Item2vec과 RNN_LSTM을 이용한 시간별 상품 추천 및 연관성 분석

이재상*, 김종국*, 신민정**, 허대현**, 이병권***,유재헌****
*동국대학교 산업시스템공학과
**동국대학교 전자전기공학부
***동국대학교 멀티미디어공학과
****이마트 24 전략마케팅팀
e-mail : jake0403@naver.com

Time to Time Product Recommendation and Correlation Analysis Based on Recurrent Neural Networks and Item2vec

Jae-Sang Lee*
*Dept of Industrial System Engineering, Dongguk University

요 약

데이터의 분석이 추세임에 따라 많은 기업의 마케팅이 개인 맞춤화 되어가고 있다. 이에 따라 Word2vec을 이용한 Item2vec을 활용하고, 시간대 별로 상품 추천을 위해 시계열 데이터를 분석하고 학습시키는 인공신경망인 RNN(Recurrent Neural Network)의 LSTM(Long Short-Term Memory) 구조를 이용하여 관련된 상품을 추천해주는 연구를 진행하고자 한다.

1. 서론

빅 데이터의 시대의 도래로 인해 모든 행동과 발생하는 일들은 수많은 데이터를 생산한다. 이러한 데이터의 홍수 속에서 다양한 분석과 연구들이 활발하며, 데이터들 사이의 상관관계를 파악하고 이를 활용하면 새로운 데이터로 인해 얻을 수 있는 효과들은 무궁무진하다. 고객 상품 추천 및 연관성 분석에서 주로 사용하는 Apriori 알고리즘은 자주 사용되는 연관성 분석 중 하나이지만 시간대 별 연관성 있는 상품을 추천해주기에 한계점이 있고, 후보 데이터군 생성 시 아이템 개수가 많아지면 계산의 복잡성은 증가하게 된다는 단점이 있다. 따라서 본 논문에서는 소비자들이 생산하는 구매 시계열 데이터로 딥러닝 기반 인공신경망인 RNN(Recurrent Neural Networks)과 LSTM을 이용해 상품을 구매한 시간대 별로 연관된 상품을 추천해주는 연구를 진행하고자 한다.

2. 관련연구

매출을 올리기 위해 다양한 고객 추천 알고리즘 들이 개발되어 왔고, Collaborative Filtering과 Apriori 알고리즘의 '장바구니 분석'과 같은 방법들이 개발 되어왔다. 하지만 데이터는 시간이 지날수록 많아지고, 이에 따라 계산의 복잡성은 증가하게 된다. 따라서 Word2vec을 응용한 Item2vec으로 상품의 연관성을 파악한다.

Word2vec이란, 2013년 Tomas Mikolov 등이 'Efficient Estimation of Word Representations in Vetor Space'을 제안한 논문으로 NNLM(Feed-Foward Neural Net Language Model)을 기반으로 단어의 벡터 값을 구하고 그 벡터 값을 통해 단어와의 관계를 파악하는 방법으로서 관련 모델로는 CBOW(Continuous Bag Of Words)모델과 Skip-gram 모델이 있다.[1] 본 연구에서는 Word2vec에 상품을 넣어 상품의 벡터 값을 만들고, Item2vec으로서 연구를 진행할 것이다.

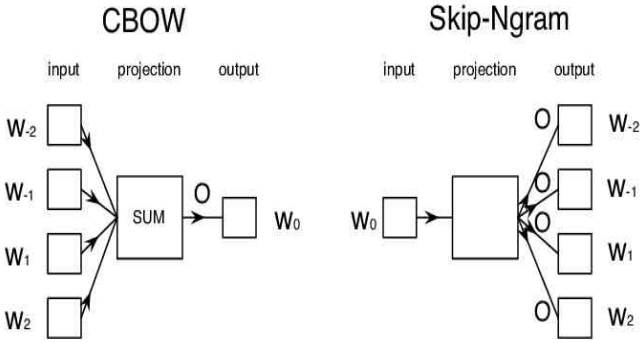

그림 1. CBOW와 Skip-gram 모델의 구조

상품 유사관계를 파악한 벡터 값을 구하게 되면 인공신경망인 RNN(Recurrent Neural Networks)의 LSTM(Long Short-Term Memory) 구조를 통해 상품 추

천을 시간대 별로 구매한 상품들을 학습시킨다.[2][3] 여기서 Test data와 Training data를 나누어 Training 시킨 결과가 Test 한 결과와 유의한지 판단한다.

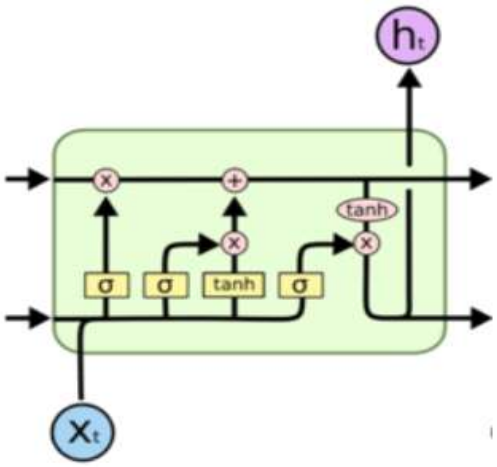

그림 2. LSTM(Long Short-Term Memory)의 구조

3. 시스템 구조

Item2vec과 RNN_LSTM을 이용한 시간별 상품 추천 및 연관성 분석은 Python을 기반으로 그림 3의 과정으로 분석 및 구현을 한다.

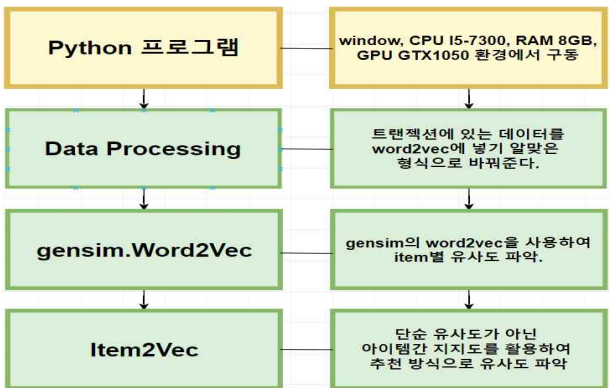

그림 3. Item2Vec 분석 방식

먼저 Python의 라이브러리인 Numpy, Pandas로 데이터 구조를 Raw data에서 Structured data로 변환을 시킨다. 라이브러리 Scikit-Learn의 train-test-split함수로 train data와 test data의 비율을 7:3으로 분할 시켜준다. 그 후 라이브러리 gensim의 Word2Vec을 import 시켜준다. train data만을 가지고 Word2Vec 함수를 이용하여 생성할 벡터의 크기, 주변 단어의 개수, 상품의 최소 출현 숫자, 컴퓨터 CPU의 코어 사용 개수, 반복학습 숫자, CBOW 또는 Skip-Gram 방식선택을 해준다. 그 후 상품별 Vector 값과 연관성 및 유사도를 구한다. 각 상품의 Vector값을 얻은 다음 상품 유사도에 따른 추천 상품이 아닌 구매한 상품에 따른 추천형식으로 Item2Vec의 형태로 변환시켜준다.

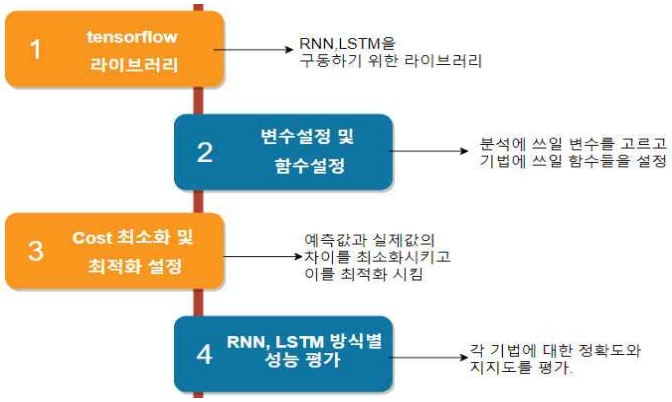

그림 4. RNN, LSTM 분석 과정

이제 시계열 데이터를 위의 그림 4.과정에 따라 RNN과 LSTM으로 분석을 한다. Python 라이브러리인 tensorflow를 import 시켜준 뒤 Data를 Numpy Array형식으로 변환시켜준다. Hidden Layer와 Weight, Batch size를 설정하고, tensorflow의 Placeholder에 shape을 입력해 준다. 모델 정의에 필요한 변수들을 설정해 준 뒤, tanh와 Relu함수로 Hidden Layer에서 계산을 시켜주고, 마지막 Output Layer에서는 Softmax 함수를 적용시켜 준다. cross-entropy를 설정하여 cost값을 정의한 뒤 Gradient-Decent-Optimizer 방식으로 cost값을 최소화 시켜준다. LSTM은 여기에 forget gate를 추가시켜 진행하면 된다.

이제 위의 과정으로 train data를 학습시킨 뒤에 test data로 RNN과 LSTM의 성능을 검증하고 더 나은 방식을 이용하여 상품 추천을 한다.

그림 5.는 Item2Vec과 K-means 군집화 방법을 사용하여 시간대별 가장 많이 사간 상품을 시각화하여 나타낸 것이다.[4]

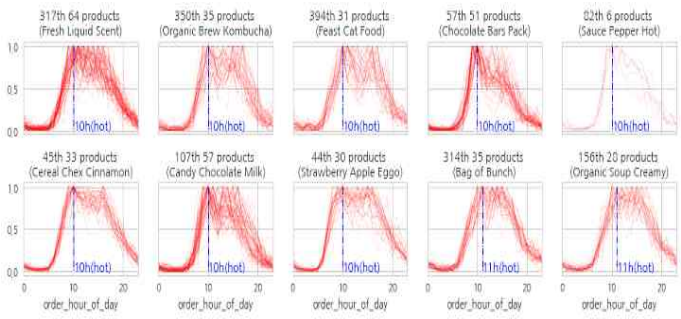

그림 5. Kmeans 군집화 방식 Item2Vec 분석

4. 결론

기존 상품 추천 방식은 연관도 분석을 통한 지지도와 향상도, 신뢰도를 바탕으로 한 일차원적인 추천 방식이었다. 하지만 연관도 분석만으론 시간에 따른 상품 선호도를 고려하지 않으면서 다양한 상품이 존재할 때에 상품을 추천하는 방법에는 적합하지 않다. 본 연구에서는 Word2Vec을 응용한 Item2Vec을 이용하여 RNN_LSTM

을 이용한 시간별 상품 추천 및 연관성 분석을 구성하였다.

향후 연구방향으로 실제 데이터를 입력하여 Score값을 구성한 뒤 추천 상품이 얼마나 예측을 잘하는지에 대한 연구가 필요하다.

본 연구는 과학기술정보통신부 및 정보통신기술진흥센터의 SW중심대학지원사업의 연구결과로 수행되었음.(2016-0-00017)

참고문헌

[1] Barkan, Oren, and Noam Koenigstein. "Item2vec: neural item embedding for collaborative filtering." 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP). IEEE, 2016.

[2] 김정미, 이주홍. "Word2vec을 활용한 RNN기반의 문서 분류에 관한 연구." 한국지능시스템학회 논문지, 27.6 (2017)

[3] Greff, Klaus, et al. "LSTM: A search space odyssey." IEEE transactions on neural networks and learning systems 28.10 (2017): 2222-2232.

[4] 최규민 “Instacart data exploratory 2” https://brunch.co.kr/@goodvc78/18 (2017)