점진적 EM 알고리즘에 의한 잠재토픽모델의 학습 속도 향상

(Accelerated Learning of Latent Topic Models by Incremental EM Algorithm)

장 정 호 † 이 종 우 †† 엄 재 홍 ††
(Jeong-Ho Chang) (Jong-Woo Lee) (Jae-Hong Eom)

요 약 잠재토픽모델(latent topic model)은 데이타에 내재된 특징적 패턴이나 데이타 정의 자질들 간의 상호 관련성을 확률적으로 모델링하고 자동 추출하는 모델로서 최근 텍스트 문서로부터의 의미 자질 자동 추출, 이미지를 비롯한 멀티미디어 데이타 분석, 생물정보학 분야 등에서 많이 응용되고 있다. 이러한 잠재토픽모델의 대규모 데이타에 대한 적용 시 그 효과 증대를 위한 중요한 이슈 중의 하나는 모델의 효율적 학습에 관한 것이다. 본 논문에서는 대표적 잠재토픽모델 중의 하나인 PLSA (probabilistic latent semantic analysis) 기법을 대상으로 점진적 EM 알고리즘을 활용한, 기본 EM 알고리즘 기반의 기존 학습에 대한 학습속도 증진 기법을 제안한다. 점진적 EM 알고리즘은 토픽 추론 시 전체 데이타에 대한 일괄적 E-step 대신에 일부 데이타에 대한 일련의 부분적 E-step을 수행하는 특징이 있으며 이전 데이타 일부에 대한 학습 결과를 바로 다음 데이타 학습에 반영함으로써 모델 학습의 가속화를 기대할 수 있다. 또한 이론적인 측면에서 지역해로의 수렴성이 보장되고 기존 알고리즘의 큰 수정 없이 구현이 용이하다는 장점이 있다. 논문에서는 해당 알고리즘의 기본적인 응용과 더불어 실제 적용과정 상에서의 가능한 데이타 분할법들을 제시하고 모델 학습 속도 개선 면에서의 성능을 실험적으로 비교 분석한다. 실세계 뉴스 문서 데이타에 대한 실험을 통해, 제안하는 기법이 기존 PLSA 학습 기법에 비해 유의미한 수준에서 학습 속도 증진을 달성할 수 있음을 보이며 추가적으로 모델의 병렬 학습 기법과의 조합을 통한 실험 결과를 간략히 제시한다.

키워드 : 잠재토픽모델, PLSA, EM 알고리즘, 점진적 학습

Abstract Latent topic models are statistical models which automatically captures salient patterns or correlation among features underlying a data collection in a probabilistic way. They are gaining an increased popularity as an effective tool in the application of automatic semantic feature extraction from text corpus, multimedia data analysis including image data, and bioinformatics. Among the important issues for the effectiveness in the application of latent topic models to the massive data set is the efficient learning of the model. The paper proposes an accelerated learning technique for PLSA model, one of the popular latent topic models, by an incremental EM algorithm instead of conventional EM algorithm. The incremental EM algorithm can be characterized by the employment of a series of partial E-steps that are performed on the corresponding subsets of the entire data collection, unlike in the conventional EM algorithm where one batch E-step is done for the whole data set. By the replacement of a single batch E-M step with a series of partial E-steps and M-steps, the inference result for the previous data subset can be directly reflected to the next inference process, which can enhance the learning speed for the entire data set. The algorithm is advantageous also in that it is

· 저자 장정호는 정보통신부 해외유학지원사업(2005년도)의 지원으로 본 연구를 수행하였음

† 정 회 원 : Fraunhofer IAIS.KD 연구원
jeongho.chang@gmail.com
†† 학생회원 : 서울대학교 전기컴퓨터공학부
jwlee@bi.snu.ac.kr
jheom@bi.snu.ac.kr
(Corresponding author임)

논문접수 : 2007년 9월 27일
심사완료 : 2007년 11월 8일

정보과학회논문지 : 소프트웨어 및 응용 제34권 제12호(2007.12)

guaranteed to converge to a local maximum solution and can be easily implemented just with slight modification of the existing algorithm based on the conventional EM. We present the basic application of the incremental EM algorithm to the learning of PLSA and empirically evaluate the acceleration performance with several possible data partitioning methods for the practical application. The experimental results on a real-world news data set show that the proposed approach can accomplish a meaningful enhancement of the convergence rate in the learning of latent topic model. Additionally, we present an interesting result which supports a possible synergistic effect of the combination of incremental EM algorithm with parallel computing.

Key words : Latent topic model, PLSA, EM algorithm, Incremental learning

1. 서 론

잠재토픽모델(latent topic model)은 데이타에 내재된 특징적 패턴이나 데이타를 정의하는 자질들 간의 상호 관련성을 확률적으로 모델링하고 추출하는 모델로서, 대표적으로는 PLSA(probabilistic latent semantic analysis) [1], LDA(latent Dirichlet allocation)[2] 등을 들 수 있다. 이러한 잠재토픽모델은 주어진 데이타로부터 유의미한 중요 특성을 요약적으로 제시하는 그 자체로도 유용성이 클 뿐 아니라 분류작업과 같은 최종 작업을 위한 전처리 단계에서도 효과적으로 적용 가능하다. 이와 같은 효용성에 기초하여 잠재토픽모델은 대규모 텍스트 문서에 대한 의미 자질의 자동 추출 및 군집화/가시화 등의 응용에 널리 활용되었으며, 최근에는 관련 모델에 대한 관심과 연구가 증대되면서 이미지를 비롯한 멀티미디어 데이타 분석, 생물정보학 분야 등에서도 많이 응용되고 있다.

잠재토픽모델을 실제 대규모 데이타에 대해 적용할 때 그 효용성을 증대시키기 위한 중요한 이슈 중의 하나는 모델의 효율적 학습에 관한 것이며, 본 논문에서는 점진적 학습 알고리즘에 의한 잠재토픽모델의 학습 속도 개선 기법을 제시한다. 잠재토픽모델은 기본적으로 가상의 랜덤변수(random variable) 즉, 직접 관찰되지 않는 내부적 토픽을 가정하는데 이러한 특성상 모델의 학습을 위해 EM 알고리즘[3] 계열의 기법들이 주로 활용되고 있다. EM 알고리즘 에서는 E-step에서 전체 데이타에 대한 일괄적인 랜덤변수 추정이 이루어진 후 추론 결과를 바탕으로 M-step에서 모델의 매개변수 추정이 이루어지는데, 논문에서는 이러한 일괄적인 추정 대신에 이 과정을 점진적으로 진행하는 점진적 EM 알고리즘[4]을 적용하여 잠재토픽모델 학습의 수렴속도를 증진시키고자 한다.

대표적인 잠재토픽모델인 PLSA 모델을 대상으로 점진적 EM 알고리즘에 의한 학습 기법을 제시하며 기존 일반 군집화 모델과의 차이를 고려하여 점진적 EM 알고리즘의 적용 과정상에서의 가능한 데이타 분할법들을 비교·분석한다. 그리고 실세계 뉴스 문서 데이타에 대해 잠재변수 및 데이타 블록 수에 따른 PLSA 모델의 학습 속도 증진 정도를 실험적으로 제시하고 그 결과를 학습 과정상의 특징과 연관하여 분석한다.

1.1 논문의 구성

본 논문은 다음과 같이 구성된다. 먼저 2장에서는 대표적 잠재토픽모델 중의 하나인 PLSA와 EM을 이용한 PLSA 학습에 대해서 다룬다. 3장에서는 점진적 EM 알고리즘의 기본 아이디어 및 이를 이용한 모델 학습 방법에 대하여 다루고, PLSA 모델의 학습에 점진적 EM 알고리즘을 적용하는 방법에 대해 기술한다. 4장에서는 Reuter 뉴스 문서집합을 이용한 실험 및 결과에 대해서 기술한다. 실험 결과는 모델 학습 속도와 학습 품질 면의 고찰을 함께 고려하여 기술한다. 마지막으로 5장에서는 논문의 결론, 본 논문에서 다룬 방법의 한계 및 전반적인 성능 향상을 위한 향후 연구과제에 대해 다룬다.

2. Probabilistic Latent Semantic Analysis (PLSA)

PLSA 모델[1]은 공기(co-occurrence) 데이타나 히스토그램 데이타에 대해 효과적인 분석이 가능한, 확률적 잠재변수모델 기반의 방법론으로서 특히 텍스트 문서 관련 응용(언어모델링, 정보검색, 문서분류 등의 후 작업을 위한 의미자질 추출 등)에서 유용하게 응용되고 있다. 하나의 문서에 대해 PLSA는 (가상의) 내부적 토픽에 의한 생성을 가정하는데, 기존의 군집화 모델과 같은 unigram mixture model과 달리 각 문서를 토픽의 조합적 구성으로 파악하고 하나의 문서가 아닌 각 (문서-단어)쌍에 대한 mixture model을 가정하는 특징이 있다. 그림 1은 그래프 모델에 의해 PLSA 모델을 요약적으로 제시한다. 그림에서 d와 w는 각각 문서 및 단어

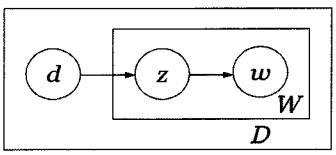


그림 1 그래프모델에 의한 PLSA 표현

를 의미하며 z는 문서 내 각 단어에 대한 잠재토픽을 의미한다.

PLSA 모델에서 전체 데이타에 대한 모델의 로그우도함수는 식 (1)과 같이 정의되며,

$$L(\theta;D)=\log\prod_{d=1}^{D}\prod_{w=1}^{W}p(w|d)^{n(d,w)} \\ =\sum_{d=1}^{D}\sum_{w=1}^{W}n(d,w)\log(\sum\nolimits_{z}p(w|z)p(z|d)) \quad (1)$$

모델의 학습은 이 함수 값을 최대화하는 모델 매개변수, 즉 최대우도추정치(MLE; Maximum Likelihood Estimates of Parameters) $p(w|z)$와 $p(z|d)$를 구하는 것이다. 이의 추정은 EM 알고리즘에 의해 이루어지며 E-step과 M-step은 다음과 같이 정의된다.

$$\text{E-step: } p(z|d,w)=\frac{p(w|z)p(z|d)}{\sum_{z'}p(w|z')p(z'|d)} \quad (2)$$

$$\text{M-step: } p(z|d)\propto\sum\nolimits_{w}n(d,w)p(z|d,w) \\ p(w|z)\propto\sum\nolimits_{d}n(d,w)p(z|d,w) \quad (3)$$

EM 알고리즘을 통해 식 (2)와 (3)의 과정을 교대로 반복적으로 수행함으로써 (지역)최적해를 추정할 수 있다. 여기서, E-step은 전체 데이타 즉 모든 문서-단어쌍에 대해 일괄적으로 이루어지는데 3절에서는 이러한 기본 EM 알고리즘에 의한 학습 방법에 기초하여 논문에서 제시하고자 하는 점진적 EM 알고리즘에 의한 PLSA 모델 학습 기법에 대해 설명한다.

3. 점진적 EM 알고리즘에 의한 PLSA 학습

3.1 점진적 EM 알고리즘

Neal과 Hinton[4]은 기본 EM 알고리즘을 통계물리학에서의 자유에너지(free energy)개념에 기반한 함수의 최적화 과정으로 해석하고, 이에 기초하여 기본 EM 알고리즘의 변형으로서 점진적 알고리즘을 제시하였으며 해당 알고리즘이 기본 알고리즘과 마찬가지로 (지역)최적해에 수렴함을 제시하였다.

결측치(缺測値; missing value)를 포함한 데이타 집합 $D=\{(x_i,z_i)\}$에 대해 모델의 특정 매개변수 θ에 대한 로그우도함수는 다음과 같이 (음의) 자유에너지 형태로 표현할 수 있다(x와 z는 각각 입력변수 및 미관찰(未觀察) 랜덤변수를 의미하며, 각 데이타 개체는 독립이고 동일한 분포를 따른다고 가정).

$$L(\theta;D)=\log p(D|\theta) \\ \geq\sum_{i=1}^{N}E_q\left[\log p(x_i,z_i|\theta)\right]+H(q)=F(q,\theta) \quad (4)$$

식 (4)에서 등식은 $q=p(z_i|x_i,\theta)$일 때 성립하며 $H(q)=-\sum q\log(q)$는 분포 q의 엔트로피(entropy)를 의미한다. 기본 EM 알고리즘은 E-step과 M-step을 반복적으로 수행함으로써 모델에 대한 (지역)최적해를 추정해 나가는데, 로그우도함수의 식 (4)에 의한 형식화를 통해 E-step과 M-step은 각각 q와 θ에 대한 함수 F의 최적화 과정으로 재해석할 수 있다. 즉,

$$\text{E-step: } q^{n+1}=\arg\max_{q}F(q,\theta^{n}) \quad (5)$$

$$\text{M-step: } \theta^{n+1}=\arg\max_{\theta}F(q^{n+1},\theta) \quad (6)$$

점진적 EM 알고리즘은 기본 EM 알고리즘에 대한 이러한 관점으로부터 유도되는데, 해당 알고리즘 수행은 표 1과 같이 요약할 수 있다. 점진적 EM 알고리즘은 E-step 면에서 기본 EM 알고리즘과 차별화되는데, 즉 전체 데이타에 대한 일괄적 E-step 대신 데이타 일부에 대한 일련의 부분적(partial) E-step을 수행한다는 점이다. 그리고 M-step은 각 부분적 E-step 수행 후 기본 EM 알고리즘에서와 같이 전체 데이타에 대해 적용된다. 따라서

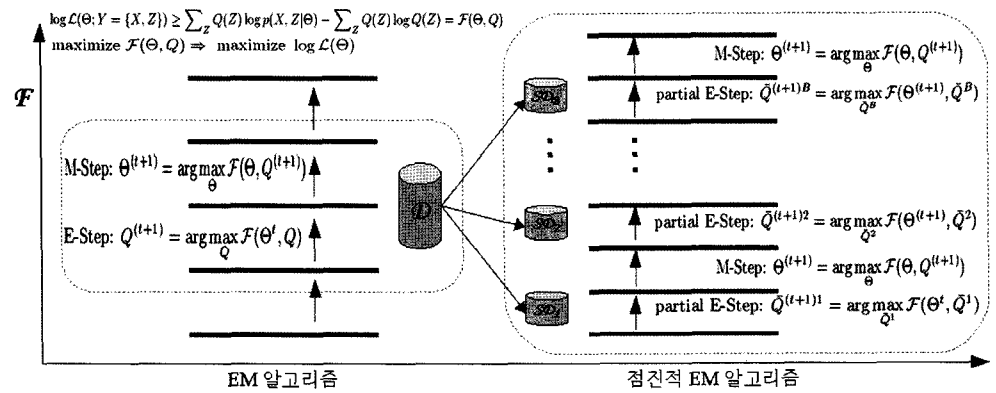


그림 2 기본 EM 알고리즘과 점진적 EM 알고리즘의 동작방식 비교 (점진적 EM 알고리즘은 E-Step을 부분 E-Step으로 분할하여 수행한다)

표 1 점진적 EM 알고리즘에 의한 모델 학습 개요

입력: ($\boldsymbol{D}=\{d_1,d_2,...,d_N\}$, $\boldsymbol{B}$)

$\boldsymbol{D}$: 총 N개의 데이타를 가지는 데이타 집합.

$\boldsymbol{B}$: 분할 데이타 블록 수.

출력: ($\boldsymbol{\theta^{MLE}}$)

$\boldsymbol{\theta^{MLE}}$: 최적 매개변수 집합 θ.

학습 알고리즘

1. 초기화 단계:

1-1. 전체 데이타 D를 B개의 서브 데이타 블록으로 분할

$D=\{SD_1,SD_2,...,SD_B\}$

1-2. 학습 알고리즘을 위한 인덱스 초기화

$t \leftarrow 1$,

$n \leftarrow 1$

1-3. 모델을 초기화

2. 학습 단계:

2-1. for $b=1,2,...,B$ (각 서브 데이타 블록에 대해)

2-1-1. 부분 E-step 수행

$q^{tb}=q^t(SD_b)=\arg\max_q F(q(SD_b),\theta^n)$

2-1-2. M-step 수행

$\theta^{n+1}=\arg\max_\theta F(q^n,\theta)$

(where, $q^n=(q^{t1},...,q^{tb},q^{(t-1)(b+1)},...,q^{(t-1)B})$)

2-1-3. $n \leftarrow n+1$

end

2-2. (a) 수렴(converge)되지 않은 경우 아래의 갱신작업 후 (2-1) 반복

$t \leftarrow t+1$

(b) 수렴된 경우 점진적 EM 학습알고리즘 종료, $\boldsymbol{\theta^{MLE}}$ 반환

전체 데이타를 B개의 데이타 블록으로 나눌 경우, 전체 데이타에 대한 매 스캔과정상에서 총 B번의 부분적 E-step 및 M-step을 수행하게 된다. 그림 2는 이러한 점진적 EM 알고리즘의 특성을 도식적으로 나타낸다.

점진적 EM 알고리즘은 기본 EM 알고리즘과 비교하여, 매 스캔과정에서 E-Step에 따른 추가 계산 비용은 발생하지 않으나 M-Step에 대해서는 데이타 블록 수에 비례하는 계산 비용이 추가된다. 하지만, 점진적 EM 알고리즘은 부분적 E-step에서의 현재 데이타 블록에 대한 결측치 추정 및 매개변수 학습 결과를 이후 데이타 블록에 대한 학습과정에서 바로 이용함으로써 전체 학습 과정의 수렴 속도(전체 데이타 반복적 접근 횟수 면에서)를 향상시킬 수 있다. 결론적으로 학습 시 보다 적은 반복 횟수에 모델의 수렴을 달성할 수 있다면 매 데이타 스캔에 대한 추가 계산 비용을 고려하더라도 학습 시간을 단축시킬 수 있다. 실제 [5]에서는 실세계 대규모 데이타에 대한 mixture model 학습 시 점진적 EM 알고리즘을 적용함으로써 전체 학습 소요 시간 면에서 상당한 개선을 달성할 수 있음을 보였다.

3.2 PLSA 모델에 대한 점진적 EM 알고리즘 적용

텍스트 문서 집합 D에 대한 PLSA의 로그우도함수(식 (1))는 식 (4)에 의해 다음과 같이 자유에너지 형태로 표현할 수 있다.

$$L(\theta;D) \geq F(q,\theta) = \sum_{d=1}^{D}\sum_{w=1}^{W} n(d,w)\sum_{z} q(z|d,w)\log(p(z|d)p(w|z)) + H(q) \quad (7)$$

또한, 3.1절에서 제시된 형태의 점진적 EM 알고리즘을 이용해 전체 데이타에 대한 매 스캔마다 식 (2)와 (3)의 학습 방법은 다음과 같이 달성되며, 그림 3은 기존 EM 알고리즘과 점진적 EM 알고리즘에 의한 PLSA 모델 학습과정을 요약적으로 제시한다.

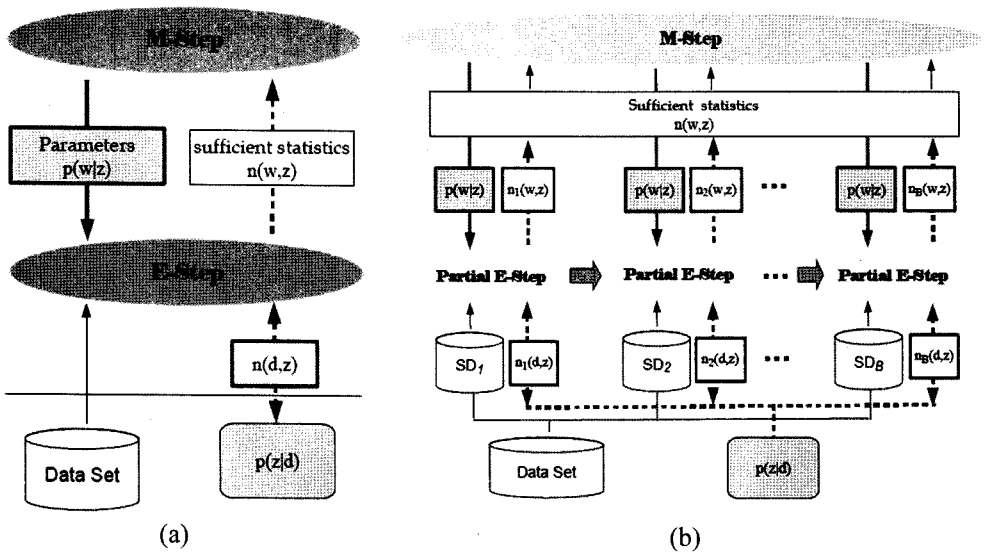


그림 3 EM 알고리즘에 의한 PLSA 모델 학습: (a) 기본 EM 알고리즘, (b) 점진적 EM 알고리즘

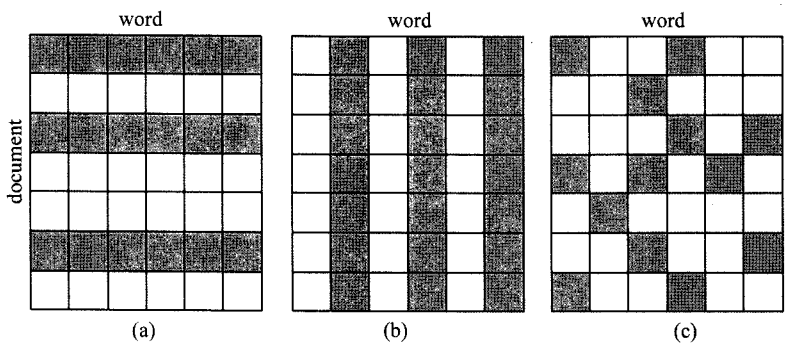


그림 4 PLSA의 점진적 EM 알고리즘을 이용한 학습 시 데이타 블록 생성: (a) 문서 단위 (b) 단어 단위 (c) 문서-단어쌍 단위

Partial E-step:

$$s^t_{zw}(SD_b) = \sum_{d \in SD_b} n(d,w)p(z \mid d,w)$$
$$s^t_{dz}(SD_b) = \sum_{w \in SD_b} n(d,w)p(z \mid d,w)$$

M-step: (8)

$$p(w \mid z) = \frac{s^n_{zw}}{\sum_l s^n_{zl}},\ p(z \mid d) = \frac{s^n_{dz}}{\sum_l s^n_{lz}}$$
$$s^n_{zw} = (s^t_{zw}(SD_1),...,s^t_{zw}(SD_b), s^{(t-1)}_{zw}(SD_{(b+1)}),...,s^{(t-1)}_{zw}(SD_B))$$
$$s^n_{dz} = (s^t_{dz}(SD_1),...,s^t_{dz}(SD_b), s^{(t-1)}_{dz}(SD_{(b+1)}),...,s^{(t-1)}_{dz}(SD_B))$$

점진적 EM 알고리즘은 데이타를 먼저 분할하고 각 데이타 블록에 대해 독립적으로 일련의 부분적 E-step을 수행하는 특성상, 데이타 블록을 생성하는 방법에 따라 그 성능이 영향을 받을 수 있다. 특히, PLSA 모델은 기존의 군집화 모델과 같은 unigram mixture model과는 달리 문서 내 각 단어에 대한 mixture model을 가정하므로 분할의 단위는 각 문서-단어 쌍이라고 할 수 있다. 이러한 특성을 고려하여, 개별 문서 단위의 분할 외에 개별 어휘 단위 분할과 문서-단어쌍 단위 분할을 추가적으로 고려할 수 있는데, 그림 4는 이와 같은 세 가지 형태의 데이타 분할법을 보인다. 4장에서의 실험을 통해 각 데이타 분할법에 의한 성능을 경험적으로 비교한다.

4. 실험 및 결과

Reuter 뉴스 문서 집합(RCV1-v2)[1] [6]에 대해 실험을 하였다. RCV1-v2는 문서 분류를 위한 표준 문서 집

1) http://www.ai.mit.edu/projects/jmlr/papers/volume5/lewis04a/lyrl2004_rcv1v2_README.htm에서 가공된 형태의 데이타를 이용 가능

합의 하나이며, 실험을 위해 그 중 23,149개의 문서들로 구성된 학습용 데이타 집합을 이용하였다. 해당 문서 집합의 어휘집 크기는 47,089, 전체 단어 빈도수는 2,772,838이다.

4.1 모델 학습 속도

먼저 기본 및 점진적 EM 알고리즘의 반복에 따른 학습 속도 면에서의 비교를 위해 반복 횟수 및 시간에 따른 학습 경향을 검토하였다. 그림 5의 (a)-(c)는 PLSA 모델의 잠재 토픽 수를 20으로 설정하고 기본 알고리즘 및 세 가지 데이타 분할법에 의한 점진적 EM 알고리즘의 초기 반복 수 50에 이르기까지의 로그우도 값의 변화를 보인다. 데이타 블록의 수는 $B=2$와 $B=4$에 대해 실험하였다. 데이타 블록들은 랜덤 방식으로 생성된다. 일단 학습 알고리즘의 반복횟수 측면에서 볼 때, 점진적 EM 알고리즘이 모든 데이타 분할법에 대해서 기본 EM 알고리즘에 비해 보다 더 진행된 학습을 달성함을 알 수 있다. 그리고 전체 데이타를 보다 더 많은 수의 블록으로 분할할수록 학습 진행 속도가 더욱 증진되는 것을 볼 수 있다. 하지만 이는 반복 횟수 면에서의 비교일 뿐이며, 점진적 EM 알고리즘은 전체데이타에 대한 매 반복마다 추가적인 M-step 비용이 소요되므로 실제 응용차원에서는 소요 시간 면에서의 비교가 중요하다.

그림 5의 (d)는 실제 소요시간 면에서의 문서-단어쌍 기반의 데이타 분할에 의한 점진적 EM 알고리즘과 기본 알고리즘의 학습 속도를 제시한다. 그림에서도 볼 수 있듯이, 동일 반복 횟수 대비 점진적 EM이 기본 EM에 의한 학습 시간보다 다소 더 소요되며 또한 데이타 분할 수가 클수록 소요시간도 증대된다. 즉 반복횟수 면에서는 학습 진행이 더 가속화되지만 반복 회당 소요시간이 크기 때문에 점진적 EM 알고리즘에 의한 속도 향상이 이루어지기 위해서는 기본 EM에 의한 학습보다 적

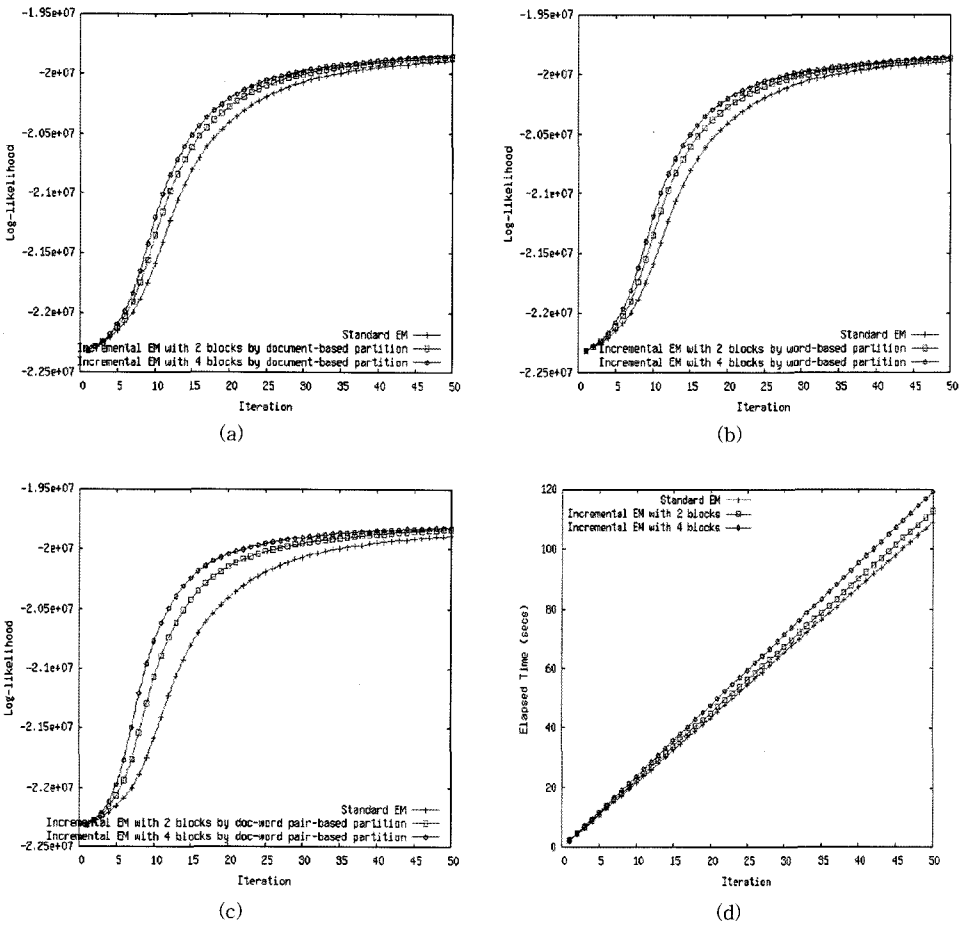


그림 5 기본 및 점진적 EM 알고리즘에 의한 PLSA 학습 진행 비교: (a) 문서 단위 데이타 분할, (b) 단어 단위 데이타 분할, (c) 문서-단어쌍 단위 데이타 분할, (d) 소요시간 면에서의 비교

표 2 점진적 EM 알고리즘에 의한 학습 시간 비용 개선 정도(괄호안의 숫자는 해당 성능이 도출된 데이타 블록 수를 의미)

잠재토픽 수	문서 단위 분할	단어 단위 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)

은 반복횟수 내에 수렴이 이루어져야 한다. 이를 확인하기 위해 여러 잠재토픽 수 및 블록 수 조합에 대해 기본 EM과 점진적 EM의 소요시간을 비교하였다.

PLSA의 잠재토픽 수는 10, 20, 40, 60, 80, 100으로 설정하였으며 각 잠재토픽 수에 대해 점진적 EM에서의 블록 수를 2, 4, 6, 8, 12, 16로 변화시켜 나가면서 실험하였다. 학습은 로그우도함수 값의 이전 대비 상대적 증가율이 5×10^{-6} 이하가 될 때까지 진행하였다.[2] 표 2는 학습 소요시간 속도 향상 측면에서 그 실험결과를 요약·제시한다.

표 2 내의 각 숫자는 기본 EM 알고리즘의 수렴 시 로그우도 값에 대한 점진적 EM 알고리즘의 실제 소요시간 면에서의 향상률을 나타낸다. T_S, T_I를 각각 해당 알고리즘이 소요된 시간이라 할 때, 향상률 r은 $r = T_s/T_I$로 계산되며 따라서 $r>1.0$인 경우 점진적 EM에 의해 학습시간이 감소하였음을, 그리고 $r<1.0$인 경우 오히려 학습시간이 증가하였음을 나타낸다. 대부분의 경우에서 (잠재토픽 수 80에 대한 문서기반 분할, 100에 대한 단어 단위 분할 제외), 정도의 차이는 있지만 의미 있는 수준의 학습 속도 증진이 이루어졌음을 알 수 있다.

보다 세부적으로 먼저 데이타 분할 방법 측면에서 볼 때 단어 단위 및 문서-단어쌍 단위 분할법이 문서기반 분할법에 비해 학습률 개선 면에서 더 우수한 경향을 보였다. 이는 PLSA 모델 학습 과정상의 특징에 기인한 것으로 판단된다. 즉, 각 문서의 토픽구성을 의미하는 $p(z|d)$는 해당 문서에 한정하여 추정되기 때문에 문서 단위 분할법의 경우 하나의 데이타 블록에 소속된 문서의 경우, 일단 해당 블록에 대한 학습 후에는 동일 반복 과정 내의 나머지 블록에 대한 일련의 학습 과정 중에는 전혀 갱신되지 않는다. 이에 반해 나머지 두 분할법에서는 하나의 문서가 서로 다른 블록에 중첩적으로 포함될 수 있기 때문에 각 문서에 대한 $p(z|d)$ 갱신과정이 $p(w|z)$ 갱신과 더불어 동일 반복 내의 나머지 블록에 대한 학습과정에서도 지속적으로 이루어질 수 있다.

다음으로 단어 단위 분할법과 문서-단어쌍 단위 분할법을 비교해 보면, 먼저 그림 5에서 문서-단어쌍 단위 분할이 단어단위 분할법에 비해 반복 횟수당 학습 속도 면에서 진행이 더 빠름을 볼 수 있지만, 표 1에서 실제 학습시간 면에서는 그 반대의 상황을 관찰할 수 있는데 이는 문서/단어쌍 단위 분할의 경우가 M-Step에 필요한 추가소요시간이 더 많기 때문으로 판단된다. 즉, 어휘집 크기를 W라 하고 전체 데이타를 B의 블록으로 분할할 경우 단어 단위 분할의 경우, 각 부분적 E-step 결과로부터의 M-step에서의 $p(w|z)$갱신 시 평균적으로 W/B개의 단어만 고려하면 되지만, 문서-단어쌍 분할법의 경우 각 쌍을 임의로 추출하기 때문에 M-step에서 평균적으로 이보다 많은 수의 단어에 대한 갱신이 필요하며 이는 전체적인 학습 비용을 증가시킬 수 있다. 실례로 그림 6은 각 분할법의 블록 수에 따른 블록 내 서로 다른 문서 및 단어들의 수를 도시하는데, 문서-단어쌍 분할법은 갱신이 필요한 문서 및 단어 개수 면에서 가장 많은 경향을 보임을 알 수 있다.

하지만 이는 동시에 해당 분할법은 단일 블록 내에서 가장 많은 개수의 서로 다른 $p(z|d)$와 $p(w|z)$에 대한 갱신이 이루어짐을 의미하므로 보다 적은 반복 횟수 내에 수렴이 가능함을 의미하며 이것이 그림 5의 반복횟수 면에서의 경향을 설명할 수 있다. 그리고 표 2에서 볼 수 있듯이, 해당 분할법은 단어 단위 분할법과 더불어 세 가지 데이타 블록 수에 대해 가장 높은 수준의 학습 속도 증진을 달성할 수 있었다. 결과적으로 각 데이타 블록 내에서 갱신되는 매개변수의 다양성 확보와 추가적 소요시간 사이의 적절한 trade-off가 필요하다고 할 수 있을 것이다.

이와 같이 점진적 EM 알고리즘에 의해 잠재토픽모델의 학습 속도를 의미 있는 수준으로 개선할 수 있었는데, 또 다른 관점에서 모델 학습을 증진시키는 방안은 학습의 병렬화를 통해서일 것이다. 이의 확인을 위해 점진적 EM 알고리즘과 병렬화 기법을 조합한 실험 결과를 간략하게 제시하고자 한다. 실험을 위해 잠재토픽의 수는 60으로 설정하였으며, 2-core CPU 상에서 멀티쓰레딩(multi-threading)을 통해 병렬화를 구현하였다. 성능 개선은 기본 EM 알고리즘의 단일 프로세스 상에서

2) 점진적 EM 알고리즘의 경우, (음의)자유에너지 값에 대해 같은 기준을 적용하였다.

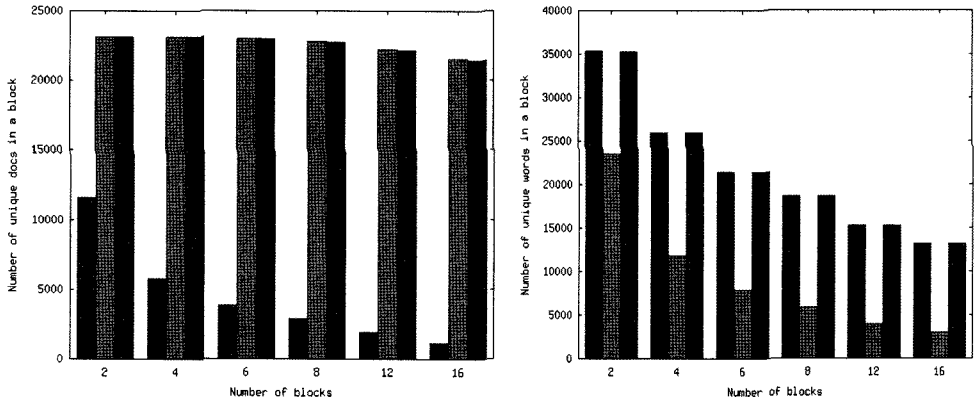


그림 6 데이타 분할법에 따른 각 데이타 블록 내 서로 다른 문서 및 단어의 수. 각 막대 그룹 내에서 순서대로 문서 단위, 단어 단위, 문서-단어쌍 단위 분할에 의한 결과이다.

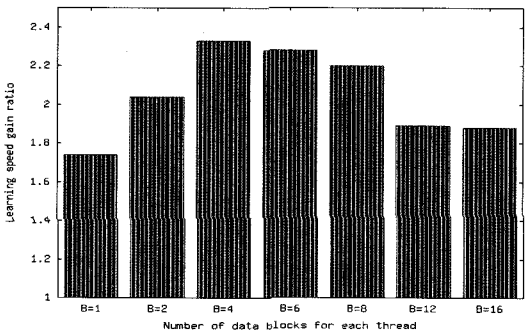


그림 7 병렬화를 통한 단일 프로세스/기본 EM 대비 잠재토픽모델 학습 속도 향상 추이(2개의 프로세스를 이용한 결과이며, B는 각 프로세스 내 데이타 블록 수이고 B=1인 경우는 기본 EM의 병렬화를 의미한다)

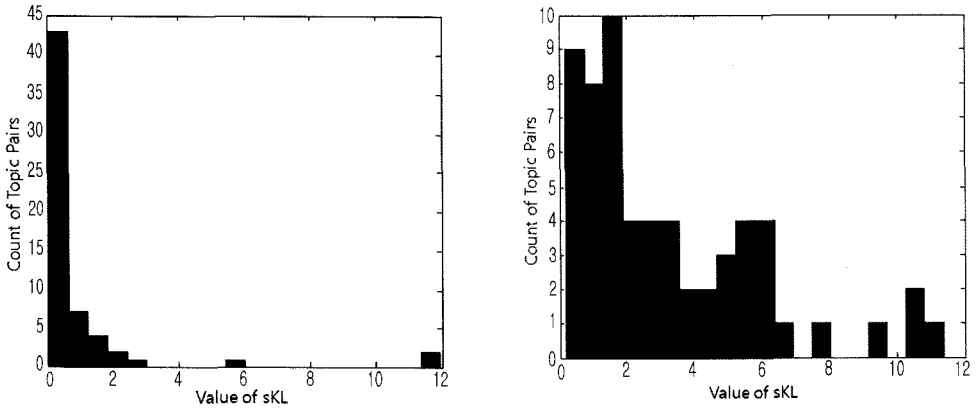
의 결과를 기본으로 하여 향상된 정도를 측정하였으며, 그림 7은 그 결과를 제시한다. 단순히 기본 EM 알고리즘을 N개의 프로세스에 의해 병렬화 할 경우 일반적으로 N배 이상의 성능 개선을 기대할 수 없다(스레드 생성과 동기화 등에 따른 비용을 고려할 때). 실험에서도 해당 병렬 구현은 약 1.7배 정도의 성능 향상에 그쳤다. 이에 반해 각 프로세스에 할당된 데이타에 대해 점진적 EM 알고리즘을 적용한 경우, 단일 프로세스 상의 기본 EM에 비해 최고 약 2.3배의 성능 향상을 이룰 수 있었다. 비록 두 프로세스 상에서의 실험 결과이지만, 이는 점진적 EM과 병렬화의 조합적 접근법의 유용성을 제시한다고 할 수 있다.

4.2 모델 학습 결과 비교

PLSA 모델의 로그 우도 값 측면에서의 비교에 더하여, 기본 EM 및 점진적 EM에 의한 학습 결과를 추출된 토픽 집합 측면에서 비교하는 것도 의미가 있을 것이다. 먼저 두 알고리즘에 의한 토픽 집합들 간의 유사성 여부를 이분그래프매칭(bipartite graph matching)을 통해 정량적으로 비교하였으며, 구체적으로 Hungarian 매칭 기법[7]을 이용하였다. 두 토픽 집합 $p_i(w|z_k)_{k=1}^K$, $p_j(w|z_k)_{k=1}^K$에 대해 각 집합으로부터의 토픽쌍 $(p_i(w|z_k), p_j(w|z_l))$의 거리 $C_{ij}(k,l)$는 다음과 같이 대칭형 KL divergence(symmetric Kullback-Leibler divergence)로 계산하였다.

$$\begin{aligned} C_{ij}(k,l) &= sKL(p_i(w|z_k), p_j(w|z_l)) \\ &= \frac{1}{2}\{KL(p_i(w|z_k),\ p_j(w|z_l)) \\ &\quad + KL(p_j(w|z_l),\ p_i(w|z_k))\} \end{aligned} \tag{9}$$

그림 8은 추출된 토픽 집합 측면에서 비교 결과를 제시한다. 점진적 EM 알고리즘의 경우 문서-단어쌍 기반의 데이타 분할을 이용하였다. 그림에서, 동일한 초기화에 대해 점진적 EM과 기본 EM에 의한 결과(■)는 임의 초기화에 따른 기본 EM 결과들 간의 변동(box plot에 의해 표현)에 비교해 볼 때, 그 차이가 상당히 미미하며 PLSA 모델 학습 시 설정된 토픽 수와 상관없이 이러한 경향이 지속적으로 관찰됨을 알 수 있다. 그림 9는 기본 EM과 점진적 EM 기반의 PLSA 모델 학습 결과(토픽 수는 60)에 대한 이분그래프매칭 시 매칭 된 토픽쌍 간의 매칭 비용의 분포를 보이며 더불어 기본 EM의 임의초기화에 의한 실행 결과들 간의 매칭 비용을 제시한다. 동일 초기화에 대한 점진적 EM에 의한 결과와 기본 EM에 의한 결과들의 비교(그림 9(a))를 보면 대부분 그 거리 값이 1.0 이하이며 이는 임의초기화를 통한 기본 EM들 간의 매칭 결과(그림 9(b))와 대비해 볼 때 의미 있는 결과이다. 따라서 PLSA 모델을



(a) (b)

그림 9 이분 그래프 매칭 결과 토픽 간 매칭 비용 분포: (a) 점진적 EM과 기본 EM에 의한 결과 (b) 임의초기화에 의한 기본 EM들 간의 비교 결과

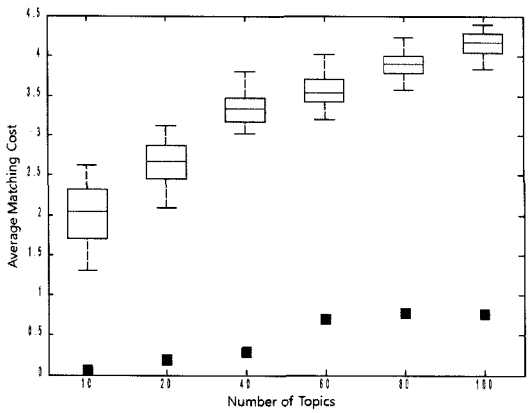


그림 8 PLSA 모델 학습에 의한 잠재토픽 집합들의 이분 그래프 매칭 결과. Box plot은 기본 EM에 대해 10회 임의 초기화에 의해 추출된 토픽 집합들 간의 이분그래프 매칭 비용을 보이며, ■ 기호는 동일 초기화에 대해 기본 EM과 점진적 EM 결과간의 매칭 비용을 나타낸다.

점진적 EM 알고리즘으로 학습함으로써 얻는 수렴해와 기본 EM에 의한 수렴해의 차이는 크지 않으며, 개별 토픽별로 분석해 볼 때 점진적 EM에 의한 토픽들이 기본 EM에 의한 토픽들과 대부분 유사하다는 것을 알 수 있다. 표 3은 기본 EM과 점진적 EM 학습 결과로부터 매칭 된 토픽쌍 중에서 그 비용이 가장 작은 쌍과 가장 큰 쌍을 제시한다.

추가적으로, 점진적 EM과 기본 EM에 의한 PLSA 모델 학습 결과를 문서 간 유사도 측정에 적용하여 그 성능을 비교하였다. 이를 위해 먼저 Reuter 뉴스 문서 집합 중 90%를 학습 데이타로 하여 PLSA 모델을 학습하였으며, 학습된 결과를 이용하여 나머지 10% 문서들에 대해 각 문서쌍 간의 거리를 측정하였다. 두 문서 d1, d2간의 거리 측정 시 각 문서를 PLSA 모델에 의해 정의되는 토픽 공간으로 사상한 다음 앞서 언급한 대칭형 KL divergence를 이용하여 계산하였다. 즉,

$$C(d_1, d_2) = sKL(p(z|d_1), p(z|d_2)). \quad (10)$$

그리고 계산된 거리 값들을 오름차순으로 정렬한 다음, 정렬된 각 쌍에 대해 문서에 대해 미리 정의된 범주(category) 정보를 이용하여 다음의 식 (11)과 같이 pair-wise precision을 측정하였다.[3)]

표 3 점진적 EM과 기본 EM에 의한 PLSA 모델 학습 결과 추출된 토픽집합 중 최저/최고 비용 매칭 토픽 쌍(전체 토픽 수는 60, 토픽별로 $p(w|z)$값이 큰 상위 10 단어를 순서대로 나열하였다)

최저 비용 매칭 토픽 쌍		최고 비용 매칭 토픽 쌍	
점진적 EM	기본 EM	점진적 EM	기본 EM
clinton	clinton	research	trad
presid	presid	chemic	tend
dole	dole	drug	pric
republ	republ	study	bid
convent	convent	alert	septemb
compaign	bill	buy	cent
bill	compaign	produc	buy
demo	demo	anal	octob
hous	hous	system	market
americ	americ	institut	oil

3) Reuter 뉴스 문서 집합(RCV1-v2) 의 경우, 하나의 문서가 여러 개의 범주에 속하는 경우가 다수 존재하는데, 실험에서는 동일 범주를 하나라도 공유할 경우 두 문서의 매칭이 올바른 것으로 설정하였다.

$$pr(i,j)=\frac{\left|\{(k,l)\,|\,(cat(d_k)=cat(d_i))\wedge(rank(k,l)\le rank(i,j))\}\right|}{rank(i,j)} \quad (11)$$

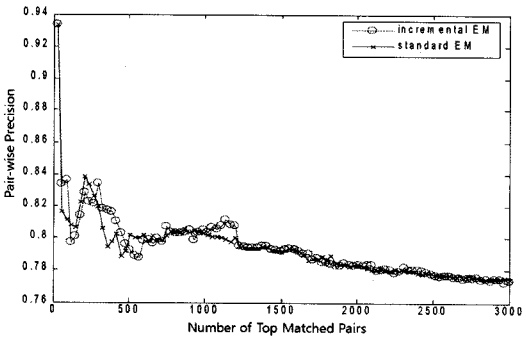


그림 10 점진적 EM과 기본 EM에 의해 학습된 PLSA 모델을 이용한 문서 유사도 측정 성능 비교 결과

식 (11)의 정의에 의해 전체적으로 pair-wise precision 값이 클수록 해당 학습 결과가 문서집합 내에 내재된 토픽 구조를 보다 더 잘 반영한다고 할 수 있다. 그림 10은 그 결과를 제시하며, 기본 EM에 의한 결과와 비교해 볼 때 정확도 측정 지점별로 약간의 차이가 있지만 점진적 EM에 의한 학습 결과가 거의 동일한 성능을 보임을 관찰할 수 있다.

이와 같은 결과에 기초하여, 결론적으로 PLSA 모델 학습 시 점진적 EM은 기본 EM에 비해 보다 적은 시간 내에 학습의 수렴을 달성할 수 있으며 초기화가 동일할 경우 학습 결과 역시 추출되는 토픽 집합 면에서나 미 관찰 데이타에 대한 성능 면에서 기본 EM과 거의 동일한 수준을 달성한다고 할 수 있다.

5. 결 론

본 논문에서는 점진적 EM 알고리즘을 이용한 텍스트 문서에 대한 잠재토픽모델의 학습 속도 증진 기법을 제시하였다. 데이타 블록들에 대해 일련의 부분적 EM 알고리즘을 연속적으로 수행하는 점진적 EM 알고리즘은 기본 EM 알고리즘에 비해 매 전체 데이타 접근마다의 단위 시간 비용은 크지만, 대체로 데이타 접근 횟수 즉 전체 반복수 면에서 보다 적은 횟수에 (지역)해에 수렴하는 특성이 있음을 확인할 수 있었다. 따라서 전체적으로는 잠재토픽 모델의 학습 속도를 증진시킬 수 있으며, 대표적 잠재토픽 모델인 PLSA 모델의 뉴스 문서 집합에 대한 학습 실제 실험을 통해 유의미한 수준의 학습 속도 개선이 이루어짐을 확인할 수 있었다. 그리고 제안된 점진적 학습 기법과 병렬 학습 기법을 결합함으로써 모델 학습의 가속화 면에서 시너지 효과를 기대할 수 있음을 간단한 실험을 통해 제시하였다. 해당 결과는 PLSA 모델에 한정되지만 목표함수 및 학습 알고리즘 상에서 연관 있는 다른 잠재토픽모델들에 대해 비슷한 수준의 결과를 얻을 수 있을 것으로 기대된다. 구체적으로 NMF (non-negative matrix factorization)[8]와 LDA를 들 수 있다.

알고리즘의 특성상 점진적 EM 알고리즘은 E-Step에 소요되는 계산 비용이 기본 EM 알고리즘에서와 같기 때문에, 잠재변수 값의 추정 단계인 E-step이 전체 모델 학습 비용의 대부분을 차지하는 경우에 더욱 더 효과적일 것이다. 예를 들어 최근 발표된, 잠재 토픽들 간의 상관관계 모델링이 가능한 correlated topic model (CTM)[9]은 E-step에서 반복적 수치최적화(numerical optimization)에 의존하기 때문에 매개변수를 closed-form으로 계산 가능한 M-step에서의 시간 비용은 상대적으로 미미하다. 따라서 점진적 EM 알고리즘을 적용할 때 추가적인 M-step 비용은 그다지 크지 않을 것이다. 향후에는 이러한 특성을 지닌 잠재토픽 모델들에 대해 점진적 EM 알고리즘의 적용을 연구하고자 한다. 그리고 다른 측면에서 잠재토픽 모델의 병렬학습과 점진적 알고리즘과의 효과적인 결합 기법에 대해 심화된 연구를 진행하고자 한다. 학습의 병렬화에 관한 연구는, 최근 병렬컴퓨팅이 일반 데스크톱(desktop)에서도 가능할 정도로 일반화되어 가고 있는 환경에서 대규모 데이타의 효율적 처리를 위한 의미 있는 연구가 될 것이다.

참 고 문 헌

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

[2] D. Blei, A. Ng, and M. Jordan, Latent Dirichlet allocation, *Journal of Machine Learning Research*, **3**, pp. 993-1022, 2003.

[3] A. P. Dempster, N. Laird, and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm (with discussion), *Journal of the Royal Statistical Society, Series B*, **39**, pp. 1-38, 1977.

[4] R. Neal and G. Hinton, A View of the EM Algorithm that Justifies Incremental, Sparse, and other Variants. In Michael I. Jordan (editor), *Learning in Graphical Models*, pp. 355-368, MIT Press, Cambridge, MA., 1999.

[5] B. Thiesson, C. Meek and D. Heckerman, Accelerating EM for large databases, *Machine Learning*, **45**, pp. 279-299, 2001.

[6] 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**, pp. 361-397, 2004.

[7] H. W. Kuhn, The Hungarian method for the

assignment problem, *Naval Research Logistics Quarterly*, **2**, pp. 83-97, 1955.
[8] D. D. Lee and H. S. Seung, Algorithms for non-negative matrix factorization, In *Advances in Neural Information Processing Systems (Proc. NIPS 2000)*, **13**, pp. 556-562, MIT Press, Cambridge, MA., 2001.
[9] D. Blei and J. Lafferty, Correlated topic models, In *Advances in Neural Information Systems (Proc. NIPS 2005)*, **18**, pp. 147-154, MIT Press, Cambridge, MA., 2006.

장 정 호

1995년 서울대학교 컴퓨터공학과 학사
1997년 서울대학교 컴퓨터공학과 석사
2005년 서울대학교 컴퓨터공학부 박사
2005년 11월~2006년 6월 Fraunhofer IPSI 박사후 연수. 2006년 7월~2007년 11월 Fraunhofer IAIS 박사후 연수. 관심분야는 기계학습, 잠재변수모델, 텍스트마이닝, 생물정보학

이 종 우

1994년 서울대학교 컴퓨터공학과 학사
1996년 서울대학교 컴퓨터공학과 석사
2007년 현재, (주)코난테크놀로지 연구원. 관심분야는 기계학습, 텍스트마이닝, 이미지 인식

엄 재 홍

1999년 강원대학교 컴퓨터공학과 학사
2001년 서울대학교 전기·컴퓨터공학부 석사. 2007년 현재 서울대학교 전기·컴퓨터공학부 박사과정. 관심분야는 기계학습, 텍스트마이닝, 데이타마이닝, 생물정보학