공학석사 학위논문

세션 기반 순환신경망 추천 모델의 성능 향상을 위한 데이터 증강 기법

A data augmentation method for session-based recommendation model

2022년 2월

서울과학기술대학교 일반대학원

데이터사이언스학과

이 경 찬

세션 기반 순환신경망 추천 모델의 성능 향상을 위한 데이터 증강 기법

A data augmentation method for session-based recommendation model

지도교수 김경옥

이 논문을 공학석사 학위논문으로 제출함

2022년 2월

서울과학기술대학교 일반대학원

데이터사이언스학과

이 경 찬

이경찬의 공학석사 학위논문을 인준함

2022년 2월

심사위원장 황상흠 (인)

심사위원 김경옥 (인)

심사위원 이영훈 (인)

요 약

제 목 : 세션 기반 순환신경망 추천 모델의 성능 향상을 위한 데이터 증강 기법

세션 기반 추천 모델(Session-based recommendation)은 세션 내에 포함된 아이템들의 연속적으로 나열된 정보를 이용하여 다음에 클릭할 아이템을 예측하는 추천시스템으로 최근 활발히 연구되고 있는 분야이다. 세션 기반 추천 모델은 유저의 시간적 선호도를 모델링하여 추천하는 순차적 추천시스템(Sequential recommendation) 방법의 한 종류로서, 이에 대한 선행 연구는 주로 연속적인 정보를 모델링하는데 적합한 순환신경망(Recurrent neural network, RNN) 구조를 채택해왔다.

세션 기반 추천시스템은 유저의 동적 선호도를 정확히 모델링하기 위하여 다양한 방향으로 발전되어왔으나, 명시적 피드백(Explicit feedback)(평점 등)이 없이 유저의 암시적 행동(Implicit behavior)(방문 등)을 사용하기 때문에 데이터가 희소하고 유저가 적은 상황에서 활용될 가능성이 높다. 만약 이처럼 데이터의 양이 적은 경우에 세션 기반 순환신경망 추천시스템을 사용한다면 적은 학습데이터에 과적합되어 일반화 성능이 저하되기 쉽다.

본 논문에서는 세션 데이터가 부족한 상황에서 순환신경망을 이용한 세션 기반 추천시스템의 성능을 향상시킬 수 있는 세션 데이터 증강 기법을 제안한다. 세션 내 아이템의 출현은 순차적 정보 외에도 아이템 사이의 관계 정보를 포함하고 있다. 제안 방법은 이러한 세션 데이터의 특성을 고려하고 유저 식별 정보 없이 세션 내 아이템의 시퀀스 정보만을 이용하여 관측 데이터와 유사한 데이터를 생성할 수 있다. 결과적으로 제안 방법을 통해 학습데이터의 양을 증가시킴으로써 세션 기반 추천 모델의 성능을 향상시킬 수 있다. 본 연구에서는 실제 데이터에 대한 실험을 통하여 세션 기반 추천시스템의 성능 향상을 입증하였다.

목 차

표 목 차

그림목차

Ⅰ. 서 론

기존의 협업필터링(Collaborative filtering) 방법 및 내용 기반(Content-based) 추천 시스템은 유저와 아이템 간의 과거 상호작용 정보를 이용하여 유저의 가장 과거부터 현재까지의 전체적인 선호도를 기반으로 추천하거나 개별 유저 및 아이템의 특징을 기반으로 유사도를 측정하여 추천한다. 이러한 방법은 긴 시간 동안 축적된 정적(static)인 선호도만을 반영하기 때문에 시간의 흐름에 따라 변화하는 동적(dynamic)인 선호도를 반영하지 못한다는 문제점을 안고 있다[1]. 유저의 가장 최근 선호도를 반영함으로써 추천 정확도를 향상시키기 위하여 순차적 추천시스템(Sequential recommendation system) 알고리즘이 개발되었다. 순차적 추천시스템이란 시간에 따라 변하는 유저의 선호도를 모델링함으로써 현재 시점에 가장 흥미 있는 아이템을 예측하는 방법이다. 순차적 추천이라는 용어는 유저의 시간적 선호도를 모델링하여 추천하는 방법론을 넓은 범위에서 일컫는 용어이며, 대부분의 순차적 추천시스템에 대한 연구는 유저 및 아이템 상호작용 데이터에서 시간 정보를 추가하는 방식으로 진행된다.

세션 기반 추천시스템(Session-based recommendation system, SBR)이란 순차적 추천시스템의 한 종류로써, 소비 과정 동안 생성된 세션으로부터 유저의 선호도를 학습하여 다음에 올 상호작용(interaction) 또는 부분적인 세션을 예측하는 추천 알고리즘이다. 세션(session)이란 방문자가 웹사이트에 접속했을 때 서버에 저장되는 방문자에 대한 정보이며, 각 세션은 연속적으로 발생한 여러 개의 유저-아이템 상호작용으로 구성되어 있다. 일반적으로, 세션은 구매된 상품들의 장바구니 또는 클릭 정보처럼 몇 분 ~ 몇 시간 동안 지속된다[2]. 대부분의 세션 기반 추천시스템 분야에서 상호작용은 아이템 자체만을 의미하며 유저 및 아이템의 개별적인 식별 정보 또는 특성 정보 없이 오직 세션으로만 다음 상호작용을 예측하는 알고리즘이 다수 연구되어 왔다. 세션 기반 추천시스템은 신규 방문자가 첫 상호작용을 발생시키는 순간 추천시스템이 동작하는 시나리오를 가정하며, 과거에 생성되어 사용 가능한 세션들과 현재 세션 안에 있는 상호작용들의 순차적 정보를 활용하여 추천하기 때문에 상호작용의 시퀀스(sequence)를 모델링하는 것이 핵심이다. 시퀀스를 모델링하기 위하여 마르코프 체인(Markov Chain) 모델을 비롯한 다양한 방법론들이 제안되었고 그 중 순환신경망을 기반으로 한 방법론들이 다수를 차지하고 있다.

순환신경망을 포함한 딥러닝 기반의 모델들은 매개변수의 수를 늘리는 방향으로 발전하여 높은 성능을 달성해왔기 때문에 데이터가 부족한 상황에서는 파라미터가 관측된 데이터에만 과적합(overfitting)되어 일반화 성능이 저하되기 쉽다. 그래서 데이터 수집 시에 비용이 많이 들거나 정답 라벨(ground truth)을 수집하기 어려운 경우(ex. 의료영상 데이터), 관측 데이터에 머신러닝 기법 등을 적용하여 이와 비슷한 새로운 데이터를 생성함으로써 데이터의 양을 증가시키기도 한다. 이를 데이터 증강(Data augmentation) 기법이라고 하며 이를 잘 활용하면 모델의 과적합을 막고 일반화 성능을 향상시킬 수 있다. 데이터 증강은 주로 컴퓨터 비전 분야에서 사용되며 flipping, cropping, rotation 등 이미지 데이터의 증강을 위한 다양한 기법들이 제안되어왔다.

시퀀스 데이터(Sequence data)에 대한 데이터 증강 방법은 주로 자연어 처리 분야에서 제안되어 왔으며, 문서 분류 등의 문제에서 정답 라벨이 부족한 경우 딥러닝 성능을 향상시키기 위한 다양한 데이터 증강 방법들이 존재한다. 주로 word embedding을 이용하여 코사인 유사도(cosine similarity)를 기준으로 높은 유사도를 갖는 단어로 대체하는 방법이 사용된다. 순차적 추천시스템에서도 데이터 증강 방법이 연구되어 왔다. [3]은 하나의 세션을 다섯 가지의 증강 방법(Random 방식 3개, Informative 방식 2개) 중 두 가지의 방법을 임의로 채택하여 한 세션 당 두 가지의 증강을 생성하여 대조 학습에 활용하였다. [4]는 SBR의 성능을 높일 수 있는 방법 중 하나로 데이터 증강 기법을 제안하여 SBR 성능을 향상시켰다.

그러나 세션 기반 추천시스템은 명시적 피드백(Explicit feedback)(평점 등) 데이터 대신 유저의 암시적 행동(Implicit behavior) 데이터를 이용한 추천시스템이므로 데이터가 희소하고 유저가 적은 상황에서 활용될 가능성이 높다. 예를 들어 신규 서비스를 시작한 지 얼마 안 된 초창기의 온라인 쇼핑몰 사이트는 아이템 수가 많을지라도 방문자 수가 적고 가입자는 더욱 적을 것이다. 현실적으로도 많은 경우에는 로그인을 하지 않았거나 유저 식별 정보(나이, 성별 등)가 밝혀지지 않은 유저가 세션 데이터를 생성한다. 그러므로 대규모 회원을 보유한 경우가 아니라면 세션 기반 추천시스템이 활용될 확률이 높고 데이터의 양이 적은 경우에는 성능이 낮아지게 된다.

이렇듯 데이터의 양이 적어 과적합이 우려되는 경우에는 데이터 증강 기법을 사용함으로써 추천 성능을 향상시킬 수 있다. 그러나 기존의 시퀀스 데이터에

대한 증강 기법은 주로 word embedding을 기준으로 하여 유사한 요소로 대체하는 등의 방식을 취한다. 그러나, 세션 데이터는 같은 시퀀스 데이터일지라도 자연어 데이터와는 출현 양상과 생성 환경이 다르기 때문에 세션 데이터에 기존 방식을 그대로 적용하기에는 무리가 있다. 세션 데이터는 순차적 정보 외에도 같은 아이템의 반복적 출현, 소수 인기 아이템의 높은 출현 빈도, 짧은 길이 등의 특성을 갖고 있기 때문에 시퀀스 데이터 증강 기법 중 널리 알려진 자연어 데이터 증강 기법을 그대로 적용하면 세션 데이터의 특성이 반영되지 않아 증강에 의한 효과적인 성능 향상을 기대할 수 없다. 따라서 세션 데이터에 특화된 증강 기법이 필요하지만 세션 기반 추천시스템을 위한 데이터 증강 기법에 대한 연구는 거의 없다.

[3]에서 제안된 random 방식은 무작위성이 존재하기 때문에 길이가 짧은 세션에 적용될 경우 원본 데이터의 분포를 변형시킬 가능성이 높고, informative 방식은 유저가 식별된 데이터에 대해서만 아이템 유사도를 계산할 수 있으므로 세션 데이터에 적용할 수 없다는 한계가 있다. 또한 [4]의 dropout 방식 역시 길이가 짧은 세션에 적용할 경우 적게 등장하는 아이템 자체를 누락할 수도 있다는 한계가 존재한다.

본 연구에서는 기존에 연구된 자연어처리 분야와 순차적 추천시스템 분야에서 사용된 증강 방법에서 착안하여 세션 데이터의 특성을 고려한 데이터 증강 방법을 제안한다. 이 방법은 유저 식별 정보 없이 아이템 시퀀스만을 이용하여 아이템간의 유사도를 구하고 세션의 정보를 최대한으로 보존하는 대체/삽입을 통해 증강데이터를 생성한다. 본 연구에서 제안하는 방법을 이용하여 원본 데이터와 유사한 데이터를 신규로 생성함으로써 학습데이터의 양을 증가시키고, 결과적으로 세션 기반 추천 시스템의 성능을 향상시킬 수 있다.

본 논문은 서론을 포함하여 총 5장으로 구성된다. 2장에서는 순차적 추천시스템 및 데이터 증강에 대한 연구들과 함께 세션 기반 추천시스템에서의 데이터 증강 방법에 대한 연구들에 대해 소개한다. 3장에서는 본 연구에서 제안하는 데이터 증강 기법에 대해 자세히 설명한다. 4장에서는 실험에 사용된 데이터 소개, 효과를 검증할 기존 SBR 알고리즘, 평가지표, 실험 설계에 대하여 소개하고 결과를 고찰한다. 마지막으로 5장에서는 연구에 대한 전체적인 요약 및 한계점과 추후 연구에 대하여 기술한다.

Ⅱ. 관련 연구

추천시스템이란 유저들에게 그들이 흥미 있어 할만한 아이템을 제안하는 지능형 시스템이다. 추천시스템의 목적은 아직 평가되지 않은 아이템 또는 유저에 대하여 평점을 예측하여 제공하거나 높은 확률로 예측된 아이템 중 상위 N개의 아이템을 유저에게 제공하는 것이다[5]. 추천시스템은 더욱 정확한 예측을 위하여 다양한 형태로 발전해왔으며 추천 성능 향상을 위한 데이터 증강 기법 역시 다양한 방향으로 연구되어왔다.

본 장에서는 순차적 추천시스템, 세션 기반 추천시스템, 그리고 이러한 추천시스템 분야에서 사용되는 데이터 증강 기법에 관련된 기존의 연구들에 대하여 소개한다.

1. 순차적 추천시스템

추천시스템 알고리즘 중에서 협업필터링 방법, 내용 기반 추천시스템 등이 주요 흐름으로 연구되어왔고 추천시스템을 대표하는 알고리즘의 자리를 차지하고 있다. 두 방법론은 비교적 간단한 아이디어로 높은 성능을 달성하였지만, 유저와 아이템 사이의 정적인 상관관계를 분석하기 때문에 오직 유저의 장기간 선호도밖에 반영할 수 없다는 한계가 있다[6].

시간에 따라 변화하는 선호도를 반영하기 위하여 마르코프 체인, 신경망 모델 등을 활용한 순차적 추천시스템 연구들이 진행되었다[7][8][9][10]. 순차적 추천시스템은 연속적인 구매와 같이 유저와 아이템간의 상호작용이 순서 정보를 갖고 있을 때 순차적 정보를 모델링하여 다음에 구매 또는 선택(상호작용)할 아이템을 예측하는 것을 목표로 한다. 초기의 순차적 추천시스템은 마르코프 체인[11]이나 세션 기반 KNN[12] 등의 머신러닝 기법을 사용하여 유저 행동의 시간적 패턴에 대한 모델링을 시도하였다. 최근에는 자연어 처리 분야에서 높은 성능을 달성한 딥러닝 기반의 순환신경망 모델을 순차적 추천시스템에 적용한 연구가 다수 제안되었으며 높은 추천 성능을 보여주고 있다[13].

2. 세션 기반 추천시스템

세션 기반 추천시스템은 순차적 추천시스템과 매우 연관되어있다. 세션 기반

추천시스템은 순차적 추천시스템의 하위 종류로 볼 수 있으며 순차적 추천시스템이라는 용어가 더욱 포괄적인 영역을 지칭한다[13].

순차적 추천시스템이 명확한 순서로 정렬된 과거 상호작용 리스트를 사용한다면 세션 기반 추천시스템은 더욱 동시 발생적으로 생성된 세션 데이터에 특화된 추천시스템이다. 세션이란 다수의 유저-아이템 상호작용으로 구성된 데이터를 의미하며, 예를 들면 한 번의 방문 안에서 발생한 상품들의 장바구니 같이 일반적으로 몇 분~몇 시간 동안 기록된다[2]. 시퀀스 데이터가 명시적, 내재적 상호작용 정보와 함께 시간 정보를 함께 사용한다면 세션 기반 추천시스템은 유저가 명시되지 않은 내재적 유저 행동만을 사용한다는 점에서 순차적 추천시스템 관련 연구들에서 사용한 데이터보다 더욱 한정된 정보의 데이터를 이용한다고 할 수 있다.

아이템 시퀀스를 고차원 공간에서 모델링하는 순환신경망 방법이 세션 기반 추천시스템에 적용되어 큰 성능 향상을 이루었다. 순환신경망을 이용한 세션 기반 추천 알고리즘은 세션 내의 아이템 시퀀스를 이용하여 해당 세션을 임베딩하고, 임베딩 벡터로 전체 아이템 집합에 대해서 다음에 위치할 아이템으로서의 확률을 계산한다. [14]는 현실 세계에서 오직 짧은 session-based 데이터(예를 들면 작은 스포츠웨어 웹사이트)에만 기반해서 추천해야하는 문제를 해결하기 위하여 RNN을 추천시스템에 도입한 GRU4REC을 제안하였다. GRU4REC은 각 세션의 마지막 아이템을 타겟 아이템으로 설정하고 GRU를 이용하여 세션의 시퀀스를 모델링한 뒤 추천 후보군 아이템에 대하여 타겟 아이템을 예측하는 방식으로 동작한다. 이에 대한 후속 연구로서 세션에서의 연속적인 정보뿐만이 아닌 세션 내에서의 주요 목적까지 모델링 하기 위하여 순환신경망과 어텐션(Attention) 메커니즘을 결합한 연구가 제안되었다[15]. 최근에는 현재 세션뿐만 아니라 유사한 세션이나 다른 세션들 내의 시퀀스까지 고려하여 아이템 시퀀스의 전역적인 정보를 활용하는 연구들이 제안되었다. 대표적으로 현재 세션의 정보와 비슷한 세션(이웃 세션)들의 정보를 함께 결합하여 예측하는 협업식(collaborative) 방법[16][17]과 아이템 사이의 전이(transition) 정보를 그래프 기반 신경망(Graph neural network)로 임베딩하여 활용하는 방법[18][19][20]이 있다.

[18]은 이전 연구들이 세션의 표현(representation)에 집중했던 것과는 달리 아이템 사이의 복잡한 전이로부터 추출된 아이템 표현에 집중하였다. 이 방법

은 미니배치에 속한 세션들의 아이템으로 세션 그래프를 구축하여 아이템 임베딩을 학습하여 이를 세션 임베딩에 구축에 활용하였다. 그래프 신경망 도입으로 큰 성능 향상이 있었고 이후 그래프 신경망을 기반으로 한 세션 기반 추천시스템이 다수 연구되었다[19][20][21]. [19]는 세션 그래프 구축 후 어텐션 모듈을 통해 과거 유저 행동들과 타겟 아이템에 대한 관계를 모델링함으로써 아이템 전이 정보와 함께 타겟 아이템에 대한 유저의 흥미도 함께 활용되었다. [20]은 세션 그래프를 구축하기 위하여 랜덤 미니배치가 아닌 이웃 세션을 이용하였다. 개별 아이템에 대하여 세션 내부에서의 다른 아이템과의 관계와 다른 세션 안에 있는 아이템 간의 관계를 결합하여 세션 표현을 구축하였다. [21]은 세션 그래프를 통해 아이템 전이 정보를 임베딩한 후 최근 아이템에 대한 흥미와 시간에 따라 변하는 다양한 선호도를 모델링하는 2단계 구조를 통해 성능을 향상시켰다.

3. 시퀀스 데이터 증강

대량의 데이터셋을 이용할 수 있는 상황은 실제 상황에서 많지 않다. 그리고 딥러닝 기반의 모델을 작은 데이터셋으로 학습시킬 경우 과적합되기 쉬워지고 이로 인해 관측되지 않은 데이터에 대한 일반화 성능이 저하된다[22]. 이러한 상황에서 해결책으로 사용할 수 있는 방법 중 하나로 데이터 증강((Data augmentation)이 있다. 데이터 증강 방법을 사용하면 관측된 데이터를 기반으로 인공적으로 데이터셋을 증폭시켜서 모델의 일반화 성능을 최적화할 수 있다.

데이터 증강 기법은 컴퓨터 비전 분야에서 활발히 연구되어왔지만, 시퀀스 데이터에 대한 연구 또한 다수 존재한다. 그 중 자연어 데이터의 증강에 대한 기법이 대부분을 차지하고 있다[3][23][24][25][26]. [3]은 문장을 대조 학습시키기 위한 증강 방법으로 단어 삭제, 묶음 삭제, 재배치, 동의어 대체 등 4가지 방법을 제시하였다. [23]은 Word2vec 모델을 이용해 학습 데이터 문장에 있는 모든 명사를 동의어로 치환함으로써 데이터를 증강하였으며 [24]는 다중 클래스 감성분석을 위한 라벨링 된 이메일 텍스트 데이터의 수집이 어렵다는 점을 해결하기 위하여 k-NN(Nearest Neighborhood) 알고리즘과 동의어 사전을 이용하여 쓰임 양상이 유사하고 품사가 동일한 단어로 대체함으로써 데이터를 증강하였다. 또한 [25]는 트위터 내 혐오 발언 텍스트 데이터 자체가 부족하여 야기되는 다수 범주 과적합 문제를 해결하기 위하여 워드 임베딩(word

embedding)이 유사한 단어로 대체하여 소수 범주의 데이터를 증강하였다. 또한 [26]은 주어진 문장에서 여러 개의 단어를 임의로 선택하여 동의어 대체, 무작위 삽입, 무작위 뒤바꿈, 무작위 삭제하는 변형을 통해 증강하는 방법을 제안하였다. 자연어 데이터를 증강하는 기법은 대부분 단어 사이의 유사도와 품사를 기반으로 동의어로 대체하거나 순서를 바꾸는 등 방법론 자체는 큰 차이가 존재하지 않는다.

자연어 데이터 증강 기법과 유사한 방식으로, 순차적 추천시스템의 성능을 향상시키기 위한 아이템 시퀀스 데이터 증강 기법이 연구되었다[4][27][28]. [4]는 세션 기반 추천시스템의 성능 향상을 위하여 세션을 마지막 아이템부터 하나씩 제외하여 나머지 세션을 사용하는 방법과 세션 내 임의로 선택한 아이템을 지워서 새로운 데이터로 사용하는 방법을 제안하였다. [27]은 트랜스포머를 시퀀스의 역방향으로 사전 학습시킨 후 시퀀스 맨 앞자리에 트랜스포머로 예측된 아이템을 추가함으로써 데이터를 증강하였다. [28]은 대조 학습을 이용한 순차적 추천시스템과 random방식과 informative방식의 데이터 증강 방법을 제안하였다. Random 방식으로는 임의로 선택된 아이템을 마스킹(mask)하거나, 더 짧은 길이로 잘라내거나(crop), 임의로 선택된 두 아이템의 순서를 뒤바꾸어 재배치(reorder)하는 방식을 제안하였다. Informative 방식으로는 유저-아이템 상호작용을 기반으로 아이템 사이에 유사도를 계산하여 가장 유사한 아이템으로 변형하는 substitute, insert 방식을 제안하였다.

시퀀스 중 일부를 변형하여 새로운 데이터를 생성하는 연구들의 대부분은 아이템 선택 시 시퀀스 내에서 무작위로 선택하는 방식을 사용한다[4][26][28]. 그러나 무작위 선택에 의한 증강이 모델의 성능을 향상시킬 수 있을지라도, 성능 향상에 도움 되지 않는 아이템까지 변형에 참여하였기 때문에 증강에 의한 성능 향상이 최대로 이루어졌다고 할 수 없다. 또한 자연어 데이터와 달리 세션 데이터는 길이가 대부분 짧기 때문에 [26]과 [28]의 무작위 재배치, 무작위 삭제와 같은 방식은 순차적 정보를 파괴하거나 더욱 짧은 세션을 생성하여 순환신경망 학습에 도움 되지 않는 증강을 포함할 수도 있다. 그리고 [28]의 substitute, insert 방식은 유저-아이템 상호작용 데이터로부터 아이템 간의 유사도를 계산하는 방식이기 때문에 유저 식별 정보를 사용하지 않는 세션 기반 순환신경망을 위한 데이터 증강 방식으로 적합하지 않다. 아이템 시퀀스 정보만으로 아이템 사이의 유사도를 계산하여 데이터를 증강하는 세션 데이터 증강 기법은 아직 연구되지 않았다.

본 논문에서는 아이템 시퀀스를 기반으로 아이템 간의 유사도를 측정하고 세션 정보의 손실을 최소화하는 변형 방식 통해 세션 데이터를 증강하는 기법을 제안한다. 그리고 전체 아이템 중에서 증강에 도움이 되는 아이템들과 그렇지 않은 아이템들의 특성을 분석하기 위하여 개별 아이템 특성에 따라 아이템들을 분류하여 증강하고 비교함으로써 위에서 언급한 무작위성 선택 방식보다 아이템 특성 정보에 기반한 선택 방식이 더욱 효과적임을 보인다.

Ⅲ. 연구 방법

세션 데이터를 증강하기 위해서는 기존의 세션과 유사한 세션을 생성해야 한다. 즉, 기존의 세션들로부터 알 수 있는 정보를 기반으로 새로운 세션을 생성하되, 새로 생성된 세션은 실제로 관측될만한 데이터여야 한다. 이러한 기존 데이터의 분포를 벗어나지 않고 원본 데이터와 유사한 세션은 원본 세션 안의 아이템 일부를 변형함으로써 얻어질 수 있다. 기존의 세션을 변형하기 위해서는 우선 변형할 아이템의 선택 기준 및 변형 방식을 정의해야 한다. 본 장에서는 변형할 아이템의 선택 기준과 변형 방식에 대하여 설명한다. 본 논문에서는 개별 아이템을 v로, 모든 개별 아이템의 집합을 V라고 표기한다.

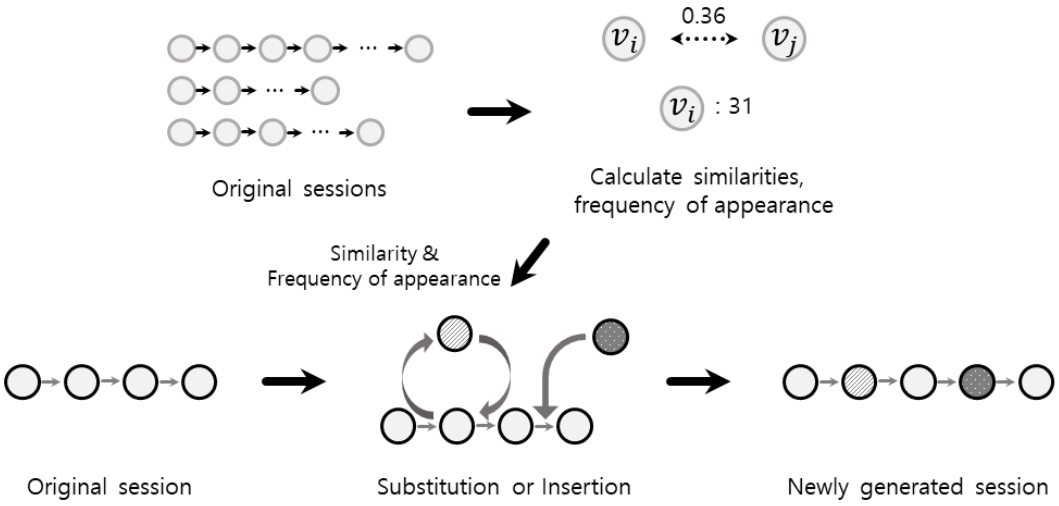

Fig. 3.1 원본 세션으로부터 새로운 세션을 생성하는 데이터 증강 과정

본 논문에서 제안하는 세션 데이터 증강 과정은 다음과 같다. 우선 모든 아이템의 출현 수에 대한 변형 아이템의 수의 비율인 p를 설정한다. 개별 아이템에 대하여 0~1의 난수를 생성하여 p보다 크고 선택 조건에 만족한다면 변형을 취하고, p보다 작으면 그대로 둔다. 아이템 변형으로 인하여 원본 세션과 달라진 세션이 생긴다면 이를 새로운 증강 데이터로써 사용한다. 즉, 원본 데이터와 증강 세션 모두 모델에 학습시킨다.

2장에서 살펴본 기존 시퀀스 데이터 증강 방법이 가진 무작위성은 개별 아이템의 특성을 고려하지 않기 때문에 무작위로 선택한다면 출현 빈도가 높은 아이템들 위주로 선택될 확률이 높다. 따라서 본 연구에서는 세션 데이터에서

추출할 수 있는 가장 큰 특성인 아이템 사이의 유사도 값과 출현 빈도를 선택 조건으로 사용하였다. 증강 전 원본 데이터셋으로부터 모든 아이템 사이의 유사도를 측정하고 개별 아이템마다 출현 빈도를 센다. 그 이후 아이템 특성 값에 기반한 선택 조건에 따라 아이템을 선택하여 해당 아이템을 가장 유사한 아이템으로 대체하거나 자기 자신의 위치 바로 앞에 삽입하는 방식을 취하였다. 전체적인 세션 데이터 증강 과정은 Fig. 3.1과 같다.

1. 아이템간의 유사도 측정

아이템을 가장 비슷한 아이템으로 대체 또는 삽입하기 위해서는 아이템 사이의 유사도를 계산해야 한다. 시퀀스 데이터에서는 요소들의 나열 순서와 근접한 위치에 출현한 빈도가 유사도를 측정하는데 중요한 정보로 활용된다. 아이템 시퀀스에서는 개별 아이템 입장에서 자기 자신과 가장 유사한 아이템은 유사도 측정 지표에 따라서 달라질 수 있다. 따라서 본 논문에서는 단어, 벡터, 집합 등의 요소들 사이의 유사도를 측정하는 데 사용되는 여러 가지 유사도 측정 지표를 적용하여 아이템 간의 유사도를 측정하고 비교하였다.

1) 동시 발생 빈도(Frequency of co-occurrence)

동시 발생 빈도란 서로 다른 두 요소가 일정 거리 내에 함께 출현한 빈도이다. 동시 발생 빈도가 높을수록 근접한 위치에 출현한 경우가 많다는 뜻이므로, 서로의 유사도가 높아지도록 측정하는 방식이다. 세션 데이터 내의 아이템에 대해서도 동시 발생 빈도를 이용하여 유사도 측정이 가능하다. 특정 관심 분야를 가진 유저가 하나의 세션을 생성했다고 가정할 때, 해당 세션 내에 위치한 아이템끼리는 동일한 카테고리에 속할 것이며 가까이 위치할수록 유사한 아이템이라고 볼 수 있다. 또한 하나의 세션 안에 있다고 무조건 유사하다고 할 수는 없다. 왜냐하면 세션의 길이가 길 때는 멀리 떨어져 있는 아이템끼리는 유사하지 않을 수 있기 때문이다. 아이템 집합 V내의 모든 아이템 쌍에 대하여 동시 발생 빈도를 계산할 때 동시 발생 행렬을 이용하면 직관적으로 이해하기 수월해질 수 있다. 본 논문에서는 일정 크기의 window 내에 함께 등장하는 아이템끼리 유사도를 높게 측정하는 window based co-occurrence matrix[29] 방식을 사용하였다. 해당 논문에서 제안하는 동시 발생 빈도 기반의 유사도 측정 방식은 다음과 같다:

(1) $|V| \times |V|$ 크기의 동시 발생 행렬 C를 만든다.

(2) window 크기 k를 정한다.
(3) 아이템 v_i 기준 좌우로 거리 window 내에 출현한 아이템들의 집합 W_i를 구한다.
(4) $C[v_i, v_j]$, $v_j \in W_i$를 1만큼 증가시킨다.
(5) 모든 아이템에 대하여 3번과 4번을 반복한다.
(6) 아이템 v_i와 가장 유사한 아이템은 $C[v_i]$에서 가장 높은 값의 인덱스이다.

동시 발생 행렬의 생성 과정을 그림으로 나타내면 Fig. 3.2와 같다. Fig. 3.2에서 예시로 주어진 세션을 이용해 동시 발생 빈도를 측정할 경우, 아이템 v_4와 가장 유사한 아이템은 v_1이 된다. 즉, 물리적으로 가까운 위치에서 빈번하게 등장할수록 가장 유사해진다.

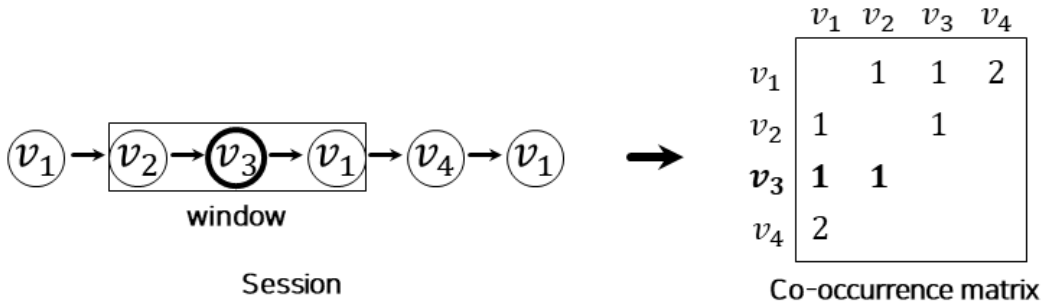

Fig. 3.2 하나의 세션으로부터 동시 발생 행렬 생성 예시

하나의 세션 내에 등장하는 모든 아이템들을 세지 않고 window를 기반으로 하여 측정하는 이유는 길이가 긴 세션 안에서 멀리 있는 아이템 사이의 유사도가 높게 나오는 것을 방지하기 위해서이다. 그러므로 window의 크기 k는 사용자가 세션 길이의 분포를 고려하여 실험에 따라 적절하게 설정할 수 있다.

2) PMI(Point-wise mutual information)

동시 발생 빈도는 주변에 많이 출현한 아이템을 가장 유사하다고 판단하는 직관적인 지표이다. 그러나 개별 아이템의 출현 빈도는 고려하지 않는다는 단점이 있다. 만약 아이템 v_i의 출현 빈도가 매우 높다면 동시 발생 빈도를 기반으로 유사도를 계산할 시 다른 아이템들의 입장에서는 주변에 단순히 많이 등장한다는 이유로 v_i가 가장 유사한 아이템으로 선택된다는 문제가 발생한다 (Fig. 3.3).

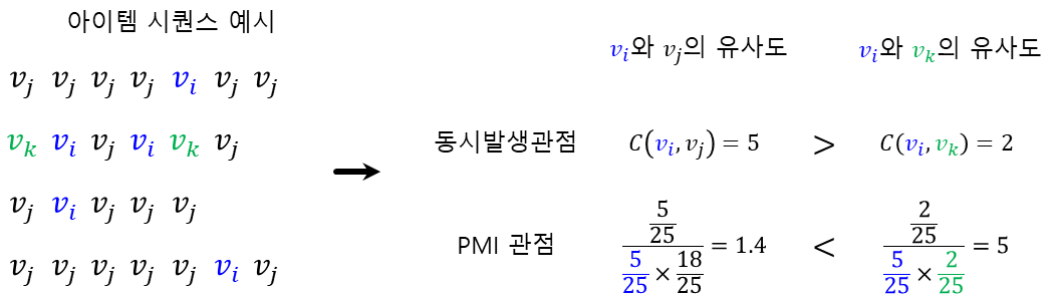

Fig. 3.3 동일한 시퀀스에 대한 동시 발생 빈도와 PMI의 차이 (window 크기 1)

PMI 지표는 이러한 문제점을 고려하기 위하여 동시 발생 빈도 값이 아닌 두 아이템의 분포를 이용한다. 정확히는 서로 다른 두 아이템의 결합 분포를 개별 아이템의 확률 곱으로 나눈 후 log를 취한다. 이를 통해 단순 동시 발생 빈도가 아닌 출현 빈도에 대한 동시 발생 수의 비율이 되므로, 한 아이템의 입장에서는 무조건 동시 발생 빈도가 높은 아이템이 아니라 적게 출현했더라도 본인 출현 빈도 중 대부분을 자신과 동시 발생한 아이템과의 PMI가 높아진다. 아이템 v_i와 v_j 사이의 PMI 값은 다음과 같이 계산할 수 있다:

$$PMI(v_i, v_j) = \log_2 \frac{P(v_i, v_j)}{P(v_i)P(v_j)}$$

$$= \log_2 \frac{\frac{C(v_i, v_j)}{N}}{\frac{C(v_i)}{N}\frac{C(v_j)}{N}} \quad \text{[Equation 3.1]}$$

$$= \log_2 \frac{C(v_i, v_j) \times N}{C(v_i)C(v_j)}$$

$$where \quad C(v_i, v_j) = co-occurrence\ matrix[v_i, v_j],$$
$$C(v_i) = \sum_{v_x \in V} co-occurrence\ matrix[v_x, i]$$

N은 데이터 내 모든 아이템의 출현 빈도 합을 의미한다. $P(v_i)$는 N에 대한 아이템 v_i의 출현 빈도의 비율이며 $P(v_i, v_j)$는 N에 대한 아이템 v_i와 아이템 v_j의 동시 발생 빈도의 비율이다. Fig. 3.3에서 아이템 v_i는 동시 발생 빈도 관점에서는 v_j와 더 유사하지만, PMI 관점에서는 v_k와 더 유사하다.

3) Jaccard index

자카드 계수(Jaccard index)는 집합의 유사도를 측정하는 지표 중 하나로, 두 집합의 교집합의 크기를 합집합의 크기로 나눈 값으로 정의되며 0에서 1 사이의 값을 갖는다[30]. 본 논문에서 아이템 간의 유사도를 계산하기 위하여 [31]에서의 동시 인용(co-citation) 지표를 활용한 문서 간 유사도 측정 방식을 사용하였다. [31]에서 사용하는 하나의 문서라는 개념을 본 논문에서는 하나의 window에 대응시키고, 해당 문서가 인용한 문서들의 집합을 window 내 아이템들의 집합에 대응시켰다. X_i를 아이템 v_i가 포함된 window들의 집합이라고 할 때, 아이템 v_i와 아이템 v_j 사이의 자카드 계수 $J(v_i,v_j)$는 다음과 같이 정의된다:

$$J(v_i,v_j)=\frac{|X_i\cap X_j|}{|X_i|+|X_j|-|X_i\cap X_j|} \quad \text{[Equation 3.2]}$$

위와 같은 정의에 의하여 아이템 v_i와 아이템 v_j는 함께 출현한 window의 수가 개별 아이템이 출현한 window의 수에서 차지하는 비중이 많을수록 유사하다고 계산된다. 즉, window의 교집합의 수를 합집합의 수로 나눈 값이 된다. 세션이 너무 긴 경우에는 먼 거리의 아이템끼리는 유사도가 낮을 것이기 때문에 동시 발생 빈도 계산 시와 마찬가지로 사용자가 데이터에 적합한 window의 크기를 설정하여야 한다.

4) Tanimoto index

타니모토 계수(Tanimoto index)는 자카드 계수로부터 파생한 유사도 측정 지표로서, Tanimoto(1957)에 의하여 nonbinary 경우에 대하여 더욱 정교해진 지표이다. 본 논문에서는 [31]에서 소개된 동시 인용 지표를 활용한 타니모토 계수 측정 방식을 사용하였다. C^{win}을 window 기반의 동시 발생 행렬이라고 정의할 때, C^{win}와 동시 발생 행렬 C는 차이점이 존재한다. 동시 발생 행렬은 개별 아이템을 중심으로 window 안에 등장한 아이템을 세는 반면, window 기반의 동시 발생 행렬 C^{win}은 중심 아이템이라는 설정 없이 각 window마다 함께 등장한 모든 쌍을 센다는 점에서 다르다. C^{win}의 주 대각선이 해당 아이템이 등장한 window의 개수라고 할 때, C^{win}으로부터 계산할 수 있는 아이템 간의 타니모토 계수는 다음과 같이 정의 된다:

$$T(v_i, v_j) = \frac{C^{win}[v_i, v_j]}{\sum_{v_x \in V - \{v_i\}} C^{win}[v_i, v_x] + \sum_{v_x \in V - \{v_j\}} C^{win}[v_j, v_x] - C^{win}[v_i, v_j]}$$

[Equation 3.3]

위 식에 의하면 타니모토 계수는 두 아이템이 동시 출현한 window의 수를 각각의 아이템이 다른 모든 아이템들과 동시 출현한 window의 합집합 수로 나눈 값이다. 즉, 아이템 v_j의 입장에서는 아이템 v_i가 다른 모든 아이템들과의 동시 출현 중에서 자기 자신과 많이 동시 출현할수록 유사해진다.

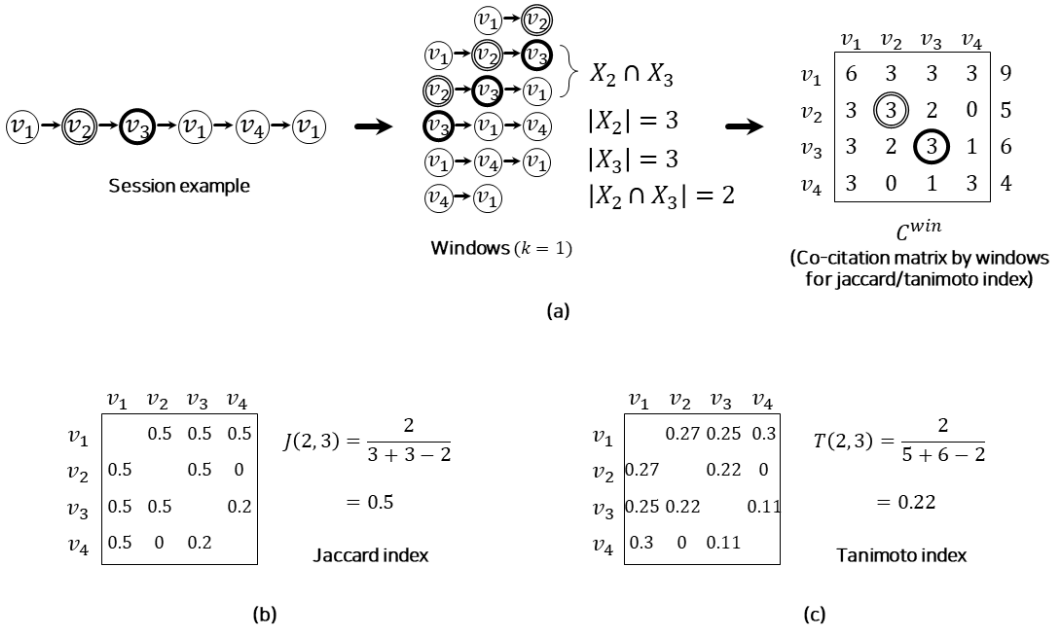

Fig. 3.4 하나의 세션으로부터 C^{win}을 생성한 후 자카드 인덱스와 타니모토 인덱스 계산하는 예시(v_2와 v_3을 중심으로). (a) 예시 세션으로부터 C^{win}을 생성하는 과정. (b) C^{win}으로부터 자카드 인덱스 계산 예시와 (c) 타니모토 인덱스 계산 예시.

자카드 계수와 타니모토 계수는 모두 C^{win}로부터 계산될 수 있다(Fig. 3.4). 그림 Fig. 3.4(a)처럼 세션으로부터 C^{win}을 생성한 후, 주 대각선 $C^{win}[v_i, v_i]$를 아이템 v_i가 등장한 window의 수라고 할 때, 자카드 계수는 개별 아이템이 등장한 window의 합집합의 수를 사용하는 반면 타니모토 계수는 개별 아이템이 다른 모든 아이템들과 동시 발생한 window의 합집합의 수를 사용한다는 면에서 두 지표가 다르다.

5) Cosine 유사도

Cosine 유사도는 두 벡터 사이의 각도에 따라 0°일 때 1, 180°일 때 -1을 부여하는 방식으로 -1에서 1 사이의 값으로 나타내는 벡터 사이의 유사도 지표 중 하나이다. 두 벡터 $\vec{a}$, $\vec{b}$의 Cosine 유사도 s_{ij}는 다음과 같이 정의된다:

$$cosine\ similarity(\vec{a},\vec{b}) = \frac{\vec{a} \cdot \vec{b}}{|a||b|} \quad \text{[Equation 3.4]}$$

본 논문에서는 [32]에서 비교한 유사도 측정 지표인 동시 발생 기반의 Cosine 유사도를 활용하였다. 만약 아이템 v_i가 등장하는 window의 수를 n_i라고 하고, v_i와 v_j를 모두 포함하는 window의 수를 n_{ij}라고 하면, 두 아이템 사이의 동시 발생 기반 코사인 유사도 s_{ij}는 다음과 같이 계산될 수 있다.

$$s_{ij} = \frac{n_{ij}}{\sqrt{n_i n_j}} \quad \text{[Equation 3.5]}$$

이 방식은 PMI와 계산 방식이 유사하지만 0에서 1 사이의 값을 가진다는 차이점이 있다. 동시 발생한 window의 수와 각 아이템이 출현한 window의 수를 이용하여 유사도를 계산한다는 점에서 자카드 인덱스, 타니모토 인덱스와 성격이 비슷하다. 다만 자카드 인덱스는 합집합의 개수, 즉 n_i와 n_j의 합에 대한 비율인 반면 Cosine 유사도는 곱을 이용한다는 점에서 Cosine 유사도가 유사도 값 사이의 차이가 적은 경향이 있다. 코사인 유사도 또한 C^{win}으로부터 다음과 같이 계산될 수 있다:

$$\begin{aligned} s_{ij} &= \frac{n_{ij}}{\sqrt{n_i n_j}} \\ &= \frac{C^{win}[v_i, v_j]}{\sqrt{C^{win}[v_i] \cdot C^{win}[v_j]}} \end{aligned} \quad \text{[Equation 3.6]}$$

6) Word2vec

Word2vec 방법은 자연어 처리 분야에서 주변 단어의 출현 정보를 이용하여 신경망을 통해 분산표현으로 나타내는 방법이다. 세션 데이터에서의 개별 아이템 또한 시퀀스이기 때문에 앞, 뒤 아이템들에 의하여 정의될 수 있다. 즉, Word2vec 방법을 아이템 시퀀스에 적용하면 두 아이템은 주변 아이템들의 출현이 비슷해질수록 유사하게 측정된다. 본 논문에서는 gensim 패키지를 이용

하여 개별 아이템을 임베딩하였다. 개별 아이템의 가장 유사한 아이템은 Equation 3.4의 임베딩 벡터 간의 계산식을 통해 가장 높은 값을 갖는 아이템으로 찾아질 수 있다.

2. 변형 아이템 선택

원본 세션과 비슷하고 실제로 관측될만한 세션을 생성하기 위해서는 원본 세션에 적절한 변형을 가해야 한다. 그러나 세션 내 모든 아이템에 다른 아이템으로 대체하거나 삽입하는 변형을 가한다면 해당 세션은 원본 데이터 분포와는 너무 멀어진 데이터가 된다. 마찬가지로 세션 내에서 너무 적은 수의 아이템만 변형한다면 원본 데이터와 너무 유사한 세션 여러 개가 학습되어 모델이 오히려 과적합될 수 있다. 그러므로 세션 내 아이템 중에서 어떤 아이템을 변형할지 선택하는 기준이 매우 중요하다.

1) 유사도 기준

본 논문에서는 선택된 아이템을 해당 아이템과 가장 높은 유사도를 갖는 아이템으로 변형하기 때문에 최고 유사도 값에 대해서만 논한다. 여기서 최고 유사도 값이란 개별 아이템 기준으로 다른 모든 아이템들과의 유사도를 계산했을 때 유사도 값 중 최고값을 의미한다. 최고 유사도 값이 높고 낮음에 따라 다음과 같이 해석할 수 있다.

아이템 v_i의 최고 유사도 값이
- 높다 : 아이템 v_i와 확실하게 유사한 아이템이 존재한다.
- 낮다 : 아이템 v_i와 확실하게 유사한 아이템이 존재하지 않는다.

만약 임의로 변형할 아이템을 선택하여 최고 유사도 값이 낮은 아이템을 변형하게 된다면 해당 세션은 관측하기 어려운 세션이 된다고 할 수 있다. 왜냐하면 실제 유저는 해당 아이템과 낮은 유사도를 갖는 아이템을 대신 선택하거나 연이어 선택할 확률이 낮기 때문이다. 따라서 최고 유사도 값이 높은 아이템을 선택해야 원본데이터와 유사하면서도 관측될만한 세션이 생성될 것이라고 할 수 있다. 본 연구에서는 최고 유사도 값이 높은 아이템과 낮은 아이템을 구별하여 각각의 아이템을 선택하여 증강했을 때 추천 성능이 어떻게 변화하는지 확인한다.

2) 출현 빈도 기준

개별 아이템은 출현 빈도가 높고 낮은지에 따라 다음과 같이 해석될 수 있다.

아이템 v_i의 출현 빈도가

- 높다 : 한 세션 내에서 임의로 하나의 아이템을 선택했을 때 아이템 v_i가 선택될 확률이 높다. 또한 정답 라벨에도 많이 등장한다.
- 낮다 : 한 세션 내에서 임의로 하나의 아이템을 선택했을 때 아이템 v_i가 선택될 확률이 낮다. 또한 정답 라벨에도 적게 등장한다.

아이템의 출현 빈도에 대해서는 다음 장에서 수치와 함께 자세히 다룰 것이지만, 세션 데이터는 소수 인기 아이템들의 출현 빈도가 나머지 다른 아이템들의 출현 빈도에 비하여 상당히 높다는 특징이 있다. 만약 데이터 증강을 위하여 한 세션 내에서 임의로 아이템을 선택한다면 이러한 인기 있는 소수의 아이템이 많이 선택될 것이며 소수의 아이템들만이 역시 대체되거나 삽입되어 변형된 결과도 획일적이게 된다. 이렇게 증강된 세션 역시 소수의 인기 아이템으로 구성될 확률이 높다. 즉, 무작위로 선택하여 증강하면 소수의 아이템만이 반복적으로 선택되어 특정 아이템으로 대체되거나 삽입되는 과정도 반복적으로 일어나기 때문에 원본 데이터 분포와 너무 유사한 세션이 생성되거나 생성된 세션들끼리도 유사한 세션이 많아지는 경향이 있다. 이는 결국 과적합에 의한 일반화 성능이 저하로 이어질 수 있다.

따라서 전체 데이터 내에서의 출현 빈도 중 높은 비중을 차지하는 인기 아이템을 위주로 증강하게 된다면 추천 결과는 소수 아이템에 집중되어 유저가 관심을 가질 만한 다른 카테고리 또는 다른 유형으로 유도하고자 하는 길 자체가 차단될 수 있다. 따라서 비인기 아이템 중에서 유저가 클릭할 만한 아이템 또한 증강하여 학습시켜야 일반화 성능을 향상시킬 수 있다는 점도 주의해야 한다.

최고 유사도 값과 출현 빈도에 대해서 분석한 것과 같이, 증강을 위해 세션 내 모든 아이템을 변형하거나 임의로 선택하여 변형하면 원본 데이터의 분포를 크게 벗어나거나 원본 데이터와 너무 유사한 데이터가 더욱 많아질 수 있다. 또한 무조건적으로 최고 유사도 값이 높은 아이템만 변형하는 것이 좋다고 할 수 없다. 그 이유는 최고 유사도 값이 높은 아이템은 대부분 적게 출현

한 아이템이므로, 이러한 아이템 위주의 세션은 원본 데이터 분포와 너무 먼 세션이 될 수 있기 때문이다. 출현 빈도가 낮은 아이템만 선택하여 증강하는 것도 좋은 방법이라고 할 수 없다. 그 이유는 출현 빈도가 낮은 아이템이 너무 많아진다면 해당 데이터의 정답 라벨 아이템에도 비인기 아이템이 많아지기 때문이다.

이처럼 최고 유사도 값과 출현 빈도에 따라 개별 아이템은 증강 시 꼭 필요한 아이템이 되기도 하고 성능을 더욱 악화시키는 아이템이 될 수도 있다. 따라서 본 논문에서는 데이터셋으로부터 추출할 수 있는 출현 빈도와 유사도 정보를 고려하여 선택하고 증강 효과를 비교함으로써 관측될만한 세션을 생성하는 데 도움이 되는 아이템과 그렇지 못한 데이터의 특성에 대하여 고찰한다.

3. 증강 방법

[28]의 informative 방식은 임의로 선택된 아이템을 가장 유사한 아이템으로 대체하는 방식과 해당 아이템 앞에 가장 유사한 아이템을 삽입하는 2가지 방식을 의미한다. 이 방식은 기존 세션의 순서 정보를 파괴하지 않는다는 장점이 있다. 또한 실제 상황에서 유저는 선택된 아이템 대신에 가장 유사한 아이템을 클릭할 수도 있기 때문에 informative 방식으로 생성된 신규 세션은 실제로 관측될만하고 아직 관측되지 않은 데이터라고 할 수 있다.

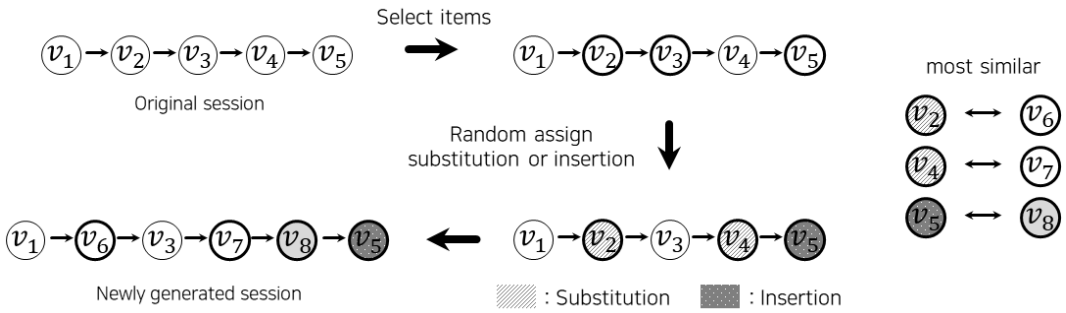

Fig. 3.5 원본 세션으로부터 변형 아이템을 선택하고 대체 및 삽입하여 신규 세션을 생성하는 과정

본 논문에서는 이러한 informative 증강 방식을 차용하여 세션 내에서 증강의 효과를 최대화할 수 있는 아이템들을 선택한 후 자기 자신과 가장 유사한 아이템으로 대체하거나 자기 자신의 앞자리에 가장 유사한 아이템을 삽입하는

방식을 취하였다. 대체나 삽입 중 하나의 방식으로만 증강되는 것을 막기 위하여 선택된 아이템에 대한 대체와 삽입 방식의 선택은 임의로 정해지도록 하였다. 이를 통해 충분한 정보를 기반으로 선택된 아이템에 한하여 임의성을 부여함으로써 다양한 변형 형태가 생성되도록 하였다. 증강 방법을 그림으로 나타내면 Fig. 3.5와 같다.

Ⅳ. 실험 설계 및 결과

본 장에서는 3장에서 제안한 세션 데이터 증강 방법을 현실 세계의 데이터에 적용하여 데이터 증강에 의한 세션 기반 추천시스템 성능의 향상을 검증하고 다양한 조건에 의한 데이터 증강 결과들을 서로 비교 및 분석한다.

1. 실험 데이터

본 연구에서는 세션 기반 추천 모델 연구에서 가장 널리 쓰이는 두 가지의 데이터에 대하여 실험을 수행하였다. 하나는 RecSys Challenge 2015에서 사용된 Yoochoose 데이터이고 다른 하나는 Diginetica 데이터이다. Yoochoose 데이터는 아이템에 대한 사용자의 클릭 이벤트로 구성되어 있고 Diginetica 데이터는 클릭 및 구매 로그에서 추출된 사용자 세션 데이터이다.

2. 평가지표

실제 추천시스템은 유저에게 N개의 아이템을 제공하는 방식으로 작동하기 때문에 순차적 추천시스템처럼 여러 개의 아이템 중 정답 아이템이 하나밖에 없는 경우에는 다음에 출현할 확률이 높은 top-N개의 아이템 안에 정답 아이템이 포함되는지 여부로 성능을 측정한다. 본 논문에서는 개별 세션별로 추천 모델이 마지막 아이템으로 예측한 확률 상위 20개의 아이템 중에서 정답 아이템이 포함될 경우에는 1, 안 될 경우에는 0을 부여하여 전체 세션에 대해 평균을 취하는 Recall@20 지표를 사용하였다. 그러나 Recall은 성능 측정 시 단순 이진(binary) 방식으로만 계산하기 때문에 top-N 안에서 정답 아이템이 얼마나 높은 순위에 위치했는지를 고려하지 않는다. 따라서 MRR@20 (Mean Reciprocal Rank) 지표를 추가로 사용한다. MRR은 N개의 아이템 내에서 정답 아이템이 위치한 순위를 측정하므로 추천 순위가 중요한 시나리오에서 더욱 정확한 지표가 될 수 있다. MRR@20은 상위 20개 아이템 중 정답 아이템 순위의 역수를 사용하며, 20개 안에 포함되지 않을 경우에는 0을 부여한 후 전체 세션에 대하여 평균을 취함으로써 계산된다.

3. 활용 모델

본 논문에서는 데이터 증강의 효과를 검증하기 위하여 세션 기반 추천시스템에

서 가장 대표적인 두 모델 NARM(Neural Attentive Recommendation Machine)[15]과 SR-GNN(Session-based Recommendation with Graph Neural Network)[18]을 사용하였다.

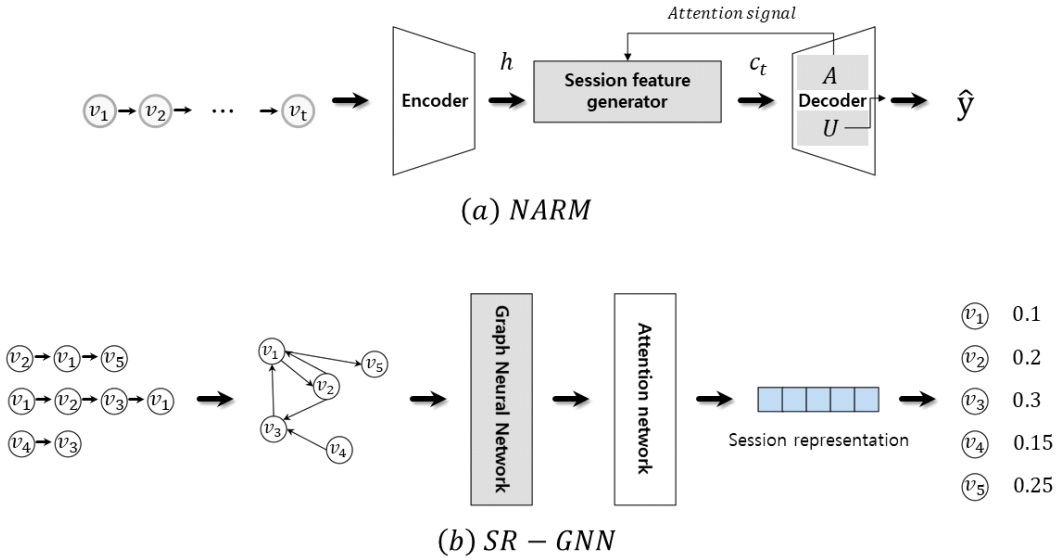

Fig. 4.1 NARM(a)과 SR-GNN(b)의 아키텍처

NARM은 neural 인코더-디코더 구조를 이용하여 세션의 고차원 hidden representation으로부터 모든 아이템에 대하여 현재 세션 안에서 발생할 수 있는 랭킹 리스트를 생성한다. 인코더와 디코더 사이에 세션 representation에 대하여 어텐션 구조를 적용함으로써 기본 순환신경망 모델을 사용한 세션 기반 추천시스템에서 큰 성능 향상을 달성하였다.

SR-GNN은 아이템 간의 복잡한 전이(transition) 정보를 더욱 잘 추출할 수 있는 그래프 신경망을 이용하였다. 개별 세션이 아닌 여러 개의 세션에 포함된 아이템끼리 directed graph를 구성하고 그래프 신경망을 통해 아이템 representation을 더욱 정교하게 생성함으로써 추천 성능을 향상시켰다.

4. 실험 설계

본 절에서는 원본 데이터의 양, 유사도 지표, 증강 아이템 선택 기준에 따른 데이터 증강의 효과를 분석하기 위하여 수행한 실험의 설계 과정에 대해 구체적으로 기술한다.

1) 원본 데이터의 양에 따른 데이터 증강 효과

본 연구에서는 세션 데이터가 충분하지 않은 상황에서 세션 기반 추천시스템의 성능을 향상시키기 위하여 관측할만한 데이터를 추가적으로 생성하는 데이터 증강이 목표이다. 현실 세계에서는 데이터 양에 대한 다양한 환경이 존재하며 데이터의 양은 각자가 처한 상황에 따라 상대적으로 인식된다. 본 실험에서는 데이터가 부족한 여러 상황을 가정하고 데이터의 양에 따른 증강 효과를 관찰하기 위하여 세션 기반 추천 알고리즘 검증에 널리 사용되는 Yoochoose 1/64, Diginetica 데이터 뿐만 아니라 데이터 양을 더욱 줄인 다양한 크기로 실험을 진행하였다. 두 데이터와 이로부터 더욱 적은 양만 추출한 데이터들에 대한 통계량은 Table 4.1과 같다.

	yoochoose 1/64		yoochoose 1/128		yoochoose 1/256		yoochoose 1/512	
	train	test	train	test	train	test	train	test
세션 수	124,472	15,172	62,236	15,172	31,118	15,016	15,559	14,646
아이템 수	17,597	6,095	15,147	6,095	11,814	5,508	7,959	4,421
	diginetica		diginetica 1/3		diginetica 1/6		diginetica 1/12	
	train	test	train	test	train	test	train	test
세션 수	186,670	15,963	62,223	15,644	31,111	15,000	15,555	13,898
아이템 수	43,097	21,131	36,427	19,651	29,129	17,063	21,297	13,594

Table 4.1 실험에 사용한 8개 데이터셋의 세션 수와 아이템 수

본 연구에서는 데이터 증강 기법의 성능 향상을 검증하기 위하여 [15]과 [18]에서 사용한 Yoochoose 1/64와 Deginetica 데이터셋을 사용하였다. 그리고 현실 세계에서 세션 데이터가 부족한 상황을 다양하게 가정하기 위하여 두 데이터셋을 각각 절반씩 3번 더 나누어 Yoochoose의 1/64, 1/128, 1/256, 1/512, Diginetica의 원본, 1/3, 1/6, 1/12 총 8개의 데이터셋으로 실험을 진행하여 데이터셋 크기에 따른 데이터 증강 효과의 차이를 확인하였다.

전처리를 위해 원본 데이터에서 우선 길이가 1인 세션을 제거하고, 출현 빈도가 5회 이하인 아이템을 제거하였다. 그 이후 다시 길이가 1인 세션을 제거하여 최소 2개 이상의 아이템이 존재하는 세션만을 사용하였다. Yoochoose 1/64와 Diginetica는 모두 가장 최근 7일 동안의 세션을 테스트 데이터로 사용하였고 Yoochoose 1/64는 나머지 세션 중 가장 최근 1/64를, Diginetica는 나머지 세션을 학습 데이터

로 사용하였다. 테스트 데이터에 등장하는 아이템 중에서 학습데이터에 등장하지 않는 아이템은 테스트 데이터에서 제거하였고 이로 인해 길이가 1이 된 세션은 다시 제거하였다. 나머지 6개의 데이터(뒤에 분수가 표기된 데이터)의 경우, 각각 Yoochoose 1/64와 Diginetica의 학습데이터에서 해당하는 분수만큼의 최근 세션들을 학습데이터로 사용하였다. 그리고 Yoochoose 1/64와 Diginetica의 테스트 데이터를 그대로 사용하되 마찬가지로 테스트 데이터에 포함된 아이템들 중에서 학습데이터에 포함되어있지 않으면 해당 아이템을 삭제한 후 길이가 1이 된 세션을 제거하였다.

2) 유사도 지표 간 비교

3장에서 살펴본 바와 같이 유사도 지표에 따라 개별 아이템의 가장 유사한 아이템이 달라질 수 있다. 즉, 같은 아이템을 선택해서 변형할지라도 유사도 지표에 따라 생성되는 세션이 달라진다. 유사도 지표가 데이터 증강 의도에 적합하다면 변형된 세션은 원본 데이터와 유사하지만 유사도 지표가 정확하지 않다면 변형된 세션은 관측하기 어려운 세션이 된다. 따라서 본 실험에서는 3장 1절에서 소개한 동시 발생 빈도, PMI, 자카드 인덱스, 타니모 인덱스, Cosine 유사도, 그리고 Word2vec 등 6가지 유사도 지표를 사용하여 증강해보고 이에 따른 성능 비교를 통해 세션 데이터에 가장 적합한 유사도 지표가 무엇인지 밝히고자 한다.

Word2vec 방법을 포함한 6가지 유사도 지표 모두 window 크기를 사전에 사용자가 설정해야 한다. Window의 크기 k를 설정한다는 것은 k 거리 안에 출현하는 아이템끼리 유사한 아이템이라고 정의한다는 의미가 있다. Yoochoose 1/64와 Diginetica 데이터의 세션 길이별 분포를 확인했을 때 길이가 10 이하인 데이터가 두 데이터 모두 90% 이상을 차지하는 것을 고려하여 본 논문에서는 window의 크기를 5로 설정하였다(Table 4.2, 4.3). 즉, 길이가 6 이하인 세션의 경우 세션 내 모든 아이템이 함께 유사하다고 기록되고, 길이가 11인 세션의 정 가운데 아이템은 해당 세션 내 모든 아이템과 유사하다고 측정된다.

Word2vec은 gensim 패키지의 Word2vec 객체를 이용하였으며 임베딩 크기는 100, window는 5, 최소 출현 빈도는 1로 설정하고 CBOW(Continuous Bag of Words) 모델을 이용하였다.

session 길이	개수	비율	누적 비율
2	54,766	0.44	0.44
3	23,825	0.19	0.63
4	14,479	0.12	0.75
5	8,564	0.07	0.82
6	5,799	0.05	0.86
7	3,836	0.03	0.89
8	2764	0.02	0.92
9	2067	0.02	0.93
10	1578	0.01	**0.95**
11	1231	0.01	0.96
12	938	0.01	0.97
13	731	0.01	0.97
14	625	0.01	0.97
15	463	0.01	0.98

Table 4.2 Yoochoose 1/64 데이터의 세션 길이별 분포

session 길이	개수	비율	누적 비율
2	52,946	0.28	0.28
3	35,860	0.19	0.48
4	25,313	0.14	0.61
5	18,377	0.10	0.71
6	13,485	0.07	0.78
7	9,917	0.05	0.84
8	7,502	0.04	0.88
9	5418	0.03	0.90
10	4232	0.02	**0.93**
11	3209	0.02	0.94
12	2307	0.01	0.96
13	1850	0.01	0.97
14	1362	0.01	0.97
15	1084	0.01	0.98

Table 4.3 Diginetica 데이터의 세션 길이별 분포

3) 최고 유사도 값과 출현 빈도에 따른 아이템 선택

변형할 아이템을 선택하는 과정에도 사용자의 선택 기준이 개입되어야 한다. 본 실험에서는 최고 유사도 값과 출현 빈도를 이용하여 다양한 선택 조건을 마련하고 선택 조건별로 구별하여 실험을 진행하였다. 최고 유사도 값과 출현 빈도에 대해 각각 높음과 낮음으로 구별하여 Fig. 4.2와 같이 네 가지 경우로 전체 아이템을 분류하였다.

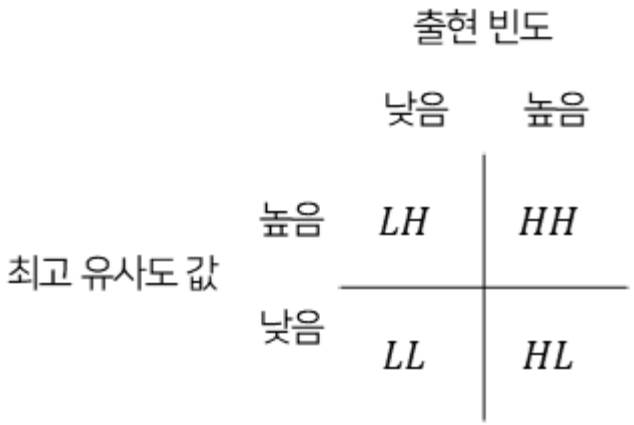

Fig. 4.2 아이템 분류 클래스

네 가지 경우에 대한 분석은 다음과 같다.

(1) HH : 출현 빈도가 높고 최고 유사도 값도 높은 아이템

이 경우에는 아이템 집합 중 소수의 아이템이 선택될 것이며 선택된 아이템

과 확실하게 유사한 아이템이 존재하므로 변형해도 별 문제가 없을 것이다. 그러나 생성된 세션이 원본 데이터의 분포와 상당히 유사할 가능성이 크다.

(2) LH : 출현 빈도는 낮고 최고 유사도 값은 높은 아이템
출현 빈도가 낮은 아이템들이 선택되면 아이템 집합 중 다수 아이템들이 선택될 것이며, 선택된 아이템들은 확실하게 유사한 아이템들이 존재한다. 즉, 적게 출현했지만 확실히 유사한 아이템이 존재한다는 것은 비인기 아이템임에도 특정 군집을 형성하고 있을 것으로 판단된다.

(3) HL : 출현 빈도가 높고 최고 유사도 값은 낮은 아이템
이 경우는 아이템 집합 중 소수의 아이템이 선택될 것이며 선택된 아이템들과 확실하게 유사한 아이템이 존재하지 않으므로 생성된 세션은 원본 데이터 분포에서 멀어질 것이다.

(4) LL : 출현 빈도가 낮고 최고 유사도 값도 낮은 아이템
출현 빈도가 낮은 아이템들이 선택되면 아이템 집합 중 다수 아이템들이 선택될 것이며, 이 아이템들과 확실하게 유사한 아이템이 존재하지 않을 것이기 때문에 원본 데이터 분포와 가장 멀어진 세션이 생성될 것이다.

본 논문에서는 네 가지 경우를 각각 클래스라고 지칭하며 특정 클래스에 속하는 아이템을 선택하여 증강하는 실험들을 서로 비교함으로써 각 클래스와 증강 효과 사이의 관계에 대해서 분석한다. 그리고 최고 유사도 값, 출현 빈도의 4가지 클래스를 분류하는 기준은 전체 아이템의 출현 빈도 합에서 1/4에 가깝게 차지하도록 조정하였다. 예를 들어 전체 아이템 집합 V의 크기가 100이고 전체 아이템의 출현 빈도 합이 1,000회라면, 하나의 클래스에 속하는 아이템만 선택하는 실험에서는 1,000의 1/4인 250개에 가까운 아이템이 선택되어 변형되도록 최고 유사도 값과 출현 빈도의 기준값을 설정하였다. 이렇게 모든 아이템의 출현 빈도의 합을 기준으로 클래스를 분류할 경우, 3절에서 설명한 것처럼 소수의 인기 아이템이 차지하는 빈도가 대부분이기 때문에 아이템 집합 V 내에서는 클래스별 아이템의 개수가 매우 불균형해지게 된다(Fig. 4.3). 또한 본 실험에서는 특성값 기준을 최대한 균형 있게 설정한 후 진행하였기 때문에 $p=1$로 설정하였다.

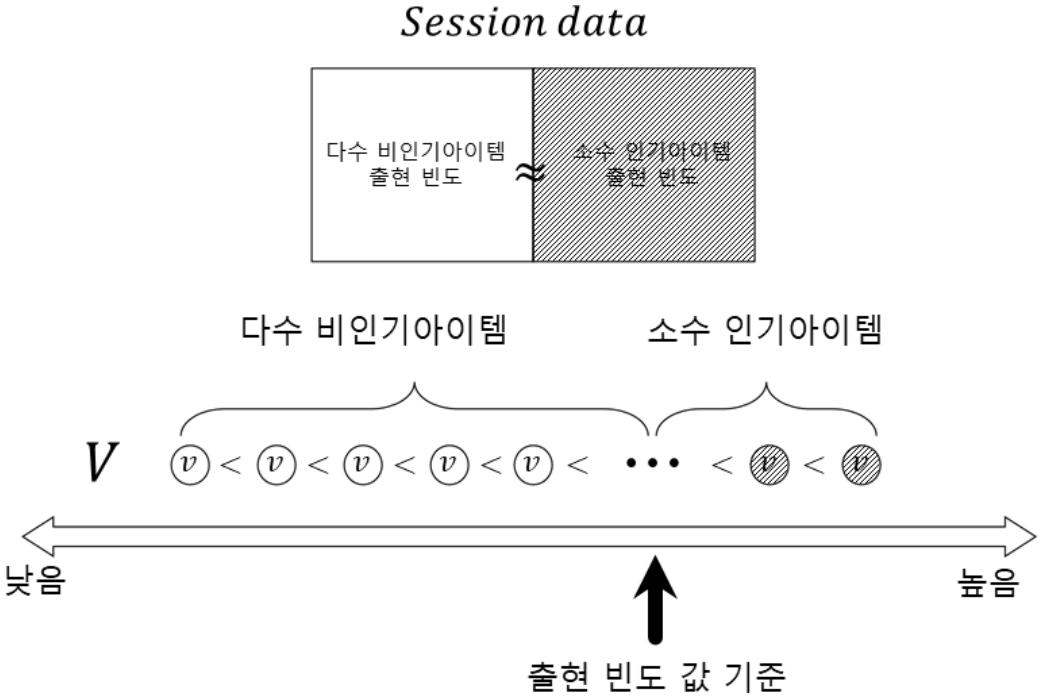

Fig. 4.3 출현 빈도 값 기준은 전체 아이템 출현 빈도의 합에 대하여 양 집단이 서로 유사해지도록 설정

또한 세션 내 전체 아이템을 증강한 경우와 전체 아이템 출현 빈도의 합에서 1/4만큼의 아이템을 무작위로 선택하여 증강한 경우를 추가적으로 실험하여 무작위 선택에 의한 증강과 아이템 특성 정보에 기반한 선택 증강을 비교하였다.

유사도 지표 중 동시 발생 빈도를 사용했을 때는 클래스 분류를 위한 최고 유사도 값 기준을 정규화하여 사용하였다. 예를 들어 만약 아이템 v_i와 가장 유사한 아이템 v_j의 동시 발생 빈도가 $C(v_i,v_j)$이고 v_i가 출현한 window의 수를 $C(v_i)$라고 할 때, v_i는 $\frac{C(v_i,v_j)}{C(v_i)}$가 최고 유사도 값 기준 이상인지 미만인지에 따라서 클래스가 분류되도록 하였다. 그 이유는 출현 빈도가 높아질수록 최고 유사도 값(동시 발생 빈도)도 비례해서 커지기 때문에 클래스 분류가 어려워지고 HH와 LL만 많아지는 현상 때문이다. 따라서 개별 아이템 기준으로 가장 많이 동시 발생한 아이템과의 동시 발생 빈도를 자기 자신의 출현 빈도로 나누어 모든 아이템들의 유사도 척도를 통일시켰다.

5. 실험 결과

본 절에서는 8개 데이터셋에 대하여 각 클래스와 유사도 지표에 따라 증강한 결과를 원본 데이터와 비교한 실험 결과를 살펴본다. NARM과 SR-GNN은 각각 성능 변화가 거의 없는 지점인 100, 15 에폭만큼 학습하였다.

Figs. 4.4~4.7은 각각 Yoochoose 1/64, 1/128, 1/256, 1/512에 포함된 각 아이템의 최고 유사도 값과 출현 빈도를 산점도로 나타낸 것이다. 마찬가지로 Figs. 4.8~4.11은 각각 Diginetica, 1/3, 1/6, 1/12에 포함된 각 아이템의 최고 유사도 값과 출현 빈도를 산점도로 나타낸 것이다. 각각의 점은 하나의 아이템을 의미하며, 점의 크기는 출현 빈도를 나타내도록 하여 점들의 면적이 전체 아이템의 출현 빈도의 합에서 차지하는 비율을 직관적으로 알 수 있도록 표현하였다. 점의 색깔은 빨간색이 HH, 초록색이 LH, 파란색이 HL, 노란색이 LL 클래스를 의미한다. 데이터셋과 유사도 지표별 각 클래스를 분류하는 최고 유사도 값과 출현 빈도의 기준 값은 Table 4.4와 같다. 지면 한계상 최고 유사도 값 기준 값을 a, 출현 빈도 기준 값을 b로 표기하였다.

			Co-occurrence (normalized)	PMI	Jaccard	Tanimoto	Cosine	W2v
Yoochoose	1/64	a	0.15	9	0.2	0.035	0.14	0.99
		b	500	300	400	500	400	150
	1/128	a	0.16	9.5	0.22	0.037	0.14	0.993
		b	150	150	130	130	130	130
	1/256	a	0.16	9	0.22	0.038	0.15	0.997
		b	110	100	100	100	100	100
	1/512	a	0.165	8.5	0.24	0.038	0.16	0.999
		b	80	80	80	80	80	80
Diginetica	1/1	a	0.065	10.2	0.15	0.022	0.06	0.998
		b	40	40	40	40	40	40
	1/3	a	0.08	10.8	0.2	0.035	0.5	0.077
		b	15	18	18	18	18	17
	1/6	a	0.09	10.8	0.25	0.042	0.1	0.4
		b	9	10	10	10	10	10
	1/12	a	0.11	10.5	0.33	0.055	0.39	0.13
		b	8	8	6	6	6	6

Table 4.4 데이터셋별 클래스 분류를 위한 최고 유사도 값과 출현 빈도의 기준 값

1) 원본 데이터 양에 따른 효과

데이터의 양이 클수록 증강에 의한 효과는 적은 경향을 보였다. Table 4.5, 4.6은 Yoochoose 1/64의 데이터 증강 결과를 나타낸다. 진하고 파랗게 표시된 결과는 원

본데이터의 결과보다 더 높은 경우이다. 데이터의 크기가 큰 Yoochoose 1/64의 경우 Recall은 NARM의 경우 최대 약 0.1, MRR은 소폭의 향상이 있었으며, SR-GNN의 경우 Recall의 소폭 향상, MRR는 최대 약 0.2의 향상이 있었다.

데이터의 양이 작아질수록 Recall과 MRR의 성능 향상 폭이 더욱 증가했다. Table 4.7, 4.8은 Yoochoose 1/128의 데이터 증강 결과를 나타낸다. 데이터 크기가 1/64에 비하여 절반으로 더욱 작아진 상황에서는 Recall이 약 0.1 ~ 0.3이 상승하였고 MRR은 0.13 ~ 0.35이 상승했다.

데이터의 양이 더욱 작은 Yoochoose 1/256과 Yoochoose 1/512에서는 상승 폭이 더욱 증가하여 1/256은 Recall이 0.2 ~ 2.4, MRR이 0.1 ~ 1.4만큼 상승하였고(Table 4.9, 4.10), 1/512는 Recall이 0.5 ~ 6, MRR이 0.1 ~ 3.9만큼 상승하였다(Table 4.11, 4.12).

Table 4.13, 4.14는 Diginetica의 데이터 증강 결과를 나타낸다. 데이터셋의 크기가 큰 Diginetica 데이터의 경우에는 성능이 향상된 경우가 없었는데, Diginetica 데이터가 Yoochoose 1/64보다 아이템의 수가 더 많고 세션의 양은 비슷하다는 점을 고려해보았을 때, Diginetica는 아이템 간 유사도를 계산할 수 있을 만큼의 개별 아이템의 충분한 출현 빈도가 부족했다고 해석할 수 있다.

Table 4.15, 4.16은 Diginetica 1/3의 데이터 증강 결과를 나타낸다. Diginetica에서는 데이터 증강에 의한 성능 향상이 없었지만 데이터의 양이 1/3만큼 줄어든 약 60,000개의 세션 데이터로부터 증강하여 실험한 결과 Recall은 최대 0.3, MRR은 최대 0.08만큼 상승하였다. 데이터의 양이 작을수록 데이터 증강 효과가 나타나는 것을 알 수 있다.

데이터의 크기가 더욱 작은 Diginetica 1/6과 Diginetica 1/12에서는 상승 폭이 더욱 증가하여 1/6은 Recall이 0.1 ~ 2, MRR이 0.08 ~ 1만큼 상승하였고(Table 4.17, .18), 1/12는 Recall이 0.7 ~ 5.7, MRR이 0.1 ~ 3만큼 상승하였다(Table 4.19, 4.20).

특히 Yoochoose 1/512와 Diginetica 1/12에서는 데이터 증강에 의한 성능이 각각 자기 자신보다 두 배의 세션을 가진 Yoochoose 1/256, Diginetica 1/6과 비슷해질 정도로 증가했다.

본 연구에서 제안하는 데이터 증강 방법을 사용할 경우, 실험으로부터 확인할 수 있는 확실한 성능 향상이 보장되는 세션의 수는 약 60,000개 이하일 때이다. 이러한 상황에서는 비교적 간단한 데이터 증강 기법을 사용함으로써 Recall과 MRR을 최소 0.1 이상 향상시킬 수 있음을 확인하였다. 세션의 수가 더욱 적은 약 30,000개 이하의 데이터 크기를 가진 상황이라면 데이터 증강 기법을 통하여 두 배의 데이터를 확보했을 때 만큼의 성능를 얻을 수 있다. 그러나 60,000개 이상의 세션이 있다면 데이터의 크기가 커질수록 성능 향상의 폭이 점점 작아지고, Diginetica 데이터에서 성능 향상이 이루어지지 않은 것처럼 데이터 증강 효과가 나타나지 않을 수 있다.

2) 유사도 지표 간 비교

Figs. 4.4~4.11를 통하여 8개 데이터셋 내 아이템들의 최고 유사도 값과 출현 빈도의 분포를 확인할 수 있다. 산점도를 통해 1) 원본 데이터 양에 따른 효과에서 예상할 수 있었던 Yoochoose와 Diginetica의 아이템별 출현 빈도의 차이를 확인할 수 있다. Yoochoose 1/64에서는 출현 빈도가 1,000회 ~ 5,000회에 속하는 아이템도 다수 존재하지만(Fig. 4.4) Diginetica 데이터에서는 대부분 아이템이 400회 이하 범주에 존재한다(Fig. 4.8). Diginetica에서는 데이터 증강에 의한 성능 향상이 없었지만 Yoochoose 1/64에서 존재한다는 것은 적은 수의 아이템들이 다수 출현하는 데이터셋에서 아이템 간의 유사도 측정이 정확해지고 증강 효과도 더욱 커질 수 있다고 해석할 수 있다.

Figs. 4.4~4.11를 통하여 유사도 지표별 최고 유사도 값의 분포를 살펴보면 Word2vec은 0.9 이상, PMI는 6~12, 타니모토 인덱스는 0~0.2, 나머지는 0~0.4에 존재하는 것을 알 수 있다. 타니모토 인덱스는 아이템 사이의 최고 유사도 값 차이가 매우 적고 동시 발생 빈도, PMI, 자카드 인덱스, Cosine 유사도의 경우에는 그에 비하여 분산이 더 크다는 특징이 있다.

성능 면에 있어서는 Word2vec을 제외한 모든 지표가 비슷한 성능 향상을 보였다. Table 4.5, 4.6을 통해 Yoochoose 1/64에서는 동시 발생 빈도, PMI, 자카드 인덱스, 타니모토 인덱스를 중심으로 성능이 향상되었다는 것을 확인할 수 있다. 특히 동시 발생 빈도와 자카드 인덱스는 거의 모든 실험 결과에서 가장 높은 성능 향상을 보여주었고, 6개 지표 중 가장 안정적인 성능 향상이 보장되는 지표라고 볼 수 있다.

다만 동시 발생 빈도, PMI, 자카드 인덱스, 타니모토 인덱스, Cosine 유사도 지표 등 5개의 지표는 데이터와 알고리즘에 따라 성능 향상 순위가 뒤바뀌기는 하지만 큰 차이가 존재하지 않았다. 이는 5가지 지표가 모두 동시 발생 행렬로부터 파생되었기 때문에 스케일의 변화만 있을 뿐 가장 유사한 아이템이 바뀌는 경우는 많지 않았다고 볼 수 있다.

Word2vec은 나머지 지표에 비하여 성능 향상을 보여주는 경우가 거의 없었는데, 이는 아이템 시퀀스에서 아이템 사이의 유사도를 측정할 때 문맥을 기반으로 측정하는 방식보다 물리적으로 가까운 위치에 빈번하게 등장할수록 유사하다고 측정하는 방식이 더욱 정확한 방식이라고 해석할 수 있다.

3) 최고 유사도 값과 출현 빈도에 따른 아이템 선택

거의 모든 데이터셋에서 LH 클래스의 성능 향상이 가장 큰 것으로 나타났다. Yoochoose 1/64(Table 4.5, 4.6), Diginetica 1/3(Table 4.15, 4.16)의 경우에는 다른 클래스에서 성능 향상이 없을 때 오직 LH에서만 성능 향상이 있었다. 모든 클래스에서 성능 향상을 보이는 경우에도, 그 폭이 가장 큰 경우는 모두 LH 클래스였다(Table 4.10, 4.12, 4.17~4.20). 클래스별 성능 향상 정도를 비교하였을 때 HH와 HL은 증강에 가장 좋지 않은 클래스였으며, LL보다도 성능 향상이 적은 경우가 더 많았다. 즉, 출현 빈도가 높은 아이템을 최대한 선택하지 않는 것이 데이터 증강에 의한 성능 향상에 도움이 된다고 해석할 수 있다.

최고 유사도 값 - 출현 빈도 산점도에서 확인할 수 있는 또 다른 점은 출현 빈도가 높은 아이템들이 최고 유사도 값이 높은 경우는 거의 없고, 적게 출현한 아이템들은 최고 유사도 값이 높은 경우가 비교적 많이 존재한다는 것이다. 즉, Figs. 4.4~4.11에서 초록색 점으로 표시된 LH 클래스 아이템들이 같은 최고 유사도 값 분류 기준을 적용한 빨간색의 HH 클래스 아이템들보다 최고 유사도 값의 평균이 높다. 즉, 최고 유사도 값이 높은 아이템 중에서도 출현 빈도에 따라 클래스 간의 성격이 매우 달라지게 된다.

LH 클래스 아이템들은 아이템 집합 내에서 다수를 차지하는 동시에 확실하게 유사한 아이템을 가진 아이템이다. 그렇기 때문에 LH 클래스 아이템을 변형했을 때 생성되는 세션은 다양성이 크고 관측될만한 세션이라고 할 수 있다. LH가 LL보다 더 좋은 성능을 보인 이유는 최고 유사도 값이 높은 아이템을 선택했기 때문이라고 볼 수 있고, HH와 HL에 비하여 LH가 좋은 성능을 보인 이유는 소수의 아이템

이 아니라 다수의 아이템을 중심으로 증폭되었기 때문에 일반화 성능이 더욱 향상된 것이라고 해석할 수 있다.

무작위로 선택하여 증강한 실험이 HH, HL, LL보다 성능이 좋은 경우도 있었지만 LH 클래스는 무작위로 선택한 실험과 모든 아이템을 변형한 실험보다 항상 좋은 결과를 보여주었다. 이는 증강에 도움이 되는 아이템과 그렇지 않은 아이템이 명확하게 구분될 수 있다는 점을 시사한다.

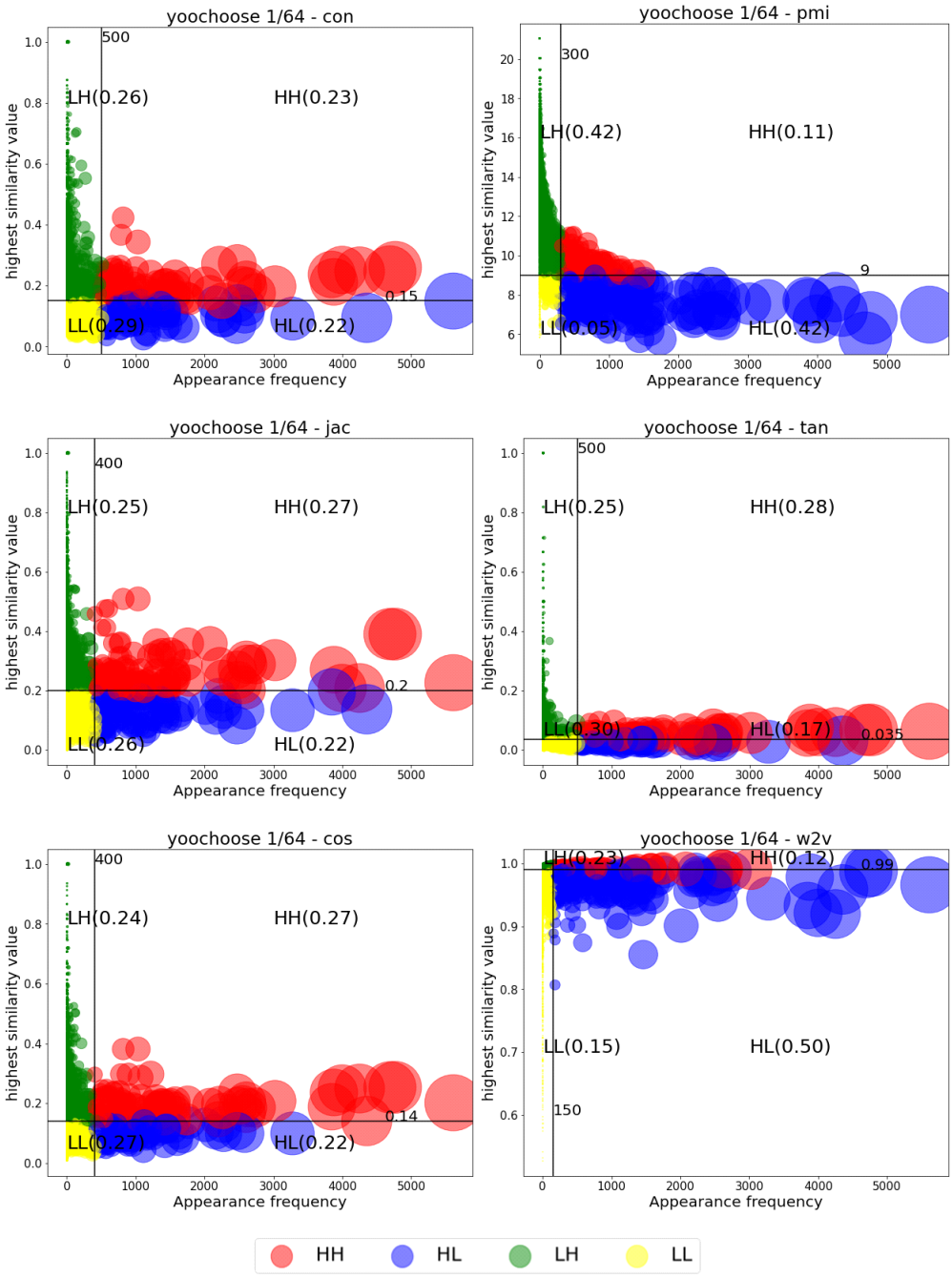

Fig. 4.4 Yoochoose 1/64의 유사도 지표별 출현 빈도와 최고 유사도 값 산점도

Table 4.5 Result of Yoochoose 1/64 - NARM

유사도 지표	HH		LH		HL		LL		random		all		original	
	Recall @20	MRR @20	Recall @20	MRR @20	Recall @20	MRR @20	Recall @20	MRR @20	Recall @20	MRR @20	Recall @20	MRR @20	Recall @20	MRR @20
co-occurrence	+51,212		+53,882		+54,019		+56,258		+45,844		+123,648			
	70.2337	30.1530	**70.7067**	30.0783	69.9380	29.7393	69.9290	29.6681	69.8482	29.5182	69.4381	29.1247		
	(-0.3716)	(-0.0342)	**(+0.1014)**	(-0.1089)	(-0.6673)	(-0.4480)	(-0.6762)	(-0.5192)	(-0.7571)	(-0.6690)	(-1.1672)	(-1.0626)		
PMI	+33,642		+66,770		+74,260		+13,325		+45,844		+123,648			
	69.8269	30.0266	70.0527	30.0747	69.7122	30.1357	69.8556	30.0628	69.2078	30.0097	70.1638	**30.2641**		
	(-0.7784)	(-0.1607)	(-0.5526)	(-0.1126)	(-0.8931)	(-0.0515)	(-0.7497)	(-0.1245)	(-1.3975)	(-0.1775)	(-0.4415)	**(+0.0769)**		
jaccard	+55,629		+45,801		+54,127		+57,785		+45,844		+123,648			
	69.9882	29.7913	**70.7246**	30.0781	69.9273	29.8580	70.1476	29.9411	69.4307	29.7868	69.3324	29.1273	세션 수 : 124,472	
	(-0.6171)	(-0.3959)	**(+0.1193)**	(-0.1091)	(-0.6780)	(-0.3292)	(-0.4576)	(-0.2461)	(-1.1746)	(-0.4004)	(-1.2729)	(-1.0599)		
tanimoto	+28,779		+41,624		+87,056		+37,854		+45,844		+124,472			
	70.1297	30.0711	**70.6727**	30.1291	69.7714	29.6521	70.0563	29.9381	69.7088	29.6707	69.4273	29.1295	70.6053	30.1872
	(-0.4756)	(-0.1162)	**(+0.0674)**	(-0.0581)	(-0.8339)	(-0.5351)	(-0.5490)	(-0.2491)	(-0.8965)	(-0.5165)	(-1.1780)	(-1.0577)		
cosine	+59,092		+46,088		+52,424		+53,243		+45,844		+123,648			
	70.0598	30.0625	70.508	29.9524	69.7982	29.5747	70.0814	29.9775	69.5672	29.6681	69.3109	29.2277		
	(-0.5454)	(-0.1247)	(-0.0545)	(-0.2348)	(-0.8071)	(-0.6125)	(-0.5239)	(-0.2097)	(-1.0381)	(-0.5191)	(-1.2944)	(-0.9596)		
w2v	+28,779		+41,624		+87,056		+37,854		+45,844		+124,472			
	70.2552	**30.2593**	69.8162	29.6421	69.1263	28.6467	70.1674	29.6356	69.4087	29.7833	68.3056	28.1598		
	(-0.3501)	**(+0.0721)**	(-0.7891)	(-0.5451)	(-1.4790)	(-1.5405)	(-0.4379)	(-0.5517)	(-1.1966)	(-0.4039)	(-2.2997)	(-2.0274)		

Table 4.6 Result of Yoochoose 1/64 - SR-GNN

co-occurrence	+51,212		+53,882		+54,019		+56,258		+45,844		+123,648			
	69.8452	30.8733	70.7401	**31.1409**	69.3562	29.9753	69.8326	30.4055	69.5882	29.4824	69.3851	29.5573		
	(-1.3884)	(-0.0797)	(-0.4935)	**(+0.1879)**	(-1.8773)	(-0.9777)	(-1.4010)	(-0.5476)	(-1.6454)	(-1.4706)	(-1.8485)	(-0.1954)		
PMI	+33,642		+66,770		+74,260		+13,325		+45,844		+123,648			
	69.5475	30.7576	70.2692	**31.1591**	69.3310	30.7046	69.8596	30.7201	69.8421	30.4401	70.4406	**31.4645**		
	(-1.6861)	(-0.1954)	(-0.9644)	**(+0.2061)**	(-1.9026)	(-0.2484)	(-1.3740)	(-0.2329)	(-1.3915)	(-0.5129)	(-0.7930)	**(+0.5115)**		
jaccard	+55,629		+45,801		+54,127		+57,785		+45,844		+123,648			
	69.6774	30.3580	70.8502	**30.9802**	69.5132	30.3692	70.2908	30.7932	69.7582	30.7781	69.3779	29.5737	세션 수 : 124,472	
	(-1.5562)	(-0.5950)	(-0.3834)	**(+0.0272)**	(-1.7204)	(-0.5838)	(-0.9427)	(-0.1598)	(-1.4754)	(-0.1749)	(-1.8557)	(-1.3793)		
tanimoto	+28,779		+41,624		+87,056		+37,854		+45,844		+124,472			
	70.0671	30.7386	70.8718	30.9298	69.6142	30.1813	69.9818	30.8465	69.4814	30.4381	69.3093	29.6365	71.2336	30.9530
	(-1.1665)	(-0.2144)	(-0.3618)	(-0.0232)	(-1.6193)	(-0.7717)	(-1.2518)	(-0.1065)	(-1.7522)	(-0.5149)	(-1.9243)	(-1.3165)		
cosine	+59,092		+46,088		+52,424		+53,243		+45,844		+123,648			
	69.6817	30.6481	70.7538	30.8861	69.4478	30.3483	69.8178	30.8132	69.7828	30.6817	69.3185	29.4865		
	(-1.5519)	(-0.3049)	(-0.4798)	(-0.0669)	(-1.7858)	(-0.6047)	(-1.4158)	(-0.1398)	(-1.4508)	(-0.2713)	(-1.9151)	(-1.4665)		
w2v	+28,779		+41,624		+87,056		+37,854		+45,844		+124,472			
	69.9625	30.8545	69.9101	30.8367	68.4884	29.1815	68.6510	29.0485	68.8118	29.7521	67.1588	29.1875		
	(-1.2711)	(-0.0985)	(-1.3234)	(-0.1163)	(-2.7452)	(-1.7715)	(-2.5826)	(-1.9045)	(-2.4218)	(-1.2009)	(-4.0748)	(-1.7655)		

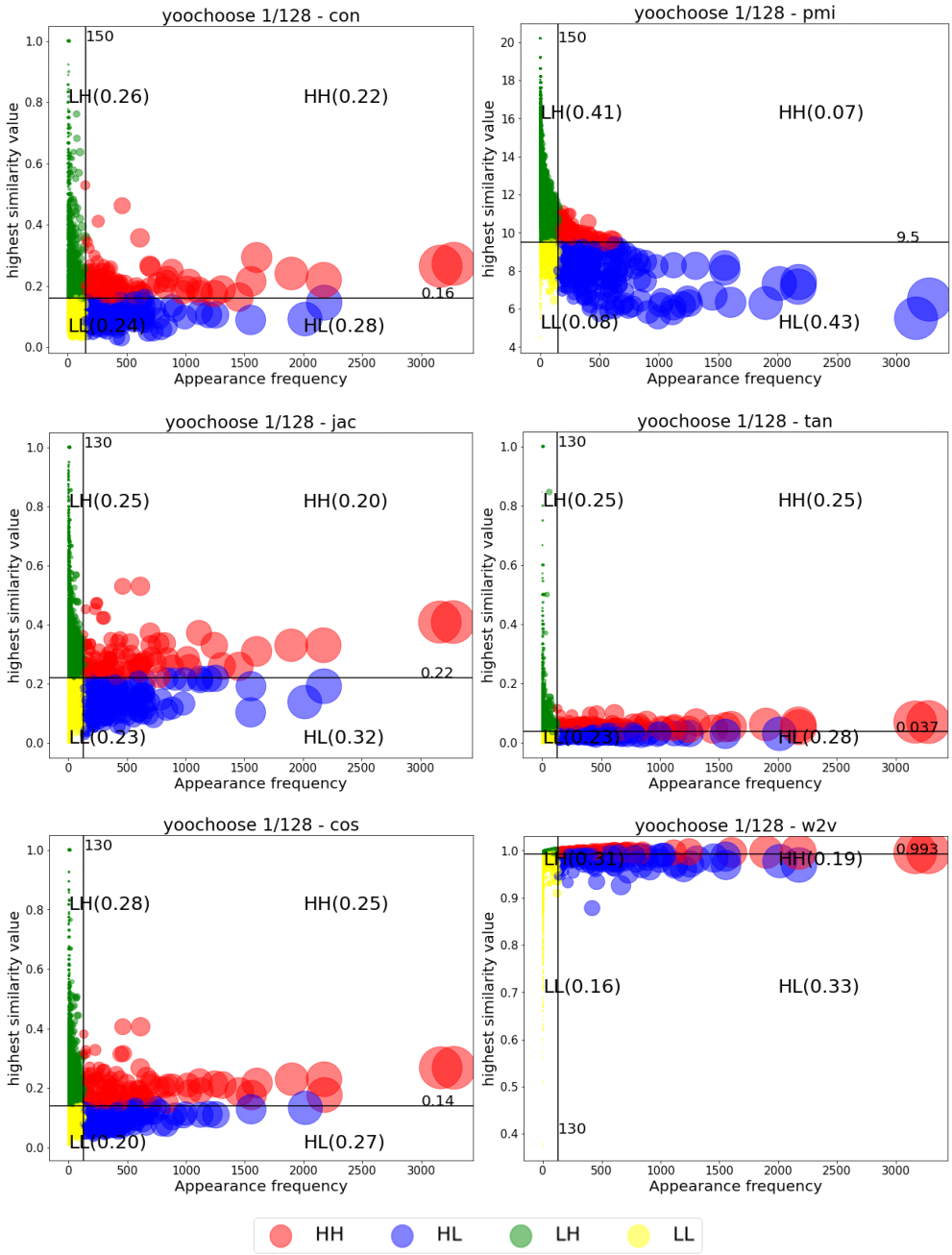

Fig. 4.5 Yoochoose 1/128의 유사도 지표별 출현 빈도와 최고 유사도 값 산점도

Table 4.7 Result of Yoochoose 1/128 - NARM

유사도 지표	HH		LH		HL		LL		random		all		original	
	Recall @20	MRR @20	Recall @20	MRR @20	Recall @20	MRR @20	Recall @20	MRR @20	Recall @20	MRR @20	Recall @20	MRR @20	Recall @20	MRR @20
co-occurrence	+25,522		+26,943		+29,500		+25,767		+24,604		+61,544			
	69.5204	29.8684	**70.0702**	29.7512	69.1135	29.1372	**69.8604**	29.5377	69.3218	29.6411	69.0321	28.9398		
	(-0.2376)	(-0.1203)	**(+0.3121)**	(-0.2375)	(-0.6445)	(-0.8515)	**(+0.1024)**	(-0.4510)	(-0.4362)	(-0.3475)	(-0.7259)	(-1.0488)		
PMI	+12,573		+35,355		+36,173		+10,963		+24,604		+61,544			
	69.1316	29.6892	69.4969	29.5357	69.0864	**30.1261**	69.0086	29.8375	69.0548	29.1025	69.5367	30.0360		
	(-0.6264)	(-0.2995)	(-0.2611)	(-0.4530)	(-0.6717)	**(+0.1375)**	(-0.7494)	(-0.1512)	(-0.7032)	(-0.8861)	(-0.2214)	(+0.0474)		
jaccard	+22,666		+23,783		+33,305		27,080		+24,604		+61544			
	69.5873	**30.0938**	**69.9056**	29.8221	68.5746	28.9420	**69.9146**	29.8144	69.1832	29.1685	68.7826	28.8709	세션 수 : 62,236	
	(-0.1707)	**(+0.1052)**	**(+0.1476)**	(-0.1665)	(-1.1834)	(-1.0466)	**(+0.1566)**	(-0.1742)	(-0.5748)	(-0.8201)	(-0.9755)	(-1.1177)		
tanimoto	+26,884		+24,841		+28,453		+23654		+24,604		+61544			
	69.3522	29.9851	**70.1606**	29.8240	68.6704	29.1751	**69.8785**	29.9184	69.1568	29.0572	68.8223	29.0020	69.7580	29.9886
	(-0.4058)	(-0.0035)	**(+0.4026)**	(-0.1647)	(-1.0876)	(-0.8135)	**(+0.1204)**	(-0.0703)	(-0.6012)	(-0.9314)	(-0.9357)	(-0.9867)		
cosine	+27,712		+25,995		+28,039		+22,089		+24,604		+61,544			
	69.4318	29.9341	**70.1751**	29.8535	68.7500	29.1315	69.5023	29.7025	68.9518	29.8410	68.7789	28.8509		
	(-0.3262)	(-0.0545)	**(+0.4170)**	(-0.1351)	(-1.0080)	(-0.8572)	(-0.2557)	(-0.2862)	(-0.8062)	(-0.1476)	(-0.9791)	(-1.1377)		
w2v	+21,858		+28,944		+34,570		+22,355		+24,604		+62,236			
	69.3956	29.1070	68.8802	29.3913	68.5330	28.8062	69.2618	29.5402	68.1896	29.4832	67.5763	27.4831		
	(-0.3624)	(-0.8816)	(-0.8778)	(-0.5973)	(-1.2250)	(-1.1825)	(-0.4962)	(-0.4484)	(-1.5684)	(-0.5054)	(-2.1817)	(-2.5056)		

Table 4.8 Result of Yoochoose 1/128 - SR-GNN

co-occurrence	+25,522		+26,943		+29,500		+25,767		+24,604		+61,544			
	69.2515	30.1275	**70.6940**	**30.5566**	68.4484	29.2085	69.8219	30.3060	69.8481	30.0815	68.8863	29.3007		
	(-0.9906)	(-0.2955)	**(+0.4520)**	**(+0.1335)**	(-1.7936)	(-1.2145)	(-0.4201)	(-0.1170)	(-0.3939)	(-0.3416)	(-1.3558)	(-1.1223)		
PMI	+12,573		+35,355		+36,173		+10,963		+24,604		+61,544			
	69.1370	30.2319	70.0618	**30.6063**	68.3884	**30.4464**	69.0243	30.1102	69.1123	29.1875	68.0848	29.4089		
	(-1.1050)	(-0.1912)	(-0.1802)	**(+0.1832)**	(-1.8536)	**(+0.0234)**	(-1.2177)	(-0.3128)	(-1.1297)	(-1.2356)	(-2.1572)	(-1.0142)		
jaccard	+22,666		+23,783		+33,305		27,080		+24,604		+61544			
	69.1838	30.2182	**70.5821**	**30.5788**	68.3482	30.1180	69.1882	30.1304	69.9828	30.3309	69.0298	29.5256	세션 수 : 62,236	
	(-1.0582)	(-0.2049)	**(+0.3401)**	**(+0.1557)**	(-1.8938)	(-0.3051)	(-1.0538)	(-0.2927)	(-0.2592)	(-0.0922)	(-1.2122)	(-0.8974)		
tanimoto	+26,884		+24,841		+28,453		+23654		+24,604		+61544			
	69.1370	30.2445	**70.5160**	**30.5885**	68.1595	29.3801	69.8528	30.4377	69.1531	30.1164	68.7536	29.6184	70.2420	30.4231
	(-1.1050)	(-0.1786)	**(+0.2740)**	**(+0.1654)**	(-2.0825)	(-1.0430)	(-0.3892)	(+0.0147)	(-1.0889)	(-0.3067)	(-1.4884)	(-0.8046)		
cosine	+27,712		+25,995		+28,039		+22,089		+24,604		+61,544			
	69.0407	30.3227	**70.5959**	**30.7761**	68.3158	29.5670	69.1821	30.2483	69.8163	30.1413	69.1814	29.5318		
	(-1.2013)	(-0.1004)	**(+0.3539)**	**(+0.3531)**	(-1.9262)	(-0.8561)	(-1.0599)	(-0.1748)	(-0.4257)	(-0.2818)	(-1.0606)	(-0.8913)		
w2v	+21,858		+28,944		+34,570		+22,355		+24,604		+62,236			
	69.3550	29.8418	69.0480	29.9369	68.0977	28.8848	69.2151	30.1367	69.4375	30.1813	67.2075	27.9214		
	(-0.8870)	(-0.5813)	(-1.1940)	(-0.4862)	(-2.1443)	(-1.5383)	(-1.0269)	(-0.2864)	(-0.8045)	(-0.2418)	(-3.0345)	(-2.5017)		

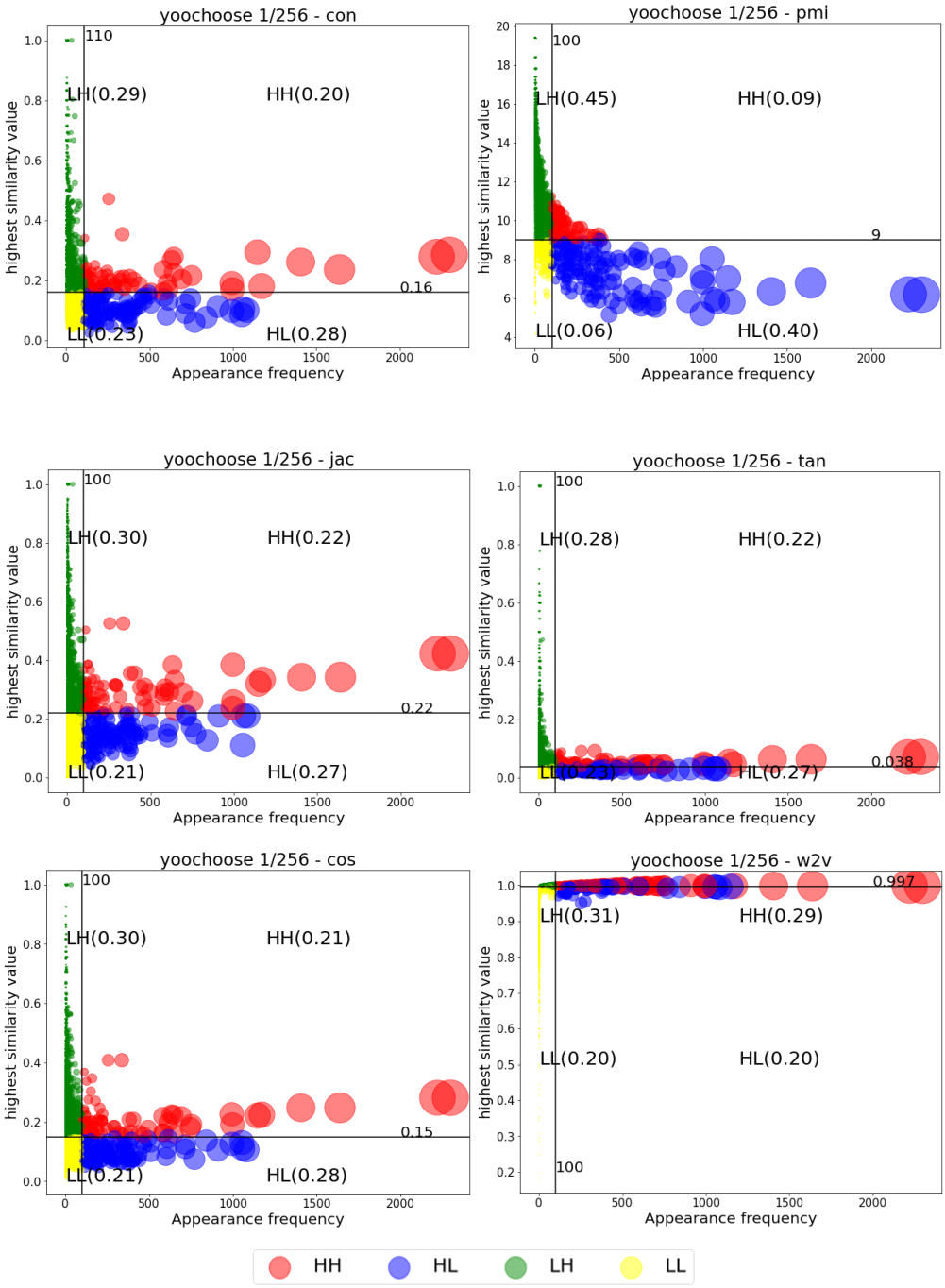

Fig. 4.6 Yoochoose 1/256의 유사도 지표별 출현 빈도와 최고 유사도 값 산점도

Table 4.9 Result of Yoochoose 1/256 - NARM

유사도 지표	HH		LH		HL		LL		random		all		original	
	Recall @20	MRR @20	Recall @20	MRR @20	Recall @20	MRR @20	Recall @20	MRR @20	Recall @20	MRR @20	Recall @20	MRR @20	Recall @20	MRR @20
co-occurrence	+11,661		+15,383		+14,518		+13,380		+13,383		+30,678			
	68.0292	29.4758	**68.7647**	29.3919	67.5110	28.4055	**68.3584**	29.4576	**68.6335**	29.3406	68.1148	28.5690		
	(-0.1244)	(-0.1042)	**(+0.6112)**	(-0.1881)	(-0.6426)	(-1.1746)	**(+0.2048)**	(-0.1224)	**(+0.6329)**	(-0.0826)	(-0.0388)	(-1.0111)		
PMI	+8,103		+19,603		+16,607		+4,492		+13,383		+30,678			
	67.5125	29.2831	68.0411	29.2776	67.7942	**29.7135**	67.7712	29.3530	67.6389	**29.5600**	**68.7601**	**29.7500**		
	(-0.6411)	(-0.2969)	(-0.1125)	(-0.3025)	(-0.3594)	**(+0.1335)**	(-0.3824)	(-0.2270)	(-0.3618)	**(+0.1368)**	**(+0.6065)**	**(+0.1699)**		
jaccard	+11,648		+14,152		+15,035		+13,733		+13,383		+30,678			
	68.0539	29.3298	**68.7998**	29.3944	67.3207	28.6377	**68.4570**	29.5131	68.5819	**29.5221**	67.9172	28.5058	세션 수 : 31,118	
	(-0.0997)	(-0.2502)	**(+0.6462)**	(-0.1856)	(-0.8329)	(-0.9423)	**(+0.3034)**	(-0.0669)	(+0.5813)	**(+0.0988)**	(-0.2363)	(-0.1742)		
tanimoto	+12,590		+13,810		+13,846		+13,131		+13,383		+30,678			
	67.9744	29.5244	**68.9309**	29.5159	67.2969	28.8584	**68.2839**	29.5375	**68.6610**	**29.5629**	67.8823	28.5401	68.1536	29.5800
	(-0.1791)	(-0.0556)	**(+0.7773)**	(-0.0641)	(-0.8567)	(-0.7216)	**(+0.1303)**	(-0.0425)	**(+0.6604)**	**(+0.1396)**	(-0.2713)	(-1.0399)		
cosine	+12,317		+14,200		+14,034		+12,306		+13,383		+30,678			
	68.1319	29.3520	**68.8592**	29.3668	67.3480	28.4790	68.1361	**29.5880**	**68.6260**	**29.5330**	67.9752	28.5806		
	(-0.0217)	(-0.2280)	**(+0.7056)**	(-0.2133)	(-0.8056)	(-1.1010)	(-0.0175)	**(+0.0079)**	**(+0.6254)**	**(+0.1097)**	(-0.1783)	(-0.9994)		
w2v	+14,583		+15,421		+13,209		+13,894		+13,383		+31,118			
	67.6050	28.6032	67.1673	28.9144	67.6292	28.4863	67.7037	29.2002	67.8878	29.3391	65.8964	27.6220		
	(-0.5486)	(-0.9768)	(-0.9863)	(-0.6656)	(-0.5244)	(-1.0937)	(-0.4499)	(-0.3798)	(-0.1129)	(-0.0842)	(-2.2572)	(-1.9580)		

Table 4.10 Result of Yoochoose 1/256 - SR-GNN

co-occurrence	+11,661		+15,383		+14,518		+13,380		+13,383		+30,678			
	66.9503	**29.0840**	**69.5623**	**30.1207**	66.6697	28.5740	**68.4031**	**29.7022**	**67.4812**	**29.0565**	**67.8641**	**29.0832**		
	(-0.2219)	**(+0.2638)**	**(+2.3901)**	**(+1.3005)**	(-0.5025)	(-0.2462)	**(+1.2309)**	**(+0.8820)**	**(+0.3090)**	**(+0.2363)**	**(+0.6919)**	**(+0.2630)**		
PMI	+8,103		+19,603		+16,607		+4,492		+13,383		+30,678			
	66.9282	28.7302	**68.8092**	**29.8337**	66.6217	**29.5855**	**67.3749**	**29.2475**	**67.8135**	**28.8513**	**68.2166**	**30.3724**		
	(-0.2440)	(-0.0901)	**(+1.6370)**	**(+1.0134)**	(-0.5505)	**(+0.7653)**	**(+0.2027)**	**(+0.4273)**	**(+0.6413)**	**(+0.0311)**	**(+1.0444)**	**(+1.5522)**		
jaccard	+11,648		+14,152		+15,035		+13,733		+13,383		+30,678			
	67.1663	**29.0599**	**69.6380**	**30.2208**	66.7620	**28.8722**	**68.5249**	**29.7399**	67.1538	**29.0561**	**29.1469**	**29.1469**	세션 수 : 31,118	
	(-0.0059)	**(+0.2397)**	**(+2.4658)**	**(+1.4006)**	(-0.4102)	**(+0.0520)**	**(+1.3527)**	**(+0.9196)**	(-0.0184)	**(+0.2359)**	**(+0.3267)**	**(+0.3267)**		
tanimoto	+12,590		+13,810		+13,846		+13,131		+13,383		+30,678			
	67.4838	**29.3617**	**69.1581**	**29.8689**	66.5091	**68.4215**	**29.7886**	**29.7886**	**67.6482**	**29.0874**	**67.8068**	**29.1310**	67.1722	28.8202
	(+0.3116)	**(+0.5414)**	**(+1.9859)**	**(+1.0487)**	(-0.6631)	**(+1.2493)**	**(+0.9684)**	**(+0.9684)**	**(+0.4760)**	**(+0.2672)**	**(+0.6346)**	**(+0.3107)**		
cosine	+12,317		+14,200		+14,034		+12,306		+13,383		+30,678			
	67.5115	**29.5313**	**69.5919**	66.4205	66.4205	28.5396	**68.0763**	**29.6941**	**67.3655**	**29.1873**	**67.6407**	**29.2179**		
	(+0.3393)	**(+0.7110)**	**(+2.4197)**	(-0.7517)	(-0.7517)	(-0.2806)	**(+0.9041)**	**(+0.8739)**	**(+0.1933)**	**(+0.3671)**	**(+0.4685)**	**(+0.3977)**		
w2v	+14,583		+15,421		+13,209		+13,894		+13,383		+31,118			
	66.7067	28.5588	66.9762	**28.9149**	66.9448	28.4125	**67.3029**	**29.3163**	66.7486	29.8483	65.3037	27.5062		
	(-0.4655)	(-0.2614)	(-0.1960)	**(+0.0947)**	(-0.2274)	(-0.4077)	**(+0.1307)**	**(+0.4960)**	(-0.4236)	(+1.0281)	(-1.8685)	(-1.3140)		

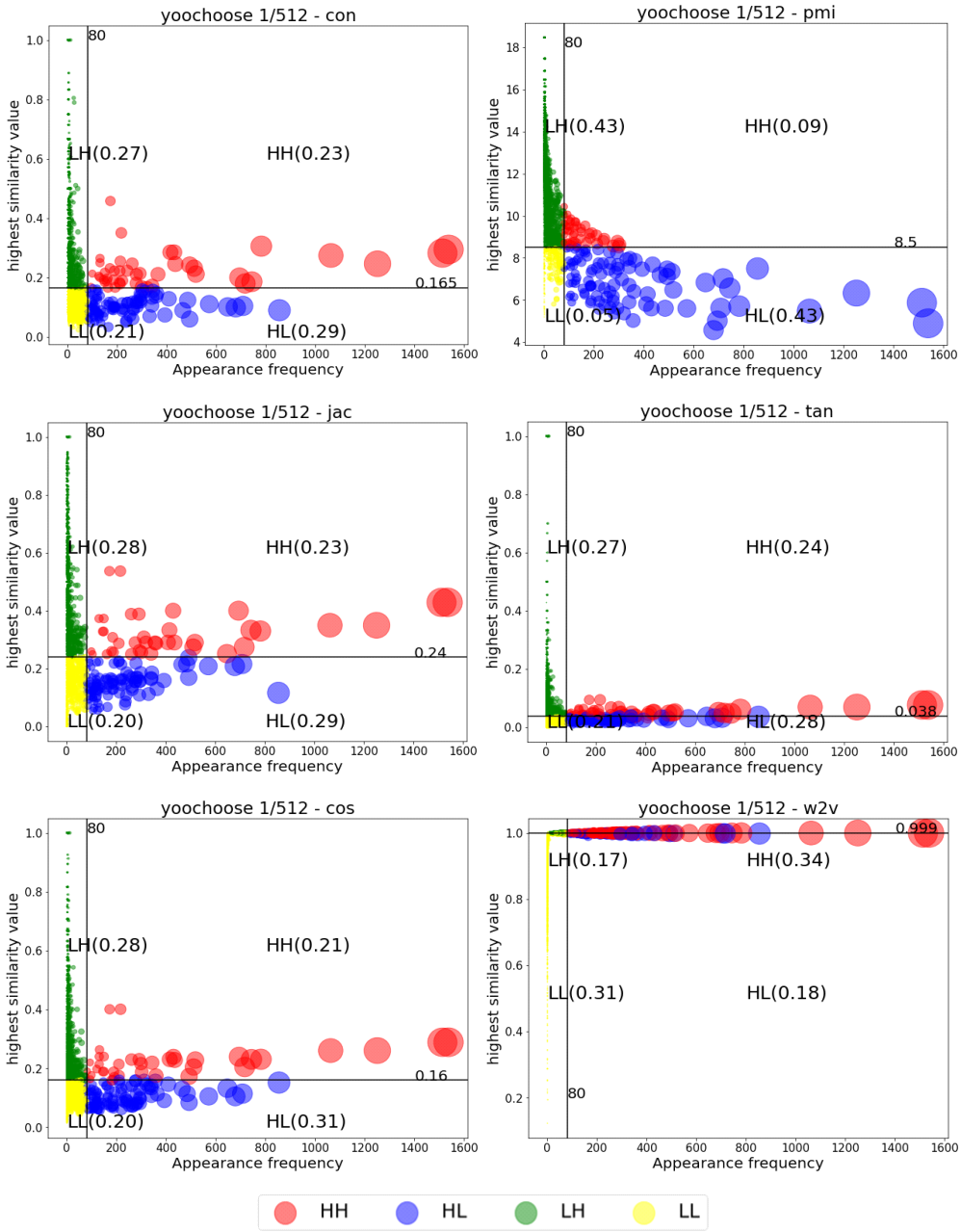

Fig. 4.7 Yoochoose 1/512의 유사도 지표별 출현 빈도와 최고 유사도 값 산점도

Table 4.11 Result of Yoochoose 1/512 - NARM

유사도 지표	HH		LH		HL		LL		random		all		original	
	Recall @20	MRR @20	Recall @20	MRR @20	Recall @20	MRR @20	Recall @20	MRR @20	Recall @20	MRR @20	Recall @20	MRR @20	Recall @20	MRR @20
co-occurrence	+6,722 65.5078 (-0.0527)	28.7473 (-0.0814)	+6,462 **67.0420** **(+1.4814)**	**28.8559** **(+0.0271)**	+7,497 65.3761 (-0.1845)	27.6440 (-1.1848)	+6,293 **66.4513** **(+0.8907)**	28.7347 (-0.0941)	+6,901 **66.2825** **(+0.6906)**	28.6027 (-0.2495)	+15,319 **66.7632** **(+1.2027)**	28.1361 (-0.6926)		
PMI	+3,915 65.1542 (-0.4063)	28.5494 (-0.2793)	+8,754 **66.2759** **(+0.7153)**	28.7726 (-0.0561)	+8,947 65.4795 (-0.0811)	28.7238 (-0.1049)	+2,126 65.4339 (-0.1267)	28.8258 (-0.0029)	+6,901 **65.7287** **(+0.1368)**	28.7837 (-0.0685)	+15,319 **66.7547** **(+1.1941)**	**28.8559** **(+0.0271)**		
jaccard	+6,237 **65.9156** **(+0.3550)**	**28.8295** **(+0.0007)**	+5,948 **67.0562** **(+1.4956)**	**28.9628** **(+0.1341)**	+8,207 65.2519 (-0.3087)	28.0218 (-0.8069)	+6,621 **66.4769** **(+0.9163)**	**28.8609** **(+0.0322)**	+6,901 **66.6086** **(+1.0167)**	28.9539 (+0.1018)	+15,319 **66.7869** **(+1.2264)**	28.0489 (-0.7799)	세션 수 : 15,559	
tanimoto	+6,788 **66.0559** **(+0.4953)**	**28.8866** **(+0.0579)**	+6,193 **67.0400** **(+1.4795)**	**29.1786** **(+0.3499)**	+7,470 65.4747 (-0.0858)	28.1452 (-0.6835)	+6,104 **66.3981** **(+0.8375)**	**29.0272** **(+0.1985)**	+6,901 **66.5518** **(+0.9598)**	28.8000 (-0.0522)	+15,319 **66.6077** **(+1.0472)**	28.2561 (-0.5727)	65.5606	28.8288
cosine	+6,164 **65.8890** **(+0.3284)**	28.7953 (-0.0334)	+6,228 **66.9746** **(+1.4141)**	**28.8897** **(+0.0610)**	+7,722 65.3164 (-0.2442)	27.9302 (-0.8985)	+5,829 **66.1099** **(+0.5493)**	28.8022 (-0.0266)	+6,901 **66.5602** **(+0.9683)**	**28.9671** **(+0.1149)**	+15,319 **66.3223** **(+0.7617)**	28.0476 (-0.7811)		
w2v	+8,048 65.3420 (-0.2186)	28.1700 (-0.6588)	+5,102 65.2746 (-0.2860)	28.3048 (-0.5240)	+6,662 65.4681 (-0.0925)	28.0510 (-0.7777)	+8,083 64.7190 (-0.8416)	28.4576 (-0.3711)	+6,901 **65.6197** **(+0.0278)**	28.4669 (-0.3853)	+15,559 64.0070 (-1.5536)	27.3073 (-1.5214)		

Table 4.12 Result of Yoochoose 1/512 - SR-GNN

co-occurrence	+6,722 **64.1932** **(+2.8852)**	**27.3818** **(+2.1609)**	+6,462 **67.5348** **(+6.2268)**	**29.0550** **(+3.8341)**	+7,497 **63.0743** **(+1.7663)**	**26.0106** **(+0.7896)**	+6,293 **64.8539** **(+3.5459)**	**27.3599** **(+2.1389)**	+6,901 **63.8460** **(+2.5380)**	**27.8433** **(+2.6223)**	+15,319 **65.8392** **(+4.5312)**	**28.2462** **(+3.0252)**		
PMI	+3,915 **62.5759** **(+1.2679)**	**26.0681** **(+0.8471)**	+8,754 **66.6278** **(+5.3198)**	**28.9362** **(+3.7153)**	+8,947 **62.9845** **(+1.6765)**	**27.2315** **(+2.0105)**	+2,126 **61.7052** **(+0.3972)**	**25.4808** **(+0.2598)**	+6,901 **63.4810** **(+2.173)**	**26.1008** **(+0.8798)**	+15,319 **66.0225** **(+4.7145)**	**29.3868** **(+4.1658)**		
jaccard	+6,237 **63.5784** **(+2.2704)**	**27.0507** **(+1.8297)**	+5,948 **66.9582** **(+5.6502)**	**28.9433** **(+3.7223)**	+8,207 **63.5707** **(+2.2627)**	**26.9292** **(+1.7082)**	+6,621 **65.7514** **(+4.4434)**	**28.1479** **(+2.9270)**	+6,901 **62.9518** **(+1.6438)**	**27.0843** **(+1.8633)**	+15,319 **66.0817** **(+4.7737)**	**28.8514** **(+3.6304)**	세션 수 : 15,559	
tanimoto	+6,788 **62.9903** **(+1.6823)**	**26.7570** **(+1.5360)**	+6,193 **67.3878** **(+6.0798)**	**29.1931** **(+3.9721)**	+7,470 **62.5664** **(+1.2584)**	**26.3071** **(+1.0861)**	+6,104 **65.8908** **(+4.5828)**	**28.3165** **(+3.0955)**	+6,901 **63.8381** **(+2.5301)**	**26.4812** **(+1.2602)**	+15,319 **66.5801** **(+5.2721)**	**28.8655** **(+3.6445)**	61.3080	25.221
cosine	+6,164 **62.8490** **(+1.5410)**	**26.5198** **(+1.2988)**	+6,228 **66.8875** **(+5.5795)**	**28.8537** **(+3.6327)**	+7,722 **62.9597** **(+1.6517)**	**26.3622** **(+1.1412)**	+5,829 **64.9169** **(+3.6089)**	**27.9378** **(+2.7169)**	+6,901 **62.8541** **(+1.5461)**	**26.8101** **(+1.5891)**	+15,319 **66.2268** **(+4.9188)**	**28.7864** **(+3.5654)**		
w2v	+8,048 **61.9782** **(+0.6702)**	**25.7715** **(+0.5506)**	+5,102 **63.5287** **(+2.2207)**	**26.9386** **(+1.7176)**	+6,662 **62.7917** **(+1.4837)**	**26.0820** **(+0.8611)**	+8,083 **63.5574** **(+2.2494)**	**27.3056** **(+2.0847)**	+6,901 **61.7786** **(+0.4706)**	**26.0872** **(+0.8662)**	+15,559 **61.6975** **(+0.3895)**	**25.9875** **(+0.7665)**		

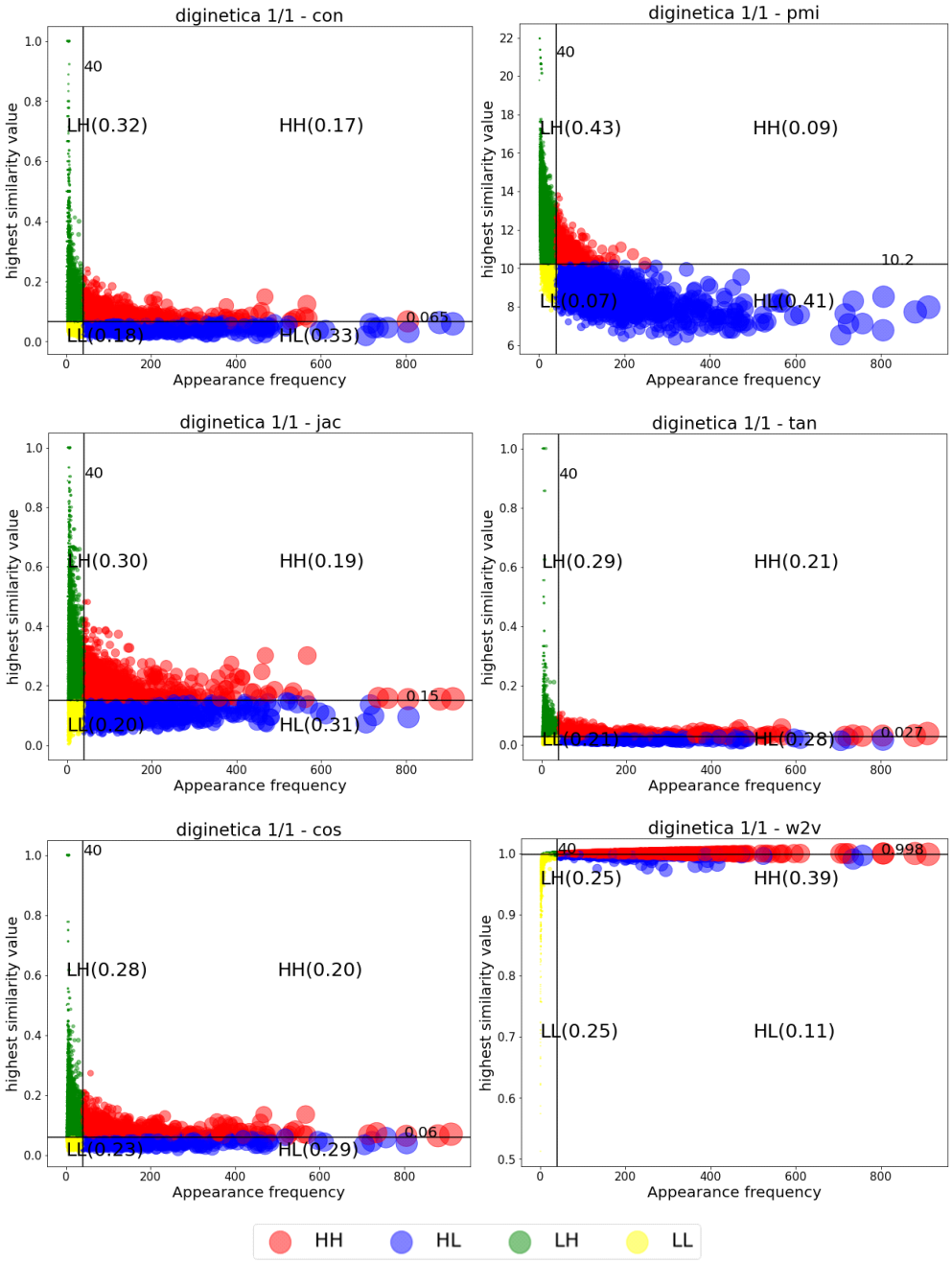

Fig. 4.8 Diginetica의 유사도 지표별 출현 빈도와 최고 유사도 값 산점도

Table 4.13 Result of Diginetica - NARM

유사도 지표	HH		LH		HL		LL		random		all		original	
	Recall @20	MRR @20	Recall @20	MRR @20	Recall @20	MRR @20	Recall @20	MRR @20	Recall @20	MRR @20	Recall @20	MRR @20	Recall @20	MRR @20
co-occurrence	+78,607 52.6643 (-1.1336)	 17.6853 (-0.8892)	+114,822 52.9551 (-0.8428)	 18.1273 (-0.4471)	+115,180 51.4857 (-2.3122)	 17.0980 (-1.4764)	+82,020 52.6857 (-1.1122)	 17.9198 (-0.6547)	+97,842 52.8084 (-0.9895)	 17.1631 (-1.4114)	+186,628 50.2305 (-3.5674)	 16.2989 (-2.2755)		
PMI	+50,753 52.4362 (-1.3617)	 17.9747 (-0.5998)	+128,816 52.5144 (-1.2835)	 17.9962 (-0.5783)	+121,459 50.4484 (-3.3495)	 17.2888 (-1.2856)	+38,170 52.5443 (-1.2536)	 18.2668 (-0.3076)	+97,842 52.0884 (-1.7095)	 17.1366 (-1.4379)	+186,628 50.8581 (-2.9397)	 17.0199 (-1.5546)		
jaccard	+83,371 52.3053 (-1.4926)	 17.5825 (-0.9919)	+99,461 53.3344 (-0.4635)	 18.1468 (-0.4276)	+106,208 51.5765 (-2.2214)	 17.2701 (-1.3044)	+94,339 52.5932 (-1.2047)	 18.0749 (-0.4996)	+97,842 52.0843 (-1.7136)	 17.6346 (-0.9399)	+186,628 50.9569 (-2.8410)	 16.5881 (-1.9863)	세션 수 : 186,670	
tanimoto	+92,029 52.2724 (-1.5255)	 17.4747 (-1.0998)	+98,579 53.2578 (-0.5401)	 18.2115 (-0.3630)	+97,285 51.6412 (-2.1567)	 17.2992 (-1.2752)	+88,331 52.5012 (-1.2966)	 17.9607 (-0.6138)	+97,842 52.4833 (-1.3146)	 17.7832 (-0.7913)	+186,628 50.7394 (-3.0585)	 16.6056 (-1.9689)	53.7979	18.5745
cosine	+90,678 52.2658 (-1.5321)	 17.5294 (-1.0451)	+95,804 53.3432 (-0.4547)	 18.0649 (-0.5096)	+99,758 51.5274 (-2.2705)	 17.3777 (-1.1967)	+95,253 52.4403 (-1.3576)	 18.0050 (-0.5695)	+97,842 52.1853 (-1.6126)	 17.6483 (-0.9262)	+186,628 50.6347 (-3.1632)	 16.7443 (-1.8302)		
w2v	+127,020 49.5321 (-4.2658)	 16.4322 (-2.1423)	+95,883 51.1356 (-2.6623)	 17.8316 (-0.7428)	+52,520 52.3176 (-1.4803)	 17.7588 (-0.8157)	+106,696 51.3941 (-2.4038)	 18.0824 (-0.4921)	+97,842 50.4355 (-3.3624)	 16.6972 (-1.8773)	+186,670 46.3665 (-7.4313)	 15.5713 (-3.0031)		

Table 4.14 Result of Diginetica - SR-GNN

co-occurrence	+78,607 54.3138 (-1.1693)	 18.5462 (-0.6220)	+114,822 55.2138 (-0.2693)	 18.9683 (-0.1999)	+115,180 54.9654 (-0.5177)	 18.1048 (-1.0634)	+82,020 53.8138 (-1.6693)	 18.8621 (-0.3061)	+97,842 54.8432 (-0.6399)	 18.2658 (-0.9024)	+186,628 54.8495 (-0.6336)	 18.6843 (-0.4839)		
PMI	+50,753 54.2788 (-1.2043)	 18.1768 (-0.9914)	+128,816 55.4158 (-0.0673)	 18.8782 (-0.2900)	+121,459 54.7832 (-0.6999)	 18.1245 (-1.0437)	+38,170 53.7328 (-1.7503)	 18.1087 (-1.0595)	+97,842 54.7832 (-0.6999)	 18.4567 (-0.7115)	+186,628 53.8640 (-1.6191)	 18.7452 (-0.4230)		
jaccard	+83,371 54.7861 (-0.6970)	 17.8531 (-1.3151)	+99,461 55.3841 (-0.0990)	 18.9088 (-0.2594)	+106,208 54.5677 (-0.9154)	 17.8655 (-1.3027)	+94,339 53.1057 (-2.3774)	 18.5207 (-0.6475)	+97,842 54.6710 (-0.8121)	 18.7754 (-0.3928)	+186,628 54.6710 (-0.8121)	 18.2757 (-0.8925)	세션 수 : 186,670	
tanimoto	+92,029 54.6488 (-0.8343)	 17.3883 (-1.7799)	+98,579 55.4318 (-0.0513)	 18.4668 (-0.7014)	+97,285 54.8432 (-0.6399)	 17.8438 (-1.3244)	+88,331 53.4301 (-2.053)	 18.1584 (-1.0098)	+97,842 53.4738 (-2.0093)	 18.5664 (-0.6018)	+186,628 53.4304 (-2.0527)	 18.1384 (-1.0298)	55.4831	19.1682
cosine	+90,678 54.8348 (-0.6483)	 17.8468 (-1.3214)	+95,804 54.8932 (-0.5899)	 18.5238 (-0.6444)	+99,758 53.8651 (-1.6180)	 17.1885 (-1.9797)	+95,253 53.4755 (-2.0076)	 17.3884 (-1.7798)	+97,842 53.3865 (-2.0966)	 18.5747 (-0.5935)	+186,628 52.4198 (-3.0633)	 18.1257 (-1.0425)		
w2v	+127,020 52.3642 (-3.1189)	 15.4205 (-3.7477)	+95,883 53.4288 (-2.0543)	 16.7105 (-2.4577)	+52,520 51.2277 (-4.2554)	 15.6007 (-3.5675)	+106,696 52.5530 (-2.9301)	 15.7058 (-3.6675)	+97,842 52.0785 (-3.4046)	 15.5007 (-3.6675)	+186,670 51.7804 (-3.7027)	 14.2251 (-4.9431)		

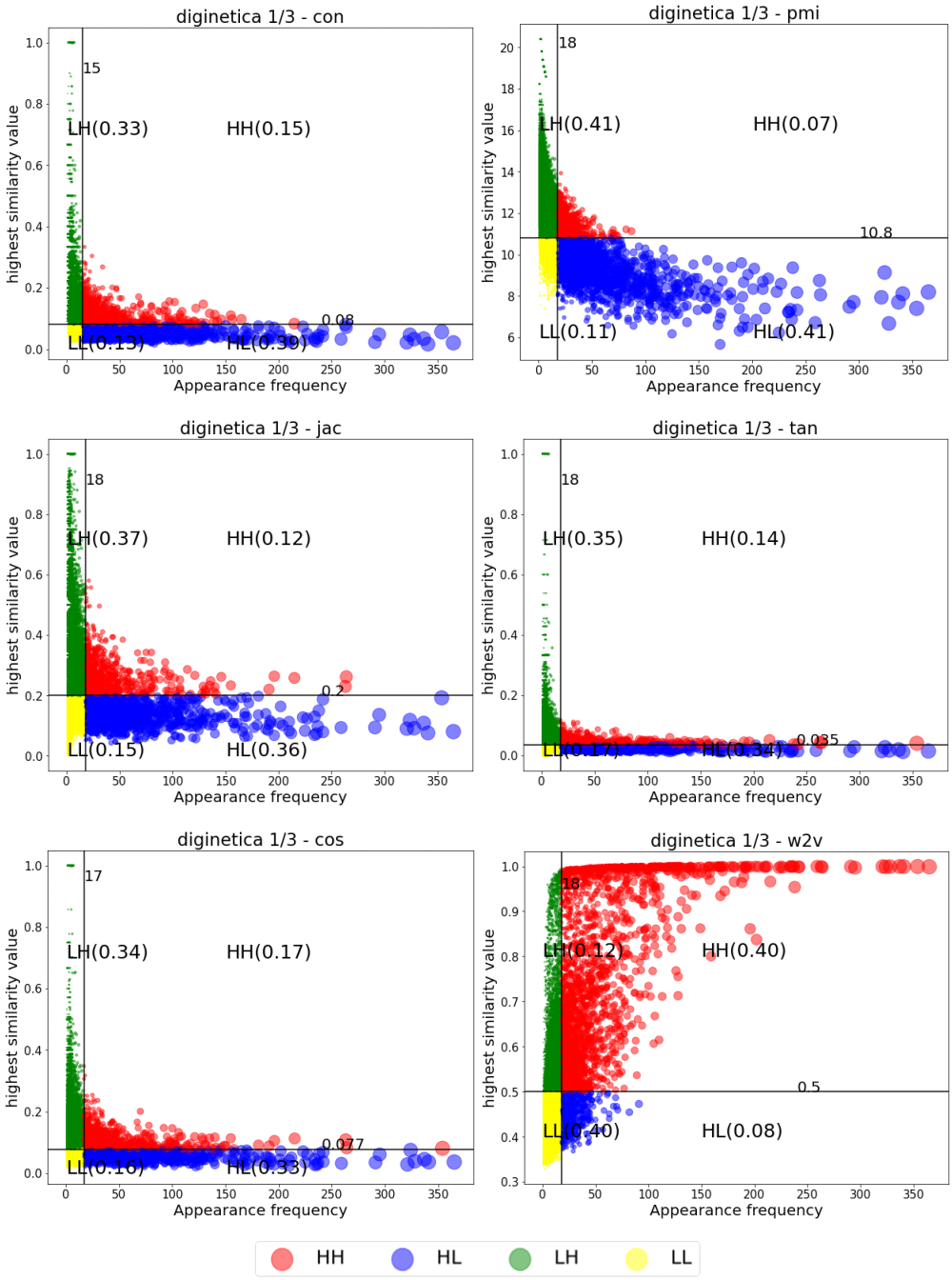

Fig. 4.9 Diginetica 1/3의 유사도 지표별 출현 빈도와 최고 유사도 값 산점도

Table 4.15 Result of Diginetica 1/3 - NARM

유사도 지표	HH		LH		HL		LL		random		all		original	
	Recall @20	MRR @20	Recall @20	MRR @20	Recall @20	MRR @20	Recall @20	MRR @20	Recall @20	MRR @20	Recall @20	MRR @20	Recall @20	MRR @20
co-occurrence	+23,200 50.1765 (-0.6163)	 17.1412 (-0.7302)	+39,010 **51.0868** **(+0.2940)**	 17.6011 (-0.2704)	+41,249 49.4904 (-1.3024)	 16.6020 (-1.2694)	+22,702 50.3164 (-0.4764)	 17.3294 (-0.5420)	+32,944 50.0436 (-0.7492)	 17.8257 (-0.0458)	+62,090 50.4290 (-0.3638)	 16.7927 (-1.0787)		
PMI	+14,434 49.8663 (-0.9265)	 17.3407 (-0.5307)	+41,072 50.6521 (-0.1407)	 17.5953 (-0.2761)	+39,905 48.6973 (-2.0955)	 16.9996 (-0.8719)	+18,064 49.6127 (-1.1801)	 17.4373 (-0.4341)	+32,944 49.8561 (-0.9367)	 17.4542 (-0.4173)	+62,090 50.0888 (-0.7040)	 17.4699 (-0.4015)		
jaccard	+19,688 50.2082 (-0.5846)	 17.1515 (-0.7200)	+36,914 **51.1271** **(+0.3343)**	 17.6041 (-0.2674)	+37,523 49.4892 (-1.3036)	 16.8960 (-0.9755)	+27,061 49.9994 (-0.7934)	 17.3183 (-0.5532)	+32,944 49.2834 (-1.5094)	 17.4855 (-0.3860)	+62,090 49.9009 (-0.8919)	 16.7775 (-1.0940)	세션 수 : 62,223	
tanimoto	+22,103 50.3271 (-0.4657)	 17.1792 (-0.6922)	+36,450 **51.1571** **(+0.3643)**	 17.7141 (-0.1574)	+35,423 49.3567 (-1.4361)	 16.7450 (-1.1265)	+25,818 50.1601 (-0.6327)	 17.4601 (-0.4114)	+32,944 49.7544 (-1.0384)	 17.8544 (-0.0171)	+62,090 49.9427 (-0.8500)	 16.6466 (-1.2249)	50.7928	17.8715
cosine	+26,812 50.0242 (-0.7686)	 17.1341 (-0.7374)	+35,909 **50.9731** **(+0.1803)**	 17.6028 (-0.2687)	+34,406 49.2634 (-1.5294)	 16.8629 (-1.0086)	+25,667 49.9750 (-0.8178)	 17.3995 (-0.4720)	+32,944 50.1514 (-0.6414)	 17.2607 (-0.6108)	+62,090 49.3529 (-1.4399)	 16.5912 (-1.2803)		
w2v	+41,174 47.7954 (-2.9974)	 16.3388 (-1.5327)	+21,091 48.7511 (-2.0417)	 17.1912 (-0.6802)	+14,664 49.1212 (-1.6716)	 17.3559 (-0.5156)	+42,604 46.9825 (-3.8103)	 17.1747 (-0.6968)	+32,944 47.3677 (-3.4251)	 16.0423 (-1.8292)	+62,223 44.9847 (-5.8081)	 16.2498 (-1.6217)		

Table 4.16 Result of Diginetica 1/3 - SR-GNN

co-occurrence	+23,200 50.7552 (-0.5931)	 17.1838 (-0.9542)	+39,010 51.1897 (-0.1586)	 17.9486 (-0.1894)	+41,249 50.4894 (-0.8589)	 17.8684 (-0.2696)	+22,702 50.3085 (-1.0398)	 17.2108 (-0.9272)	+32,944 50.1458 (-1.2025)	 17.0082 (-1.1298)	+62,090 49.4832 (-1.8651)	 17.6621 (-0.4759)		
PMI	+14,434 50.1832 (-1.1651)	 17.0583 (-1.0797)	+41,072 50.1783 (-1.1700)	 17.0844 (-1.0536)	+39,905 50.1868 (-1.1615)	 16.4838 (-1.6542)	+18,064 50.0452 (-1.3031)	 17.1108 (-1.0272)	+32,944 50.2201 (-1.1282)	 16.8748 (-1.2632)	+62,090 50.1865 (-1.1618)	 16.1404 (-1.9976)		
jaccard	+19,688 50.1831 (-1.1652)	 17.1831 (-0.9549)	+36,914 **51.6366** **(+0.2883)**	 18.0628 (-0.0752)	+37,523 50.1458 (-1.2025)	 17.1590 (-0.9790)	+27,061 50.1185 (-1.2298)	 17.0381 (-1.0999)	+32,944 50.2151 (-1.1332)	 17.2682 (-0.8698)	+62,090 50.1113 (-1.2370)	 16.7583 (-1.3797)	세션 수 : 62,223	
tanimoto	+22,103 50.0486 (-1.2997)	 17.0868 (-1.0512)	+36,450 **51.5878** **(+0.2395)**	 **18.2218** **(+0.0838)**	+35,423 50.1474 (-1.2009)	 17.1832 (-0.9548)	+25,818 50.2698 (-1.0785)	 17.0065 (-1.1315)	+32,944 50.0047 (-1.3436)	 17.2023 (-0.9357)	+62,090 50.4835 (-0.8648)	 16.7854 (-1.3526)	51.3483	18.1380
cosine	+26,812 50.4834 (-0.8649)	 17.2845 (-0.8535)	+35,909 51.2678 (-0.0805)	 17.9905 (-0.1475)	+34,406 50.3786 (-0.9697)	 17.1583 (-0.9797)	+25,667 50.7832 (-0.5651)	 17.3815 (-0.7565)	+32,944 50.3512 (-0.9971)	 17.1833 (-0.9547)	+62,090 50.1583 (-1.1900)	 17.4032 (-0.7348)		
w2v	+41,174 48.4831 (-2.8652)	 16.1450 (-1.9930)	+21,091 49.1821 (-2.1662)	 16.7248 (-1.4132)	+14,664 47.1588 (-4.1895)	 15.1831 (-2.9549)	+42,604 47.6852 (-3.6631)	 15.7838 (-2.3542)	+32,944 47.8981 (-3.4502)	 15.8765 (-2.2615)	+62,223 46.7485 (-4.5998)	 14.8954 (-3.2426)		

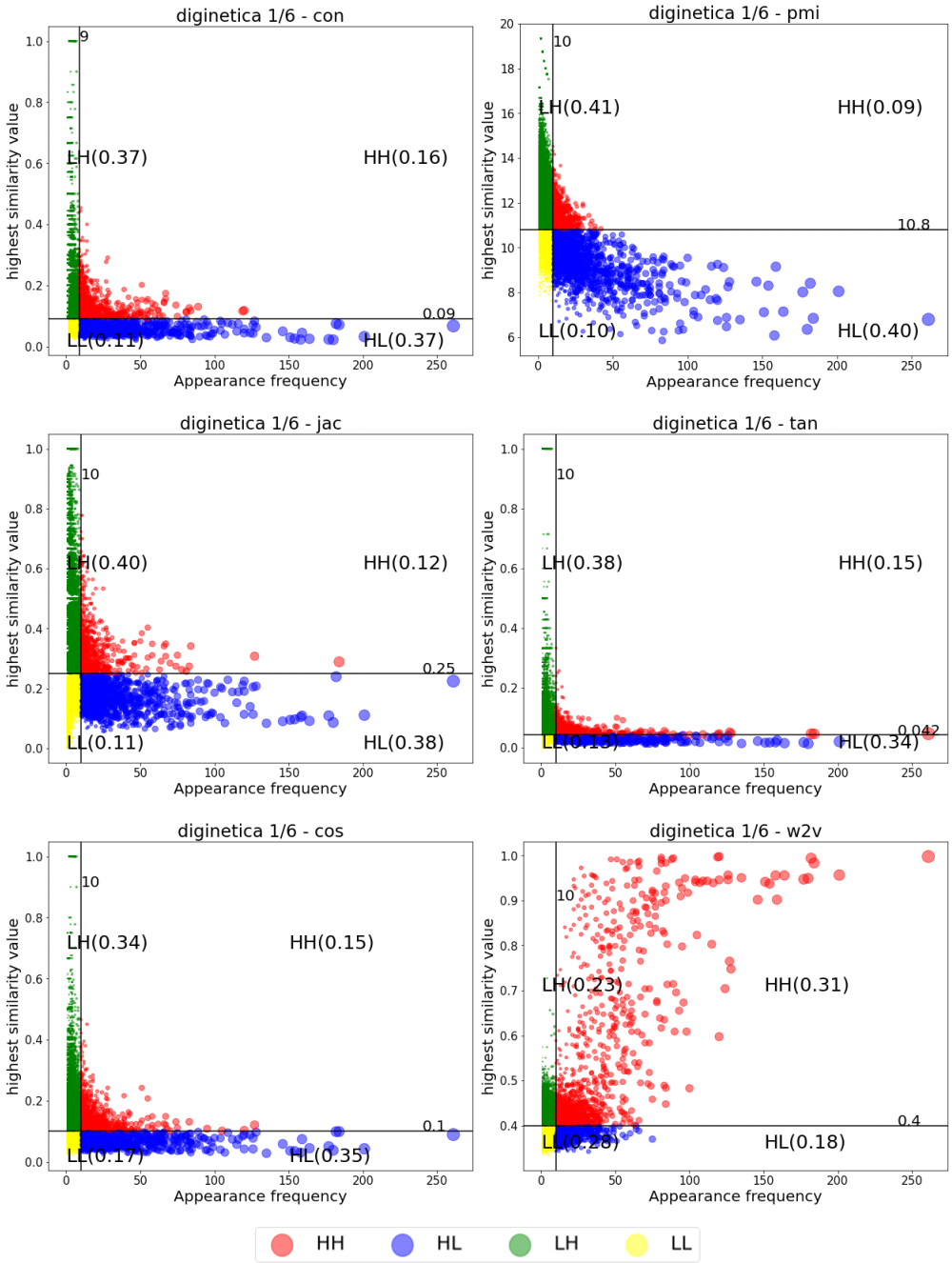

Fig. 4.10 Diginetica 1/6의 유사도 지표별 출현 빈도와 최고 유사도 값 산점도

Table 4.17 Result of Diginetica 1/6 - NARM

유사도 지표	HH		LH		HL		LL		random		all		original	
	Recall @20	MRR @20	Recall @20	MRR @20	Recall @20	MRR @20	Recall @20	MRR @20	Recall @20	MRR @20	Recall @20	MRR @20	Recall @20	MRR @20
co-occurrence	+12,190 47.5139 (+0.6110)	16.7192 (-0.4887)	+20,610 48.7312 (+1.8282)	17.2956 (+0.0877)	+19,933 47.4662 (+0.5633)	16.1791 (-1.0287)	+9,560 47.4174 (+0.5144)	16.8642 (-0.3437)	+16,108 47.9246 (+1.0216)	16.8975 (-0.3104)	+30,962 48.8047 (+1.9017)	16.5718 (-0.6360)		
PMI	+8,676 47.1141 (+0.2112)	16.7924 (-0.4154)	+20,192 48.6863 (+1.7833)	17.1948 (-0.0130)	+19,177 45.6393 (-1.2636)	16.3531 (-0.8547)	+8,426 46.8186 (-0.0844)	16.8047 (-0.4032)	+16,108 47.4202 (+0.5172)	16.8603 (-0.3476)	+30,962 47.1721 (+0.2692)	16.8049 (-0.4029)		
jaccard	+8,981 47.2984 (+0.3954)	16.6527 (-0.5552)	+18,949 48.7892 (+1.8863)	17.1338 (-0.0740)	+19,200 47.2776 (+0.3746)	16.4492 (-0.7587)	+11,298 47.3974 (+0.4944)	17.0816 (-0.1262)	+16,108 47.8120 (+0.9090)	16.7912 (-0.4166)	+30,962 49.0382 (+2.1353)	16.6565 (-0.5514)	세션 수 : 31,111	
tanimoto	+11,304 47.3373 (+0.4343)	16.7922 (-0.4156)	+19,066 48.9134 (+2.0104)	17.2069 (-0.0009)	+17,548 47.1316 (+0.2286)	16.3586 (-0.8493)	+10,604 47.4080 (+0.5050)	16.9060 (-0.3019)	+16,108 47.8610 (+0.9580)	16.8491 (-0.3587)	+30,962 48.8917 (+1.9888)	16.4110 (-0.7968)	46.9030	17.2078
cosine	+11,615 47.4165 (+0.5136)	16.6829 (-0.5250)	+17,814 48.6057 (+1.7027)	17.1690 (-0.0389)	+17,590 46.6442 (-0.2587)	16.4072 (-0.8006)	+12,717 47.4151 (+0.5121)	16.7786 (-0.4292)	+16,108 47.8935 (+0.9905)	16.8544 (-0.3534)	+30,962 48.0158 (+1.1128)	16.5015 (-0.7063)		
w2v	+17,580 44.7470 (-2.1559)	16.0092 (-1.1986)	+16,543 45.0449 (-1.8581)	16.6358 (-0.5720)	+13,881 44.8254 (-2.0776)	16.3160 (-0.8918)	+18,878 44.9732 (-1.9297)	16.7830 (-0.4248)	+16,108 45.9534 (-0.9495)	16.7210 (-0.4869)	+31,111 40.9925 (-5.9104)	15.7069 (-1.5009)		

Table 4.18 Result of Diginetica 1/6 - SR-GNN

co-occurrence	+12,190 46.4838 (+0.6477)	16.3824 (-0.5171)	+20,610 47.0568 (+1.2207)	16.7682 (-0.1313)	+19,933 46.1565 (+0.3204)	16.0148 (-0.8847)	+9,560 45.9384 (+0.1023)	16.0838 (-0.8157)	+16,108 46.1088 (+0.2727)	16.7821 (-0.1174)	+30,962 46.4852 (+0.6491)	16.7862 (-0.1133)		
PMI	+8,676 46.0782 (+0.2421)	16.3821 (-0.5174)	+20,192 47.4283 (+1.5922)	16.8381 (-0.0614)	+19,177 46.2820 (+0.4459)	15.9401 (-0.9594)	+8,426 45.8668 (+0.0307)	16.1587 (-0.7408)	+16,108 46.1807 (+0.3446)	16.2320 (-0.6675)	+30,962 46.8483 (+1.0122)	16.6818 (-0.2177)		
jaccard	+8,981 46.4821 (+0.6460)	16.2832 (-0.6163)	+18,949 47.4832 (+1.6471)	17.9107 (+1.0112)	+19,200 45.0252 (-0.8109)	16.1135 (-0.7860)	+11,298 45.4853 (-0.3508)	16.0505 (-0.849)	+16,108 46.6875 (+0.8514)	16.9788 (0.0793)	+30,962 46.7489 (+0.9128)	16.3467 (-0.5528)	세션 수 : 31,111	
tanimoto	+11,304 46.7836 (+0.9475)	16.1964 (-0.7031)	+19,066 47.6958 (+1.8597)	16.7828 (-0.1167)	+17,548 45.4628 (-0.3733)	16.0845 (-0.8150)	+10,604 45.6731 (-0.163)	15.8685 (-1.031)	+16,108 46.7486 (+0.9125)	16.4865 (-0.4130)	+30,962 46.4622 (+0.6261)	16.2085 (-0.6910)	45.8361	16.8995
cosine	+11,615 46.4836 (+0.6475)	16.3786 (-0.5209)	+17,814 47.4612 (+1.6251)	17.0716 (+0.1721)	+17,590 45.8486 (0.0125)	16.1820 (-0.7175)	+12,717 45.4863 (-0.3498)	16.1445 (-0.755)	+16,108 46.1047 (+0.2686)	16.3489 (-0.5506)	+30,962 46.1775 (+0.3414)	16.6478 (-0.2517)		
w2v	+17,580 44.8838 (-0.9523)	14.4833 (-2.4162)	+16,543 45.7824 (-0.0537)	14.7628 (-2.1367)	+13,881 44.8181 (-1.018)	13.3210 (-3.5785)	+18,878 44.0086 (-1.8275)	13.4568 (-3.4427)	+16,108 44.4862 (-1.3499)	14.4833 (-2.4162)	+31,111 42.6132 (-3.2229)	12.2208 (-4.6787)		

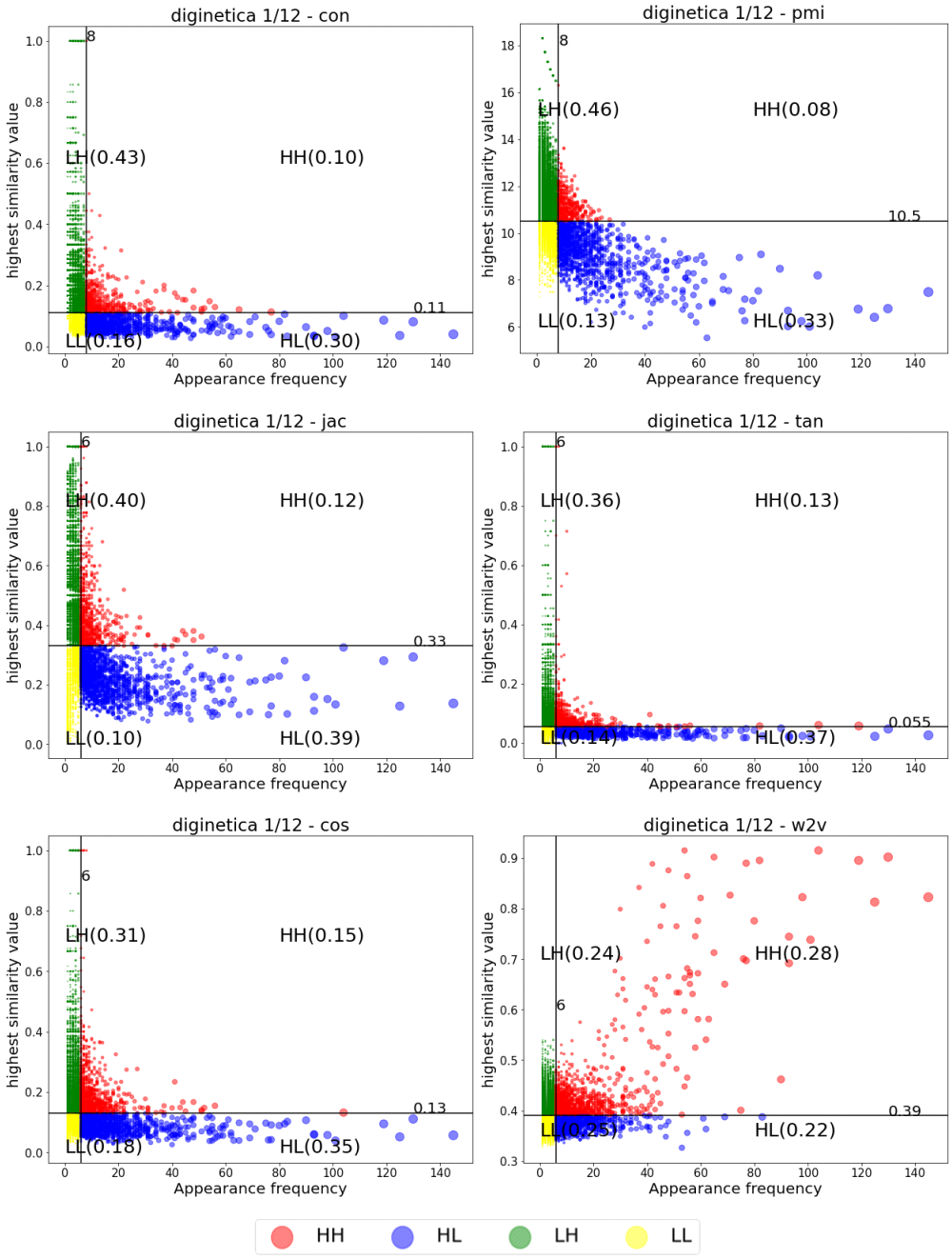

Fig. 4.11 Diginetica 1/12의 유사도 지표별 출현 빈도와 최고 유사도 값 산점도

Table 4.19 Result of Diginetica 1/12 - NARM

유사도 지표	HH		LH		HL		LL		random		all		original	
	Recall @20	MRR @20	Recall @20	MRR @20	Recall @20	MRR @20	Recall @20	MRR @20	Recall @20	MRR @20	Recall @20	MRR @20	Recall @20	MRR @20
co-occurrence	+4,044 42.4732 (+2.2925)	 16.5734 (+0.0182)	+11,228 45.9361 (+5.7554)	 17.3830 (+0.8277)	+8,654 43.5619 (+3.3812)	 16.1699 (-0.3853)	+6,219 44.2559 (+4.0752)	 16.7256 (+0.1704)	+7,866 44.3182 (+4.1375)	 16.4905 (-0.0647)	+15,411 47.0823 (+6.9016)	 16.6242 (+0.0690)		
PMI	+3,572 41.8173 (+1.6366)	 16.4984 (-0.0568)	+10,637 45.5649 (+5.3842)	 17.2122 (+0.6569)	+8,273 42.1082 (+1.9275)	 16.1665 (-0.3887)	+4,765 42.7225 (+2.5418)	 16.3780 (-0.1772)	+7,866 43.6159 (+3.4352)	 16.4379 (-0.1174)	+15,411 45.1016 (+4.9209)	 16.8043 (+0.2490)		
jaccard	+4,266 42.8376 (+2.6569)	 16.4026 (-0.1526)	+9,093 45.3301 (+5.1494)	 17.2390 (+0.6838)	+9,838 43.8805 (+3.6998)	 16.1472 (-0.4080)	+5,365 42.6401 (+2.4594)	 16.6488 (+0.0936)	+7,866 44.0162 (+3.8355)	 16.6256 (+0.0703)	+15,411 46.7811 (+6.6004)	 16.6483 (+0.0931)	세션 수 : 15,555	
tanimoto	+5,021 42.8687 (+2.6879)	 16.6629 (+0.1077)	+9,061 45.2345 (+5.0538)	 17.1468 (+0.5915)	+9,114 43.6845 (+3.5038)	 16.1245 (-0.4307)	+5,348 43.1761 (+2.9954)	 16.6112 (+0.0560)	+7,866 43.9131 (+3.7324)	 16.5480 (-0.0073)	+15,411 46.7908 (+6.6101)	 16.7213 (+0.1661)	40.1807	16.5552
cosine	+5,734 43.3630 (+3.1823)	 16.5010 (-0.0542)	+8,257 44.5766 (+4.3959)	 17.0673 (+0.5120)	+8,967 43.0557 (+2.8749)	 16.1555 (-0.3998)	+6,537 43.9990 (+3.8183)	 16.8706 (+0.3154)	+7,866 44.1167 (+3.9360)	 16.4987 (-0.0565)	+15,411 45.7042 (+5.5235)	 16.3666 (-0.1887)		
w2v	+8,693 40.9286 (+0.7479)	 15.9574 (-0.5979)	+8,632 40.8068 (+0.6261)	 16.4985 (-0.0568)	+7,958 40.2493 (+0.0686)	 16.1149 (-0.4403)	+8,850 40.9488 (+0.7681)	 16.3902 (-0.1650)	+7,866 41.3996 (+1.2189)	 16.2140 (-0.3413)	+15,555 37.5530 (-2.6277)	 16.0455 (-0.5097)		

Table 4.20 Result of Diginetica 1/12 - SR-GNN

co-occurrence	+4,044 41.8186 (+3.3868)	 16.1580 (+1.1708)	+11,228 43.8382 (+5.4064)	 17.8742 (+2.887)	+8,654 41.4862 (+3.0544)	 16.6680 (+1.6808)	+6,219 42.4032 (+3.9714)	 16.0838 (+1.0966)	+7,866 41.1078 (+2.6760)	 16.8952 (+1.9080)	+15,411 42.8969 (+4.4651)	 16.8992 (+1.9120)		
PMI	+3,572 40.0785 (+1.6467)	 16.0475 (+1.0603)	+10,637 43.6611 (+5.2293)	 16.9815 (+1.9943)	+8,273 42.1850 (+3.7532)	 16.8769 (+1.8897)	+4,765 42.6502 (+4.2184)	 16.8473 (+1.8601)	+7,866 40.8682 (+2.4364)	 16.9452 (+1.9580)	+15,411 44.3851 (+5.9533)	 16.0862 (+1.0990)		
jaccard	+4,266 41.6875 (+3.2557)	 16.8624 (+1.8752)	+9,093 44.0525 (+5.6207)	 17.4832 (+2.4960)	+9,838 42.1082 (+3.6764)	 16.1468 (+1.1596)	+5,365 42.3872 (+3.9554)	 16.8084 (+1.8212)	+7,866 40.6955 (+2.2637)	 16.8942 (+1.9070)	+15,411 43.9207 (+5.4889)	 16.0821 (+1.0949)	세션 수 : 15,555	
tanimoto	+5,021 41.8762 (+3.4444)	 16.6438 (+1.6566)	+9,061 44.1567 (+5.7249)	 18.0214 (+3.0342)	+9,114 41.9274 (+3.4956)	 16.8381 (+1.8509)	+5,348 42.7211 (+4.2893)	 16.8532 (+1.8660)	+7,866 40.4682 (+2.0364)	 16.0855 (+1.0983)	+15,411 44.0217 (+5.5899)	 16.9548 (+1.9676)	38.4318	14.9872
cosine	+5,734 42.2382 (+3.8064)	 16.4884 (+1.5012)	+8,257 43.9058 (+5.4740)	 17.4072 (+2.4200)	+8,967 41.6784 (+3.2466)	 16.0148 (+1.0276)	+6,537 42.9891 (+4.5573)	 16.0087 (+1.0215)	+7,866 40.7401 (+2.3083)	 16.0071 (+1.0199)	+15,411 43.6410 (+5.2092)	 16.1875 (+1.2003)		
w2v	+8,693 39.5612 (+1.1294)	 16.7892 (+1.8020)	+8,632 41.3631 (+2.9313)	 16.0526 (+1.0654)	+7,958 40.5220 (+2.0902)	 16.8321 (+1.8449)	+8,850 40.3132 (+1.8814)	 16.0874 (+1.1002)	+7,866 40.3356 (+1.9038)	 16.1431 (+1.1559)	+15,555 41.0087 (+2.5769)	 16.8072 (+1.8200)		

Ⅴ. 결 론

1. 연구 요약

본 연구에서는 세션 기반 추천시스템 활용 시 세션 데이터를 증강할 수 있는 기법을 제안하고 이를 이용한 추천 성능 향상이 가능함을 보였다. 데이터 증강 시 신규로 생성된 데이터는 원본 데이터와 유사하고 관측될 만한 데이터이어야 한다는 조건을 만족하기 위하여 일부 아이템을 가장 유사한 아이템으로 대체하거나 삽입하는 방식을 이용하였다. 또한 유저와 아이템 간의 상호작용 데이터 및 특징 정보 없이 아이템 시퀀스만으로 유사도를 측정하기 위하여 동시 발생 빈도 기반의 유사도 지표와 문맥 기반의 유사도 지표를 비교하였다. 이를 통해 다른 데이터와는 달리 추천시스템에서 활용되는 아이템 시퀀스 데이터에는 동시 발생 기반의 유사도 지표가 더욱 정확하다는 것을 보였다.

현실 세계에서는 가입자 및 방문자에 대한 정보를 얻기가 쉽지 않고, 이러한 경우 세션 기반 추천시스템이 활용될 것이므로 세션 데이터의 양이 부족한 상황을 다양한 데이터셋의 크기로 상응시켜 실험을 진행하였다. 약 60,000개 이상의 세션 데이터에서는 일부 선택 조건을 만족하는 아이템에 대하여 제안 증강 방식을 적용했을 때 성능이 향상되었고 그 이하의 데이터에 대해서는 증강된 데이터에 의한 성능 향상 폭이 원본 데이터의 증가에 따른 성능 향상 폭과 비슷해질 정도로 증강 효과가 나타났다.

또한 데이터셋으로부터 추출된 아이템 특성 정보에 기반하여 아이템을 분류하고 실험 결과를 비교함으로써 시퀀스 데이터 증강 시 무작위로 선택하는 것보다 아이템 특성 정보에 기반한 선택이 더욱 효과적이라는 것을 보였다. 또한 이를 통해 전체 아이템 중에서도 증강에 도움이 되는 아이템과 그렇지 않은 아이템이 존재한다는 사실을 확인하였다.

2. 한계점 및 추후 연구

본 연구에서 진행한 실험에 따르면 출현 빈도가 적고 확실히 유사한 아이템들 위주로 변형하여 신규 세션을 생성할 경우 가장 등장할 만한 데이터가 생성된다는 사실을 확인할 수 있다. 그러나 본 실험에서 진행한 방법은 파라미

터 설정에 있어서 매우 제한적인 범위만을 검증하였기 때문에 최고 유사도 값과 출현 빈도라는 정보를 사용했을 때 데이터 증강에 의한 성능 향상이 최대로 이루어지는 조건을 찾았는지에 대해서는 확답을 내릴 수 없다는 한계가 있다. 이는 아이템 특성 정보 추출부터 아이템 선택, 증강에 이르기까지의 과정이 정형적이지 않고 파라미터를 실험 데이터에 의존적인 방식으로 설정했기 때문에 드러난 한계점이라고 할 수 있다.

따라서 추후 연구로는 본 연구에서 밝힌 분석 결과를 바탕으로 전체적인 증강 과정을 더욱 정형적인 과정으로 일반화시켜 정의하는 것이 필요하다. 또한 각 클래스의 다양한 조합을 생성하여 증강 효과를 더욱 증가시킬 수도 있으므로, 데이터 증강 효과를 향상시키기 위해서는 더 많은 다양한 가설 설정과 검증이 필요하다.

참고문헌

[1] 천우진, & 강필성. (2020). 전이 확률 기반 벡터를 이용한 추천 시스템 성능 향상. 대한산업공학회지, 46(4), 393-403.

[2] Wang, S., Cao, L., Wang, Y., Sheng, Q. Z., Orgun, M. A., & Lian, D. (2021). A survey on session-based recommender systems. ACM Computing Surveys (CSUR), 54(7), 1-38.

[3] Wu, Z., Wang, S., Gu, J., Khabsa, M., Sun, F., & Ma, H. (2020). Clear: Contrastive learning for sentence representation. arXiv preprint arXiv:2012.15466.

[4] Tan, Y. K., Xu, X., & Liu, Y. (2016, September). Improved recurrent neural networks for session-based recommendations. In Proceedings of the 1st workshop on deep learning for recommender systems (pp. 17-22).

[5] Sharma, L., & Gera, A. (2013). A survey of recommendation system: Research challenges. International Journal of Engineering Trends and Technology (IJETT), 4(5), 1989-1992.

[6] Xu, M., Liu, F., & Xu, W. (2019, December). A Survey on Sequential Recommendation. In 2019 6th International Conference on Information Science and Control Engineering (ICISCE) (pp. 106-111). IEEE.

[7] Rendle, S., Freudenthaler, C., & Schmidt-Thieme, L. (2010, April). Factorizing personalized markov chains for next-basket recommendation. In Proceedings of the 19th international conference on World wide web (pp. 811-820).

[8] Chen, X., Xu, H., Zhang, Y., Tang, J., Cao, Y., Qin, Z., & Zha, H. (2018, February). Sequential recommendation with user memory networks. In Proceedings of the eleventh ACM international conference on web search and data mining (pp. 108-116).

[9] Kang, W. C., & McAuley, J. (2018, November). Self-attentive sequential recommendation. In 2018 IEEE International Conference on Data Mining (ICDM) (pp. 197-206). IEEE.

[10] Sun, F., Liu, J., Wu, J., Pei, C., Lin, X., Ou, W., & Jiang, P. (2019, November). BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In Proceedings of the 28th ACM international conference on information and knowledge management (pp. 1441-1450).

[11] He, R., & McAuley, J. (2016, December). Fusing similarity models with

markov chains for sparse sequential recommendation. In 2016 IEEE 16th International Conference on Data Mining (ICDM) (pp. 191-200). IEEE.
[12] Huang, J., Zhao, W. X., Dou, H., Wen, J. R., & Chang, E. Y. (2018, June). Improving sequential recommendation with knowledge-enhanced memory networks. In The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (pp. 505-514).
[13] Fang, H., Zhang, D., Shu, Y., & Guo, G. (2020). Deep learning for sequential recommendation: Algorithms, influential factors, and evaluations. ACM Transactions on Information Systems (TOIS), 39(1), 1-42.
[14] Hidasi, B., Karatzoglou, A., Baltrunas, L., & Tikk, D. (2015). Session-based recommendations with recurrent neural networks. arXiv preprint arXiv:1511.06939.
[15] 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).
[16] Wang, M., Ren, P., Mei, L., Chen, Z., Ma, J., & de Rijke, M. (2019, July). A collaborative session-based recommendation approach with parallel memory modules. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 345-354).
[17] Luo, A., Zhao, P., Liu, Y., Zhuang, F., Wang, D., Xu, J., ... & Sheng, V. S. (2020, July). Collaborative Self-Attention Network for Session-based Recommendation. In IJCAI (pp. 2591-2597).
[18] 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).
[19] Yu, F., Zhu, Y., Liu, Q., Wu, S., Wang, L., & Tan, T. (2020, July). TAGNN: Target attentive graph neural networks for session-based recommendation. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 1921-1924).
[20] Zheng, Y., Liu, S., Li, Z., & Wu, S. (2020, November). DGTN: Dual-channel Graph Transition Network for Session-based Recommendation. In 2020 International Conference on Data Mining Workshops (ICDMW) (pp. 236-242). IEEE.
[21] Hu, D., Wei, L., Zhou, W., Huai, X., Fang, Z., & Hu, S. (2021). PEN4Rec: Preference Evolution Networks for Session-based Recommendation. arXiv preprint

arXiv:2106.09306.
[22] Taylor, L., & Nitschke, G. (2018, November). Improving deep learning with generic data augmentation. In 2018 IEEE Symposium Series on Computational Intelligence (SSCI) (pp. 1542-1547). IEEE.
[23] 임장혁, 유기윤, & 김지영. (2018). 순환 신경망과 데이터 증강을 이용한 재난 정보 탐지 방법. 대한공간정보학회지, 26(4), 29-36.
[24] Liu, S., Lee, K., & Lee, I. (2020). Document-level multi-topic sentiment classification of email data with bilstm and data augmentation. Knowledge-Based Systems, 197, 105918.
[25] Rizos, G., Hemker, K., & Schuller, B. (2019, November). Augment to prevent: short-text data augmentation in deep learning for hate-speech classification. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management (pp. 991-1000).
[26] Wei, J., & Zou, K. (2019). Eda: Easy data augmentation techniques for boosting performance on text classification tasks. arXiv preprint arXiv:1901.11196.
[27] Liu, Z., Fan, Z., Wang, Y., & Yu, P. S. (2021). Augmenting Sequential Recommendation with Pseudo-Prior Items via Reversely Pre-training Transformer. arXiv preprint arXiv:2105.00522.
[28] Liu, Z., Chen, Y., Li, J., Yu, P. S., McAuley, J., & Xiong, C. (2021). Contrastive self-supervised sequential recommendation with robust augmentation. arXiv preprint arXiv:2108.06479.
[29] Socher, R., Bengio, Y., & Manning, C. (2012). Deep learning for NLP. Tutorial at Association of Computational Logistics (ACL).
[30] Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bull Soc Vaudoise Sci Nat, 37, 547-579.
[31] Leydesdorff, L. (2008). On the normalization and visualization of author co-citation data: Salton's Cosine versus the Jaccard index. Journal of the American Society for Information Science and Technology, 59(1), 77-85.
[32] Kadowaki, N., & Kishida, K. (2020). Empirical Comparison of Word Similarity Measures Based on Co-Occurrence, Context, and a Vector Space Model. Journal of Information Science Theory and Practice, 8(2), 6-17.

Abstract

A data augmentation method for session-based recommendation model

Lee, Kyeong Chan
(Supervisor Kim, Kyoung ok)
Dept. of Data Science
Graduate School
Seoul National University of Science and Technology

Session-based recommendation system(SBRS) predicts the next item to click by using successively listed information of items contained within a session. SBRS is a type of sequential recommendation method that models user's temporal preferences. Prior researches have mainly adopted recurrent neural network(RNN) which is suitable for modeling sequential information.

Session-based recommendation systems have evolved to accurately model user's dynamic preferences. However, SBRS are likely to be utilized in environment where large data is not available because SBRS uses only implicit behaviors (such as visits, clicks) not explicit feedback (such as ratings). If the SBRS with RNN are used in these situations, generalization performance is low due to overfitting to small training data.

In this paper, we propose a data augmentation method that can improve the performance of SBRS in a situation where session data is insufficient. The item sequences in sessions are including relationship information between items as well as sequential information. The proposed method considers the characteristics of the session data and generate sessions similar to the observed data without user profiles. As a result, the performance of the SBRS can be improved by inflating training data through the proposed method. This study demonstrated the improvement of the performance of SBRS through a variety of experiments with real-world data.

감사의 글

‘데이터 사이언티스트’라는 직업이 멋있어 보여서 석사과정을 시작하기로 마음먹었는데 벌써 그로부터 2년이 지났습니다. 새로운 시작이 주었던 두근거림은 아직도 여전하고 앎의 기쁨을 느꼈던 순간들도 잊지 못할 것 같습니다. 알아야 할 것이 산더미 같고 갑자기 주어진 낯선 시간들에 고민만으로 지나쳐 보낸 하루들이 많았음에도 무사히 졸업하게 되어 다행이라고 생각합니다. 하지만 학문의 세계가 광활하다는 사실을 알게 되며 절로 겸손해질 수밖에 없는 시간이었다는 점이 더욱 다행이라고 생각합니다.

석사학위논문을 마무리하면서, 헤매는 데에 정신없었던 석사과정 동안 저에게 힘이 돼주셨던 분들께 마지막 페이지 감사의 글 형식을 빌려 감사 인사를 드리고 싶습니다.

우선 부모님께 가장 감사드립니다. 갑작스러운 대학원 진학에도 좋은 선택이라며 관심 가져주시고, 응원해주시고, 항상 걱정해주시는 두 분이 계셔서 저는 아무런 걱정 없이 하고 싶었던 공부를 마음껏 할 수 있었습니다. 말로 하기는 부끄럽지만 두 분을 항상 존경하며 꼭 부모님 같은 부모님이 되고 싶다는 것을 알아주셨으면 좋겠습니다. 그리고 멀리서도 격려해준 형과 형수, 지난 10월에는 저를 삼촌으로 만들어준 로한이에게도 감사드립니다.

누구보다도 본 논문은 물론이거니와 지도교수님의 위치에서 올바른 방향을 가르쳐주시고 연구자의 자세를 길러주신 김경옥 교수님께 감사드립니다. 짧은 시간이었지만 학문적인 어려움을 가장 믿고 기댈 수 있는 분께 연구 활동의 기본을 배우게 되어 정말 감사하게 생각합니다. 졸업할 때 즈음에는 훌륭한 제자가 되고 싶었는데 아직도 시간이 많이 필요할 것 같습니다. 하지만 앞으로 좋은 연구자가 될 수 있도록 노력해나가겠습니다. 감사합니다.

그리고 학위논문 심사를 맡아주신 황상흠 교수님, 이영훈 교수님, 열정적인 강의와 많은 세미나를 통해 학문의 저변을 넓혀 주신 데이터사이언스학과 교수님들, 가장 가까운 곳에서 서로 응원하고 지지해주고 도와준 청록, 호현, 예랑, 주현, 배울 것이 많은 학과 모든 동료에게도 감사드립니다.

마지막으로, 원하지는 않았지만, 석사과정 시작과 동시에 2년간 공부에만 매진할 수 있도록 지구 전체를 조용하게 만들어준 코로나19에게도 감사하다고 전하고 싶습니다.

2021년 12월 30일 이경찬 올림