



(19) 대한민국특허청(KR)  
(12) 공개특허공보(A)

(11) 공개번호 10-2020-0026351  
(43) 공개일자 2020년03월11일

(51) 국제특허분류(Int. Cl.)  
*G06F 16/00* (2019.01)

(52) CPC특허분류  
*G06F 16/36* (2019.01)  
*G06F 16/35* (2019.01)

(21) 출원번호 10-2018-0102046

(22) 출원일자 2018년08월29일

심사청구일자 2018년08월29일

(71) 출원인

동국대학교 산학협력단

서울특별시 중구 필동로1길 30 내 (필동3가, 동국대학교)

연세대학교 산학협력단

서울특별시 서대문구 연세로 50 (신촌동, 연세대학교)

(72) 발명자

이영섭

서울특별시 송파구 올림픽로 99 164동 1601호 (잠실동, 잠실엘스아파트)

박홍주

서울특별시 성동구 독서당로39길 38-23 (옥수동)

박태영

서울특별시 성동구 행당동 왕십리로 241, 102동 905호

(74) 대리인

특허법인이지

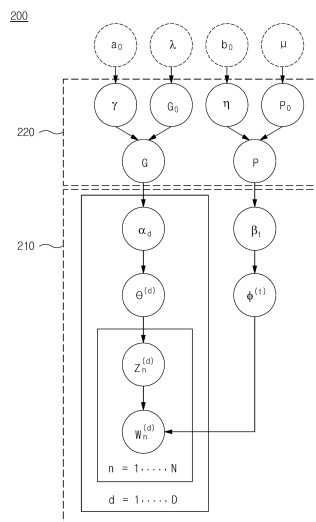
전체 청구항 수 : 총 11 항

(54) 발명의 명칭 향상된 잠재 디리클레 할당 모델을 이용한 토픽 분석 장치 및 방법

### (57) 요약

본 발명은 향상된 잠재 디리클레 할당 모델을 이용한 토픽 분석 방법에 관한 것으로, 토픽 분석 방법에 있어서, 문서에 포함된 텍스트 데이터를 전처리하는 전처리 단계; 및 잠재 디리클레 할당(Latent Dirichlet allocation, LDA)모델 구조에 계층적 디리클레 프로세스(Dirichlet Process, DP) 구조를 추가한 향상된 잠재 디리클레 할당 모델을 이용하여 상기 전처리된 텍스트 데이터에서 상기 토픽을 분석하는 단계;를 포함하되, 상기 향상된 잠재 디리클레 할당 모델에 대해 부분 붕괴된 깁스 샘플러(Partially Collapsed Gibbs Samplers, PCG)를 이용해 샘플링을 실행한다.

대표도 - 도2



(52) CPC특허분류  
*G06F 40/205* (2020.01)

이 발명을 지원한 국가연구개발사업

과제고유번호 1365002640

부처명 기상청

연구관리전문기관 한국기상산업기술원

연구사업명 기상·지진See-At기술개발연구

연구과제명 기상 관측·예보 분야의 비정형데이터 분석 기술 개발

기 여 율 1/1

주관기관 동국대학교 산학협력단

연구기간 2017.09.01 ~ 2018.08.31

---

## 명세서

### 청구범위

#### 청구항 1

토픽 분석 방법에 있어서,

문서에 포함된 텍스트 데이터를 전처리하는 전처리 단계; 및

잠재 디리클레 할당(Latent Dirichlet allocation, LDA)모델 구조에 계층적 디리클레 프로세스(Dirichlet Process, DP) 구조를 추가한 향상된 잠재 디리클레 할당 모델을 이용하여 상기 전처리된 텍스트 데이터에서 상기 토픽을 분석하는 단계;를 포함하되,

상기 향상된 잠재 디리클레 할당 모델에 대해 부분 붕괴된 깁스 샘플러(Partially Collapsed Gibbs Samplers, PCG)를 이용해 샘플링을 실행하는 것

을 특징으로 하는 향상된 잠재 디리클레 할당 모델을 이용한 토픽 분석 방법.

#### 청구항 2

제1항에 있어서,

상기 향상된 잠재 디리클레 할당 모델의 하이퍼파라미터(hyperparameter)인  $\alpha$  및  $\beta$ 는 사전에 지정된 상수가 아닌 것

을 특징으로 하는 향상된 잠재 디리클레 할당 모델을 이용한 토픽 분석 방법.

#### 청구항 3

제1항에 있어서,

상기 향상된 잠재 디리클레 할당 모델의 하이퍼파라미터인  $\alpha$  및  $\beta$ 는 자동으로 추정되는 것

을 특징으로 하는 향상된 잠재 디리클레 할당 모델을 이용한 토픽 분석 방법.

#### 청구항 4

제1항에 있어서,

상기 향상된 잠재 디리클레 할당 모델의 하이퍼파라미터인  $\alpha$  및  $\beta$ 는 디리클레 분포의 서로 다른 차원 파라미터인 것

을 특징으로 하는 향상된 잠재 디리클레 할당 모델을 이용하는 토픽 분석 방법.

#### 청구항 5

제1항에 있어서,

상기 향상된 잠재 디리클레 할당 모델은 이질적인 문서의 분석을 위해 상기 이질적인 문서의 텍스트 데이터를 군집화하여 각 군집마다 하이퍼파라미터를 할당하는 것

을 특징으로 하는 향상된 잠재 디리클레 할당 모델을 이용한 토픽 분석 방법.

## 청구항 6

문서에 포함된 텍스트 데이터를 전처리하는 전처리부;

상기 전처리된 텍스트 데이터를 잠재 디리클레 할당모델 구조에 계층적 디리클레 프로세스 구조를 추가한 향상된 잠재 디리클레 할당 모델을 이용하여 토픽을 분석하는 토픽 분석부; 및

상기 향상된 잠재 디리클레 할당 모델에 대해 부분 붕괴된 깁스 샘플러(Partially Collapsed Gibbs Samplers, PCG)를 이용해 샘플링을 실행하는 샘플링부;를 포함하는 것

을 특징으로 하는 향상된 잠재 디리클레 할당 모델을 이용한 토픽 분석 장치.

## 청구항 7

제6항에 있어서,

상기 향상된 잠재 디리클레 할당 모델의 하이퍼파라미터인  $\alpha$  및  $\beta$ 는 사전에 지정된 상수가 아닌 것

을 특징으로 하는 향상된 잠재 디리클레 할당 모델을 이용한 토픽 분석 장치.

## 청구항 8

제6항에 있어서,

상기 향상된 잠재 디리클레 할당 모델의 하이퍼파라미터인  $\alpha$  및  $\beta$ 는 자동으로 추정되는 것

을 특징으로 하는 향상된 잠재 디리클레 할당 모델을 이용한 토픽 분석 장치.

## 청구항 9

제6항에 있어서,

상기 향상된 잠재 디리클레 할당 모델의 하이퍼파라미터인  $\alpha$  및  $\beta$ 는 디리클레 분포의 서로 다른 차원 파라미터인 것

을 특징으로 하는 향상된 잠재 디리클레 할당 모델을 이용하는 토픽 분석 장치.

## 청구항 10

제6항에 있어서,

상기 향상된 잠재 디리클레 할당 모델은 이질적인 문서의 분석을 위해 상기 이질적인 문서의 텍스트 데이터를 군집화하여 각 군집마다 하이퍼파라미터를 할당하는 것

을 특징으로 하는 향상된 잠재 디리클레 할당 모델을 이용한 토픽 분석 장치.

## 청구항 11

제1항 내지 제5항 중 어느 한 항에 따른 향상된 잠재 디리클레 할당 모델을 이용한 토픽 분석 방법을 수행하기 위한 컴퓨터 프로그램이 기록된 컴퓨터로 판독 가능한 기록 매체.

**발명의 설명**

**기술 분야**

[0001] 본 발명은 향상된 잠재 디리클레 할당 모델을 이용한 토픽 분석 장치 및 방법에 관한 것이다.

## 배경 기술

[0003] 인터넷의 발달로 웹 문서 양이 급격하게 증가함에 따라, 인터넷에서 생성되는 수많은 대용량의 문서를 토픽별로 분류하는 토픽 분석 기술은 최근 가장 많은 주목을 받고 있는 분야로 이에 대한 연구가 활발히 진행되고 있다.

[0005] 특히, 토픽 분석 기술 중에서 잠재 디리클레 할당(Latent Dirichlet Allocation, LDA) 모델이 가장 널리 이용되고 있다.

[0007] 도 1은 종래 잠재 디리클레 할당 모델의 구조이다.

[0009] 도 1을 참조하면, 종래 잠재 디리클레 할당 모델은 하이퍼파라미터(hyperparameters)라고 불리는  $\alpha$  및  $\beta$  값을 사전에 지정해 주어야 하는데,  $\alpha$  및  $\beta$  값은 문서마다 다를 수 있기 때문에 잘못된  $\alpha$  및  $\beta$  값의 지정은 잘못된 분석을 야기하는 문제점이 있다.

[0011] 또한, 종래의 잠재 디리클레 할당 모델은 이질적인 문서의 텍스트 데이터를 동질적인 텍스트 데이터로 취급하여 이질적인 문서의 토픽을 분석하는데 한계를 보이는 문제점이 있다.

## 선행기술문헌

### 특허문헌

[0013] (특허문헌 0001) 한국등록특허 제10-1616544호(2016.04.28 공고)

## 발명의 내용

### 해결하려는 과제

[0014] 본 발명이 해결하고자 하는 기술적 과제는, 하이퍼파라미터 값을 자동으로 찾을 수 있는 향상된 잠재 디리클레 할당 모델을 이용한 토픽 분석 장치 및 방법을 제공하는 데 있다.

[0016] 본 발명이 해결하고자 하는 다른 기술적 과제는, 이질적인 문서의 토픽을 분석할 수 있는 향상된 잠재 디리클레 할당 모델을 이용한 토픽 분석 장치 및 방법을 제공하는 데 있다.

### 과제의 해결 수단

[0018] 상기와 같은 기술적 과제를 해결하기 위해, 본 발명의 바람직한 일 측면에 따르면, 토픽 분석 방법에 있어서, 문서에 포함된 텍스트 데이터를 전처리하는 전처리 단계; 및 잠재 디리클레 할당(Latent Dirichlet allocation, LDA)모델 구조에 계층적 디리클레 프로세스(Dirichlet Process, DP) 구조를 추가한 향상된 잠재 디리클레 할당 모델을 이용하여 상기 전처리된 텍스트 데이터에서 상기 토픽을 분석하는 단계;를 포함하되, 상기 향상된 잠재 디리클레 할당 모델에 대해 부분 붕괴된 깁스 샘플러(Partially Collapsed Gibbs Samplers, PCG)를 이용해 샘플링을 실행하는 향상된 잠재 디리클레 할당 모델을 이용한 토픽 분석 방법을 제공한다.

- [0020] 여기서, 상기 향상된 잠재 디리클레 할당 모델의 하이퍼파라미터(hyperparameter)인  $\alpha$  및  $\beta$ 는 사전에 지정된 상수가 아닐 수 있다.
- [0022] 여기서, 상기 향상된 잠재 디리클레 할당 모델의 하이퍼파라미터인  $\alpha$  및  $\beta$ 는 자동으로 추정될 수 있다.
- [0024] 여기서, 상기 향상된 잠재 디리클레 할당 모델의 하이퍼파라미터인  $\alpha$  및  $\beta$ 는 디리클레 분포의 서로 다른 차원 파라미터일 수 있다.
- [0026] 여기서, 상기 향상된 잠재 디리클레 할당 모델은 이질적인 문서의 분석을 위해 상기 이질적인 문서의 텍스트 데이터를 군집화하여 각 군집마다 하이퍼파라미터를 할당할 수 있다.
- [0028] 본 발명의 바람직한 다른 측면에 따르면, 문서에 포함된 텍스트 데이터를 전처리하는 전처리부; 상기 전처리된 텍스트 데이터를 잠재 디리클레 할당모델 구조에 계층적 디리클레 프로세스 구조를 추가한 향상된 잠재 디리클레 할당 모델을 이용하여 토픽을 분석하는 토픽 분석부; 및 상기 향상된 잠재 디리클레 할당 모델에 대해 부분 붕괴된 깁스 샘플러(Partially Collapsed Gibbs Samplers, PCG)를 이용해 샘플링을 실행하는 샘플링부;를 포함하는 향상된 잠재 디리클레 할당 모델을 이용한 토픽 분석 장치를 제공한다.
- [0030] 본 발명의 바람직한 또 다른 측면에 따르면, 상기 향상된 잠재 디리클레 할당 모델을 이용한 토픽 분석 방법을 수행하기 위한 컴퓨터 프로그램이 기록된 컴퓨터로 판독 가능한 기록 매체를 제공할 수 있다.

### 발명의 효과

- [0032] 본 발명은 향상된 잠재 디리클레 할당 모델을 이용한 토픽 분석 장치 및 방법을 통해 최적의 하이퍼파라미터 값을 자동으로 찾을 수 있는 효과가 있다.
- [0034] 또한, 본 발명은 향상된 잠재 디리클레 할당 모델을 이용한 토픽 분석 장치 및 방법을 통해 이질적인 문서의 토픽을 정교하게 분석할 수 있는 효과가 있다.

### 도면의 간단한 설명

- [0036] 도 1은 종래 잠재 디리클레 할당 모델의 구조이다.
- 도 2는 본 발명의 일 실시예에 따른 향상된 잠재 디리클레 할당 모델을 이용한 토픽 분석 방법에서 향상된 잠재 디리클레 할당 모델의 구조이다.
- 도 3은 본 발명의 일 실시예에 따른 향상된 잠재 디리클레 할당 모델을 이용한 토픽 분석 방법 순서도이다.
- 도 4는 본 발명의 일 실시예에 따른 향상된 잠재 디리클레 할당 모델을 이용한 토픽 분석 방법과 종래 방법의 문서에 대한 토픽 기여도의 사후 분포를 비교한 그림이다.
- 도 5는 본 발명의 일 실시예에 따른 향상된 잠재 디리클레 할당 모델을 이용한 토픽 분석 방법과 종래 다양한 방법의  $\theta^{(d)}$ 와  $\phi^{(t)}$ 에 대한 평균 제곱 오차(MSE)의 성능을 비교한 그래프이다.
- 도 6은 본 발명의 일 실시예에 따른 향상된 잠재 디리클레 할당 모델을 이용한 토픽 분석 방법과 종래 다양한 방법의 로그-가능도(log-likelihood) 및 퍼플렉시티(perplexity)와 관련한 시뮬레이션 성능을 비교한 그래프이다.

도 7은 본 발명의 일 실시예에 따른 향상된 잠재 디리클레 할당 모델을 이용한 토픽 분석 방법과 종래 다양한 방법의 실제 기사에 대한 로그-가능도 및 퍼플렉시티와 관련한 시뮬레이션 성능을 비교한 그래프이다.

도 8은 본 발명의 일 실시예에 따른 향상된 잠재 디리클레 할당 모델을 이용한 토픽 분석 방법과 종래 다양한 방법의 실제 기사에 대한 하이퍼파라미터  $\alpha_d$ 의 클러스터링과 관련한 시뮬레이션 성능을 비교한 그래프이다.

도 9는 본 발명의 다른 실시예에 따른 향상된 잠재 디리클레 할당 모델을 이용한 토픽 분석 장치의 구성도이다.

### 발명을 실시하기 위한 구체적인 내용

- [0037] 본 발명은 다양한 변경을 가할 수 있고 여러 가지 실시 예를 가질 수 있는바, 특정 실시 예들을 도면에 예시하고 상세한 설명에 상세하게 설명하고자 한다. 그러나 이는 본 발명을 특정한 실시 형태에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술범위에 포함되는 모든 변경, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다.
- [0038] 제1, 제2 등과 같이 서수를 포함하는 용어는 다양한 구성요소들을 설명하는데 사용될 수 있지만, 해당 구성요소들은 이와 같은 용어들에 의해 한정되지는 않는다. 이 용어들은 하나의 구성요소들을 다른 구성요소로부터 구별하는 목적으로만 사용된다.
- [0040] 어떤 구성요소가 다른 구성요소에 '연결되어' 있다거나, 또는 '접속되어' 있다고 언급된 때에는, 그 다른 구성요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 할 것이다. 반면에, 어떤 구성요소가 다른 구성요소에 '직접 연결되어' 있다거나, '직접 접속되어' 있다고 언급된 때에는, 중간에 다른 구성요소가 존재하지 않는 것으로 이해되어야 할 것이다.
- [0042] 본 출원에서 사용한 용어는 단지 특정한 실시예를 설명하기 위해 사용된 것으로, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 출원에서, '포함한다' 또는 '가지다' 등의 용어는 명세서상에 기재된 특징, 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.
- [0044] 도 2는 본 발명의 일 실시예에 따른 향상된 잠재 디리클레 할당 모델을 이용한 토픽 분석 방법에서 향상된 잠재 디리클레 할당 모델의 구조이다.
- [0046] 도 2를 참조하면, 향상된 잠재 디리클레 모델(200)의 구조는 잠재 디리클레 할당모델 구조(210)에 계층적 디리클레 프로세스(Dirichlet Process, DP) 구조(220)가 추가된 것이다.
- [0048] 여기서,  $N$ 은  $n$ 번째 문서의 단어 수,  $D$ 는 말뭉치 전체 문서 개수,  $w_n^{(d)}$ 는  $n$ 번째  $d$ 문서의 단어,  $z_n^{(d)}$ 는 문서  $d$ 의  $n$ 번째 단어에 대한 토픽 색인,  $\theta^{(d)}$ 는 매개변수  $\alpha_d$ 를 가진 디리클레 분포에 뒤따르는  $T$ 차원 무작위 변수로  $d$ 문서에서 토픽의 확률,  $\phi^{(t)}$ 는  $T \times M$  차원 행렬로 토픽  $t$ 에서 단어의 확률,  $\alpha_d$ 에서  $\alpha$ 는 하이퍼파라미터 중 하나인 디리클레 분포의  $T$ 차원 파라미터로 사전에 지정된 상수가 아닌  $\theta^{(d)}$ 의 디리클레 분포의 스칼라 파라미터,  $\beta_t$ 의  $\beta$ 는 하이퍼파라미터 중 하나인 디리클레 분포의  $M$ 차원 파라미터로 사전에 지정된 상수가 아닌  $\phi^{(t)}$ 의 디리클레 분포의 스칼라 파라미터,  $G$  및  $P$ 는  $\alpha$ 와  $\beta$  각각의 농도 파라미터(concentration parameters),  $G_0$  및  $P_0$ 는  $\alpha$ 와  $\beta$ 의 기본 분포(base measure),  $\gamma$  및  $\eta$ 는 디리클레 프로세스 이전 각각의  $\alpha$ 와  $\beta$ 이며,  $a_0 = 1$ ,  $b_0$

= 1이다.

[0050]  $\alpha$ 는 독립적이고 미지의 사전 분포  $G$ 에 따라 동일하게 분포되며,  $G$ 는 정밀도 매개 변수  $\gamma$  및  $\lambda$ 의 밀변 분포  $G_0$ 를 갖는 디리클레 프로세스로부터 유도되며,  $\beta$ 는 독립적이고 미지의 사전 분포  $P$ 에 따라 동일하게 분포되며,  $P$ 는 정밀도 매개 변수  $\eta$  및  $\mu$ 의 밀변 분포  $P_0$ 를 갖는 디리클레 프로세스로부터 유도되어,  $\alpha$  및  $\beta$ 는 자동으로 추정된다.

[0052] 향상된 잠재 디리클레 할당 모델(200)은 서로 문서 종류가 다른 이질적인 문서의 텍스트 데이터를 군집화하여 각 군집마다 하이퍼파라미터를 할당해 이질적인 문서의 분석을 할 수 있는데, 이는 하이퍼파라미터인  $\alpha$ 와  $\beta$ 가 사전에 지정된 상수가 아닌 사전 디리클레 분포의 디리클레 프로세스의 혼합물로 유도되기 때문이다.

[0054] 향상된 잠재 디리클레 할당 모델(200)의 목표 분포는  $p(Z, S, U, \Theta, \Phi, \alpha^*, \beta^*, \gamma, \eta|W)$ 이며, 목표 분포를 기반으로  $\Theta, \Phi$ 는  $Z$ 에 대한 조건부 분포에서 부분적으로 붕괴 될 수 있기 때문에 부분 붕괴된 깁스 샘플러 (Partially Collapsed Gibbs Samplers, PCG)를 실행해  $\Theta, \Phi$ 를 붕괴시켜 샘플링 한다. 여기서, 부분 붕괴된 깁스 샘플러는 공지된 기술이므로 자세한 설명은 생략하도록 한다.

[0056] 부분 붕괴된 깁스 샘플러의 샘플링은 변수추출을 위한 6개의 과정으로 구성되어 있다.

[0058] 과정 1은 확률로 분리된  $p(z_n^{(d)} | Z_{-(n,d)}, S, U, \alpha^*, \beta^*, \gamma, \eta, W)$ 로부터  $z_n^{(d)}$ 를 추출하는데, 추출식은 아래 식 1과 같다.

[0060] 식 1

$$p(z_n^{(d)} = t^* | Z_{-(n,d)}, S, U, \alpha^*, \beta^*, \gamma, \eta, W) \propto \frac{C_{-(n,d),t^*}^{(w_n^{(d)})} + \beta_{S_t^s}^*}{C_{-(n,d),t^*}^{(+)} + M\beta_{S_t^s}^*} \cdot \frac{C_{-(n,d),t^*}^{(d)} + \alpha_{S_t^s}^*}{C_{-(n,d),+}^{(d)} + T\alpha_{S_t^s}^*}, t^* = 1, \dots, T$$

[0063] 여기서,  $Z_{-(n,d)} = Z \setminus \{z_n^{(d)}\}$ 는 문서  $d$ 의  $n$  번째 단어를 제외한 토픽 색인의 집합을 나타내며,  $C_{-(n,d),t}^{(w_n^{(d)})}$ 는 단어  $w$ 가  $Z_{-(n,d)}$ 에서 토픽  $t$ 에 할당 된 횟수를 나타내고,  $C_{-(n,d),t}^{(d)}$ 는 문서  $d$  내의 단어가 토픽  $t$ 에 할당 된 횟수를

나타낸다.  $W$ 는 말뭉치,  $Z$ 는 토픽  $t$ 에서 단어  $n$ 이 나올 확률,  $S = \{S^\alpha, S^\beta\}$ ,  $S^\alpha = \{S_d^\alpha\}_{d=1}^D$ ,  $S^\beta = \{S_t^\beta\}_{t=1}^T$ ,  $\{S_d^\alpha\}=i$ 는  $\alpha_d$ 가 문서 클러스터  $i$ 에 속하는 것을,  $\{S_t^\beta\}=j$ 는  $\beta_t$ 가 토픽 클러스터  $t$ 에 속하는 것을 나타내며,  $U = \{U^\alpha, U^\beta\}$ ,  $U^\alpha = \{U_d^\alpha\}_{d=1}^D$ ,  $U^\beta = \{U_t^\beta\}_{t=1}^T$ ,  $M$ 은 고유 단어,  $T$ 는 잠재적인 토픽이다.



[0065] 과정 2는  $p(\Phi, \Theta | Z, S, U, \alpha^*, \beta^*, \gamma, \eta, W)$ 로부터  $(\Theta, \Phi)$ 를 추출하는데, 추출식은 아래 식 2 및 식 3과 같다. 식 2는 독립 T차원 디리클레 분포, 식 3은 T 독립 M 차원 디리클레 분포에 대한 것이다.

[0067] 식 2

$$[0068] \quad \Theta^{(d)} | (Z, S, U, \alpha^*, \beta^*, \gamma, \eta, W) \stackrel{\text{ind}}{\sim} \text{Dirichlet} \left( C_1^{(d)} + \alpha_{S_d^1}^*, \dots, C_T^{(d)} + \alpha_{S_d^T}^* \right), \quad d = 1, \dots, D$$

[0070] 식 3

$$[0071] \quad \Phi^{(t)} | (Z, S, U, \alpha^*, \beta^*, \gamma, \eta, W) \stackrel{\text{ind}}{\sim} \text{Dirichlet} \left( C_t^{(w_1)} + \beta_{S_t^1}^*, \dots, C_t^{(w_M)} + \beta_{S_t^M}^* \right), \quad t = 1, \dots, T$$

[0073] 여기서,  $C_t^{(d)}$  는 문서 d 내의 단어가 토픽 t에 할당 된 횟수를 나타내며,  $C_t^{(w)}$  는 단어 집합 w가 토픽 색인 집합 Z에서 토픽 t에 할당되는 횟수를 나타낸다.

[0075] 과정 3은  $p(S|Z, U, \Phi, \Theta, \alpha^*, \beta^*, \gamma, \eta, W)$ 로부터 S를 추출하는데, 추출식은 아래 식 4 및 식 5와 같다. 식 4는 독립 이산 분포 D, 식 5는 독립 이산 분포 T에 관한 것이다.

[0077] 식 4

$$[0078] \quad p(S_d^\alpha = i | Z, U, \Phi, \Theta, \alpha^*, \beta^*, \gamma, \eta, W) \propto U_i^\alpha \prod_{l < i} (1 - U_l^\alpha) \cdot \frac{\Gamma(T\alpha_i^*)}{\Gamma(\alpha_i^*)^T} \prod_{t=1}^T (\theta_t^{(d)})^{\alpha_i^* - 1}, \quad i = 1, \dots, I$$

[0080] 식 5

$$[0081] \quad p(S_t^\beta = j | Z, U, \Phi, \Theta, \alpha^*, \beta^*, \gamma, \eta, W) \propto U_j^\beta \prod_{l < j} (1 - U_l^\beta) \cdot \frac{\Gamma(M\beta_j^*)}{\Gamma(\beta_j^*)^M} \prod_{m=1}^M (\phi_m^{(t)})^{\beta_j^* - 1}, \quad j = 1, \dots, J$$

[0083] 과정 4는  $p(U|Z, S, \Phi, \Theta, \alpha^*, \beta^*, \gamma, \eta, W)$ 로부터 U를 추출하는데, 추출식은 아래 식 6 및 식 7과 같다. 식 6은 독립 베타 분포 I, 식 7은 독립 베타 분포 J에 관한 것이다.

[0085] 식 6

[0086]

$$U_i^\alpha | (Z, S, \Phi, \Theta, \alpha^*, \beta^*, \gamma, \eta, W) \stackrel{\text{ind}}{\sim} \text{Beta} \left( 1 + \sum_{d=1}^D \delta_i(S_d^\alpha), \gamma + \sum_{d=1}^D \sum_{k=i+1}^I \delta_k(S_d^\alpha) \right), \quad i = 1, \dots, I-1$$

[0088]

식 7

[0089]

$$U_j^\beta | (Z, S, \Phi, \Theta, \alpha^*, \beta^*, \gamma, \eta, W) \stackrel{\text{ind}}{\sim} \text{Beta} \left( 1 + \sum_{t=1}^T \delta_j(S_t^\beta), \eta + \sum_{t=1}^T \sum_{k=j+1}^J \delta_k(S_t^\beta) \right), \quad j = 1, \dots, J-1$$

[0091]

여기서,  $U_I^\alpha = 1$ ,  $U_J^\beta = 1$  이다.

[0093]

과정 5는  $p(\alpha^*, \beta^* | Z, S, U, \Phi, \Theta, \gamma, \eta, W)$ 로부터  $(\alpha^*, \beta^*)$ 를 추출하는데, 추출식은 아래 식 8과 같다.

[0095]

식 8

[0096]

$$p(\alpha^*, \beta^* | Z, S, U, \Phi, \Theta, \gamma, \eta, W) = \prod_{i=1}^I p(\alpha_i^* | Z, S, U, \phi, \theta, \gamma, \eta, W) \prod_{j=1}^J p(\beta_j^* | Z, S, U, \phi, \theta, \gamma, \eta, W)$$

이며,

[0098]

$\alpha_i^*$  는 식 9로 추출된다.

[0099]

식 9

[0100]

$$\left( \frac{\Gamma(T\alpha_i^*)}{\Gamma(\alpha_i^*)^T} \right)^{n_i^\alpha} \exp \left( -\alpha_i^* \left( \lambda - \sum_{d: S_d^\alpha = i} \sum_{t=1}^T \log \theta_t^{(d)} \right) \right)$$

[0102]

$\beta_j^*$  는 식 10으로 추출된다.

[0103]

식 10

[0104]

$$\left( \frac{\Gamma(M\beta_j^*)}{\Gamma(\beta_j^*)^M} \right)^{n_j^\beta} \exp \left( -\beta_j^* \left( \mu - \sum_{t: S_t^\beta = j} \sum_{m=1}^M \log \phi_m^{(t)} \right) \right)$$

$$n_i^\alpha = \sum_{d=1}^D \delta_i(S_d^\alpha), n_j^\beta = \sum_{t=1}^T \delta_j(S_t^\beta)$$

[0105] 여기서, 이다.

[0107] 과정 6은 독립적인 감마 분포의 산물을  $p(\gamma, \eta | Z, S, U, \Phi, \Theta, \alpha^*, \beta^*, W)$ 로부터  $(\gamma, \eta)$ 를 추출하는데, 추출식은 아래 식 11과 식 12와 같다.

[0109] 식 11

$$\gamma | (Z, S, U, \Phi, \Theta, \alpha^*, \beta^*, W) \sim \text{Gamma}(I, a_0 - \sum_{i=1}^{I-1} \log(1 - U_i^\alpha))$$

[0110]

[0112] 식 12

$$\eta | (Z, S, U, \Phi, \Theta, \alpha^*, \beta^*, W) \sim \text{Gamma}(J, b_0 - \sum_{j=1}^{J-1} \log(1 - U_j^\beta))$$

[0113]

[0115] 도 3은 본 발명의 일 실시예에 따른 향상된 잠재 디리클레 할당 모델을 이용한 토픽 분석 방법 순서도이다.

[0116]

[0117] 도 3을 참조하면, S310단계에서 문서에 포함된 텍스트 데이터를 전처리한다. 문서에 포함된 데이터는 텍스트 데이터이므로, 토픽을 분석하기 위해서는 문서의 전처리를 하여야 한다.

[0119] 구체적으로, 문서에 포함된 텍스트 데이터에서 문장을 분리하여 형태소별로 태깅할 수 있다. 형태소란, 뜻을 가진 가장 작은 말을 뜻한다. 또한, 형태소별로 태깅된 결과 중 명사만 추출할 수 있으며, 추출된 명사 중에서 불용어를 제거할 수 있다.

[0121] S320단계에서는 전처리된 텍스트 데이터를 향상된 잠재 디리클레 할당 모델을 이용하여 토픽을 분석한다. 즉, 전처리된 텍스트 데이터를 향상된 잠재 디리클레 할당 모델을 이용하여 토픽을 분석함으로써, 단어의 집합으로 표현할 수 있다.

[0123] 이때, 향상된 잠재 디리클레 할당 모델에 대해 부분 붕괴된 깃스 샘플러를 이용해 샘플링을 실행한다.

[0125] 본 발명의 일 실시예에 따른 향상된 잠재 디리클레 할당 모델을 이용한 토픽 분석 방법은 애플리케이션으로 구현되거나 다양한 컴퓨터 구성요소를 통하여 수행될 수 있는 프로그램 명령어의 형태로 구현되어 컴퓨터 판독 가능한 기록 매체에 기록될 수 있다.

[0127] 컴퓨터 판독 가능한 기록 매체는 프로그램 명령어, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 컴퓨터 판독 가능한 기록 매체에 기록되는 프로그램 명령어는, 본 발명을 위한 특별히 설계되고

구성된 것들이거나와 컴퓨터 소프트웨어 분야의 당업자에게 공지되어 사용 가능한 것일 수도 있다.

- [0129] 컴퓨터 판독 가능한 기록 매체의 예에는, 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체, CD-ROM, DVD와 같은 광기록 매체, 플롭티컬디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media) 및 ROM, RAM, 플래시 메모리 등과 같은 프로그램 명령어를 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다.
- [0131] 프로그램 명령어의 예에는, 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드도 포함된다. 하드웨어 장치는 본 발명에 따른 처리를 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.
- [0133] 도 4는 본 발명의 일 실시예에 따른 향상된 잠재 디리클레 할당 모델을 이용한 토픽 분석 방법과 종래 방법의 문서에 대한 토픽 기여도의 사후 분포를 비교한 그림이다. 도 4(a)는 문서에 대한 토픽 기여도의 실제 분포, 도 4(b)는 향상된 잠재 디리클레 할당 모델을 이용한 토픽 분석 방법에 기반한 토픽 기여도의 사후 분포, 도 4(c-1) 내지 도 4(c-9)는 종래 잠재 디리클레 할당 모델을 이용한 토픽 분석 방법의 토픽 기여도의 사후 분포 그림이다.
- [0135] 도 4는 100개의 문서로 구성된 말뭉치를 시물레이션한 결과로, 각 문서에는 평균 300개의 포아송(Poisson) 분포에서 생성된 단어 시퀀스가 포함되어 있으며, 100개의 고유 단어와 3개의 토픽을 가지고 있다. 단색점은 말뭉치의 모든 문서에 대해 예상되는 토픽 기여도를 나타낸다.
- [0137] 도 4를 참조하면, 그림 4(a)의 실제 분포는 매우 복잡적으로 나타났는데, 이는 도 4 (b)의 향상된 잠재 디리클레 할당 모델을 이용한 토픽 분석 방법에 의해 잘 예측된 것을 확인할 수 있다.
- [0139] 반면, 도 4(c-1) 내지 도 4(c-9)는 종래 잠재 디리클레 할당 모델을 이용한 토픽 분석 방법으로 하이퍼파라미터인  $\alpha$ 와  $\beta$ 를 사전에 지정된 상수로 사용하는 한편, 붕괴된 깃스 샘플러(Collapsed Gibbs Samplers, CG)를 이용했기 때문에,  $\alpha = 0.5$ 와  $\beta = 0.001$ 인 도 4(c-4)만 토픽 기여도 사후 분포가 명백하게 편향되지 않은 추정치를 산출하고 나머지 다른 경우는 단봉적(unimodal)으로 나타난 것을 확인할 수 있다.
- [0141] 이는 종래 잠재 디리클레 할당 모델을 이용한 토픽 분석 방법의 경우 이질적인 문서의 토픽 분석을 위한 멀티 모드 토픽 기여도 분포가 적절하게 매핑 되지 않을 수 있다는 것이다. 즉, 붕괴된 깃스 샘플러의 성능은 고정된 하이퍼파라미터 변수 선택에 따라 달라지기 때문에 큰 편향이 발생할 수 있다.
- [0143] 이와는 대조적으로, 도 4 (b)의 향상된 잠재 디리클레 할당 모델을 이용한 토픽 분석 방법은 토픽 기여도의 사후 분포를 유연하게 모델링하며, 단어 기여도의 분포에 유연성을 부여하고, 데이터가 자동으로 하이퍼파라미터 값을 추정하도록 함으로써 강력한 결과를 산출할 수 있다.
- [0145] 도 5는 본 발명의 일 실시예에 따른 향상된 잠재 디리클레 할당 모델을 이용한 토픽 분석 방법과 종래 다양한 방법의  $\theta^{(d)}$ 와  $\phi^{(t)}$ 에 대한 평균 제곱 오차(MSE)의 성능을 비교한 그래프이다.
- [0147] 도 6은 본 발명의 일 실시예에 따른 향상된 잠재 디리클레 할당 모델을 이용한 토픽 분석 방법과 종래 다양한 방법의 로그-가능도(log-likelihood) 및 퍼플렉시티(perplexity)와 관련한 시물레이션 성능을 비교한 그래프이

다.

- [0149] 도 5 및 도 6을 참조하면, 향상된 잠재 디리클레 할당 모델을 이용한 토픽 분석 방법(PCG)이 VB, CVB 및 종래 잠재 디리클레 할당 모델을 이용한 토픽 분석 방법(CG)과 비교해 평균 제곱 오차(MSE)와 퍼플렉시티(perplexity)가 낮고, 로그-가능도(log-likelihood)가 높아 상대적으로 우수한 성능을 보이는 것을 확인할 수 있다.
- [0151] 여기서, 도 5 및 도 6은 390개의 문서로 구성된 말뭉치를 시뮬레이션한 결과로, 각 문서에는 평균 1000개의 포아송 분포에서 생성된 단어 시퀀스가 포함되어 있으며, 1200개의 고유 단어와 10개의 토픽을 가지고 있다.
- [0153] 또한, 390개의 문서는 300개의 교육 자료와 90개의 테스트 문서로 나뉘며, 시뮬레이션은 100번 반복했다.
- [0155] VB 및 CVB 방법은 수렴 될 때까지 실행되었으며, VB는 Newton-Raphson 방법을 사용하여 하이퍼파라미터 값을 계산하고 CVB는  $\alpha = 0.05$ ,  $\beta = 0.005$ 를 사용했다.
- [0157] 종래 잠재 디리클레 할당 모델을 이용한 토픽 분석 방법(CG) 및 향상된 잠재 디리클레 할당 모델을 이용한 토픽 분석 방법(PCG)은 500번의 번인 반복으로 2000번의 반복 작업을 수행했다.
- [0159] 또한, 종래 잠재 디리클레 할당 모델을 이용한 토픽 분석 방법은 CG (0.01), CG (0.05), CG (0.2), CG (0.5)로 표현된  $\alpha = 0.01, 0.05, 0.2$  및  $0.5$ 를  $\beta = 0.005$ 로 설정했으며, 향상된 잠재 디리클레 할당 모델은  $\theta^{(d)}$ 와  $\phi^{(t)}$ 의 하이퍼파라미터에 대해 사전 디리클레 분포의 디리클레 프로세스의 혼합물로 유도되었고, 기본 분포  $G_0$  및  $P_0$ 는 각각 1과 10의 지수 분포로 설정했다.
- [0161] 도 7은 본 발명의 일 실시예에 따른 향상된 잠재 디리클레 할당 모델을 이용한 토픽 분석 방법과 종래 다양한 방법의 실제 기사에 대한 로그-가능도 및 퍼플렉시티와 관련한 시뮬레이션 성능을 비교한 그래프이다.
- [0163] 도 8은 본 발명의 일 실시예에 따른 향상된 잠재 디리클레 할당 모델을 이용한 토픽 분석 방법과 종래 다양한 방법의 실제 기사에 대한 하이퍼파라미터  $\alpha_d$ 의 클러스터링과 관련한 시뮬레이션 성능을 비교한 그래프이다.
- [0165] 여기서, 실제 기사는 문서의 토픽과 크기면에서 다양한 스펙트럼을 얻기 위해 토픽과 크기가 다른 기사 388 개로 구성된 2000-2002년의 신경 정보 처리 시스템(NIPS) 회의 데이터와 BBC 기사 중 2004-2005년 스포츠 및 기술 섹션의 912개 기사를 분석했다.
- [0167] 도 7을 참조하면, 향상된 잠재 디리클레 할당 모델을 이용한 토픽 분석 방법이 다른 방법과 비교해 가장 높은 로그-가능도를 보이는 것을 확인할 수 있다.
- [0169] 또한, 향상된 잠재 디리클레 할당 모델을 이용한 토픽 분석 방법이 다른 방법과 비교해 가장 작은 퍼플렉시티를 제공 하였지만, 토픽의 수가 증가함에 따라 다른 방법들과 차이는 줄어드는 것을 확인할 수 있다.

- [0171] 특히, 토픽의 수가 많을 때,  $\alpha = 0.05$  또는  $0.2$  인 종래 잠재 디리클레 할당 모델을 이용한 토픽 분석 방법이 향상된 잠재 디리클레 할당 모델을 이용한 토픽 분석 방법과 비슷한 성능을 보여  $\alpha = 0.05$  및  $0.2$ 는 최적에 가깝다는 것을 확인할 수 있는데, 이는 향상된 잠재 디리클레 할당 모델을 이용한 토픽 분석 방법이 자동으로 선택한 하이퍼파라미터  $\alpha$ 가 최적에 가깝다는 것을 알 수 있다.
- [0173] 도 8을 참조하면, 하이퍼파라미터 클러스터링은 사전 디리클레 혼합물에서 샘플링 된  $\alpha_d$ 개의 원자가 유사한 클러스터 크기를 갖는 배수인 것을 보여주는데, 즉  $\theta^{(d)}$ 의 추정 분포가 여러 개의 최빈값을 갖는 경향이 있음을 보여준다.
- [0175] 샘플링 된  $\alpha_d$ 의 가중 평균은  $0.1$ 와  $0.2$  사이이므로 종래 잠재 디리클레 할당 모델을 이용한 토픽 분석 방법의 최적  $\alpha$ 는 향상된 잠재 디리클레 할당 모델을 이용한 토픽 분석 방법의 가중 평균에 가깝고, 서로 일치해 향상된 잠재 디리클레 할당 모델을 이용한 토픽 분석 방법이 데이터에서 최적의 하이퍼파라미터를 자동으로 추정한다는 것을 확인할 수 있다.
- [0177] 도 9는 본 발명의 다른 실시예에 따른 향상된 잠재 디리클레 할당 모델을 이용한 토픽 분석 장치의 구성도이다.
- [0179] 도 9를 참조하면, 향상된 잠재 디리클레 할당 모델을 이용한 토픽 분석 장치(900)는 전처리부(910), 토픽 분석부(920) 및 샘플링부(930)를 포함한다.
- [0181] 전처리부(910)는 문서에 포함된 텍스트 데이터를 전처리한다. 문서에 포함된 데이터는 텍스트 데이터이므로, 토픽을 분석하기 위해서는 문서의 전처리를 하여야 한다. 즉, 전처리부(910)는 토픽을 분석하기 위해 텍스트 데이터를 전처리한다.
- [0183] 구체적으로, 전처리부(910)는 문서에 포함된 텍스트 데이터에서 문장을 분리하여 형태소별로 태깅 할 수 있다. 형태소란, 뜻을 가진 가장 작은 말을 뜻한다. 또한, 형태소별로 태깅된 결과 중 명사만 추출할 수 있으며, 추출된 명사 중에서 불용어를 제거할 수 있다.
- [0185] 토픽 분석부(920)는 전처리부(910)를 통해 전처리된 텍스트 데이터를 입력받아 향상된 잠재 디리클레 할당 모델을 이용하여 토픽을 분석한다. 즉, 토픽 분석부(920)가 전처리된 텍스트 데이터를 향상된 잠재 디리클레 할당 모델을 이용하여 토픽을 분석함으로써, 단어의 집합으로 표현할 수 있다.
- [0187] 이때, 샘플링부(930)는 향상된 잠재 디리클레 할당 모델에 대해 부분 붕괴된 깃스 샘플러를 이용해 샘플링을 실행한다.
- [0189] 이상에서 본 발명에 따른 실시 예들이 설명되었으나, 이는 예시적인 것에 불과하며, 본 발명의 속하는 기술분야에서 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 범위의 실시예가 가능하다는 점을 이해할 것이다. 따라서 본 발명의 진정한 기술적 보호범위는 다음의 청구범위에 의해서 정해져야 할 것이다.

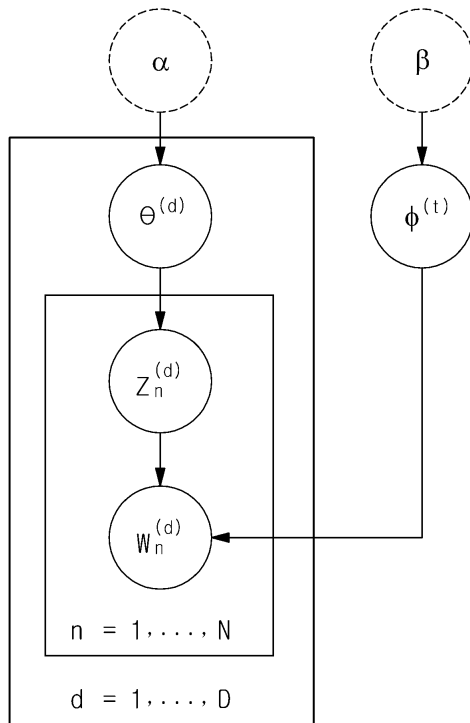
## 부호의 설명

- [0191] 910 : 전처리부    920 : 토픽 분석부

930 : 샘플링부

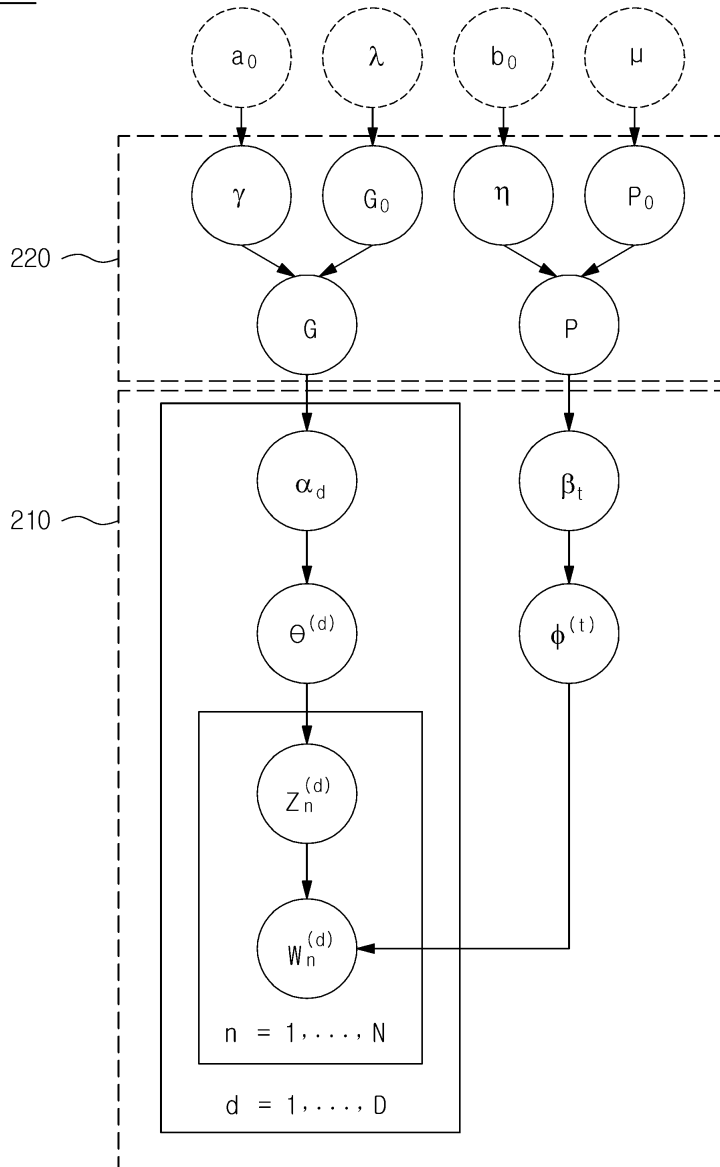
도면

도면1



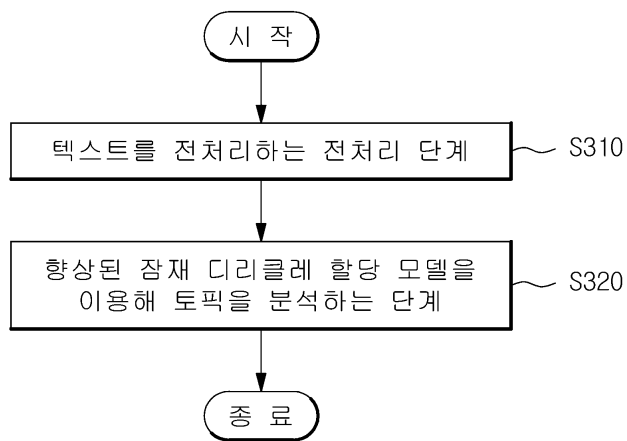
도면2

200

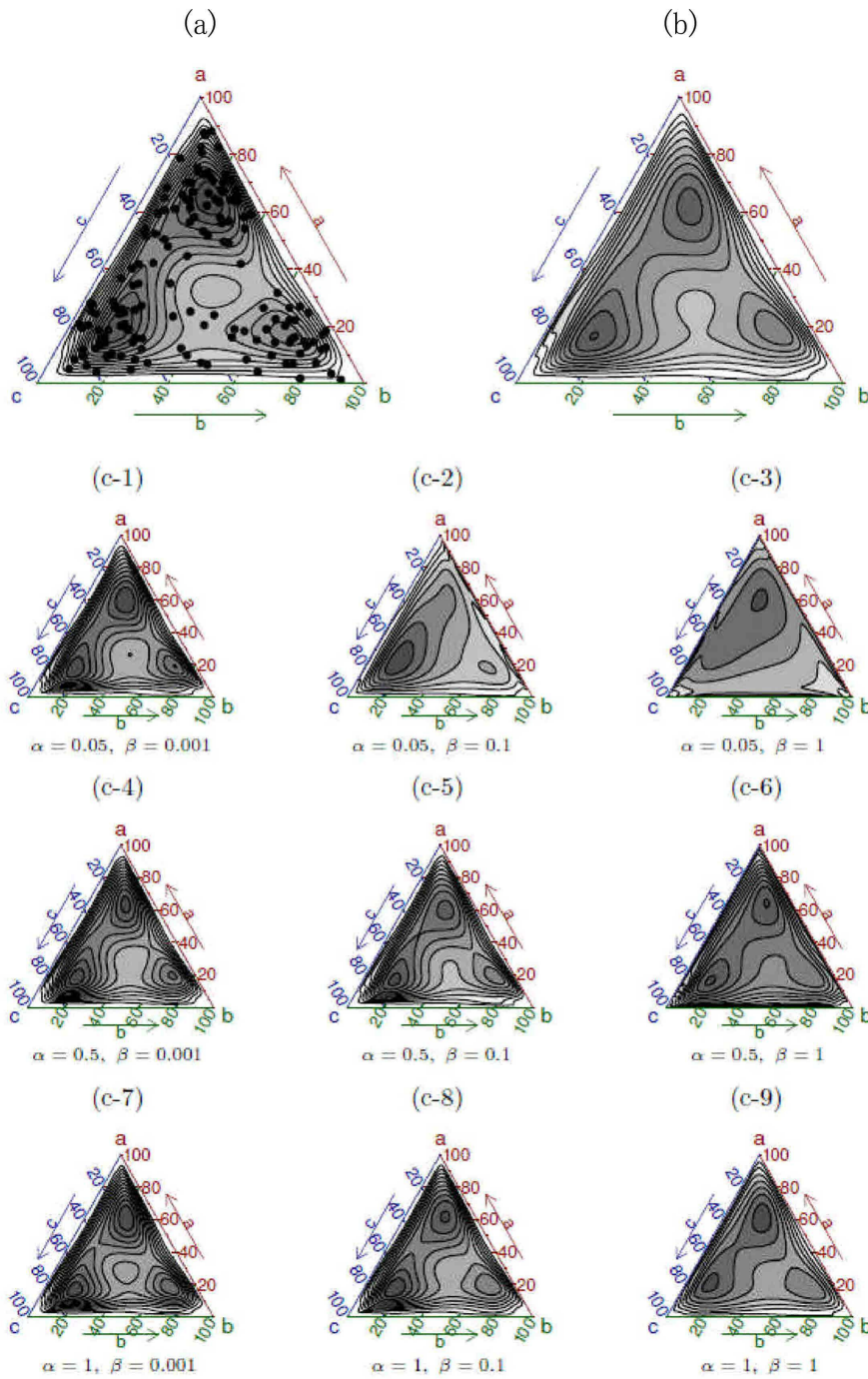




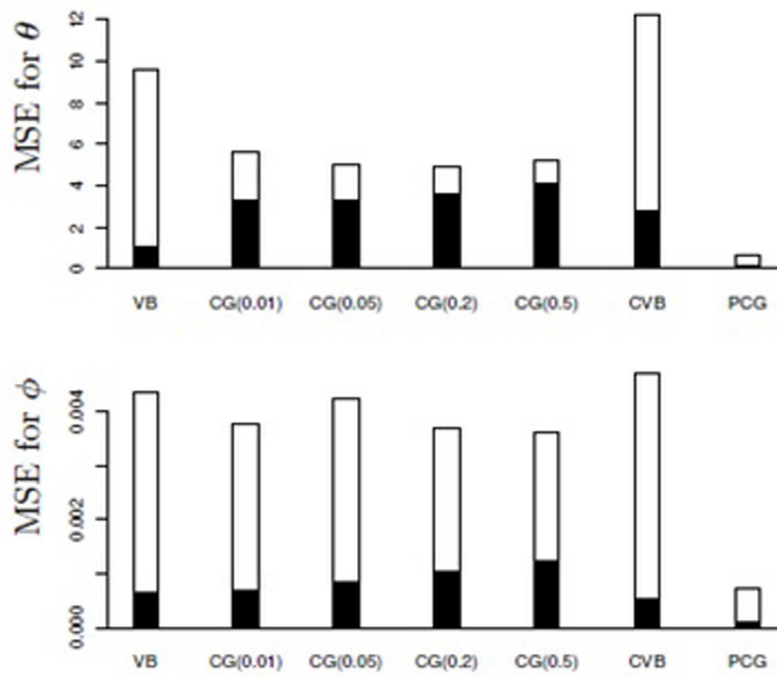
도면3



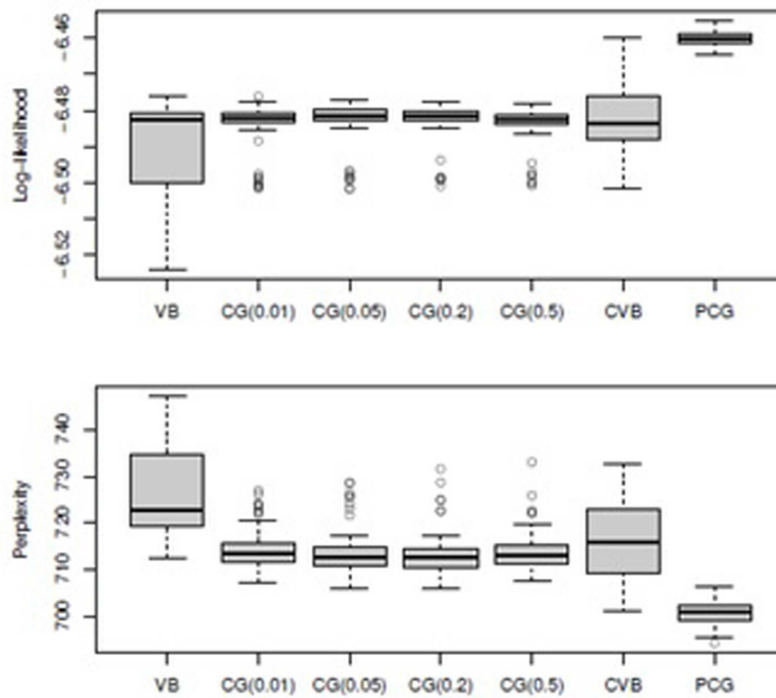
도면4



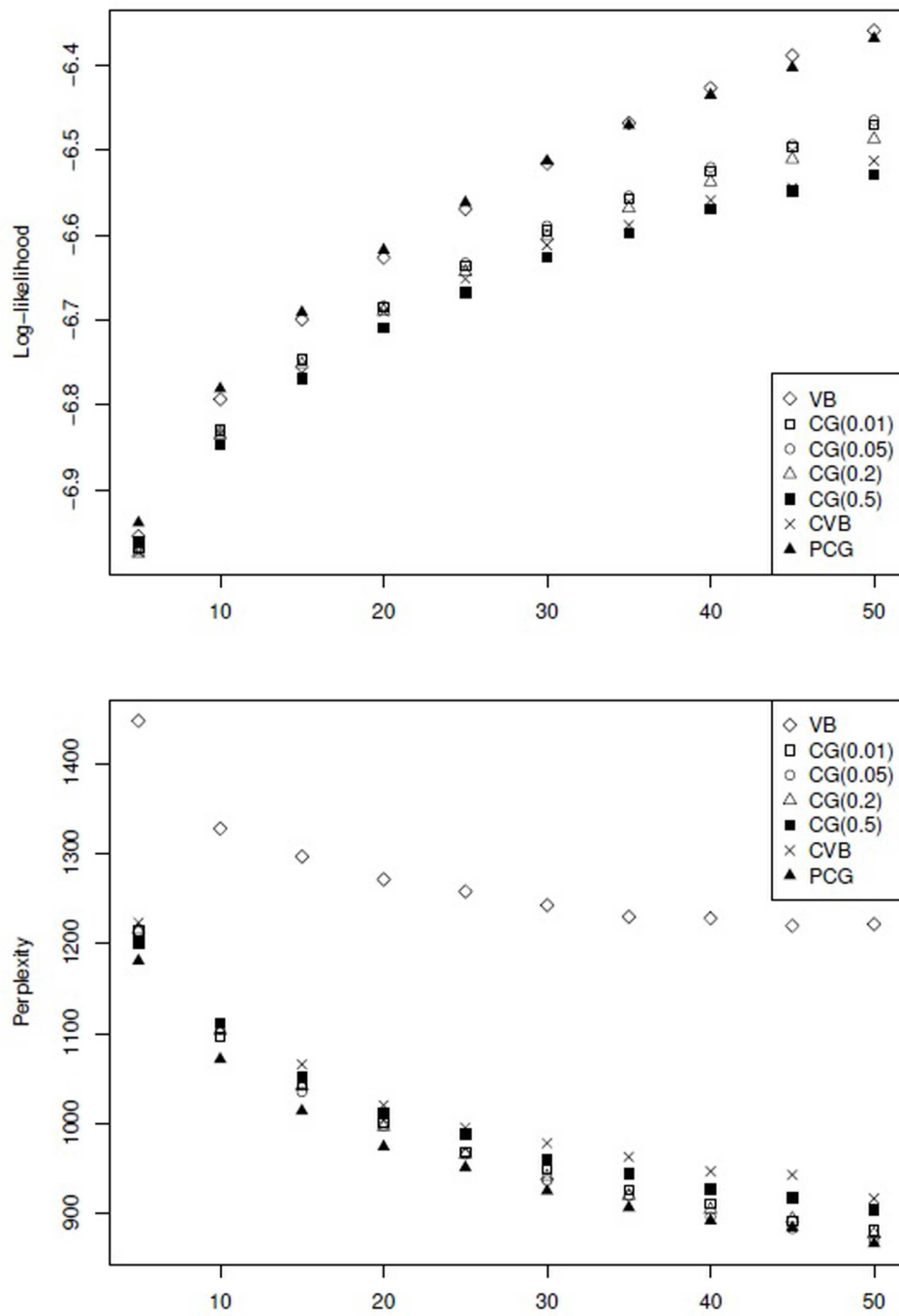
도면5



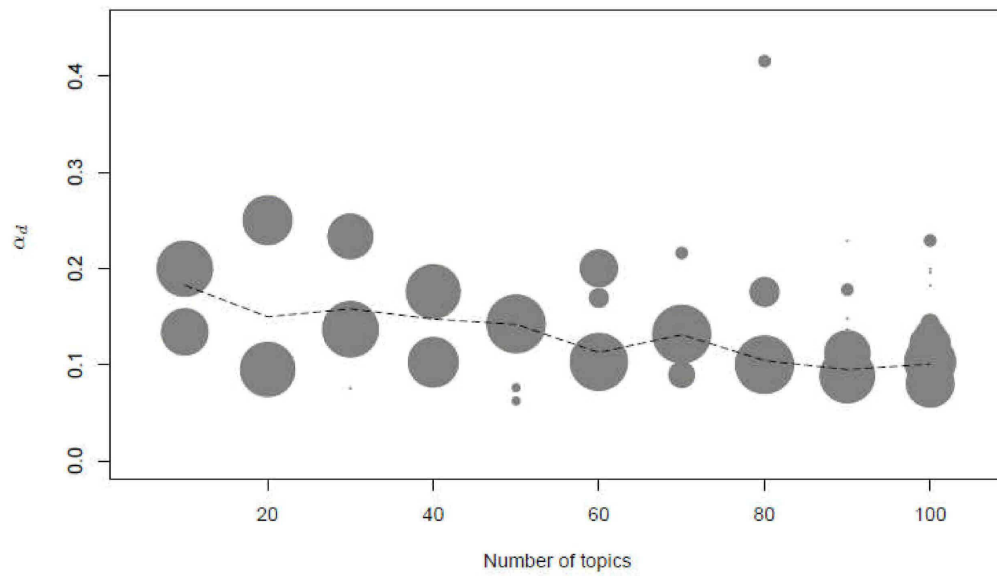
도면6



도면7



도면8



도면9

900

