토픽 모델링을 이용한 유사 시청 사용자 그룹핑 및 TV 프로그램 추천 알고리듬

표신지[1)] 김은희[2)] 김문철[1,2)]

한국과학기술원 정보통신공학과[1)], 전기및전자공학과[2)]

sjpyo@kaist.ac.kr, lins77@kaist.ac.kr, mkim@ee.kaist.ac.kr

Topic modeling based similar user grouping and TV program recommendation for Smart TV

Pyo, Shinjee Kim, EunHui Kim, Munchurl

Dept. of Information and Communications Engineering[1)], Dept. of Electrical Engineering[2)]

Korea Advanced Institute of Science and Technology

요약

본 논문에서는 토픽 모델링 기반 TV 프로그램 유사 시청 사용자 그룹핑 및 이를 이용한 TV 프로그램 콘텐츠 추천 알고리듬을 제안하였다. 제안 기술은 토픽 모델링 기법 중 Latent Dirichlet Allocation(LDA) 방법을 이용하여 TV프로그램 시청 기록 내에서 은닉된 유사 사용자들을 그룹핑하고 이러한 유사 시청 사용자 그룹 정보를 이용하여 사용자에게 선호 TV 프로그램 콘텐츠를 자동으로 추천하는 알고리듬이다. 제안된 자동 추천 알고리듬의 성능평가를 위해 실제 TV 시청기록 데이터를 이용하여 훈련 기간과 검증 기간을 나누어 훈련 기간 동안 제안한 알고리듬을 이용하여 사용자 개인에 대한 추천 TV 프로그램 콘텐츠 목록을 생성하여 검증 기간 동안에 실제 추천된 TV프로그램을 얼마나 시청했는지를 측정하여 추천 정확도를 검증하였다.

1. 서론

최근 스마트 TV의 등장으로 TV 단말에서의 웹 콘텐츠 시청, 웹 서핑, TV 어플리케이션 사용 등이 자연스러워 졌다. 또한 양질의 콘텐츠와 다매체 서비스인 IPTV 서비스, 케이블 TV서비스들로 인해 사용자들은 TV 단말을 통해 이전보다 훨씬 더 많은 양의 콘텐츠들을 제공받게 되었다. 하지만 이러한 콘텐츠 양의 기하급수적인 증가는 사용자 자신이 원하는 콘텐츠를 찾아 시청하고자 할 경우 많은 부담감으로 작용하고 있다. 따라서 본 논문에서는 사용자의 과거의 시청 기록 데이터를 바탕으로 토픽 모델링 기법을 이용하여 사용자와 유사한 시청 형태를 갖는 유사 시청 사용자 그룹을 추론하고 이를 통해 사용자가 원하는 콘텐츠들을 자동으로 추천하는 알고리듬을 제안한다. 제안한 알고리듬의 성능 검증을 위해 실제 시청 기록 데이터를 이용하여 검증 기간 동안의 사용자 시청 기록 데이터에 대해 추천 정확도를 측정하여 검증한다.

II. 관련 연구

본 논문과 관련된 연구로는 TV프로그램 시청 기록을 LDA에 적용하여 TV프로그램들로 구성된 토픽을 생성하고 이를 기반으로 사용자에게 해당 토픽에 존재하는 TV 프로그램들을 자동 추천하는 연구가 있었다[1]. 관련 연구에서는 LDA를 적용하기 위한 데이터 구성방법이 사용자를 문서로, 사용자가 시청한 TV 프로그램 콘텐츠들을 단어들로 정의하여 LDA를 적용하였으며, 하이퍼 파라미터인 α와 β를 업데이트하여 비대칭 파라미터로 모델을 구성했을 때와 대칭 파라미터로 모델을 구성했을 때의 성능 비교를 통해 비대칭 파라미터로 모델을 구성했을 때가 성능이 향상됨을 확인하였다. 본 논문에서는 이와 달리 유사 시청 사용자 그룹핑을 통하여 사용자가 속한 유사 시청 사용자 그룹을 추출하고, 이를 기반으로 TV 프로그램 콘텐츠를 추천하는 자동 추천 알고리듬을 제안하였다.

또한 본 논문에서 제시하는 알고리듬은 토픽 모델링 기법에 기반한 알고리듬으로서 토픽 모델링 기법은 본래 문서 내에서 은닉 주제들을 찾아내기 위해 개발된 통계 추론 모델이다[2]. 문서들이 특정 주제들을 가지고 작성되었다는 가정 하에 전체 문서들에 분포된 은닉 주제들을 찾아내는 방법이다. 이러한 토픽 모델링기법 중 가장 대표적인 방법으로 LDA(Latent Dirichlet Allocation) 알고리듬이 있으며, LDA 알고리듬은 생성 모델로서 문서 내의 은닉 주제들을 찾아내는 알고리듬이다[2]. LDA에서는 파라미터 사이의 확률 관계를 Dirichlet 분포와 다항 분포로 정의하여 표현하는데, 이는 Dirichlet 분포와 다항 분포 사이에 conjugate prior 관계가 성립하여 사후 확률을 계산할 때 사전 확률과 likelihood 곱으로 나타나는 수식이 사전 확률로 주어지는 Dirichlet 분포와 동일한 수식의 모양으로 나타나는 이점이 있다. 생성 모델은 실제 문서를 작성하는 과정으로 보고 문서를 작성하기 위해 각 문서에 어떠한 주제들을 포함시킬 것인지, 또 그에 따라 어떤 단어들을 어떤 주제에서 선택하여 배치할 것인지를 각각의 파라미터로 모델링한다. 따라서 관찰 데이터로서 주어진 각 문서 내 단어들의 발생 빈도를 이용하여 생성 모델의 파라미터들을 추론하는 과정을 통해 전체 문서 집합의 은닉 주제들과 각 문서내의 주제 분포, 각 단어들이 각 주제에 포함될 확률들을 알아낼 수 있다. 이러한 과정을 그림 1과 같은 그래프 모델로 표현할 수 있다. 그림 1의 α와 β 는 하이퍼 파라미터로서 전체 문서 집합에 대해 동일한 값을 갖는다. α는 각 문서가 어떠한 주

제 비율로 구성될지를 나타내는 θ값을 결정하는 파라미터이다. θ는 Dirichlet 분포를 따르는 값이며, 따라서 α값에 따라 θ가 분포하게 될 Dirichlet 분포의 형태가 결정된다. 마찬가지로 β도 각 단어가 어떠한 주제들의 비율로 구성될지를 나타내는 ϕ 값을 결정하는 파라미터로서 β값에 따라 ϕ가 분포하게 될 Dirichlet 분포의 형태가 결정된다. 또한 θ는 각 문서에 대한 주제 비율 값으로서 Dirichlet 분포를 따르며 θ값에 따라 문서 내에 존재하는 단어들의 주제 z가 결정된다. z는 다항 분포를 따르는 은닉 확률 변수로서 다항 분포 파라미터가 θ가 된다. 각 단어의 주제를 나타내는 z값과 각 단어에 대한 전체 주제에 대한 비율 값 ϕ값 에 따라 단어 w가 결정된다.

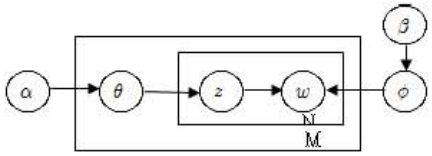

그림 1 LDA의 그래프 모델

그림 1의 LDA 그래프 모델에 대한 결합 분포를 나타내면 다음과 같다.

$$p(\boldsymbol{\theta},\boldsymbol{z},\boldsymbol{w},\boldsymbol{\phi}|\boldsymbol{\alpha},\boldsymbol{\beta}) = \prod_{m=1}^{M}\prod_{n=1}^{N_m} p(z_{m,n}|\theta_m)p(\theta_m|\boldsymbol{\alpha})\,p(w_{m,n}|z_{m,n},\boldsymbol{\phi})p(\phi|\beta) \quad (1)$$

전체 문서(M개)와 각 문서에 포함된 전체 단어(N_m개)들의 확률 값의 곱으로 결합 분포를 나타낼 수 있다. 하지만 수식 (1)에서 실제로 우리가 알고 있는 확률 값과 변수는 없다. 단지 각 문서에 포함되어 있는 단어와 각 단어의 빈도만을 알고 있기 때문에 다양한 approximation 방법을 이용하여 은닉 변수인 θ,z,ϕ, α, β들을 추론해야 한다. 은닉 변수를 추론하는 방법에는 variational Bayesian inference[2], Mean field variational approximation[3], Gibbs sampling[4] 등이 존재하며 이러한 추론 과정을 통해 은닉 변수와 은닉 주제를 알 수 있다. 다음 절에서는 이러한 LDA를 기반으로 본 논문에서 제안하는 유사 시청 사용자 그룹핑에 대해 설명한다.

Ⅲ. 토픽 모델링 기반 TV 프로그램 유사 시청 사용자 그룹핑 및 TV프로그램 추천 알고리듬

본 논문에서 제안하는 추천 알고리듬은 다음과 같은 과정으로 이루어져있다.

과정 1) 토픽 모델링 기법을 이용하여 유사 시청 사용자들을 그룹핑하는 과정

과정 2) 유사 시청 사용자 그룹을 기반으로 각 사용자의 추천 TV 프로그램 목록을 생성하는 과정

과정 1은 TV프로그램 시청 기록 데이터에 문서 검색 및 문서 내의 주제 추론 기법인 토픽 모델링 기법을 적용하여 유사 시청 사용자들의 그룹을 생성하는 단계이고, 과정 2는 이를 바탕으로 각 사용자가 속하는 그룹 정보를 이용하여 각 사용자에 적합한 추천 TV프로그램 목록을 생성하여 추천한다.

3.1 유사 시청 사용자 그룹핑

기존의 LDA에서는 문서와 문서 내의 단어들을 입력 값으로 하여 알고리듬의 연산을 수행하였다. 본 논문에서는 이러한 문서와 문서 내의 단어들을 시청 TV 프로그램과 해당 TV 프로그램을 시청한 사용자들로 정의하여 토픽 모델링에 적용한다. 즉, 각각의 TV 프로그램 콘텐츠가 하나의 문서가 되고, 각 TV프로그램을 시청한 사용자들과 그 사용자가 해당 TV 프로그램을 시청한 횟수가 하나의 문서를 구성하는 단어와, 단어의 빈도가 되도록 하였다. 표 1은 TV프로그램 시청 데이터의 LDA 적용을 위한 데이터 구성을 나타낸다. 표 1과 같이 데이터를 구성하고 LDA 알고리듬에 적용하였다. 은닉 파라미터 추론은 Gibbs sampling 기법을 이용하여 full conditional probability값을 통한 샘플링과 파라미터 업데이트를 통한 추론으로 θ, ϕ, α, β 값을 추론하였다[5].

표 1. TV프로그램과, 해당 TV프로그램을 시청 사용자들과 시청 횟수로 구성한 데이터 예시

TV 프로그램	시청한 사용자수	사용자ID:시청횟수
KBS뉴스9	1875	101021301:71 101021302:38 102020301:61 102020302:60...
시사매거진 2580	1139	101020702:3 101020705:2 101021302:4 104023001:5…
일일연속극<웃어라동해야>	1675	101020702:23 101021301:62 101021302:33 101021602:8 ...

샘플링과 파라미터 업데이트를 위한 수식은 다음과 같다.

$$P(z_i = k|z_{-i},w) \propto \frac{\beta_{w_i} + n_{k,-i}^{(w_i)}}{\sum_{t=1}^{V}\beta_t + n_{k,-i}^{(t)}} \frac{\alpha_k + n_{k,-i}^{(m_i)}}{\sum_{k'=1}^{K}\alpha_{k'} + n_{k',-i}^{(m_i)}} \quad (2)$$

$$\theta_k^{(m)} = P(z_i = k|m) = \left(\alpha_k + n_k^{(m)}\right) / \left(\sum_{k'=1}^{K}\alpha_{k'} + n_{k'}^{(m)}\right) \quad (3)$$

$$\phi_k^{(w)} = P(w_i = w|z_i = k) = \left(\beta_w + n_k^{(w)}\right) / \left(\sum_{t=1}^{V}\beta_t + n_k^{(t)}\right) \quad (4)$$

k 인덱스는 토픽에 대한 인덱스이고, i 인덱스는 문서 내의 각 단어에 대한 인덱스이다. 또한 V는 전체 단어의 종류 수이고, K는 전체 토픽의 개수, t 는 단어 사전에 저장된 단어의 인덱스이며, m은 문서에 대한 인덱스이다. 따라서 $P(z_i = k|z_{-i},w)$ 는 문서 내의 단어 w_i 에 대한 토픽 z_i를 제외한 모든 단어에 대한 토픽 값과, 모든 단어가 주어졌을 때, 문서 내의 단어 w_i 에 대한 토픽 z_i가 k일 확률이며 이 값을 sampling 시에 이용하여 가장 높은 확률을 갖는 k값을 z_i에 할당하도록 하였다. 또한 $\theta_k^{(m)}$는 문서 m 의 k 토픽에 대한 확률 값이고, $\phi_k^{(w)}$는 특정 단어 w 의 k토픽에 대한 확률 값이며, 이러한 파라미터들의 전개 과정은 본 논문에서는 생략하도록 한다(수식에 대한 자세한 내용과 α와 β의 업데이트 수식은[5] 참고). 표 1과 같이 구성된 데이터를 입력 데이터로 이용하고, 토픽 개수 K 값을 입력 값으로 하여 Gibbs sampling을 수행하게 되면, 은닉 파라미터들이 추론되어 ϕ값을 통해 각 토픽을 어떠한 단어들이 구성하고 있는지를 알 수 있다. 이렇게 생성된 토픽은 문서 내에서 유사한 발생 형태를 갖는 단어들로서 본 논문에서 정의한 문서와 단어의 개념을 고려할 때, 유사 시청 사용자 그룹으로 볼 수 있다. 즉, 생성되는 토픽 자체가 유사 시청 사용자 그룹이 되는 것이다. 다음 절에서는 생성된 유사 시청 사용자 그룹을 기반으로 사용자에게 추천하는 알고리듬에 대해 설명한다.

3.2 유사 시청 사용자 그룹 기반 추천 알고리듬

LDA 알고리듬과 Gibbs sampling을 통해 생성된 유사 시청 사용자 그룹들이 존재할 때, 본 논문에서 제안하는 유사 시청 사용자 그룹 기반 추천 알고리듬을 이용하여 각 사용자들에 대하여 추천 TV 프로그램 목록을 생성할 수 있다. 유사 시청 사용자 그룹 기반 추천 알고리듬은 다음과 같은 과정으로 진행된다.

과정 1) 파라미터 ϕ를 이용한 사용자 별 사용자 그룹 정렬

과정 2) 파라미터 θ를 이용한 사용자 그룹 별 TV 프로그램 콘텐츠 정렬

과정 3) 과정 1과 2를 이용한 각 사용자 별 사용자 그룹을 기반으로 추천 TV프로그램 콘텐츠 목록 형성

각 단계에 대해 설명하면 다음과 같다. 과정 1은 먼저, 각 사용자들의 각 토픽(유사 사용자 그룹)에 대한 확률 값인 ϕ값을 이용하여 각 사용자들이 어떠한 토픽에 포함될 확률이 높은 순서로 정렬한다. 정렬된 토픽의 인덱스 값을 저장하여 각 사용자가 포함될 확률이 높은 토픽 순서대로 ϕ'행렬에 저장하고, 또한 토픽의 인덱스는 ϕ'_{idx} 행렬에 저장한다. ϕ'와 ϕ'_{idx} 는 $V\times K$ 행렬이고, V는 전체 사용자의 수(단어 수), K는 전체 토픽 수(사용자 그룹 수)이다. 따라서 $\phi'_{idx_{i,j}}$는 i 번째 사용자가 j 번째로 높은 확률 값을 갖는 토픽의 인덱스 값, $\phi'_{i,j}$는 해당 인덱스에 대한 확률 값을 뜻한다. 또한 이것은 i 번째 사용자가 j 번째로 포함될 확률이 높은 사용자 그룹을 뜻한다.

과정 2에서는 그 θ값을 이용하여 각 토픽을 어떠한 문서가 많이 포함하고 있는지, 비율 값이 큰 순서대로 정렬한다. 앞서 설명한 대로 θ값은 각 문서에 포함된 토픽들의 비율 값으로서, 각 문서에 대한 θ_m 값은 전체 토픽의 개수인 K 차원의 벡터 값이며 전체 토픽에 대하여 이들을 합했을 때 1이 된다. 따라서 전체 문서에 대한 θ값은 $M\times K$ 행렬로 나타내고, M은 전체 문서의 개수이다. 이러한 θ값을 토픽을 기준으로 해당 토픽을 가장 많이 포함하고 있는 문서 순서대로 해당 문서의 인덱스를 θ'_{idx} 행렬에 저장하고 정렬된 확률 값을 θ'에 저장한다. θ'와 θ'_{idx}는 $K\times M$ 행렬이고, $\theta'_{idx_{k,m}}$는 전체 문서 중, k 번째 토픽을 m 번째로 많이 포함하고 있는 문서의 인덱스 값, $\theta'_{k,m}$는 해당 문서 인덱스에 대한 확률 값을 나타낸다.

과정 3에서는 과정 1과 2에서 정렬한 ϕ', ϕ'_{idx}와 θ', θ'_{idx}를 이용하여 사용자 별 추천 TV프로그램 콘텐츠 목록을 형성한다. 즉, ϕ'와 ϕ'_{idx}를 이용하여 각 사용자의 선호 토픽을 알 수 있고, θ'와 θ'_{idx}를 이용하여 해당 토픽에 대한 문서를 알 수 있다. 따라서 각 사용자가 포함된 선호 사용자 그룹을 기준으로 해당 사용자 그룹을 많이 포함하고 있는 TV 프로그램 콘텐츠들을 각 사용자의 추천 TV 프로그램 목록으로 생성할 수 있다. 이를 정량적인 값으로 정렬하기 위해 다음과 같은 수식으로 콘텐츠 목록의 순위를 정할 수 있다.

$$S_{i,\theta'_{idx,k_{ij},m}} = \phi'_{i,j} \times \theta'_{k_{i,j},m}\ ,\ \ k_{i,j} = \phi'_{idx_{i,j}} \qquad (5)$$

$S_{i,\theta'_{idx,k_{ij},m}}$은 사용자 i 가 문서 인덱스 $\theta'_{idx_{k_{ij},m}}$에 대한 점수로서 사용자 i 가 j 번째로 포함될 확률이 높은 토픽에 대한 확률 $\phi'_{i,j}$값과 해당 토픽에 대한 정렬된 θ'중에서 m 번째로 해당 토픽을 많이 포함하고 있는 문서의 비율 값을 곱한 결과이다. 여기서 사용자 i 가 포함될 확률이 높은 순서대로 top N의 토픽을 고려 할 수 있으며, 이는 j 값을 1부터 N까지로 제한하여 계산할 수 있다. 또한 각 토픽별 정렬된 문서도 마찬가지로 top M개의 문서로 제한하여 계산할 수 있으며, 이는 m 값을 1부터 M까지로 제한하여 계산할 수 있다. 이로써 사용자 별 top N개의 토픽과 해당 토픽 별 top M개의 문서들에 대한 총 $N\times M$개의 $S_{i,\theta'_{idx,k_{ij},m}}$ 값이 계산되며, $S_{i,\theta'_{idx,k_{ij},m}}$값을 큰 순서대로 정렬하여 사용자에게 최종적으로 추천 콘텐츠 목록을 제공할 수 있다.

V. 실험 결과

4.1 실험 설계 및 데이터 처리

3.1 절의 표1과 같이 보유한 사용자 시청 기록 데이터에서 시청률 높은 순서대로 TV 프로그램을 추출하고, 해당 TV프로그램을 시청한 사용자들과 사용자들의 시청 횟수를 추출하였다. 보유한 데이터는 국내 시청률 조사기관인 TNmS로부터 구매한 데이터로써 2011년 1월부터 7월까지 7개월간 수도권의 2300여명의 시청자 패널로부터 수집한 시청 기록 데이터이다. 시청자는 id로 구분되며, 시청 TV 프로그램 제목, 시청 시작 시간과 종료 시간, 시청 날짜, 시청 요일, 시청 TV 프로그램 세부 정보 등이 포함되어 있다. 실험을 위해 시청률 상위 62개의 TV프로그램을 선별하였으며, 훈련 기간은 2011년 1월부터 4월 30일까지, 검증 기간은 2011년 5월부터 2011년 7월 31일까지로 정하여 실험을 수행하였다. 전체 문서의 개수 M는 전체 TV프로그램 콘텐츠 개수인 62가 되며, 전체 단어의 종류 수 V는 TV 프로그램을 시청한 전체 사용자 수인 2033명이 된다. 또한 Gibbs sampling을 통한 파라미터 추론을 위해 총 iteration 횟수는 2000회, 그 중 burn-in 기간은 500회까지, sampling lag는 10회 마다 실행하는 것으로 설정하여 추론하였으며, α, β의 초기 값은 α=2 β=0.1로 정하여 수행하였다. 생성할 토픽의 개수 K는 10, 20, 30, 40, 50, 60, 70, 80 까지 변화시켜가면서 수행하였다.

4.2 토픽 모델링 수행 결과

토픽 모델링을 수행한 결과, 유사 시청 사용자 그룹이 생성되었다. 생성된 사용자 그룹 내의 사용자들이 시청한 TV 프로그램들이 실제 유사한 것을 확인할 수 있었다.

표 2. K=50일 때 생성된 그룹(토픽)내의 사용자와 사용자의 시청 프로그램 목록

K	토픽#	사용자ID	시청 프로그램 목록 (시청 횟수로 정렬)
50	7	106023703	KBS뉴스9 MBC뉴스데스크 일일연속극<웃어라동해야>생방송오늘아침 ...
		105023602	KBS뉴스9 MBC뉴스데스크 SBS8시뉴스 일일연속극<웃어라동해야> ...
		205024102	KBS뉴스9 일일연속극<웃어라동해야> MBC뉴스데스크 일일시트콤<몽땅내사랑> ...
	34	111029003	KBS뉴스9 일일연속극<웃어라동해야>일일아침연속극<장미의전쟁> 일일아침드라마<사랑하길잘했어> ...
		111029004	KBS뉴스9 일일연속극<웃어라동해야> 일일아침연속극<장미의전쟁> 일일아침드라마<사랑하길잘했어> ...
		112026402	KBS뉴스9 일일연속극<웃어라동해야> 아침마당 KBS뉴스네트워크 6시내고향 주말연속극<사랑을믿어요> ...
	43	261022405	짱구는못말려10 일요일이좋다 SBS8시뉴스 MBC뉴스데스크 ..
		113021403	꼬마버스타요 MBC뉴스데스크 로보카폴리 캐니멀 ..
		121027504	캐니멀 로보카폴리 짱구는못말려10 모여라!딩동댕 ...

4.3 사용자 별 추천 콘텐츠 목록

4.2절의 LDA수행 결과, 생성된 유사 시청 사용자 그룹과 추론된 ϕ, θ 값과 각 사용자의 콘텐츠 별 점수 $S_{i,\theta'_{idx,k_{ij},m}}$에 따라 각 사용자의 추천 콘텐츠 목록을 생성하였다. 표 3은 각각의 K 값에 따라 추출된 사용자 별 추천 콘텐츠 목록을 나타낸다. 표 3와 같이 생성된 각 사용자 별 추천 콘텐츠 목록을 이용하여 추천 목록의 상위 5개, 10개, 15개, 20개로 사용자에게 추천하여 제공할 수 있다.

표 3. 사용자 별 추천 콘텐츠 목록

K	사용자ID	추천 콘텐츠 목록
10	202020102	유재석김원희의놀러와 특별기획<시크릿가든> 해피투게더 ...
	107022801	시청자칼럼-우리사는세상 KBS뉴스네트워크 6시내고향 ...
	121022702	전국노래자랑 바른말고운말 도전1000곡 ...

30	107022801	특선만화 짱구는못말려10 우리들의일밤 ...
	107021902	일일아침연속극<장미의전쟁> 일일아침드라마<사랑하길 잘했어> 아침드라마<당신참예쁘다>...
	231022902	인간극장<KBS1> 바른말고운말 극한직업 ...
50	204003102	시청자칼럼우리사는세상 KBS뉴스네트워크 6시내고향 ...
	110006502	위기탈출넘버원 주말극장<웃어요엄마> 일일시트콤<몽땅내사랑> ...
	223021102	로보카폴리 캐니멀 짱구는못말려10 ...

4.4 추천 정확도

추천 정확도의 측정을 위해서는 추천한 TV프로그램에 대해 실제로 얼마나 사용자가 시청했는지를 측정해야 하나 이는 많은 시간과 비용이 요구된다. 따라서 본 논문에서는 4.3 절에서 생성한 사용자 별 추천 목록으로 사용자에게 추천했을 시, 검증 기간 동안에 해당 추천 콘텐츠들을 사용자가 얼마나 시청되었는지를 측정함으로써 간접적으로 제안 알고리듬의 추천 성능을 측정한다. 추천 정확도는 다음의 수식으로 계산한다.

$$\mathrm{Precision}_i = \frac{1}{N}\sum_{n=1}^{N} C_{i,n} \ , \ C_{i,n} = \begin{cases} 1, Rec_{i,n} \in TVprog_i \\ 0, otherwise \end{cases} \quad (6)$$

여기서 $\mathrm{Precision}_i$는 사용자 i 의 추천 정확도이고, N은 추천 TV 프로그램 콘텐츠의 개수이다. $C_{i,n}$는 사용자 i 가 사용자 i 의 n번째 추천 프로그램 콘텐츠인 $Rec_{i,n}$를 시청했는지에 대한 표시자로써, 사용자 i 의 테스트 기간 동안의 시청 프로그램 목록 중에 $Rec_{i,n}$가 포함될 경우 1, 그렇지 않을 경우 0의 값을 갖는다. 따라서 $C_{i,n}$를 전체 추천 콘텐츠에 대해여 더한 값을 추천 콘텐츠 개수로 나누어 추천 정확도를 계산한다. 토픽의 개수에 따른 추천 정확도 비교를 위해 임의로 선별한 70명의 사용자들의 추천 정확도를 계산하였으며, 70명의 사용자들의 추천 정확도의 평균을 계산하였다. 표 4는 토픽 개수 별 사용자의 추천 정확도와 평균 추천 정확도이다. 표 4에서와 같이, 유사 시청 사용자 그룹의 개수인 K를 어떻게 정하느냐에 따라 추천 정확도가 달라지는 것을 확인할 수 있었다. 그리고 상위 5개, 상위 10개로 각각 추천해 보았을 때에, 추천 정확도에 차이가 나는 것을 알 수 있었다. 정확도가 가장 높은 경우는 사용자 그룹의 개수가 50이고, 상위 5개로 추천 해주었을 때, 사용자들의 평균 추천 정확도가 0.7212로 가장 높았으며, 가장 낮은 경우는 사용자 그룹의 개수가 10이고 상위 10개로 추천해주었을 때 0.4216으로 가장 낮았다. 또한 전반적으로 상위 5개로 추천해 주었을 때의 추천 정확도가 상위 10개로 추천해 주었을 때보다 높았으며, 사용자 그룹의 개수가 50일 때가 다른 사용자 그룹의 개수 일 때보다 추천 정확도가 높게 나타났다. 즉, 사용자 그룹의 개수가 50일 때 사용자들을 가장 적절하게 그룹핑 하여 ϕ와 θ값을 이용하여 생성한 사용자 별 TV 프로그램 콘텐츠 추천 목록에 대한 추천 정확도가 가장 높게 나타난 것으로 볼 수 있다. 또한 전반적인 평균 추천정확도가 낮게 나타난 것은 전체 문서의 개수, 즉 전체 TV프로그램 콘텐츠의 개수가 62개로 제한적이어서 모델의 안정성이 떨어져 추천 정확도에 영향을 미친 것으로 분석된다.

Ⅵ. 결론

본 논문에서는 LDA 알고리듬을 이용하여 유사 시청 사용자 그룹핑을 수행하고, 생성된 유사 시청 사용자 그룹들과 각 TV프로그램 별 그룹의 비율을 이용하여 각 사용자의 추천 TV 프로그램 목록을 생성하고 테스트 기간 동안의 시청 기록과 비교하여 추천 정확도를 계산함으로 성능을 검증하였다. 향후 계획으로 현재 구성된 62개의 TV프로그램 콘텐츠에 대한 데이터를 추가 보충하여 모델의 안정성을 높여 성능 고도화를 진행할 계획이며, TV프로그램 콘텐츠와 해당 콘텐츠에 대한 description 정보, 웹 콘텐츠와 해당 콘텐츠에 대한 description 정보를 이용하여 TV콘텐츠 뿐 만아니라 웹 콘텐츠도 함께 추천이 가능한 모델로 확장할 계획이다.

표 4. K=10, 30, 50, 70일 때 Top 5, Top 10 추천 시 사용자들의 평균 추천 정확도와 세 명의 사용자의 추천 정확도

Top 5 로 추천했을 시			
K	평균 precision	사용자 id	precision
10	0.7157	112022901	1
		251026205	1
		107900202	0.8
30	0.6138	112022901	0.4
		251026205	0.8
		107900202	0.8
50	0.7212	112022901	1
		251026205	1
		107900202	1
70	0.7012	112022901	0.667
		251026205	1
		107900202	0.6
Top 10 로 추천했을 시			
K	평균 precision	사용자 id	precision
10	0.4216	112022901	0.6
		251026205	0.9
		107900202	0.7
30	0.6013	112022901	0.7
		251026205	0.8
		107900202	0.7
50	0.6863	112022901	1
		251026205	0.89
		107900202	0.875
70	0.5601	112022901	0.375
		251026205	0.625
		107900202	0.6

Acknowledgement

이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임(2012-01120197)

본 연구는 지식경제부 및 한국산업기술평가관리원의 IT산업융합원천기술개발사업의 일환으로 수행하였음. [10039161, 스마트 TV의 UX 향상을 위한 UI 핵심 기술 연구]

참고문헌

[1] 김은희, 표신지, 김문철, "협업필터링 Latent Topic기반 Automatic TV Recommendation," 한국방송공학회 추계학술대회, pp.62-65, 2011년 11월.

[2] David M. Blei, Andrew Y. Ng, Michael I. Jordan, "Latent Dirichlet Allocation," Journal of Machine Learning Research 3, pp.993-1022, 2003.

[3] Blei, D. M., Lafferty, J. D.,"Correlated topic models," Advances in Neural Information Processing Systems 18. MIT Press, Cambridge, MA.

[4] Tomas L. Griffiths, Mark Steyvers, "Finding scientific topics," Proceedings of the National Academy of Sciences, vol. 101, suppl. 1, pp. 5228-5235, April, 2004.

[5] G. Heinrich. Parameter estimation for text analysis. Technical report, 2005.