Bi-Source 토픽 모델 기법을 이용한 기사-상품 연관 검색1)

김병희○ 이바도 하성종 조남익 장병탁
서울대학교 전기컴퓨터공학부
{bhkim○, bdlee}@bi.snu.ac.kr, oanchovy@ispl.snu.ac.kr, nicho@snu.ac.kr, btzhang@bi.snu.ac.kr

Bi-Source Topic Modeling Method and Its Application to Article-Goods Associative Search

Byoung-Hee Kim○[1], Bado Lee[1], Seong Jong Ha[2], Namik Cho[2], Byoung-Tak Zhang[1*]
[1]School of Computer Sci. and Eng., Seoul National University
[2]Department of Electrical Engineering, Seoul National University

1. 서 론

디지털 컨버전스가 진행됨에 따라 글, 그림 기반의 전통적인 온라인 콘텐츠뿐만 아니라 동영상, 사용자의 선호도 및 사용 이력 등 연관성이 큰 다양한 모달리티가 혼재된 형태로 데이터가 쏟아져 나오고 있다. 사용자 중심의 검색 및 추천 서비스를 위해서는 이러한 멀티모달 데이터에서의 정보 추출 및 연관성 분석 기법이 필수적이다.

본 논문에서는 멀티모달 데이터의 연관성 분석의 한 경우로서, 다양한 출처에서 생성된 관측 데이터를 기반으로 출처 간의 의미적 연관성을 학습하는 문제를 다룬다. 출처가 다른 데이터 간의 연관성을, 데이터의 특성값 공간(feature space)을 단일화하여 특성값 수준에서 표현하기보다는, 출처별 특성을 반영한 개별 특성값 집합을 정의한 후 두 출처 공통의 맥락을 표현하는 은닉 변수를 두어, 은닉 변수 단계에서 연관성을 추출하고자 한다. 이러한 목적으로, LDA (latent Dirichlet allocation) [1] 모델을 기반으로 한 이중 출처(bi-source) 토픽 모델 기법(BSTM)을 제시한다.

BSTM 모델을 이용하면 두 출처에서 얻은 데이터 간의 연관성을 토픽 분포의 유사도를 기반으로 계량화할 수 있다. BSTM을 이용한 기사-상품 연관 검색 및 추천은 기사와 상품을 표현하는 특성값(feature)을 구분하여 정의하고 공통의 토픽을 추출하여 기사 대비 유사 토픽 분포를 가지는 상품을 선별하는 절차를 거치게 된다(그림 1). 모델의 기본적 적용 사례로서, 사진 정보만을 이용한 기사-상품 연관 검색에의 적용 예를 보인다. 온라인상에서 젊은 여성층을 대상으로 서비스 중인 잡지 기사와 쇼핑몰 상품을 선별하고, 두 출처의 사진 정보만을 이용한 연관 검색 테스트 결과 상품의 카테고리 적중률 최대 60% 및 다양한 잠재적 연관성을 표현함을 확인할 수 있었다.

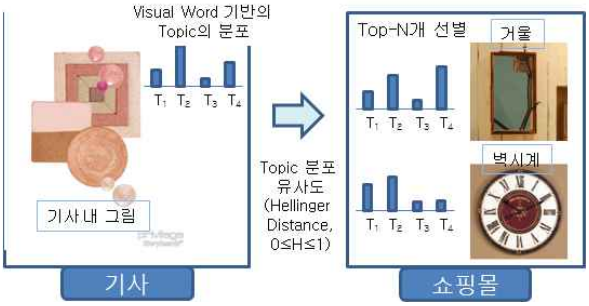

그림 1. BSTM 기반의 기사-상품 연관 검색 개념도

2. 본 론

두 출처(bi-source)에서 수집된 데이터 간의 연관성 학습을 위해 출처별 특성값을 구분하여 정의하고, 두 출처간 공통의 컨텍스트를 표현하는 토픽을 가정하는 경우 [그림 3]과 같은 BSTM 모델을 설정할 수 있다. LDA에서와 마찬가지로 토픽을 표현하는 은닉변수가 각 특성값의 bag-of-words 형태의 이산화된 표현이 가능한 경우 BSTM 모델 하에서의 데이터 집합(데이터 개수 D, 두 출처의 전체 특성값(단어) 수 N, 토픽의 수

BSTM에서의 생성 절차

1. 토픽 혼합율 추출 $\theta \sim Dirichlet(\alpha)$
2. 토픽 배정율 추출 $z_n|\theta \sim Multinomial(\theta)$
3. 출처 A의 단어 추출 $I_{An}|z_n \sim Multinomial(\pi_A)$
4. 출처 P의 단어 추출 $I_{Pn}|z_n \sim Multinomial(\pi_P)$

그림 2. BSTM에서의 생성 절차

1) 본 연구는 지식경제부 및 한국산업기술평가관리원의 IT산업원천기술개발사업의 일환으로 수행하였으며 (KI002138, MARS), 교육과학기술부의 재원으로 국가연구재단의 지원을 받아 수행된 연구(314-2008-1-D00377, Xtran) 및 교육과학기술부의 BK21-IT사업에 의해 일부 지원되었음.

K)의 생성 절차는 [그림 2와] 같다.

BSTM에서는 [2]에서 제시한 바와 같이 사후확률 $p(z|I_A, I_P)$를 계산하기 위해 깁스 샘플링(Gibbs sampling) 기법을 기반으로 다음의 확률 계산 과정을 반복한다:

$$p(z_i = j|\mathbf{z}_{-i}, \mathbf{I_A}, \mathbf{I_P}) \propto \frac{n_{-i,j}^{(i_A)} + \pi_A}{n_{-i,j}^{(\cdot)} + N\pi_A} \frac{n_{-i,j}^{(i_P)} + \pi_P}{n_{-i,j}^{(\cdot)} + N\pi_P} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + K\alpha}.$$

$n_j^{(w)}$는 단어 w가 토픽 배정 벡터 $\mathbf{z}$의 j번째 토픽에 할당된 회수이고, $n_j^{(d)}$는 문서 d에서 단어가 j번째 토픽에 할당된 회수이다. $n_{-i}^{(\cdot)}$는 현재 토픽 z_i에 할당되지 않은 단어의 수이다.

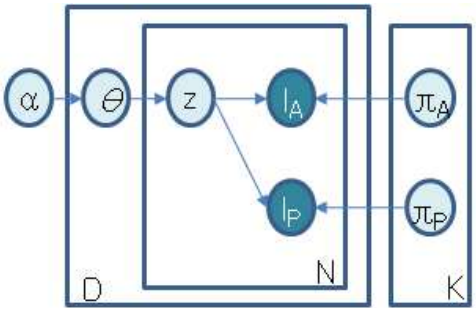

그림 3. LDA를 확장한 Bi-Source Topic Model (BSTM) 기법의 확률 그래프 모델 표현

학습된 BSTM모델에서 θ는 각 예제를 토픽의 분포로 표현한다. 모델에서 K개의 토픽이 결정되면, 두 예제 간의 유사도의 기준으로, 각 이미지의 토픽 분포 간의 거리를 Hellinger distance[3]를 이용하여 다음과 같이 계량화한다:

$$d(\theta_i, \theta_j) = H(\theta_i, \theta_j) = \frac{1}{2} \sum_{k \in K} \left(\sqrt{\theta_i(k)} - \sqrt{\theta_j(k)} \right)^2.$$

그림 데이터에 BSTM 모델을 적용하기 위해 각 그림을 시각단어의 집합으로 표현한다. 본 논문에서는 1) SURF 기반 특성값 추출, 2) 평균이동-클러스터링을 통한 시각단어 정의, 3) SVDD를 이용한 특성값별 시각단어 할당의 과정을 거쳐 그림 데이터를 Bag-of-Visual-Words로 표현하였다.

BSTM 모델을 이용한 기사-상품 연관 검색 성능을 살펴보기 위해 젊은 여성층을 대상으로 한 잡지와 쇼핑몰 상품을 선별하였다. 잡지로는 스토리서치™에서 온라인으로도 제공되고 있는 뷰티라이프, 마이웨딩, 프리빌리지, 세 종류의 2007~2008년도 컨텐츠를, 쇼핑몰 상품 사진은 스토리샵™ 에서 판매 중인 리빙/주방/육아 관련 37개 카테고리의 상품을 선별하였다. 수집한 잡지기사 사진은 4,816개, 쇼핑몰 상품 사진은 5,375개이며, 이 사진 데이터를 기반으로 1,545개의 시각단어(visual word)를 선별하였다. BSTM 모델 학습을 위한 깁스 샘플링 기반 추론 및 사진 간의 유사도 측정은 토픽 모델링 툴박스[4]를 수정하여 프로그램으로 구현하였다. 세 하이퍼파라미터의 값은 α=1.0, π_A=π_P=0.01로 설정하였다.

성능 평가를 위한 데이터로 5개 카테고리(귀걸이, 벽시계, 수저포크나이프, 쿠션대쿠션, 플레이트접시)를 선별하고 각 카테고리 별로 50개씩, 250개의 상품의 사진을 지정하였다. 평가데이터 내에서 각 그림을 질의로 제시하고 Hellinger distance 기준 상위 N개 상품 중 동일 카테고리 상품의 포함 비율이 N/2 이상이면 '성공'으로 판별한다(그림 4). 평가 결과 60% 전후의 성공률을 얻었다. 또한, 학습 데이터 내에서 기사 사진을 질의로 연관 상품의 사진을 검색한 결과 질의와 검색 결과의 다양한 의미적 연상 관계가 추출되는 사례를 다수 확인하였으며, 이는 다양한 사용자의 선호를 반영할 수 있다는 점에서 고무적인 결과로 보인다.

그림 4. 연관검색 성능 평가용 데이터 내에서의 질의용 사진 및 연관검색 결과의 예

3. 결 론

본 논문에서는 멀티모달 데이터 간의 연관성 분석 및 검색을 위한 모델로서 Bi-Source 토픽 모델(BSTM) 기법을 제안하고, 기사-상품 연관 검색에의 응용 사례를 보였다. 한국어 잡지 기사 사진 및 쇼핑몰 상품의 사진 데이터에 BSTM을 적용하여, 기사와 상품 간 공통의 컨텍스트를 표현하는 토픽을 추출하고 두 출처에서 얻은 사진 간의 연관관계를 토픽 분포의 유사도를 기반으로 계량화할 수 있음을 보였다. 상품 사진 내의 연관 검색 결과 최대 60% 이상의 적중률을 얻을 수 있었으며, 기사 사진을 질의로 한 관련 상품 검색 결과 다양한 잠재적 연관성을 추출할 수 있음을 확인하였다.

향후 연구 과제로 토픽 간의 연관성까지 고려하는 correlated LDA [3] 기반의 신규 모델을 개발하고, 상품의 카테고리 정보 또는 기사의 텍스트 컨텐츠를 반영하도록 모델을 확장하고자 한다.

참고문헌

[1] D. Blei, A. Ng, and M. I. Jordan, Latent Dirichlet allocation, *J. Mach. Learn. Res.*, **3**:993-1022, 2003.
[2] T. L. Griffiths and M. Steyvers, Finding scientific topics. *P. NATL. ACAD. SCI. USA,* 101:5228-5235, 2004.
[3] D. Blei and J. Lafferty, A correlated topic model of Science, *Annals of Applied Statistics,* 1:1 17-35, 2007.
[4] M. Steyvers & T. Griffiths, Matlab Topic Modeling Toolbox 1.3.2, http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm