The Korean Journal of Applied Statistics (2020)
33(6), 833–842
DOI: https://doi.org/10.5351/KJAS.2020.36.6.833

Hierarchical grouping recommendation system based on the attributes of contents: a case study of 'The Movie Dataset'

Yoon Kyoung Kim[a] · In-Kwon Yeo[a,1]

[a]Department of Statistics, Sookmyung Women's University

(Received November 3, 2020; Revised November 17, 2020; Accepted November 18, 2020)

Abstract

Global platforms such as Netflix, Amazon, and YouTube have developed a precise recommendation system based on various information from large set of customers and many of the items recommended here are leading to actual purchases. In this paper, a cluster analysis was conducted according to the attribute of the content, expecting that there would be a difference in user preferences according to the attribute of the recommended content. Gower distance was used for use regardless of the type of variables. In this paper, using the data of movie rating site 'The Movie Dataset', the users were grouped hierarchically and recommended movies based on genre, director and actor variables. To evaluate the recommended systems proposed, user group was divided into train set and test set to examine the precision. The results showed that proposed algorithms have far higher precision than UBCF.

Keywords: clustering, Gower's distance, precision

1. 서론

인터넷이나 온라인상에서 물건을 구매하거나 영화를 예매할 때, 원하는 것을 직접 검색하며 찾기도 하지만 해당 사이트에서 추천하는 아이템을 살펴보거나 구매하기도 한다. 구매자가 선호할 만한 것들을 모아서 보여준다면 구매자는 원하는 물건을 찾기 위한 시간을 절약할 수 있고 추가 구매를 진행할 수도 있다. 다양한 분야에서 사용자에게 맞는 최적의 구매정보를 제공하는 추천 시스템이 사용되고 있다. 대표적으로 아마존, 유튜브, 넷플릭스 등과 같은 대형 플랫폼에서 방대한 데이터를 바탕으로 사용자들의 선호에 맞는 제품을 제공하는 추천 시스템을 이용하고 있다.

기존에 알려진 추천시스템 알고리즘을 살펴보면 다음과 같이 분류할 수 있다.

- Content-based recommendation: 사용자의 과거 구매이력 등을 바탕으로 선호하는 아이템을 파악한 후, 미리 분류된 아이템 범주와의 유사도를 계산해 추천하는 방식
- Collaborative filtering (CF): 사용자가 과거에 구매했던 아이템 이력을 바탕으로 유사한 사용패턴을 가지는 사용자들을 찾아내 추천하는 방식
- Hybrid approaches: content-based recommendation과 collaborative filtering을 융합하여 추천하는 방식

[1] Corresponding author: Department of Statistics, Sookmyung Women's University, 86-1 cheongpa-ro 47-gil 100, Yongsan-gu, Seoul 04310, Korea. E-mail: inkwon@sookmyung.ac.kr

우리나라에서도 추천시스템에 대한 연구가 컴퓨터, 정보, 경영 분야에서 오래전부터 이루어지고 있으며 Lee와 Lee (2001), Kim 등 (2002)이 그 예이다. Choi 등 (2012)은 영화 추천 시스템 연구에서 사용자들의 정보 부족 문제를 해결하고자 장르의 상관관계를 이용한 알고리즘을 제안했다. K-평균 알고리즘(k-means algorithm)과 같은 군집분석을 기반으로 한 추천 시스템에 대해서도 다양한 연구가 진행되었다. 군집분석을 통해 군집화된 사용자를 이용한 추천시스템이 기존 CF와 비슷한 성능을 가지며 온라인 성능을 향상시키는 것으로 나타났다. Wang 등 (2014)은 PCA를 이용하여 특징정보(feature information)를 낮은 차원의 밀집된 공간으로 집중시키고, 이렇게 변형된 사용자 공간을 분리하기 위해 K-평균 군집화(k-means clustering)와 유전알고리즘(genetic algorithms)을 결합한 혼합모형기반(hybrid model-based) 영화추천시스템을 제안하였다. 이들 연구에서 군집화을 기반으로 한 CF보다 높은 예측도를 보여주며, 데이터 부족으로 발생하는 완전시작(cold-start) 문제에 대해서도 효과적인 것을 확인하였다. 사용자 기반(user-based)의 추천 시스템의 확장성 문제를 해결하기 위해 Sarwar 등 (2002)은 아이템 간 유사도를 바탕으로 평점을 예측하는 항목기반(item-based) CF를 소개하였다. 항목 간에 유사도를 구하는 다양한 방법과 추천에 필요한 예측 평점을 구하는 다양한 방법을 소개하고 추천시스템을 비교하였다. R이나 Python에서는 위의 알고리즘을 바탕으로 한 여러 추천 시스템 패키지가 존재한다. 이 방법 이외에도 국내외적으로 딥러닝 기법을 추천시스템에 적용하는 방법도 많이 연구되고 있는데 Lee 등 (2019), Moon 등 (2020), Zhang 등 (2019)을 참고하기 바란다.

추천시스템을 구축하는데 있어 가장 중요한 것은 해당 시스템에서 학습시킬 자료가 얼마나 풍부하게 있는가이며 이들 자료를 수집하기 위해서는 상당한 재정적 뒷받침이 있어야 한다. 또한 개인정보 활용에 있어 법적 제약이 큰 상황에서 고객의 인구생태학적 정보를 활용한 추천시스템은 제한적으로만 사용될 수 밖에 없다. 이 논문에서는 추천 대상인 콘텐츠를 속성에 따라 군집화하고 군집화된 콘텐츠의 선호도에 차이가 있는 그룹과 아닌 그룹으로 나누어 추천을 달리하는 방법을 제안한다. 콘텐츠의 여러 속성을 동시에 고려하는 경우 그에 따른 경우의 수가 많아져 분석에 한계가 있을 수 있다. 그리고 각 속성에 따른 사용자들의 특징을 파악하기 어렵다. 따라서 속성들을 단계적으로 고려해 사용자들의 특징을 최대한 파악하고자 한다. 사용자들 간에 동질적인 패턴을 파악할 수 있다면 비슷한 취향의 사용자들을 대상으로 좀 더 정확한 추천을 할 수 있을 것이라 기대한다. 더 나아가 그룹화를 실시할 때 사용자들의 개인 정보를 이용하는 것이 아닌 추천 콘텐츠의 기본적인 정보를 활용함으로써 개인적인 정보가 없는 데이터의 경우에도 적용해 볼 수 있도록 한다.

2. 이론적 배경

이 논문에서는 유사한 취향을 가진 사용자들은 선호하는 콘텐츠도 비슷하다는 전제 하에 사용자들을 그룹화하고 그룹별로 선호할 가능성이 높은 콘텐츠를 추천한다. 사용자를 그룹화할 때 활용할 수 있는 정보는 어떤 콘텐츠인가에 따라 다르다. 사용자의 개인 정보 등을 사용할 수 있으나 상황에 따라 해당 정보가 수집되지 않거나 사용자별로 정보의 양에 차이가 커 활용에 있어 한계가 있을 수 있다. 예를 들어, 이 논문에서 분석할 Kaggle의 ‘The Movie Dataset(https://www.kaggle.com/rounakbanik/the-movies-dataset)’ (TMDS)에서는 관객에 대한 개인 정보가 별도로 주어지지 않고 있다.

이 논문에서는 사용자의 정보 대신 사용자의 구매했던 콘텐츠의 속성을 활용하여 사용자들을 먼저 그룹화한다. 여기서 속성이란 콘텐츠가 만들어질 때 필수적으로 부여되는 정보를 의미하는데 예를 들어 영화의 경우 장르, 감독, 배우, 음악의 경우 장르, 아티스트, 음원 길이, 식품의 경우 원산지, 등급, 가격 등이 될 수 있다. 일반적으로 콘텐츠의 속성이 수치자료(공변량)이면 해당 속성을 하나의 변수로 표현할 수 있으

나 범주형 자료(요인)이면 가변수로 전환하여 변수 개수는 수준의 수에 비례하게 많아진다. 이 때 요인의 수준은 서로 배타적으로 이루어져야 한다. 하지만 영화나 음악의 장르의 경우 한 콘텐츠에 여러 가지 장르가 복합적으로 포함되어 있는 경우가 많다. TMDS의 경우에도 표시 가능한 장르는 총 20가지이며 영화별로 해당 장르의 유무를 표시하는 20개의 장르 변수를 사용해야 한다. 이들 변수를 통한 가능한 장르 조합은 $2^{20}-1$로 백만 가지 이상 된다.

속성에 따라 분류된 그룹의 수가 많아지면 그에 따른 사용자의 그룹도 많아지고 세분화된 추천시스템을 만들 수 있으나 그만큼 관리 및 운영이 어려워 실제 추천시스템을 구축할 때에는 적정 수의 그룹으로 나눌 필요가 있다. 속성 자료의 그룹화 과정에는 이전에 정해 놓은 별도의 라벨이 없기 때문에 군집분석과 같은 비지도학습 방법을 적용해야 한다. 만약 속성이 수치자료로 이루어져 있다면 거리 기반 계층적 군집분석이나 k-평균 군집분석과 같은 방법을 적용할 수 있으나 질적자료나 이진자료의 경우 별도의 방법이 필요하다. 이 논문에서는 이진자료나 질적자료에 대해서도 군집분석을 하기 위해 Gower (1971)가 제안한 방법으로 콘텐츠를 그룹화하는 방법을 고려하였다.

데이터의 속성은 수치자료와 범주형 자료가 혼재되어 있는 경우가 대부분이기 때문에 수치자료에서 사용되는 유클리드 거리로 데이터 속성을 군집화 할 수 없다. 수치자료와 범주형 자료가 혼합되어 있는 경우 사용할 수 있는 방법 중 Gower 거리가 있다. Gower (1971)는 이분적(dichotomous) 변수, 질적(qualitative) 변수, 양적(quantitative) 변수에 대해서도 유사성 또는 비유사성의 정도를 계산하는 방법을 제안하였다. 각 관측값은 p개의 수치자료와 q개의 범주형 자료로 구성되어 있고 i번째 관측값을 다음과 같이 표시한다고 하자.

$$x_i = \left(z_{i1}, \dots, z_{ip}, c_{i,1}, \dots, c_{iq}\right),$$

여기서 $z_i = (z_{i1}, \dots, z_{ip})$는 수치자료, $c_i = (c_{i,1}, \dots, c_{iq})$는 범주형 자료를 의미한다. Gower 비유사성 계수(Gower's dissimilarity coefficent)는 각 변수의 측도를 0에서 1사이로 정규화한 후 거리의 가중 평균으로 계산되는데, Everitt 등 (2011)과 Tuerhonf과 Kim (2014)에서는 두 관측값 x_i와 x_j의 Gower 비유사성 계수를 다음과 같이 계산한다.

$$D_{x_i,\ x_j} = \frac{\sum_{k=1}^{p} W_{z_i,z_j,k}\ D_{z_i,z_j,k}}{\sum_{k=1}^{p} W_{z_i,z_j,k}} + \frac{\sum_{k=1}^{p} W_{c_i,c_j,k}\ D_{c_i,c_j,k}}{\sum_{k=1}^{q} W_{c_i,c_j,k}}$$

여기서 $W_{z_i,z_j,k}$와 $W_{c_i,c_j,k}$는 수치자료와 범주형 자료에 대한 가중치를 의미하며, $z_k^* = (z_{1k}, \dots, z_{nk})$를 k번째 수치변수의 전체 자료라고 했을 때,

$$D_{z_i,z_j,k} = \frac{\left|z_{ik} - z_{jk}\right|}{\max(z_k^*) - \min(z_k^*)},$$

$$D_{c_i,c_j,k} = \begin{cases} 0, & c_{ik} = c_{jk} \\ 1, & c_{ik} \neq c_{jk} \end{cases},$$

이다. Gower 거리에서의 가중치 W를 계산하는 방법은 여러 가지가 있는 자세한 내용은 Bektas와 Schumann (2019), van den Hoven (2016)을 참조하기 바란다. Gower 거리를 이용한 비유사성 행렬은 R 패키지 cluster의 daisy 함수나 StatMatch의 gowwe.dist함수를 이용하여 계산할 수 있다.

3. 사례분석

본 논문에서는 데이터 분석 및 머신러닝 학습 플랫폼인 Kaggle에서 제공하는 TMDS의 영화 평점 데이터

를 사용해 고객들이 선호할만한 영화를 추천하고자 한다. 각 영화에 대해 20개 장르의 포함여부를 확인하고 군집분석을 통해 영화를 적정수의 군집으로 나눈다. 관객도 장르에 영향을 받는 관객과 받지 않는 관객으로 나누고 장르에 영향을 받지 않는 경우 감독이나 배우에 영향을 받는지를 확인한다. 장르, 감독, 배우에 영향을 받지 않는 관객에 대해서는 특별히 추천하지 않는다. 제안 추천방법의 적절성을 알아보기 위해 데이터를 훈련자료(train data)와 검증자료(test data)로 나누어 훈련자료에 의해 추천된 영화를 검증자료의 관객들이 얼마나 높은 평점을 주었는지로 평가해 보았다. 이에 대한 자세한 분석 과정은 다음과 같다.

3.1. 데이터 전처리

TMDS에는 영화의 특징을 나타내는 메타데이터(movies_metadata.csv), 관객별로 관람한 영화의 평점(ratings.csv를 포함 4개의 파일) 자료, 각 영화별로 출연진과 제작진의 정보(credits.csv) 자료로 구성되어 있고 각 관객와 영화에 대해 고유 아이디가 부여되어 있다. 아래의 표는 id가 862번인 Toy Story의 메터데이터 형태와 credits 자료에 포함된 출연진과 제작진의 정보를 정리한 것이다. 영화 메타데이터에는 장르, 줄거리, 제작비용 및 수익, 개봉년도 등의 정보와 더불어 평균평점과 평가한 관객 수가 기록되어 있다. 이 논문에서는 영화를 선택하는데 있어서 가장 중요하다고 생각되는 장르, 감독, 배우 변수만을 사용했으며 분석의 편의를 위해 대표적인 감독과 배우에 대해서만 분석하였다.

변수	값
id	862
original_title	Toy Story
genres	[{'id': 16, 'name': 'Animation'}, {'id': 35, 'name': 'Comedy'}, {'id': 10751, 'name': 'Family'}]
vote_average	7.7
vote_count	5415

변수	값
id	862
cast	[{'cast_id': 14, 'character': 'Woody (voice)', 'credit_id': '52fe4284c3a36847f8024f95', 'gender': 2, 'id': 31, 'name': 'Tom Hanks', 'order': 0, 'profile_path': '/pQFoyx7rp09CJTAb932F2g8Nlho.jpg' }, ...]
crew	[{'credit_id': '52fe4284c3a36847f8024f49', 'department': 'Directing', 'gender': 2, 'id': 7879, 'job': 'Director', 'name': 'John Lasseter', 'profile_path': '/7EdqiNbr4FRjIhKHyPPdFfEEEFG.jpg' }, ...]

원자료에서 평점을 매긴 관객은 총 270,896명이며, 각 관객들이 1개 이상의 영화에 대해 0.5점에서 5점까지 0.5점 단위로 평점을 부여했다. 장르 정보가 있는 영화 42,980개였으며 해당 영화에 대해 평점이 없는 경우에는 분석자료에서 제외하였으며 최종 분석 데이터는 7,191개의 영화와 265,848명의 평점 정보로 구성되었다.

원자료에는 복수응답처럼 여러 개의 장르가 한 영화에 텍스트 형태로 보관되어 있어 이를 분석가능한 형태로 변환하는 과정이 필요했다. 표시 가능한 장르는 총 20개(로맨스, 액션, 호러 등)으로 각 영화에 대해 해당 장르에 포함되면 1, 아니면 0의 값을 가지는 20개의 이진변수로 만들어 Gower 거리를 계산하여 분석했다.

3.2. 장르 군집

군집분석을 위한 비유사도를 구하기 위해 각 영화별로 고유 아이디와 20개 장르 변수로 이루어진 행렬을 생

성한다. 이 행렬을 R 패키지 'cluster'의 'daisy( )'함수에 적용하여 Gower 거리를 구한다. 함수 옵션으로 'method=gower'를 지정하면 혼합 변수에 대한 Gower 거리 행렬을 구할 수 있으며 장르 변수를 0과 1로 표시한 이분적(asymmetric binary) 형태와 수준이 두 개인 질적 변수(symmetric binary) 형태로 지정할 수 있다.

장르 변수를 바탕으로 구한 영화들의 비유사도를 partitioning around medoids (PAM)에 적용하여 군집 분석을 실시한다. PAM은 K-평균 군집분석과 마찬가지로 비계층적 군집분석이며 지정된 수의 군집으로 비용함수가 최소화 될 때까지 재할당하는 작업을 반복한다. 여기서 비용함수는 각 군집에 속해 있는 대표 객체와 나머지 관측값의 거리로 이루어져 있는다. K-평균 알고리즘에서는 대표 객체가 데이터들의 무게 중심인 평균이지만 PAM에서는 데이터 특성을 잘 반영할만한 관측 데이터를 대표 객체로 사용한다. PAM은 k-평균 알고리즘보다 이상점에 로버스트하다는 장점을 가진다. K-평균 알고리즘이나 PAM을 적용하기 위해서는 적절한 군집수 k를 결정해야 한다. Charrad 등 (2014)은 R 패키지 'NbClust'에서 제공하는 군집 적절성을 평가하는 다양한 지수 방법을 정리했으며 이 지수를 활용하여 적정한 k를 선정할 수 있다. 이 논문에서는 군집수(k)를 2~50으로 바꾸어가며 분석하였으며 $k = 17$를 설명력이 높은 군집 수로 설정했다.

3.3. 관객 분류

TMDS의 자료에서는 관객에 대한 특별한 정보가 없어 관객이 봤던 영화의 정보를 활용하여 관객을 분류할 수 밖에 없다. 관객을 분류하는데 있어 먼저 장르에 영향을 받는지 여부에 따라 1차 분류하고 영향을 받지 않는 경우 감독 및 배우에 영향을 받는지를 확인하였다. 장르는 영화의 특성을 객관적으로 반영하는 정보라 판단되어 장르를 우선 분류 기준으로 사용하였다.

장르에 영향을 받는 관객인지 아닌지를 평가하기 위해 다음과 같은 방법을 적용했다. 어떤 한 관객의 평점을 장르 군집별로 나누어 군집별로 평균평점에 차이가 있는지를 알아보기 위해 분산분석을 실시한다. 분산분석결과 유의수준 5%에서 유의한 차이가 있으면 이 관객은 장르에 영향을 받는 관객으로 분류한다. 장르 군집이 총 17개이기 때문에 수준의 수가 최대 17인 일원배치 분산분석이지만 대부분의 고객은 5개 이하의 수준에서 분석이 이루어졌다. 장르에 영향을 받는 경우에는 어떤 장르 군집에서 평점이 높은지를 보고 해당 장르 군집으로 관객으로 분류하였다.

장르에 영향을 받지 않는 경우 특정 감독이나 배우의 영화와 아닌 영화의 평균 평점에 차이가 있는지를 알아보기 위해 이표본 t-검정을 실시한다. 검정결과 유의한 차이가 있는 경우 그 관객을 해당 감독 또는 배우에 영향을 받는다고 분류한다. 이 논문에서는 대표적인 4명의 감독(Steven Spielberg, Tim Burton, James Cameron, Christopher Nolan)과 2명의 배우(Leonardo DiCaprio, Natalie Portman)를 선정하여 분석했다.

3.4. 추천 방법

일반적으로 추천시스템에서는 추천 대상인 특정 사용자가 구매하지 않은 아이템에 대해 평점을 예측하고, 그 결과를 바탕으로 가장 높은 예측 평점을 가진 아이템을 추천한다. 군집 분석을 바탕으로 한 기존 추천시스템 연구에서는 군집 분석으로 유사 고객를 정의하고, 다음과 같은 식을 통해 특정 고객의 각 아이템에 대한 예측 평점을 구한다.

$$r_{c,i} = \frac{1}{n} \sum_{c^* \in \hat{C}} r_{c^*,i} ,$$

$$r_{c,i} = \frac{\sum_{c^* \in \hat{C}} s(c, c^*) r_{c^*,i}}{\sum_{c^* \in \hat{C}} |s(c, c^*)|},$$

$$r_{c,i} = \bar{r}_c + \frac{\sum_{c^* \in \hat{C}} s(c, c^*)\left(r_{c^*,i} - \bar{r}_{c^*}\right)}{\sum_{c^* \in \hat{C}} |s(c, c^*)|},$$

여기서 $r_{c,i}$는 고객 c의 아이템 i에 대한 예측 평점이다. $\hat{C}$는 전체 고객 n명 중 고객 c와 유사하고 아이템 i의 평가정보가 존재하는 고객들의 집합이다. $\bar{r}_c$는 모든 아이템에 대한 $r_{c,i}$의 평균을 나타내고 고객 c와 c^* 사이의 유사도인 $s(c, c^*)$는 피어슨의 상관계수, 코사인 유사도 등으로 구할 수 있다. 특정 아이템과 유사한 아이템에 부여한 고객의 평점을 가지고 예측 평점을 구할 수도 있다.

$$r_{c,i} = \frac{\sum_{i^* \in \hat{I}} s(i, i^*) r_{c,i^*}}{\sum_{i^* \in \hat{I}} |s(i, i^*)|},$$

여기서 $\hat{I}$는 아이템 i와 유사한 모든 아이템을 의미하며 r_{c,i^*}는 고객 c의 아이템 i^*의 평점을 의미한다. 여기서 r_{c,i^*}을 원평점 대신 회귀식을 통해 구한 평점을 사용할 수도 있다. 군집분석을 기반으로 한 추천시스템에서는 각 고객과 아이템에 대해 위의 예측평점이 계산하고 높은 예측평점의 아이템을 추천한다.

본 논문에서는 장르와 감독 및 배우 순으로 분류된 그룹의 관객이 관람한 영화 목록을 활용하여 추천 목록을 생성한다. 영화의 장르를 바탕으로 군집분석하여 17개의 영화 군집을 생성하였으며 각 군집에 속해 있는 영화들은 유사 장르의 영화라고 판단한다. 이를 바탕으로 장르에 영향을 받는 관객의 경우 각 영화 군집의 평균 평점과 관람 횟수를 이용하여 추천하였다. 예를 들어, 군집1에 높은 평점을 부여한 관객(이하 군집1 관객)들을 대상으로 다음과 같은 방법으로 추천 목록을 각각 생성하였다.

[추천1] 군집1 관객이 관람한 군집1 영화 중 평균 평점이 4점 이상인 영화

[추천2] 군집1 관객이 최다 관람한 200개 영화와 [추천1] 조건을 동시에 만족하는 영화

[추천2]는 관객수를 고려하여 해당 영화의 화제성을 반영한 것이다. 다른 장르의 영화도 추천할 수 있도록 전체 영화에 대해 추천한 영화와 유사도가 높은 영화를 선정하는 방법도 고려했으나 분석결과가 좋지 않아 결과비교 및 해석에서 제외하였다.

감독 및 배우에 영향을 받는 관객의 경우에는 영화의 평점 평균과 유사도를 이용하여 추천 목록을 생성하였다.

[추천3] 해당 감독이 감독하거나 배우가 출연한 영화가 속한 군집에서 평균 평점이 4점 이상인 영화

[추천4] 해당 감독이나 배우가 참여한 영화와 유사도가 높은 영화와 [추천3] 조건을 동시에 만족하는 영화

위와 같이 추천된 영화는 기존 예측 평점의 순위로 추천되는 영화 목록과는 달리 사용자들에게 어떠한 이유로 추천되었는지 명확한 설명이 가능해진다.

3.5. 평가

신규 사용자가 어떤 영화 속성에 영향을 받는지 알고 있다는 전제로 추천시스템을 평가하기 위해 그룹별로 사용자를 일종의 훈련자료(train data) 70%, 검증자료(test data) 30%로 나누었다. 각 사용자 그룹의 훈련자료에 대해 제안 방법을 적용하여 추천 목록을 생성한다. 평가의 안정성을 위해 각 그룹별로 훈련자료와 검증자료를 10번 반복 생성하여 분석하였다.

Table 3.1. Precisions of proposed recommendation algorithms

그룹	[추천1]			[추천2]		
	평균 T	평균 R	정밀도	평균 T	평균R	정밀도
군집1	76.9	50.9	0.662	4.3	4.3	1.0
군집2	99.1	70.1	0.708	5.9	5.9	1.0
군집3	87.3	63.4	0.726	10.0	10.0	1.0
군집4	172.2	117.0	0.679	13.6	13.6	1.0
군집5	150.5	108.1	0.719	16.0	16.0	1.0
군집6	119.2	84.0	0.705	10.8	10.8	1.0
군집7	144.0	102.3	0.712	13.0	13.0	1.0
군집8	297.5	187.7	0.631	30.5	30.5	1.0
군집9	74.3	48.6	0.655	11.4	11.4	1.0
군집10	187.9	137.0	0.729	12.0	12.0	1.0
군집11	74.3	48.2	0.651	7.3	7.3	1.0
군집12	136.3	98.1	0.720	13.2	13.2	1.0
군집13	139.5	95.2	0.683	12.7	12.7	1.0
군집14	99.5	72.4	0.729	7.0	7.0	1.0
군집15	105.5	63.1	0.599	6.0	6.0	1.0
군집16	82.5	57.9	0.702	6.0	6.0	1.0
군집17	44.0	29.4	0.670	6.6	6.6	1.0

추천시스템을 평가하는데 있어 고객이 추천한 아이템을 얼마나 구매했는지로 측정할 수도 있으나 장기적으로 고객의 추천시스템 활용도를 높이기 위해서는 추천받은 아이템에 얼마나 만족하는지를 평가하는 것이 더 중요할 수 있다. 이런 관점에서 이 논문에서는 옳게 추천한 영화란 검증자료에서 해당 영화를 본 관객의 평균평점이 4 이상인 영화로 정의하였다.

추천시스템에 대한 평가측도는 어떤 목적을 가지고 평가하느냐에 따라서도 달라질 수 있다. 본 논문에서는 추천한 영화에 대해 얼마나 만족하는지를 평가하기 위해 다음과 같은 정밀도(precision)로 평가하였다.

$$\text{정밀도} = \frac{\text{옳게 추천한 아이템 수}}{\text{추천한 전체 아이템 수}} = \frac{R}{T}.$$

Table 3.1에는 [추천1]과 [추천2]방식으로 훈련자료에 의해 생성된 추천 목록(추천한 전체 이이템)과 이들 목록에 대해 검증자료의 관객들이 관람하고 평균 평점이 4 이상으로 표시한 영화 목록(옳게 추천한 아이템)의 수로 계산된 평균 정밀도가 주어졌다. 여기서 평균 T는 10번의 반복과정에서 추천된 평균 영화의 수를, 평균 R은 옳게 추천된 평균 영화 수를 의미한다.

[추천1]의 경우 17개의 군집에서 정밀도가 대략 0.6 ~ 0.8 사이에 있고 [추천2]는 전체 군집에서 정밀도가 1인 것으로 나타났다. [추천2]는 추천 수가 상대적으로 적은 반면 많은 관객이 본 영화로 평균 평점이 4점 이상이라는 것은 상당히 많은 관객으로부터 높은 평점을 받았기에 검증 자료에서도 높은 평점을 받을 수 밖에 없다고 예상할 수 있다. 이에 비해 [추천1]은 관객의 수를 반영하지 않았기에 평균 평점의 변동성을 알 수 없고 이에 따라 [추천2]보다는 낮은 정밀도를 보이고 있으나 아래에서 비교할 'user-based collaborative filtering (UBCF)'에 비해 정밀도가 매우 높은 것으로 분석되었다. 이는 영화 그룹에 따라 관객이 잘 분류되어 각 관객 그룹별로 공통적인 장르 취향을 뚜렷하게 나타내고 있다고 생각할 수 있다.

Table 3.2는 예시로 일부 감독 및 배우를 선정하고, 영화 장르에 영향을 받지 않는 사용자들 중 특정 감독

Table 3.2. Precisions of proposed recommendation algorithms based on director and actor

감독 및 배우	추천3			추천4		
	평균 T	평균 R	정밀도	평균 T	평균 R	정밀도
Steven Spielberg	421.9	211.6	0.501	37.5	21.3	0.565
Tim Burton	308.7	174.1	0.564	4.5	4.3	0.963
James Cameron	55.7	26.8	0.481	-	-	-
Christopher Nolan	116.9	78.7	0.680	2.0	2.0	1.0
Leonardo DiCaprio	86.5	46.9	0.543	13.2	8.3	0.626
Natalie Portman	293.6	143.3	0.485	48.6	29.2	0.600

Table 3.3. Comparison of the proposed algorithm and UBCF

추천방법	[추천1]	[추천3]	UBCF_c	UBCF_p
평균 정밀도	0.687	0.542	0.083	0.093

및 배우에 영향을 받는 사용자들을 분류하여 추천한 결과이다. 결과를 보면 [추천3]은 [추천1]과 비교해 조금 낮은 정밀도를 가지는 것으로 나타났으며 [추천4]는 [추천3]보다 적은 추천 수를 가지는 반면 상대적으로 높은 정밀도를 가지며 감독이 누구인가에 따라 정밀도에 차이가 있는 것으로 나타났다. Cameron 감독 그룹에서와 같이 [추천4]에 해당하는 영화가 없는 상황이 발생할 수 있다.

제안 방법과 기존 방법을 비교해 보기 위해 사용자 간 유사도를 바탕으로 평점을 예측하는 UBCF을 적용하여 정밀도를 계산해 보았다. UBCF에서는 코사인 유사도(UBCF_c)와 피어슨 상관계수(UBCF_p)를 사용했으며 동일한 비교를 위해 무작위로 관객을 훈련자료와 검증자료로 나누어 10회 반복하여 평균 정밀도를 계산하였다. Table 3.3에서의 [추천1]과 [추천3]의 정밀도는 Table 3.1와 Table 3.2에서의 정밀도 평균으로 UBCF의 평균 정밀도와는 상당한 차이가 있는 것을 확인할 수 있다.

4. 결론

본 연구에서는 콘텐츠에 대한 명확한 속성이 있는 경우 그 속성에 따라 콘텐츠 사용자들을 계층적으로 분류하고 분류된 그룹에 적절한 아이템을 추천하는 방법을 제안하였다. 이 방법을 TMDS 영화 평점 자료에 적용하여 추천시스템을 구축해 보고 평가하였다. 영화의 특징을 이용하여 사용자들을 분류하고, 분류된 사용자들의 정보를 바탕으로 추천을 수행하였다.

기존 추천시스템들은 각 영화에 대한 예측 평점을 구하고, 높은 평점 순으로 추천 영화 리스트를 생성하는 방식이다. 하지만 본 논문에서는 각 영화에 대한 예측 평점을 구하지 않고, 영화 군집별로 분류된 관객이 관람한 영화 목록을 이용하여 추천하였다. 신규 관객에게 같은 그룹의 관객이 관람한 영화를 바탕으로 영화를 추천하는 방식이다.

영화를 객관적으로 나타낼 수 있는 특징으로 장르, 감독, 배우 등이 있다. 이 정보들은 관객들에 의해서 만들어진 정보가 아닌, 영화가 제작되면 필수적으로 부여되는 정보이다. 따라서 이러한 정보를 사용한다면 관객에 대한 정보가 적은 경우에도 영화추천이 가능하다. 본 논문에서는 장르, 감독, 배우 정보를 가지고 관객을 계층적으로 분류하였다. 계층적으로 분류되기 때문에 관객들이 어떤 기준으로 영화를 선택하는지 구체적인 설명이 가능해진다. 더 나아가 추천 영화 리스트가 어떤 이유로 관객에게 추천되었는지 타당한 근거를 제시할 수 있으므로 관객의 신뢰도 역시 높아질 수 있을 것이다.

추후 연구로 분석에 이용되는 속성의 순서를 고려할 필요가 있다. 현재 장르 기반으로 분류한 그룹의 정밀도가 높은 것으로 나타났는데 이는 장르 기반으로 관객을 먼저 분류했기 때문일 수 있다. 그러므로 여러 순서에 대해 분류해보고 최적의 결과를 내는 경우를 이용한다면 더 정확한 추천시스템이 될 것이라 기대한다. 데이터 측면에서 사용자들의 기본적인 개인 정보를 활용하여 좀 더 정교한 추천시스템을 제시해볼 수 있다. 성별과 나이 등의 정보가 추가로 주어진다면 영화 자체만의 정보와 더불어 특정 성별이나 연령대의 사람들이 어떤 영화 패턴을 시청하는지까지 파악할 수 있을 것이다.

References

Bektas, A. and Schumann, R. (2019). How to optimize Gower distance weights for the k-medoids clustering algorithm to obtain mobility profiles of the Swiss population, *IEEE 6th Swiss Conference on Data Science (SDS)*.

Charrad, M., Ghazzali, N., Boiteau, V., and Niknafs, A. (2014). NbClust: An R package for determining the relevant number of clusters in a data set, *Journal of Statistical Software*, **61**, 1–36.

Choi, S. M., Ko, S. K., and Han, Y. S. (2012). A movie recommendation algorithm based on genre correlations, *Expert Systems with Applications*, **39**, 8079–8085.

Everitt, B. S., Landau, S., Leese, M., and Stahl, D. (2011). *Cluster analysis* (5th ed). Wiley.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties, *Biometrics*, **27**, 857–874.

Kim, Y. S., Kim, B. C., and Yoon, B. J. (2002). Design and implementation of e-commerce applications using improved recommender systems, *KIPS Transactions on Computer and Communication Systems*, 9-D, 329–336.

Lee, H. G. and Lee, S. Y.(2001). The design and implementation of an adaptive Information recommendation agent system, *Journal of Information Technology Applications and Management*, **3**, 77–89.

Lee, R. K., Chung, N., and Hong, T. (2019). Developing the online reviews based recommender models for multi-attributes using deep learning, *The Journal of Information Systems*, **28**, 97–114.

Moon, H., Lim, J., Kim, D., and Cho, Y. (2020). A Deep learning based recommender system using visual information, *Knowledge Management Research*, **21**, 27–44.

Sarwar, B. M., Karypis, G., Konstan, J., and Riedl, J. (2002). Recommender systems for large-scale e-commerce: Scalable neighborhood formation using clustering. In *Proceedings of the Fifth International Conference on Computer and Information Technology*, **1**, 291–324.

Tuerhong, G. and Kim, S. B. (2014). Gower distance-based multivariate control charts for a mixture of continuous and categorical variables, *Expert Systems with Applications*, **41**, 1701–1707.

van den Hoven, J. (2016). Clustering with Optimized Weights for Gower's Metric, University Amsterdam. Available from: http://www.few.vu.nl/ sbhulai/papers/thesis-vandenhoven.pdf

Wang, Z., Yu, X., Feng, N., and Wang, Z. (2014). An improved collaborative movie recommendation system using computational intelligence, *Journal of Visual Languages & Computing*, **25**, 667–675.

Zhang, S., Yao, L., Sun, A., and Tay, Y. (2019). Deep learning based recommender system: A survey and new perspectives, *ACM Computing Surveys*, **52**, Article No 5, 1–35.

콘텐츠 속성에 따른 계층적 그룹화 추천시스템: 'The Movie Dataset' 분석사례연구

김윤경[a] · 여인권[a,1]

[a]숙명여자대학교 통계학과

(2020년 11월 3일 접수, 2020년 11월 17일 수정, 2020년 11월 18일 채택)

요 약

넷플릭스, 아마존, 유튜브 등 대형 플랫폼에서는 고객의 다양한 정보를 활용하여 정밀한 추천시스템을 마련하고 여기서 추천된 상당수의 아이템이 실제 구매로 이어지고 있다. 본 논문에서는 추천 컨텐츠의 속성에 따라 사용자의 선호도에 차이가 있을 것이라고 예상하고 콘텐츠의 속성에 따라 군집분석을 실시하였다. 속성의 형태와 관계없이 사용할 수 있도록 Gower 거리를 사용했다. 본 논문에서는 영화 평점 사이트인 'The Movie Dataset'의 자료를 이용하여 영화의 기본정보인 장르, 감독 및 배우 변수를 바탕으로 사용자를 계층적으로 분류하고 영화를 추천하였다. 본 논문에서 제안한 추천 시스템을 평가하기 위하여 각 사용자 그룹별로 훈련자료와 검증자료로 나누어 정밀도를 살펴보았다. 그 결과 UBCF보다 월등히 높은 정밀도를 갖는 것으로 나타났다.

주요용어: Gower 거리, 군집분석, 장밀도

[1] 교신저자: (04310) 서울시 용산구 청파로47길 100, 숙명여대 통계학과. E-mail: inkwon@sookmyung.ac.kr