한국컴퓨터정보학회 동계학술대회 논문집 제23권 제1호 (2015. 1)

# 연령 및 프로그램 줄거리를 활용한 콘텐츠 기반 TV 프로그램 추천 시스템

방한별°, 이혜우*, 이지형*
°*성균관대학교 정보통신대학
e-mail:{hbyul91, ddxplus, john}@skku.edu°*

# A Content-based TV Program Recommendation System Using Age and Plots

Hanbyul Bang°, HyeWoo Lee*, Jee-Hyong Lee*
°*College of Information and Communication Engineering, Sungkyunkwan University

● 요 약 ●

추천 시스템의 대표적인 연구 중 하나인 콘텐츠 기반 추천 시스템 연구는 TV 프로그램이나 영화의 줄거리, 장르, 리뷰 등의 콘텐츠의 메타데이터를 이용한다. 그러나 이러한 연구들은 콘텐츠 관련 정보에만 의존할 뿐, 시청자의 프로파일과 콘텐츠의 정보를 함께 고려하지 않는다. 본 논문에서는 시청자의 프로파일 중 연령과 콘텐츠의 정보인 프로그램의 줄거리를 활용한 TV 프로그램 추천 시스템을 제안한다. 본 추천 시스템은 시청자를 연령에 따라 분류한 후, LDA 알고리즘을 이용하여 시청자의 시청 TV 프로그램의 줄거리를 분류된 나이에 따라 각각의 줄거리 토픽 모델로 생성한다. 이를 기준으로 시청자가 원하는 시간대에 방송되는 프로그램들의 줄거리 토픽벡터와 시청자의 선호도 토픽벡터의 유사도를 비교해 가장 유사도가 높은 TV 프로그램을 시청자에게 추천하는 방식이다. 본 논문에서는 연구의 효용성을 검증하기 위해 줄거리만을 사용한 경우와 줄거리와 연령을 동시에 활용한 경우를 비교 실험하였다. 실험을 통해 프로그램의 줄거리만을 사용한 경우보다 연령을 동시에 활용한 경우의 추천 시스템 성능이 개선된 것을 확인할 수 있었다.

**키워드:** 콘텐츠 기반 추천 시스템(content-based recommendation system), 프로그램 줄거리(program plot), 연령(age), 잠재 디리클레 할당(Latent Dirichlet Allocation)

## I. 서 론

디지털 방송의 보편화가 이루어짐에 따라 시청자들은 더 많은 채널을 접할 수 있게 되었다. 이로 인해 시청자들이 시청할 수 있는 프로그램 수는 더욱 많아졌으나, 많은 프로그램들 중 무엇을 시청할지 결정하고 선택하는 것은 더욱 어려워졌다. 이와 같은 어려움을 극복하기 위하여, 최근 TV 프로그램 추천 시스템에 대한 연구가 활발히 진행되고 있다[1,2,3].

대표적인 연구 중 하나인 콘텐츠 기반 추천 시스템 연구는 TV 프로그램이나 영화의 줄거리, 장르, 리뷰 등의 콘텐츠의 메타데이터를 이용한다[1,2]. 그러나 이러한 연구들은 콘텐츠 관련 정보에만 의존할 뿐, 시청자의 프로파일과 콘텐츠의 정보를 함께 고려하지 않는다.

2004년 통계청의 사회통계조사에 따르면 청년층은 주로 오락 장르의 프로그램들을 선호하였으며, 중장년층에서는 뉴스, 노년층에서는 연속극을 선호하는 것으로 조사되었다.[1] 또한, 드라마 장르의 프로그램은 시청자의 성별이나 학력보다는 시청자의 연령대가 가장 큰 영향을 미친다는 연구결과가 발표된 바 있다[3]. 이처럼 TV 프로그램 선호도는 시청자의 연령에 따라 크게 달라질 수 있다. 따라서 시청자의 만족도를 높일 수 있는 TV 프로그램 추천 시스템을 개발하기 위해서는 TV 프로그램 관련 정보뿐만 아니라, 시청자의 정보 또한 고려하여야 한다.

본 논문에서는 시청자의 연령과 프로그램의 줄거리를 활용한 TV 프로그램 추천 시스템을 제안한다. 시청자를 연령에 따라 분류한 후 LDA 알고리즘을 이용하여 TV 프로그램의 줄거리를 모델링한다. 그 다음 해당 시청자에 연령에 대응되는 모델을 이용해 시청자의 선호 프로그램과 가장 유사도가 높은 TV 프로그램을 추천한다.

논문의 구성은 다음과 같다. 2장에서는 추천 시스템과 LDA 알고리즘에 대하여 설명한다. 다음으로 3장에서는 본 논문에서 제안하는 "연령 및 프로그램 줄거리를 활용한 콘텐츠 기반 TV 프로그램 추천 시스템"에 대해 설명한다. 4장에서는 실험을 통하여 해당 시스템에 대해 분석한다. 마지막으로 5장에서는 본 논문에 대한 결론을 제시한다.

1) www.forwoman.or.kr

## II. 관련 연구

### 1. 관련연구

#### 1.1 추천 시스템

추천 시스템은 대표적으로 협업 필터링 추천 시스템(Collaborative Filtering Recommendation System)과 콘텐츠 기반 추천 시스템(Content-based Recommendation System)으로 나눌 수 있다[4].

협업 필터링 추천 시스템은 자신과 취향이 비슷한 사람들의 평가를 기준으로 추천해주는 시스템으로 가장 많이 사용되는 추천 시스템이다. 그러나 자신과 취향이 비슷한 사람의 평가 내역을 이용하기 위해서는 충분한 양의 데이터가 수집되어야 하기 때문에 추천 시스템 초반에는 정확한 추천이 어려우며 정확한 추천을 위해서는 충분한 데이터 수집을 위한 시간이 요구된다[1].

콘텐츠 기반 추천 시스템은 정보 검색분야를 기본으로 하는 추천 시스템으로 시청자에 대한 정보나 추천하고자 하는 프로그램에 대한 정보를 분석하여 추천하는 시스템이다. [1]에서는 TV프로그램의 줄거리, 장르 등을 아이템으로 사용한 콘텐츠 기반 추천 시스템을 제안하였으며, [2]에서는 영화의 리뷰, 줄거리 등을 이용한 콘텐츠 기반 추천 시스템을 제안했다. 그러나 이와 같은 연구들은 단순히 장르나, 줄거리와 같은 콘텐츠의 메타데이터만을 이용하였을 뿐, 시청자의 정보를 함께 고려하지 않았다.

#### 1.2 Latent Dirichlet Allocation (LDA)

LDA는 문서에 대해 어떤 주제가 존재하는지를 확률적으로 모델링 할 수 있는 알고리즘이다[5]. LDA를 사용해 주어진 문서들에 대해 어떠한 주제와 단어들로 구성되어 있는지에 대한 분석과 모델링이 이루어진다. LDA은 그림 1과 같이 Graphical Model로 표현이 가능하다.

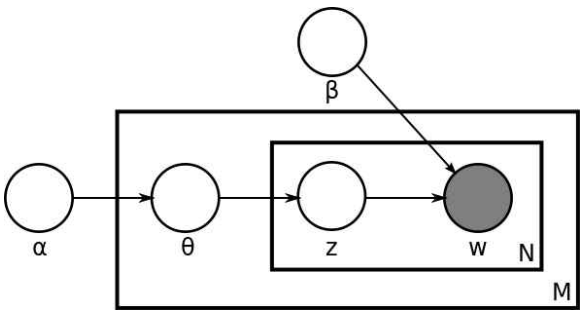


그림 46. LDA 그래프 모델
Fig. 1. Graphical Model of LDA

- $\alpha$ : Dirichlet 문서(document)-주제(topic) 분포의 매개변수
- $\beta$ : Dirichlet 주제(topic)-단어(word) 분포의 매개변수
- $\theta$ : 혼합된 여러 문서(document)의 주제(topic)에 대한 가중치
- z : 가중치에 의해 선택되는 주제(topic)
- w : 특정 주제(topic)에 의해 선택되는 단어(word)

그림 1의 Graphical Model은 단어 w가 주제 z에 포함될 확률을 나타낸 것이다.

## III. 제안방법

본 논문에서는 TV 프로그램의 줄거리와 시청자의 연령대를 동시에 고려한 추천 시스템을 제안한다. 제안 방법은 시청자를 연령에 따라 그룹으로 분류하여, 각 그룹의 시청 TV 프로그램의 줄거리를 LDA 알고리즘을 이용하여 모델링한 후, 이를 이용하여 시청자에게 TV 프로그램을 추천한다. 제안방법은 그림 2와 같다.

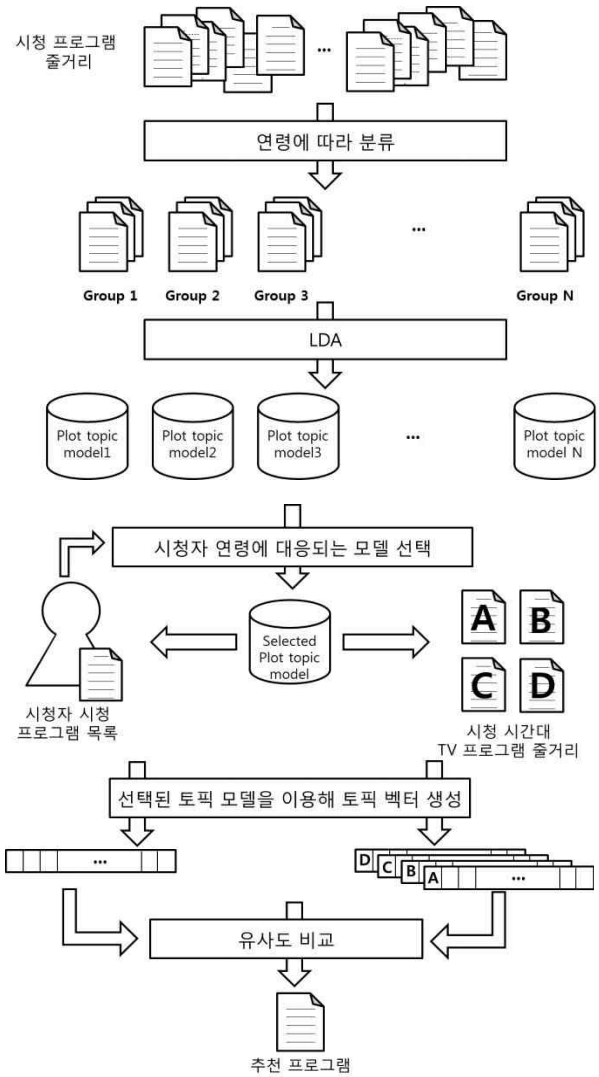


그림 2. 연령 및 프로그램 줄거리를 활용한 콘텐츠 기반 TV 프로그램 추천 시스템
Fig. 2. A Content-based TV Program Recommendation System Using Age and Plots

먼저 시청자의 연령 정보와 시청 TV 프로그램 줄거리를 수집한 후 연령을 기준으로 분류한다. 다음으로, 나눠진 그룹들의 시청기록에 해당하는 줄거리를 문서로 취급하여, LDA 알고리즘을 통해 각각의 토픽모델을 생성한다. 이후 시청자의 연령에 대응되는 줄거리 토픽 모델을 선택한다.

그 다음 시청자가 추천 받기를 원하는 시간과 동 시간대 방송되고 있는 프로그램들의 줄거리 토픽 벡터를 선택하여 이 토픽 벡터와

시청자의 선호도 토픽 벡터와의 유사도를 구한다. 이는 시청자가 선호하는 정도와의 유사도를 비교하기 위함이다. 유사도를 비교해 그 중 가장 유사도가 높게 나온 프로그램을 선정해 시청자에게 추천한다. 이 때 유사도를 계산하기 위하여 코사인 유사도(Cosine similarity)를 사용하였다. 코사인 유사도는 두 벡터간의 각도의 코사인 값을 이용하여 유사도를 구하는 방법으로 다차원 공간의 벡터를 다룰 때 유용하기 때문에 이를 사용하였다. 코사인 유사도는 식 (1)과 같이 정의된다.

$$Similarity(\vec{A}, \vec{B}) = \frac{\vec{A} \cdot \vec{B}}{\| \vec{A} \| \| \vec{B} \|} \quad (1)$$

식 (1)에서 $\vec{A}$는 시청자의 선호도 토픽 벡터이며, $\vec{B}$는 시청자가 추천 받기 원하는 시간대 프로그램의 토픽 벡터이다. $\vec{B}$를 바꿔가며 가장 높은 유사도를 가지는 프로그램을 추천한다.

## IV. 실험 및 결과

### 1. 실험 데이터

실험 데이터는 전국 멀티미디어 통합 조사 기관인 TNmS[2]가 수집한 6개월간의 3,318명의 시청자 개인 시청 기록과 1,187개의 TV 프로그램 데이터를 사용하였다. 채널은 KBS1, KBS2, MBC, SBS 4개의 지상파 채널을 사용하였다.

### 2. 실험

본 논문에서는 각 개인 시청기록의 80%를 트레이닝 데이터로 사용하였고 20%를 테스트 데이터로 사용하였다. 20% 테스트 데이터를 해당 시청자의 연령에 맞는 줄거리 토픽 모델을 기준으로 시청자의 시청정보와 동 시간대에 방송되고 있는 TV 프로그램 줄거리 데이터를 이용하여 줄거리 토픽 벡터를 만든다.

실험에서는 청년층의 기준인 30세를 전후로 하여 시청자를 두 그룹으로 분류하여 토픽 벡터를 만들었다. 이 때 토픽 개수는 20으로 하였다. 그 후 시청자의 선호도 토픽 벡터와의 유사도를 구해 가장 유사도가 높은 프로그램을 골라 시청자가 실제 시청한 프로그램과 비교하여 예측 정확도를 구했다. 베이스라인으로는, 연령대를 고려하지 않았을 경우의 줄거리 기반의 콘텐츠 추천방법을 사용하였다.

베이스라인과의 성능 비교를 위해 평가 척도로 예측 정확도를 사용하였다. 예측 정확도는 식 (2)와 같다.

$$precision = \frac{tp}{tp + fp} \quad (2)$$

$precision$은 추천한 TV 프로그램 중 실제로 추천이 성공할 확률을 나타낸다. $tp$는 추천이 성공한 경우의 횟수이며, $fp$는 추천이 실패한 경우의 횟수이다.

2) http://www.tnms.tv

표 1. 예측 정확도
Table 1. Prediction Accuracy

| | 줄거리 | 줄거리 + 연령 |
|---|---|---|
| 예측 정확도 | 27.74% | 27.89% |

표 1에서 줄거리만을 이용한 베이스라인보다 시청자의 연령정보를 함께 고려한 제안방법의 예측 정확도가 향상된 것을 확인할 수 있었다.

## V. 결 론

본 논문에서는 연령 및 프로그램 줄거리를 활용한 콘텐츠 기반 TV 프로그램 추천 시스템을 제안하였다. 이는 연령을 기준으로 시청자를 분류해 시청자의 시청 프로그램 줄거리를 여러 개의 모델로 만들어 시청자의 연령에 따라 다른 줄거리 토픽 모델을 사용하게 한다. 그 후 시청자가 시청을 원하는 시간에 방송되는 프로그램들의 줄거리 토픽벡터와 시청자의 선호도 토픽 벡터간의 유사도를 구해 TV 프로그램을 추천하는 추천 시스템이다. TV 프로그램의 줄거리 정보와 연령을 사용한 LDA 모델링 실험에서 베이스라인보다 연령을 적용시켰을 때 정확도가 상승한 것을 확인할 수 있었다. 향후 연구에 있어 분류에 기준이 되는 나이로 프로그램 선호를 구분 짓는 연령을 사용하면 더욱 정확한 예측을 할 수 있을 것으로 판단된다.

## Acknowledgement

본 연구는 정부(미래창조과학부) 및 한국산업기술평가관리원의 SW컴퓨팅산업융합원천기술개발사업의 일환으로 수행된 연구임 (2014-044-024-002). 또한, 본 연구는 미래창조과학부 및 정보통신기술진흥센터의 산업융합원천기술개발사업(정보통신)의 일환으로 수행하였음(10041244, 스마트TV 2.0 소프트웨어 플랫폼).

## 참고문헌

[1] Sangwon Yoo, Hongrae Lee, Hyungdong Lee, and Hyungjoo Kim, “A Content-based TV Program Recommender,” The Korea Information Science Society, pp. 638-692, 2003.

[2] Shinhyun Ahn, and Chung-kon Shi, “Movie Recommendation System Based on Cultural Metadata,” The Korea Information Science Society, pp. 78-79, 2008.

[3] Ji Hoon Bae, Eun Ji Park, Da Seul Lee, Ji Won Goh, Hyunmin Kim, and Young-Hyuk Kim, “Recommendation of TV Program Based on Personalized Websurfing Information,” The Korea Information Science Society, Vol. 37, No. 2(C), pp. 217-221, 2010.

[4] Francesco Ricci, Lior Rokach, Bracha Shapira and Paul B. Kantor, “Recommender Systems Handbook,” Springer Verlag Gmbh, 2010.

[5] David M. Blei, Andrew Y. Ng, and Michael I. Jordan, “Latent

Dirichlet Allocation," the Journal of machine Learning research, pp. 993-1022, 2003.