텍스트 문서의 주제어 추출을 위한 확률적 그래프 모델의 학습

신형주[0] 장병탁 김영택
서 울 대 학 교 컴 퓨 터 공 학 부
hjshin@scai.snu.ac.kr
{btzhang,ytkim}@cse.snu.ac.kr

Learning Probabilistic Graph Models for Extracting Topic Words in a Collection of Text Documents

Hyung-Joo Shin[0] Byoung-Tak Zhang Yung Tak Kim
School of Computer Science and Engineering, Seoul National University

요 약

본 논문에서는 텍스트 문서의 주제어를 추출하고 문서를 주제별로 분류하기 위해 확률적 그래프 모델을 사용하는 방법을 제안하였다. 텍스트 문서 데이터를 문서와 단어의 쌍으로(dyadic)표현하여 확률적 생성 모델을 학습하였다. 확률적 그래프 모델의 학습에는 정의된 likelihood를 최대화하기 위한 EM(Expected Maximization) 알고리즘을 사용하였다. TREC-8 Ad Hoc 텍스트 데이터에 대하여 학습된 확률 그래프 모델의 성능을 실험적으로 평가하였다. 이로부터 찾아 낸 문서에 대한 주제어가 사람이 제시한 주제어와 유사한 지와, 사람이 각 주제에 대해 분류한 문서가 이 확률모델로부터의 분류와 유사한 지를 실험적으로 검토하였다.

1. 서론

텍스트 문서에서 사람이 주제어를 찾아내고 문서를 주제별로 분류하는 데에는 일정한 규칙이 있을 것이다. 즉, 사람이 문서의 주제를 파악하고 문서를 주제별로 분류할 때는 문서와 단어와의 관계에 대한 미리 학습되어 있는 모델을 사용할 것이다. 그러므로 방대한 양의 문서의 주제를 사람이 분류하는 것처럼 합당하게 자동으로 파악하고 분류하기 위해서는 같은 주제에 포함되는 문서들이 가지는 모델의 특징을 찾아내는 것이 중요하다.

문서의 숨은 의미를 파악하여 문서 색인(indexing)에 적용하거나[4][5], 문서의 주제를 파악하고(learning topic) 분류(clustering)하는 것[7], 또는 문서의 차원을 줄이는 것[6]에 관한 연구가 최근에 많이 이루어지고 있다. 이 중에서[5][7]에서 사용하는 모델은 latent variable로 표현되는 모델의 mixture이다.

텍스트 문서와 같이 sparse한 특징을 가지는 데이터로부터 모델을 학습하기 위해서 generative model을 사용하는 연구 또한 활발히 이루어지고 있다[3][8].

본 논문에서는 이러한 연구들에 기초하여 Aspect Model로 문서 데이터를 정의하고 EM으로 학습[3]한 결과가 사람이 문서에 대해 파악한 주제어, 그리고 주제별 분류와 유사하다는 것을 보였다.

기본적으로 문서를 문서와 단어 간의 co-occurrence로 정의한다. 이 데이터와 latent variable에 대한 통계적 혼합 모델(statistical mixture models), 즉 aspect model [3]을 기반으로 EM(Expectation Maximization)을 이용하여 문서와 단어의 확률분포를 학습(fitting)하였다. 모델은 단어, latent variable, 문서의 세 층(layer)으로 구성된다. 이 모델에 대한 가정은 문서와 단어는 어떤 latent variable로부터 확률적으로 생성(generate)된다는 것이다. 이 때 latent variable이 결국 문서의 주제라고 가정할 때 각 주제를 표현하는 주제어의 집합과, 또 각 Topic에 속하는 문서의 집합이 사람이 하는 작업과 얼마나 유사한지를 보인다.

2. 문서의 숨은 의미 추출을 위한 확률적 생성 모델의 학습

Aspect model은 알려지지 않은 class variable $z_k \in \mathbf{Z}=\{z_1,\cdots.,z_K\}$로부터 단어 $w_n \in \mathbf{W}=\{w_1,\cdots.,w_M\}$와 문서 $d_n \in \mathbf{D}=\{d_1,\cdots.,d_N\}$가 확률적으로 생성된다는 가정 하에 제시된 모델이다[3][5]. 이 때 사용하는 데이터는 각 문서 d_n 에 나타난 단어 w_m 의 빈도수 (d_n, w_m), $n=1,..,N$, $m=1,..,M$ 이다. 이것은 generative model로서 다음과 같이 정의된다[3].

$$P(d_n,w_m)=P(d_n)P(w_m\mid d_n), \tag{1}$$

$$P(w_m\mid d_n)=\sum_{k=1}^{K}P(w_k\mid d_n)P(z_k\mid d_n) \tag{2}$$

(1), (2)에 Bayes' rule을 적용하면 다음과 같다.

$$P(d_n,w_m)=\sum_{k=1}^{K}P(z_k)P(w_m\mid z_k)P(d_n\mid z_k) \tag{3}$$

이 때 기본적인 가정은 각각의 (d_n,w_m)가 iid (identically independently distributed)이며 단어 w는 latent variable에만 의존하고 문서 d에는 독립적으로 생성된다는 것이다. 이 모델에서는 다음과 같은 log-likelihood를 최대화하는 데에 일반적으로 EM알고리즘을 사용한다[1][2][5].

$$L=\sum_{n=1}^{N}\sum_{m=1}^{M}n(d_n,w_m)\log P(d_n,w_m) \tag{4}$$

여기서 $n(d_n, w_m)$는 문서 d_n 에 나타난 단어 w_m 의 빈도수를 나타낸다.

(4)을 국지적으로 최대화(local maximum)하는 EM알고리즘의 E-step은 다음과 같다.

$$P(z_k \mid d_n, w_m) = \frac{P(z_k)P(d_n \mid z_k)P(w_m \mid z_k)}{\sum_{k=1}^{K} P(z_k)P(d_n \mid z_k)P(w_m \mid z_k)} \quad (5)$$

M-step은 다음과 같다.

$$P(w_m \mid z_k) = \frac{\sum_{n=1}^{N} n(d_n, w_m)P(z_k \mid d_n, w_m)}{\sum_{m=1}^{M}\sum_{n=1}^{N} n(d_n, w_m)P(z_k \mid d_n, w_m)} \quad (6)$$

$$P(d_n \mid z_k) = \frac{\sum_{m=1}^{M} n(d_n, w_m)P(z_k \mid d_n, w_m)}{\sum_{m=1}^{M}\sum_{n=1}^{N} n(d_n, w_m)P(z_k \mid d_n, w_m)} \quad (7)$$

$$、 P(z_k) = \frac{1}{R}\sum_{m=1}^{M}\sum_{n=1}^{N} n(d_n, w_m)P(z_k \mid d_n, w_m), \quad (8)$$

$$R \equiv \sum_{m=1}^{M}\sum_{n=1}^{N} n(d_n, w_m)$$

위에서 설명한 Aspect model을 그림으로 표현하면 그림 1과 같다. $P(w_m|z_k)$는 latent variable z_k 를 소주제로 포함하는 문서들에 단어 w_m가 나타날 확률로 해석될 수 있다. 마찬가지로 $P(d_n|z_k)$는 문서 d_n이 소주제로 z_k를 포함할 확률로 해석될 수 있다. 이 모델과 가정이 사람이 문서를 분류하는 기준과 유사하다면 모델의 학습이 끝난 후에 다음과 같은 결과가 나올 것이다. 첫째, K=L, 즉 latent variable의 개수를 주어진 topic의 개수와 같게 한다면 각 latent variable z_k(k=1,…,K)에 대해 $P(w_m|z_k)$가 큰 단어 w의 집합은 사람이 topic c_l (l=1,…,L)을 표현할 때 사용한 단어의 집합과 유사할 것이다. 둘째, 각 주제 c_l에 대해 $I(d_n \in c_l)$=1인 문서, 즉 c_l에 속한 문서들은 c_l에 속하지 않은 문서들보다 $P(d_n|z_k)$값이 클 것이다.

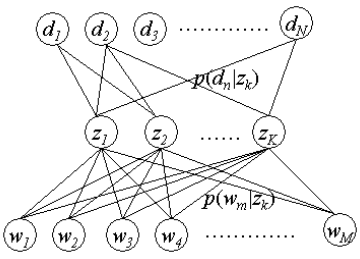

그림 1. Aspect Model

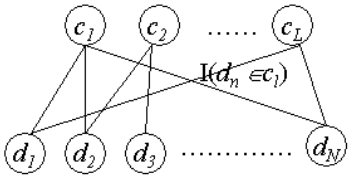

그림 2. 문서의 주어진 주제별 분류

3. 실험 및 결과

실험에는 TREC-8 [1]의 adhoc task에 사용되는 데이터를 사용했다. 여기에는 DTDS, FR94, FT, FBIS, LATIMES의 문서와 50개의 topics 401-450가 있다. 각 topic에 대해 relevant한 문서가 (사람에 의해)정해져 있는데 이 중 relevant한 문서의 수가 많은 434(347개), topic 401(300개)에 relevant한 문서 중 FBIS, FT, LATIMES에 속한 문서를 사용하였다.(여기에 속한 문서 중 문서의 크기가 6KB이상, 16KB이하인 169개의 문서를 사용하였다.) 이를 Porter 알고리즘으로 stemming하고 524개의 stopwords를 제거하여 7826개의 단어를 뽑아냈다. 7826개의 단어 중에 frequency의 합이 5번 이하이거나 1개 이하의 문서에만 나오거나 7000개 이상의 문서에 나오는 5610개의 단어를 제외한 2216개의 단어를 사용하였다. 169개의 문서와 2216개의 단어에 대해 각 문서에 나타난 단어의 빈도수를 값으로 하는 문서-단어 행렬을 만들었다. [2]

TREC-8 adhoc데이터에서 topic 434, 401이 그림 3과 같다.

Topic 434
Title: *Estonia*, *economy*
Description: What is the *state* of the economy of Estonia?
Relevant Document: **Documents that give concrete *economic* information such as economic *statistics*, entering economic *unions* and *treaties*, or *monetary* performance are relevant, as are discussions of economic issues such as *transportation* or *pollution*.**
Topic 401
Title: foreign *minorities*, *Germany*
Description: What *language* and *cultural* *differences* impede the *integration* of foreign minorities in Germany?
Relevant Document: A relevant document will focus on the causes of the lack of integration in a significant way; that is, the mere mention of *immigration* *difficulties* is not relevant. Documents that discuss immigration problems unrelated to Germany are also not relevant.

그림 3. Topic 434, 401의 Definition

(밑줄은 stopwords를 제거하고 stemming한 후 뽑아 낸 단어임. 밑줄 친 단어 중 이탤릭 체로 된 것은 Topic의 특징을 정확히 표현하는 단어임)

Latent variable을 topic의 수와 같은 2개(K=L=2)로 하여 EM알고리즘을 100 iteration만큼 돌려 (6), (7), (8)을 계산해 냈다. 표 1은 k=0,1에 대해 $P(w_m|z_k)$가 큰 단어 10개를 $P(w_m|z_k)$의 내림차순으로 정렬한 것이다. 표 2는 각 topic에 중요하다고 여겨지는 (그림 3에서 이텔릭체로 되어 있고 밑줄 친 단어) 단어에 대해 $P(w_m|z_k)$ 의 값이다. 표 3은 각 topic에 주어진 단어의 집합에 대해, 그 외의 모든 단어에 대해 $P(w_m|z_k)$의 최대, 평균, 최소값이다.

표 1과 그림 3을 비교해 보면 확률적으로 찾아 낸 두 개의 주제에 대한 주제어가 사람이 제시한 Topic Definition과 매우 유사함을 볼 수 있다. 또한 표 2, 3에서 볼 수 있듯이 Topic Definition에 사용되는 중요한 단어들에 대해서 해당 Topic에 대한 단어의 확률이 해당하지 않는 Topic에 대한 확률보다 높다.

표 4 (a)는 k=0,1에 대해 $P(d_n|z_k)$가 큰 문서 상위 80개, 89개 중 Topic 401, Topic434에 속한 개수로 이 방법이 문서의 clustering에 응용될 수 있음을 보이고 있다. 즉, N개의 문서를 K'개로 clustering할 때 latent variable의 수 K=K'로 두고 위의 알고리즘 알고리즘을 수행하여 각 k'(k'=1,…,K')에 대해 일정 개수만큼 $P(d_n|z_{k'})$이 큰 문서를 cluster k'에 할당하면 된다. 표 4 (b)는 Topic401에 속한 문서 80개 중 $P(d_n|z_1) > P(d_n|z_0)$인 문서의 개수와 Topic434에 속한 문서 89개 중 $P(d_n|z_0) > P(d_n|z_1)$인 문서의 개수로 이 방법이 문서의 classification에 응용될 수 있음을 보이고 있다.

표 5는 각 topic에 relevant한 문서의 집합에 대해, 그 외 모든 문서에 대해, 그리고 모든 문서에 대해 $P(d_n|z_k)$의 최대, 최소, 평균값이다. 이는 각 Topic에 relevant한 문서들이 relevant하지 않은 문서들보다 Topic에 대한 확률이 더 높다는 것을 보여준다.

---

[1] Text Retrieval Conference 8, http://trec.nist.gov
[2] 이를 위해 McCallum 등이 만든 *Bow*library 를 이용하였다. 이는 http://www.cs.cmu.edu/~mccallum/bow 에서 구할 수 있다.

k=0	k=1
estonia	germani
percent	immigr
state	integ
russian	minor
estonian	union
bank	cultur
russia	foreign
baltic	asian
econom	language
invest	unity

표 1. k=0,1에 대해 $P(w_m|z_k)$가 큰 단어 10개를 $P(w_m|z_k)$의 내림차순으로 정렬한 것

	k=0	k=1
minor	0	0.0010
germani	0.0010	0.0170
language	0.0010	0.0010
cultur	0	0.0010
differ	0.0010	0.0010
integ	0.0010	0.0010
immigr	0	0.0050
estonia	0.0160	0
economi	0.0050	0
state	0.0120	0.0050
econom	0.0050	0
statist	0.0010	0
union	0.0030	0.0030
treati	0	0.0010
monetary	0.0020	0
tranport	0.0010	0
pollution	0.0010	0

표 2. 각 topic에 중요하다고 여겨지는 단어에 대해 $P(w_m|z_k)$ (이탤릭 체는 Topic401의 topic definition에 해당하는 단어이다.)

	Topic 434	*Topic 401*
최대	0.0160/0.0160	0.0050/0.0050
평균	0.0029/0.0004	0.0021/0.0003
최소	0/0	0/0

표 3. 각 topic에 주어진 단어의 집합에 대해, 모든 단어의 집합에 대해 $P(w_m|z_k)$ 의 최대, 평균, 최소값

	Topic434	Topic 401t
K=0	89 개	0 개
K=1	4 개	76 개

(a)

	$P(d_n\|z_0) > P(d_n\|z_1)$	$P(d_n\|z_1) > P(d_n\|z_0)$
Topic434	87 개	2 개
Topic 401	3 개	77 개

(b)

표 4. (a) k=0,1에 대해 $P(d_n|z_k)$가 큰 문서 상위 80개, 89개 중 Topic 401, Topic434에 속한 개수로 이 방법이 문서의 clustering에 응용될 수 있음을 보인다. (b) Topic401에 속한 문서 80개 중 $P(d_n|z_1) > P(d_n|z_0)$인 문서의 개수와 Topic434에 속한 문서 89개 중 $P(d_n|z_0) > P(d_n|z_1)$인 문서의 개수로 이 방법이 문서의 classification에 응용될 수 있음을 보인다.

4. 결론

서론에서 이야기 한 것처럼 사람이 문서의 주제를 파악하고 문서를 주제별로 분류할 때는 문서와 단어와의 관계에 대한 미리 학습되어 있는 모델을 사용할 것이다.

	Topic434 relevant	Topic 434 not relevant	Topic 434에 속한 모든 문서
최대	0.0240	0.0190	0.0240
평균	0.0107	0.0031	0.0085
최소	0.0030	0.0000	0.0000

	Topic401 relevant	Topic 401 not relevant	Topic 401에 속한 모든 문서
최대	0.0260	0.0120	0.0260
평균	0.0085	0.0012	0.0059
최소	0.0000	0.0000	0.0000

표 5. 각 topic에 relevant한 문서의 집합에 대해, 그 외 모든 문서에 대해, 그리고 모든 문서에 대해 $P(d_n|z_k)$의 최대, 최소, 평균값

실험결과를 통해 확률적 그래프 모델로 문서의 주제어($P(w_m|z_k)$)와 분류($P(d_n|z_k)$)를 학습한 결과가 사람이 문서에 대해 파악한 주제어, 그리고 주제별 분류와 유사하다는 것을 알 수 있다.

Latent variable을 분류(classification)의 한 class, 또는 clustering의 prototype으로 볼 수 있는지의 여부는 아직 논쟁의 여부가 남아 있다. 하지만 본 논문을 통해 텍스트 문서 데이터에 대해 latent variable이 문서의 주제를 나타내고, 각 latent variable에 대한 $P(w_m|z_k)$가 주제어가 될 수 있으며 $P(d_n|z_k)$값을 가지고 문서를 분류할 수 있음을 알 수 있다. 이는 문서분류 뿐 아니라 문서의 clustering, 그리고 어떤 문서의 집합에서 문서들의 주제어를 찾아내는 데에도 이용될 수 있다.

하지만 이 논문의 실험은 Topic을 2개밖에 사용하지 않았으며 사용한 단어와 문서의 수도 현실에서 접하는 방대한 양의 문서에 비하면 매우 작다. 그러므로 많은 양의 텍스트 문서에 대한 검증이 필요하다.

감사의 글

본 연구는 정보통신부가 시행하고 있는 대학 기초연구기술 지원사업(98-199)에 의해 일부 지원되었음.

참고 문헌

[1] Dempster, A.P., N.M.Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm (with discussion)", *J. Roy. Stat. Soc.,* B39, 1-38, 1977.

[2] Jeff A. Bilmes, "A Gentle Tutorial of EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models", International Computer Science Institute, 1998.

[3] Thomas Hofmann, Jan Puzicha, "Unsupervised Learning from Dyadic Data", in *Advances in Neural Information Processing System,* 1998

[4] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R., "Indexing by latent semantic analysis", *Journal of the American Society for Information Science,* 1990.

[5] Thomas Hoffmann, "Probabilistic Latent Semantic Indexing", in *Proc. of the 22nd International Conference on Research and Development in Information Retrieval (SIGIR'99),* 1999.

[6] Yiming, Yang., "Noise Reduction in a Statistical Approach to Text Categorization,", in *Proceedings of the 18th Annual international ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'95),* pp. 256-263, 1995

[7] Thomas Hoffmann, "Learning and Representing Topic. A Hierarchcal Mixture Model for Word Occurrences in Document Databases", in *Conferences for Automated Learning and Discovery,* 1998.

[8] Hinton, G. E. and Ghahramani, Z., "Generative Models for Discovering Sparse Distributed Representations. *Phil. Trans. Roy. Soc. London B,* 352:1177-1190, 1997.