쇼핑몰 상품 카테고리 분류를 위한 Hyperbolic Interaction Model의 적용과 분석

김시현[1], 이은지[1,2], 박성원[1,2], 차미영[2,1], 김선동[2]
[1] 한국과학기술원 전산학부 [2] 기초과학연구원 데이터 사이언스 그룹
{sihk, mk35471, psw0416 }@kaist.ac.kr, {mcha,sundong}@ibs.re.kr

Hyperbolic Interaction Model for Category Classification in E-Commerce

Sihyeon Kim[1], Eunji Lee[1,2], Sungwon Park[1,2], Meeyoung Cha[2,1], Sundong Kim[2]
[1] School of Computing, Korea Advance Institute of Science and Technology
[2] Data Science Group, Institute for Basic Sceince

요 약

온라인 쇼핑을 하는 소비자들은 사이트의 물품 분류 목록에서 원하는 물품을 찾고 비교하기 때문에 정확하고 통일성 있게 분류된 상품은 소비자들의 구매를 돕는다. 본 연구에서는 정확도 높은 쇼핑몰 상품 분류기를 만들기 위해 계층 분류에 효과적인 모델인 HyperIM을 카카오 쇼핑몰 데이터셋에 적용해 보았다. HyperIM 임베딩 결과를 통해 계층적 특징이 잘 반영되었음을 확인하였고, 실험 결과 분석을 통해 분류 모델의 성능을 높이기 위한 방향을 제안한다.

1. 서론

최근 인터넷 기술의 발달과 인터넷 쇼핑의 편리성, 가격의 이점으로 인터넷 쇼핑몰의 사용량이 증가했다[1]. 한 인터넷 쇼핑몰의 아이템은 수억 개에 이르기도 하여, 물품의 정확한 카테고리 분류는 소비자들의 검색에 큰 역할을 한다. 반면 공급자 별 동일한 상품이라도 다르게 분류될 수 있으며, 데이터 오류로 데이터 범위를 밖의 값이 존재할 수 있다. 이 연구는 소비자의 직관적이고 효율적인 쇼핑을 위한 물품 분류 모델을 제시한다. 이 논문의 구성으로 먼저 2장에서는 사용된 카카오 쇼핑몰 데이터를 소개하고, 3장에서는 계층적(hierarchical) 데이터 분류를 위한 Hyperbolic Interaction Model[2] (이하 HyperIM) 및 이를 기반으로 설계된 실험방법을 기술한다. 4-6장에서는 실험의 결과 및 고찰과 향후 연구 방향, 결론을 정리한다.

2. 카카오 쇼핑몰 데이터셋

이 연구에는 카카오 아레나에서 주최하는 쇼핑몰 상품 카테고리 분류 대회의 쇼핑몰 데이터[1]를 활용하였다. 분류의 목적은, 예로 상품명 '맛있는 제주차 3종세트 …' 및 기타 정보를 활용하여 상품의 카테고리인 음료/생수/커피 → 차/티백 → 차 선물세트 (대/중/소분류)를 예측하는 분류기를 만드는 것이다. 표 1은 데이터의 속성을 소개하며, 총 12개 중 4개 속성은 상품의 카테고리 분류 결과를 의미하고, 대/중/소/세의 네 가지 분류 위계를 가진다. 일부 상품에 대해 상위 분류에 대한 정보만 존재하기도 하며, 각 분류는 57, 552, 3,190, 404개의 라벨이 존재했다. 총 8,134,818개의 학습 데이터와 507,783개의 검증 데이터 및 1,526,523개의 테스트 데이터로 이루어져 있는 빅데이터에 해당한다.

표 1. 카카오 쇼핑몰 데이터셋

이름	상세	타입	비고
pid	상품 ID	String	Input feature (x)
product	상품명	String	
brand	브랜드명	String	
model	정제된 상품명	String	
maker	제조사	String	
price	가격	Numeric	
updttm	상품 업데이트 시간	String	
img_feat	ResNet50을 활용한 상품 이미지 임베딩	String	
bid (l_b)	대분류ID (57개)	String	Target label
mid (l_m)	중분류ID (552개)	String	
sid (l_s)	소분류ID (3,190개)	String	
did (l_d)	세분류ID (404개)	String	

그림 1을 통해 쇼핑몰 상품의 카테고리가 심한 불균형성(class imbalance)을 띄는 점을 확인할 수 있으며, 이러한 데이터 특성을 반영하여 효과적인 분류기의 개발이 이 연구가 지닌 어려운 점이다.

[1] https://arena.kakao.com/c/5/data

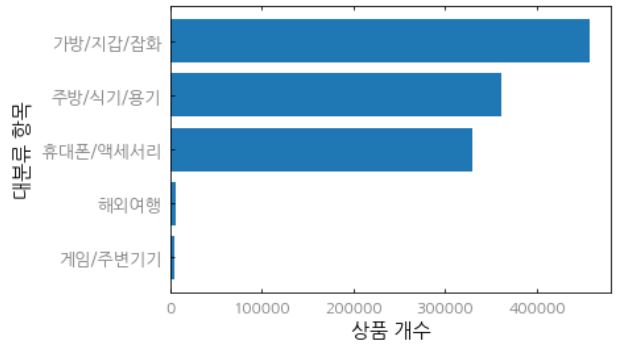

그림 1. 상품 분류 분포의 불균형성

3. 계층 분류를 위한 HyperIM 활용

3.1. HyperIM 개요

이 연구는 계층적 구조를 가지는 데이터를 분류하기에 적합한 모델을 사용하기 위해 카테고리 및 물품 설명에 활용된 단어의 임베딩을 쌍곡 공간(hyperbolic space)의 푸앵카레 원판(Poincare disk)에 학습한다. 쌍곡 기하학의 특성은 유클리드 공간에 비해 계층 데이터의 임베딩을 학습하기에 적합하다[3]. HyperIM은 각 카테고리와 아이템 설명 간의 푸앵카레 거리를 활용하여 분류를 진행한다.

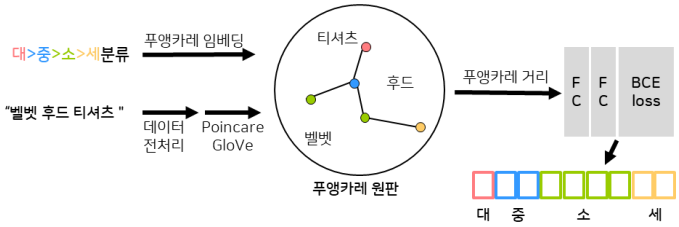

그림 2. HyperIM: 푸앵카레 원판에 학습된 상품의 설명 및 카테고리 간의 유사도를 계산하여 계층 학습 진행

3.2. 데이터 전처리

상품 분류를 위한 입력 값의 인자로 상품명(product) 속성을 활용하였다. 상품명의 경우 조사가 존재하지 않는 불완전한 문장의 형태를 띈다. 토큰화는 띄어쓰기를 기준으로 하고 한 글자의 단어, 숫자, 기호는 제거하였다. 토큰화 된 상품명에서 최대 16개의 단어를 사용하며, 단어 16개 미만의 상품명은 0으로 패딩했다.

3.3. 임베딩

쌍곡 공간에서의 카테고리 라벨 임베딩은 Gensim 라이브러리의 Poincare Model[2]을 이용하였다. 입력값으로 학습 데이터의 대/중/소/세분류에서 나타난 대-중, 중-소, 소-세분류의 부모-자식 노드의 관계를 받아, 각 카테고리별 임베딩을 생성한다.

단어 임베딩은 Poincare Glove를 이용하였다. 이 방법은 단어의 동시 발생 횟수를 변수로 가지는 손실 함수(loss function)을 이용하여 단어 간 관계를 반영한다. 입력값은 전처리 된 상품명 데이터이다. 라벨과 단어 임베딩은 12차원이며, 가장 많이 사용된 200,000개의 단어를 활용하였다.

그림 3은 임베딩 결과를 2차원 공간에 시각화 한다. 왼편 그림은 2차원의 푸앵카레 원판에 임베딩 된 라벨 대/중/소/세분류를 나타낸다. 루트를 기준으로 트리 구조를 그리고 있으며, 분류 체계의 계층적 구조가 잘 반영된 것을 볼 수 있다. 오른편 그림은 물품 설명에 사용된 단어 임베딩과 카테고리 라벨 임베딩을 t-SNE로 차원 축소(dimension reduction)하여 2차원 유클리드 공간에 나타낸 결과이다. '대분류-등산/캠핑/낚시'와 이의 하위 라벨 '중분류-등산의류'의 임베딩이 유클리드 공간에 가까이 위치한 것을 볼 수 있다. 상품명에 활용된 단어의 경우, 임베딩 공간 상에서 두 그룹으로 분리되었다.

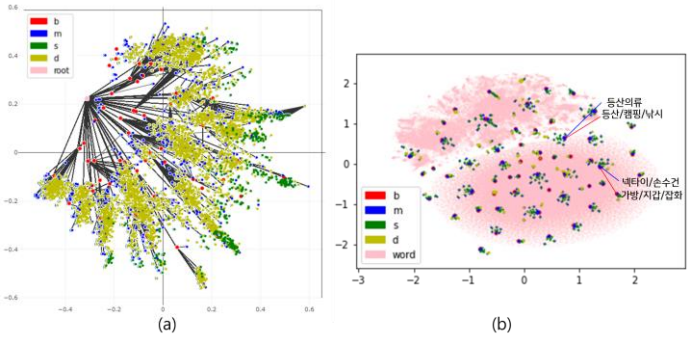

그림 3. 임베딩 시각화 (a) 푸앵카레 원판에서의 라벨 분포 (b) 유클리드 공간에서의 라벨/단어 분포

3.4. 분류 값 예측

HyperIM f를 활용하여 각 상품 정보 x를 학습하여 대/중/소/세분류에 속할 확률 벡터 P_b, P_m, P_s, P_d를 받는다.

$$P_b(x), P_m(x), P_s(x), P_d(x) = f(x), \sum P(x) = 1.$$

상품 x의 대분류 예측값 $l_b(x)$는 확률 벡터 중 가장 큰 softmax값을 가지는 인덱스인 $argmax(P_b(x))$가 된다. 다음으로, 상품 x의 중분류 $l_m(x)$은 $\mathrm{argmax}(P'_m(x))$로 얻는데, $P'_m(x)$는 HyperIM으로 계산된 중분류 확률 벡터 $P_m(x)$ 중에서, 예측된 대분류 $l_b(x)$와 부모-자식 관계가 있는 중분류들에 대한 확률 벡터를 뜻한다. 소/세분류 $l_s(x)$, $l_d(x)$의 예측은 동일한 방식으로 진행된다.

4. 실험 결과 및 고찰

연구에서는 평가 지표로 카카오 대회 방식에서 제시된 방법을 따랐다. 이는 각 라벨 l에 대한 정확도 a_l와 계층 구조에 따른 가중치 w_l를 종합하여 최종 점수를 계산하

[2] https://radimrehurek.com/gensim/models/poincare.html

는 것이다. 하위 분류에 대한 값을 맞출수록 더 높은 가중치 w_l가 부여된다. 각 계층에 대한 분류의 가중치는 대분류 1.0, 중분류 1.2, 소분류 1.3, 그리고 세분류 1.4에 해당한다. 만점은 1.225로, HyperIM을 사용한 점수는 0.77로 확인할 수 있었다.

$$\text{score} = \sum_{l \in \{b,m,s,d\}} (a_l * w_l) / 4$$

가중치	w_b	w_m	w_s	w_d
값	1.0	1.2	1.3	1.4

카카오 아레나의 리더보드에 현재까지 등록된 기존 모델의 최고 점수[3]가 1.08임을 감안했을 때 제안된 모델은 다소 아쉬운 점수의 성능을 보여주었는데, 그 이유는 다음과 같이 분석한다.

주어진 데이터에는 카테고리를 제외하고 8가지 속성이 있지만, 이번 실험에서는 상품명만을 실험 모델에 사용하였다. 브랜드명을 사용한다면 상품과 라벨 간 관계를 강화하여 카테고리 라벨을 더 정확하게 특정할 수 있고, 이미지 정보를 사용한다면 상품을 파악하는 성능을 높일 수 있다. LSTM을 사용한 모델은 상품명과 이미지 정보 모두를 이용했다.

HyperIM을 제안한 논문[2]에서 사용된 데이터는 위키나 기사처럼 문장으로 있어 Poincare Glove 임베딩을 사용했을 때 단어 사이의 관계에서 의미 있는 정보를 얻을 수 있다. 반면 본 실험에서 사용한 상품명 데이터는 주어나 동사, 목적어 없이 명사의 조합으로 이루어져 있어 단어 간 관계가 약하다. 이는 그림 3(b)에서 단어 임베딩이 두 그룹으로 나누어진 원인과 관련이 있다고 사료된다.

마지막으로 HyperIM은 한 아이템에 대한 라벨이 여러 개인 다중 라벨 분류(multi-label classification)을 위한 모델이다. HyperIM은 손실함수로 BCELoss와 시그모이드 레이어(Sigmoid layer)의 결합인 BCEWithLogitsLoss를 사용한다. 쇼핑몰 카테고리 분류 문제는 단 하나의 라벨에 대해서만 오차(loss)를 계산하면 되지만, 이 손실 함수는 각 계층에서 여러 개의 라벨을 고려하기 때문에 단일 예측 문제에서 성능이 떨어질 수 있다.

5. 향후 연구 방향

모델의 성능을 높이기 위한 방법으로 다음을 제안한다. 먼저, 상품 이미지, 상품의 가격 등 추가 속성을 모델 학습에 활용할 수 있다. 서로 다른 분포를 따르는 속성값을 최적으로 학습하기 위한 속성별 중요도를 결정하는 컴포넌트의 역할이 중요할 것이다.

또한, 라벨의 하위에 해당 라벨과 관련 있는 단어를 함께 그래프에 임베딩하는 것을 제안한다. 불완전 문장에서 부족한 단어 간 관련성을 보완하기 위한 방법이다. 유사한 모델로 계층-텍스트를 함께 표현한 HIMECat[5]이 있다. 쇼핑몰 데이터셋의 특징상 라벨 이름이 상품명에 포함되어 있는 경우처럼 라벨과 상품명의 유사도가 비교적 높은 특징을 가진다. 라벨의 이름 또한 단어 임베딩에 추가하면 학습 효과를 높일 수 있다.

마지막으로, 계층 분류 모델을 적용하였을 때, 카테고리별 분류 성능 이외의 다양한 지표—고객의 상품 클릭율, 구매율, 재구매율 등의 향상 정도를 파악해볼 수 있다. eBay에서 진행한 연구[4]에서 활용한 데이터셋을 추가로 활용하면 고객의 구매 내역, 함께 열람한 상품 내역, 대신 추천된 상품 정보 등 고객과 상품 간의 관계를 밀도 있게 파악하여, 쇼핑몰 환경에서의 상품 추천을 시뮬레이션해볼 수 있다.

6. 결론

발달하는 인터넷 쇼핑몰 시장에서 필요한 쇼핑몰 상품 분류기를 제공하기 위한 연구를 진행하였다. 계층적인 분류 체계를 효과적으로 파악하기 위해 HyperIM을 사용했고, 대/중/소/세분류를 계층적 특징이 반영되도록 임베딩하는데 성공하였다. 향후, 단어 임베딩을 개선하고 추가 데이터를 이용하여 쇼핑몰 상품 추천에서의 계층 모델의 효용성을 알아보고자 한다.

사사문구

이 논문은 기초과학연구원(IBS-R029-C2)과 한국연구재단 기초연구사업(NRF-2017R1E1A1A01076400)의 지원을 받아 수행된 연구임.

참 고 문 헌

[1] 통계청, 2021년 2월 온라인쇼핑 동향, 2021.

[2] Boli Chen et al., “Hyperbolic Interaction Model for Hierarchical Multi-Label Classification”, AAAI 2020.

[3] Benjamin Paul Chamberlain et al., “Neural Embeddings of Graphs in Hyperbolic Space”, KDD Workshop 2017.

[4] Da Xu et al., “Knowledge-aware Complementary Product Representation Learning”, WSDM 2020.

[5] Yu Zhang et al., “Hierarchical Metadata-Aware Document Categorization under Weak Supervision”, WSDM 2021.

[3] https://github.com/lime-robot/product-categories-classification