

CNN 기반의 국내 스타트업-해외 바이어간 추천시스템 설계

최 정석, 문 남미
호서대학교 벤처대학원 융합공학과
20205162@vision.hoseo.ac.kr, nammee.moon@gmail.com

Designing a Recommendation System between Korean Start-ups and Foreign Buyers based on Convolutional Neural Network

Jungsuk Choi, Nammee Moon
Dept. of Convergence Engineering, Hoseo Graduate School of Venture

요 약

본 논문은 국내 스타트업의 상품/서비스에 적합한 해외 바이어를 찾아 맞춤형으로 추천해주는 시스템을 설계하고자 한다. 추천 알고리즘은 CNN기반의 Word2Vec과 Doc2Vec 알고리즘을 활용하며, 정확도를 높이기 위해 시각정보를 활용한다. 추천 시스템에 사용되는 데이터는 비정형 데이터인 회사 소개 및 상품/서비스 소개 문장 데이터이며, 제품 사진을 시각정보로 이용한다. 유사도가 높은 순으로 추천하기 위해 문장데이터를 키워드 리스트로 변환하고, Word2vec 모델에 이식시켜 키워드 좌표를 만들어 벡터화한다. 그리고, 문장의 중심점간 거리를 계산해 기업간 유사성 및 연관성을 도출한다. 이를 바탕으로 국내 스타트업의 문장데이터 및 시각정보와 유사도가 높은 순으로 해외바이어를 추천한다.

키워드 : CNN, word2vec, doc2vec, 시각정보, 추천시스템

1. 서론

국내 스타트업이 스케일업을 위해서는 작은 국내 시장에 머물기보다 글로벌 시장으로 진출해 시장을 확장해 나가는 마케팅 활동이 필요하다. 본 논문에서는 국내 스타트업이 적격 해외 바이어에게 상품 및 서비스를 수출할 수 있도록 CNN(Convolutional Neural Network) 기반의 Word2Vec과 Doc2Vec 알고리즘을 활용해 맞춤형으로 해외바이어를 추천해주는 시스템을 설계했다.

제안하는 추천 시스템에 활용되는 데이터는 비정형 데이터인 회사 소개 및 제품 설명서등 문장 데이터와 제품 사진등의 시각정보다. 추천을 위해 CNN기반의 Word2Vec 알고리즘을 활용해 이들 데이터간 유사도를 계산한 후, 유사도가 높은 순으로 해외바이어를 추천한다. 단어와 문장으로 제품을 완전하게 설명하기 어려운 Word2Vec 알고리즘의 단점을 보완하기 위해 시각정보인 제품 사진을 통해 성능을 개선하고자 한다. 2장에서는 관련연구를 요약했고, 3장에서는 제안하는 추천시스템의 프로세스와 기본구조를 설명했다. 마지막으로 4장에서는 결론 및 향후 과제에 대해 정리했다.

2. 관련 연구

추천시스템과 관련한 주요 사전연구는 다음과 같다. 선행연구[1]은 Word2Vec과 사용자벡터, 영화벡터를 2채널로 구성하고, 2채널 합성곱 신경망 학습을 통해 추천시스템의 정확도를 높였다. 선행연구[2]는 시각정보를 활용한 딥러닝 기반 추천 시스템을 통해 추천시스템의 정확도를 높이고, 활용 데이터의 영역을 시각정보로까지 넓혔다. 선행연구[3]은 텍스트와 이미지 태그 데이터에 기반한 추천 시스템을 개발했다. 선행연구[4]는 Word2Vec과 더불어 Doc2Vec을 함께 사용하는 것이 문장분류에 효과적임을 검증했다.

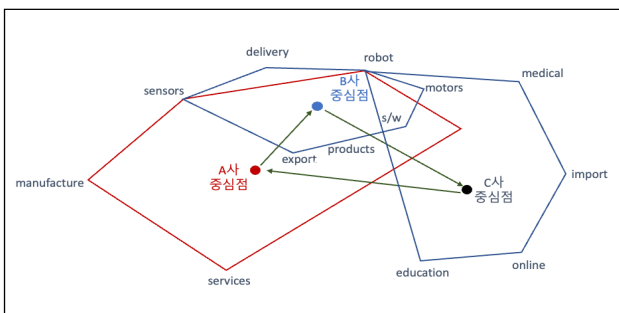
3. 제안하는 CNN기반 추천 시스템

국내 스타트업과 해외 바이어의 회사소개, 취급 상품/서비스 등을 소개하는 키워드와 문장(영문)을 수집, 전처리 과정을 거쳐 하둡(HDFS)에 저장한다.

비정형 데이터인 회사소개 자료의 문장을 키워드 리스트로 변환하고, 이를 Word2Vec 모델에 이식시켜 좌표로 만들어 벡터화한다. 그리고 키워드 리스트의 좌표 중심점간 거리를 계산해 기업간 유사성 및 연관

성을 도출한다.

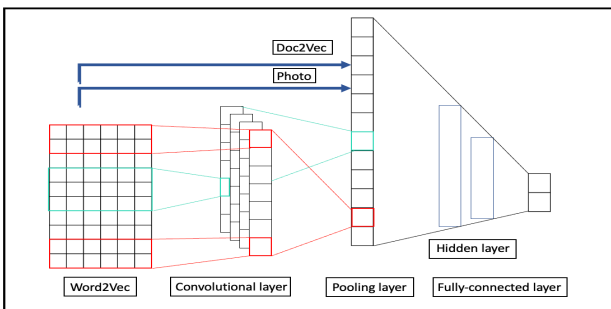
예를 들어 A사 홈페이지의 ‘We manufacture sensors, robots for medical service’라는 문장에서 키워드로 ‘manufacture, sensors, robots, medical, service’를 추출한다. B사 브로슈어의 ‘We export sensors, motors, s/w for delivery robots’라는 문장에서 ‘export, sensors, motors, s/w, delivery, robot’라는 키워드를 추출한다. 그리고 C사 자료에서 ‘we import robots for online medical education’이라는 문장에서 ‘import, robot, online, medical education’이라는 키워드를 추출한다.



<그림 1> 문장 중심점간 거리 측정

위 그림 1 에서 거리측정 결과 A-B사간 거리는 ‘robot, sensor’라는 동일한 키워드와 ‘product, s/w, motor’라는 키워드의 유사성으로 인해 0.9로 가장 가깝고, B-C사간 거리는 ‘robot’이라는 동일 키워드와 ‘medical, product’라는 키워드의 유사성에 따라 1.4로 나타나고, A-C사간 거리 ‘robot’이라는 동일 키워드를 제외하고는 유사성이 낮아 1.9로 가장 멀게 나타난다.

이 거리를 바탕으로 A,B,C사의 제품 사진의 유사도를 분석해 거리 편차를 조정한다. CNN기반으로 Word2Vec알고리즘에 Doc2Vec 알고리즘을 함께 사용하고, 제품 사진을 시각정보로 활용한다, 이를 위해 은닉층(hidden layer)를 2개 보완한다.



<그림 2> 추천시스템의 구조 설계

이와 같은 과정을 거쳐 Word2Vec과 시각정보 이미지의 유사도를 측정해 국내 스타트업에 적합한 해외 바이어를 찾아낼 수 있다. 그리고, 추천된 해외 바이어와 유사한 X국가, Y국가, Z국가의 바이어를 추천함으로써 여러나라의 바이어를 추천할 수 있다.

4. 결론 및 향후과제

가. 결론

본 논문은 국내 스타트업의 글로벌 시장 진출을 통한 스케일업을 지원하기 위해 Word2Vec 알고리즘을 사용해 문장의 유사도를 파악해 유사도가 높은 순으로 추천해주는 매칭 시스템을 설계했다. 단어간 유사도 측정을 위한 Word2Vec 알고리즘, 그리고 문장간 유사도 측정을 위한 Doc2Vec 알고리즘, 여기에 시각정보인 제품/서비스 사진의 유사도를 측정해 국내 스타트업에게 해외 적격 바이어를 추천하는 시스템의 효과성을 높이고자 했다. 또한, 키워드 중심의 Word2Vec 알고리즘은 제품을 완전히 설명하는데 부족한 점이 있으므로, 이런 문제점을 보완하기 위해 사진 등 시각정보를 추가로 활용해 정확도를 높이고, 데이터의 영역을 넓혔다.

나. 향후 과제

한편, 국내 스타트업에게 해외 바이어를 추천한 후 얼마만큼의 효과가 있는지 검증하는 시스템이 필요하다. 국내 스타트업을 대상으로 설문조사 등을 통해 추천된 바이어와 거래가 성사되었는지를 파악해 시스템에 피드백한다면 추천시스템의 성능을 높일 수 있을 것이다.

[참고문헌]

- [1] 강부식(2021), “word2vec과 2채널 합성곱 신경망을 활용한 영화추천시스템이 정확도 개선”. 「한국지식정보기술학회논문지」, vol.16, no2, pp. 255-263
- [2] 문현실.조윤희(2020), “시각정보를 활용한 딥러닝 기반 추천시스템”. 「지식경영연구」, pp.27-43
- [3] 정인경.김보미.김수경.유건아(2020), “텍스트와 이미지 태그 데이터에 기반한 뉴스기사 추천시스템”. 「디지털 콘텐츠학회논문지」, 21(3), pp. 479-486
- [4] 김도우.구명완(2017), “Doc2Vec과 Word2Vec을 활용한 Convolutional Neural Network 기반 한국어 신문 기사 분류”. 「정보과학회논문지 44(7)」, pp.742-747.