한국컴퓨터정보학회 동계학술대회 논문집 제31권 제1호 (2023. 1)

Item2vec과 LSTM을 사용한 추천 시스템 설계

차민수O, 우지영*
O순천향대학교 ICT융합학과,
*순천향대학교 ICT융합학과
e-mail: ckalstn0522@gmail.comO, jywoo@sch.ac.kr*

Recommender System Design with Item2vec and LSTM

Minsu ChaO, Jiyoung Woo*
ODept. of ICT Convergence, SoonChunHyang University,
*Dept. of ICT Convergence, SoonChunHyang University

● 요 약 ●

본 논문에서는 최대 규모의 게임 플랫폼인 Steam에서 수집한 유저 정보 데이터 셋에 Item2vec과 LSTM을 사용하여 추천 시스템을 구현한다. 수집한 유저 정보 데이터 셋에 Item2vec을 적용하여 각각의 유저들이 보유하고 있는 고유한 Appid들을 200차원의 벡터로 변환한다. 그 후 데이터 셋을 기간에 따라 4단계의 시퀀스로 나눈 후 LSTM을 사용하여 유저별로 최대 5가지의 추천 리스트를 생성한다. 유저 정보 데이터 셋은 액티브한 유저 정보를 얻기 위해 Steam 게임 리뷰 항목에서 리뷰를 남긴 유저들의 데이터를 api를 사용해 수집했으며 LSTM을 사용한 실험의 성능 평가 지표는 RMSE를 사용했고 이때의 성능은 0.1357을 얻을 수 있었다.

키워드: Item2vec, 추천 시스템(Recommender System), Skip-Gram with Negative Sampling(SGNS)

I. Introduction

오늘날에는 OTT 서비스나 뉴스, 게임 플랫폼 등을 사용해 누구나 수많은 콘텐츠를 쉽게 접할 수 있게 되었다. 하지만 그만큼 방대한 콘텐츠들 사이에서 만족할 만한 콘텐츠를 찾기에는 쉽지 않은 일이 되어버렸다. 그중에서도 게임 관련 콘텐츠를 향한 소비자들의 니즈는 날이 갈수록 늘어나고 있다. 2016년 기준 25억 명 이상의 사람들이 비디오게임을 즐겨 할 정도로 게임 산업은 거대해졌으며 코로나 사태로 인해 자택에서 보내는 시간이 늘어난 현재의 상황에서는 게임 산업의 중요성이 더 커졌다고 볼 수 있다. 게임 산업은 최근 몇 년간 수많은 장르와 플랫폼의 개발로 높은 다양화를 이루어냈으며 이로 인해 Steam이 최대 규모의 게임 플랫폼으로 탄생하게 되었다.

Steam은 매일 천만 명이 넘게 접속하며 게임 구매뿐만 아니라 다양한 커뮤니티, 의견 공유 등이 가능하다. 하지만 이렇게 다양한 제품과 많은 사용자들이 있다는 사실은 유저들이 특정 새로운 게임을 선택하기 어렵게 만든다. 실제로 2014년 Steam 레지스트리에 따르면 구매한 게임 중 37%가 한 번도 플레이 된 적 이 없다는 결과가 나왔다.[1] 이처럼 낭비되는 콘텐츠의 최소화를 위해 추천 시스템의 중요성은 날이 갈수록 올라가고 있으며 이를 위해 본 논문에서는 Steam 유저 정보 데이터 셋을 사용하여 Item2vec과 LSTM(Long Short-Term Memory)을 적용한 추천 시스템을 제안한다.

II. The Proposed Scheme

실험에 사용한 데이터는 활발한 활동을 보이는 유저들의 정보를 얻기 위해 게임 리뷰를 남긴 유저들로부터 Steam api 사용해 유저 정보 데이터를 수집했다. 유저 정보 데이터 셋에는 유저 별로 부여받은 고유한 Steamid, 보유하고 있는 게임의 식별 Appid, 보유 게임의 총 Playtime, 그리고 최근 2주 동안의 Playtime이 포함되어 있다. 먼저 2개월 동안의 데이터 셋을 사용해 유저들의 2개월 Playtime을 구한 후 이를 통해 Item2vec을 적용하여 데이터 셋에 존재하는 Appid 들을 200차원의 벡터로 변환한다.[2]

Item2vec은 Neural word embedding 방식 중 Skip-Gram with Negarive Sampling 을 Item-based Collaborative filtering에 적용하여 Item을 Embedding vector로 변환하는 기법이다.[3] 이렇게 Appid를 200차원의 벡터로 변환한 2개월간의 데이터를 2주 간격으로 나눠 총 4개의 시퀀스를 만든다. 그리고 한 개의 시퀀스마다 유저별로 시퀀스 기간 내의 Playtime이 가장 높은 상위 5개의 200차원 변환 벡터를 가지도록 인풋 데이터를 생성한다.

이렇게 만들어진 4D 인풋 데이터를 LSTM에 사용하기 위해 ConvLSTM(Convolutional LSTM)을 사용한다.

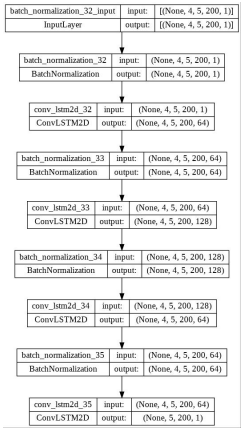

Fig. 1. 설계 모델의 구조

ConvLSTM 층을 지날 때마다 배치 정규화를 진행하며 최적화기는 Adam optimizer를 사용한다. 아웃풋 데이터는 최대 5개의 추천 Appid의 200차원 벡터가 나오게 되며 비교 대상이 되는 타겟 데이터 셋은 인풋 데이터 셋 기준 한 달 후의 유저 정보 데이터 셋을 사용한다. 평가 지표는 RMSE(Root Mean Squared Error)를 사용하며 32, 64, 128개의 배치 사이즈별로 평가 지표를 산출해 가장 우수한 모델을 선정한다.

III. Conclusions

Table 1. 배치 사이즈별 성능평가

Batch size	Validation loss	Validation RMSE
32	**0.0184**	**0.1357**
64	0.0184	0.1358
128	0.0188	0.1371

배치 사이즈 32는 Validation loss 0.0184, Validation RMSE 0.1357이 나왔다. 배치 사이즈 64는 Validation loss 0.0184, Validation RMSE 0.1358로 배치 사이즈 32와 큰 성능 차이가 존재하지는 않는다. 배치 사이즈 128은 Validation loss 0.0188, Validation RMSE 0.1371로 위의 두 모델보다 성능이 떨어지는 모습을 볼 수 있다. 따라서 가장 우수한 성능을 보여주는 배치 사이즈 32모델을 채택한다.

Table 2. 테스트 성능평가

batch size	test loss	test rmse
32	0.0184	0.1357

마지막으로 모델의 결과로 나온 200차원 아웃풋 데이터를 기반으로 코사인 유사도를 사용해 아웃풋 데이터와 가장 유사한 벡터를 가지고 있는 게임을 추천 리스트로 생성한다.

Fig. 2. 유저 데이터 기반 추천 리스트 생성

유저 데이터를 기반으로 추천 리스트를 생성했을 때 분석 기간 기준 한 달 뒤 플레이 한 게임과 유사한 게임을 추천해 주는 추천 리스트를 생성하였다.

ACKNOWLEDGEMENT

본 연구는 교육부의 지자체-대학협력기반지역혁신사업(1345341784)의 지원을 받아 수행되었음

REFERENCES

[1] Germán Cheuque, José Guzmán, Denis Parra. "*Recommender Systems for Online Video Game Platforms: the Case of STEAM,*" WWW '19: Companion Proceedings of The 2019 World Wide Web Conference, pp. 763-771.

[2] Zi Yin, Yuanyuan Shen "*On the dimensionality of word embedding,* " In Advances in Neural Information Processing Systems, pp. 894-905.

[3] Oren Barkan, Noam Koenigstein. *"Item2Vec: Neural Item Embedding for Collaborative Filtering,"* In Machine Learning for Signal Processing (MLSP), 2016 IEEE 26th International Workshop on. IEEE, 1-6.