



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2020-0078159
(43) 공개일자 2020년07월01일

- | | |
|---|--|
| <p>(51) 국제특허분류(Int. Cl.)
 <i>G06F 16/93</i> (2019.01) <i>G06F 16/35</i> (2019.01)
 <i>G06F 40/20</i> (2020.01)</p> <p>(52) CPC특허분류
 <i>G06F 16/93</i> (2019.01)
 <i>G06F 16/35</i> (2019.01)</p> <p>(21) 출원번호 10-2018-0167786
 (22) 출원일자 2018년12월21일
 심사청구일자 없음</p> | <p>(71) 출원인
 삼성전자주식회사
 경기도 수원시 영통구 삼성로 129 (매탄동)
 고려대학교 산학협력단
 서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)</p> <p>(72) 발명자
 강경필
 서울특별시 성북구 안암로 145(안암동5가)
 최성재
 서울특별시 성북구 안암로 145(안암동5가)
 (뒷면에 계속)</p> <p>(74) 대리인
 정홍식, 김태현</p> |
|---|--|

전체 청구항 수 : 총 14 항

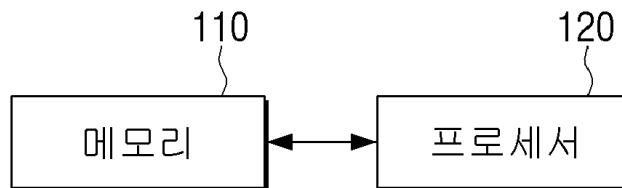
(54) 발명의 명칭 전자 장치 및 이의 대표 문서 선정 방법

(57) 요약

전자 장치가 개시된다. 본 전자 장치는 적어도 하나의 인스트럭션을 포함하는 메모리 및 메모리와 연결되어 전자 장치를 제어하는 프로세서를 포함하고, 프로세서는 적어도 하나의 인스트럭션을 실행함으로써, 토픽 모델링을 통해 문서 집합을 계층적인 구조를 갖는 복수의 토픽 노드에 따라 분할하고, 토픽 노드에 포함된 문서의 개수에 기초하여 복수의 토픽 노드 중에서 도미넌트 토픽 노드를 결정하고, 도미넌트 토픽 노드에 포함된 복수의 문서 중에서 도미넌트 토픽 노드에 포함된 복수의 문서로부터 획득된 복수의 형태소에 기초하여 결정된 하나의 문서를 문서 집합의 대표 문서로 결정한다.

대표도 - 도1

100



(52) CPC특허분류

G06F 40/268 (2020.01)

(72) 발명자

주재걸

서울특별시 성북구 안암로 145(안암동5가)

이형욱

경기도 수원시 영통구 삼성로 129(매탄동)

강민구

경기도 수원시 영통구 삼성로 129(매탄동)

박동훈

경기도 수원시 영통구 삼성로 129(매탄동)

서봉주

경기도 수원시 영통구 삼성로 129(매탄동)

정윤성

경기도 수원시 영통구 삼성로 129(매탄동)

명세서

청구범위

청구항 1

전자 장치에 있어서,

적어도 하나의 인스트럭션(instruction)을 포함하는 메모리; 및

상기 메모리와 연결되어 상기 전자 장치를 제어하는 프로세서;를 포함하고,

상기 프로세서는, 상기 적어도 하나의 인스트럭션을 실행함으로써,

토픽 모델링을 통해 문서 집합을 계층적인 구조를 갖는 복수의 토픽 노드에 따라 분할하고, 토픽 노드에 포함된 문서의 개수에 기초하여 상기 복수의 토픽 노드 중에서 도미넌트(dominant) 토픽 노드를 결정하고, 상기 도미넌트 토픽 노드에 포함된 복수의 문서 중에서 상기 도미넌트 토픽 노드에 포함된 복수의 문서로부터 획득된 복수의 형태소에 기초하여 결정된 하나의 문서를 상기 문서 집합의 대표 문서로 결정하는, 전자 장치.

청구항 2

제1항에 있어서,

상기 프로세서는,

상기 문서 집합에 포함된 복수의 문서에 기초하여 획득된 텀-도큐먼트(term-document) 매트릭스에 비음수 행렬 분해(Nonnegative Matrix Factorization, NMF)를 수행하여, 상기 문서 집합에 대응되는 노드를 기설정된 개수의 제1 토픽 노드들로 분할하며,

상기 텀-도큐먼트 매트릭스는, 상기 문서 집합에 포함된 복수의 문서에서 획득된 복수의 형태소가 상기 복수의 문서 각각에 포함된 빈도수에 기초하여 생성되는, 전자 장치.

청구항 3

제2항에 있어서,

상기 프로세서는,

상기 제1 토픽 노드들을 각각 분할하였을 때 생성되는 자식 토픽 노드들에 기초하여 상기 제1 토픽 노드들 각각의 스코어를 산출하고, 상기 산출된 스코어에 기초하여 상기 제1 토픽 노드들 중에서 분할될 토픽 노드를 결정하고, 상기 결정된 토픽 노드를 기설정된 개수의 제2 토픽 노드들로 분할하는, 전자 장치.

청구항 4

제3항에 있어서,

상기 프로세서는,

상기 제1 토픽 노드들 중 분할되지 않는 토픽 노드, 및 상기 제2 토픽 노드들을 각각 분할하였을 때 생성되는 자식 토픽 노드들에 기초하여 상기 분할되지 않는 토픽 노드 및 상기 제2 토픽 노드들 각각의 스코어를 산출하고, 상기 산출된 스코어에 기초하여 상기 분할되지 않는 토픽 노드 및 상기 제2 토픽 노드들 중에서 분할될 토픽 노드를 결정하고, 상기 결정된 토픽 노드를 기설정된 개수의 제3 토픽 노드들로 분할하는, 전자 장치.

청구항 5

제4항에 있어서,

상기 프로세서는,

기설정된 조건을 만족하는 토픽 노드를 제외하면서 상기 분할 과정을 수행하여 상기 계층적인 구조를 갖는 복수의 토픽 노드를 생성하며,

상기 기설정된 조건은,

토픽 노드를 분할하였을 때 생성되는 자식 토픽 노드들에 기초하여 산출된 스코어가 음수인 조건, 및 토픽 노드 및 자식 토픽 노드 간 또는 자식 토픽 노드들 간에 기설정된 개수 이상의 동일한 빈도수 상위 형태소가 존재하는 조건을 포함하는, 전자 장치.

청구항 6

제1항에 있어서,

상기 프로세서는,

상기 계층적인 구조를 갖는 복수의 토픽 노드 중에서 리프(leaf) 노드에 해당하는 복수의 토픽 노드 각각에 포함된 문서의 개수를 판단하고, 상기 리프 노드에 해당하는 복수의 토픽 노드 중에서 문서의 개수가 가장 많은 토픽 노드를 상기 도미넌트 토픽 노드로 결정하는, 전자 장치.

청구항 7

제1항에 있어서,

상기 프로세서는,

상기 도미넌트 토픽 노드에 포함된 복수의 문서로부터 획득된 복수의 형태소가 상기 복수의 문서 각각에 포함된 빈도수 및 상기 도미넌트 토픽 노드에서의 상기 복수의 형태소의 가중치에 기초하여 상기 도미넌트 토픽 노드에 포함된 복수의 문서의 스코어를 산출하고, 상기 산출된 스코어에 기초하여 상기 도미넌트 토픽 노드에 포함된 복수의 문서 중에서 하나의 문서를 선택하고, 상기 선택된 문서를 상기 대표 문서로 결정하는, 전자 장치.

청구항 8

전자 장치의 대표 문서 선정 방법에 있어서,

토픽 모델링을 통해 문서 집합을 계층적인 구조를 갖는 복수의 토픽 노드에 따라 분할하는 단계;

토픽 노드에 포함된 문서의 개수에 기초하여 상기 복수의 토픽 노드 중에서 도미넌트(dominant) 토픽 노드를 결정하는 단계; 및

상기 도미넌트 토픽 노드에 포함된 복수의 문서 중에서 상기 도미넌트 토픽 노드에 포함된 복수의 문서로부터 획득된 복수의 형태소에 기초하여 결정된 하나의 문서를 상기 문서 집합의 대표 문서로 결정하는 단계;를 포함하는, 대표 문서 선정 방법.

청구항 9

제8항에 있어서,

상기 분할하는 단계는,

상기 문서 집합에 포함된 복수의 문서에 기초하여 획득된 텀-도큐먼트(term-document) 매트릭스에 비음수 행렬 분해(Nonnegative Matrix Factorization, NMF)를 수행하여, 상기 문서 집합에 대응되는 노드를 기설정된 개수의 제1 토픽 노드들로 분할하며,

상기 텀-도큐먼트 매트릭스는, 상기 문서 집합에 포함된 복수의 문서에서 획득된 복수의 형태소가 상기 복수의 문서 각각에 포함된 빈도수에 기초하여 생성되는, 대표 문서 선정 방법.

청구항 10

제9항에 있어서,

상기 분할하는 단계는,

상기 제1 토픽 노드들을 각각 분할하였을 때 생성되는 자식 토픽 노드들에 기초하여 상기 제1 토픽 노드들 각각의 스코어를 산출하고, 상기 산출된 스코어에 기초하여 상기 제1 토픽 노드들 중에서 분할될 토픽 노드를 결정하고, 상기 결정된 토픽 노드를 기설정된 개수의 제2 토픽 노드들로 분할하는, 대표 문서 선정 방법.

청구항 11

제10항에 있어서,

상기 분할하는 단계는,

상기 제1 토픽 노드들 중 분할되지 않는 토픽 노드, 및 상기 제2 토픽 노드들을 각각 분할하였을 때 생성되는 자식 토픽 노드들에 기초하여 상기 분할되지 않는 토픽 노드 및 상기 제2 토픽 노드들 각각의 스코어를 산출하고, 상기 산출된 스코어에 기초하여 상기 분할되지 않는 토픽 노드 및 상기 제2 토픽 노드들 중에서 분할될 토픽 노드를 결정하고, 상기 결정된 토픽 노드를 기설정된 개수의 제3 토픽 노드들로 분할하는, 대표 문서 선정 방법.

청구항 12

제11항에 있어서,

상기 분할하는 단계는,

기설정된 조건을 만족하는 토픽 노드를 제외하면서 상기 분할 과정을 수행하여 상기 계층적인 구조를 갖는 복수의 토픽 노드를 생성하며,

상기 기설정된 조건은,

토픽 노드를 분할하였을 때 생성되는 자식 토픽 노드들에 기초하여 산출된 스코어가 음수인 조건, 및 토픽 노드 및 자식 토픽 노드 간 또는 자식 토픽 노드들 간에 기설정된 개수 이상의 동일한 빈도수 상위 형태소가 존재하는 조건을 포함하는, 대표 문서 선정 방법.

청구항 13

제8항에 있어서,

상기 도미넌트 토픽 노드를 결정하는 단계는,

상기 계층적인 구조를 갖는 복수의 토픽 노드 중에서 리프(leaf) 노드에 해당하는 복수의 토픽 노드 각각에 포함된 문서의 개수를 판단하고, 상기 리프 노드에 해당하는 복수의 토픽 노드 중에서 문서의 개수가 가장 많은 토픽 노드를 상기 도미넌트 토픽 노드로 결정하는, 대표 문서 선정 방법.

청구항 14

제8항에 있어서,

상기 대표 문서를 결정하는 단계는,

상기 도미넌트 토픽 노드에 포함된 복수의 문서로부터 획득된 복수의 형태소가 상기 복수의 문서 각각에 포함된 빈도수 및 상기 도미넌트 토픽 노드에서의 상기 복수의 형태소의 가중치에 기초하여 상기 도미넌트 토픽 노드에 포함된 복수의 문서의 스코어를 산출하고, 상기 산출된 스코어에 기초하여 상기 도미넌트 토픽 노드에 포함된 복수의 문서 중에서 하나의 문서를 선택하고, 상기 선택된 문서를 상기 대표 문서로 결정하는, 대표 문서 선정 방법.

발명의 설명

기술 분야

[0001] 본 개시는 전자 장치 및 이의 대표 문서 선정 방법에 관한 것으로, 더욱 상세하게는 복수의 문서에서 대표 문서를 결정하는 전자 장치 및 이의 대표 문서 선정 방법에 관한 것이다.

배경 기술

[0002] 전자 기술의 발전으로, 사용자들은 인터넷 등을 통해 기사 등을 접할 수 있게 되었다.

[0003] 하지만, 사용자에게 제공되는 기사의 양이 방대한 관계로, 사용자는 이들 기사 중에서 대표가 되는 기사를 찾기 위해 많은 시간을 소모하고 있다.

[0004] 이에 따라, 전체 기사 중에서 대표성 있는 기사를 선별하여, 사용자에게 제공하기 위한 방안의 모색이 요청된다.

발명의 내용

해결하려는 과제

[0005] 본 개시는 상술한 문제점을 해결하기 위해 안출된 것으로, 본 개시의 목적은 복수의 문서를 구분하기 위해 토픽 모델링을 통해 복수의 토픽 노드를 생성하고, 이들 토픽 노드들 중에서 도미넌트 토픽 노드를 판단하고, 도미넌트 토픽 노드에 포함된 복수의 문서 중에서 대표 문서를 결정하는 전자 장치 및 이의 대표 문서 선정 방법을 제공함에 있다.

과제의 해결 수단

[0006] 본 개시의 일 실시예에 따른, 전자 장치는, 적어도 하나의 인스트럭션을 포함하는 메모리 및 상기 메모리와 연결되어 상기 전자 장치를 제어하는 프로세서를 포함하고, 상기 프로세서는 상기 적어도 하나의 인스트럭션을 실행함으로써, 토픽 모델링을 통해 문서 집합을 계층적인 구조를 갖는 복수의 토픽 노드에 따라 분할하고, 토픽 노드에 포함된 문서의 개수에 기초하여 상기 복수의 토픽 노드 중에서 도미넌트 토픽 노드를 결정하고, 상기 도미넌트 토픽 노드에 포함된 복수의 문서 중에서 상기 도미넌트 토픽 노드에 포함된 복수의 문서로부터 획득된 복수의 형태소에 기초하여 결정된 하나의 문서를 상기 문서 집합의 대표 문서로 결정한다.

[0007] 이 경우, 상기 프로세서는 상기 문서 집합에 포함된 복수의 문서에 기초하여 획득된 텀-도큐먼트 매트릭스에 비음수 행렬 분해를 수행하여, 상기 문서 집합에 대응되는 노드를 기설정된 개수의 제1 토픽 노드들로 분할하며, 상기 텀-도큐먼트 매트릭스는 상기 문서 집합에 포함된 복수의 문서에서 획득된 복수의 형태소가 상기 복수의 문서 각각에 포함된 빈도수에 기초하여 생성될 수 있다.

[0008] 또한, 상기 프로세서는 상기 제1 토픽 노드들을 각각 분할하였을 때 생성되는 자식 토픽 노드들에 기초하여 상기 제1 토픽 노드들 각각의 스코어를 산출하고, 상기 산출된 스코어에 기초하여 상기 제1 토픽 노드들 중에서 분할될 토픽 노드를 결정하고, 상기 결정된 토픽 노드를 기설정된 개수의 제2 토픽 노드들로 분할할 수 있다.

[0009] 그리고, 상기 프로세서는 상기 제1 토픽 노드들 중 분할되지 않는 토픽 노드, 및 상기 제2 토픽 노드들을 각각 분할하였을 때 생성되는 자식 토픽 노드들에 기초하여 상기 분할되지 않는 토픽 노드 및 상기 제2 토픽 노드들 각각의 스코어를 산출하고, 상기 산출된 스코어에 기초하여 상기 분할되지 않는 토픽 노드 및 상기 제2 토픽 노드들 중에서 분할될 토픽 노드를 결정하고, 상기 결정된 토픽 노드를 기설정된 개수의 제3 토픽 노드들로 분할할 수 있다.

[0010] 또한, 상기 프로세서는 기설정된 조건을 만족하는 토픽 노드를 제외하면서 상기 분할 과정을 수행하여 상기 계층적인 구조를 갖는 복수의 토픽 노드를 생성하며, 상기 기설정된 조건은 토픽 노드를 분할하였을 때 생성되는 자식 토픽 노드들에 기초하여 산출된 스코어가 음수인 조건, 및 토픽 노드 및 자식 토픽 노드 간 또는 자식 토픽 노드들 간에 기설정된 개수 이상의 동일한 빈도수 상위 형태소가 존재하는 조건을 포함할 수 있다.

[0011] 한편, 상기 프로세서는 상기 계층적인 구조를 갖는 복수의 토픽 노드 중에서 리프 노드에 해당하는 복수의 토픽 노드 각각에 포함된 문서의 개수를 판단하고, 상기 리프 노드에 해당하는 복수의 토픽 노드 중에서 문서의 개수가 가장 많은 토픽 노드를 상기 도미넌트 토픽 노드로 결정할 수 있다.

[0012] 또한, 상기 프로세서는 상기 도미넌트 토픽 노드에 포함된 복수의 문서로부터 획득된 복수의 형태소가 상기 복수의 문서 각각에 포함된 빈도수 및 상기 도미넌트 토픽 노드에서의 상기 복수의 형태소의 가중치에 기초하여 상기 도미넌트 토픽 노드에 포함된 복수의 문서의 스코어를 산출하고, 상기 산출된 스코어에 기초하여 상기 도미넌트 토픽 노드에 포함된 복수의 문서 중에서 하나의 문서를 선택하고, 상기 선택된 문서를 상기 대표 문서로 결정할 수 있다.

[0013] 한편, 본 개시의 일 실시예에 따른, 전자 장치의 대표 문서 선정 방법은 토픽 모델링을 통해 문서 집합을 계층적인 구조를 갖는 복수의 토픽 노드에 따라 분할하는 단계, 토픽 노드에 포함된 문서의 개수에 기초하여 상기 복수의 토픽 노드 중에서 도미넌트 토픽 노드를 결정하는 단계 및 상기 도미넌트 토픽 노드에 포함된 복수의 문

서 중에서 상기 도미넌트 토픽 노드에 포함된 복수의 문서로부터 획득된 복수의 형태소에 기초하여 결정된 하나의 문서를 상기 문서 집합의 대표 문서로 결정하는 단계를 포함한다.

[0014] 이 경우, 상기 분할하는 단계는 상기 문서 집합에 포함된 복수의 문서에 기초하여 획득된 텀-도큐먼트 매트릭스에 비음수 행렬 분해를 수행하여, 상기 문서 집합에 대응되는 노드를 기설정된 개수의 제1 토픽 노드들로 분할하며, 상기 텀-도큐먼트 매트릭스는 상기 문서 집합에 포함된 복수의 문서에서 획득된 복수의 형태소가 상기 복수의 문서 각각에 포함된 빈도수에 기초하여 생성될 수 있다.

[0015] 또한, 상기 분할하는 단계는 상기 제1 토픽 노드들을 각각 분할하였을 때 생성되는 자식 토픽 노드들에 기초하여 상기 제1 토픽 노드들 각각의 스코어를 산출하고, 상기 산출된 스코어에 기초하여 상기 제1 토픽 노드들 중에서 분할될 토픽 노드를 결정하고, 상기 결정된 토픽 노드를 기설정된 개수의 제2 토픽 노드들로 분할할 수 있다.

[0016] 그리고, 상기 분할하는 단계는 상기 제1 토픽 노드들 중 분할되지 않는 토픽 노드, 및 상기 제2 토픽 노드들을 각각 분할하였을 때 생성되는 자식 토픽 노드들에 기초하여 상기 분할되지 않는 토픽 노드 및 상기 제2 토픽 노드들 각각의 스코어를 산출하고, 상기 산출된 스코어에 기초하여 상기 분할되지 않는 토픽 노드 및 상기 제2 토픽 노드들 중에서 분할될 토픽 노드를 결정하고, 상기 결정된 토픽 노드를 기설정된 개수의 제3 토픽 노드들로 분할할 수 있다.

[0017] 또한, 상기 분할하는 단계는 기설정된 조건을 만족하는 토픽 노드를 제외하면서 상기 분할 과정을 수행하여 상기 계층적인 구조를 갖는 복수의 토픽 노드를 생성하며, 상기 기설정된 조건은 토픽 노드를 분할하였을 때 생성되는 자식 토픽 노드들에 기초하여 산출된 스코어가 음수인 조건, 및 토픽 노드 및 자식 토픽 노드 간 또는 자식 토픽 노드들 간에 기설정된 개수 이상의 동일한 빈도수 상위 형태소가 존재하는 조건을 포함할 수 있다.

[0018] 한편, 상기 도미넌트 토픽 노드를 결정하는 단계는 상기 계층적인 구조를 갖는 복수의 토픽 노드 중에서 리프 노드에 해당하는 복수의 토픽 노드 각각에 포함된 문서의 개수를 판단하고, 상기 리프 노드에 해당하는 복수의 토픽 노드 중에서 문서의 개수가 가장 많은 토픽 노드를 상기 도미넌트 토픽 노드로 결정할 수 있다.

[0019] 또한, 상기 대표 문서를 결정하는 단계는 상기 도미넌트 토픽 노드에 포함된 복수의 문서로부터 획득된 복수의 형태소가 상기 복수의 문서 각각에 포함된 빈도수 및 상기 도미넌트 토픽 노드에서의 상기 복수의 형태소의 가중치에 기초하여 상기 도미넌트 토픽 노드에 포함된 복수의 문서의 스코어를 산출하고, 상기 산출된 스코어에 기초하여 상기 도미넌트 토픽 노드에 포함된 복수의 문서 중에서 하나의 문서를 선택하고, 상기 선택된 문서를 상기 대표 문서로 결정할 수 있다.

발명의 효과

[0020] 이상과 같은 본 개시의 다양한 실시 예에 따르면, 문서 집합에 포함된 복수의 문서를 대표하는 하나의 문서를 효과적으로 선정할 수 있게 된다.

도면의 간단한 설명

[0021] 도 1은 본 개시의 일 실시 예에 따른 전자 장치의 블록도를 설명하기 위한 도면,
 도 2는 본 개시의 일 실시 예에 따라 계층적인 구조를 갖는 복수의 토픽 노드를 생성하는 방법을 설명하기 위한 도면,
 도 3은 본 개시의 일 실시 예에 따라 계층적인 구조를 갖는 복수의 토픽 노드를 생성하는 방법을 설명하기 위한 도면,
 도 4는 본 개시의 일 실시 예에 따라 계층적인 구조를 갖는 복수의 토픽 노드를 생성하는 방법을 설명하기 위한 도면,
 도 5는 본 개시의 일 실시 예에 따라 계층적인 구조를 갖는 복수의 토픽 노드를 생성하는 방법을 설명하기 위한 도면,
 도 6은 본 개시의 일 실시 예에 따라 계층적인 구조를 갖는 복수의 토픽 노드를 생성하는 방법을 설명하기 위한 도면,
 도 7은 본 개시의 일 실시 예에 따라 계층적인 구조를 갖는 복수의 토픽 노드를 생성하는 방법을 설명하기 위한

도면,

도 8은 본 개시의 일 실시 예에 따라 계층적인 구조를 갖는 복수의 토픽 노드를 생성하는 방법을 설명하기 위한 도면,

도 9는 본 개시의 일 실시 예에 따라 계층적인 구조를 갖는 복수의 토픽 노드를 생성하는 방법을 설명하기 위한 도면,

도 10은 본 개시의 일 실시 예에 따라 계층적인 구조를 갖는 복수의 토픽 노드를 생성하는 방법을 설명하기 위한 도면,

도 11은 본 개시의 일 실시 예에 따라 계층적인 구조를 갖는 복수의 토픽 노드를 생성하는 방법을 설명하기 위한 도면,

도 12는 본 개시의 일 실시 예에 따라 계층적인 구조를 갖는 복수의 토픽 노드를 생성하는 방법을 설명하기 위한 도면,

도 13은 본 개시의 일 실시 예에 따라 계층적인 구조를 갖는 복수의 토픽 노드를 생성하는 방법을 설명하기 위한 도면,

도 14는 본 개시의 일 실시 예에 따라 계층적인 구조를 갖는 복수의 토픽 노드를 생성하는 방법을 설명하기 위한 도면,

도 15는 본 개시의 일 실시 예에 따라 계층적인 구조를 갖는 복수의 토픽 노드를 생성하는 방법을 설명하기 위한 도면,

도 16은 본 개시의 일 실시 예에 따라 계층적인 구조를 갖는 복수의 토픽 노드를 생성하는 방법을 설명하기 위한 도면,

도 17은 본 개시의 일 실시 예에 따라 계층적인 구조를 갖는 복수의 토픽 노드를 생성하는 방법을 설명하기 위한 도면,

도 18은 본 개시의 일 실시 예에 따라 도미넌트 토픽 노드를 결정하는 방법을 설명하기 위한 도면,

도 19는 본 개시의 일 실시 예에 따라 도미넌트 토픽 노드에 포함된 복수의 문서 중에서 대표 문서를 결정하는 방법을 설명하기 위한 도면,

도 20은 본 개시의 일 실시 예에 따라 도미넌트 토픽 노드에 포함된 복수의 문서 중에서 대표 문서를 결정하는 방법을 설명하기 위한 도면,

도 21은 본 개시의 일 실시 예에 따른 전자 장치의 구성을 설명하기 위한 블록도,

도 22는 본 개시의 일 실시 예에 따라 대표 문서를 제공하는 방법을 설명하기 위한 도면,

도 23은 본 개시의 일 실시 예에 따라 대표 문서를 제공하는 방법을 설명하기 위한 도면, 그리고

도 24는 본 개시의 일 실시 예에 따른 전자 장치의 대표 문서 선정 방법을 설명하기 위한 흐름도이다.

발명을 실시하기 위한 구체적인 내용

[0022] 이하, 본 개시의 다양한 실시 예가 첨부된 도면을 참조하여 기재된다. 그러나, 이는 본 개시에 기재된 기술을 특정한 실시 형태에 대해 한정하려는 것이 아니며, 본 개시의 실시 예의 다양한 변경(modifications), 균등물(equivalents), 및/또는 대체물(alternatives)을 포함하는 것으로 이해되어야 한다. 도면의 설명과 관련하여, 유사한 구성요소에 대해서는 유사한 참조 부호가 사용될 수 있다.

[0023] 본 개시에서, "가진다," "가질 수 있다," "포함한다," 또는 "포함할 수 있다" 등의 표현은 해당 특징(예: 수치, 기능, 동작, 또는 부품 등의 구성요소)의 존재를 가리키며, 추가적인 특징의 존재를 배제하지 않는다.

[0024] 본 개시에서, "A 또는 B," "A 또는/및 B 중 적어도 하나," 또는 "A 또는/및 B 중 하나 또는 그 이상"등의 표현은 함께 나열된 항목들의 모든 가능한 조합을 포함할 수 있다. 예를 들면, "A 또는 B," "A 및 B 중 적어도 하나," 또는 "A 또는 B 중 적어도 하나"는, (1) 적어도 하나의 A를 포함, (2) 적어도 하나의 B를 포함, 또는 (3) 적어도 하나의 A 및 적어도 하나의 B 모두를 포함하는 경우를 모두 지칭할 수 있다.

- [0025] 본 개시에서 사용된 "제1," "제2," "첫째," 또는 "둘째," 등의 표현들은 다양한 구성요소들을, 순서 및/또는 중요도에 상관없이 수식할 수 있고, 한 구성요소를 다른 구성요소와 구분하기 위해 사용될 뿐 해당 구성요소들을 한정하지 않는다.
- [0026] 어떤 구성요소(예: 제1 구성요소)가 다른 구성요소(예: 제2 구성요소)에 "(기능적으로 또는 통신적으로) 연결되어((operatively or communicatively) coupled with/to)" 있다거나 "접속되어(connected to)" 있다고 언급된 때에는, 어떤 구성요소가 다른 구성요소에 직접적으로 연결되거나, 다른 구성요소(예: 제3 구성요소)를 통하여 연결될 수 있다고 이해되어야 할 것이다. 반면에, 어떤 구성요소(예: 제1 구성요소)가 다른 구성요소(예: 제2 구성요소)에 "직접 연결되어" 있다거나 "직접 접속되어" 있다고 언급된 때에는, 어떤 구성요소와 다른 구성요소 사이에 다른 구성요소(예: 제 3 구성요소)가 존재하지 않는 것으로 이해될 수 있다.
- [0027] 본 개시에서 사용된 표현 "~하도록 구성된(또는 설정된)(configured to)"은 상황에 따라, 예를 들면, "~에 적합한(suitable for)," "~하는 능력을 가지는(having the capacity to)," "~하도록 설계된(designed to)," "~하도록 변경된(adapted to)," "~하도록 만들어진(made to)," 또는 "~를 할 수 있는(capable of)"과 바꾸어 사용될 수 있다. 용어 "~하도록 구성된(또는 설정된)"은 하드웨어적으로 "특별히 설계된(specifically designed to)" 것만을 반드시 의미하지 않을 수 있다. 대신, 어떤 상황에서는, "~하도록 구성된 장치"라는 표현은, 그 장치가 다른 장치 또는 부품들과 함께 "~할 수 있는" 것을 의미할 수 있다. 예를 들면, 문구 "A, B, 및 C를 수행하도록 구성된(또는 설정된) 부프로세서"는 해당 동작을 수행하기 위한 전용 프로세서(예: 임베디드 프로세서), 또는 메모리 장치에 저장된 하나 이상의 소프트웨어 프로그램들을 실행함으로써, 해당 동작들을 수행할 수 있는 범용 프로세서(generic-purpose processor)(예: CPU 또는 application processor)를 의미할 수 있다.
- [0028] 도 1은 본 개시의 일 실시 예에 따른 전자 장치의 블록도를 설명하기 위한 도면이다.
- [0029] 도 1의 전자 장치(100)는 스마트폰, 태블릿 PC, 이동 전화기, 영상 전화기, 데스크탑 PC, 랩탑 PC, 넷북 컴퓨터, TV, 서버 등과 같은 다양한 타입의 디바이스로 구현될 수 있다.
- [0030] 메모리(110)는 적어도 하나의 인스트럭션(instruction)을 포함할 수 있다. 또한, 메모리(110)는 전자 장치(100)의 적어도 하나의 다른 구성요소에 관계된 데이터, 소프트웨어, 프로그램 등을 저장할 수 있다.
- [0031] 이 경우, 메모리(110)는 비휘발성 메모리, 휘발성 메모리, 플래시메모리(flash-memory), 하드디스크 드라이브(HDD) 또는 솔리드 스테이트 드라이브(SSD) 등으로 구현될 수 있다. 메모리(110)는 프로세서(120)에 의해 액세스되며, 프로세서(120)에 의한 데이터의 독취/기록/수정/삭제/갱신 등이 수행될 수 있다. 본 개시에서 메모리라는 용어는 메모리(110), 프로세서(120) 내 롬(미도시), 램(미도시) 또는 전자 장치(100)에 장착되는 메모리 카드(미도시)(예를 들어, micro SD 카드, 메모리 스틱)를 포함할 수 있다.
- [0032] 프로세서(120)는 메모리(110)와 연결되어 전자 장치(100)를 제어할 수 있다. 특히, 프로세서(120)는 메모리(110)에 저장된 적어도 하나의 인스트럭션을 실행함으로써, 문서 집합에 포함된 복수의 문서 중에서 대표 문서를 판단할 수 있다.
- [0033] 여기에서, 문서는 기사, 게시글, 논문, 법률 사례 등과 같은 다양한 형태의 글을 포함할 수 있다. 다만, 이는 일 예일 뿐이며, 문서는 적어도 하나의 단어 또는 적어도 하나의 문장 등과 같은 텍스트를 포함하는 다양한 형태의 글을 포함할 수 있다. 이 경우, 문서는 전자 장치(100)의 메모리(110)에 저장되어 있거나, 외부의 서버 또는 다른 전자 장치로부터 수신될 수 있다.
- [0034] 먼저, 프로세서(120)는 토픽 모델링을 통해 문서 집합을 계층적인 구조를 갖는 복수의 토픽 노드에 따라 분할할 수 있다.
- [0035] 구체적으로, 본 개시의 일 실시 예에 따르면, 비음수 행렬 분해(Nonnegative Matrix Factorization, NMF)에 기반한 토픽 모델링을 통해 계층적인 구조를 갖는 복수의 토픽 노드에 따라 문서 집합에 포함된 복수의 문서를 분할할 수 있다. 이 경우, 기존과 달리 토픽의 개수를 사전에 설정하는 것이 아니라, 주어진 문서 집합에 맞게 동적으로 적합한 개수만큼의 토픽들을 생성하여, 계층적인 구조를 갖는 복수의 토픽 노드를 생성할 수 있다.
- [0036] 이를 위해, 프로세서(120)는 문서 집합에 포함된 복수의 문서에 기초하여 획득된 텀-도큐먼트(term-document) 매트릭스에 비음수 행렬 분해를 수행하여, 문서 집합에 대응되는 노드를 기설정된 개수의 제1 토픽 노드들로 분할할 수 있다.
- [0037] 여기에서, 텀-도큐먼트 매트릭스는 문서 집합에 포함된 복수의 문서에서 획득된 복수의 형태소(또는, 키워드)가

복수의 문서 각각에 포함된 빈도수(즉, 출현 빈도)에 기초하여 생성될 수 있다.

- [0038] 그리고, 기설정된 개수는 비음수 행렬 분해의 랭크(rank)에 기초하여 결정될 수 있으며, 본 개시의 일 실시 예에 따르면, 기설정된 개수는 2 일 수 있다. 즉, 랭크가 2인 비음수 행렬 분해를 통해 문서 집합에 대응되는 노드를 2 개의 토픽 노드들로 분할할 수 있다.
- [0039] 이를 위해, 먼저, 프로세서(120)는 복수의 문서로부터 복수의 형태소를 획득할 수 있다.
- [0040] 이 경우, 프로세서(120)는 토큰나이저(tokenizer) 엔진을 통해 복수의 문서를 분석할 수 있다.
- [0041] 여기에서, 토큰나이저 엔진은 문서에 포함된 텍스트를 띄어쓰기 등에 따라 구분하고 구분된 텍스트에서 특수문자, 조사, 관사 등을 제외하여, 문서로부터 명사, 동사 등과 같이 의미를 갖는 형태소를 획득하는 엔진을 포함할 수 있다. 다만, 이는 일 예일 뿐이고, 토큰나이저 엔진은 다양한 방식으로 문서에서 텍스트를 구분하여 형태소를 획득할 수 있음은 물론이다. 한편, 이러한 토큰나이저 엔진은 메모리(110)에 저장될 수 있다.
- [0042] 이에 따라, 프로세서(120)는 토큰나이저 엔진을 통해 복수의 문서로부터 복수의 형태소를 획득할 수 있다.
- [0043] 그리고, 프로세서(120)는 복수의 문서 각각에 복수의 형태소가 포함된 빈도수에 대한 정보를 판단하여, 텀-도큐먼트 매트릭스를 생성할 수 있다.
- [0044] 구체적으로, 프로세서(120)는 복수의 문서 전체로부터 획득된 복수의 형태소 각각이 복수의 문서 각각에 포함된 빈도수를 산출하고 산출된 빈도수를 매트릭스 형태로 표현하여, 텀-도큐먼트 매트릭스를 생성할 수 있다.
- [0045] 예를 들어, 도 2와 같이, 문서 1(210), 문서 2(220), 문서 3(230)이 존재하고, 문서 1(210)에서 획득된 형태소가 "aaa", "bbb", "ccc", "ddd"(215)이고, 문서 2(220)에서 획득된 형태소가 "bbb", "ddd", "eee"(225)이고, 문서 3(230)에서 획득된 형태소가 "aaa", "bbb", "fff", "ggg", "ggg"(235)인 경우를 가정한다.
- [0046] 이와 같이, 문서 1 내지 문서 3(210, 220, 230)으로부터 획득된 형태소가 "aaa", "bbb", "ccc", "ddd", "eee", "fff", "ggg"와 같은 경우, 프로세서(120)는 문서 1(210)에 "aaa", "bbb", "ccc", "ddd", "eee", "fff", "ggg"가 포함된 빈도수를 판단하고, 문서 2(220)에 "aaa", "bbb", "ccc", "ddd", "eee", "fff", "ggg"가 포함된 빈도수를 판단하고, 문서 3(230)에 "aaa", "bbb", "ccc", "ddd", "eee", "fff", "ggg"가 포함된 빈도수를 판단할 수 있다.
- [0047] 이 경우, 문서 1(210)에는 "aaa", "bbb", "ccc", "ddd"가 1 개 존재하고, 다른 형태소는 존재하지 않는다. 또한, 문서 2(220)에는 "bbb", "ddd", "eee"가 1 개 존재하고, 다른 형태소는 존재하지 않는다. 또한, 문서 3(230)에는 "ggg"가 2 개, "aaa", "bbb", "fff"가 1 개 존재하고, 다른 형태소는 존재하지 않는다.
- [0048] 이에 따라, 프로세서(120)는 문서 별로 형태소가 포함된 빈도수에 기초하여, 텀-도큐먼트 매트릭스(240)를 생성할 수 있다. 이 경우, 텀-도큐먼트 매트릭스(240)에서 행(row)는 형태소를 나타내고, 열(column)은 문서를 나타내며, 매트릭스에 포함된 값들은 문서 별로 형태소가 포함된 갯수를 나타낸다. 예를 들어, 텀-도큐먼트 매트릭스(240)의 첫 번째 열을 참조하면, 문서 1(210)에는 형태소 "aaa", "bbb", "ccc", "ddd"가 1 개씩 존재하고, 형태소 "fff", "ggg"는 존재하지 않는다는 것을 알 수 있다.
- [0049] 한편, 도 2에서는 문서가 3 개인 것으로 도시하였는데, 이는 설명의 편의를 위한 일 예에 불과할 뿐이고, 문서의 개수는 2 이상일 수 있고, 또한, 문서로부터 획득되는 형태소의 개수도 다양할 수 있음은 물론이다.
- [0050] 이후, 프로세서(120)는 텀-도큐먼트 매트릭스에 비음수 행렬 분해를 수행하여, 기설정된 개수의 제1 토픽 노드들을 생성할 수 있다.
- [0051] 예를 들어, 프로세서(120)는 문서 집합을 최상위 노드(root node)로 하여, 문서 집합에 포함된 복수의 문서에 대한 텀-도큐먼트 매트릭스에 랭크가 2인 비음수 행렬 분해를 수행하여 2 개의 자식 토픽 노드를 생성할 수 있다.
- [0052] 구체적으로, 주어진 문서 집합 $T = \{t_1, t_2, \dots, t_M\}$ 과 문서 집합에서 나온 전체 단어(즉, 형태소) 집합 $V = \{w_1, w_2, \dots, w_N\}$ 에 대해, 각 문서에 대해 산출된 단어의 출현 빈도를 나타내는 텀-도큐먼트 매트릭스 $A \in \mathbb{R}^{N \times M}$ 는 하기의 수학적 식 1과 같이 정의될 수 있다.

수학식 1

[0053] $a_{ij} = \text{freq}(t_i, w_j)$

[0054] 이때, $\text{freq}(t_i, w_j)$ 는 i 번째 문서 t_i 에서 j 번째 단어 w_j 가 언급된 횟수를 나타낸다.

[0055] 이 경우, 비음수 행렬 분해를 통해 K 개(즉, 랭크가 K)의 토픽을 추출한다고 할 때, 하기의 수학식 2를 통해 텀-도큐먼트 매트릭스는 $W \in \mathbb{R}^{N \times K}$, $H \in \mathbb{R}^{M \times K}$ 와 같이 분해될 수 있다.

수학식 2

[0056] $W, H = \underset{W, H}{\operatorname{argmin}} \|A - WH^T\|_F^2$

[0057] 여기에서, W 는 텀-토픽 가중치를 나타내며, H 는 토픽-도큐먼트 가중치를 나타낸다.

[0058] 이 경우, W, H 를 산출하기 위해, 프로세서(120)는 수학식 2에서 각 매트릭스에 대해 편미분을 계산하는 하기와 같은 수학식 3, 4를 각 반복 단계마다 한 매트릭스를 고정하고, 다른 한 매트릭스를 업데이트하는 방식(coordinate gradient descent update algorithm)으로 수행하여, W 와 H 매트릭스를 산출하게 된다. 이때, 프로세서(120)는 프로베니우스 놈(Frobenious Norm)의 값이 Tolerance($=10^{-4}$) 이하일 경우 최소가 된다고 가정하여, 전술한 과정을 종료하고 W 와 H 매트릭스를 산출하게 된다.

수학식 3

[0059] $H \leftarrow H \frac{(W^T A)}{W^T W H}$

수학식 4

[0060] $W \leftarrow W \frac{(A H^T)}{W H H^T}$

[0061] 결과적으로, 프로세서(120)는 비음수 행렬 분해를 통해 텀-도큐먼트 매트릭스를 텀-토픽 가중치를 나타내는 매트릭스와 토픽-도큐먼트 가중치를 나타내는 매트릭스로 분해될 수 있다.

[0062] 예를 들어, 텀-도큐먼트 매트릭스 $A(310)$ 가 도 3과 같은 경우를 가정한다. 이 경우, 텀-도큐먼트 매트릭스 $A(310)$ 의 값들은 M 개의 문서에 대해, 문서 별로, N 개의 형태소가 포함된 빈도수를 나타낸다. 여기에서, N 개의 형태소는 M 개의 문서로부터 획득된 형태소이다.

[0063] 예를 들어, a_{11} 는 문서 1에 형태소 1이 언급된 횟수, a_{21} 는 문서 1에 형태소 2가 언급된 횟수, ..., a_{N1} 는 문서 1에 형태소 N 이 언급된 횟수, ..., a_{1M} 는 문서 M 에 형태소 1이 언급된 횟수, a_{2M} 는 문서 M 에 형태소 2가 언급된 횟수, ..., a_{NM} 는 문서 M 에 형태소 N 이 언급된 횟수를 나타낸다.

[0064] 한편, 텀-도큐먼트 매트릭스 $A(310)$ 에 대해 랭크가 K 인 비음수 행렬 분해를 수행하면, 텀-토픽 가중치 $W(320)$ 및 토픽-도큐먼트 가중치 $H(330)$ 가 생성될 수 있다.

[0065] 이 경우, 텀-토픽 가중치 $W(320)$ 는 매트릭스 형태로 표현될 수 있으며, 매트릭스의 값들은 K 개의 토픽에 대해, 각 토픽에서의 N 개의 형태소의 가중치를 나타낸다.

- [0066] 예를 들어, w_{11} 는 토픽 1에서의 형태소 1의 가중치, w_{21} 는 토픽 1에서의 형태소 2의 가중치, ..., w_{N1} 는 토픽 1에서의 형태소 N의 가중치, ..., w_{1K} 는 토픽 K에서의 형태소 1의 가중치, w_{2K} 는 토픽 K에서의 형태소 2의 가중치, ..., w_{NK} 는 토픽 K에서의 형태소 N의 가중치를 나타낸다.
- [0067] 한편, 토픽-도큐먼트 가중치 H(330)는 매트릭스 형태로 표현될 수 있으며, 매트릭스의 값들은 M 개의 문서에 대해, 각 문서에서의 K 개의 토픽의 가중치를 나타낸다.
- [0068] 예를 들어, w_{11} '는 문서 1에서의 토픽 1의 가중치, w_{21} '는 문서 1에서의 토픽 2의 가중치, ..., w_{K1} '는 문서 1에서의 토픽 K의 가중치, ..., w_{1M} '는 문서 M에서의 토픽 1의 가중치, w_{2M} '는 문서 M에서의 토픽 2의 가중치, ..., w_{KM} '는 문서 M에서의 토픽 K의 가중치를 나타낸다.
- [0069] 한편, 본 개시의 일 실시 예에 따르면, 랭크가 2(즉, $K=2$)인 비음수 행렬 분해를 이용한다는 점에서, 프로세서(120)는 문서 집합에 대응되는 최상위 노드에 대해 2 개의 자식 토픽 노드를 생성할 수 있다.
- [0070] 이후, 프로세서(120)는 토픽-도큐먼트 가중치에 기초하여 각 토픽 노드 각각에 속하는 문서를 결정하여, 문서 집합에 포함된 복수의 문서를 복수의 토픽에 따라 분할할 수 있다.
- [0071] 이 경우, 프로세서(120)는 문서가 상대적으로 큰 가중치를 갖는 토픽 노드에 포함되도록, 문서가 속하는 토픽 노드를 결정할 수 있다.
- [0072] 구체적으로, 토픽-도큐먼트 매트릭스에 대해 랭크가 2인 비음수 행렬 분해를 수행하여, 2 개의 토픽 자식 노드(K_1 , K_2)가 생성된 경우를 가정한다. 이 경우, 프로세서(120)는 하기의 수학식 5을 이용하여 문서를 토픽-도큐먼트 가중치가 상대적으로 큰 토픽에 할당하는 방식으로 복수의 문서를 클러스터링하여, 각 토픽에 포함되는 하위 문서 집합 (D_1 , D_2)를 생성할 수 있다.

수학식 5

$$D(d_i) = \underset{j}{\operatorname{argmax}} H_{i,j}$$

- [0073]
- [0074] 예를 들어, 토픽-도큐먼트 가중치 H(410)가 도 4와 같은 경우를 가정한다. 이 경우, 문서 1에서의 토픽 1의 가중치는 w_{11} '이고, 문서 1에서의 토픽 2의 가중치는 w_{21} '이다. 이 경우, 이들 가중치 w_{11} ' 및 w_{21} ' 중에서 w_{11} '가 큰 값을 갖는 경우, 프로세서(120)는 문서 1은 토픽 1에 속하는 것으로 분류할 수 있다. 또한, 문서 2에서의 토픽 1의 가중치는 w_{12} '이고, 문서 2에서의 토픽 2의 가중치는 w_{22} '이다. 이 경우, 이들 가중치 w_{12} ' 및 w_{22} ' 중에서 w_{22} '가 큰 값을 갖는 경우, 프로세서(120)는 문서 2는 토픽 2에 속하는 것으로 분류할 수 있다. 또한, 문서 3에서의 토픽 1의 가중치는 w_{13} '이고, 문서 3에서의 토픽 2의 가중치는 w_{23} '이다. 이 경우, 이들 가중치 w_{13} ' 및 w_{23} ' 중에서 w_{13} '가 큰 값을 갖는 경우, 프로세서(120)는 문서 3은 토픽 1에 속하는 것으로 분류할 수 있다. 또한, 문서 M-1에서의 토픽 1의 가중치는 w_{1M-1} '이고, 문서 M-1에서의 토픽 2의 가중치는 w_{2M-1} '이다. 이 경우, 이들 가중치 w_{1M-1} ' 및 w_{2M-1} ' 중에서 w_{2M-1} '가 큰 값을 갖는 경우, 프로세서(120)는 문서 M-1은 토픽 2에 속하는 것으로 분류할 수 있다. 또한, 문서 M에서의 토픽 1의 가중치는 w_{1M} '이고, 문서 M에서의 토픽 2의 가중치는 w_{2M} '이다. 이 경우, 이들 가중치 w_{1M} ' 및 w_{2M} ' 중에서 w_{1M} '가 큰 값을 갖는 경우, 프로세서(120)는 문서 M은 토픽 1에 속하는 것으로 분류할 수 있다.
- [0075] 결국, 도 5와 같이, 프로세서(120)는 문서 집합에 대응되는 최상위 노드(510)에 포함된 M 개의 문서를 2 개의 토픽에 따라 분류할 수 있다. 이 경우, 토픽 1에 포함된 하위 문서 집합 D_1 (520)은 문서 1, 문서3, ..., 문서 M을 포함하고, 토픽 2에 포함된 하위 문서 집합 D_2 (530)은 문서 2, ..., 문서 M-1을 포함할 수 있다.
- [0076] 이후, 프로세서(120)는 전술한 방식을 이용하여 각 토픽 노드를 분할하여, 계층적인 구조를 갖는 복수의 토픽 노드를 생성할 수 있다.

- [0077] 이때, 프로세서(120)는 복수의 토픽 노드 중에서 우선적으로 먼저 분할된 토픽 노드를 결정하고, 결정된 토픽 노드를 분할한다. 그리고, 프로세서(120)는 토픽 노드에서 분할된 자식 토픽 노드, 및 분할되지 않은 토픽 노드 중에서 우선적으로 먼저 분할된 토픽 노드를 결정하고, 결정된 토픽 노드를 분할할 수 있다.
- [0078] 구체적으로, 프로세서(120)는 제1 토픽 노드들을 각각 분할하였을 때 생성되는 자식 토픽 노드들에 기초하여 제1 토픽 노드들 각각의 스코어를 산출하고, 산출된 스코어에 기초하여 제1 토픽 노드들 중에서 분할될 토픽 노드를 결정하고, 결정된 토픽 노드를 기설정된 개수의 제2 토픽 노드들로 분할할 수 있다.
- [0079] 이 경우, 프로세서(120)는 제1 토픽 노드들 중에서 산출된 스코어가 가장 높은 토픽 노드를 분할될 토픽 노드로 결정할 수 있다.
- [0080] 그리고, 프로세서(120)는 제1 토픽 노드들 중 분할되지 않는 토픽 노드, 및 제2 토픽 노드들을 각각 분할하였을 때 생성되는 자식 토픽 노드들에 기초하여 분할되지 않는 토픽 노드 및 제2 토픽 노드들 각각의 스코어를 산출하고, 산출된 스코어에 기초하여 분할되지 않는 토픽 노드 및 제2 토픽 노드들 중에서 분할될 토픽 노드를 결정하고, 결정된 토픽 노드를 기설정된 개수의 제3 토픽 노드들로 분할할 수 있다.
- [0081] 이하에서는 최상위 노드에서 분할된 2 개의 토픽 노드를 분할하여, 계층적인 구조를 갖는 복수의 토픽 노드를 생성하는 방법을 보다 구체적으로 설명하도록 한다.
- [0082] 예를 들어, 도 6와 같이, 문서 집합에 대응되는 최상위 노드(610)에 포함된 복수의 문서를 2 개의 토픽 노드(620, 630)에 따라 분류한 경우를 가정한다.
- [0083] 이 경우, 도 7과 같이, 프로세서(120)는 토픽 노드(620)에 포함된 복수의 문서들에 기초하여 톰-도큐먼트 매트릭스를 생성하고, 톰-도큐먼트 매트릭스에 대해 랭크가 2인 비음수 행렬 분해를 수행하여 토픽 노드(620)에 대한 2 개의 자식 토픽 노드(621, 622)를 생성한다. 그리고, 프로세서(120)는 토픽 노드(620)에 포함된 복수의 문서들을 클러스터링하여, 각 자식 토픽 노드에 포함되는 하위 문서 집합을 생성할 수 있다.
- [0084] 한편, 자식 토픽 노드를 생성하는 방법 및 각 자식 토픽 노드에 포함되는 문서를 결정하는 방법은 최상위 노드에 대해 이용된 방법과 동일한 방법이 이용될 수 있다는 점에서, 구체적인 설명은 생략하도록 한다.
- [0085] 그리고, 프로세서(120)는 2 개의 자식 토픽 노드(621, 622) 각각에 속한 복수의 문서에 포함된 형태소에 기초하여, 토픽 노드(620)에 대한 스코어를 산출할 수 있다. 여기에서, 스코어는 토픽 노드(620)를 분할하였을 때 생성되는 자식 토픽 노드(621, 622)가 토픽 노드(620)를 잘 설명할 수 있는지를 나타내는 값에 해당한다. 한편, 스코어를 산출하는 구체적인 방식은 후술하도록 한다.
- [0086] 한편, 프로세서(120)는 토픽 노드(630)에 대해서도 전술한 과정을 반복하여 토픽 노드(630)에 대한 스코어를 산출할 수 있다.
- [0087] 구체적으로, 도 8과 같이, 프로세서(120)는 토픽 노드(630)에 포함된 복수의 문서들에 기초하여 톰-도큐먼트 매트릭스를 생성하고, 톰-도큐먼트 매트릭스에 대해 랭크가 2인 비음수 행렬 분해를 수행하여 토픽 노드(630)에 대한 2 개의 자식 토픽 노드(631, 632)를 생성한다. 그리고, 프로세서(120)는 토픽 노드(630)에 포함된 복수의 문서들을 클러스터링하여, 각 자식 토픽 노드에 포함되는 하위 문서 집합을 생성할 수 있다.
- [0088] 한편, 자식 토픽 노드를 생성하는 방법 및 각 자식 토픽 노드에 포함되는 문서를 결정하는 방법은 최상위 노드에 대해 이용된 방법과 동일한 방법이 이용될 수 있다는 점에서, 구체적인 설명은 생략하도록 한다.
- [0089] 그리고, 프로세서(120)는 2 개의 자식 토픽 노드(631, 632) 각각에 속한 복수의 문서에 포함된 형태소에 기초하여, 토픽 노드(630)에 대한 스코어를 산출할 수 있다. 여기에서, 스코어는 토픽 노드(630)를 분할하였을 때 생성되는 자식 토픽 노드(631, 632)가 토픽 노드(630)를 잘 설명할 수 있는지를 나타내는 값에 해당한다. 한편, 스코어를 산출하는 구체적인 방식은 후술하도록 한다.
- [0090] 이후, 프로세서(120)는 토픽 노드(620, 630)에 대해 산출된 스코어에 기초하여, 토픽 노드(620, 630) 중에서 분할될 토픽 노드를 결정하고, 결정된 토픽 노드를 2 개의 자식 토픽 노드로 분할할 수 있다.
- [0091] 이 경우, 프로세서(120)는 스코어가 상대적으로 더 큰 토픽 노드를 분할될 토픽 노드로 결정할 수 있다.
- [0092] 예를 들어, 토픽 노드(620)에 대한 스코어가 토픽 노드(630)에 대한 스코어보다 큰 경우, 도 9와 같이, 프로세서(120)는 토픽 노드(620, 630) 중에서 분할될 토픽 노드는 토픽 노드(620)인 것으로 결정하고, 토픽 노드(620)를 분할하여 2 개의 자식 토픽 노드(621, 622)를 생성할 수 있다.

- [0093] 이후, 프로세서(120)는 토픽 노드(621, 622, 630)에 대해 전술한 과정을 반복적으로 수행하여, 토픽 노드(621, 622, 630) 중에서 분할될 토픽 노드를 결정하고, 결정된 토픽 노드를 2 개의 자식 토픽 노드로 분할할 수 있다.
- [0094] 구체적으로, 도 10과 같이, 프로세서(120)는 토픽 노드(621)에 포함된 복수의 문서들에 기초하여 톰-도큐먼트 매트릭스를 생성하고, 톰-도큐먼트 매트릭스에 대해 랭크가 2인 비음수 행렬 분해를 수행하여 토픽 노드(621)에 대한 2 개의 자식 토픽 노드(623, 624)를 생성한다. 그리고, 프로세서(120)는 토픽 노드(621)에 포함된 복수의 문서들을 클러스터링하여, 각 자식 토픽 노드에 포함되는 하위 문서 집합을 생성할 수 있다.
- [0095] 한편, 자식 토픽 노드를 생성하는 방법 및 각 자식 토픽 노드에 포함되는 문서를 결정하는 방법은 최상위 노드에 대해 이용된 방법과 동일한 방법이 이용될 수 있다는 점에서, 구체적인 설명은 생략하도록 한다.
- [0096] 그리고, 프로세서(120)는 2 개의 자식 토픽 노드(623, 624) 각각에 속한 복수의 문서에 포함된 형태소에 기초하여, 토픽 노드(621)에 대한 스코어를 산출할 수 있다. 여기에서, 스코어는 토픽 노드(621)를 분할하였을 때 생성되는 자식 토픽 노드(623, 624)가 토픽 노드(621)를 잘 설명할 수 있는지를 나타내는 값에 해당한다. 한편, 스코어를 산출하는 구체적인 방식은 후술하도록 한다.
- [0097] 한편, 프로세서(120)는 토픽 노드(622)에 대해서도 전술한 과정을 반복하여 토픽 노드(622)에 대한 스코어를 산출할 수 있다.
- [0098] 구체적으로, 도 11과 같이, 프로세서(120)는 토픽 노드(622)에 포함된 복수의 문서들에 기초하여 톰-도큐먼트 매트릭스를 생성하고, 톰-도큐먼트 매트릭스에 대해 랭크가 2인 비음수 행렬 분해를 수행하여 토픽 노드(622)에 대한 2 개의 자식 토픽 노드(625, 626)를 생성한다. 그리고, 프로세서(120)는 토픽 노드(622)에 포함된 복수의 문서들을 클러스터링하여, 각 자식 토픽 노드에 포함되는 하위 문서 집합을 생성할 수 있다.
- [0099] 한편, 자식 토픽 노드를 생성하는 방법 및 각 자식 토픽 노드에 포함되는 문서를 결정하는 방법은 최상위 노드에 대해 이용된 방법과 동일한 방법이 이용될 수 있다는 점에서, 구체적인 설명은 생략하도록 한다.
- [0100] 그리고, 프로세서(120)는 2 개의 자식 토픽 노드(625, 626) 각각에 속한 복수의 문서에 포함된 형태소에 기초하여, 토픽 노드(622)에 대한 스코어를 산출할 수 있다. 여기에서, 스코어는 토픽 노드(622)를 분할하였을 때 생성되는 자식 토픽 노드(625, 626)가 토픽 노드(622)를 잘 설명할 수 있는지를 나타내는 값에 해당한다. 한편, 스코어를 산출하는 구체적인 방식은 후술하도록 한다.
- [0101] 한편, 프로세서(120)는 토픽 노드(630)에 대해서도 전술한 과정을 반복하여 토픽 노드(630)에 대한 스코어를 산출할 수 있다.
- [0102] 구체적으로, 도 12와 같이, 프로세서(120)는 토픽 노드(630)에 포함된 복수의 문서들에 기초하여 톰-도큐먼트 매트릭스를 생성하고, 톰-도큐먼트 매트릭스에 대해 랭크가 2인 비음수 행렬 분해를 수행하여 토픽 노드(630)에 대한 2 개의 자식 토픽 노드(631, 632)를 생성한다. 그리고, 프로세서(120)는 토픽 노드(630)에 포함된 복수의 문서들을 클러스터링하여, 각 자식 토픽 노드에 포함되는 하위 문서 집합을 생성할 수 있다.
- [0103] 한편, 자식 토픽 노드를 생성하는 방법 및 각 자식 토픽 노드에 포함되는 문서를 결정하는 방법은 최상위 노드에 대해 이용된 방법과 동일한 방법이 이용될 수 있다는 점에서, 구체적인 설명은 생략하도록 한다.
- [0104] 그리고, 프로세서(120)는 2 개의 자식 토픽 노드(631, 632) 각각에 속한 복수의 문서에 포함된 형태소에 기초하여, 토픽 노드(630)에 대한 스코어를 산출할 수 있다. 여기에서, 스코어는 토픽 노드(630)를 분할하였을 때 생성되는 자식 토픽 노드(631, 632)가 토픽 노드(630)를 잘 설명할 수 있는지를 나타내는 값에 해당한다. 한편, 스코어를 산출하는 구체적인 방식은 후술하도록 한다.
- [0105] 이후, 프로세서(120)는 토픽 노드(621, 622, 630)에 대해 산출된 스코어에 기초하여, 토픽 노드(621, 622, 630) 중에서 분할될 토픽 노드를 결정하고, 결정된 토픽 노드를 2 개의 자식 토픽 노드로 분할할 수 있다.
- [0106] 이 경우, 프로세서(120)는 스코어가 상대적으로 더 큰 토픽 노드를 분할될 토픽 노드로 결정할 수 있다.
- [0107] 예를 들어, 토픽 노드(622)에 대한 스코어가 토픽 노드(621) 및 토픽 노드(630)에 대한 스코어보다 큰 경우, 도 13과 같이, 프로세서(120)는 토픽 노드(621, 622, 630) 중에서 분할될 토픽 노드는 토픽 노드(622)인 것으로 결정하고, 토픽 노드(622)를 분할하여 2 개의 자식 토픽 노드(625, 626)를 생성할 수 있다.
- [0108] 결국, 프로세서(120)는 토픽 노드들에 대해, 전술한 과정을 반복적으로 수행하여 토픽 노드들 중에서 분할될 토픽 노드를 결정하고, 결정된 토픽 노드를 2 개의 자식 토픽 노드로 분할하여, 계층적인 구조를 갖는 복수의 토픽

픽 노드를 생성할 수 있다.

[0109] 한편, 토픽 노드에 대해 스코어를 산출하는 방법은 다음과 같다.

[0110] 구체적으로, 스코어는 토픽 노드를 분할하여 자식 토픽 노드를 생성하였을 때, 자식 토픽 노드가 그들의 부모인 토픽 노드를 얼마나 설명할 수 있는지에 대한 우선 순위 스코어를 나타낸다.

[0111] 이와 관련하여, 토픽 노드 K_t 에 대한 스코어 f_t 를 산출하기 위해, 먼저, 프로세서(120)는 토픽 노드 K_t 에 대한 자식 토픽 노드 K_{t1} , K_{t2} 를 생성할 수 있다. 그리고, 프로세서(120)는 토픽 노드 K_t 및 자식 토픽 노드 K_{t1} , K_{t2} 각각에 대한 텀-도큐먼트 매트릭스를 생성하고, 하기의 수학적 식 6에 기초하여 토픽 노드 및 자식 토픽 노드 각각에서의 형태소를 빈도수에 따라 정렬할 수 있다.

수학적 식 6

$$P_t = \{w_{t,1}, w_{t,2}, \dots, w_{t,N}\} = \underset{w}{\operatorname{argsort}} \sum_{d \in D_t} \operatorname{freq}(d_t, w)$$

[0113] 예를 들어, 도 14와 같이, 토픽 노드 K_t 에 2 개의 자식 토픽 노드 K_{t1} , K_{t2} 가 분할된 경우를 가정한다.

[0114] 이 경우, 프로세서(120)는 토픽 노드 K_t 에 대응되는 문서 집합 D_t 에 포함된 복수의 문서에 기초하여 텀-도큐먼트 매트릭스(1410)를 생성할 수 있다.

[0115] 그리고, 프로세서(120)는 복수의 형태소 $w_1, w_2, w_3, \dots, w_N$ 각각에 복수의 문서에 포함된 빈도수들의 합을 산출할 수 있다.

[0116] 예를 들어, 형태소 w_1 의 경우, 문서 1에 a_{11} 번 언급되고, 문서 2에 a_{12} 번 언급되고, ..., 문서 M에 a_{1M} 번 언급된다. 이에 따라, 형태소 w_1 의 빈도수들의 합 b_1 는 $b_1 = a_{11} + a_{12} + \dots + a_{1M}$ 이다. 또한, 형태소 w_2 의 경우, 문서 1에 a_{21} 번 언급되고, 문서 2에 a_{22} 번 언급되고, ..., 문서 M에 a_{2M} 번 언급된다. 이에 따라, 형태소 w_2 의 빈도수들의 합 b_2 은 $b_2 = a_{21} + a_{22} + \dots + a_{2M}$ 이다. 또한, 형태소 w_N 의 경우, 문서 1에 a_{N1} 번 언급되고, 문서 2에 a_{N2} 번 언급되고, ..., 문서 M에 a_{NM} 번 언급된다. 이에 따라, 형태소 w_N 의 빈도수들의 합 b_N 은 $b_N = a_{N1} + a_{N2} + \dots + a_{NM}$ 이다.

[0117] 이후, 프로세서(120)는 빈도수들의 합이 큰 순으로 형태소를 정렬하여, 수학적 식 6과 같이 P_t 즉, 토픽 노드 K_t 에 포함된 형태소를 빈도수의 합이 큰 순으로 정렬한 형태소 집합을 획득할 수 있다.

[0118] 전술한 예에서, $b_3 > b_2 > b_N > \dots > b_1$ 인 경우, $P_t = \{w_{t,1}, w_{t,2}, \dots, w_{t,N}\} = \{w_3, w_2, w_N, \dots, w_1\}$ 와 같다. 여기에서, $w_{t,i}$ 는 토픽 노드 K_t 에서 빈도수의 합이 i 번째로 높은 형태소를 나타낸다.

[0119] 한편, 프로세서(120)는 전술한 과정을 자식 토픽 노드 K_{t1} , K_{t2} 에 대해 수행하여, 자식 토픽 노드 K_{t1} 에 대한 P_{t1} (즉, 자식 토픽 노드 K_{t1} 에 포함된 형태소를 빈도수의 합이 큰 순으로 정렬한 형태소 집합) 및 자식 토픽 노드 K_{t2} 에 대한 P_{t2} (즉, 자식 토픽 노드 K_{t2} 에 포함된 형태소를 빈도수의 합이 큰 순으로 정렬한 형태소 집합)을 획득할 수 있다.

[0120] 이후, 프로세서(120)는 토픽 노드에 대한 스코어를 산출할 수 있다.

[0121] 이 경우, 프로세서(120)는 정보 검색에서 주로 사용되는 척도인 nDCG(normalized Discounted Cumulative Gain)을 이용하여, 스코어를 산출하게 된다. 여기에서, nDCG는 검색된 문서들이 얼마나 높은 순위에 관련성 있게 반환되는지를 평가하는 척도이다.

[0122] 구체적으로, 프로세서(120)는 하기의 수학적 식 7에 기초하여, nDCG를 산출하게 된다.

수학식 7

$$\begin{aligned} \text{rel}_i(K_{\text{child}}, K_{\text{parent}}) &= \text{rank}(w_{\text{child}, i}, P_{\text{parent}}) \\ \text{DCG}_p(K_{\text{child}}, K_{\text{parent}}) &= \sum_{i=1}^p \frac{\text{rel}_i(K_{\text{child}}, K_{\text{parent}})}{\log_2(i+1)} \\ \text{IDCG}_p(K_{\text{child}}, K_{\text{parent}}) &= \sum_{i=1}^{|\text{REL}|} \frac{\text{rel}_i(K_{\text{child}}, K_{\text{parent}})}{\log_2(i+1)} \\ \text{nDCG}_p(K_{\text{child}}, K_{\text{parent}}) &= \frac{\text{DCG}_p(K_{\text{child}}, K_{\text{parent}})}{\text{IDCG}_p(K_{\text{child}}, K_{\text{parent}})} \end{aligned}$$

[0123]

[0124] 여기에서, rel은 relevance로 $\text{rank}(w_{\text{child}, i}, P_{\text{parent}})$ 에 의해 정의되며, $\text{rank}(w_{\text{child}, i}, P_{\text{parent}})$ 는 자식 토픽 노드의 i 번째 형태소가 부모 노드에서 가지는 순위를 의미한다. p는 자식 토픽 노드에 포함된 전체 형태소의 개수를 나타낸다.

[0125] IDCG(즉, ideal DCG)는 자식 토픽 노드가 이상적으로 정렬되었을 때의 스코어가 된다. 즉, IDCG는 부모 토픽 노드에 포함된 형태소를 빈도수에 따라 정렬할 때, 정렬된 순서를 기준으로 자식 토픽 노드에 포함된 형태소를 재정렬한 값에서 계산된 DCG를 의미한다.

[0126] 구체적으로, IDCG는 부모 토픽 노드에 포함된 형태소를 빈도수가 큰 순으로 정렬한 경우, 자식 토픽 노드에 포함된 형태소의 순위를 부모 토픽 노드에서의 순위와 같다고 가정하고, DCG를 계산한 것을 의미한다. 따라서,

$$\text{IDCG}_p(K_{\text{child}}, K_{\text{parent}}) = \sum_{i=1}^p \frac{i}{\log_2(i+1)}$$

IDCG는

[0127] 이에 따라, 프로세서(120)는 하기의 수학식 8에 기초하여, 토픽 노드에 대한 스코어를 산출하게 된다.

수학식 8

$$\text{priority}_t = \text{nDCG}_p(K_{t1}, K_t) * \text{nDCG}_p(K_{t2}, K_t)$$

[0128]

[0129] 즉, 프로세서(120)는 자식 토픽 노드 K_{t1} 가 토픽 노드 K_t 를 얼마나 잘 설명하고 있는지를 나타내는 $\text{nDCG}_p(K_{t1}, K_t)$ 및 자식 토픽 노드 K_{t2} 가 토픽 노드 K_t 를 얼마나 잘 설명하고 있는지를 나타내는 $\text{nDCG}_p(K_{t2}, K_t)$ 를 산출하고, 이들을 곱하여 토픽 노드 K_t 에 대한 스코어 priority_t 를 산출할 수 있다.

[0130] 결국, 프로세서(120)는 전술한 방법을 통해 토픽 노드에 대한 스코어를 산출할 수 있다.

[0131] 한편, 프로세서(120)는 전술한 바와 같이 토픽 노드를 분할하여, 계층적인 구조를 갖는 복수의 토픽 노드를 생성하는데, 이때, 기설정된 조건을 만족하는 토픽 노드를 제외하면서 분할 과정을 수행할 수 있다.

[0132] 여기에서, 기설정된 조건은 토픽 노드를 분할하였을 때 생성되는 자식 토픽 노드들에 기초하여 산출된 스코어가 음수인 조건, 토픽 노드 및 자식 토픽 노드 간 또는 자식 토픽 노드들 간에 기설정된 개수 이상의 동일한 빈도수 상위 형태소가 존재하는 조건을 포함할 수 있다.

[0133] 여기에서, 먼저, 자식 토픽 노드들 간에 기설정된 개수 이상의 동일한 빈도수 상위 형태소가 존재하는 조건은 다음과 같다.

[0134] 예를 들어, 토픽 노드에서 2 개의 자식 토픽 노드들 즉, 제1 자식 토픽 노드 및 제2 자식 토픽 노드가 분할된 경우를 가정한다.

[0135] 여기에서, 프로세서(120)는 제1 자식 토픽 노드에 속한 복수의 문서들에 포함된 형태소를 빈도수 순위에 따라 정렬하여 상위 10 개의 형태소를 판단한다. 그리고, 프로세서(120)는 제2 자식 토픽 노드에 속한 복수의 문서들

에 포함된 형태소를 빈도수 순위에 따라 정렬하여 상위 10 개의 형태소를 판단한다. 이때, 프로세서(120)는 제1 및 제2 자식 토픽 노드 각각에서 획득된 상위 10 개 내의 형태소 간에 기설정된 개수 이상 동일한 형태소가 겹치는 경우, 해당 토픽 노드를 분할 과정에서 제외하여, 더 이상 분할하지 않는다. 여기에서, 기설정된 개수는 가령, 3 개가 될 수 있다. 다만, 이 값은 변경될 수 있음은 물론이다.

[0136] 한편, 토픽 노드와 자식 토픽 노드들 간에 기설정된 개수 이상의 동일한 빈도수 상위 형태소가 존재하는 조건은 다음과 같다.

[0137] 예를 들어, 토픽 노드에서 2 개의 자식 토픽 노드들 즉, 제1 자식 토픽 노드 및 제2 자식 토픽 노드가 분할된 경우를 가정한다.

[0138] 여기에서, 프로세서(120)는 토픽 노드에 속한 복수의 문서들에 포함된 형태소를 빈도수 순위에 따라 정렬하여 상위 10 개의 형태소를 판단한다.

[0139] 그리고, 프로세서(120)는 제1 자식 토픽 노드에 속한 복수의 문서들에 포함된 형태소를 빈도수 순위에 따라 정렬하여 상위 10 개의 형태소를 판단한다. 이후, 프로세서(120)는 토픽 노드 및 제1 자식 토픽 노드 각각에서 획득된 상위 10 개의 형태소 간에 겹치는 형태소가 존재하는지를 판단하고, 겹치는 형태소가 존재하는 경우, 겹치는 형태소의 개수를 판단할 수 있다.

[0140] 또한, 프로세서(120)는 제2 자식 토픽 노드에 속한 복수의 문서들에 포함된 형태소를 빈도수 순위에 따라 정렬하여 상위 10 개의 형태소를 판단한다. 이후, 프로세서(120)는 토픽 노드 및 제2 자식 토픽 노드 각각에서 획득된 상위 10 개의 형태소 간에 겹치는 형태소가 존재하는지를 판단하고, 겹치는 형태소가 존재하는 경우, 겹치는 형태소의 개수를 판단할 수 있다.

[0141] 이후, 프로세서(120)는 토픽 노드 및 제1 자식 토픽 노드 간의 겹치는 형태소의 개수와 토픽 노드 및 제2 자식 토픽 노드 간의 겹치는 형태소의 개수에 대한 평균 값이 기설정된 값 이상인 경우, 해당 토픽 노드를 분할 과정에서 제외하여, 더 이상 분할하지 않는다. 여기에서, 기설정된 값은 가령, 3일 수 있다. 다만, 이 값은 변경될 수 있음은 물론이다.

[0142] 한편, 프로세서(120)는 토픽 노드에 포함된 복수의 문서의 개수가 기설정된 제1 비율 이상인 경우, 해당 토픽 노드를 분할하는 반면, 토픽 노드에 포함된 복수의 문서의 개수가 기설정된 제2 비율 이하인 경우에는 해당 토픽 노드를 분할 과정에서 제외할 수 있다.

[0143] 구체적으로, 프로세서(120)는 토픽 노드에 대해, (토픽 노드에 포함된 복수의 문서의 개수)/(최상위 노드에 포함된 복수의 문서의 개수) $\geq 1/3$ 이상인 경우, 해당 토픽 노드를 분할할 수 있다. 다만, 프로세서(120)는 토픽 노드에 대해, (토픽 노드에 포함된 복수의 문서의 개수)/(최상위 노드에 포함된 복수의 문서의 개수) $\leq 1/10$ 이하인 경우, 해당 토픽 노드를 분할 과정에서 제외하고, 해당 토픽 노드는 더 이상 분할하지 않을 수 있다. 다만, 이들 비율은 변경될 수 있음은 물론이다.

[0144] 이하에서는, 도 15 내지 도 17을 참조하여, 분할 과정에서 제외되는 즉, 더 이상 분할되지 않는 토픽 노드에 대해 설명하도록 한다.

[0145] 예를 들어, 도 15와 같은 경우, 프로세서(120)는 토픽 노드들(1510, 1520, 1530, 1540)에서 분할될 토픽 노드를 결정할 수 있다.

[0146] 이때, 프로세서(120)는 최상위 노드에 포함된 복수의 문서의 개수에 대한 토픽 노드(1510)에 속한 복수의 문서의 개수의 비율이 $1/10$ 이하인 경우, 토픽 노드(1510)를 분할 과정에서 제외할 수 있다.

[0147] 또한, 프로세서(120)는 토픽 노드(1530)의 자식 토픽 노드들에 대해, 빈도수 상위 형태소 간에 동일한 형태소가 3 개 이상 존재하는 경우, 토픽 노드(1530)를 분할 과정에서 제외할 수 있다.

[0148] 이에 따라, 프로세서(120)는 토픽 노드들(1520, 1540)에 대한 스코어를 산출하고, 토픽 노드(1520)에 대해 산출된 스코어가 양수이고 토픽 노드(1540)에 대한 스코어보다 큰 경우, 도 16과 같이 토픽 노드(1520)를 2 개의 자식 토픽 노드(1550, 1560)으로 분할할 수 있다.

[0149] 이후, 프로세서(120)는 토픽 노드들(1540, 1550, 1560)에서 분할될 토픽 노드를 결정할 수 있다.

[0150] 이때, 프로세서(120)는 토픽 노드(1550) 및 토픽 노드(1560)의 자식 토픽 노드들에 대해, 빈도수 상위 형태소 간에 동일한 형태소가 평균 3 개 이상 존재하는 경우, 토픽 노드(1550)를 분할 과정에서 제외할 수 있다.

- [0151] 이에 따라, 프로세서(120)는 토픽 노드들(1540, 1560)에 대한 스코어를 산출할 수 있다.
- [0152] 이때, 프로세서(120)는 토픽 노드(1560)에 대해 산출된 스코어가 음수인 경우, 토픽 노드(1550)를 분할 과정에서 제외할 수 있다.
- [0153] 한편, 프로세서(120)는 토픽 노드들(1540)에 대해 산출된 스코어가 양수인 경우, 도 17과 같이 토픽 노드(1540)를 2 개의 자식 토픽 노드(1570, 1580)으로 분할할 수 있다.
- [0154] 이후, 만일, 토픽 노드(1570, 1580)에 대해 산출된 스코어가 모두 음수인 경우, 프로세서(120)는 토픽 노드(1570, 1580)를 분할 과정에서 제외할 수 있으며, 결국, 최상위 노드를 분할하여 도 17과 같은 계층적인 구조를 갖는 복수의 토픽 노드를 생성할 수 있다.
- [0155] 결국, 이러한 방식에 따른 계층적인 토픽 모델링을 통해 계층적인 구조를 갖는 복수의 토픽 노드가 생성될 수 있다.
- [0156] 이와 관련하여, 기존의 토픽 모델링 기법의 경우, 토픽의 개수를 사전에 설정하여야 하는데, 분석하려는 문서 집합의 크기 및 주제에 따라 토픽의 개수는 다양하게 존재할 수 있다는 점에서, 기존의 방법으로는 적절한 개수의 토픽을 설정하는데 어려움이 따르며, 토픽의 개수에 따라 결과가 달라지게 된다는 문제점이 있었다. 하지만, 본 개시의 일 실시 예에 따르면, 전술한 방법을 통해, 주어진 문서 집합에 맞게 동적으로 적합한 개수만큼의 토픽을 생성할 수 있다는 점에서, 보다 효과적인 토픽 모델링이 가능해진다.
- [0157] 한편, 프로세서(120)는 토픽 노드에 포함된 문서의 개수에 기초하여 복수의 토픽 노드 중에서 도미넌트(dominant) 토픽 노드를 결정할 수 있다.
- [0158] 구체적으로, 프로세서(120)는 계층적인 구조를 갖는 복수의 토픽 노드 중에서 리프(leaf) 노드에 해당하는 복수의 토픽 노드 각각에 포함된 문서의 개수를 판단하고, 리프 노드에 해당하는 복수의 토픽 노드 중에서 문서의 개수가 가장 많은 토픽 노드를 도미넌트 토픽 노드로 결정할 수 있다.
- [0159] 예를 들어, 도 18과 같은 계층적인 구조를 갖는 복수의 토픽 노드가 생성된 경우를 가정한다.
- [0160] 이 경우, 토픽 노드들(1810, 1820, 1830, 1840, 1850, 1860, 1870)이 트리 구조의 노드에서 리프 노드에 해당한다.
- [0161] 이에 따라, 프로세서(120)는 토픽 노드들(1810, 1820, 1830, 1840, 1850, 1860, 1870)에 포함된 문서의 개수를 판단하고, 이들 토픽 노드들 중에서 하나를 도미넌트 토픽 노드로 결정할 수 있다.
- [0162] 이 경우, 토픽 노드(1810)에 포함된 문서는 41 개로, 토픽 노드들(1810, 1820, 1830, 1840, 1850, 1860, 1870)에서 가장 많은 문서를 포함하고 있다는 점에서, 프로세서(120)는 토픽 노드(1810)를 도미넌트 토픽 노드로 결정할 수 있다.
- [0163] 다만, 리프 노드에 해당하는 복수의 토픽 노드 중에서 적어도 2 개 이상의 토픽 노드에 포함된 복수의 문서의 개수가 다른 토픽 노드에 포함된 문서들의 개수보다 많고, 이들이 서로 동일한 개수의 문서를 포함하고 있는 경우, 프로세서(120)는 이들 토픽 노드에 대해 획득된 팀-토픽 가중치에 기초하여 하나의 토픽 노드를 도미넌트 토픽 노드로 결정할 수 있다.
- [0164] 구체적으로, 프로세서(120)는 이들 토픽 노드 각각에 대해 팀-토픽 가중치를 획득하고, 팀-토픽 가중치의 합이 더 큰 토픽 노드를 도미넌트 토픽 노드로 결정할 수 있다.
- [0165] 이후, 프로세서(120)는 도미넌트 토픽 노드에 포함된 복수의 문서 중에서 하나를 문서 집합의 대표 문서로 결정할 수 있다.
- [0166] 구체적으로, 프로세서(120)는 도미넌트 토픽 노드에 포함된 복수의 문서 중에서 도미넌트 토픽 노드에 포함된 복수의 문서로부터 획득된 복수의 형태소에 기초하여 결정된 하나의 문서를 문서 집합의 대표 문서로 결정할 수 있다.
- [0167] 즉, 프로세서(120)는 도미넌트 토픽 노드에 포함된 복수의 문서로부터 획득된 복수의 형태소에 기초하여 도미넌트 토픽 노드에 포함된 복수의 문서 중에서 하나의 문서를 선택하고, 선택된 문서를 문서 집합의 대표 문서로 결정할 수 있다.
- [0168] 이 경우, 프로세서(120)는 도미넌트 토픽 노드에 포함된 복수의 문서로부터 획득된 복수의 형태소가 복수의 문

서 각각에 포함된 빈도수 및 도미넌트 토픽 노드에서의 복수의 형태소의 가중치에 기초하여 도미넌트 토픽 노드에 포함된 복수의 스코어를 산출하고, 산출된 스코어에 기초하여 도미넌트 토픽 노드에 포함된 복수의 문서 중에서 하나의 문서를 선택할 수 있다.

- [0169] 이때, 프로세서(120)는 도미넌트 토픽 노드에 포함된 복수의 문서 중에서 가장 높은 스코어를 갖는 문서를 선택할 수 있다.
- [0170] 이를 위해, 프로세서(120)는 도미넌트 토픽 노드에 포함된 복수의 문서에 대한 텀-토픽 매트릭스 및 텀-토픽 가중치에 대한 정보를 획득할 수 있다.
- [0171] 구체적으로, 프로세서(120)는 도미넌트 토픽 노드의 부모에 해당하는 토픽 노드에 대한 텀-토픽 매트릭스 및 텀-토픽 가중치에 기초하여, 도미넌트 토픽 노드의 텀-토픽 매트릭스 및 텀-토픽 가중치에 대한 정보를 획득할 수 있다.
- [0172] 예를 들어, 도 19와 같이, 토픽 노드(1930)가 복수의 토픽 노드 중에서 도미넌트 토픽 노드에 해당하고, 토픽 노드(1930)의 부모 토픽 노드가 토픽 노드(1910)인 경우를 가정한다. 이 경우, 토픽 노드(1910)에 대한 텀-토픽 매트릭스(1911)과 같고, 텀-토픽 가중치를 나타내는 매트릭스(1912)와 같이 나타낼 수 있다.
- [0173] 이때, 토픽 노드(1930)에 포함된 문서가 문서 2, 문서 3, ..., 문서 y-1과 같은 경우를 가정한다.
- [0174] 이 경우, 프로세서(120)는 부모 토픽 노드(1910)에 대한 텀-토픽 매트릭스(1911)에서 토픽 노드(1930)에 포함되는 문서에 대응되는 열들(즉, 문서 2에 대한 열, 문서 3에 대한 열, ..., 문서 y-1에 대한 열)을 획득하고, 획득된 열들을 통해 토픽 노드(1930)에 대한 텀-도큐먼트 매트릭스(1931)를 획득할 수 있다.
- [0175] 그리고, 프로세서(120)는 부모 토픽 노드(1910)에 대한 텀-토픽 가중치를 나타내는 매트릭스(1912)에서 토픽 노드(1930)에 대응되는 행(즉, 토픽 2에 대한 행)을 획득하고, 획득된 행을 통해 토픽 노드(1930)에 대한 텀-토픽 가중치를 나타내는 매트릭스(1932)를 획득할 수 있다.
- [0176] 이후, 프로세서(120)는 도미넌트 토픽 노드에 포함된 복수의 문서에 대한 텀-토픽 매트릭스 및 텀-토픽 가중치를 이용하여, 도미넌트 토픽 노드에 포함된 복수의 문서 각각의 스코어를 산출할 수 있다.
- [0177] 이를 위해, 프로세서(120)는 도미넌트 토픽 노드에 대한 텀-토픽 가중치를 내림 차순에 따라 정렬할 수 있다. 즉, 프로세서(120)는 도미넌트 토픽 노드에 대한 텀-토픽 가중치를 크기가 큰 순으로 정렬할 수 있다.
- [0178] 그리고, 프로세서(120)는 하기의 수학식 9에 기초하여, 도미넌트 토픽 노드에 포함된 복수의 문서 각각의 스코어를 산출할 수 있다.

수학식 9

$$\text{score}_j = \frac{\sum_{k=1}^K S_{jk} \times (1/\log(k+1)) \times W_k}{\sum_{k=1}^{10} S_{jk} \times W_k}$$

- [0179]
- [0180] 여기에서, score_j 는 도미넌트 토픽 노드에 포함된 복수의 문서 중에서 j 번째 문서의 스코어를 나타낸다. 여기서, $j=\{1,2,\dots,d\}$ 이고, d는 도미넌트 토픽 노드에 포함된 문서의 개수이다.
- [0181] 그리고, W_k 는 내림 차순에 따라 정렬된 텀-토픽 가중치 중에서 k 번째 가중치를 나타내고, S_{jk} 는 j 번째 문서에 포함된 형태소의 빈도수 중에서, k 번째 가중치에 대응되는 형태소의 빈도수를 나타낸다. 또한, K는 기설정된 값으로, 10 보다 작거나 같은 값이 될 수 있으며, 일 예로, K는 3이 될 수 있다. 다만, K 값은 변경 가능함은 물론이다.
- [0182] 예를 들어, 도 19에 도시된 텀-토픽 가중치를 크기가 큰 순으로 정렬할 때, 재정렬된 텀-토픽 가중치가 도 20의 매트릭스(2020)와 같이 나타내어지는 경우를 가정한다.
- [0183] 이때, K가 3인 경우, 도미넌트 토픽 노드에 포함된 복수의 문서 각각에 포함된 문서의 스코어는 도 20과 같이 나타낼 수 있다.
- [0184] 이와 같이, 프로세서(120)는 도미넌트 토픽 노드에 포함된 복수의 문서 각각에 대한 스코어를 산출할 수 있다.

- [0185] 그리고, 프로세서(120)는 도미넌트 토픽 노드에 포함된 복수의 문서 중에서 가장 높은 스코어를 갖는 문서를 선택하고, 선택된 문서를 문서 집합의 대표 문서로 결정할 수 있다. 전술한 도 20의 예에서, 문서 2의 스코어가 가장 높은 경우, 프로세서(120)는 문서 2를 문서 집합에 포함된 복수의 문서 중 대표 문서로 결정할 수 있다.
- [0186] 이와 같은 방법에 따라 산출된 스코어에 기초하여 대표 문서를 결정할 경우, 도미넌트 토픽 노드에 대한 매칭률이 가장 높은 문서가 도미넌트 토픽 노드에서 선택되고, 해당 문서가 문서 집합의 대표 문서가 될 수 있다.
- [0187] 결국, 본 개시의 일 실시 예에 따르면, 전술한 방법을 통해 문서 집합에 포함된 복수의 문서 중에서 대표 문서를 판단할 수 있게 된다.
- [0188] 이러한 방법에 따라 대표 문서를 결정하게 되면, 대표 문서가 전체 문서 집합의 요약으로서의 역할이 가능해진다.
- [0189] 구체적으로, 토픽은 문서 집합 내에 존재하는 형태소들에 대한 확률 분포로 정의되는데, 특정 형태소의 확률 값이 높을수록 그 토픽의 의미가 해당 형태소를 많이 반영하는 것으로 볼 수 있다. 한편, 주어진 문서 집합에 대해 토픽 모델링을 수행하는 경우, 여러 개의 서로 다른 토픽이 나오게 되고, 여러 문서들은 그들 내에 존재하는 세부 주제 즉, 각 토픽 별로 문서군을 형성하게 된다. 즉, 이러한 세부 주제 하나하나가 각 토픽이 되는 것이고, 해당 세부 주제를 가장 잘 나타내는 형태소들이 높은 확률을 할당받게 되는 식으로 특정 토픽의 확률 분포가 결정된다. 또한, 각 문서는 내용 상 가장 관련도가 높은 토픽으로 분류되게 된다. 따라서, 문서 집합에서 도출된 복수의 토픽에서 도미넌트 토픽을 결정하고, 도미넌트 토픽에서 대표 문서를 선택하는 경우, 선택된 대표 문서가 전체 문서 집합의 요약으로서의 역할이 가능해진다.
- [0190] 도 21은 본 개시의 일 실시 예에 따른 전자 장치의 구성을 상세히 도시한 블록도이다.
- [0191] 도 21에 도시된 바와 같이, 전자 장치(100)는 메모리(110) 및 프로세서(120), 통신 인터페이스(130)를 포함할 수 있다. 이 경우, 프로세서(120)는 이들 구성요소와 전기적으로 연결되어 전자 장치(100)의 전반적인 동작 및 기능을 제어할 수 있다. 한편, 전자 장치(100)의 구현 예에 따라 도 21에 도시된 구성요소 이외에 다른 구성요소가 추가될 수 있음은 물론이다. 한편, 도 21에 도시된 메모리(110), 프로세서(120)는 전술한 바 있다는 점에서, 중복되는 설명은 생략하기로 한다.
- [0192] 프로세서(120)는 중앙처리장치, 어플리케이션 프로세서 및 커뮤니케이션 프로세서 중 하나 또는 그 이상을 포함할 수 있다. 프로세서(120)는 전자 장치(100)를 제어할 수 있다. 예를 들어, 프로세서(120)는 전자 장치(100)의 적어도 하나의 구성요소들의 제어 및 연산이나 데이터 처리 등을 실행할 수 있다.
- [0193] 통신 인터페이스(130)는 외부 전자 장치와 통신을 수행하기 위한 구성으로 회로(circuitry)를 포함할 수 있다. 이 경우, 통신 인터페이스(130)가 외부 전자 장치와 통신 연결되는 것은 제3 기기(예로, 중계기, 허브, 액세스 포인트, 서버 또는 게이트웨이 등)를 거쳐서 통신하는 것을 포함할 수 있다.
- [0194] 예를 들어, 통신 인터페이스(130)는 LTE, WiFi(wireless fidelity) 등과 같은 무선 통신 및, USB(universal serial bus), HDMI(high definition multimedia interface) 등과 같은 유선 통신을 위한 구성요소를 포함할 수 있으며, 무선 통신 또는 유선 통신이 수행되는 네트워크는 텔레커뮤니케이션 네트워크, 예를 들면, 컴퓨터 네트워크(예: LAN 또는 WAN), 인터넷, 또는 텔레폰 네트워크 중 적어도 하나를 포함할 수 있다.
- [0195] 한편, 전자 장치(100)가 검색 서비스를 제공하는 서버로 구현되는 경우, 프로세서(120)는 외부 전자 장치에 입력된 사용자 커맨드(command)에 기초하여 복수의 문서 중에서 대표 문서를 결정하고, 대표 문서에 대한 정보(가령, 대표 문서의 이미지, 요약 내용, 타이틀 등)를 통신 인터페이스(130)를 통해 외부 전자 장치로 전송할 수 있다. 이 경우, 프로세서(120)는 전술한 방법을 통해 복수의 문서에서 대표 문서를 판단할 수 있다.
- [0196] 이하에서는, 도 22 내지 도 26을 참조하여, 본 개시의 일 실시 예에 따라 대표 문서를 제공하는 방법을 설명하도록 한다.
- [0197] 프로세서(120)는 외부 전자 장치에 입력된 키워드를 통신 인터페이스(130)를 통해 외부 전자 장치로부터 수신받을 수 있다. 예를 들어, 사용자는 도 22와 같이, 외부 전자 장치(가령, 핸드폰)를 통해 전자 장치(100)에서 제공하는 웹 페이지(2210)에 접속하고, 웹 페이지(2210)에 포함된 검색 창(2220)에 자신이 검색하고자 하는 키워드를 입력할 수 있다.
- [0198] 이 경우, 프로세서(120)는 복수의 기사가 저장된 데이터베이스에서 키워드와 관련된 문서를 검색하고, 통신 인터페이스(130)를 통해 검색된 문서에 대한 정보를 외부 전자 장치로 전송할 수 있다. 이때, 외부 전자 장치에는

도 23과 같은 검색 결과 화면이 표시될 수 있다. 검색 결과 화면에는 기사 별로, 기사에 대한 이미지, 기사 제목 및 기사의 일부 내용이 포함될 수 있다.

- [0199] 그리고, 프로세서(120)는 검색 결과 화면에서 하나의 기사를 선택하는 사용자 커맨드가 외부 전자 장치로부터 수신되면, 선택된 기사의 전체 내용을 외부 전자 장치로 전송할 수 있다. 이 경우, 외부 전자 장치에는 도 24와 같이, 전체 기사 내용을 포함하는 화면에 표시될 수 있다.
- [0200] 한편, 도 24와 같이, 전체 기사 내용을 포함하는 화면은 대표 기사에 대한 정보를 요청하기 위한 오브젝트(2410)를 포함할 수 있다.
- [0201] 이에 따라, 오브젝트(2410)를 선택하는 사용자 커맨드가 외부 전자 장치로부터 수신되면, 프로세서(120)는 전술한 방법을 이용하여, 키워드에 따라 검색된 복수의 기사에서 대표 기사를 판단하고, 대표 기사에 대한 정보를 외부 전자 장치로 전송할 수 있다.
- [0202] 이때, 프로세서(120)는 시간에 따라 기사가 등록된 횟수를 고려하여 검색된 기사를 복수의 그룹으로 구분하고, 복수의 그룹 각각에 포함된 복수의 기사에서 대표 기사를 판단할 수 있다.
- [0203] 이를 위해, 프로세서(120)는 검색된 기사가 데이터베이스에 등록된 시간을 판단할 수 있다. 그리고, 프로세서(120)는 시간 별로 등록된 기사의 개수를 판단하고, 기사가 등록된 개수가 피크가 되는 적어도 하나의 시간을 판단할 수 있다.
- [0204] 예를 들어, 시간에 따라 외부의 서버에 등록된 문서의 개수가 도 25와 같은 경우, 프로세서(120)는 기설정된 시간 구간에서 등록된 문서가 가장 많은 시점(2510, 2520, 2530)을 판단할 수 있다.
- [0205] 그리고, 프로세서(120)는 복수의 피크 지점을 기준으로 복수의 시간 구간을 설정하고, 각 시간 구간에 포함된 문서들 별로 문서 집합을 구성할 수 있다.
- [0206] 즉, 도 25와 같은 경우, 프로세서(120)는 시점(2510)을 포함하는 제1 시간 구간(2540), 시점(2620)을 포함하는 제2 시간 구간(2550), 시점(2530)을 포함하는 제3 시간 구간(2560)을 각각 설정하고, 각 시간 구간에 포함되는 문서들을 이용하여 문서 집합을 구성할 수 있다. 가령, 프로세서(120)는 제1 시간 구간(2540)에 포함되는 복수의 문서로 하나의 문서 집합을 구성하고, 제2 시간 구간(2550)에 포함되는 복수의 문서로 하나의 문서 집합을 구성하고, 제3 시간 구간(2560)에 포함되는 문서로 하나의 문서 집합을 구성할 수 있다.
- [0207] 그리고, 프로세서(120)는 전술한 방법을 통해, 각 시간 구간에 포함되는 복수의 문서에서 대표 문서를 각각 판단하고, 각 시간 구간에 대해 판단된 대표 문서에 대한 정보를 통신 인터페이스(130)를 통해 외부 전자 장치로 전송할 수 있다.
- [0208] 이 경우, 도 26과 같이, 외부 전자 장치에는 복수의 시간 구간 각각에서 판단된 대표 기사의 이미지 및 대표 기사의 타이틀이 타임 라인 상에 배열되어 표시될 수 있다.
- [0209] 이에 따라, 본 개시의 일 실시 예에 따르면, 사용자는 키워드와 관련된 추가적인 검색 없이도, 깊이 있는 정보를 제공받을 수 있게 된다.
- [0210] 한편, 전술한 예뿐만 아니라, 문서 집합은 다양하게 획득될 수 있다. 가령, 특정 년도(가령, 2018년, 2017년)에 발행된 복수의 기사가 하나의 문서 집합을 구성할 수 있고, 특정한 주제(가령, 스포츠 등)에 대한 복수의 기사가 하나의 문서 집합을 수도 있다.
- [0211] 또한, 전술한 예에서는 전자 장치가 외부 전자 장치로 대표 문서에 대한 정보를 제공하는 것으로 설명하였으나, 이는 일 예에 불과하다.
- [0212] 구체적으로, 프로세서(120)는 키워드에 따라 검색된 복수의 문서를 외부 전자 장치로부터 수신하고, 수신된 복수의 문서에서 대표 문서를 판단하고, 판단된 대표 문서를 전자 장치(100)의 디스플레이(미도시)에 표시할 수도 있다.
- [0213] 이를 위해, 전자 장치(100)는 디스플레이(미도시) 및 입력 인터페이스(미도시)를 포함할 수 있다.
- [0214] 이 경우, 디스플레이(미도시)는 다양한 화면을 표시할 수 있다. 그리고, 입력 인터페이스(미도시)는 다양한 사용자 커맨드(command)을 입력받기 위한 구성으로, 회로(circuitry)를 포함할 수 있다. 이 경우, 입력 인터페이스(미도시)는 입력된 사용자 커맨드를 프로세서(120)로 전달할 수 있다. 입력 인터페이스(미도시)는 예를 들면, 터치 패널, 키 등을 포함할 수 있다.

- [0215] 도 27은 본 개시의 일 실시 예에 따른 전자 장치의 대표 문서 선정 방법을 설명하기 위한 흐름도이다.
- [0216] 먼저, 토픽 모델링을 통해 문서 집합을 계층적인 구조를 갖는 복수의 토픽 노드에 따라 분할한다(S2710).
- [0217] 이후, 토픽 노드에 포함된 문서의 개수에 기초하여 복수의 토픽 노드 중에서 도미넌트 토픽 노드를 결정한다(S2720).
- [0218] 그리고, 도미넌트 토픽 노드에 포함된 복수의 문서 중에서 도미넌트 토픽 노드에 포함된 복수의 문서로부터 획득된 복수의 형태소에 기초하여 결정된 하나의 문서를 문서 집합의 대표 문서로 결정한다(S2730).
- [0219] 한편, S2710 단계는 문서 집합에 포함된 복수의 문서에 기초하여 획득된 텀-도큐먼트 매트릭스에 비음수 행렬 분해를 수행하여, 문서 집합에 대응되는 노드를 기설정된 개수의 제1 토픽 노드들로 분할하며, 텀-도큐먼트 매트릭스는 문서 집합에 포함된 복수의 문서에서 획득된 복수의 형태소가 상기 복수의 문서 각각에 포함된 빈도수에 기초하여 생성될 수 있다.
- [0220] 또한, S2710 단계는 제1 토픽 노드들을 각각 분할하였을 때 생성되는 자식 토픽 노드들에 기초하여 제1 토픽 노드들 각각의 스코어를 산출하고, 산출된 스코어에 기초하여 제1 토픽 노드들 중에서 분할될 토픽 노드를 결정하고, 결정된 토픽 노드를 기설정된 개수의 제2 토픽 노드들로 분할할 수 있다. 그리고, S2410 단계는 제1 토픽 노드들 중 분할되지 않는 토픽 노드, 및 제2 토픽 노드들을 각각 분할하였을 때 생성되는 자식 토픽 노드들에 기초하여 분할되지 않는 토픽 노드 및 제2 토픽 노드들 각각의 스코어를 산출하고, 산출된 스코어에 기초하여 분할되지 않는 토픽 노드 및 제2 토픽 노드들 중에서 분할될 토픽 노드를 결정하고, 결정된 토픽 노드를 기설정된 개수의 제3 토픽 노드들로 분할할 수 있다.
- [0221] 이 경우, S2710 단계는 기설정된 조건을 만족하는 토픽 노드를 제외하면서 분할 과정을 수행하여 계층적인 구조를 갖는 복수의 토픽 노드를 생성하며, 기설정된 조건은 토픽 노드를 분할하였을 때 생성되는 자식 토픽 노드들에 기초하여 산출된 스코어가 음수인 조건, 및 토픽 노드 및 자식 토픽 노드 간 또는 자식 토픽 노드들 간에 기설정된 개수 이상의 동일한 빈도수 상위 형태소가 존재하는 조건을 포함할 수 있다.
- [0222] 한편, S2720 단계는 계층적인 구조를 갖는 복수의 토픽 노드 중에서 리프 노드에 해당하는 복수의 토픽 노드 각각에 포함된 문서의 개수를 판단하고, 리프 노드에 해당하는 복수의 토픽 노드 중에서 문서의 개수가 가장 많은 토픽 노드를 도미넌트 토픽 노드로 결정할 수 있다.
- [0223] 한편, S2730 단계는 도미넌트 토픽 노드에 포함된 복수의 문서로부터 획득된 복수의 형태소가 복수의 문서 각각에 포함된 빈도수 및 도미넌트 토픽 노드에서의 복수의 형태소의 가중치에 기초하여 도미넌트 토픽 노드에 포함된 복수의 문서의 스코어를 산출하고, 산출된 스코어에 기초하여 도미넌트 토픽 노드에 포함된 복수의 문서 중에서 하나의 문서를 선택하고, 선택된 문서를 대표 문서로 결정할 수 있다.
- [0224] 한편, 문서 집합에서 대표 문서를 결정하는 구체적인 방법에 대해서는 전술한 바 있다.
- [0225] 한편, 본 개시에서 사용된 용어 "부" 또는 "모듈"은 하드웨어, 소프트웨어 또는 펌웨어로 구성된 유닛을 포함하며, 예를 들면, 로직, 논리 블록, 부품, 또는 회로 등의 용어와 상호 호환적으로 사용될 수 있다. "부" 또는 "모듈"은, 일체로 구성된 부품 또는 하나 또는 그 이상의 기능을 수행하는 최소 단위 또는 그 일부가 될 수 있다. 예를 들면, 모듈은 ASIC(application-specific integrated circuit)으로 구성될 수 있다.
- [0226] 본 개시의 다양한 실시예들은 기기(machine)(예: 컴퓨터)로 읽을 수 있는 저장 매체(machine-readable storage media)에 저장된 명령어를 포함하는 소프트웨어로 구현될 수 있다. 기기는, 저장 매체로부터 저장된 명령어를 호출하고, 호출된 명령어에 따라 동작이 가능한 장치로서, 개시된 실시예들에 따른 전자 장치(예: 전자 장치(100))를 포함할 수 있다. 명령이 프로세서에 의해 실행될 경우, 프로세서가 직접, 또는 프로세서의 제어 하에 다른 구성요소들을 이용하여 상기 명령에 해당하는 기능을 수행할 수 있다. 명령은 컴파일러 또는 인터프리터에 의해 생성 또는 실행되는 코드를 포함할 수 있다. 기기로 읽을 수 있는 저장매체는, 비일시적(non-transitory) 저장매체의 형태로 제공될 수 있다. 여기서, '비일시적'은 저장매체가 신호(signal)를 포함하지 않으며 실재(tangible)하다는 것을 의미할 뿐 데이터가 저장매체에 반영구적 또는 임시적으로 저장됨을 구분하지 않는다.
- [0227] 일시예에 따르면, 본 개시에 개시된 다양한 실시예들에 따른 방법은 컴퓨터 프로그램 제품(computer program product)에 포함되어 제공될 수 있다. 컴퓨터 프로그램 제품은 상품으로서 판매자 및 구매자 간에 거래될 수 있다. 컴퓨터 프로그램 제품은 기기로 읽을 수 있는 저장 매체(예: compact disc read only memory (CD-ROM))의 형태로, 또는 어플리케이션 스토어를 통해 온라인으로 배포될 수 있다. 온라인 배포의 경우에, 컴퓨터 프로그램 제품의 적어도 일부는 제조사의 서버, 어플리케이션 스토어의 서버, 또는 중계 서버의 메모리와 같은 저장 매체

에 적어도 일시 저장되거나, 임시적으로 생성될 수 있다.

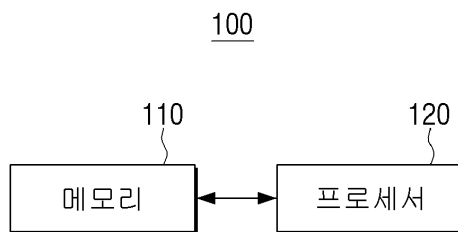
[0228] 다양한 실시예들에 따른 구성 요소(예: 모듈 또는 프로그램) 각각은 단수 또는 복수의 개체로 구성될 수 있으며, 전술한 해당 서브 구성 요소들 중 일부 서브 구성 요소가 생략되거나, 또는 다른 서브 구성 요소가 다양한 실시예에 더 포함될 수 있다. 대체적으로 또는 추가적으로, 일부 구성 요소들(예: 모듈 또는 프로그램)은 하나의 개체로 통합되어, 통합되기 이전의 각각의 해당 구성 요소에 의해 수행되는 기능을 동일 또는 유사하게 수행할 수 있다. 다양한 실시예들에 따른, 모듈, 프로그램 또는 다른 구성 요소에 의해 수행되는 동작들은 순차적, 병렬적, 반복적 또는 휴리스틱하게 실행되거나, 적어도 일부 동작이 다른 순서로 실행되거나, 생략되거나, 또는 다른 동작이 추가될 수 있다.

부호의 설명

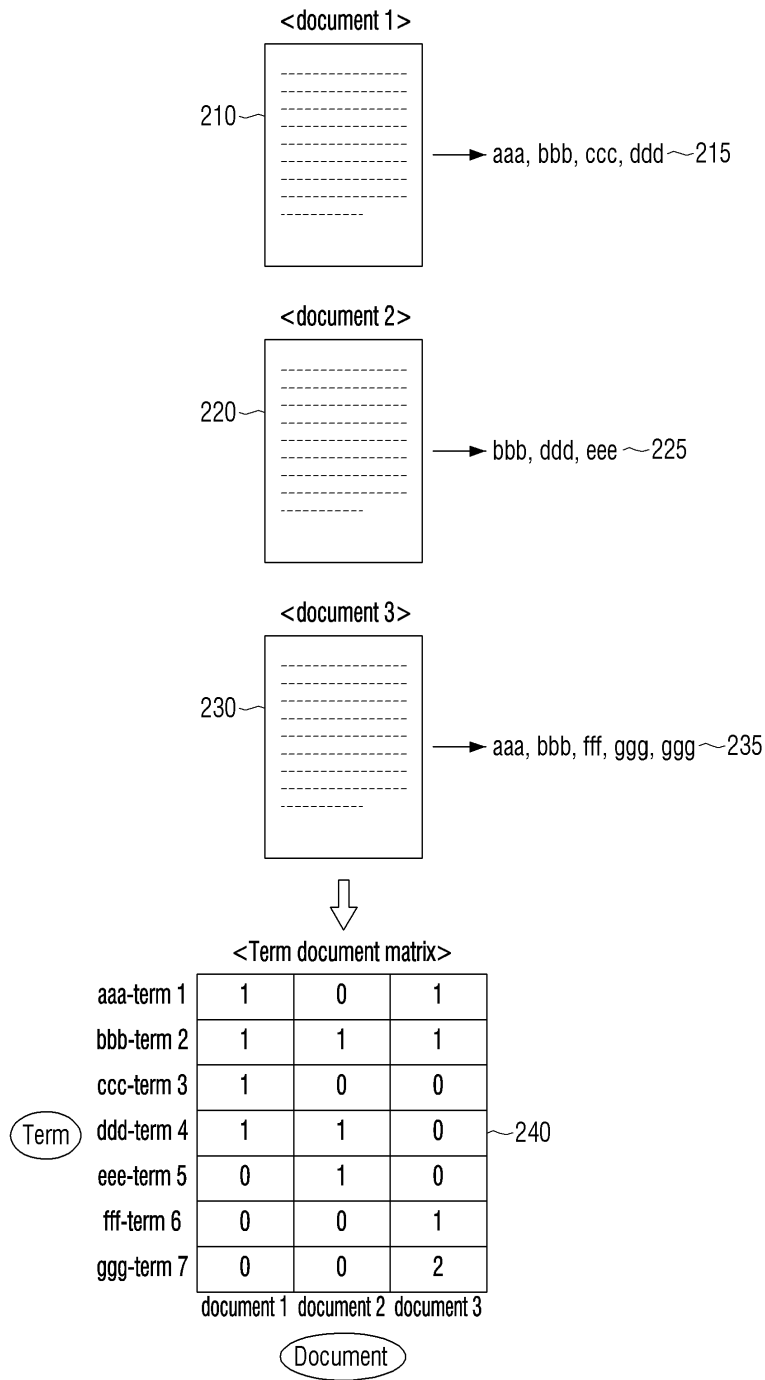
[0229] 100 : 전자 장치 110 : 메모리
120 : 프로세서

도면

도면1



도면2



도면3

term 1	a_{11}	a_{12}	...	a_{1M}
term 2	a_{21}	a_{22}	...	a_{2M}
term 3	a_{31}	a_{32}	...	a_{3M}
⋮	⋮	⋮		⋮
term N	a_{N1}	a_{N2}	...	a_{NM}
	document 1	document 2	...	document M

≫ 비음수 행렬 분해

term 1	w_{11}	...	w_{1K}
term 2	w_{21}	...	w_{2K}
term 3	w_{31}	...	w_{3K}
⋮	⋮		⋮
term N	w_{N1}	...	w_{NK}
	topic 1	...	topic K

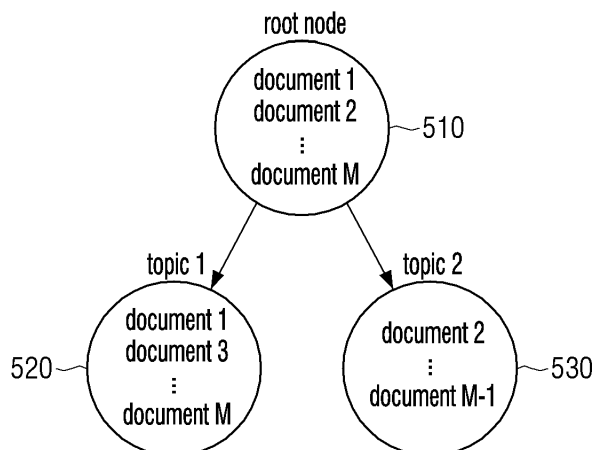
×

topic 1	w'_{11}	w'_{12}	...	w'_{1M}
⋮	⋮	⋮		⋮
topic K	w'_{K1}	w'_{K2}	...	w'_{KM}
	document 1	document 2	...	document M

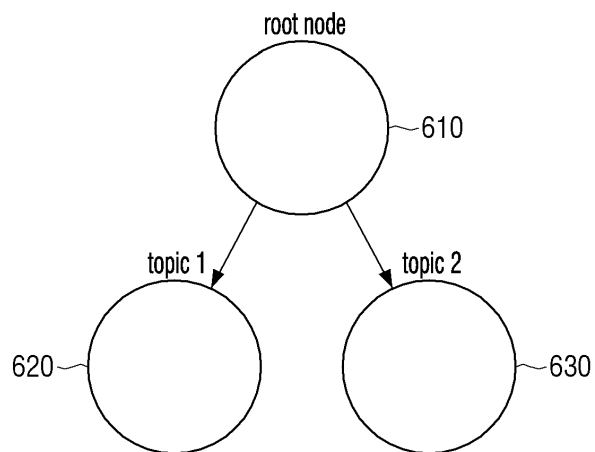
도면4

topic 1	w'_{11}	w'_{12}	w'_{13}	...	w'_{1M-1}	w'_{1M}
topic 2	w'_{21}	w'_{22}	w'_{23}	...	w'_{2M-1}	w'_{2M}
	document 1	document 2	document 3	...	document M-1	document M

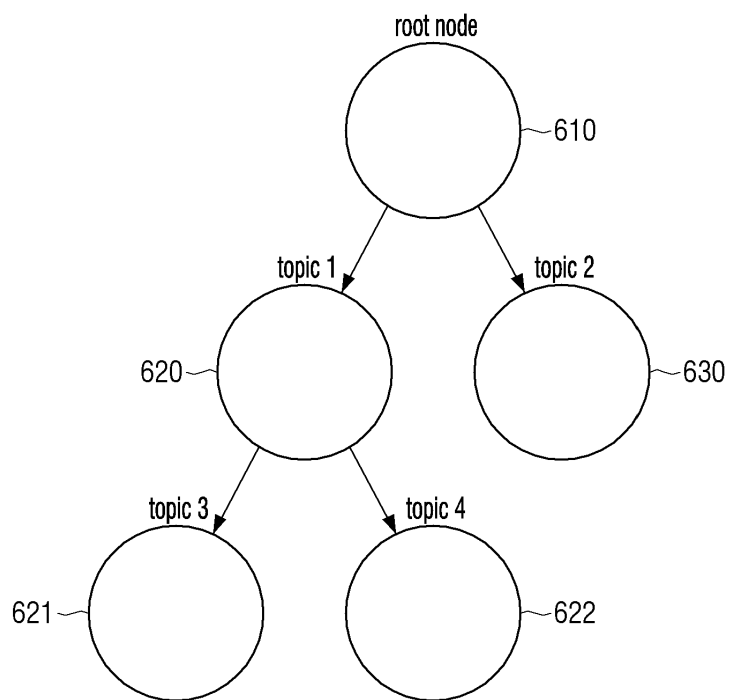
도면5



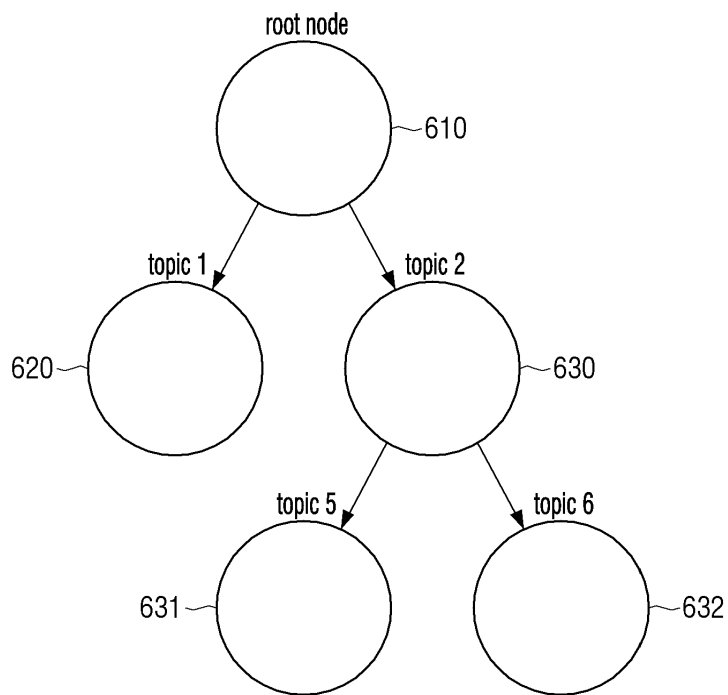
도면6



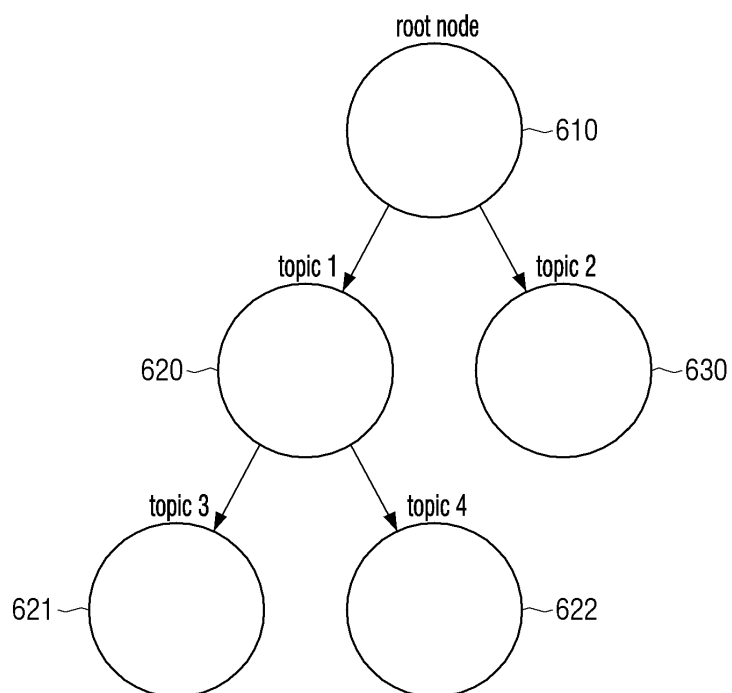
도면7



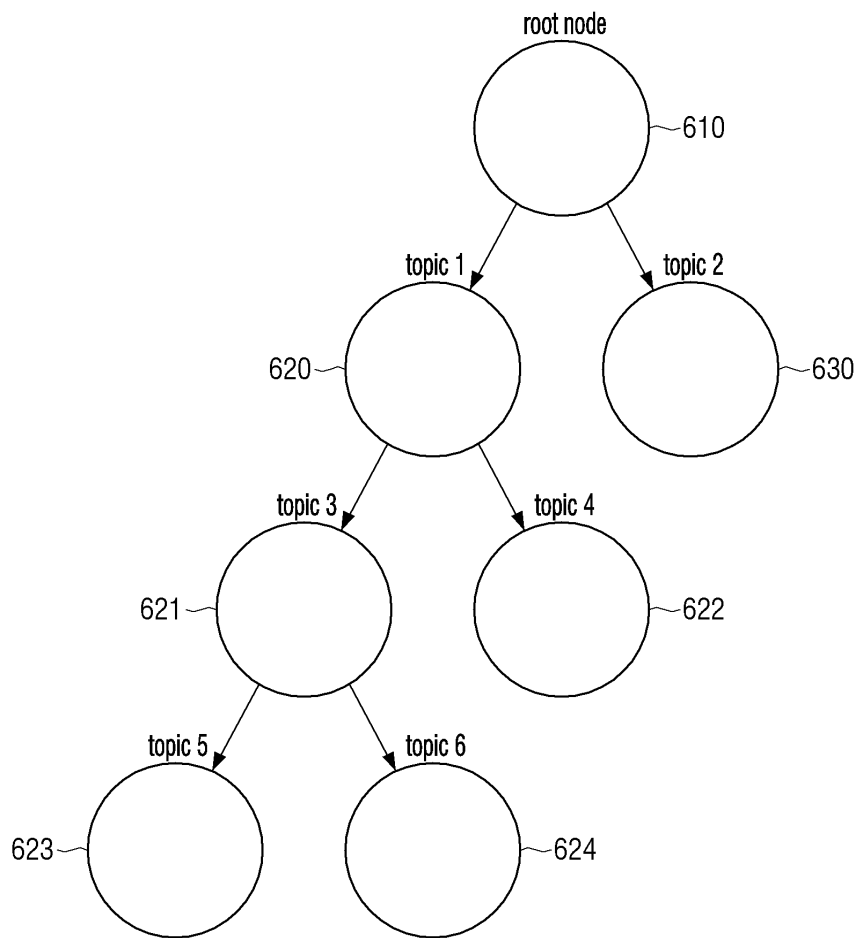
도면8



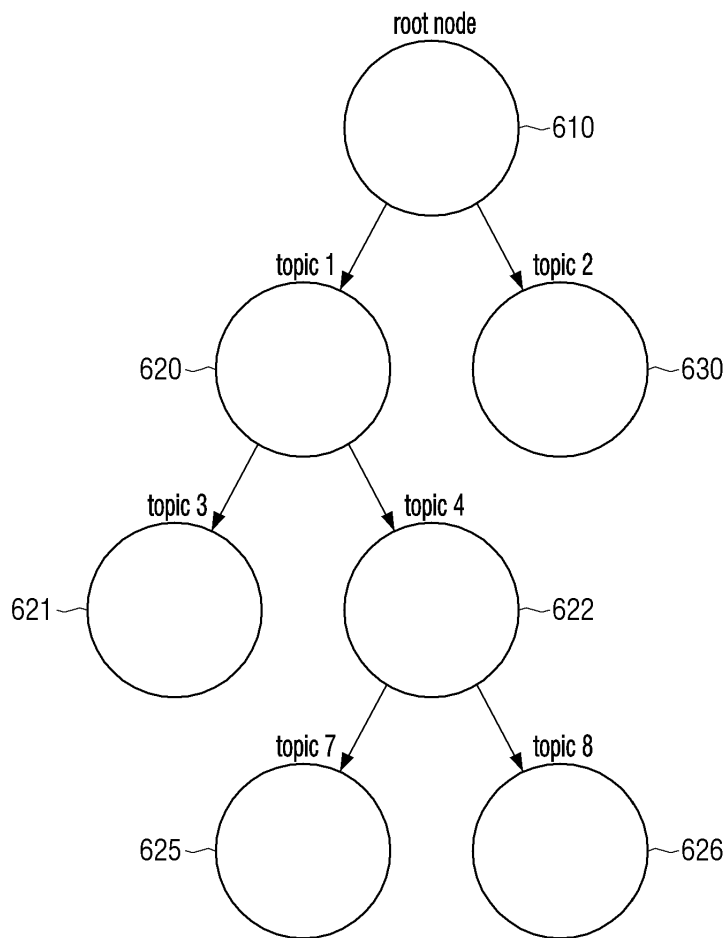
도면9



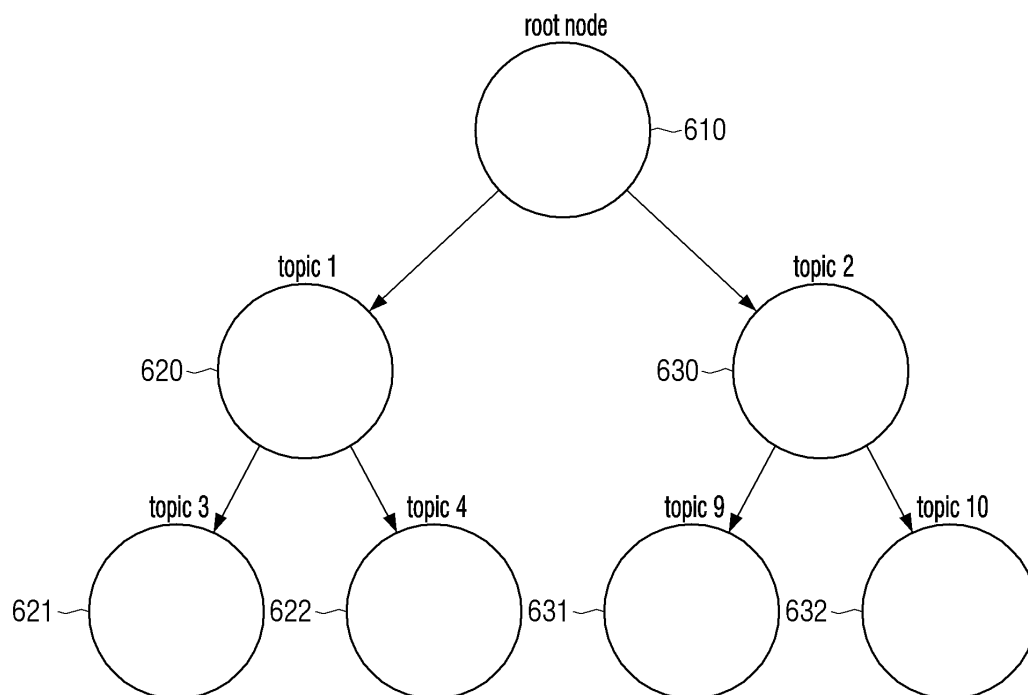
도면10



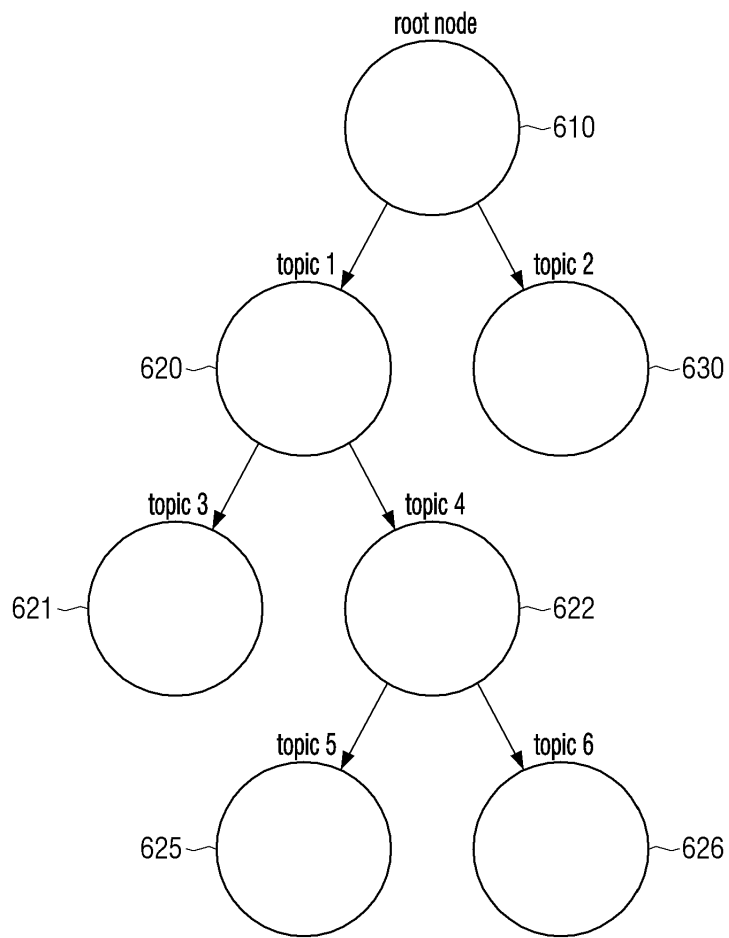
도면11



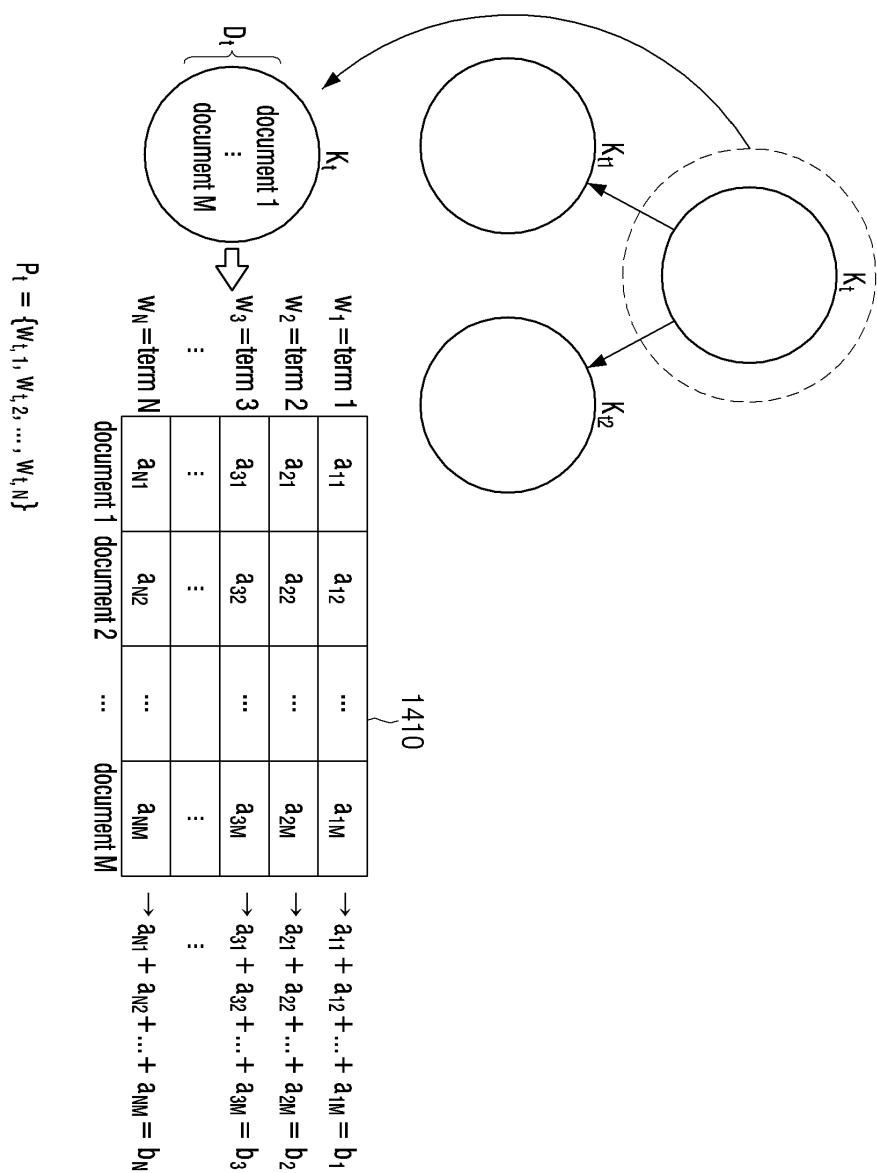
도면12



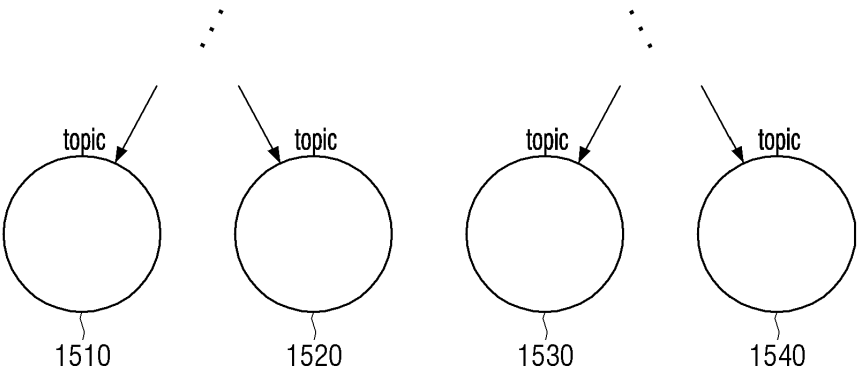
도면13



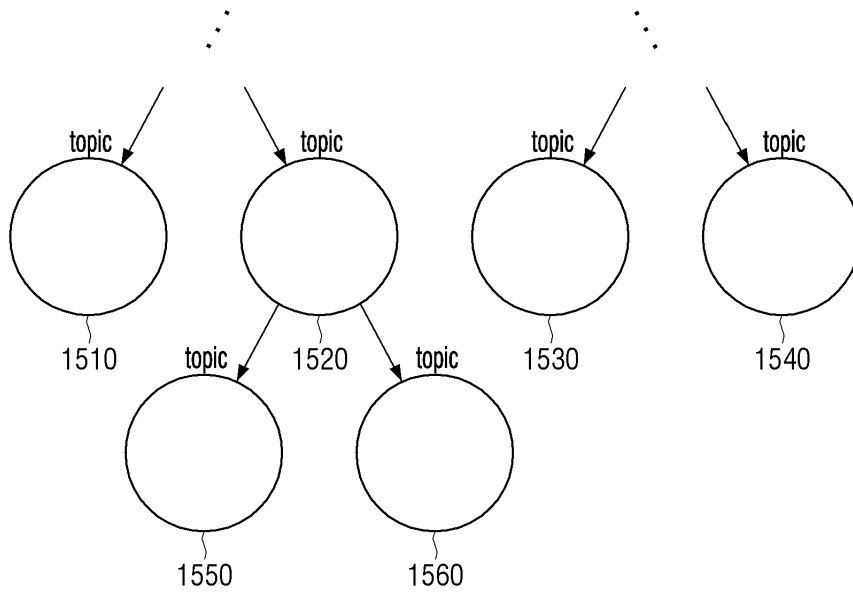
도면14



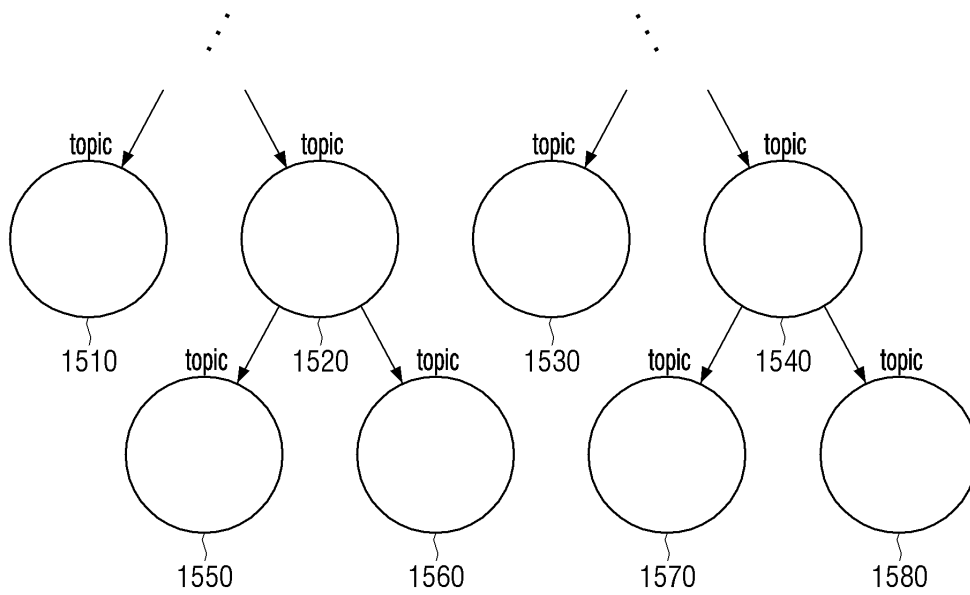
도면15



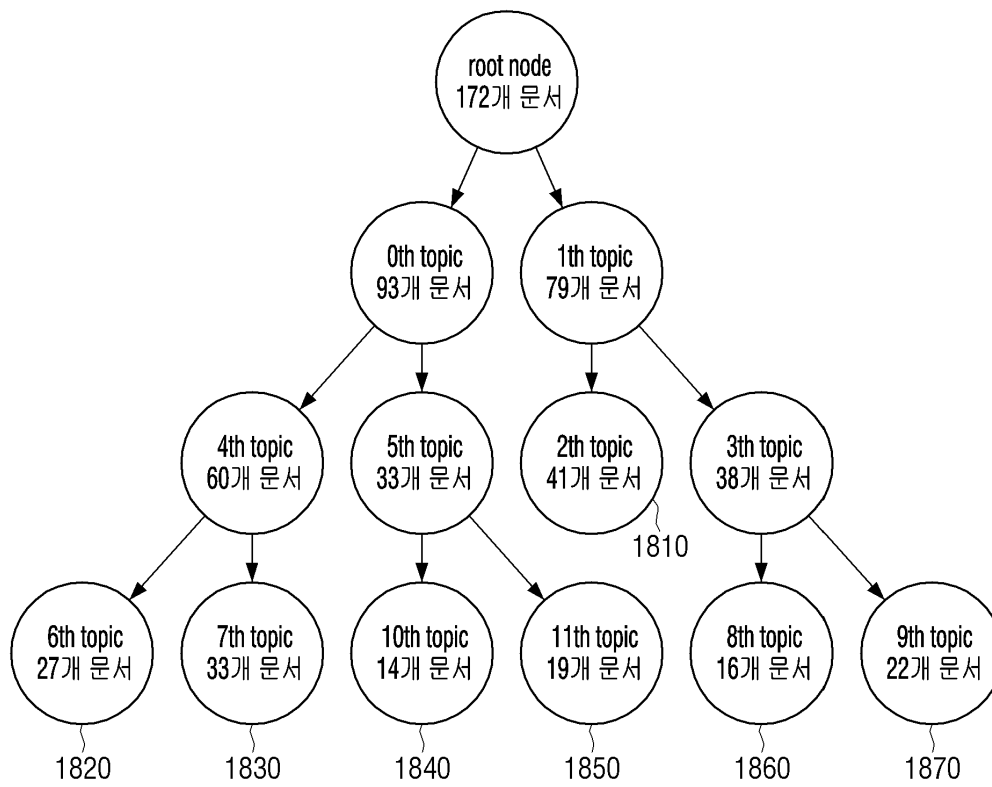
도면16



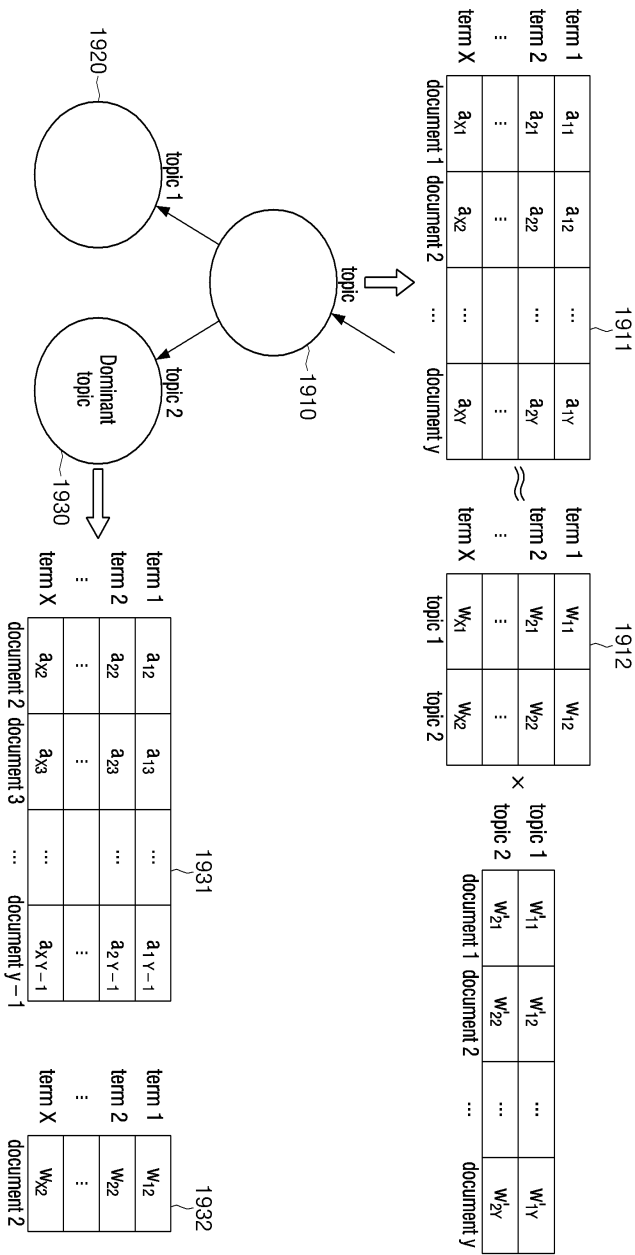
도면17



도면18



도면19



도면20

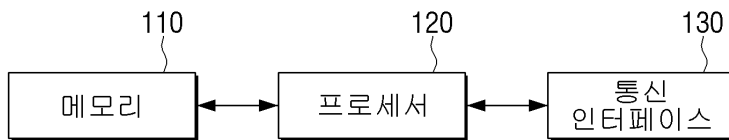
	2010					2020	
term 1	a_{12}	a_{13}	...	a_{1Y-1}	term 4	w_{42}	
term 2	a_{22}	a_{23}	...	a_{2Y-1}	term 8	w_{82}	
⋮	⋮	⋮		⋮	term 2	w_{22}	
term X	a_{X2}	a_{X3}	...	a_{XY-1}	⋮	⋮	
	document 2	document 3	...	document y-1	term 5	w_{52}	
						topic 2	

$$\text{문서2의 스코어} = \frac{(a_{42} \times \frac{1}{\log_2} \times w_{42}) + (a_{82} \times \frac{1}{\log_3} \times w_{82}) + (a_{22} \times \frac{1}{\log_4} \times w_{22})}{(a_{42} \times w_{42}) + (a_{82} \times w_{82}) + \dots + (a_{52} \times w_{52})}$$

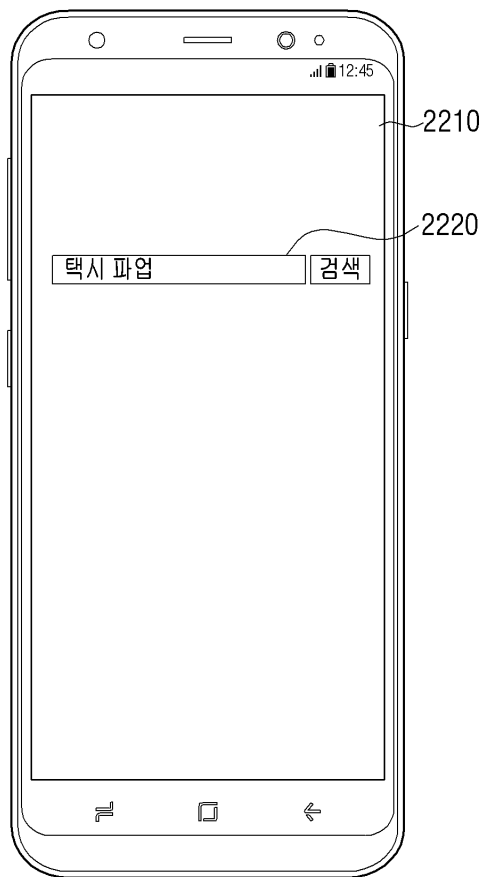
$$\text{문서3의 스코어} = \frac{(a_{43} \times \frac{1}{\log_2} \times w_{42}) + (a_{83} \times \frac{1}{\log_3} \times w_{82}) + (a_{23} \times \frac{1}{\log_4} \times w_{22})}{(a_{43} \times w_{42}) + (a_{83} \times w_{82}) + \dots + (a_{53} \times w_{52})}$$

$$\text{문서y-1의 스코어} = \frac{(a_{4Y-1} \times \frac{1}{\log_2} \times w_{42}) + (a_{8Y-1} \times \frac{1}{\log_3} \times w_{82}) + (a_{2Y-1} \times \frac{1}{\log_4} \times w_{22})}{(a_{4Y-1} \times w_{42}) + (a_{8Y-1} \times w_{82}) + \dots + (a_{5Y-1} \times w_{52})}$$

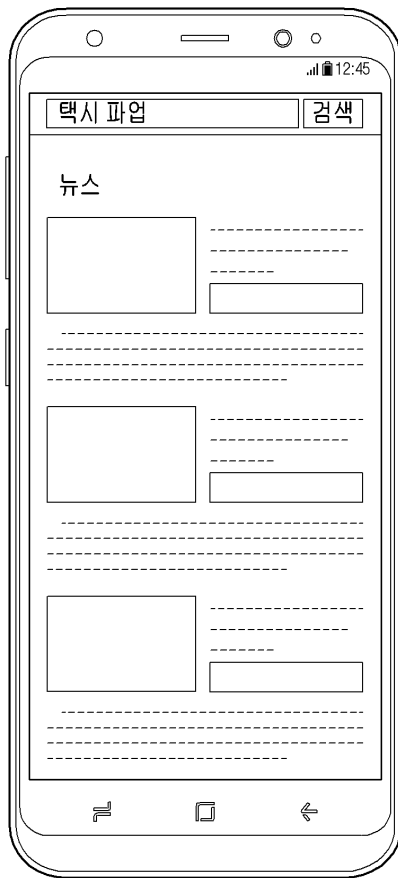
도면21



도면22



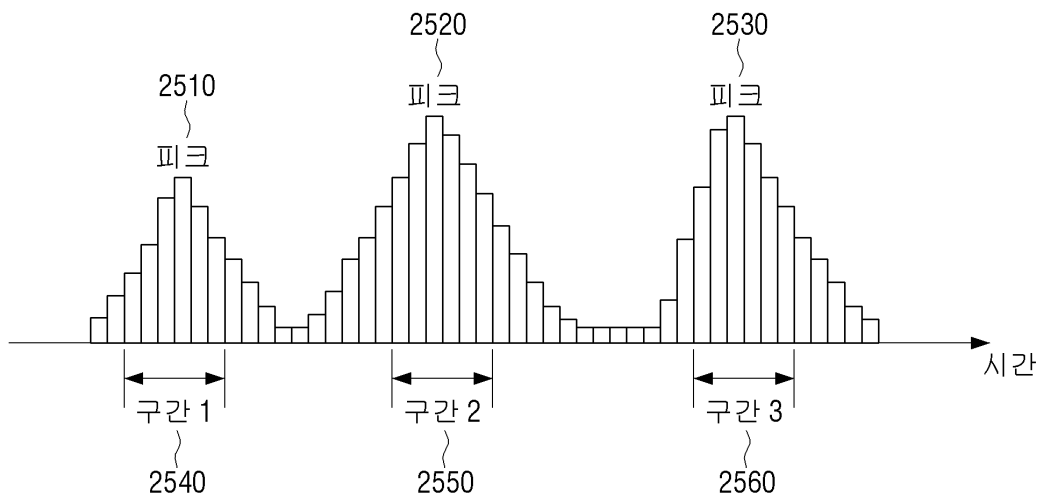
도면23



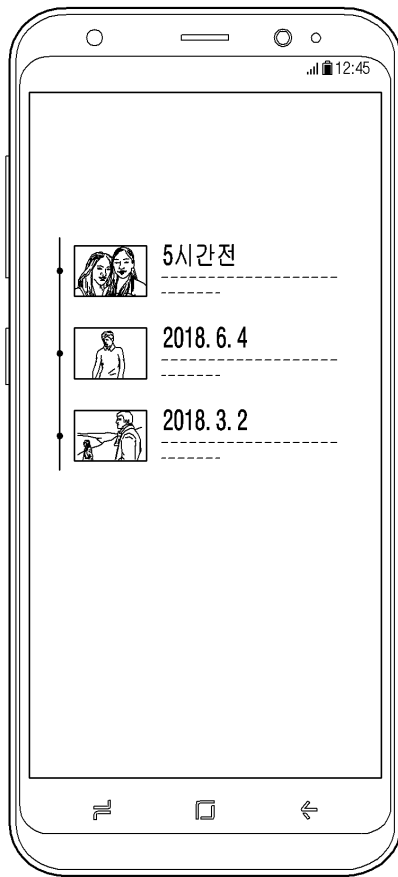
도면24



도면25



도면26



도면27

