점진적 EM 알고리즘에 의한 잠재토픽모델의 학습 속도 향상

장정호* 이종우** 엄재홍**O
*Fraunhofer IAIS.KD **서울대학교 컴퓨터공학부
jeongho.chang@gmail.com jwlee@bi.snu.ac.kr jheom@bi.snu.ac.kr

Accelerated Learning of Latent Topic Models by Incremental EM Algorithm

Jeong-Ho Chang* Jong-Woo Lee** Jae-Hong Eom**O
*Fraunhofer IAIS.KD **School of Computer Science & Engineering, Seoul National University

잠재토픽모델(latent topic model)은 데이터에 내재된 특징적 패턴이나 데이터 정의 자질들간의 상호 관련성을 확률적으로 모델링하고 자동 추출하는 모델로서 텍스트 문서로부터의 의미 자질 자동 추출을 비롯하여 최근 이미지를 비롯한 멀티미디어 데이터 분석, 생물정보학 분야 등에서 많이 응용되고 있다. 이러한 잠재토픽모델의 대규모 데이터에 대한 적용시 그 효과 증대를 위한 중요한 이슈 중의 하나는 모델의 효율적 학습에 관한 것이다. 본 논문에서는 대표적 잠재토픽모델 중의 하나인 PLSA 모델[1]을 대상으로 점진적 EM 알고리즘을 활용한, 기본 EM 알고리즘 기반의 기존 학습에 대한 학습속도 증진 기법을 제안한다. 점진적 EM 알고리즘은 토픽 추론시 전체 데이터에 대한 일괄적 E-step 대신에 일부 데이터에 대한 일련의 부분적 E-step을 수행하는 특징이 있으며 이전 데이터 일부에 대한 학습 결과를 바로 다음 데이터 학습에 반영함으로써 모델 학습의 가속화를 기대할 수 있을 뿐 아니라 이론적인 측면에서 지역해로의 수렴성이 보장되고 기존 알고리즘의 큰 수정없이 구현이 용이하다는 장점이 있다[2].

PLSA 모델은 공기(co-occurrence) 데이터나 히스토그램 데이터에 대해 효과적인 확률적 잠재변수모델 기반의 분석 방법론으로서 특히 언어모델링이나 정보검색 등의 텍스트 문서 관련 응용에서 유용하게 응용되고 있다. 하나의 텍스트 문서 $d = (w_1, w_2, \ldots w_{N_d})$에 대해 PLSA의 기본적인 모델링은 가상의 잠재토픽을 도입하여 $p(w_n|d) = \sum_{k=1}^{K} p(w_n|z_k)p(z_k|d)$에 이루어지며, 모델의 학습은 다음과 같은 EM 알고리즘에 의해 달성된다.

$$E-Step: \quad p(z_k|d_m, w_n) = \frac{p(w_n|z_k)p(z_k|d_m)}{\sum_{k=1}^{K} p(w_n|z_k)p(z_k|d_m)} \qquad M-Step: \quad p(z_k|d_m) \propto \sum_{n=1}^{N_d} n(d_m, w_n)p(z_k|d_m, w_n)$$

$$p(w_n|z_k) \propto \sum_{m=1}^{M} n(d_m, w_n)p(z_k|d_m, w_n)$$

이와 같은 PLSA 모델의 기본적인 학습 알고리즘을 점진적 EM 알고리즘으로 대체하여 모델의 학습시 수렴 속도를 향상시킬 수 있다. 그림 1은 PLSA 모델에 대한 점진적 EM 알고리즘을 기본 EM 알고리즘과 대비해 요약적으로 제시한다.

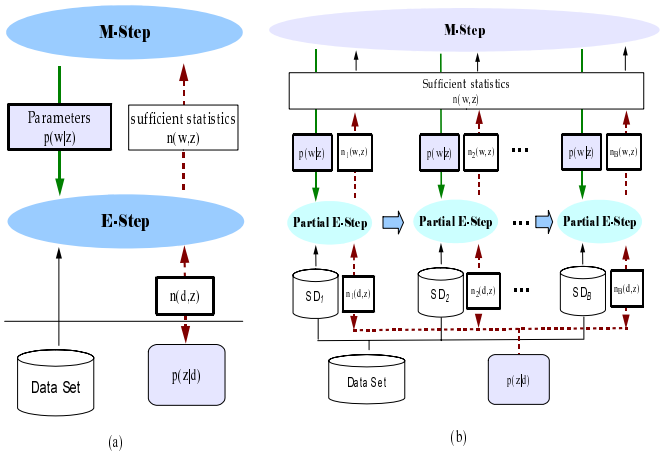

그림 1. PLSA 모델 학습: (a) 기본 EM 알고리즘 (b) 점진적 EM 알고리즘

그림에서와 같이 점진적 EM 알고리즘은 데이터를 먼저 분할하고 각 데이터 블록에 대해 독립적으로 일련의 부분적 E-step을 수행하는 점에서 기본 EM과 차이가 있는데, 이러한 특성상, 데이터 블록을 생성하는 방법에 따라 그 수렴상의 특성도 달라질 수 있다. PLSA 모델은 일반적인 unigram mixture model과 달리 문서 내 각 단어에 대한 mixture model을 가정하므로 데이터 분할 단위는 각 문서-단어 쌍이라고 할 수 있다. 이러한 사실에 기반하여 점진적 EM에 의한 PLSA 모델 학습시 개별 문서 단위 분할, 개별 단어 단위 분할, 문서-단어쌍 단위 분할의 세가지 데이터 분할 방법을 고려하

였다.

Reuter 뉴스 문서 집합(RCV1-v2)[3]에 대한 실험 및 분석을 통해 제안하는 점진적 EM 기반 PLSA 모델 학습의 유용성을 검증하였다. 총 문서 개수는 23,149개이며 전체 어휘집 크기는 47,089이다. 그림 2는 EM 알고리즘의 반복 횟수 측면에서의 모델 학습 속도를 비교하여 제시하며, 점진적 EM 알고리즘이 기본 EM 알고리즘에 비해 반복 횟수 면에서 보다 빠르게 학습이 진행됨을 알 수 있다. 하지만 단위 반복당 점진적 EM 알고리즘의 시간 비용이 더 크기 때문에 학습의 최종 수렴 시까지의 실제 소요 시간 측면에서의 비교가 필요하다. 표 1은 세 가지 데이터 분할법을 채용한 점진적 EM 알고리즘의 기본 EM 알고리즘 대비 학습 시간 비용 개선율을 제시한다. 데이터 분할법과 상관없이 대체적으로 점진적 EM 알고리즘에 의한 PLSA 모델 학습이 기본 EM 기반 학습보다 유의미한 속도 개선을 달성함을 알 수 있다.

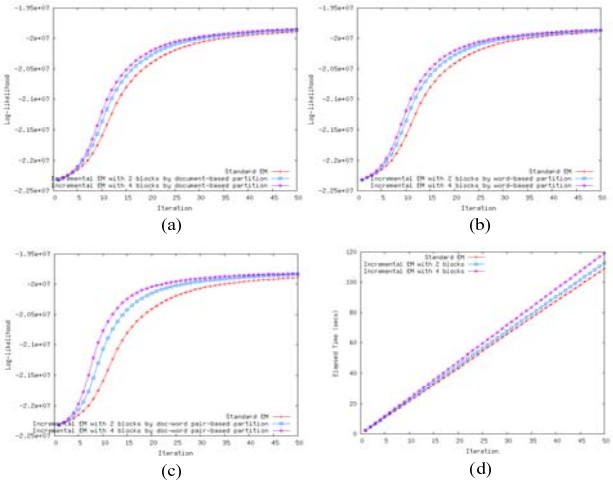

그림 2 기본 및 점진적 EM 알고리즘에 의한 PLSA 모델 학습 진행 비교: (a) 문서 단위 분할 (b) 단어단위 분할 (c) 문서-단어쌍 단위 분할 (d) 실제 소요시간 면에서의 비교

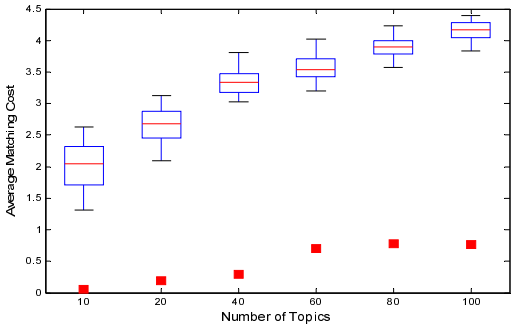

그림 3 PLSA 모델 학습에 의한 잠재토픽 집합들의 이분그래프 매칭 결과. Box plot은 임의초기화에 의한 기본EM간의 비교이며, ■기호는 동일초기화에 대한 점진적 EM 및 기본 EM의 비교 결과

이와 같은 PLSA 모델의 목적 함수 값 도달 측면에서의 비교에 더하여, 기본 EM 및 점진적 EM 알고리즘에 의한 학습 결과를, 추출된 토픽 집합 측면에서 비교해 보았다. 두 알고리즘에 의한 토픽 집합들간의 유사성 여부를 이분그래프 매칭(bipartite graph matching)을 통해 정량적으로 비교하였으며 구체적으로는 Hungarian matching 기법을 이용하였다. 단어들에 대한 다항 분포로 정의되는 두 토픽들간의 거리는 대칭형 KL divergence를 이용하여 계산하였으며 그림 3은 그 결과를 제시한다. 동일한 초기화에 대해 점진적 EM과 기본 EM에 의한 결과의 차이는 임의 초기화에 따른 기본 EM 결과들간의 변동에 비하여 그 차이가 상당히 미미하며 모델 학습 시 설정된 토픽 수와 상관없이 이러한 경향이 지속적으로 관찰됨을 알 수 있다. 이러한 결과들에 기초해 볼 때 결론적으로, PLSA 모델 학습시 점진적 EM 알고리즘은 기본 EM 알고리즘 기반의 학습에 비해 보다 적은 시간 내에 학습의 수렴을 달성하며 초기화가 동일할 경우 학습 결과 역시, 추출되는 토픽을 분석해 볼 때 기본 EM과 거의 동일한 수준을 달성한다고 할 수 있다. 이러한 점진적 EM 알고리즘을 병렬화 기법과 결합한다면 모델의 보다 효율적인 학습이 가능할 것이다.

표 1 점진적 EM 알고리즘에 의한 학습 시간 비용 개선 정도 (괄호 안은 해당 성능이 도출된 데이터 블록 수)

잠재토픽 수	문서단위 분할	단위단위 분할	문서-단어쌍 단위 분할
10	1.27 (4)	1.53 (8)	1.69 (6)
20	1.62 (6)	1.91 (6)	1.45 (4)
40	1.31 (6)	1.29 (16)	1.53 (6)
60	1.19 (4)	1.61 (6)	1.38 (12)
80	0.93 (2)	1.49 (12)	1.25 (8)
100	1.27 (4)	0.83 (4)	1.30 (4)

참고문헌

[1] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis, *Machine Learning*, 42(1-2):177-196, 2001.

[2] R. Neal and G. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in Graphical Models* pages 355-371. Kluwer Academic Publishers, 1998.

[3] D. D. Lewis, Y. Yang, T. Rose, and F. Li. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research* 5:361-397, 2004.