추천 시스템에서의 컨텐츠 기반 필터링 기법의 효과

Effectiveness of Content-based Filtering Technique on Recommendation System

저자
(Authors)

홍동균, 이연창, 김상욱, 이종욱
Dong-Gyun Hong, Yeon-Chang Lee, Sang-wook Kim, Jongwuk Lee

출처
(Source)

한국정보과학회 학술발표논문집 , 2017.06, 266-267 (2 pages)

발행처
(Publisher)

한국정보과학회
KOREA INFORMATION SCIENCE SOCIETY

URL

http://www.dbpia.co.kr/Article/NODE07207203

APA Style

홍동균, 이연창, 김상욱, 이종욱 (2017). 추천 시스템에서의 컨텐츠 기반 필터링 기법의 효과. 한국정보과학회 학술발표논문집, 266-267.

추천 시스템에서의 컨텐츠 기반 필터링 기법의 효과*

홍동균[O1], 이연창[1], 김상욱[1], 이종욱[2]
[1]한양대학교 컴퓨터·소프트웨어학과
[2]성균관대학교 소프트웨어학과
{dg4271, lyc0324, wook}@hanyang.ac.kr, jongwuklee@skku.edu

Effectiveness of Content-based Filtering Technique on Recommendation System*

Dong-Gyun Hong[O1], Yeon-Chang Lee[1], Sang-wook Kim[1], Jongwuk Lee[2]
[1]Dept. of Computer and Software, Hanyang University
[2]Dept. of Software, Sungkyunkwan University

요 약

협업 필터링 중심의 추천 시스템은 cold-start 문제를 완화하기 위해 컨텐츠 기반 필터링 기법을 추가로 활용한다. 기존 컨텐츠 필터링 방법들은 아이템의 내용 정보로부터 잠재 내용 벡터를 추출하여 평점이 부족한 아이템을 추천하는데 활용 한다. 우리는 실제로 이러한 잠재 내용 벡터가 해당 아이템을 잘 표현하고 있는지 확인하기 위하여, 영화 도메인에서 잠재 내용 벡터로 얻은 이웃 영화의 각 영화들의 관련성을 확인한다.

1. 서론

추천 시스템은 한 사용자의 이용 기록을 분석하여 해당 사용자가 다음으로 선호할 만한 아이템을 찾아서 제공해준다. 추천 시스템은 크게 추천 받을 타겟 사용자와 유사한 취향의 사용자가 선호했던 아이템을 추천하는 협업 필터링 (collaborative filtering) 기법과 추천 받을 사용자가 선호 했던 아이템과 유사한 내용 정보를 갖는 아이템을 추천하는 컨텐츠 기반 필터링 (content-based filtering) 기법으로 나누어진다. Netflix 추천 경진 대회에서 협업 필터링 기법이 놀라운 성능을 보여준 이후, 컨텐츠 기반 필터링 기법에 비해 협업 필터링 기법이 더 활발하게 연구되고 있다.

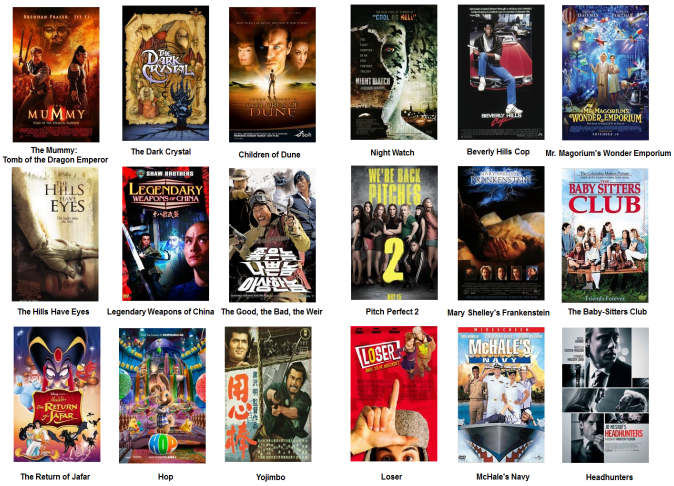
그러나 협업 필터링 기법 중심의 추천 시스템은 사용자의 선호도 기록이 거의 없거나 전혀 없는 아이템을 추천해주는 데에 어려움을 겪는데, 이를 cold-start 아이템 문제라고 한다. Cold-start 아이템 문제를 완화하기 위한 대표적인 연구로 collaborative topic regression (이하 CTR) [1], collaborative deep learning (이하 CDL)이 있다[2]. 위의 연구들은 각각 토픽 모델링과 딥러닝 기술을 기반으로 아이템들의 내용 정보를 분석하여 생성된 cold-start 아이템의 잠재 내용 벡터를 기존 협업 필터링 기법에 활용하는 하이브리드 방식으로 cold-start 아이템 문제를 완화하고자 하였다.

본 논문에서는 실제로 이러한 컨텐츠 기반 필터링 기법으로 분석하여 얻은 cold-start 아이템의 잠재 내용 벡터가 해당 아이템을 제대로 표현하는지 확인하고자 한다. 이를 위해, 우리는 먼저 영화 도메인에서 신작 영화가 갖는 내용 정보 (줄거리, 장르, 출연진 정보)를 기반으로 얻은 잠재 내용 벡터를 활용하여 해당 영화와 유사한 이웃 영화들을 찾는다. 그 후, 우리는 설문조사 형태로 신작 영화와 해당 영화의 이웃 영화들 간의 관련성을 평가함으로써, 잠재 내용 벡터가 해당 영화를 얼마나 잘 표현하는지 확인해보고자 한다.

2. 컨텐츠 기반 필터링 기법

문서 정보를 효과적으로 표현하는 토픽 모델링 및 딥러닝과 같은 방법론들이 대두됨에 따라 해당 방법론들을 추천 시스템에 활용하고자 하는 고찰이 이루어졌고 대표적인 연구로 CTR, CDL, ConvMF [3] 가 있다. CTR은 내용 정보를 아이템의 프로파일에 반영하기 위해 대표적인 토픽 모델링 기법인 Latent Dirichlet Allocation (LDA)를 행렬 분해 기법인 Probabilistic Matrix Factorization (PMF)에 통합시킨 연구이다. CTR은 평점 정보 외의 LDA 기술을 통해 추출한 아이템의 토픽 정보를 추가로 활용함으로써 기존 평점 정보만으로는 파악이 어려웠던 cold-start 아이템의 특성을 파악할 수 있다. CTR과 달리, CDL과 ConvMF는 각각 딥러닝 기술인 Stacked Denoising Auto-Encoder (SDAE), Convolutional Neural Network (CNN)을 PMF에 통합시킨 연구이다. CDL은 CTR이 부족한 내용 정보를 사용하는 상황에서 어려움을 겪는 단점을, 딥러닝 방법을 통해 해결하여 정확도를 향상시킨 연구이다. ConvMF은 텍스트를 입력으로 갖는 CNN을 활용하여 'Bag of words' 방식인 LDA, SDAE보다 문서의 문맥을 고려해서 더욱 효과적으로 내용 정보를 분석할 수 있음을 보인 연구이다.

* 이 성과는 2017년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. NRF-2015R1C1A1A01055442, 2017R1A2B3004581)



(a) LDA (b) CNN

그림1. 신작 영화에 대한 (a) LDA 기반 이웃 영화와 (b) CNN 기반 이웃 영화

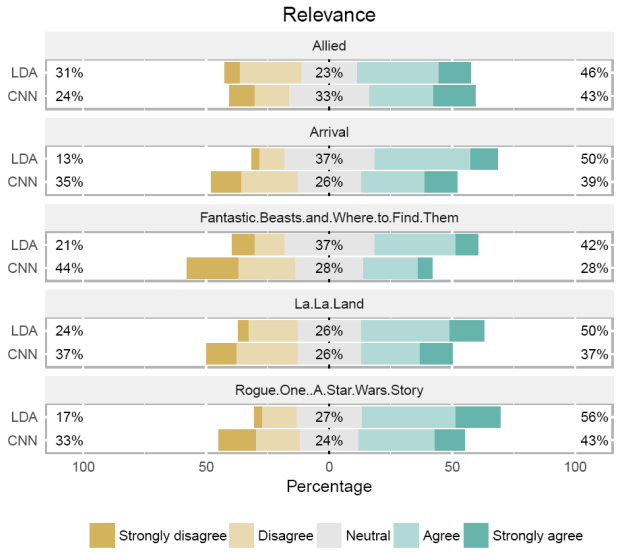


그림2. 내용 기반 이웃 아이템의 관련성 평가 결과

3. 실험

우리는 먼저 IMDB*에서 신작 영화(cold-start 아이템) 5편을 선정하였다. 그 후, 해당 영화들의 이웃 영화들을 찾기 위해, 추천 시스템의 성능 평가를 위한 벤치마크 데이터로 자주 사용되는 MovieLens**의 ml-latest-small 데이터를 사용하였다. 우리는 영화들의 내용 정보를 얻기 위해 각 영화의 Wikipedia 페이지에서 줄거리 정보를 크롤링하여 사용하였다.

우리는 LDA와 CNN을 각각 사용하여5편의 신작 영화와 ml-latest-small 데이터에 포함된 모든 영화들에 대해 잠재 내용 벡터를 추출하였다. 이후 각 신작 영화의 잠재 내용 벡터와 코사인 유사도가 가장 높은 상위 9편의 영화를 찾고, 이들을 각 신작 영화의 이웃 영화들로 선정하였다.

그림 1은 신작 영화 <신비한 동물 사전>에 대해 LDA(그림1.(a))와 CNN (그림1.(b))을 기반으로 선정한 이웃 영화 결과들을 보여준다. <신비한 동물 사전>은 신비한 동물을 마법으로 부릴 수 있는 남자에 대한 줄거리를 갖는 영화이다. 그러나 이웃 영화의 각 영화를 살펴보면 해당 영화와 관련성이 부족해 보이는 <좋은놈, 나쁜놈, 이상한놈>, <헤드 헌터>와 같은 영화들이 다수 포함된 것을 확인할 수 있다.

우리는 각 신작 영화와 이웃 영화들 간의 관련성을 정량적으로 평가하기 위하여 100명의 사람들에게 설문조사를 진행하였다. 응답자들은 각 방법으로 얻은 이웃 영화와 각 신작 영화 간의 관련성을 '전혀 관련 없음'부터 '매우 관련 있음' 까지 5단계의 리커트 척도***로 평가하였다. 모든 응답자들의 평가를 종합한 결과는 그림 2와 같다. 그림 2는 다섯 편의 영화에 대해서 각 컨텐츠 기반 필터링 기법의 결과에 대한 응답 비율을 보여준다. 그림 내의 왼쪽, 가운데, 오른쪽의 숫자(%)들은 각각 '관련 없음', '중립', '관련 있음'의 비율을 보여준다. 결과를 보면 두 방법 모두 '관련 있음'의 비율이 평균적으로 50%를 넘지 못하고 있음을 알 수 있다. 뿐만 아니라, '관련 없음'의 평균 비율도 LDA는 21%, CNN은 34%로 꽤 큰 비율의 응답자가 내용 정보를 기반으로 얻은 이웃 영화들과 각 신작 영화 간의 관련성이 부족하다고 응답했음을 확인 할 수 있다.

4. 결론

Cold-start 아이템 문제를 완화하기 위해서 기존 협업 필터링 중심의 추천 알고리즘들은 컨텐츠 기반 필터링 기법을 추가로 활용해왔다. 그러나 영화 도메인에서 영화들의 내용 정보를 분석하여 잠재 내용 벡터를 얻고, 이를 기반으로 얻은 신작 영화의 이웃 영화들을 평가해본 결과, 사람이 판단하기에 이들 간의 관련성이 높지 않음을 확인하였다. 따라서, 우리는 향후 이러한 컨텐츠 기반 필터링 기법의 한계를 개선하는 연구를 수행하고자 한다.

참고 문헌

[1] C. Wang and D. M. Blei. "Collaborative topic modeling for recommending scientific articles" *In Proc. of ACM KDD*,pages 448-456, 2011.

[2] H. Wang et al. "Collaborative deep learning for recommender systems," *In Proc. of ACM KDD*, pages 1235-1244, 2015.

[3] Kim, Donghyun, et al. "Convolutional matrix factorization for document context-aware recommendation." *In Proc. of ACM RecSys*, pages 233-240 2016.

* http://www.imdb.com/

** http://grouplens.org/datasets/movielens

*** https://en.wikipedia.org/wiki/Likert_scale