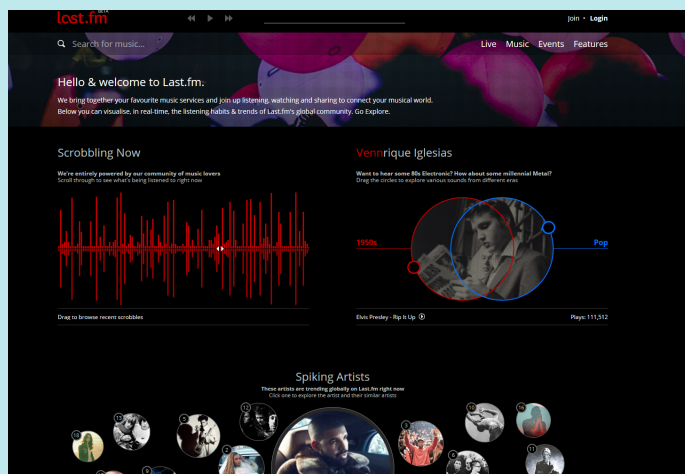


우리는 원하는 정보를 보고 있을까?

콘텐츠 추천 알고리즘의 진화

최근 세상을 깜짝 놀라게 한 알파고(AlphaGo)가 내가 좋아할 만한 영화나 드라마를 추천해준다면? 넷플릭스의 성공 신화는 콘텐츠 추천 서비스에서 비롯되었다. 이용자의 성향과 관심사를 분석하고 콘텐츠 이용 행태를 세밀히 관찰하여 개인 맞춤형 콘텐츠를 제공하는 것은 이제 콘텐츠 사업자에게 선택이 아닌 필수로 간주된다. '나의 취향과 특성을 나보다 더 잘 아는' 콘텐츠 추천 알고리즘의 작동 원리와 발전 양상을 짚어본다.

서봉원 서울대학교 융합과학기술대학원 교수



사용자의 음악 청취 행태를 바탕으로 음악을 추천하는 라스트에프엠(출처: 라스트에프엠 홈페이지)

온라인에서 추천은 이미 광범위하게 사용되고 있는 기술이다. 온라인 쇼핑 시 구매자가 관심 있을 만한 물건을 보여주거나, 뉴스 기사를 읽을 때 관련 기사의 목록을 제공하는 등 익숙한 서비스뿐만이 아니다. 페이스북에서 사용자가 알 만한 온라인 친구들을 알려주는 등 광범위한 곳에서 추천 알고리즘이 사용되고 있다. 이러한 추가적인 콘텐츠들은 때로는 너무 과다하게 제공되는 경우도 있고, 때에 따라 사용자에게 요긴하게 이용되는 경우도 있을 것이다. 이 글은 이러한 콘텐츠 추천이 어떠한 원리로 구현되는지 개략적으로 알아보고 최근 관련 기술의 동향을 살펴보고자 한다.

고전적 추천 시스템

협업 필터링

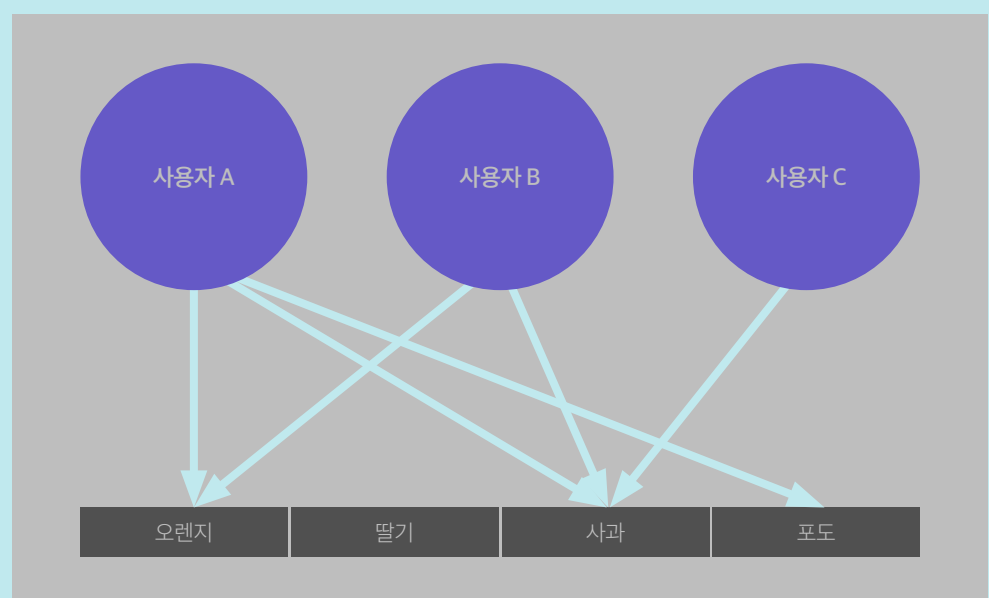
추천 시스템은 사용자가 선호할 만한 아이템을 추측함으로써 여러 가지 항목 중 사용자에게 적합한 특정 항목을 선택(information filtering)하여 제공하는 시스템을 일컫는다. 여기서 '필터링'이란 여러 가지 항목 중 적당한 항목을 선택하는 기술을 말하는 IT 용어이다. 최근의 기술 발전에 따라 여러 가지 새로운 기술이 사용되기는 하지만, 기본적인 추천 시스템은 협업 필터링(Collaborative filtering)과 콘텐츠 기반 필터링(Content-based filtering)을 기반으로 한다.

협업 필터링이란 대규모의 기존 사용자 행동 정보를 분석하여 해당 사용자와 비슷한 성향의 사용자들이 기존에 좋아했던 항목을 추천하는 기술이다. 가장 일반적인 예는 온라인 쇼핑 사이트에서 흔히 볼 수 있는 ‘이 상품을 구매한 사용자가 구매한 상품들’ 서비스이다. 예를 들어 ‘라면’을 구입한 사용자가 ‘생수’를 구입한 경우가 많으면 ‘라면’을 구입하는 구매자에게 ‘생수’를 추천하는 경우이다.

이 알고리즘은 결과가 직관적이며 항목의 구체적인 내용을 분석할 필요가 없다는 장점이 있다. 이 경우는 라면과 생수가 식품인지 아닌지, 서로 같이 사용되어야 하는 관계인지 분석할 필요가 없다. 다만 사용자가 두 제품을 같이 구매했다는 기록을 바탕으로 새로운 사용자에게 추천한다. 이러한 전략을 사용하는 경우, 비슷한 패턴을 가진 사용자나 항목을 추출하는 기술이 핵심적이며 행렬분해(Matrix Factorization), k-최근접 이웃 알고리즘 (k-Nearest Neighbor algorithm; kNN) 등의 방법이 많이 사용된다.

위의 예에서 나타나듯, 협업 필터링을 위해서는 반드시 기존 자료를 활용해야 한다. 하지만 이러한 자료들을 사용자에게 직접 요구해야만 하는 것은 아니다. 협업 필터링은 사용자들이 자연스럽게 사이트를 사용하면서 검색을 하고, 항목을 보고, 구매한 내역을 사용할 수 있는 장점이 있다. 세계 최대의 온라인 소매 업체인 아마존이 이러한 전략을 사용한 바 있으며, 유명한 음악 서비스인 라스트에프엠(Last.fm)도 사용자의 음악 청취 행태를 바탕으로 음악을 추천하고 있다. 페이스북이나 링크드인도 사용자와 친구들의 유사성을 바탕으로 새로운 친구나 그룹을 추천한다.

그러나 협업 필터링은 몇 가지 단점이 있다. 먼저 콜드 스타트(Cold Start)라고 일컬어지는 문제이다. 협업 필터링은 기존의 자료가 필요한바, 기존에 없던 새로운 항목이 추가되는 경우는 추천이 곤란해진다. 예를 들어 음악 서비스의 경우, 신곡이 발표되면 이를 추천할 수 있는 정보가 쌓일 때까지 추천이 어려워지는 것이다. 콜드 스타트란 말 그대로 ‘새로 시작할 때 곤란함’을 의미한다. 협업 필터링 외에 위키 같은 협업 시스템에서 초기 정보 부족의 문제점을 일컫기 위해 사용되기도 한다.



협업 필터링: 사용자 A와 사용자 B가 사과와 오렌지를 같이 구매했으므로, 사과를 산 사용자 C에게 오렌지를 추천한다.

두 번째, 협업 필터링은 계산량이 비교적 많은 알고리즘이므로 사용자 수가 많은 경우 효율적으로 추천할 수 없는 단점이 있다. 앞서 간단하게 소개한 행렬분해의 경우, 사용자 수가 커짐에 따라 계산이 몇 시간에서 며칠까지 걸리는 경우가 종종 생긴다.

마지막으로 롱테일(Long tail) 문제이다. 시스템 항목이 많다 하더라도 사용자들은 소수의 인기 있는 항목에만 관심을 보이기 마련이다. 따라서 사용자들의 관심이 적은 다수의 항목은 추천을 위한 충분한 정보를 제공하지 못하는 경우가 많다. 이러한 비대칭적 쏠림 현상이 일반적이라는 사실은 크리스 앤더슨(Chris Anderson)이나 클레이 셔키(Clay Shirky) 등이 일찍이 밝힌 바 있다. 다시 말해 추천 시스템이 관리하는 항목이 많은 경우, 협업 필터링은 한계가 있을 수 있다.

콘텐츠 기반 필터링

콘텐츠 기반 필터링은 위와 같은 협업 필터링과는 다른 방법으로 추천을 구현하는 방법이다. 협업 필터링이 사용자의 행동 기록을 이용하는 반면, 콘텐츠 기반 필터링은 항목 자체를 분석하여 추천을 구현한다. 예를 들어 음악을 추천하기 위해 음악 자체를 분석하여 유사한 음악을 추천하는 방식이다.

콘텐츠 기반 필터링을 위해서는 항목을 분석한 프로파일(item profile)과 사용자의 선호도를 추출한 프로파일(user profile)을 추출하여 이의 유사성을 계산한다. 유명한 음악 사이트인 판도라(Pandora)의 경우, 신곡이 출시되면 음악을 분석하여 장르, 비트, 음색 등 약 400여 항목의 특성을 추출한다. 그리고 사용자로부터는 'like'를 받은 음악의 특색을 바탕으로 해당 사용자의 프로파일을 준비한다. 이러한 음악의 특성과 사용자 프로파일을 비교함으로써 사용자가 선호할 만한 음악을 제공하게 된다.

이 기법은 콘텐츠의 내용을 분석해야 하므로 아이템 분석 알고리즘이 핵심적이며, 이를 위해 군집분석(Clustering analysis), 인공신경망(Artificial neural network), tf-idf(term frequency-inverse document frequency) 등의 기술이 사용된다.

콘텐츠 기반 필터링은 내용 자체를 분석하므로 협업 필터링에서 발생하는 콜드 스타트 문제를 자연스럽게 해결할 수 있다. 하지만 다양한 형식의 항목을 추천하기 어려운 단점이 있다. 예를 들어 음악과 사진, 비디오를 동시에 추천해야 하는 경우, 각각의 항목에서 얻을 수 있는 정보가 다르기 때문에 프로파일을 구성하기 매우 어려워진다.

이와 같은 고전적 추천 알고리즘은 2000년대 초반까지 많은 분야에서 사용되었다.

넷플릭스 프라이즈와 모델 기반 협력 필터링

최근 한국에 상륙한 넷플릭스는 1997년 DVD 대여 사업으로 시작했다. 최근에는 온라인에서 바로 시청 가능한 스트리밍 서비스를 주력으로, 전 세계 7천5백만 명의 사용자를 확보하고 미국에서만 7조 원 이상의 매출을 올리는 대표적인 온라인 콘텐츠 제공자가 됐다. 넷플릭스는 전체 영화 시청의 75%가 추천을 통해 이루어지는 것으로 알려진 만큼, 추천 시스템의 정확도가 회사의 핵심적인 경쟁력이며 이를 위해 많은 투자를 하고 있다.

이 중 100만 달러의 상금을 걸고 추천 알고리즘을 공모한 넷플릭스 프라이즈(Netflix Prize)가 많은 관심을 모았다. 실제 데이터 중에서 48만 명의 사용자가 1만8천여 개의 영화에 대해 작

NETFLIX

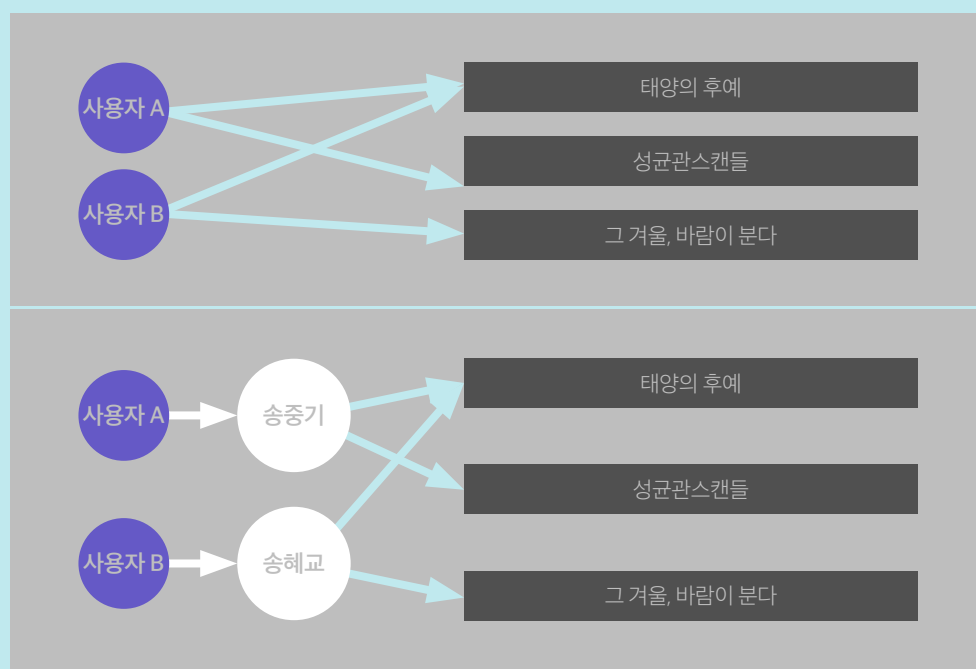
성한 1억여 개의 평가를 제공하고, 사용자가 좋아할 만한 영화를 추천하는 알고리즘을 일반 대중 대상으로 제한 없이 공모한 것이다. 기존의 알고리즘보다 10% 이상 개선을 이루는 알고리즘에 상금 100만 달러를 주는 이 상은 2007년에 시작되어 수많은 추천 알고리즘 연구를 촉발했다. 상금에서 보이듯 목표를 달성하기 매우 어려울 것으로 예상되었으나, 2년 후인 2009년에 실제로 이를 만족하는 알고리즘이 개발되었다.

수상한 알고리즘은 앞서 설명한 협력 필터링에 기반하여 이를 고도화한 것이다. 이러한 기술을 모델 기반 협력 필터링(Model-based Collaborative Filtering algorithm)이라 한다. 모델 기반 협력 필터링은 기존 항목 간 유사성을 단순히 비교하는 것에서 벗어나 자료 안에 내재한 패턴을 이용하는 기법이다.

이 중 한 방법으로, 연관되는 자료의 크기를 동적으로 변화시키는 방법이 있다. 예를 들어 영화를 추천하는 경우, ‘해리 포터’ 시리즈 2편을 추천하기 위해서는 ‘해리 포터’ 시리즈 1편, 단 한 편을 좋아했는가가 다른 무엇보다 중요한 요소이다. 하지만 <주토피아>를 추천하기 위해서는 많은 수의 유사한 영화를 고려해야 한다.

이처럼 추천을 위한 자료의 크기를 변화시키는 방법도 있으나 현재 제일 주목받고 있는 추천 알고리즘은 잠재(latent) 모델에 기반을 둔 방법이다. 잠재 모델이란 사용자가 특정 항목을 선호하는 이유를 알고리즘적으로 알아내는 기법이다.

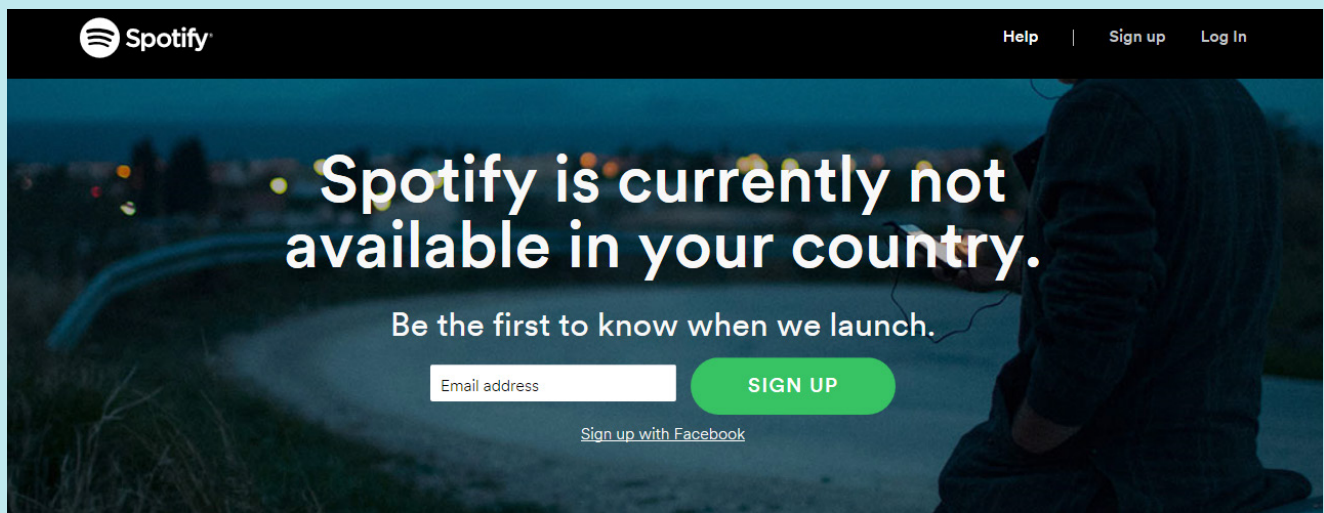
예를 들어 어느 사용자가 <태양의 후예>라는 드라마를 좋아하는 경우, 이 정보를 단순히 그대로 사용하는 것이 아니라, 주위의 정보를 이용해 선호 이유를 유추하는 것이다. 그 사용자는 <태양의 후예>를 주연배우 때문에 좋아할 수도 있고, 드라마 OST가 좋아서 선호할 수도 있으며, 액션 멜로 장르를 선호해서 선택할 수도 있다. 많은 양의 정보를 분석함으로써 이러한 이유를 알아내고, 이를 추천에 이용하는 것이다.



잠재 모델의 예: 위의 그림은 사용자 선호 정보를 그대로 사용하지만 아래 그림은 추론된 잠재 노드(latent node)를 사용함으로써 추천의 정확도를 높일 수 있다.

모델 기반 협력 필터링은 이러한 세부적 정보를 유추함으로써 높은 정확도로 항목을 추천할 수 있다. 추천의 이유를 직관적으로 사용자에게 전달함으로써 추천의 신뢰성도 높일 수 있어, 현재 활발히 연구되고 있다. 하지만 이러한 모델을 만들어내는 데는 매우 많은 계산이 필요하고, 이에 따라 즉각적인 추천이 어려울 수 있다. 모델 기반 협력 필터링은 자료에 내재되어 있는 패턴을 알아내는 것이 핵심적인 기술이며, LDA(Latent Dirichlet Allocation), 베이지안 네트워크(Bayesian Network) 등의 알고리즘이 사용된다.

이와는 다른 방식으로, 최근 주목받는 딥러닝(Deep Learning) 기술에 기반을 둔 새로운 알고리즘들이 여러 분야에서 놀라운 진전을 보이고 있다. 음악 서비스인 스포티파이(Spotify)가 협력 필터링에 딥러닝 기술을 적용한다고 알려져 있으며, 구글은 추천을 위한 텍스트를 자동으로 생성하기 위해 딥러닝 기술을 사용하고 있다. 최근에 사람을 놀라게 한 알파고의 경우에서 보듯, 딥러닝에 기반을 둔 알고리즘들이 개발될 것으로 예상된다.




Spotify is currently not available in your country.

Be the first to know when we launch.


Email address [SIGN UP](#)

[Sign up with Facebook](#)


Spotify gives you instant access to millions of songs – from old favorites to the latest hits. Just hit play to stream anything you like.




Listen everywhere
Spotify works on your computer, mobile, tablet and TV.



Unlimited, ad-free music
No ads. No interruptions. Just music.



Download music & listen offline
Keep playing, even when you don't have a connection.

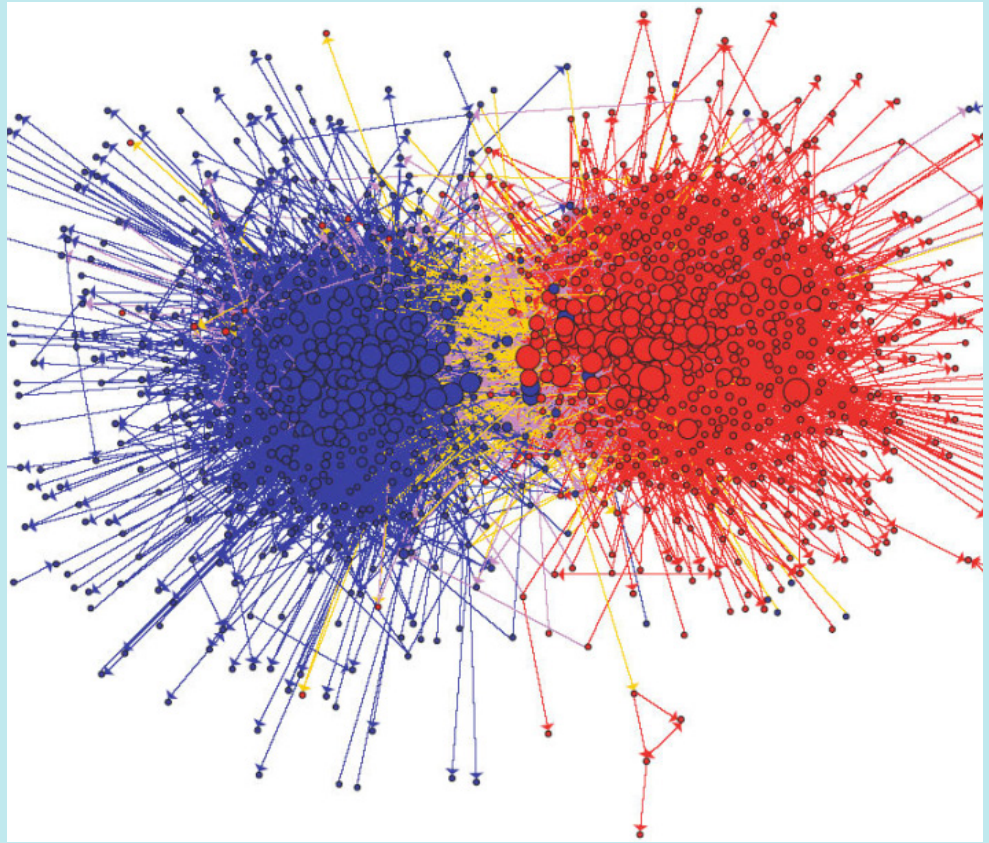


Premium sounds better
Get ready for incredible sound quality.

협력 필터링에 딥러닝 기술을 적용한다고 알려진 음악 서비스 스포티파이(출처: 스포티파이 홈페이지)

추천 알고리즘과 필터버블

추천 알고리즘은 음악, 영화, 광고, 온라인 뉴스, 친구 소개 등 거의 모든 온라인 시스템에서 사용되고 있다. 앞서 소개한 바와 같이 추천 시스템은 정보를 추려서(filtering) 사용자에게 제공한다. 다시 말하면 사용자가 전체 정보를 볼 기회를 박탈당할 수도 있다는 말이다. 추천 시스템이 고도화될수록 사용자의 입맛에 맞는 정보만 제공되고 나머지 정보는 감추어지는 위험이 생기는데, 이러한 현상을 필터버블(filter bubble)이라고 한다.



양극화된 정치 블로그. 푸른색은 미국 민주당을 지지하는 블로그이며 붉은색은 공화당을 지지하는 블로그이다.
(출처: Adamic & Glance, 2005)

상품이나 광고의 경우에는 이러한 현상의 심각성이 크지 않을 수 있다. 하지만 뉴스나 정보의 경우, 예기치 않은 정보의 차단이 큰 문제가 될 수도 있을 것이다. 예를 들어 본인의 정치적 입맛에 맞는 뉴스만 계속 추천받아 보는 경우가 가능해지는 것이다. 보고 싶은 정보만 보고, 보기 불편한 정보는 자동으로 건너뛰는 것이 기술적으로 가능해지면서 야기될 수 있는 정보의 편향적 제공은 극단적인 양극화와 같은 사회적 문제를 가져올 수도 있다. 2014년에 페이스북 연구진이 PNAS에 발표한 논문(Kramer et al. 2014)에서 보이듯, 추천되는 정보에 따라서 사용자의 감정도 조절할 수 있다는 사실이 밝혀지기도 했다.

추천 알고리즘뿐만 아니라 새롭게 개발되는 신기술들은 단순한 기술적인 문제를 해결하는 데서 벗어나 이미 우리의 생활과 정서를 좌우할 수 있는 경지에까지 이르고 있다. 추천된 항목을 얼마나 신뢰하며 어떻게 사용해야 하는가, 다시 한번 심각하게 생각해볼 시점이다. K