트위터 추천 알고리즘에 대한 소개 및 분석

Introduction and analysis on Twitter recommendations

진희원
컴퓨터공학부
호서대학교
천안시, 대한민국
jhw6525@hanmail.net

정지은
컴퓨터공학부
호서대학교
천안시, 대한민국
heyfer6867@gmail.com

이수인
컴퓨터공학부
호서대학교
아산시, 대한민국
lisuin0802@gmail.com

요 약

최근 Git Hub 에 트위터의 트윗/사용자 추천 알고리즘이 일부 공개가 됐다. 따라서 현재 공개된 일부 코드를 분석하고 트위터 추천 알고리즘의 동작 방식을 본 논문에서 연구한다. 또한 매일 게시되는 약 5 억개의 트윗 중 소수의 상위 트윗을 추출하여 사용자의 메인 페이지 타임라인에 표시하는 추천 알고리즘에 대한 방법을 소개하고자 한다. Home Timeline 에 추천 되는 트윗은 추천 파이프라인 주요기능 세가지를 사용하여 추천되며, Candidate Source 프로세스를 통해 다양한 추천 소스에서 트윗을 가져오고 기계학습모델을 사용하여 각 트윗에 순위를 매긴다. 사용자의 Home Timeline 에 노출시키기 이전에 사용자가 차단하거나, 중복되었거나, 트위터 규정을 위반한 트윗 등을 필터링 한다. 이러한 과정을 거쳐 사용자의 Home Timeline 에 관심사에 맞는 트윗 추천을 하게끔 작동한다.

키워드 : 트위터, 추천 알고리즘, Clustering, 휴리스틱 알고리즘

1. 서론

전 세계 사람들이 사용하는 대표적인 SNS 트위터의 CEO 가 창립자 잭 도시(Jack Dorsey)에서 일론 머스크(Elon Musk)로 바뀌었다. 이때 일론 머스크의 CEO 취임 시 내세운 대표적인 공약이 있는데, 바로 트위터의 일부 소스 코드를 오픈소스로 공개하는 것이다. 최근 개개인의 취향에 맞춘 커스텀 상품을 원하는 소비자 또는 사용자를 따라 SNS 내에서도 사용자 개개인의 취향을 반영하기 시작했다. 대표적으로 You Tube 추천 영상, Twitter 의 추천 트윗, Amazon 의 콘텐츠 기반 추천 시스템, 넷플릭스 추천 영화/드라마 등이 포함된다. 이 중에서 일론 머스크가 오픈소스로 공개한 트위터 추천 시스템 알고리즘 (Twitter recommendation system algorithm)의 구동 방식을 분석하였다.

추천 시스템의 구현 방식에는 협업 필터링 (Collaborative filtering)과 콘텐츠 기반 필터링 (Content-based filtering), 하이브리드 추천 시스템이 있다. 협업 필터링 시스템은 사용자 중심 방법으로, 사용자의 성향이 변하지 않았을 것이라 가정하고 선호도 분석 결과를 기반으로 유사한 성향을 가진 다른 사용자들에게 추천해주는 시스템이다. 이 시스템은 추천 콘텐츠에 대한 이해가 필요하지 않다는 장점이 있지만, 처음 적용할 경우 사용자의 평가가 없는 콘텐츠에 대해서는 추천이 어렵다. 또한 동일한 콘텐츠에 동일한 평가를 한 사용자들의 유사도 평가가 이루어지지 않는다. 콘텐츠 기반 필터링은 사용자가 평가한 콘텐츠를 기반으로 하여 해당 콘텐츠의

성질과 유사한 콘텐츠를 추천하는 방법이다. 이 방법은 콘텐츠의 변화가 거의 없을 때 활용도가 높은 방법이지만 사용자 개개인의 선호만을 고려하기에 주변 사용자들의 선호를 알아내기 어렵다는 단점이 있다. 하이브리드 추천 시스템은 협업 필터링 시스템과 콘텐츠 기반 필터링 시스템의 한계를 극복하기 위한 방법으로, 협업 필터링 시스템과 콘텐츠 기반 필터링 시스템을 독립적으로 사용할 수 있기에 상대적으로 예측 정확도가 높다는 장점이 있다. (박연지, 2020.6)

2. 추천 시스템의 구동(Features)

트위터의 추천 시스템은 세 가지 방법 중 하나를 사용한 것이라 단정하긴 어렵지만, 딥러닝(Deep learning)을 사용한 방법으로, 사용자의 성향 예측에 대한 정확도가 높다. 트윗(Tweet) 추천 과정은 아래 그림과 같다.

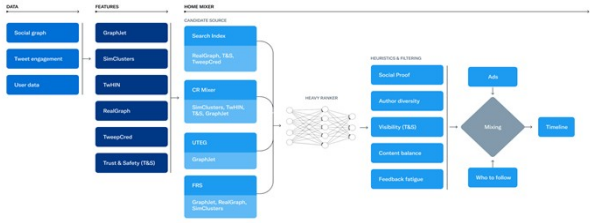

그림 1.트위터 추천 알고리즘의 구동 순서

먼저 트위터에서 사용자의 데이터(User data)와 트윗 참여도(Tweet engagement), 사회적 데이터(Social praph)를 수집한 후, 이를 GraphJet, SimClusters, TwHIN, RealGraph, TweepCred, Trust&Safety 각각의 데이터 모델을 사용하여 분류한다. 각 데이터 모델의 역할은 다음과 같다.

SimClusters Model : user1 이 user2 팔로우하는 경우, user2 가 user1 을 팔로우하는 경우, 서로 팔로우하는 경우를 가정한다. 또한 유명한 사용자들의 트윗을 기반으로 사용자들을 POP, 정치, K-POP 등의 각 커뮤니티(Community)로 분류하여 생산자(Producers)라 하고, 그들을 팔로우하거나 상호 팔로우가 된 사람들을 소비자(Consumers)라 한다. 이 모델은 이분법 그래프를 사용하여 유사한 팔로워를 가지고 있는 생산자 커뮤니티와 유명한 커뮤니티의 생산자를 분류할 수 있다.

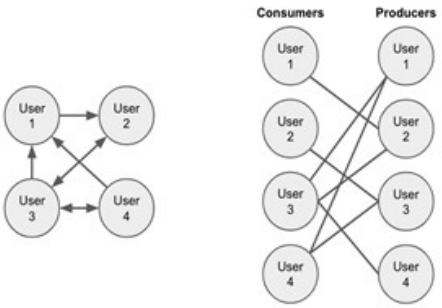

그림 2.SimClusters Model

그리고 생산자의 수(m)소비자의 수(n)의 배열로 나타낸 후, 생산자의 팔로워들 간 유사성을 계산하여 생산자-생산자 사이의 유사성을 cosine 으로 나타낸다. cosine 유사도 값에 따라 가중치를 부여하고, 이 가중치가 임계점 미만이라면 그래프에서 제거하도록 한다.

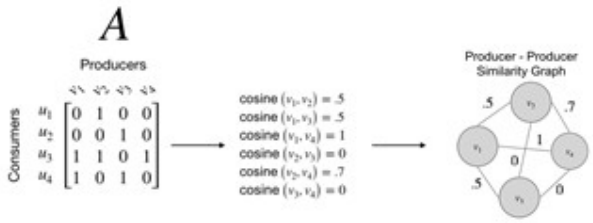

그림 3.생산자-생산자 사이의 유사성 계산

생산자(m)-생산자(m)의 유사성을 계산하였다면 이와 같은 방법으로 생산자(m)-커뮤니티(k)의 유사성을 식별할 수 있다. 결과적으로 이 두 개의 행렬을 이용해 소비자(n)-커뮤니티(k) 간의 유사성도 계산할 수 있다.

$$U_{m\times k} = A_{m\times n} \times V_{n\times k}$$

그림 4.소비자-생산자,생산자-커뮤니티의 연관성을 이용한 사용자-커뮤니티의 연관성

TwHIN : 트위터 내에서 사용자가 팔로우하는 유저, 사용자가 즐겨 찾는 트윗, 사용자가 클릭한 광고 등의 데이터를 기반으로 다양한 추천 시스템 모델을 훈련한다.

Trust & Safety : 성인/포르노 콘텐츠, 모욕, 욕설 등 트위터의 서비스 약관을 위반하는 트윗을 탐지한다.

Real Graph : 트위터에서 사용자 간의 상호작용 횟수를 계산한다. 상호작용에는 즐겨찾기, 리트윗(RT), 팔로우, 프로필 보기, 트윗 클릭 등이 포함된다.

Tweepcred : 구글에서 개발한 PageRank 알고리즘을 사용하여 다른 사용자와의 상호작용을 기반으로 트위터 사용자의 영향력을 계산한다. 이때, 트위터 사용자를 노드(node), 상호작용을 엣지(edge)로 처리하여 사용자의 PageRank 점수를 반복적으로 계산하고 업데이트한다.

2.1 추천 시스템의 구동(Candidate Source)

위와 같은 일련의 데이터 모델을 사용하여 트윗 분류를 마친 후에 Candidate Source 의 Search Index, CR Mixer, UTEG, FRS 알고리즘을 거친다.

Search Index : Apache Lucene[1] 기반으로, 실시간 검색 시스템으로 타임라인 네트워크 내 트윗 검색이 가능하다. 또한 Earlybird[2]를 사용하여 Home Timeline 네트워크 내 트윗 검색 기능을 제공한다.

CR Mixer : 트위터 사용자 개개인에 맞춘 트윗을 제공하기 위한 후보 트윗 생성한다.

UTEG(User Tweet Entity Graph) : GrapJet 프레임워크를 기반으로 구축되었으며, 트위터의 Home Timeline 에서 볼 수 있는 '좋아요' 네트워크 외부 트윗을 생성한다. 사용자의 userID 를 사용하여 트윗에 참여한 사용자 수와 그들의 가중치를 기준으로 최상위 트윗을 반환하며, 이 코드로 트위터의 Home Timeline 에서 RT 또는 '좋아요' 수가 많은 트윗이 가장 위에 뜨게 된다.

FRS(Follow Recommendations Service) : 사용자가 팔로우 할 수도 있는 계정을 제안한다.

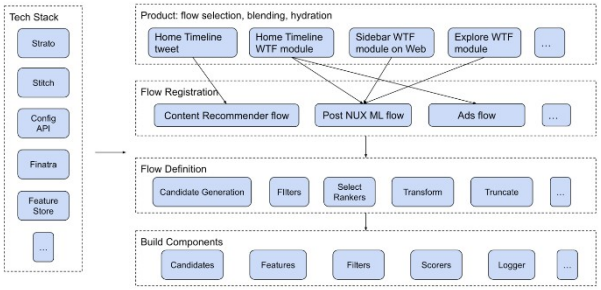

그림 5.FRS 의 개요

WTF(Who To Follow)모듈을 사용하며, FRS 의 개요는 다음 그림과 같다.

Candidate Generation 단계에서 트위터 계정 중 사용자에게 노출할 만한 후보 계정을 찾은 후, Filtering 에서 후보 계정들을 필터링하여 노출 품질을 개선한다. Ranking 에서 노출시킨 후보자들의 계정을 사용자가 팔로우 또는 클릭할 확률을 예측하여 반환하면, Transform 에서 중복된 노출을 제거하거나, 'XX 사용자가 팔로우하는 계정' 등을 추가하여 변형한다. FRS 는 이러한 과정을 구현하고 적용하는 역할을 한다. Search Index, CR Mixer, UTEG, FRS 의 과정을 거친 Candidate Source 는 딥 러닝을 통해 학습시킨 후 사용자의 트위터 Home Timeline 에 나타나게 된다. 이 모든 과정이 불과 약 1.5~2.0 초에 걸쳐 이루어지게 된다.

3. 분석결과

트위터의 추천 알고리즘은 협업 필터링, 콘텐츠 기반의 필터링, 하이브리드의 방법에 딥 러닝을 더한 방법으로 정확도가 상당히 높다는 결과를 내릴 수 있었다. 뿐만 아니라 사용자가 관심을 가질 만한 트윗 또는 다른 사용자를 추천하는 시스템을 통해 사용자가 해당 어플리케이션에 오래 머물 수 있도록 하는 기업 운영 전략도 엿볼 수 있었다.

[1] 정보 검색 라이브러리 오픈소스 소프트웨어

[2] 트위터에서 실시간 뉴스를 제공한다.

4. 결론

본 연구는 단순히 트위터의 추천 트윗 운영이 어떠한 방법으로 운영되는 지 의문을 가져 시작되었으며, 오픈 소스로 공개한 알고리즘은 오랜 기간동안 SNS 를 운영하며 쌓아온 데이터에 딥 러닝을 더한 트위터라는 기업만의 운영 전략 이라는 결과를 도출해 낼 수 있었다. 또한 트위터의 추천 알고리즘의 공개는 부정적 측면과 긍정적 측면을 가지고 있다는 시사점을 찾을 수 있었는데, 부정적 측면으로는 불특정 다수가 트위터 소스 코드를 그대로 복제하여 이름만 바꾼 버전을 출시 할 위험이 있다. 하나의 SNS 운영에 주된 영향을 미치는 코드를 오픈 소스화 하는 것이 트위터 외의 SNS 견제와 트위터의 지적 재산권에는 도움이 되지 않는다는 의미이다. 여기에 더해 코드의 오픈 소스화는 중대한 보안 관련 위험을 초래할 수 있다. 코드를 오픈 소스로 공개할 시, 취약점이 드러난다 하더라도 개발자가 즉시 알아차리기는 어렵기 때문이다. 때문에 오픈 소스로 공개하고자 할 경우 개발자의 의견을 고려하여 특정한 부분만 공개하는 것이 안전한 편이다.

긍정적인 측면으로는 이 트위터 추천 알고리즘을 바탕으로 많은 개인 또는 기업에서 사용할 가능성과 이로 인한 트위터의 새로운 경쟁사의 발굴을 노릴 수 있다는 점이 있다. 다양한 SNS 가 출시되는 시기에 새로운 경쟁사의 출현이란 기업의 발전과 소비자에게 있어 긍정적인 영향을 미치기 때문이다. 또한 공개된 코드를 이용하여 취약점을 보완하고 개선하여 사용자에게 더욱 더 개선된 사용 경험을 제공할 수 있다는 장점이 있다.

이 외에도 여러 부정적인 측면과 긍정적 측면이 존재하지만 개발자가 공개된 코드를 어떻게, 어떤 의도로 사용하는 지에 따라 무게를 두는 쪽이 다를 것이라 생각된다.

Acknowledgement

본 연구는 2023 년 과학기술정보통신부 및 정보통신기획 평가원의 SW 중심대학사업의 연구결과로 수행되었음(2019-0-01834)

참 고 문 헌

박연지, “유사도 검사 알고리즘을 이용한 추천시스템 설계 및 구현에 관하여”, 석사학위논문, pp.3~8, 2020. 6.

참고사이트
https://blog.twitter.com/engineering/en_us/topics/open-source/2023/twitter-recommendation-algorithm

https://www.technologyreview.kr/%ED%8A%B8%EC%9C%84%ED%84%B0-%EC%95%8C%EA%B3%A0%EB%A6%AC%EC%A6%98%EC%9D%84-%EC%98%A4%ED%94%88%EC%86%8C%EC%8A%A4%ED%99%94%ED%95%98%EB%A0%A4%EB%8A%94-%EC%9D%BC%EB%A1%A0-%EB%A8%B8%EC%8A%A4%ED%81%AC/