협업 필터링 Latent Topic기반 Automatic TV Recommendation

*김은희 **표신지 ***김문철

한국과학기술원

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

Automatic TV Recommendation based on collaborative filtered Latent Topic

*Kim, EunHui **Pyo, Shinjee ***Kim, Munchurl

Korea Advanced Institute Science and Technology

요약

최근 화두가 되고 있는 스마트 폰 앱의 관심으로 스마트 TV의 앱에 대한 관심도 함께 증가하고 있다. TV시청 이용자들의 편의를 위해 증가하고 있는 수많은 채널과 콘텐츠 중, 개인 사용자의 이용 습관 및 대중의 선호 프로그램을 고려하여, 편리하게 원하는 TV프로그램에 접근하도록 해 주는 TV 앱이 있다면 이는 매우 중요한 기능으로 자리 잡을 가능성이 높을 것으로 예상된다.

이에 본 논문은 사용자의 시청 이용행태를 기반으로 주제모델링 기술의 고전적 모델인 LDA을 기반으로 협업필터링을 결합한 TV 선호 프로그램 추천 알고리듬을 제안한다. 개인의 관심 선호도는 일반적으로 특정 개수로 한정지어지는 특성을 고려하여, 개인 선호도 특성이 구별 되도록 두 가지 방법을 적용하였다. 하나는 개인 선호도 프로파일의 특정 상위 주제만을 고려하는 것이고, 또 다른 하나는 개인별 주제에 대한 선호도의 다양성이 드러나도록 비대칭 하이퍼-파라미터를 갖는 LDA를 사용 하였다. 실험 결과, 두 가지 방식에 대해 사용자의 실제 TV시청 이용내역 데이터를 기반으로 추천 성능의 향상을 평균 Precision 값을 측정하여 확인하였다. 또한, 본 논문에서는 주제 모델링을 통해 학습된 각 주제의 상위 확률의 TV 프로그램들을 분석한 결과, 하나의 주제가 개인별 시청의 특성 보다는 가족단위의 시청 특성을 드러냄을 확인할 수 있었다.

1. 서론

스마트 폰 앱(App.)에 대한 관심 증가는 스마트 TV 앱에 대한 관심의 증가로 이어지고 있다. TV의 다채널을 통해 소비 가능한 방대한 양의 콘텐츠 중 사용자의 시청 선도호에 맞게 원하는 콘텐츠로 접근할 수 있는 용이성과 대중이 선호하는 프로그램 정보를 제공할 수 있는 TV 프로그램 자동추천은 방대한 TV 프로그램 정보의 홍수로부터선호 TV 프로그램을 효율적으로 소비할 수 있도록 도와주는 중요한 TV 기능으로 자리 잡을 것으로 예상된다.

본 논문에서는 주제 모델링의 기본 모델인 LDA(Latent Dirichlet Allocation) 모델[1]을 기반으로 개인의 선호도와 동시에 대중의 선호도를 고려한 협업필터링 기반의 추천모델을 제안한다. 사용자에게 직접 TV 프로그램에 대한 선호여부를 요구하지 않고, 사용자의 시청 이용 데이터를 기반으로 사용자 선호도를 학습하고 이를 기반으로 사용자 선호도가 높은 TV 프로그램을 추천하는 모델을 제시한다.

TV 프로그램에 대한 사용자들의 시청 관심이 다양하기 때문에 사용자별 특정 프로그램에 대한 관심도의 다양성을 추천 모델에 적용하여 asymmetric Dirichlet hyper-parameter를 학습하는 데이터 기반의 모델 학습 방법을 이용 하였다 [3, 4]. 해당 방법들을 기반으로 순위정렬모델의 추천 성능을 실험 제시한다.

전형적인 주제 모델링에서 문서내의 단어 발생 빈도(확률)을 기반으로 주제 모델링을 할 경우, 단어들이 하나의 주제(토픽)을 갖도록 구성되는 것이 일반적이다. 그런데, TV 시청 이용 데이터를 기반으로 토픽 모델링 하였을 경우, 하나의 주제를 구성하는 TV 프로그램들이 TV시청 이용행태를 드러냄을 제시한다.

2. 관련 연구

본 논문에서 제시하는 추천 모델은 LDA에 기반한 방법으로, LDA는 사용자가 문서 작성을 위해 주제를 정하고 해당 주제에 걸맞은 단어를 선택하는 과정을 모델화하였다. 큰 하나의 주제 범위(corpus)를 갖는 여러 문서들의 단어들의 발생빈도를 이용하여 LDA모델을 통해 학습된 결과는 문서별 은닉주제(latent topic)에 대한 확률 분포 $\boldsymbol{\theta}$ 및 전체 corpus내의 단어들의 은닉주제에 대한 확률 분포 $\boldsymbol{\Phi}$가 생성된다. 그림 1은 LDA 모델을 그림으로 도식화 한 것으로서, D는 전체 문서의 개수를, Dw는 문서별 단어의 개수를 나타낸다. K는 은닉주제의 개수이다. V는 corpus내 전체 단어의 개수이다 [1].

※ 본 연구는 2011년도 일부 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것이고 (2011-0012085), 지식경제부 및 한국산업기술평가관리원의 IT산업융합원천기술개발사업의 일환으로 수행하였음. [10039161, 스마트 TV의 UX 향상을 위한 UI 핵심 기술 연구]

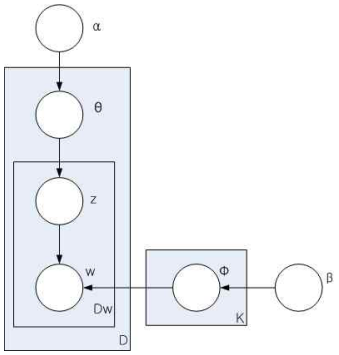

그림 1 LDA 도식 모델

Asunsion은 LDA모델의 parameter추정 방법에 대해 다양한 접근의 추론(inference)기술을 비교 실험한 바 있다. 각 기술의 최적화 hyper-parameter를 찾으면 marginal likelihood 확률 성능(perplexity 성능)에 큰 차이가 나지 않음을 실험으로 입증하였다 [2].

본 논문에서는 LDA 모델의 비대칭 하이퍼 파라미터 (asymmetric hyper-parameter)를 비롯한 모델 파라미터의 예측 값을 구하기 위해 깁스 샘플링(Gibbs sampling) 방법을 적용하였다. 문서별 각 발생 단어 토큰에 대해 은닉 주제(그림 1의 z)를 샘플링하면서 모델 파라미터 $\boldsymbol{\theta}$와 $\boldsymbol{\Phi}$를 다음 수식과 같이 업데이트 과정을 통해 계산 한다.

$$\widehat{\Phi_{kw}} = (N_{kw}^{\neg id} + \beta_w)/(N_{k.}^{\neg id} + \sum_w \beta_w) \quad (1)$$

$$\widehat{\Theta_{kd}} = (N_{kd}^{\neg id} + \alpha_k)/(N_{.d}^{\neg id} + \sum_k \alpha_k) \quad (2)$$

여기서 d 는 문서, w 는 단어(전체 단어 사전의 인덱스), i 는 단어의 문서 내 토큰, k 는 은닉 주제 인덱스를 나타낸다. $N_{kw}^{\neg id}$에서 $\neg id$는 현재의 토큰 i의 주제 할당 값을 제외하는 것을 의미한다. $N_{kw}^{\neg id}$는 현재 단어 w가 k주제에 할당된 개수이다. (현재의 주제 할당 값을 제외하는 자세한 내용은 깁스 샘플링 알고리듬 참조) [5]. $N_{k.}^{\neg id}$는 k topic에 할당된 전체 단어 w의 개수를 의미하며, 이는 $N_{kw}^{\neg id}$ 값을 전체 w에 대해 합한 값이다. $N_{kd}^{\neg id}$는 현재 문서 d에 k주제가 나타난 횟수를 나타낸다. $N_{.d}^{\neg id}$는 d문서에 할당된 전체 주제의 개수를 의미하며, 이는 $N_{kd}^{\neg id}$를 모든 주제 k에 대해 합한 값이다[5 ,6]. 따라서 $\widehat{\Phi_{kw}}$는 각 단어 w가 갖는 k 번째 은닉 주제에 대한 확률을 나타내는 모델 파리미터 Φ_{kw}의 예측치 이고, $\widehat{\Theta_{kd}}$는 각 문서 d의 k 번째 은닉 주제에 대한 확률을 나타내는 모델 파라미터 θ_{kd}에 대한 예측치를 나타낸다. $\boldsymbol{\alpha}$와 $\boldsymbol{\beta}$는 각각 $\boldsymbol{\theta}$와 $\boldsymbol{\Phi}$ 파라미터의 선행 확률 (prior)로, Dirichlet 분포(Dirichlet Distribution) 이고, $\boldsymbol{\theta}$와 $\boldsymbol{\Phi}$ 는 앞서 설명된 것과 같이 Multinomial 분포(Multinomial Distribution)이다[1]. 일반적으로 $\boldsymbol{\alpha}$와 $\boldsymbol{\beta}$는 대칭(symmetric)하게 설정된 작은 값으로 실험 제시 되어 왔다 [1, 2, 6]. Wallah는 비대칭 하이퍼 파라미터 학습을 통해 marginal likelihood 성능 향상을 실험을 통해 입증 하였다 [3, 4].

본 논문에서는 사용자 시청이용 데이터에 대해 LDA를 기반으로 학습된 추천 모델을 제시하고, 사용자의 TV 프로그램에 대한 개인별 선호도를 잘 드러내도록 개인별 선호도를 나타내는 $\boldsymbol{\theta}$의 상위 확률 만 고려하고, 비대칭 하이퍼 파라미터를 추정할 경우 추천 성능 향상됨을 확인하였다. 비대칭 하이퍼 파라미터 추정 방법은 Wallah가 제시한 Fixed Point iteration방법을 이용하였다[3]. 다음 절에서는 추천 모델의 구성에 대해 구체적으로 설명한다.

3. 제안된 추천 모델

가. 실험 데이터

본 논문에서 사용한 실험 데이터는 2002년 12월 1일부터 2003년 5월 31일까지 6개월간의 2,000명의 사용자가 6개 지상파 채널을 시청한 AGB Neilson Korea의 사용자 시청 데이터를 이용하였다. 모델 학습을 위한 훈련기간을 4개월, 모델 검증을 위한 테스트 기간을 2개월로 나누어 실험을 진행하였다.

나. TV 시청 데이터를 이용한 주제 기반의 추천 모델링

TV시청 데이터는 사용자가 시청한 내역을 기록한 데이터로, TV 프로그램이 갖는 특성으로는 여러 회 차에 걸쳐 방영된 TV 프로그램들이 주를 이루고 있다. TV 시청 데이터 기반의 추천은 협업 필터링 개념을 적용하여 대중의 선호 프로그램 추천을 하더라도, 개인의 특성을 고려한 추천이 더욱 효과적임이 실험으로 입증된바 있다 [7]. 그러므로 본 논문은 개인별 TV 콘텐츠 주제에 대한 선호도와 전체 TV프로그램의 토픽에 대한 분포를 이용하여 순위정렬 모델을 구성하고, 각 개인의 훈련기간 4개월 동안의 시청 데이터를 기준으로 테스트 기간 2개월 동안의 추천과 이에 대한 사용자 개인별 Precision을 측정하고 실험 사용자들의 평균 Precision을 기준으로 추천 모델을 검증한다.

본 논문에서는 각 TV 사용자는 문서 d에 대응되고 각 사용자가 시청한 TV 프로그램은 단어 w에 대응되도록 하였다. 사용자가 해당 TV 프로그램을 시청한 횟수를 한 문서내의 해당 단어의 발생 횟수로 간주하였다. 전체 시청 사용자 중 샘플링 된 976명의 사용자 시청 데이터를 모델 파라미터를 예측하기 위해 사용하였다.

훈련 기간 사용자의 데이터를 기준으로 깁스 샘플링 알고리듬을 적용한 식 (1), (2)를 이용하였다. 모델의 파라미터를 학습하는 전체 순서도는 그림2와 같다.

그림 2의 모델 학습 순서도에 나타난 것과 같이, 일반적인 대칭 하이퍼 파라미터의 깁스 샘플링 알고리듬을 이용한 파라미터 추정 과정은 2단계 과정으로 끝난다. 본 논문에서는 2단계 과정을 거쳐 얻어진 은닉 주제 $\mathbf{z}$ 분포(각 문서 d 의 단어 w의 은닉 주제 결정 인덱스 값으로 표현되는 matrix)와 추정된 비대칭 하이퍼 파라미터 $\boldsymbol{\alpha}$를 기준으로 수식(1), (2)를 이용하여 3단계 샘플링을 추가 수행하였다. 샘플링 시 몬테까를로(Monte Carlo) 이론에 따라, 모델 파라미터 $\widehat{\Phi_{kw}}$와 $\widehat{\Theta_{kd}}$를 샘플링 횟수 S 만큼 더한 값을 샘플링 횟수로 정규화하는 식 (3)과 같은 방법으로 Φ_{kw}와 θ_{kd}값을 구한다.

$$p = \frac{1}{S}\sum_{s=1}^{S} \hat{p^s} \quad (3)$$

샘플링 각 단계별로 반복횟수는 그림 2 순서도에 기술된 것과 같이

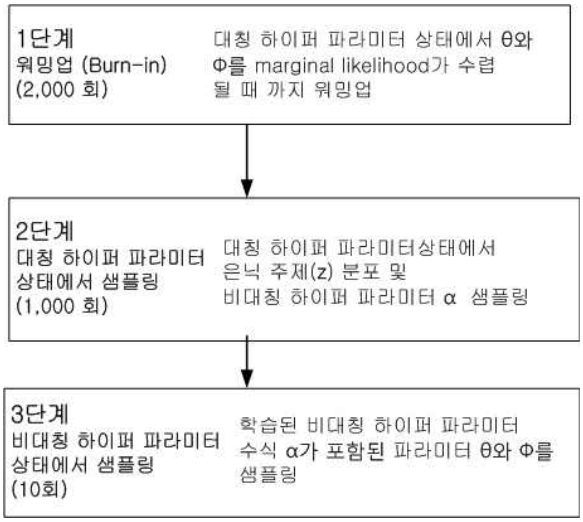

그림 2 모델 파라미터 학습 순서도

marginal likelihood값의 수렴 상태를 확인하며 수행되었다.

훈련 4개월 기간 추정된 파라미터 중 사용자 u의 각 은닉주제에 대한 선호도 θ_u를 u사용자 프로파일(사용자의 TV시청 선호도)이라 간주한다. Φ_p는 각 은닉 주제별 대중의 p프로그램에 대한 선호도를 나타낸다. 개별 사용자 u에게 추천하는 프로그램 p의 추천 순위는 식(4)과 같이 전체 주제에 대한 log marginal likelihood를 이용하여 해당 점수를 정하고 이를 기준으로 추천 후보 프로그램을 정렬한다.

$$Score_u^p \approx \sum_{k=1}^{K} \log\left(1 + \left(\theta_u \times \phi_p \times C\right)\right) \quad (4)$$

C는 적절한 상수 값을 의미한다. 본 실험에서는 100을 적용하였다.

4. 실험 성능 및 분석

가. 제시 추천 모델의 추천 성능 최적화

marginal likelihood를 최적화하는 방법으로 찾은 200개의 주제, 대칭 prior α=15, β=0.05 각각의 값을 시작으로, 각 파라미터를 학습하였다. 그림2의 2단계 SS(대칭 α, 대칭 β), 3단계 AS(비대칭 α, 대칭 β) 하이퍼 파라미터로 학습된 각각의 $\boldsymbol{\theta}$와 $\boldsymbol{\Phi}$값을 얻는다. 식(4)기반의 순위정렬 모델로 개별 사용자에게 추천된 TV프로그램을 검증기간 2개월 동안 해당 사용자의 실제 시청 TV프로그램과 비교하여 평균 Precision을 측정한 결과 표1과 같은 결과를 얻었다.

표1을 통해 확인할 수 있듯이, SS prior를 이용할 경우, 사용자 선호도를 특정 개수로 한정지어 추천하여 추천 성능의 크게 향상 되었다. 그러나 이 성능은 AS prior로 학습하여 추천한 성능에 미치지 못한다. 또한, AS prior로 학습한 개별 u 사용자 선호도 θ_u중 상위 40개의 주제에 대한 확률만 고려하는 것이 모든 주제에 대한 확률을 고려하는 것보다 더 나은 추천 성능을 보임을 확인할 수 있다.

표 1 Prior학습 방식과 θ_u 고찰에 따른 추천 성능 변화

평균 Precision					
추천개수	5	10	20	30	40
대칭 하이퍼 파라미터 (SS)					
	0.573	0.570	0.547	0.505	0.488
대칭 하이퍼 파라미터 (SS) (θ_u의 상위 40개의 확률만 고려할 경우)					
	0.772	0.722	0.684	0.655	0.624
비대칭 α, 대칭 β 하이퍼 파라미터 (AS)					
	0.732	0.718	0.689	0.663	0.640
비대칭 α, 대칭 β 하이퍼 파라미터 (AS) (θ_u의 상위 40개의 확률만 고려할 경우)					
	0.779	0.742	0.703	0.670	0.639

나. 은닉 주제의 분석

각 주제 k별 Φ_k내 상위 확률의 TV프로그램들을 정렬하여 주제들의 구성을 분석해 보았다. 각 주제 내 일관성을 장르, 채널, 요일, 방영시간대 (하루 24시간을 4시간씩 6단위로 나눔), 프로그램제목, 하위장르로 살펴보았다. 어느 정도 일관성 있는 TV프로그램들이 모여 있다. 그림 3의 예시처럼 17번 주제는 어린이(유아)장르의 만화 인형극들이 주를 이루고 있고, 0 주제는 보도, 정보 및 오락 TV 프로그램들이 있다. 그런데, 각 주제를 하나의 장르로 구분 지을 수 있는 형태가 아니다. 보통 TV는 개인시청보다는 가족단위의 시청이 이뤄지므로, 하나의 연령대 TV프로그램 보다는 어린이, 유아와 어른이 시청하는 프로그램이 한 주제 내에 공존함을 확인할 수 있다.

그림 3 주제별 상위 확률 TV 프로그램 예시

5. 결론

TV프로그램 시청 데이터를 기반으로 사용자에게 TV프로그램 시청을 돕는 추천 순위모델을 LDA기반의 모델로 정의하고, 추천 성능 향상을 위해 사용자 선도도가 사용자별로 차이가 드러나도록 prior를 AS학습한 결과를 적용하여 의미 있는 추천 성능의 결과를 확인하였으며, 제시 모델로 드러난 주제들이 사용자의 TV시청 이용행태를 드러내는 것을 확인할 수 있었다.

참 고 문 헌

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

[2] A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh. On smoothing and inference for topic models. In Proceedings of the International Conference on Uncertainty in Artificial Intelligence, 2009..

[3] H. M. Wallach. Structured Topic Models for Language. Ph.D. thesis, University of Cambridge, 2008

[4] H. M. Wallach, D. Mimno, and A. McCallum. Rethinking lda: Why priors matter. In Advances in Neural Information Processing Systems 22, pages 1973–1981, 2009.

[5] C.M. Bishop. Pattern recognition and machine learning. ch 11, Springer, 2006..

[6] Tomas L. Griffiths, Mark Steyvers, "Finding scientific topics," Proceedings of the National Academy of Sciences, vol. 101, suppl. 1, pp. 5228-5235, April, 2006.

[7] Eunhui kim, Shinjee Pyo, Eunkyung Park and Munchurl Kim, "An automatic TV Recommendation for (IP)TV Personalization," IEEE Transactions on Broadcasting, vol. 57, no.3, pp.674-684, Sept, 2011.