO 데이터셋으로 예시를 들면 오디언스가 프로그램을 본 시간(dwell time), 한 프로그램을 보다가 다

른 프로그램으로 돌릴 때까지 걸린 시간(interval time)등이 모델에 추가적으로 정보를 줄 수 있다. 따라서 원 데이터에서 사용 가능한 변수를 조금 더 고려하면 개인화 세션 기반 추천 태스크의 후속 연구에서 보다 높은 성능을 보이는 모델을 개발할 수 있을 것이다.

## 참고문헌(References)

### [국내 문헌]

- 나혜연, 남기환.(2020).사용자 선호도 변화에 따른 추천시스템의 다양성 적용. 지능정보연구, 26(4), 67-86.
- 장동수, 이청용, 김재경.(2023).딥러닝 기반 온라인 리뷰의 언어학적 특성을 활용한 추천 시스템 성능 향상에 관한 연구. 지능정보연구, 29(1), 41-63.
- 홍태호, 홍준우, 김은미, 김민수. (2022). 영화 리뷰의 상품 속성과 고객 속성을 통합한 지능형 추천시스템. 지능정보연구, 28(2), 1-18.

### [국외 문헌]

- Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555.
- Hamilton, W., Ying, Z., & Leskovec, J. (2017). Inductive representation learning on large graphs. Advances in neural information processing systems, 30.
- Hidasi, B., Karatzoglou, A., Baltrunas, L., & Tikk, D. (2015). Session-based recommendations with recurrent neural networks. arXiv preprint arXiv:1511.06939.
- Hornik, K., Stinchcombe, M., & White, H. (1989).

- Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5), 359-366.
- Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.
- Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, 42(8), 30-37.
- Li, J., Ren, P., Chen, Z., Ren, Z., Lian, T., & Ma, J. (2017, November). Neural attentive session-based recommendation. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (pp. 1419-1428).
- Liu, Q., Zeng, Y., Mokhosi, R., & Zhang, H. (2018, July). STAMP: short-term attention/memory priority model for session-based recommendation. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 1831-1839).
- Michael, K. (2020). *eCommerce Events History in Cosmetics Shop*. Released March 2020, Retrieved April 30, 2023, from <https://www.kaggle.com/datasets/mkechinov/ecommerce-events-history-in-cosmetics-shop>
- Sahoo, N., Singh, P. V., & Mukhopadhyay, T. (2012). A hidden Markov model for collaborative filtering. *MIS quarterly*, 1329-1356.
- Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001, April). Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web* (pp. 285-295).
- Shani, G., Brafman, R., & Heckerman, D. (2002, August). An MDP-based recommender system, Proc. 18th Conf. In *Uncertainty in Artificial Intelligence*.
- Shepard, C., Rahmati, A., Tossell, C., Zhong, L., & Kortum, P. (2011). LiveLab: measuring wireless networks and smartphone users in the field. *ACM SIGMETRICS Performance Evaluation Review*, 38(3), 15-20.
- Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2017). Graph attention networks. *arXiv preprint arXiv:1710.10903*.
- Wu, S., Tang, Y., Zhu, Y., Wang, L., Xie, X., & Tan, T. (2019, July). Session-based recommendation with graph neural networks. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, No. 01, pp. 346-353).

Abstract

## Personalized Session-based Recommendation for Set-Top Box Audience Targeting

Jisoo Cha\* · Koosup Jeong\* · Wooyoung Kim\* · Jaewon Yang\* · Sangduk Baek\* ·  
Wonjun Lee\* · Seoho Jang\*\* · Taejoon Park\*\*\* · Chanwoo Jeong\*\*\* · Wooju Kim\*\*\*\*

TV advertising with deep analysis of watching pattern of audiences is important to set-top box audience targeting. Applying session-based recommendation model(SBR) to internet commercial, or recommendation based on searching history of user showed its effectiveness in previous studies, but applying SBR to the TV advertising was difficult in South Korea due to data unavailabilities. Also, traditional SBR has limitations for dealing with user preferences, especially in data with user identification information. To tackle with these problems, we first obtain set-top box data from three major broadcasting companies in South Korea(SK, KT, LGU+) through collaboration with Korea Broadcast Advertising Corporation(KOBACO), and this data contains of watching sequence of 4,847 anonymized users for 6 month respectively. Second, we develop personalized session-based recommendation model to deal with hierarchical data of user-session-item. Experiments conducted on set-top box audience dataset and two other public dataset for validation. In result, our proposed model outperformed baseline model in some criteria.

**Key Words** : Session-Based Recommendation, Audience Targeting, Graph Neural Network, Recommendation System, Representation Learning

Received : May 17, 2023 Revised : June 22, 2023 Accepted : June 25, 2023

Corresponding Author : Wooju Kim

---

\* Dept. of Industrial Engineering, Yonsei University

\*\* Digital Analytics, Yonsei University

\*\*\* Korea Broadcast Advertising Corporation(KOBACO)

\*\*\*\* Corresponding Author: Wooju Kim

Department of Industrial Engineering, Yonsei University

50, Yonsei-ro, Seodaemun-gu, Seoul, Republic of Korea

Tel: +82-2-2123-7754, Fax: +82-2-2123-7754, E-mail: wkim@yonsei.ac.kr

## 저 자 소개



**차지수**

연세대학교에서 학사학위를 취득하였으며, 동 대학원 산업공학과에서 석사과정 재학 중이다. 주요 연구분야는 추천 시스템, 그래프 뉴럴 네트워크이다.



**정구섭**

동국대학교 산업시스템공학과에서 학사학위를 취득하였으며, 현재 연세대학교 산업공학과에서 석사과정을 진행 중이다. 연구관심분야는 추천 시스템이다.



**김우영**

연세대학교에서 신학과 컴퓨터과학 학부과정을 마치고 동대학원 산업공학과에서 석박사통합과정 재학 중이다. 현재 언어 인공지능과 그래프 형태로 표현된 데이터의 결합을 중점적으로 연구하고 있다.



**양재원**

고려대학교에서 바이오의공학부 학부과정을 마치고 연세대학원 산업공학과에서 석박사통합과정 재학 중이다. 현재 언어 인공지능, 그래프 인공지능, 딥러닝에 대해 연구하고 있다.



**백상덕**

연세대학교에서 경영학, 산업공학 학부과정을 마치고 동대학원 산업공학과에서 석사과정 재학 중이다. 현재 자연어처리, 시계열 예측 분야에 관심을 두고 연구하고 있다



**이원준**

인천대학교에서 학사학위를 취득하고 연세대학교 산업공학과에서 석사과정 재학중이다. 연구 관심분야는 자연어 처리, 추천 시스템이다.



**장서호**

연세대학교에서 천문학과, 수학과 학부과정을 마치고 동대학원 인공지능융합대학 DA대학원 석사과정에 재학중이다. 현재 Generative-AI, 그래프 인공지능을 연구중이며 주요 관심사는 Multimodal-AI이다.



**박태준**

서울대학교 문학사 취득 및 한국과학기술원(KAIST)에서 미래학 석사 학위 과정을 밟고 있으며 현재 한국방송광고진흥공사(KOBACO) 디지털전략팀 차장으로 재직 중이다. 주요 연구 관심분야는 타게팅 광고 기술, 추천시스템, AI기반 광고 등이다.



**정찬우**

가천대학교 공학사를 취득하였으며 현재 한국방송광고진흥공사(KOBACO) 디지털전략팀 팀원으로 재직 중이다. 주요 연구 관심분야는 타게팅 광고 기술, 추천시스템, AI기반 광고 등이다.



**김우주**

한국과학기술원(KAIST)에서 1994년 경영과학 박사 학위를 취득하였으며, 연세대학교 산업공학과에서 교수로 재직 중이다. 주요 연구관심분야는 텍스트, 그래프 등의 비정형 데이터에 대한 분석 및 정보 추출, 추천 시스템 등이다.