기계 학습에 기초한 추천시스템 알고리즘에 대한 연구

하승인(서강대학교 경영연구소)
유영범(서강대학교 경영대학원)
정예숙(서강대학교 경영대학원)

기계 학습에 기초한 추천시스템 알고리즘에 대한 연구

Ha, SeungYin(First Author)
서강대학교 경영연구소

Yoo, Yiung Bum(Co Author)
서강대학교 경영대학원

Jung, Ye Suk(Corresponding Author)
서강대학교 경영대학원

Abstract

Online content service providers are using recommendation systems as part of their efforts to increase sales. The recommendation system identifies and recommends the customer 's preferred content, and it helps the customer to increase the satisfaction and the loyalty of the service by using the content suitable for the user' s taste without searching the content.

In this study, we propose an algorithm for selecting recommendation contents for individual customers by using the Movie Lens data. The algorithms used in the existing recommendation systems have the disadvantage that they can not utilize contents that do not exist in the data since the important words are selected from the given data and the contents are selected based thereon. On the other hand, the Latent Dirichlet Allocation algorithm is that can utilize potential keywords that are not in the data.

Keywords: Recommendation System, Machine Learning, Topic Modeling, Latent Dirichlet Allocation

I. 서 론

인간이 7개 이상의 항목을 동시에 비교하는 경우 인지적 한계로 인해 기억력과 변별력이 보다 떨어지는 현상을 매직넘버 7이라 하며, 한 번에 처리할 수 있는 개체

의 한계를 나타내는 용어로 사용되고 있다(Miller et al., 1956). Iyengar et al(2000)은 대안이 많은 경우보다 적은 경우에 선택이 보다 잘 이루어지며 만족도도 높음을 확인하였는데, 두 집단에게 각각 6개와 30개의 초콜릿에서 하나의 제품을 선택하게 하고 만족도를 확인한 결과 6개가 제공된 집단의 만족도가 더 높으며, 실험 참가대가로 5달러를 받거나 초콜릿을 가져가는 것을 선택하게 한 경우에도 6개가 제공된 집단이 현금보다 초콜릿을 가져가는 비율이 훨씬 높았다. 이처럼 선택해야 하는 항목이 많은 것 보다 적절한 경우 제품의 선택 및 만족도가 높아지는 현상을 선택 대안의 과잉이라 하며, 인간은 인지능력의 한계로 복잡한 것 보다 적절하게 단순한 것을 선택하게 되는 것이다(민재형, 2014).

선택 대안의 과잉을 해소하여 매출을 높이기 위한 일환으로 온라인 콘텐츠 제공업체들은 추천시스템을 활용하고 있다. De Vriendt et al(2011)은 조사에 참여한 영국인 50%이상이 고객추천시스템에 관심을 보였으며, 고객의 43%는 만족하는 콘텐츠 추천 시 비용을 지불할 수 있음을 확인하였다. 이와 같이 추천시스템은 고객이 선호하는 콘텐츠를 파악한 후 추천하여, 고객이 별도의 콘텐츠 검색을 하지 않고 취향에 맞는 콘텐츠를 사용함으로써 만족도를 높이고 서비스에 대한 충성도를 높이는데 기여하고 있다. 아마존은 책과 CD 등 제품을 고객에게 추천하고 있으며(Adomavicius & Tuzhilin, 2005), 특히 넷플릭스는 90초 이내에 고객이 관심 있을만한 영상을 자세히 리뷰하여 시청을 지속하게 하며, 각 세부그룹별로 추천영상을 제시하거나 고객에게 특정화된 장르의 영상을 제시한다. 영상을 시청한 이후에도 영상을 잠시 중단한 것인지, 보기 싫어서 이탈했는지를 파악하여 지속적으로 시청이 가능하도록 영상을 추천해주고 있다. 이러한 넷플릭스의 추천시스템은 알고리즘의 지속적인 개선으로 추천항목의 정확도를 높이고 있다(Gomez-Uribe & Hunt, 2016).

추천시스템은 이미 광범위하게 활용되고 있으나 관련 연구들이 대부분 협업필터링에 관심이 쏠려있는 것에 비해 잠재 디리클레 할당을 통한 연구는 미비하나 실정이다. 본 연구는 기계학습에 대한 연구로, 잠재 디리클레 할당을 통해 추천시스템의 활용 가능성을 제시하고자 한다. 구체적으로 기존 군집화관련 방법론은 고객과 콘텐츠 그리고 이의 횟수 등의 변수에 대한 거리를 정의하기가 어려운 문제점이 있는 반면 잠재 디리클레 할당은 변수별 거리를 측정하지 않아도 활용 가능하며 보다 다양한 고객의 특성을 반영할 수 있을 것으로 보고 추천시스템 알고리즘으로 활용할 수 있음을 실증하고자 한다.

II. 이론적 배경

2.1 기계학습(machine learning)

기계학습은 복잡한 데이터를 이해할 수 있는 지식으로 변환하기 위한 알고리즘을 연구하는 것으로 크게 지도학습(Supervised Learning)과 비지도학습(Unsupervised Learning)으로 구분할 수 있으며, 학습방법에 따라 분류, 수치예측, 패턴파악, 군집화 등으로도 세분화할 수 있다. 지도학습은 예측하기 위한 목표변수와 다른 속성을 갖는 변수들 사이의 관계를 일반화된 모형으로 만드는 방법으로, 예측하려는 값이 존재하므로 학습의 목적이 명확하게 주어진다. 즉 지도학습으로 목표변수와 각 변수들 간의 관계를 수량화하여 범주형이나 수치형 값을 예측할 수 있다. 비지도학습은 예측을 위한 목표변수가 없어 학습의 목적이 명확하게 주어지지 않으며, 오직 데이터 사이의 관계를 나타내는 패턴을 찾아내는데 사용된다. 비지도학습은 데이터에서 유사한 그룹으로 구분되는 군집화과정에 많이 활용되며, 연구자에 따라 군집화이후 해석되는 결과가 다르게 나타날 수 있다. 이외 목표변수가 존재하는 데이터와 존재하지 않는 데이터 모두를 사용하는 준지도학습(Semi-Supervised Learning)이 있으며, 목표변수가 없는 데이터가 대부분이며 목표변수가 있는 데이터는 적은 경우 사용하는 방법이다. 또한 선택 가능한 방법들에서 성공과 실패를 반복하여 최적의 안을 탐색하는 강화학습(Reinforcement Learning)으로 구분할 수 있다.

2.2 추천시스템(recommendation system)

추천시스템은 페이스북이나 링크드인의 알 수도 있는 사람, 아마존의 추천시스템 A9, 넷플릭스의 맞춤형 영화 순위 등 다양한 서비스영역에서 활용되고 있으며 그 활용영역이 점차 확대되고 있다. 즉 고객이 온라인상에서 제공되는 여러 콘텐츠들 중 하나를 선택하려고 할 때 보다 나은 결정을 내리 수 있도록 도와주는 목적으로 고객의 이전 콘텐츠 소비기록이나 콘텐츠별 정보를 통해 개개인마다 특화된 추천정보를 제공해 주는 것이다.

이를 위한 추천시스템에서 주로 활용되는 알고리즘은 협업 필터링(collaborating filtering), 콘텐츠기반 추천(Content-Based Recommendation) 등이 활용된다. 협업 필터링은 콘텐츠 소비성향이 비슷했던 고객이라면 이후에도 비슷한 성향을 가질 것이라 전제하고 성향이 유사한 고객끼리 구성하고 선호도를 활용하여 아직 소비하지 않은 콘텐츠를 교차 추천하거나 연관성이 높은 콘텐츠를 추천하는 것이다. 콘텐츠기반 추천은 콘텐츠 자체의 특성과 고객의 콘텐츠 소비정보를 분석하여 각 고객별로 소비정보와 일치하는 콘텐츠를 추천하는 것이다. 하지만 이들 추천시스템은 신규 콘텐츠 추가로 인한 데이터의 한계로 고객 간의 유사성을 발견하기 어렵거나 다양한 항목의 콘텐츠를 추천하기 어려운 단점이 있다.

III. 연구의 설계 및 측정방법

3.1 표본 선정

정교하고 과학적인 추천을 통해 콘텐츠 소비량을 증가시키는 것은 기업의 매출 증가 이외에도 고객의 탐색시간을 줄여 고객만족을 높일 수 있는 근거가 될 수 있다. 따라서 본 연구는 실제 고객의 데이터를 기반으로 추천항목을 자동으로 제시하는 것이 목적이므로, 무비렌즈의 데이터 중 2017년 26,260명의 영화정보 1,273,631개를 최초 모집단으로 설정하였다. 이 중 1개월 동안 시청편수가 4편미만인 경우 서비스 이용도가 매우 낮아 제외하였으며, 120편 초과의 경우 정상적인 영화시청으로 간주할 수 없어 분석대상에서 제외하였다. 또한 고객성향은 서비스 이용도에 따라 차별화되어 나타날 것이므로, 일정기간동안 영화를 시청하는 편수에 따라 분석집단을 세분화하였다. 1개월 동안 4편 이상 8편 이하의 영화를 시청한 경우 매주 1편에서 2편 이하의 영화를 시청하는 것으로 간주하여 서비스이용도가 낮은 고객군으로 구한 5,443명, 8편 초과 30편 이하는 서비스이용도가 중간인 고객군으로 6,895명, 30편 초과 120편 이하는 서비스이용도가 높은 고객군으로 4,728명으로 구분하였다. 마지막으로 2017년 1월에서 5월까지의 데이터를 훈련데이터로 이후 6월에서 7월까지 2개월간의 데이터를 검증데이터로 사용하였다.

본 연구는 전략적으로 세분화된 집단을 고객성향에 따라 다시 세부적으로 구분하여 콘텐츠를 추천하는 것이 목적이므로, 이를 위한 주제어 선정을 위해 영화가 속하는 장르의 개수 및 장르, 영화 개봉년도 및 영화평점을 활용하였다.

3.2 잠재 디리클레 할당(LDA, latent dirichlet allocation)

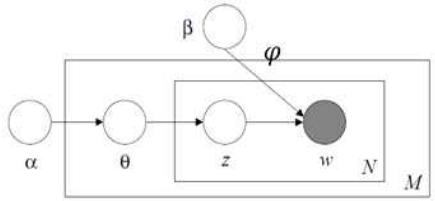
추천시스템에 활용되는 대부분의 알고리즘은 분석에 사용된 데이터 내에서만 중요 단어들을 추출하고 이를 토대로 추천 콘텐츠를 결정하게 되므로, 분석에 사용되지 않은 콘텐츠들은 추천되지 못하는 단점이 있다. 즉 새로 개봉하여 흥행에 성공한 영화는 추천 콘텐츠로 활용되지 못하는 것이다. 하지만 토픽모델링에 활용되는 잠재 디리클레 할당(LDA, Latent Dirichlet Allocation)의 경우 데이터 내에서 광범위하게 발생하는 주제 또는 특정 콘텐츠에서만 사용되는 주제를 식별해내는 알고리즘으로 이러한 추천시스템의 단점을 극복할 수 있다.

토픽 모델은 문서집합에서 단어로부터의 단서들을 이용하여 몇 개의 주제로 구분하는 기법으로 Deerwester et al.(1990)이 제안한 LSI(Latent Semantic Indexing)이 초기 모델이었으나, 이는 확률적 방법론이 아니어서 Hofmann(1999)이 PLSA(Probability Latent Semantic Indexing)을 제안하였다. 이후 Blei et al.(2003)이 PLSA을 확장하여 디리클레(Dirichlet)분포를 기반으로 하는 LDA 모형을 제안하였고

현재 LDA 모형은 토픽 모델의 기본으로써 다양한 분야에서 사용되고 있다.

LDA 모형은 단어들이 각 토픽에 속할 확률을 계산하고, 계산된 단어의 분포를 바탕으로 어떤 토픽에 속할 확률을 분석하여 최종적으로 개별 문서가 어떤 토픽을 다루고 있는지 예측하는 기법이다. Blei et al.(2003)은 LDA 추론 과정을 <그림 1>과 같이 표현하였다. k는 토픽의 개수이고 w는 어떤 문서에서의 단어이며, 각 단어는 관측할 수 없는 토픽 z로부터 생성된다고 가정한다. θ_k는 어떤 문서가 주어졌을 때 토픽이 선택될 확률이고 이는 디리클레분포를 따른다. ϕ_k는 토픽이 주어졌을 때 단어가 선택될 확률이고 이 또한 디리클레분포를 따른다고 가정하고 다음 <표 1>과 같은 생성과정을 거친다.

<그림 1> LDA 모형



<표 1> LDA 생성과정

1. 모든 토픽 각각에 대하여 $\phi_k \; Dir(\beta)$
2. 모든 문서 각각에 대하여 $\theta_k \; Dir(\alpha)$
3. 각 단어에 대하여
   (1) 토픽 $Z_n \sim Multinomail(\theta)$ 선택
   (2) 단어 $W_n \sim Multinomail(\phi_{Z_n})$

이러한 잠재 디리클레 할당은 비지도학습의 일환으로 대용량의 텍스트 문서에 존재하는 주제를 추출하는데 사용되는 알고리즘으로 주어진 데이터 내에서 확률모형을 통해 콘텐츠들을 구성하는 잠재된 주제를 선택하고 이에 해당하는 콘텐츠들을 선택하는 과정을 반복하여 임의의 주제를 찾아낼 수 있다. 또한 확률모형을 통해 콘텐츠들을 선별하므로 콘텐츠 소비 횟수의 거리나 콘텐츠간의 거리를 정의하지 않아도 되는 장점이 있다.

IV. 실증 분석

4.1 잠재 주제의 선정

잠재 디리클레 할당에서 주제개수 K는 연구자가 임의로 결정해야 하므로, K의 개수에 따른 혼잡도(perplexity)를 계산하여 최적의 수를 선정한다. 이는 주제의 단어 분포 정보와 문서의 주제비중 정보를 통해 계산하며 모든 주제에 대해 합산한 값을 활용하게 된다. 본 연구에서는 전략적으로 세분화된 그룹별로 주제개수를 바꿔가며 혼잡도를 계산하였으며, 세분화된 집단별 혼잡도는 <표 1>과 같다. 즉 혼잡도의 특성상 주제의 개수가 늘어날수록 낮아지나 그 감소폭이 크지 않은 것을 확인할 수 있다. 본 연구에서는 세분화된 집단별 고객의 수가 많지 않고 고객특성에 따른 추천 컨텐츠로 장르를 제공하는 것이 목적이므로, 주제개수 K는 5개로 선정하였다.

<표 2> 세분화 집단별 혼잡도(perplexity)

K	집단1	집단2	집단3
3	44.3171	45.5596	42.5836
5	42.8049	44.2996	41.4092
7	41.8591	43.3858	40.6211
10	40.8449	42.2182	39.8614
15	39.7847	40.7464	38.9564
20	39.0687	39.7992	38.4024
30	38.5566	38.8878	37.8640
50	38.4779	38.4979	37.5946

주) 집단1은 서비스이용도가 낮은 고객군, 집단2는 서비스이용도가 중간인 고객군, 집단3은 서비스이용도가 높은 고객군 임.

4.1.1 서비스이용도가 낮은 고객군

주제개수를 5개로 고정하여 1개월 동안 4편 이상 8편 이하의 영화를 시청하는 서비스이용도가 낮은 고객군에 대하여 분석한 결과, 주제 1에 속하는 고객들은 주로 스릴러, 미스터리 및 호러에 속하는 평점이 높은 영화로 2011년 이후 상영된 최신영화를 보는 고객군이라 할 수 있다. 또한 주제 3은 액션, 공상과학, 판타지 장르로 개봉년도가 1977에서 2003년까지 나타나며 평점이 낮게 나타나는 것으로 보아 개봉년도와 평점은 영화선택에 있어 크게 고려하지 않는 고객군이라 할 수 있다. 마지막으로 주제 4는 코메디나 만화, 판타지 등에 해당하는 영화로 평점이 높은 영화를 선호하는 고객군이라 할 수 있다. 해당 고객군에 대해 선정된 주제는 <그림 2>와 같다.

<그림 2> 서비스 이용도가 낮은 집단

주제1	주제2	주제3	주제4	주제5
thriller	drama	action	comedy	generes 1
generes 3	generes 2	sci-fi	romance	year2016
rating4	rating5	adventure	rating4	rating3.5
rating4.5	crime	generes 3	animation	rating3
drama	thriller	generes 4	generes 2	drama
mystery	war	fantasy	fantasy	year2017
horror	year1999	imax	children	year2015
year2014	generes 4	year2003	generes 5	rating2.5
year2013	year1994	rating0.5	rating3.5	rating2
year2011	year2010	year2002	year2009	documentary
year2006	year1997	year1977	generes 4	rating1.5
year2012	year1995	year1980	year2004	rating1
year2000	year1998	rating1	year2001	no genres
western	year1993	year1981	drama	year1970
film-noir	rating4.5	year1989	year2007	year1969

4.1.2 서비스이용도가 중간인 고객군

서비스이용도가 중간인 고객군에서 주제 1에 해당하는 경우 가장 최근에 개봉한 영화이나 영화가 속하는 장르가 많지 않고 평점을 고려하지 않는 고객군이라 할 수 있다. 주제 3은 개봉년도가 1995년에서 1999년에 해당하는 과거 개봉영화로 범죄, 스릴러 및 전쟁 장르에 포함되는 영화로 평점에 높은 영화를 선호하는 고객군이라 할 수 있다. 마지막으로 주제 5의 경우 2011년 이후 개봉한 영화로 평점이 높으며, 영화의 장르가 스릴러, 액션, 공상과학 및 미스터리에 해당하는 영화를 선호하는 고객군이라 할 수 있다. 서비스이용도가 중간인 고객군의 선정된 주제는 <그림 3>과 같다.

<그림 3> 서비스 이용도가 중간인 군

주제1	주제2	주제3	주제4	주제5
generes 1	adventure	drama	comedy	thriller
year2016	action	rating5	drama	action
rating3	fantasy	crime	romance	sci-fi
rating3.5	animation	generes 2	rating4	generes 3
drama	generes 4	generes 3	generes 2	rating4

rating2.5	sci-fi	thriller	generes 3	generes 4
generes 2	children	war	rating4.5	mystery
rating2	imax	year1999	rating3.5	rating4.5
year2015	generes 5	year1994	year2013	crime
year2017	generes 3	mystery	year2007	horror
horror	year2003	generes 4	year2011	year2014
comedy	year2001	rating4.5	year2006	year2012
documentary	year2008	year1998	year2004	year2013
rating1.5	year2002	year1995	year2012	rating3.5
rating1	generes 6	year1997	year2014	year2011

4.1.3 서비스이용도가 높은 고객군

마지막으로 서비스이용도가 높은 고객군에서 주제 1에 해당하는 경우 개봉년도가 1995년에서 1999년까지로 과거에 개봉한 영화중 장르가 드라마, 코메디 및 로맨스에 해당하는 영화를 선호하는 고객군이라 할 수 있으며, 주제 2는 2000년대 초반에 개봉한 영화로 평점이 높으며 스릴러, 범죄 및 미스터리 등에 포함되는 영화를 선호하는 고객군이다. 주제 5의 경우 가장 최근에 개봉한 영화로 평점이 1점에서 3.5점까지 높지 않음에도 불구하고 스릴러, 코메디 및 호러에 해당하는 영화를 선호하는 고객군이라 할 수 있다. 서비스이용도가 높은 고객군의 선정된 주제는 <그림 4>와 같다.

<그림 4> 서비스 이용도가 높은 군

주제1	주제2	주제3	주제4	주제5
drama	rating5	action	comedy	rating3
rating4	drama	sci-fi	adventure	generes 1
generes 2	thriller	adventure	fantasy	drama
comedy	crime	generes 4	animation	rating3.5
romance	generes 3	rating4	children	thriller
generes 3	rating4.5	generes 3	romance	generes 2
crime	generes 2	thriller	generes 4	year2016
rating3.5	mystery	imax	generes 5	comedy
generes 1	action	fantasy	rating4	rating2.5
rating4.5	generes 4	generes 5	generes 3	horror
year1999	sci-fi	year2014	imax	generes 3
war	war	rating3.5	musical	rating2

year1998	year2006	rating4.5	generes 6	rating0.5
year1997	year2004	year2012	year2004	year2015
year1995	year2001	crime	year2001	rating1

V. 결론

본 연구에서는 보다 다양화되는 고객의 특성에 맞는 추천시스템을 위한 알고리즘을 제언하는 목적을 갖고 있다. 이러한 연구목적을 달성하기 위해서 본 연구에서는 무비렌즈의 고객정보를 활용하여 서비스이용도에 따라 세분화된 고객을 보다 세분화할 수 있는 주제를 선정하였다. 실증분석 결과, 세분화된 고객군에 따라 차별화되는 개봉연도 및 장르에 대한 추천항목을 추출할 수 있었으며, 이를 통해 고객에게 제시되는 수많은 콘텐츠들에 대해 선택 대안의 과잉을 해소하여 매출을 높이기 위한 일환으로 추천시스템을 활용할 수 있을 것이다.

결국 본 연구는 콘텐츠 제공 업체가 고객이 선호하는 콘텐츠를 파악한 후 추천하여, 고객이 별도의 콘텐츠 검색을 하지 않고 취향에 맞는 콘텐츠를 사용함으로써 만족도를 높이고 서비스에 대한 충성도를 높이는데 기여할 수 있을 것이다.

참고문헌

Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE transactions on knowledge and data engineering*, 17(6), 734-749.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3, 993-1022.

De Vriendt, J., Degrande, N., & Verhoeyen, M. (2011). Video content recommendation: An overview and discussion on technologies and business models. *Bell Labs Technical Journal*, 16(2), 235-250.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. J*ournal of the American society for information science*, 41(6).391-407.

Gomez-Uribe, C. A., & Hunt, N. (2016). The netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems (TMIS)*, 6(4), 13-19.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. ACM. 50-57.