

Word2vec 기반 사용자 관심 카테고리 비즈니스 카테고리 매칭

유경호, 김판구
조선대학교 컴퓨터공학과
infinite_gh@naver.com, pkkim@chosun.ac.kr

Matching user interest categories and business categories based on Word2vec

KyoungHo Yu, Pankoo Kim
Department of computer engineering, Chosun University

요 약

최근 사용자들의 상품 구매 패턴은 SNS나 인터넷 방송에서 유명한 인플루언서의 사진이나 영상을 보고 그 상품을 검색하여 구매하고 있다. 온라인상에서 보았던 상품을 자신의 위치에서 인접한 상점을 방문하여 구매하고자 할 때, 그 상품을 구매하려면 어떤 상점으로 가야할지 잘 모를 수 있다. 따라서 본 연구에서는 이미지를 입력으로하여 분류된 사용자 관심 카테고리를 비즈니스 카테고리와의 매칭을 통해 사용자에게 추천하고자 한다. 사용자 관심 카테고리나 비즈니스 카테고리의 유사도 측정은 word2vec을 사용하여 벡터화한 뒤 cosine 유사도를 통해 측정하였다. 실험결과 본 연구에서 제안하는 방법으로 하였을 때 관련있는 카테고리 사이의 유사도가 높은 것을 확인 하였다.

1. 서론

최근 사용자들은 SNS나 인터넷 방송을 통해 유명해진 인플루언서(influencer)가 업로드한 이미지를 보고, 인플루언서가 착용한 상품을 구입하거나 방문했던 음식점이나 유명 관광지를 직접 찾아가는 사례가 증가하고 있다[1]. 하지만, 온라인상에서 보았던 상품이나 음식점 혹은 서비스를 사용자의 위치 주변에 있는 상점에 찾길 원한다면 어떤 상점을 방문해야할지 잘 모를 수 있다. 따라서, 사용자가 관심있어하는 이미지를 갖고 있을 때, 이 이미지가 어떤 상점 카테고리에 속하는지 딥 러닝을 사용하여 분류하는 연구가 선행되었다[2].

본 연구에서는 사용자 위치에 인접한 상점을 사용자에게 추천하기 위해 선행연구에서 수행하였던 사용자 관심 카테고리나 비즈니스 카테고리 사이의 유사도를 측정하여 가장 유사한 상점을 추천하도록 한다. 유사도 측정은 사용자 관심카테고리와 비즈니스 카테고리가 텍스트이기 때문에 word2vec을 사용하여 벡터화 한 뒤 cosine 유사도를 사용하여 측정한다.

2. 관련 연구

Word embedding이란 텍스트를 구성하는 단어를 실수 벡터로 매핑하는 것을 말한다[3]. 대표적인 word embedding의 방법으로는 Word2vec이 있으며 CBOW(Continuous Bag of Words)와 Skip-gram 두 가지 방식의 모델을 지원한다. CBOW와 Skip-gram의 차이는 '나는 학교에 버스를 타고 간다.'라는 문장에서 '버스'를 예측한다면 CBOW는 창(window) 안에 있는 모든 단어를 기준으로 예측한다. 반면 Skip-gram은 임의의 가까운 단어를 기준으로 예측한다. 대용량의 문장을 사전학습한 word2vec 모델은 얇은 신경망을 사용하므로 학습 속도가 빠르고 언어 모델과 비교하였을 때 모델의 복잡도가 낮아 자연어 처리 분야에서 단어와 단어사이의 유사성 또는 문장과 문장 사이의 유사성을 판단할 때 자주 사용되는 모

델이다.

3. 본론

3.1 Word2vec을 이용한 카테고리 유사도 측정

본 절에서는 word2vec을 사용하여 딥 러닝 네트워크를 사용하여 분류된 사용자 관심 카테고리나 비즈니스 카테고리 사이의 유사도를 측정한다. 본 연구에서 제안하는 카테고리 사이의 유사도 측정은 그림 1과 같다. 사용자 관심 이미지를 딥 러닝 네트워크에 입력하여 분류된 사용자 관심 카테고리나 비즈니스 카테고리를 word2vec을 사용하여 벡터화한다. 그 다음 벡터 값을 cosine 유사도를 사용하여 측정하게 된다. 그 후 측정된 유사도를 바탕으로 높은 유사도를 보인 상점을 사용자에게 추천하게 된다.

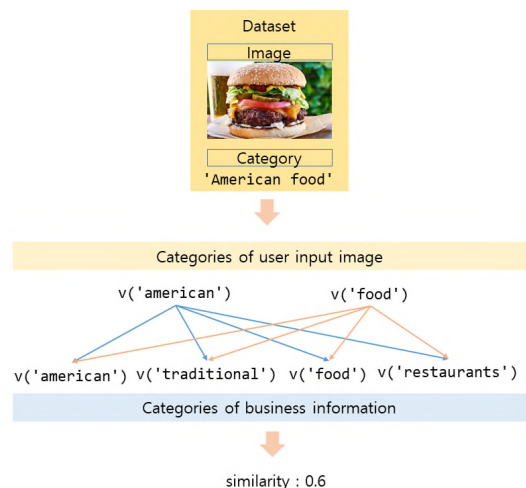


그림 1. Word2vec을 이용한 카테고리 유사도 측정

3.2 실험 결과

실험에 사용한 데이터는 웹 검색엔진을 사용하여 크롤링한 이미지를 딥 러닝 네트워크에 입력으로 하여 분류된 사용자 관심 카테고리라 yelp local business data에서 오하이오 주의 비즈니스 데이터를 사용하였다.

표 1. 'american food'와 비즈니스 카테고리 사이의 유사도

Image Categories	Business Categories	Similarity
'american', 'food'	'american', 'traditional', 'food', 'cajun', 'creole', 'fast', 'food', 'restaurants', 'chicken', 'wings'	0.6
	'paddleboarding', 'boating', 'rafting', 'kayaking', 'jet', 'skis', 'active', 'life'	0.14
	'active', 'life', 'trainers', 'gyms', 'fitness', 'instruction'	0.15
	'parks', 'active', 'life', 'hiking'	0.14
	'eyelash', 'service', 'beauty', 'spas', 'nail', 'salons'	0.17
	'eyebrow', 'services', 'hair', 'salons', 'beauty', 'spas'	0.18

표 2. 'Hair salon'과 비즈니스 카테고리 사이의 유사도

Image Categories	Business Categories	Similarity
'hair', 'salon'	'american', 'traditional', 'food', 'cajun', 'creole', 'fast', 'food', 'restaurants', 'chicken', 'wings'	0.17
	'paddleboarding', 'boating', 'rafting', 'kayaking', 'jet', 'skis', 'active', 'life'	0.13
	'active', 'life', 'trainers', 'gyms', 'fitness', 'instruction'	0.13
	'parks', 'active', 'life', 'hiking'	0.06
	'eyelash', 'service', 'beauty', 'spas', 'nail', 'salons'	0.35
	'eyebrow', 'services', 'hair', 'salons', 'beauty', 'spas'	0.48

실험 결과는 표 1과 2와 같다. 표 1은 'american food'와 비즈니스 데이터 6개를 비교한 결과이며, 표 2는 'hair salon'과 비즈니스 데이터를 비교한 것이다. 표 1에서 'american food'와 첫 번째 비즈니스 카테고리에서 높은 유사도를 보였으며 표 2에서 'hair salon'과 5, 6번째 비즈니스 카테고리에서 높은 유사도를 보였다. 두 표에서 알 수 있듯이 육안으로 판단하였을 때 유사하다고 판단되는 두 카테고리 사이에서 유사한 것을 볼 수 있다.

4. 결론

본 연구에서는 사용자가 인플루언서가 업로드한 이미지나 영상을 보고 구입하고 싶은 상품을 구입하고자 할 때, 이미지의 카테고리라와 상점 카테고리 사이의 유사도를 측정하여 높은 유사도를 보이는 비즈니스를 사용자에게 추천하는 방법을 제안하였다. 제안하는 방법은 word2vec을 사용하여 카테고리를 벡터화하고 cosine 유사도를 사용하여 측정하였다. 이 때 각 카테고리의 각 단어마다 유사도를 측정하여 그 평균 값을 사용자 관심 카테고리-비즈니스 카테고리의 유사도로 측정하였다. 실험 결과 육안상 비슷하다고 생각되는 사용자 관심카테고리라와 비즈니스 카테고리에서 높은 유사도를 보였다. 따라서, 본 연구에서 제안하는 카테고리 매칭방법을 사용하였을 때 사용자가 관심있어하는 이미지에 대한 상점을 추천할 수 있다.

감사의 글

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF-2020R1A2C2007091).

참고 문헌

- [1] Rafay, Abdul, M. Suleman, and Affan Alim. "Robust Review Rating Prediction Model based on Machine and Deep Learning: Yelp Dataset." 2020 International Conference on Emerging Trends in Smart Technologies (ICE TST). IEEE, 2020.
- [2] 유경호, 김강민, 김관구. "사용자 관심 이미지와 상점 카테고리 매칭을 위한 이미지 분류", 12회 CSMS, 2021.pp.
- [3] Mikolov, T., Chen, K., Corrado, G., & Dean, J., "Efficient estimation of word representations in vector space", arXiv preprint arXiv:1301.3781, 2013