데이터 정제를 통한 딥러닝 기반의 유저 맞춤형 음식추천시스템

김균엽○, 강상우
가천대학교 소프트웨어학과,가천대학교 소프트웨어학과
gyop0817@gc.gachon.ac.kr, swkang@gachon.ac.kr

User-specific Food Recommended System Using Data Cleaning

Gyun-Yeop Kim○, Sang-Woo Kang
Gachon University. Department of software

요 약

제품을 추천하는 기능은 사용자의 콘텐츠 또는 제품 소비량에 직결되기에 다양한 인터넷 플랫폼에서 많은 관심을 받고 있다. 이러한 제품 추천 시스템의 성능은 다양한 머신러닝 알고리즘과 딥러닝의 발전에 의해 성능을 비약적으로 개선되어왔다. 하지만 여느 딥러닝과 머신러닝 알고리즘과 마찬가지로 추천 시스템들의 성능은 빅데이터의 품질에 따라 매우 민감한 영향을 받는다. 본 논문에서는 모바일 배달 플랫폼에서 사용자들의 리뷰 데이터들을 통해 딥러닝과 빅데이터를 사용하여 음식을 추천하는 방법을 제안한다. 또한 사용자들의 리뷰 데이터들을 정제하여 데이터의 품질을 높이는 과정을 추가하여 그 결과가 성능에 얼마만큼 영향을 미치는 지를 실험을 통하여 분석한다.

주제어: 추천시스템, 자연어처리, 데이터 정제

1. 서론

최근 모바일 플랫폼 시장이 확대되면서 사용자에게 콘텐츠 또는 제품을 추천하는 기능은 대부분의 플랫폼에서 필수적인 기능으로 자리 잡아왔다. 이러한 제품 추천 기능의 성능은 사용자가 콘텐츠를 더 많이 소비할 수 있도록 유도하기에 많은 관심을 받아왔다. 또한 추천 성능을 개선하기 위한 많은 연구들이 진행되었다.

기존에는 사용자들의 일부 정보로부터 특징을 추출하여 제품을 추천하였다. 하지만 최근에는 긴 시간 동안 축적된 유저들의 빅데이터와 딥러닝의 발전이 가속화되었고 추천 시스템 역시 많은 성능 향상이 이루어졌다.

딥러닝과 빅데이터를 이용한 추천 시스템은 축적된 데이터의 품질에 크게 영향을 받는다. 하지만 모바일 플랫폼 시장의 확대와 함께 플랫폼 이용자가 증가함에 따라 플랫폼을 이용한 바이럴 마케팅, 광고, 이벤트를 이용한 댓글 유도 등 실제 사용자의 정보가 아닌 신뢰할 수 없는 정보 또한 급증하였다. 이러한 신뢰할 수 없는 정보는 딥러닝을 이용한 추천 시스템의 성능에서 학습하는 도중 악영향을 미치게 된다.

따라서 본 논문에서는 사용자의 음식 리뷰와 평점을 데이터로 사용하여 신뢰할 수 있는 정보의 분류 및 정제를 이용한 추천 시스템을 제안한다.또한 음식 리뷰 데이터의 정제를 통한 음식 추천 시스템의 성능 향상의 척도를 알아보며 제안한 프로세스에서의 적합한 음식 추천 모델과 데이터 정제 모델에 대해 알아본다.

2. 관련 연구

2.1 BERT

BERT(Bidirectional Encoder Representation from Transformers)[1]는 기존의 문맥 정보를 포함하지 않거나 단방향 문맥 정보를 포함한 언어 모델들에서 Masking 토큰을 통해 양방향 문맥 정보를 추출할 수 있는 딥러닝 모델이다.

BERT는 12개의 transformer block과 12개의 self-attention head를 통해 768차원의 hidden state를 도출해낸다. BERT는 문장 분류를 위해 문장의 첫 번째 토큰을 <CLS>토큰으로 지정하였다. <CLS>토큰의 hidden state는 전체 문장에 대한 representation을 가지고 있기 때문에 <CLS>토큰을 통해 문장 분류에 적용이 가능하다.[2]

2.2 추천 시스템

추천 시스템은 사용자의 축적된 구매 또는 사용 데이터를 기반으로 미래에 어떤 제품 또는 서비스를 선호할 것인지 예측하는 기술이다. 추천 시스템은 사용자의 평가정보를 가지고 추천하는 explicit feedback 데이터를 이용한 모델과 implicit feedback 데이터를 이용한 모델이 있다. explicit feedback 모델에서는 explicit feedback 데이터를 통해 축적된 사용자의 별점 또는 영

화 평점 등 사용자가 제품 또는 서비스를 선호하는지에 대한 정보를 기반으로 통계적으로 추천한다.[3] implicit feedback 모델[4]은 또한 시청 기록, 구매 기록 등 점수가 없이 사용자의 제품 사용 여부와 같은 이분법적인 데이터인 implicit feedback데이터를 통해 통계적으로 추천한다. 이후에는 deep learning을 사용하여 사용 순서를 기반으로 다음 순서에서 선호도를 예측하는 모델들이 제안되었다. 본 논문에서는 Wavenet[5]의 CNN 구조를 기반으로 다음 순서를 예측하는 모델을 사용하였다.

3. 제안방법

추천 시스템은 기존의 사용자가 남긴 데이터를 기반으로 사용자가 선호할 만한 제품을 추천하는 시스템이다. 그렇기에 추천 시스템의 알고리즘 또한 중요하지만 사용자가 남긴 데이터의 품질 또한 중요하다. text classification을 통해 데이터의 신뢰성 여부를 판별하고 testing 하는 과정에서 신뢰할 수 있는 데이터만을 입력함으로써 사용자의 실제 데이터만으로 추천할 수 있는 추천 시스템 프로세스를 제안한다.

제안하는 모델은 크게 두 가지의 기능으로 구성된다. 첫 번째 기능은 딥러닝을 사용한 댓글 분류를 통해 사용자가 남긴 평점과 리뷰가 신뢰할 수 있는지 여부에 대해 판별하는 데이터 정제 기능이다. 해당 정제 기능을 통해 기존의 가공되지 않은 데이터를 신뢰할 수 있는 데이터로 변경한다. 두 번째 기능은 데이터 정제 이후 신뢰할 수 있는 데이터에서 사용자의 주문과 평점 정보를 기반으로 사용자에게 적합한 음식을 추천해 주는 기능이다.

그림 1은 제안한 음식 추천 시스템의 구조이다. 추천 시스템을 학습할때는 사용자의 데이터를 통해 음식을 추천하기 이전에 사용자의 리뷰 데이터가 신뢰할 수 있는 데이터인지 확인한다. 신뢰할 수 있는 데이터인지 확인한 결과로 신뢰할 수 있는 데이터만을 정제한다. 이후 정제된 데이터를 통해 음식 추천을 하는 과정을 진행한다.

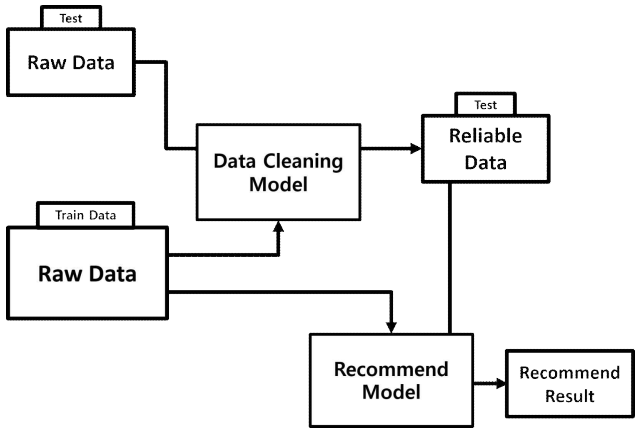
3.1 데이터 정제

데이터 정제에서는 문장 분류 분야에서 사용된 BERT(Bidirectional Encoder Representation from Transformers)를 사용하였다[1]. 사용자의 리뷰 데이터를 형태소 단위로 토큰화하여 입력으로 사용하였으며, BERT의 <CLS>토큰에 FCNN(Fully Connected Neural Network)와 softmax를 통해 해당 리뷰가 포함된 데이터가 신뢰할 수 있는 데이터인지 분류하였다.

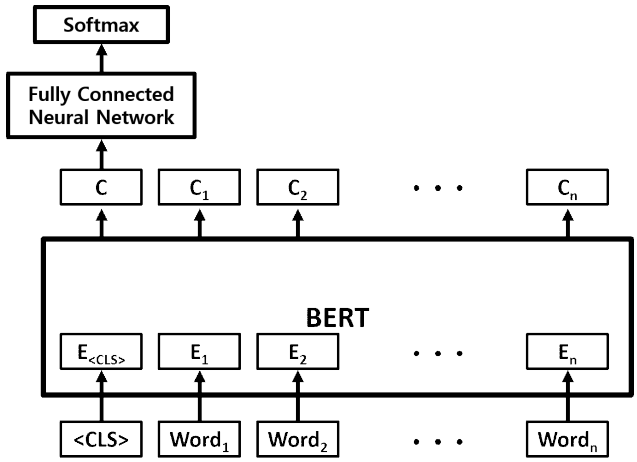
3.2 음식 추천

음식 추천을 하는 과정에서는 implicit feedback factorization model을 이용하였다. 사용자들의 음식 주문 순서를 입력으로 넣어 결과로 모든 음식에 대한 선호도를 예측하였다.

사용한 implicit feedback factorization model의 구조는 다음과 같다[7]. 먼저 음식 구매와 유저의 쌍을 이룬 데이터를 사용한다. 해당 데이터에서 음식 데이터와 유저 데이터들을 bloom embedding을 하여 유저와 음식의 vector representation을 구한다. 음식과 유저 각각의 representation을 dot production을 하였을 때 유저가 해당 음식을 선호하는지에 대한 정보를 예측할 수 있다. 이후 실제 선호도 데이터를 통해 예측한 선호도를 비교하며 학습한다.



<그림 1> 음식 추천 시스템 구조



<그림 2> 데이터 정제 모델

4. 실험

4.1 실험 환경

제안하는 모델을 실험하기 위해 모바일 배달음식 플랫폼 Y사의 음식 리뷰 데이터를 crawling 하여 사용하였다. 데이터는 60개의 음식점에 대한 25,000개의 리뷰를 추출하여 사용하였으며, 리뷰 이벤트 기간동안의 음식점의 리뷰 중 수작업을 통해 신뢰할 수 있는 데이터를 제거하여 신뢰할 수 없는 댓글을 추출하였으며 리뷰 이벤트를 한적이 없는 음식점의 리뷰를 신뢰 할 수 있는 데이터로 추출하였다. 모집한 데이터중 신뢰할 수 없는 데이터는 16,000개이고 신뢰할 수 있는 데이터는 9,000개

이다. 해당 데이터를 7:3 비율로 나누어 train dataset과 development dataset으로 사용하였다.

각각의 리뷰는 하나의 주문에 대한 리뷰이기 때문에 음식 추천을 위해 여러 음식을 주문한 경우 주문을 음식 단위로 나누어 같은 평점을 적용하였다.

5. 실험 및 평가

실험은 데이터 정제 기능과 음식 추천 기능에 대한 성능 평가와 음식 추천 시 데이터 정제 유무에 따른 성능의 변화를 측정하였다. 데이터 정제 성능 지표는 식(6)과 같은 accuracy를 사용하여 정제의 정확도를 평가하였다. 음식 추천의 성능 평가 지표는 식 (7)과 같이 MRR(Mean Reciprocal Rank)를 사용하여 실제 선호도가 높은 음식들과 예측한 선호도의 음식 순위와의 유사성을 판별하여 추천 시스템의 성능을 측정하였다.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i} \quad (7)$$

표 1에서 데이터 정제 기능의 모델 별 성능이다. 각 모델의 데이터 정제 기능의 성능을 accuracy를 통해 비교하였다. BERT 모델은 SKT Brain 팀에서 제작한 한국어로 pre-trained된 모델인 KoBERT-base를 사용하였다.[6] RNN의 마지막 hidden state에 FCNN와 softmax layer를 통해 추출한 classification 성능은 71.32이다. 하지만 BERT의 <CLS>토큰에 FCNN와 softmax layer를 통과시킨 성능은 78.13으로 기존보다 높은 수치를 보였다.

표 2에서는 정제된 데이터를 통해 추천하였을 때 각 모델의 성능을 비교하였다. 추천 모델의 경우 Spotlight 패키지를 사용하여 실험하였다.[7] 실험 결과 음식에 대한 평점을 사용한 Explicit Feedback Factorization에 비해서 이분적인 데이터인 음식 구매 여부에 대한 정보를 사용한 Implicit Feedback Factorization과 Wavenet을 이용한 음식 추천이 높은 성능을 보였다. 그중 Implicit Feedback Factorization이 다른 모델에 비해 높은 성능을 보였다.

표 3은 테스트 데이터 정제 여부에 따른 음식 추천 성능 평가이다. 표 1에서 최고 성능이 나온 BERT와 표 2에서 높은 성능을 보인 Implicit Feedback Factorization과 Wavenet을 사용하여 테스트 데이터 정제에 따른 음식 추천 성능 차이를 실험하였다. 두 추천 모델 모두 테스트 데이터 정제 이후에 정제 이전보다 증가하였으며 정제 이전 추천 모델과 정제 이후 추천 모델 모두 Implicit Feedback Factorization이 높은 성능을 보였다. 표 3의 경우 학습 데이터에 정제가 이루어지지 않아 학습 데이터는 신뢰할 수 없는 정보가 포함된 데이터임에도 불구하고 테스트 데이터의 정제만으로도 음식 추천 성능이 향상된 결과를 얻을 수 있었다.

모델	Accuracy
RNN	71.32
BERT	78.13

<표 1> 모델에 따른 데이터 정제 성능

모델	MRR
Explicit Feedback Factorization [3]	0.012
Implicit Feedback Factorization [4]	0.073
Wavenet[5]	0.024

<표 2> 모델에 따른 음식 추천 성능

모델	MRR
Wavenet[5] w/o 데이터정제	0.020
Wavenet[5] w/ test 데이터정제	0.022
IFF[4] w/o 데이터정제	0.057
IFF[4] w/ test 데이터정제	0.065

<표 3> 테스트 데이터에서의 데이터 정제에 따른 음식 추천 성능

모델	MRR
Wavenet[5] w/o 데이터정제	0.020
Wavenet[5] w/ 데이터정제	0.060
IFF[4] w/o 데이터정제	0.057
IFF[4] w/ 데이터정제	0.130

<표 4> 전체 데이터 정제에 따른 음식 추천 성능

표 4는 테스트와 학습 데이터 모두에서의 데이터 정제에 따른 음식 추천 성능 평가이다. 표 3과 같이 표 1, 표 2에서 최고 성능이 나온 데이터 정제 모델과 음식 추천 모델을 사용했다. 데이터 정제를 한 것이 데이터 정제를 하기 전보다 매우 높은 성능을 얻을 수 있었으며 테스트 데이터만 학습한 표 3에 비해서도 높은 성능을 얻을 수 있었다.

5. 결론

본 논문에서는 음식 리뷰 데이터를 통한 음식 추천 시스템을 제안하였다. 제안한 시스템은 문장 분류를 통해 음식 리뷰 데이터 중 신뢰할 수 있는 데이터를 정제하고 추출한 신뢰할 수 있는 데이터를 통해서 음식을 추천하는 시스템을 설계하였다. 실험을 통해 데이터를 정제한 것이 정제하지 않은 것보다 높은 성능을 나타낸다는 것을 증명하였고 데이터 정제 기능과 음식 추천 기능에서 제안한 모델이 효율적임을 보여주었다.

또한 음식 추천과 데이터 정제에서 효율적인 모델에 대해 알아보았으며 학습 데이터에서 정제를 거치지 않더라도 테스트 데이터의 정제를 통해 성능을 개선할 수 있음을 보여주었다.

향후 연구에서는 더 많은 도메인에서 데이터 정제에 대해 실험한 성능을 측정하며, 제시한 방법과 기존 방법의 추천 결과에 대해 사용자가 느낀 체감 성능을 조사하고 분석할 예정이다.

Acknowledgement

이 성과는 2020년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No.NRF-2019R1C1C1006299)

참고문헌

[1] Devlin, Jacob, et al. “Bert: Pre-training of deep bidirectional transformers for language understanding.” arXiv preprint arXiv:1810.04805, 2018

[2] Sun C., Qiu X., Xu Y., Huang X. How to Fine-Tune BERT for Text Classification?, arXiv preprint arXiv:1905.05583, 2019

[3] Y. Koren, R. Bell, CH. Volinsky, Matrix factorization techniques dor recommender systems, IEEE Computer, 42 (8),pp. 42-49, 2009

[4] Hu, Y. Koren, C.H. Volinsky, Collaborative filtering for implicit feedback datasets, in: IEEE International Conference on Data Mining (ICDM), pp.263, 2008

[5] A.v.d. Oord, et al., Wavenet: a generative model for raw audio, CoRR abs/1609.03499, 2016

[6] https://github.com/SKTBrain/KoBERT

[7] https://github.com/maciejkula/spotlight