한국컴퓨터정보학회 동계학술대회 논문집 제30권 제1호 (2022. 1)

비지도 학습 기반 초개인화 추천 서비스를 위한 메타데이터 추출의 중요성 고찰

백주련*, 고광호O
*평택대학교 데이터정보학과,
O평택대학교 스마트자동차학과
e-mail: {jrpaik*, kwanghoO}@ptu.ac.kr

Consideration upon Importance of Metadata Extraction for a Hyper-Personalized Recommender System on Unsupervised Learning

Juryon Paik*, Kwang-Ho KoO
*Dept. of Data Information & Statistics, Pyeongtaek University,
ODept. of Smart Automobile, Pyeongtaek University

● 요 약 ●

서비스 관점에서 구축되는 추천 시스템의 성능은 얼마나 효율적인 추천 모델을 적용하여 심층적으로 설계되었는가에 좌우된다고도 볼 수 있다. 특히, 추천 시스템의 초개인화는 세계적인 추세로 1~2년 전부터 구글, 아마존, 알리바바 등의 데이터 플랫폼 강자들이 경쟁적으로 딥 러닝 기반의 알고리즘을 개발, 자신들의 추천 서비스에 적용하고 있다. 본 연구는 갈수록 고도화되는 추천 시스템으로 인해 발생하는 여러 문제들 중 사용자 또는 서비스 정보가 부족하여 계속적으로 발생하고 있는 Cold-start 문제와 추천할 서비스와 사용자는 지속적으로 늘어나지만 실제로 사용자가 소비하게 되는 서비스의 비율은 현저하게 감소하는 데이터 희소성 문제 (Sparsity Problem)에 대한 솔루션을 모색하는 알고리즘 관점에서 연구하고자 한다. 본 논문은 첫 단계로, 적용하는 메타데이터에 따라 추천 결과의 정확성이 얼마나 차이가 나는지를 보이고 딥러닝 비지도학습 방식을 메타데이터 선정 및 추출에 적용하여 실시간으로 변화하는 소비자의 실제 생활 패턴 및 니즈를 예측해야 하는 필요성에 대해서 기술하고자 한다.

키워드: 추천시스템(recommender system), 메타데이터(metadata), 그림자데이터(digital shadow), 비지도학습(unsupervised learning)

I. Introduction

McKinsey와 Tech Emergency의 통계에 의하면, Amazon 매출의 35%, BestBuy의 23.7%가 추천 시스템에 의해 이뤄지며, Netflix의 경우는 최대 75% 그리고 YouTube는 60%까지 추천 시스템을 통해 기업 매출이 발생한다고 보고 있다[1, 3]. 국내 경우 역시 다르지 않다. 국내 최대 포털인 네이버는 딥 러닝 기반 개인 추천 기능을 앞세워 전체 이용자 중 80%를 AI 솔루션 사용자로 끌어들였으며, Netflix처럼 온라인동영상서비스를 제공하는 SK브로드밴드는 인공지능 기술을 적용하여 고도화된 개인 맞춤형 추천서비스를 제공한다. 서비스 관점에서 구축되는 추천 시스템의 성능은 얼마나 효율적인 추천 알고리즘을 적용하여 심층적으로 설계되었는가에 좌우된다고도 볼 수 있다. 특히, 추천 시스템의 초개인화는 세계적인 추세로 1~2년 전부터 구글, 아마존, 페이스북 등의 데이터 플랫폼 강자들이 경쟁적으로 딥 러닝 기반의 알고리즘을 개발, 자신들의 추천 서비스에 적용하고 있다. 그러나, 여전히 해결되지 못하고 있는 문제와 장애 요소가 존재하며 기업들은 비즈니스의 한계와 요구 사항에 따라 가장 적합한 추천 알고리즘들을 선택한다.[1, 2, 6]

디지털 섀도(Digital Shadow)란 그림자 데이터라고도 지칭되며 디지털 서비스를 제공받은 사용자가 의도했든 의도하지 않았든 네트워크상에 남기게 되는 사용자의 모든 데이터 흔적들을 의미한다. 대표적 생성 사례로 전자우편 발송, 사회관계망서비스 프로필 갱신, 신용카드 판독, ATM 사용 등에서 무수히 많은 그림자 데이터들이 편편히 만들어지며 대다수의 사용자들은 이런 데이터가 잔재하는지 여부도

인식하지 못한다. 사용자 프라이버시와 매우 밀접하게 관계되어 있으면서도 디지털 서비스를 제공받는 사용자의 주변 어디서나 그리고 끊임없이 흔적을 남기며 생성되는 서비스 관점에서 매우 중요한 데이터들이라 할 수 있다.

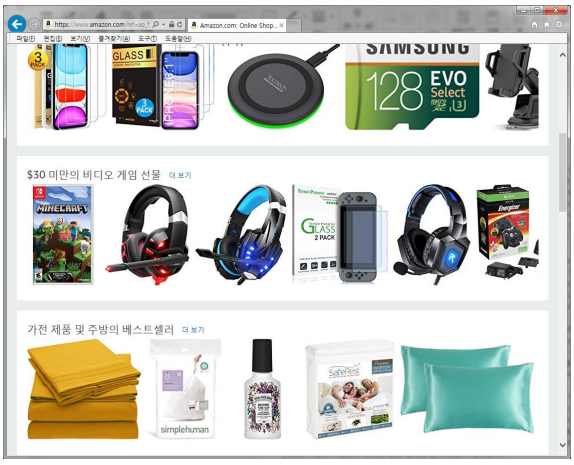
갈수록 고도화되는 추천 시스템에 비해 여전히 해결되지 못하고 있는 여러 문제들 중 신규 사용자 또는 새로운 아이템 및 서비스에 대한 데이터가 거의 없어 필연적으로 발생하는*cold-start 문제*와, 추천할 서비스 및 아이템 그리고 사용자는 지속적으로 늘어나지만 실제로 기존 소비자가 관심을 갖는 서비스의 비율은 현저하게 감소하는 *희소성 문제(sparsity problem)*에 대한 솔루션의 하나로 초개인화 데이터인 디지털 새도를 추천시스템 모델링 단계부터 적용하여 활용하는 방안을 연구하고자 한다.

본 논문은 그 첫 단계로, 적용하는 메타데이터에 따라 추천 결과의 정확성이 얼마나 차이가 나는지를 보이고 딥러닝 비지도학습 방식[7,8]을 메타데이터 선정 및 추출에 적용하여 실시간으로 변화하는 소비자의 실제 생활 패턴 및 니즈를 예측해야 하는 필요성에 대해서 기술하고자 한다.

II. The Proposed Scheme

1. Motivation

신규 사용자 또는 새로운 서비스나 아이템에 대한 데이터가 거의 없어 필연적으로 발생하는 cold-start 문제와 추천할 서비스나 및 아이템 그리고 소비자는 지속적으로 늘어나지만 실제로 기존 소비자가 관심을 갖거나 구매하려는 서비스의 비율은 현저하게 감소하는 희소성 문제 (Sparsity Problem)는 추천시스템이 지속적으로 그리고 필연적으로 만나게 되는 문제점들이다. Fig.1은 Amazon의 첫 화면을 비교한 것으로 위의 그림은 계정 없이 접속한 신규 고객에게 추천 하는 상품들이지만 해당 고객의 관심을 끌만한 상품들은 거의 없으며, 아래의 그림은 계정이 있는 기존 고객의 첫 화면이지만 이미 구매가 끝난 상품들과 연관되거나 아니면 전혀 흥미 없는 상품들(ex. 슬리퍼, 필로우 등)로 추천되어 있다.



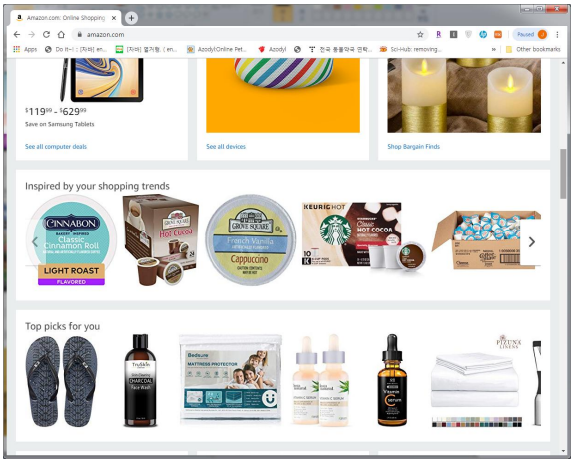


Fig. 1. Amazon Login Page : New User (Top) vs. Old User (Bottom)

기존 사용자 기반의 협업필터링 모델링을 아이템 중심의 협업필터링으로 관점을 바꾸어 추천시스템 모델링에 높은 기여를 한 Amazon 조차도 여전히 해결되지 않는 문제점을 갖고 있는 것이다.

구글 Play 스토어 추천에 사용되면서 높은 효율 개선을 보이는 구글에서 개발한 Wide & Deep 모델의 주요 개발 이유는 cold-start 문제에 대하여 좀 더 능동적으로 대처하고, 추천 리스트가 너무 구체화되거나 너무 일반화되는 것을 적절히 안배하여 판에 박힌 아이템으로만 추천이 발생하는 것을 줄이고자 하였다. 그러면서 PC ⇒ 노트북 ⇒ 모바일 순으로 사람과 정보가 연결되는 도구들이 점점 소형화개인화됨에 따라 개개인의 위치, 시간, 이동, 상황 등 초개인화 데이터를 분석해 적절한 정답을 제시하는 플랫폼 확장 및 고도화 연구도 병행하고 있다. 질의나 지도학습 없이 상황만으로 사용자의 니즈를 파악하여 최적의 서비스까지 연결해주는 추천 시스템 구축을 위해서는 디지털 새도 활용 기술 고도화 경쟁이 필연이다.

패턴을 알 수 없고 끊임없이 변화해서 명확한 레이블을 충분히 확보하기 쉽지 않은 디지털 새도 활용 영역에서는 데이터 자체의 내재된 구조를 학습해 데이터 셋이 보유한 데이터 레코드 수보다 훨씬 작은 수의 매개변수 집합으로 훈련해 유용한 특징(features)으로 매핑하려는 딥러닝 비지도학습이 더 적합한 이유이며 훈련을 위한 매개변수 추출이 매우 중요하게 이루어져야 하는 선과정인 것이다.

2. Importance of Metadata

Cold-start 문제와 Sparsity 문제를 해결하기 위한 방법들의 가장 큰 프레임은 하이브리드 추천시스템을 이용하는 것이며 최근에는 사용자와 서비스 혹은 아이템 간의 비선형적이고 복잡한 관계를 학습하는 딥러닝 모델들과의 다양한 결합한 이루어지고 있는 추세이다. 모델링 결합에 앞서 본 논문에서는 비지도학습 훈련 대상인 매개변수들에 따라 추천시스템의 정확도가 얼마나 차이가 있는지를 먼저 살펴보고자 한다.

사용한 데이터는 MovieLens에서 공개적으로 제공하는 데이터들 중 100K 데이터 셋으로 전체평점 데이터(전체평점의 개수는 약 2500백만로 62,000개의 영화에 대해서 162,000명의 사용자가 평가)

중 1700개의 영화에 대해서 1000명의 사용자가 평가한 100,000개의 평점 데이터를 이용했으며 프로그래밍 코드는 Anaconda와 Spyder를 사용한 파이썬 코드이다. MovieLens 100K 데이터의 총 23개 파일 중 3개의 파일을 사용하였는데 사용자 데이터인 u.user, 영화 데이터인 u.item 그리고 영화평가 데이터인 u.data이다.

메타데이터 추출이 추천시스템의 성능이 얼마나 중요한 영향을 끼치는지 측정하기 위해 RMSE (Root Mean Squared Error)를 계산했으며 메타데이터는 첫째는 성별(Sex)을 대상, 둘째는 직업을 대상, 그리고 세 번째는 성별과 직업을 동시에 고려한 사용자 집단으로 나누어 예측 후 그에 대한 정확도를 계산하였다.

$$SE = \frac{\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m}\left((r_{i,j} - p_{i,j})^2\right) if ri,j \neq \nu ll, 0 otherwise}}{k}$$

$$= \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y_j})}$$

m: test set에 있는 사용자의 수
m: test set에 있는 아이템의 수
$r_{i,j}$: 사용자 i의 아이템 j에 대한 실제 평점
$p_{i,j}$: 사용자 i의 아이템 j에 대한 예상 평점
k: 정확도 계산에 포함된 총 아이템 수 ($null$이 아닌 평점의 개수)

10만개의 영화평점 데이터 중 75%는 학습을 위한 train set으로 사용하고 나머지 25%는 평가를 위한 test set으로 사용하여 세 기준의 메타데이터에 따라 영화를 추천하는 간단한 모델들을 구현 후 정확도를 평가하였다. 다음은 메타데이터에 따른 3가지 추천 모델 평가를 위한 evaluation 함수 부분이다.

```
## test set 으로 추천 모델의 RMSE를 계산하여 train set으로
예측한 값의 정확도 평가하는 함수
def evaluation(model):
    # 예측 대상인 test set으로부터 (user_id,
    # movie_id) 쌍 생성
    id_pairs = zip(X_test['user_id'], X_test['movie_id'])
    # 모든 (user_id, movie_id) 쌍을 대상으로 주어진
    # 모델을 사용해서 예측한 평점 리스트 저장
    y_pred = np.array([model(user, movie) for (user,
             movie) in id_pairs])
    # 실제 평점 추출
    y_true = np.array(X_test['rating'])
    # RMSE 계산 후 반환
    return RMSE(y_true, y_pred)
```

다음 코드는 세 번째 추천 모델인 성별과 직업 메타데이터를 모두 고려하여 추천하는 모델의 코드 부분과 마지막 라인은 evaluation 함수를 적용하여 해당 모델의 정확도를 계산한다.

```
## 성별과 직업 메타데이터 사용 추천 모델
g_o_mean = merged_ratings[['movie_id', 'sex',
           'occupation', 'rating']].groupby(['movie_id',
           'sex', 'occupation'])['rating'].mean()

def cf_gender_occupation(user_id, movie_id):
    if movie_id in rating_matrix:
        gender = users.loc[user_id]['sex']
        occupation = users.loc[user_id]['occupation']
        if (gender in g_o_mean[movie_id]) &
            (occupation in g_o_mean[movie_id]):
          gen_occup_rating = g_o_mean[movie_id]
                             [gender][occupation]
        else:
            gen_occup_rating = 3.0
    else:
        gen_occup_rating = 3.0
    return gen_occup_rating

evaluatedValue = evaluation(cf_gender_occupation)
```

메타데이터에 따른 세 추천 모델의 정확도는 ①성별 메타데이터 모델 1.034, ② 직업 메타데이터 모델 1.123 ③ 성별과 직업 메타데이터 모델 1.241로 계산되었다. 메타데이터 종류가 많으면 더 정확한 추천 결과가 나오는 것이 아니라 사용자의 개인적인 성향을 얼마나 많이 반영하는 메타데이터를 선정하는지 여부가 추천 모델의 정확성을 높인다고 할 수 있다.

데이터 셋의 특징을 담고 있는 많은 메타데이터들 중 모든 사용자의 선호에 맞는 정보를 담고 있는 메타데이터는 존재하지 않는다. 대신, 가장 일반화할 수 있는 메타데이터들을 선정하여 여러 조합과 추천 모델들을 결합하여 가장 높은 정확도를 보이는 조합을 사용할 뿐이다. 메타데이터 선정에 있어서 딥러닝 비지도학습을 사용자의 디지털 새도에 적용한다면 사용자들의 개인적인 성향을 가장 잘 반영하는 메타데이터 선정이 가능하고 이를 기반으로 한 딥러닝 추천모델의 적용은 본 연구의 최종 지향점인 좀 더 초개인화에 근접한 결과를 도출할 것으로 사료한다.

III. Conclusions

본 논문은 갈수록 고도화되는 추천 시스템에 비해 여전히 해결되지 못하고 있는 여러 문제들 중 cold-start 문제와 희소성 문제에 대한 솔루션의 하나로 초개인화 데이터인 디지털 새도를 추천시스템 모델링 단계부터 적용하여 활용하는 방안의 첫 단계로, 적용하는 메타데이터에 따라 추천 결과의 정확성이 얼마나 차이가 나는지를 보이고 딥러닝 비지도학습 방식을 메타데이터 선정 및 추출에 적용하여 실시간으로 변화하는 소비자의 실제 생활 패턴 및 니즈를 예측해야 하는 필요성에 대해서 기술하였다. 현재 메타데이터 선정에 있어서 비지도학습 방안과 여러 딥러닝 모델 방식들에 대한 조합 연구를 진행하고 있다.

ACKNOWLEDGEMENT

이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. NRF-2021R1F1A1064073).

REFERENCES

[1] Shual Z., Lina Y., Aixin S., and Yi T. Deep Learning based Recommender System: A Survey and New Perspectives. ACM Computing Surveys Vol. 1, No. 1. July 2018, pp. 1-35

[2] Hyo-Jeong Lee, Ki-bum Kim, Yeon-kyung Choi. Use of AI Algorithm to Create Business Opportunities. SAMJUNG KPMG ISSUE MONITOR Vol. 84. June 2018.

[3] Markus S., Hmaed Z., and Ching-Wei C. Current Challenges and Visions in Music Recommender System Research. International Journal of Multimedia Information Retrieval Vol. 7. April 2018, pp. 95-116

[4] Raciel Y. T., Yailé C. M., and Luis M. A Recommender System for Programming Online Judges Using Fuzzy Information Modeling. Informatics Vol. 5, No. 17. April 2018, 17 pages.

[5] Balraj K. and Neeraj S. Approaches, Issues and Challenges in Recommender Systesm: A Systematic Review. Indian Journal of Science and Technology Vol. 9, No. 47. December 2016, pp. 1–12.

[6] Jieun Son, Seoung Bum Kim, Hyunjoong Kim, Sungzoon Cho. Review and Analysis of Recommender Systems. Journal of the Korean Institute of Industrial Engineers, Vol. 41, No 2. April 2015, pp. 185–208.

[7] Se Hoon Jung, Jong Chan Kim, Kim Cheeyong, Kang Soo You, Chun Bo Sim. A Study on Classification Evaluation Prediction Model by Cluster for Accuracy Measurement of Unsupervised Learning Data, Journal of Korea Multimedia Society, Vol. 21, No. 7, July 2018, pp. 779-786.

[8] Minsuk Kim, Reinforcement Learning Research and Convergence Technology Trends, IITP ITFIND, March 2021, pp.13-23.