(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2017년03월09일
(11) 등록번호 10-1713831
(24) 등록일자 2017년03월02일

(51) 국제특허분류(Int. Cl.)
*G06F 17/30* (2006.01)
(52) CPC특허분류
*G06F 17/30421* (2013.01)
(21) 출원번호 10-2016-0094921
(22) 출원일자 2016년07월26일
심사청구일자 2016년07월26일
(56) 선행기술조사문헌
이원구 외7명. 이기종 데이터간 상호운여적 분류체계 관리를 위한 분류체계 자동화 방안. 한국정보통신학회논문지. 한국정보통신학회. 2011년 12월, 제15권 제12호, 2609-2618페이지.
WO2015183098 A1
KR1020100080099 A
JP2008040985 A

(73) 특허권자
**한국과학기술정보연구원**
대전광역시 유성구 대학로 245 (어은동)
(72) 발명자
**정유철**
대전광역시 유성구 봉명로 48, 802동 403호 (원신흥동, 신안인스빌리베라)
**김광영**
대전광역시 유성구 온천북로33번길 21-13, 명일시티 301호 (봉명동)
*(뒷면에 계속)*
(74) 대리인
**김용인, 지관영**

전체 청구항 수 : 총 12 항 심사관 : 최정권

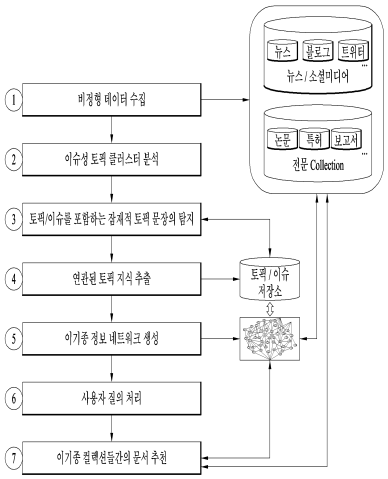
(54) 발명의 명칭 **문서추천장치 및 방법**

## (57) 요 약

본 발명은 문서추천장치 및 방법에 관한 것이다.

이를 위해, 본 발명은 DB로부터 텍스트 데이터를 수집하는 단계; 상기 수집된 텍스트 데이터를 토픽별로 클러스터링하는 단계; 상기 토픽별로 클러스터링된 텍스트 데이터에서 이벤트를 포함하는 토픽 문장을 탐지하는 단계; 상기 탐지한 토픽 문장과 연관된 지식을 추출하는 단계; 개별 텍스트 컬렉션으로부터 워드벡터들을 생성하는 단계; 상기 추출된 연관된 지식과 상기 생성된 워드 벡터들을 결합하여 이기종 정보 네트워크를 생성하는 단계; 입력단어에 대응하여 상기 생성된 네트워크 내의 타겟 텍스트 컬렉션으로부터 상기 워드벡터를 활용하여 확장단어를 생성하는 단계; 및 상기 확장단어를 기반으로 상기 타겟 컬렉션으로부터 문서를 추천하는 단계;를 포함한다.

*대 표 도* - 도1



(72) 발명자

**김진영**

대전광역시 유성구 문화원로 106, 1211호 (봉명동, 유성 캠퍼스타워1)

**오홍선**

대전광역시 유성구 어은로 57, 106동 801호 (어은동, 한빛아파트)

**서동준**

대전광역시 유성구 죽동로 251, 303동 1204호 (죽동, 푸르지오아파트)

**이석형**

대전광역시 유성구 상대남로 26, 915동 2503호 (상대동, 도안신도시9블록 트리풀시티아파트)

**이혜진**

대전광역시 유성구 반석동로 33, 505동 1801호 (반석동, 반석마을5단지아파트)

**정서영**

대전광역시 중구 동서대로 1388, 101동 603호 (목동, 금호한사랑아파트)

# 명 세 서

## 청구범위

### 청구항 1

수집모듈이 DB로부터 텍스트 데이터를 수집하는 단계;

클러스터링모듈이 상기 수집된 텍스트 데이터를 토픽별로 클러스터링하는 단계;

탐지모듈이 상기 토픽별로 클러스터링된 텍스트 데이터에서 이벤트를 포함하는 토픽 문장을 탐지하는 단계;

추출모듈이 상기 탐지한 토픽 문장과 연관된 지식을 추출하는 단계;

워드벡터 생성모듈이 개별 텍스트 컬렉션으로부터 워드벡터들을 생성하는 단계;

네트워크 생성모듈이 상기 추출된 연관된 지식과 상기 생성된 워드벡터들을 결합하여 이기종 정보 네트워크를 생성하는 단계;

확장단어 생성모듈이 입력단어에 대응하여 상기 생성된 네트워크 내의 타겟 텍스트 컬렉션으로부터 상기 워드벡터를 활용하여 확장단어를 생성하는 단계; 및

문서추천모듈이 상기 확장단어를 기반으로 상기 타겟 텍스트 컬렉션으로부터 문서를 추천하는 단계;를 포함하는 문서추천방법.

### 청구항 2

제 1 항에 있어서, 상기 클러스터링하는 단계는

상기 텍스트 데이터에 대한 토픽 모델링하는 단계;

상기 모델링된 토픽의 키워드를 정제하는 단계;

상기 정제된 키워드를 활용하여 상기 텍스트 데이터를 정제하는 단계;및

상기 정제된 텍스트 데이터의 제목을 라벨링하는 단계;를 더 포함하는 문서추천방법.

### 청구항 3

제 1 항에 있어서, 상기 연관된 지식의 추출은 상기 이벤트를 기초로 상기 탐지된 문장 내의 개체명들을 추출하는 것을 특징으로 하는 문서추천방법.

### 청구항 4

제 1 항에 있어서, 상기 이기종 정보 네트워크는 상기 개별 텍스트 컬렉션으로부터 생성된 워드벡터 내에 존재하는 단어들을 연결하여 생성하는 것을 특징으로 하는 문서추천방법.

### 청구항 5

제 1 항에 있어서, 상기 확장단어는 상기 입력단어와 인접한 상기 워드 벡터 내의 연관된 단어인 것을 특징으로 하는 문서추천방법.

### 청구항 6

제 1 항에 있어서, 상기 워드벡터는 상기 개별 텍스트 컬렉션에 개시된 연관된 단어들로 구성되는 것을 특징으로 하는 문서추천방법.

### 청구항 7

DB로부터 텍스트 데이터를 수집하는 수집모듈;

상기 수집된 텍스트 데이터를 토픽별로 클러스터링하는 클러스터링모듈;

상기 토픽별로 클러스터링된 텍스트 데이터에서 이벤트를 포함하는 토픽 문장을 탐지하는 탐지모듈;

상기 탐지한 토픽 문장과 연관된 지식을 추출하는 추출모듈;

개별 텍스트 컬렉션으로부터 워드벡터들을 생성하는 워드벡터 생성모듈;

상기 추출된 연관된 지식과 상기 생성된 워드 백터들을 결합하여 이기종 정보 네트워크를 생성하는 네트워크 생성모듈;

입력단어에 대응하여 상기 생성된 네트워크 내의 타겟 텍스트 컬렉션으로부터 상기 워드벡터를 활용하여 확장단어를 생성하는 확장단어 생성모듈; 및

상기 확장단어를 기반으로 상기 타겟 텍스트 컬렉션으로부터 문서를 추천하는 문서추천모듈;를 포함하는 문서추천장치.

**청구항 8**

제 7 항에 있어서, 상기 클러스터링모듈은,

상기 텍스트 데이터에 대한 토픽 모델링하는 모델링모듈;

상기 모델링된 토픽의 키워드를 정제하는 키워드정제모듈;

상기 정제된 키워드를 활용하여 상기 텍스트 데이터를 정제하는 텍스트 데이터 정제모듈;및

상기 정제된 텍스트 데이터의 제목을 라벨링하는 라벨링모듈;을 더 포함하는 문서추천장치.

**청구항 9**

제 7 항에 있어서, 상기 연관된 지식의 추출은 상기 이벤트를 기초로 상기 탐지된 문장 내의 개체명들을 추출하는 것을 특징으로 하는 문서추천장치.

**청구항 10**

제 7 항에 있어서, 상기 이기종 정보 네트워크는 상기 개별 텍스트 컬렉션으로부터 생성된 워드벡터 내에 존재하는 단어들을 연결하여 생성하는 것을 특징으로 하는 문서추천장치.

**청구항 11**

제 7 항에 있어서, 상기 확장단어는 상기 입력단어와 인접한 상기 워드 벡터 내의 연관된 단어인 것을 특징으로 하는 문서추천장치.

**청구항 12**

제 7 항에 있어서, 상기 워드벡터는 상기 개별 텍스트 컬렉션에 개시된 연관된 단어들로 구성되는 것을 특징으로 하는 문서추천장치.

**청구항 13**

삭제

## *발명의 설명*

### *기 술 분 야*

[0001] 본 발명은 이기종 텍스트 컬렉션을 상호 연결하여 구축된 이기종 정보 네트워크를 기반으로 문서 추천을 하기 위한 장치 및 방법에 관한 것이다.

### *배 경 기 술*

[0002] 종래에는 소셜미디어, 뉴스 등의 비정형 텍스트 데이터로부터 토픽 또는 이슈 또는 이벤트를 추출하려는 많은

연구들이 있었다. 이를 이용한 트렌드 분석, 주가예측 분석, 동향분석 시스템 등을 구현한 사례들이 있다. 따라서 뉴스에 나타난 주요 토픽에 대한 내용으로 논문, 특허, 보고서 등과 같은 이기종 컨텐트 를 추천하는 경우 종래 기술은 단순 키워드 검색 또는 기 정의된 연관 질의어 확장 정도의 수준에 머물고 있다.

## 발명의 내용

### 해결하려는 과제

[0003] 본 발명은 상기한 바와 같은 문제점을 해결하기 위한 것으로 이기종 텍스트 컬렉션을 상호 연결하여 구축된 이기종 정보 네트워크를 기반으로 문서 추천을 하기 위한 장치 및 방법에 관한 것이다.

### 과제의 해결 수단

[0004] 이와 같은 목적을 달성하기 위해, 본 발명의 일 실시예에 따른 문서 추천 방법은 DB로부터 텍스트 데이터를 수집하는 단계; 상기 수집된 텍스트 데이터를 토픽별로 클러스터링하는 단계; 상기 토픽별로 클러스터링된 텍스트 데이터에서 이벤트를 포함하는 토픽 문장을 탐지하는 단계; 상기 탐지한 토픽 문장과 연관된 지식을 추출하는 단계; 개별 텍스트 컬렉션으로부터 워드벡터들을 생성하는 단계; 상기 추출된 연관된 지식과 상기 생성된 워드벡터들을 결합하여 이기종 정보 네트워크를 생성하는 단계; 입력단어에 대응하여 상기 생성된 네트워크 내의 타겟 텍스트 컬렉션으로부터 상기 워드벡터를 활용하여 확장단어를 생성하는 단계; 및 상기 확장단어를 기반으로 상기 타겟 컬렉션으로부터 문서를 추천하는 단계;를 포함할 수 있다.

[0005] 또한 본 발명의 다른 실시예에 따른 문서 추천 장치는 DB로부터 텍스트 데이터를 수집하는 수집모듈; 상기 수집된 텍스트 데이터를 토픽별로 클러스터링하는 클러스터링모듈; 상기 토픽별로 클러스터링된 텍스트 데이터에서 이벤트를 포함하는 토픽 문장을 탐지하는 탐지모듈; 상기 탐지한 토픽 문장과 연관된 지식을 추출하는 추출모듈; 개별 텍스트 컬렉션으로부터 워드벡터들을 생성하는 워드 벡터 생성모듈; 상기 추출된 연관된 지식과 상기 생성된 워드벡터들을 결합하여 이기종 정보 네트워크를 생성하는 네트워크 생성모듈; 입력단어에 대응하여 상기 생성된 네트워크 내의 타겟 텍스트 컬렉션으로부터 상기 워드벡터를 활용하여 확장단어를 생성하는 확장단어 생성모듈; 및 상기 확장단어를 기반으로 상기 타겟 컬렉션으로부터 문서를 추천하는 문서추천모듈;를 포함할 수 있다.

### 발명의 효과

[0006] 이상 설명한 바와 같이, 본 발명에 의하면 문서 추천 장치 및 방법을 제공함으로써 뉴스를 기반으로 주요한 과학기술관련 이슈/토픽을 파악할 수 있다.

[0007] 또한 본 발명에 의하면 문서 추천 장치 및 방법을 제공함으로써 토픽 지식 템플릿을 획득할 수 있다.

[0008] 또한 본 발명에 의하면 특허/논문에 존재하는 연관단어들을 결합하여 이기종 정보 네트워크 (HIN)를 구성할 수 있다.

[0009] 또한 본 발명에 의하면 이기종 정보 네트워크는 서로 다른 특성의 텍스트 컬렉션을 상호 검색 시 동적으로 타겟 컬렉션에 적합한 연관어를 획득하여 사용자에게 보다 적합한 문서를 추천할 수 있다.

[0010] 또한 본 발명에 의하면 뉴스→특허, 논문→특허, 특허→논문 등 서로 다른 컬렉션을 효과적으로 검색하고자 하는 목적의 사용자에게 검색 편의성을 제공할 수 있다.

[0011] 또한 본 발명에 의하면 이기종 정보네트워크 자체는 연관 개체의 타입(type)을 저장하고 있기에, 연관정보 추적 및 분석에 다양하게 활용할 수 있다.

### 도면의 간단한 설명

[0012] 도 1은 문서추천시스템을 설명하는 도면이다.

도 2는 문서추천장치를 설명하는 모듈 구성도이다.

도 3은 비정형 데이터 수집을 설명하기 위한 도면이다.

도 4는 이슈성 토픽 클러스터링을 설명하기 위한 도면이다.

도 5는 토픽 문장 탐지를 설명하기 위한 도면이다.

도 6은 토픽 지식 템플릿을 설명하기 위한 도면이다.

도 7은 뉴스를 중심으로 한 동적 HIN의 구성을 설명하기 위한 도면이다.

도 8은 이기종 정보 네트워크 구성을 설명하기 위한 도면이다.

도 9는 간단한 질의에 대한 처리를 설명하기 위한 도면이다.

도 10은 문서형태의 질의에 대한 처리를 설명하기 위한 도면이다.

도 11은 사용자의 명시적 피드백을 설명하기 위한 도면이다.

도 12는 이기종 컬렉션들 간의 문서추천을 설명하기 위한 도면이다.

도 13은 문서추천방법을 설명하기 위한 흐름도이다.

## *발명을 실시하기 위한 구체적인 내용*

[0013] 본 발명의 일 실시예를 첨부된 도면들을 참조하여 상세히 설명한다. 또한, 본 발명을 설명함에 있어, 관련된 공지 구성 또는 기능에 대한 구체적인 설명이 본 발명의 요지를 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명을 생략한다.

[0014] 도 1는 문서추천시스템을 설명하는 도면이다.

[0015] 도 1를 참조하면, 문서추천시스템은 DB로부터 비정형 텍스트 데이터를 수집하여 수집된 텍스트 데이터를 토픽별로 클러스터링한 후 토픽/이슈를 포함하는 잠재적 토픽 문장을 탐지하고 이와 연관된 토픽 지식을 추출할 수 있다. 또한 추출된 연관된 토픽 지식은 저장할 수 있다. 연관된 토픽 지식을 활용하여 이기종 정보 네트워크를 생성할 수 있다. 또한 사용자로부터 질의사항을 입력받아 이기종 켈렉션들 간의 문서를 추천할 수 있다. 이와 관련된 구체적인 내용은 아래에서 설명한다.

[0016] 도 2는 문서추천장치를 설명하는 모듈 구성도이다.

[0017] 도 2를 참조하면, 문서추천장치는 수집모듈(200), 클러스터링모듈(210), 탐지모듈(220), 추출모듈(230), 워드벡터 생성모듈(240), 네트워크 생성모듈(250), 확장단어 생성모듈(260), 문서추천모듈(270)을 포함할 수 있다. 수집모듈은 DB로부터 텍스트 데이터를 수집할 수 있다. 클러스터링 모듈은 수집된 텍스트 데이터를 토픽별로 클러스터링 할 수 있다. 클러스터링 모듈은 텍스트 데이터에 대한 토픽 모델링하는 모델링모듈, 모델링된 토픽의 키워드를 정제하는 키워드정제모듈, 정제된 키워드를 활용하여 텍스트 데이터를 정제하는 텍스트 데이터 정제모듈, 정제된 텍스트 데이터의 제목을 라벨링하는 라벨링모듈을 포함할 수 있다. 탐지모듈은 토픽별로 클러스터링된 텍스트 데이터에서 이벤트를 포함하는 토픽 문장을 탐지할 수 있다. 추출모듈은 탐지한 토픽 문장과 연관된 지식을 추출할 수 있다. 또한 추출모듈은 이벤트를 기초로 탐지된 문장 내의 개체명들을 추출할 수 있다.

[0018] 워드벡터 생성모듈은 개별 텍스트 컬렉션으로부터 워드벡터들을 생성할 수 있다. 네트워크 생성모듈은 추출된 연관된 지식과 생성된 워드벡터들을 결합하여 이기종 정보 네트워크를 생성할 수 있다. 또한 이기종 정보 네트워크는 개별 텍스트 컬렉션으로부터 생성된 워드벡터 내에 존재하는 단어들을 연결하여 생성될 수 있다.

[0019] 확장단어 생성모듈은 입력단어에 대응하여 생성된 네트워크 내의 타겟 텍스트 컬렉션으로부터 워드벡터를 활용하여 확장단어를 생성할 수 있다. 문서추천모듈은 확장단어를 기반으로 타겟 컬렉션으로부터 문서를 추천할 수 있다. 전술한 확장단어는 입력단어와 인접한 워드 벡터 내의 연관된 단어인 것을 특징으로 한다. 전술한 워드벡터는 개별 컬렉션에 개시된 연관된 단어들로 구성되는 것을 특징으로 할 수 있다.

[0020] 도 3은 비정형 데이터 수집을 설명하기 위한 도면이다.

[0021] 도 3을 참조하면, 비정형 데이터의 구체적인 예시로 뉴스/소셜미디어 데이터, 논문, 특허, 보고서 등이 있다. 이와 관련하여 구체적으로 수집모듈은 시의성 있는 사회현안·이슈·토픽을 추출하기 위한 바탕이 되는 뉴스/소셜미디어 데이터를 수집함과 동시에 전문적 지식을 담고 있는 논문, 특허, 보고서 등도 입수 주기에 따라 수집하여 물리적 데이터베이스(database)에 저장할 수 있다. 수집모듈은 뉴스데이터에서 협약 및 크롤링 & 정제 의해 기사명, 기사내용, 기사날짜, 신문사 등의 정보를 획득할 수 있으며, 소셜미디어 데이터의 경우 텍스트 생성일, 내용, URL등을 저장할 수 있다. 수집모듈은 논문, 특허, 연구보고서에서 협약에 의해 입수되는 메타 정보 및 본문 텍스트를 저장할 수 있다. 저장하는 목록, 저장하는 형식, 저장하는 방법에 관해서는 설계자의 의도에

따라 변경이 가능하므로 전술한 것에 한정되지 않는다.

[0022] 도 4는 이슈성 토픽 클러스터링을 설명하기 위한 도면이다.

[0023] 도 4를 참조하면, 클러스터링 모듈은 뉴스 텍스트 문서들을 토픽별로 클러스터링하고 제목을 라벨링(labeling)하는 과정을 포함할 수 있다. 클러스터링모듈은 기 수집된 뉴스 문서들을 주간/월간/연도별로 분석하여 시의적으로 유의한 토픽들로 클러스터링 할 수 있다. 또한 클러스터링모듈은 기본적으로는 일 또는 월 단위로 주요 토픽 클러스터를 도출할 수 있다. 클러스터링모듈은 텍스트 데이터에 대한 토픽 모델링하는 모델링모듈, 모델링된 토픽의 키워드를 정제하는 키워드정제모듈, 정제된 키워드를 활용하여 텍스트 데이터를 정제하는 텍스트 데이터 정제모듈, 및 정제된 텍스트 데이터의 제목을 라벨링하는 라벨링모듈을 포함할 수 있다.

[0024] 모델링모듈은 일 또는 월별 뉴스기사들을 입력으로 하여 LDA(Latent dirichlet allocation)를 활용한 토픽 모델링을 수행할 수 있다. 키워드 정제모듈은 모델링된 토픽의 키워드를 정제할 수 있다. 텍스트 데이터 정제모듈은 정제된 키워드를 활용하여 텍스트 데이터를 정제할 수 있다. 또한 키워드 정제모듈 또는 텍스트 데이터 정제모듈은 각 토픽 클러스터에 존재하는 단어들이 부적절한 것들이 존재할 수 있기 때문에 기 정의한 불용어(stop-word) 사전 및 클러스터 내 존재확률을 고려하여 정제할 수 있다. 키워드 정제모듈 또는 텍스트 데이터 정제모듈은 정제과정 시 Word2Vec에서 구현된 유사도 계산기법 및 PMI-IR 등을 채택할 수 있으며, 유형에 따라 약간의 수정된 버전을 활용할 수 있다. 또한 키워드 정제모듈 또는 텍스트 데이터 정제모듈은 정제된 키워드를 활용하여 텍스트 데이터를 정제할 경우, 각 토픽별 정제된 키워드들을 질의어로 하여 해당기간 내 뉴스를 검색하면 대부분 일관된 주제의 뉴스가 검색되므로, 부분적으로 잘못 검색된 결과는 검색결과 클러스터링을 통해 핵심 토픽에서 벗어나는 뉴스기사들을 제외할 수 있다.

[0025] 라벨링모듈은 일관된 주제의 뉴스들에 대한 대표 제목 선정 (일종의 라벨링(labeling))을 수행하기 위해 뉴스의 제목 및 내용에서 이벤트 표현들을 추출하고 고빈도의 이벤트와 관련 어구가 등장하는 뉴스의 제목을 최종 선택할 수 있다. 라벨링 모듈은 이벤트 표현 사전을 기반으로 해당 클러스터의 뉴스들에서 주로 발생하는 핵심 이벤트 표현들을 추출하고 뉴스 제목에서의 출현 횟수를 고려하여 하여 최종 클러스터의 제목을 선정할 수 있다. 만약 뉴스의 내용은 같은데 제목이 다르게 표현된 경우, 라벨링모듈은 주요 이벤트 표현과 자주 나타나는 개체명들 (예, 인물명, 기관명, 제품명, 등)을 같이 표기하도록 설정할 수 있다. 또한 라벨링모듈은 실시간 토픽별 제목의 선정을 위해서는 트위터와 같은 소셜미디어의 시의성 있는 정보를 이용할 수 있다.

[0026] 전술한 이벤트는 행사, 사건, 대회, 정치적 안건 등 다양한 분야에 걸쳐서 존재할 수 있다. 다양한 특성의 이벤트는 사전을 조회해서 이벤트 여부를 판별하며, 각 주제 분야별 이벤트 리스트는 새로운 뉴스들이 입수됨에 따라, 계속적으로 확장할 수 있다. 최근에는, 비교사학습(unsupervised learning) 기법을 통해 이벤트명 인식기를 구현할 수 있다. 가령 예를 들어, Computer Security분야에서는 (ENTITY, DATE)을 이벤트로 정의할 수 있으며, 가령 DDoS공격의 이벤트는 (github, 2013.07.29.), (paypal, 2010.12.10.)로 표현할 수 있다. 또한 TwiCal에서의 이벤트 (ENTITY, Event Phrase, Date, Event Type) 정도로 정의할 수 있으며 예를 들어 (Steve Jobs, dided, 2011. 10.06, Death), (iPhone, Announcement, 2011.10.04.), ProductionLaunch)와 같이 표현할 수 있다. 또한 이벤트는(evnet, date)의 seed instances를 통한 비교사 학습기법을 이용하여 새로운 이벤트들을 계속적으로 확장할 수 있다. 참고로 도메인을 특정하지 않은 상태에서 뉴스 제목들로부터 획득한 이벤트 표현들은 "개발박차", "공개임박", "~열렸다", "~나선다", "~어렵다" 등의 다양한 표현들로 이뤄질 수 있다. 다양한 이벤트 표현들은 문맥을 고려한 비교사 학습에 의해 꾸준히 학습되며, 새로이 발굴된 이벤트 표현은 사전에 추가되어 계속적으로 이벤트명 인식기에 적용할 수 있다.

[0027] 도 5는 토픽 문장 탐지를 설명하기 위한 도면이다.

[0028] 도 5를 참조하면, 탐지모듈은 각 토픽별 뉴스기사에서 핵심 이벤트를 담고 있는 주요 토픽 문장을 탐지한다. 기본적으로 탐지모듈은 이벤트(event)를 중심으로 같은 문장 내에서 이벤트와 연관된 개체들 (예, 인물, 기관, 장소, 제품명, 날짜표현, 과학기술 용어 등)이 출현하는 문장을 선정할 수 있다. 이를 위해서는 탐지모듈은 기 추출된 이슈/토픽별 연관 단어 사전(dictionary)을 활용하여 연관단어들을 포함하고 있는 문장들을 선별할 수 있다. 토픽별 연관단어 사전은 주/월 단위로 자동적으로 계속 갱신되어 최신 토픽에 대한 처리가 가능하며, 문서/문장 선별을 위해서는 관련 단어의 포함 여부만을 키워드 스팟팅 (keyword spotting) 방식으로 빠르게 확인할 수 있다.

[0029] 도 6은 토픽 지식 템플릿을 설명하기 위한 도면이다.

[0030] 도 6을 참조하면, 추출모듈은 각 문서에서 잠재적 토픽 문장으로 선정된 문장들을 대상으로, 이슈성 토픽을 구

성하는 이벤트 주변의 어구에서 인물명, 기관명, 장소명, 시간표현, 제품명, 과학기술용어 등을 추출한다. 기본적으로, 추출모듈은 핵심 이벤트 어구를 탐지 후, 사전(dictionary)를 기반으로 각 문장 내에 존재하는 주요 표현들을 추출하는 방식을 사용할 수 있다. 또한 추출모듈은 주제별로 축소 및 확장이 가능한 토픽 지식 템플릿에 기반하여, 연관된 토픽 지식을 추출할 수 있다. 하지만, 새로운 문서 내에 등장하는 연관된 토픽 지식의 추출을 위해서는 이벤트가 포함된 문장 내에 존재하는 관련 개체명들이 잘 추출되어야 한다. 이를 위해 추출모듈은 기계학습에 기반한 개체명인식 (named entity recognition) 기술들이 사용되며, 중요 개체명들은 인물(person), 장소(location), 기관(organization), 날짜(date), 기술 용어(technology term), 그리고 관련 이벤트(event) 등이 있다. 같은 토픽에 속하는 여러 뉴스기사들 안에는 빈번히 동시 출현하는 [이벤트, 인물], [이벤트, 장소, 날짜], [이벤트, 기관], [인물, 기술용어], [이벤트, 제품명] 등이 있을 수 있다. 이때 임계 빈도(실험적으로 셋팅)이상 출현하는 이벤트 관련어구를 통합 연결하여, 도 6과 같이 토픽별 토픽 지식템플릿을 구성하게 된다. 정확도 높은 토픽 지식템플릿을 구성하려면 고성능 개체명 인식기의 구현이 필수이다. 일반적으로 개체명 인식기는 기 존재하는 라벨링된 학습데이터(labeled training data)의 부재로 신규 개체(음악가, 영화배우, 저자, 연구자)들에 대한 분류를 제대로 수행하지 못하는 경우가 많다. 이 문제와 더불어 노이즈가 존재하는 학습데이터에도 잘 작동하는 이상적인 세부 개체 타입 분류(fine grained named entity type classification) 시스템을 구현하기 위해, 자질과 라벨의 연결 표현(joint representation)을 학습시키는 것이 임베딩(embedding)기법이다.

[0031] 전술한 토픽 지식 템플릿은 주제도메인 (정치, 문화, 보안, 경제 등)의 특성에 따라 이벤트와 연관된 어구 (혹은 시드 인스턴스(seed instances))들은 다르게 정의할 수 있다. 예를 들어 [이벤트, 인물 OR 기관, 타겟, 주요 용어]를 기본 템플릿으로 정의하고 최근 뉴스들을 처리하면 [방문 오바마, 쿠바], [장학금지원, 건국대], [조작 여부조사, 더민주, 폴크스바겐, 신차], [개발박차, 인공지능, 군사로봇], [인기몰이, 한국폰, 이스라엘, 갤럭시 S7] 정도로 이벤트 중심의 어구들을 추출할 수 있다. 토픽 지식 템플릿으로 뉴스 텍스트에서는 기본 템플릿으로 연관어구들을 추출을 기본으로 하며, 과학기술 및 건강/생활 도메인 내에서 [이슈/이벤트, 인물, 기관, 과학기술용어] 템플릿에 의해 관련 인스턴스들(instances)을 추출하고, 이를 기반으로 향후 논문/특허/과학기술 보고서와의 연계 및 분석에 대응할 수 있다.

[0032] 도 7은 이기종 정보 네트워크 구성을 설명하기 위한 도면이다.

[0033] 도 7을 참조하면, 이기종 정보 네트워크 생성은 서로 다른 목적으로 작성된 문서 컬렉션들 각각의 특징을 분석하고, 이들을 상호 연결하기 위한 단어 네트워크를 구성하기 위한 것이다. 같은 텍스트 컬렉션을 이용하여 서로 다른 워드 벡터들 (예, Word2Vec과 DVRS)을 만든 후 이를 결합할 수 있다. 이는 워드 벡터(Word Vector)를 구성하는 알고리즘에 따라 각각의 강점이 있는데, 다른 알고리즘에 의해 생성된 서로 다른 워드벡터를 합성함으로써 단어유추 작업(Word Analogy task)에서 보다 나은 성능을 보일 수 있다. 네트워크 생성모듈은 서로 다른 컬렉션에서 파생된 워드벡터를 결합함으로써 컬렉션별로 존재하는 단어들의 문맥(context)을 파악하고 이들을 상호 연결할 수 있다. 네트워크 생성모듈은 뉴스/소셜 미디어 컬렉션뿐만 아니라 과학기술 전문 컬렉션들에서 획득한 의미적 워드벡터들(semantic word vectors)로부터 인접하게 연관된 개체타입의 종류와 관계의 종류에 국한하기 않고, 상호 연결할 수 있다. 네트워크 생성모듈은 관계명을 명시적으로 표현하지 않고 상호 연결 시 연결 선(edge)의 강도를 증가시키면서 그래프 형태의 네트워크를 구성할 수 있다. 네트워크 생성모듈은 인식된 개체명의 개체타입(entity type)은 명기할 수 있다. 네트워크 생성모듈은 뉴스에서 추출된 이벤트에 국한되지 않고, 도 7과 같이 각 컬렉션별로 생성된 워드벡터 내에서 존재하는 단어 리스트를 순회(iteration)하면서 각 단어별로 가장 유사한 단어 10~20개를 워드벡터 내 확률 값에 의해 랭킹하고 이들을 상호 연결할 수 있다. 이때 네트워크 생성모듈은 연결되는 각 단어들은 개체명 사전을 참고하여 개체 타입(entity type)도 명기할 수 있다. 기 구축된 HIN은 사용자 질의 (질의어, 타겟 컬렉션)에 따라 문서 추천에 사용될 수 있는 단어 쌍을 선별적으로 추가 획득하거나 기타 개체별 추론기능에 활용될 수 있다.

[0034] 도 8은 뉴스를 중심으로 한 동적 HIN의 구성을 설명하기 위한 도면이다.

[0035] 도 8을 참조하면, 네트워크 생성모듈은 이기종 정보 네트워크의 구성하여 생성과 활용을 위해 다음의 과정을 수행할 수 있다. 네트워크 생성모듈은 문서 컬렉션 내에서 단어별로 출현하는 맥락을 고려한 인접단어를 추출하는 방법을 이용하여 각 컬렉션별로 인접단어들 간의 쌍으로 구성되는 워드벡터를 생성할 수 있다. 예를 들어 네트워크 생성모듈은 Word2Vec를 이용하면 각 컬렉션별로 지정한 규모 (예, feature vector size=200) 규모의 워드벡터를 생성할 수 있다. 각 컬렉션별로 구성되는 벡터공간은 각 컬렉션의 특징을 나타내는데, 네트워크 생성모듈은 이종 단어 연결 네트워크의 생성을 위해 서로 다른 컬렉션별로 구성된 워드벡터를 결합할 수 있다. 첫 번째 경우, 네트워크 생성모듈은 가장 기본적으로 뉴스의 토픽 지식 템플릿을 기반으로 이기종 워드 벡터들을 결

합할 수 있다. 뉴스 -> 뉴스, 뉴스 -> 논문, 뉴스 -> 특허로의 검색을 지원하는 경우도 8과 같이 각 컬렉션별 워드벡터를 선택적으로 결합하여 존재하는 단어들로 네트워크를 연결할 수 있다. 예를 들어 네트워크 생성모듈은 뉴스의 키워드 "반도체"는 뉴스, 논문, 특허 별로 그 컬렉션 "반도체"라는 키워드와 인접하여 나타나는 워드벡터에 존재하는 단어들을 연관단어로 연결하여 네트워크를 갱신할 수 있다. 결과적으로 네트워크 생성모듈이 매스미디어의 중요 이벤트의 핵심 키워드 중심으로 네트워크를 생성하면서 연결된 단어들의 출현빈도에 따라 연결하는 가중치를 다시 계산할 수 있다. 이때 네트워크 생성모듈은 연결되는 각 단어들은 개체명 사전을 참고하여 개체 타입(entity type)도 명기할 수 있다.

[0036] 도 9는 간단한 질의에 대한 처리를 설명하기 위한 도면이다.

[0037] 도 9를 참조하면, 확장단어 생성모듈은 사용자의 질의어를 입력 받아(짧게는 평균 2~5 단어로 구성된 질의어 혹은 문서 자체가 질의가 될 수 있다.) 사용자의 의도 및 질의유형 (사용자 질의/목적, 타겟 컬렉션)을 고려하여 이기종 정보 네트워크를 통해 유의한 질의 확장어를 획득할 수 있다. 확장단어 생성모듈은 사용자 질의에 대해 사용자가 선택한 타겟 컬렉션을 대상으로 적응형 질의 확장을 수행할 수 있다. 확장단어 생성모듈은 전술한 이기종 정보 네트워크인 HIN을 사용하여 타겟 컬렉션을 대상으로 구성된 워드벡터를 이용하여 적응형 질의 확장을 수행할 수 있다.

[0038] 도 10은 문서형태의 질의에 대한 처리를 설명하기 위한 도면이다.

[0039] 도 10을 참조하면, 확장단어 생성모듈은 입력이 전술한 질의어가 아니라 유사한 논문 또는 특허인 경우 논문을 입력으로 하여 관련된 특허나 신문 기사를 찾거나, 혹은 특허를 입력으로 관련된 논문이나 신문기사를 찾고자 하는 사용자의 질의를 처리할 수 있다. 확장단어 생성모듈은 논문과 특허 각각은 고유의 특징을 지니기에 개별적인 처리 기법을 통해 목적 수준의 문서 요약정보를 획득할 수 있다. 또한, 확장단어 생성모듈은 적응화 과정을 통한 질의를 확장할 수 있다. 또한 검색/추천 단계에서 이기종 정보 네트워크인 HIN를 이용함에 있어, 크게 질의어에 속하는 단어가 HIN에 존재하는 경우와 존재하지 않는 경우로 나뉘는데, HIN에 존재하는 경우 확장단어 생성모듈은 질의어와 인접한 워드벡터 내의 단어 쌍들을 확장 질의어로 추가할 수 있다. 하지만, 확장단어 생성모듈은 사용자 질의어에 속하는 단어가 없는 경우 최초 질의어를 적응화 과정을 거쳐서 대용 가능한 인접단어를 획득하고 이를 질의어로 사용할 수 있다.

[0040] 도 11은 사용자의 명시적 피드백을 설명하기 위한 도면이다.

[0041] 도 11을 참조하면, 입력이 특허문서인 경우, 도 11과 같이 특허의 목적 수준의 요약정보 및 (추정되는) 유사 목적들을 제시·시각화하면 사용자는 이들 중 자신의 목적에 맞는 목적들(복수 개 허용) 및 타겟 컬렉션을 선택하면, 적응형 질의 확장과정으로 이어진다.

[0042] 참고로 논문은 정확한 메타정보들이 존재하는 외부 식별 시스템이 존재하기에, DOI (document object identifier) 같은 식별자로 접근하여 CrossRef (http://www. crossref.org/) 또는 인용색인서버로부터 논문의 정확한 저자, 기관, 초록, 키워드 등의 메타정보를 획득할 수 있다. 따라서 목적 수준의 문서요약은 주요 키워드와 논문의 초록에 명시된 연구의 목적/해결책들이 요약의 주 내용이 된다. 바이오메디컬(biomedical) 분야의 논문들은 초록 자체가 연구목적/실험방법/실험결과 등으로 구성되어 있어, 현존하는 자동요약기법들을 활용하면 목적을 나타내는 주요 표현을 추출할 수 있다. 기본적으로는 핵심 토픽 및 토픽 관련 표현을 잘 찾는 것이고, 추가적으로 연구 논문의 목적을 잘 추출하는 것이다. 기타 분야의 과학기술 문헌들은 분야별 목적 수준의 태깅 데이터를 확보하여 목적 수준의 표현을 추출하는 모듈을 구현함으로써 논문별로 목적요약정보를 획득할 수 있다.

[0043] 참고로 특허의 경우 출원인, 발명자 및 초록 필드에서 논문 수준의 메타정보를 획득할 수 있다. 그리고 선행기술조사문헌 정보를 통해 관련기술의 선행 기술에 대한 정보도 획득이 가능하다. 하지만 특허의 경우 회피성 표현이 많기에 자동으로 특허의 목적을 판단하기가 쉽지 않다. 목적기반의 자동태깅과 같은 태깅 기법이 특허분야에 적용되어 확장 개발된다면 자동으로 특허의 목적 수준 요약을 제공할 수 있다. 기존 연구에서는 해결한 문제의 개념을 추출하는 연구가 있는데, 주로 서론(Introduction)과 결론(conclusion)에서 주요한 정보를 찾고, 발명에 대한 중요한 표현어구를 선별하였으며, n-gram 표현까지 염두에 두었다. 본 제안 특허에서는 특허초록/청구항의 핵심문장들을 중심으로 목적 수준의 요약정보를 포함하는 주요 동사패턴을 보이는 주요 문장들을 추출하는 방법을 활용한다. 예를 들어, 국내 특허의 경우 "효과가 있다", "특징으로 한다", "위함이다", "제공한다", "구성된다", "제조한다", "수행된다", "제조방법이다", "발명인 것이다", "위한 것이다", "이용될 수 있다", "제공하는 것이다", "얻을 수 있다", "얻는다", "나타낸다" 등의 동사표현은 "본 발명은", "본 고안은", "이 고

안은" 등의 표현과 함께 빈번히 나타난다. 특허문서의 요약은 이들 패턴들 안에 나타나는 주요 단어들에서 불용어(stop words)를 제외하고 구성된다. 특히, 특허마다 갖게 되는 고유의 주제분류인 IPC(International Patent Classification) / CPC (Cooperative Patent Classification) 코드에 의거하여 같은 범주에 해당하는 특허들의 주요목적들을 획득할 수 있다.

[0044] 도 12는 이기종 컬렉션들 간의 문서추천을 설명하기 위한 도면이다.

[0045] 도 12를 참조하면, 문서추천모듈은 확장단어 생성모듈로부터 획득한 사용자 질의에 대응되는 적응형 질의 확장의 결과를 기반으로 타겟 컬렉션으로부터 문서를 추천할 수 있다. 확장된 질의를 통해 타겟 컬렉션에서 문서를 매칭하는 과정은 전통적인 Vector Space Model (VSM)모델 하에서 TF*IDF score를 기반으로 하고, 사용자 질의, 확장된 질의를 다차원적으로 나누어 관련 그 기준의 적합도(relevance)에 따라 문서들이 랭킹될 수 있다.(** Relevance Score = alpha * query_similarity + (1-alpha) * goal_similarity ) 문서추천모듈은 질의 유사도와 목적 유사도 각각을 가중치를 고려하여, 사용자의 의도에 맞는 문서를 매칭하기 위해 위 점수를 고려하며, 그 값을 내림차순 정렬하여 추출된 문서들을 반환할 수 있다.

[0046] 도 13은 문서추천방법을 설명하기 위한 흐름도이다.

[0047] 도 13을 참조하면, 문서추천방법은 DB로부터 텍스트 데이터를 수집하는 단계(S1300), 수집된 텍스트 데이터를 토픽별로 클러스터링하는 단계(S1301), 토픽별로 클러스터링 된 텍스트 데이터에서 이벤트를 포함하는 토픽 문장을 탐지하는 단계(S1302), 탐지한 토픽 문장과 연관된 지식을 추출하는 단계(S1303), 개별 텍스트 컬렉션으로부터 워드벡터들을 생성하는 단계(S1304), 추출된 연관된 지식과 생성된 워드벡터들을 결합하여 이기종 정보 네트워크를 생성하는 단계(S1305), 입력단어에 대응하여 생성된 네트워크 내의 타겟 텍스트 컬렉션으로부터 워드벡터를 활용하여 확장단어를 생성하는 단계(S1306), 및 확장단어를 기반으로 타겟 컬렉션으로부터 문서를 추천하는 단계(S1307)를 포함할 수 있다. 수집모듈(200)은 DB로부터 텍스트 데이터를 수집하는 단계(S1300)를 수행할 수 있다. 이에 대한 설명은 도1, 도2, 및 도3에서 전술한 바 있다. 클러스터링모듈(210)은 수집된 텍스트 데이터를 토픽별로 클러스터링하는 단계(S1301)를 수행할 수 있다. 이에 대한 설명은 도1, 도2, 및 도4에서 전술한 바 있다. 탐지모듈(220)은 토픽별로 클러스터링 된 텍스트 데이터에서 이벤트를 포함하는 토픽 문장을 탐지하는 단계(S1302)를 수행할 수 있다. 이에 대한 설명은 도1, 도2, 및 도5에서 전술한 바 있다. 추출모듈(230)은 탐지한 토픽 문장과 연관된 지식을 추출하는 단계(S1303)를 수행할 수 있다. 이에 대한 설명은 도1, 도2, 및 도6에서 전술한 바 있다. 워드벡터 생성모듈(240)은 개별 텍스트 컬렉션으로부터 워드벡터들을 생성하는 단계(S1304)를 수행할 수 있다. 이에 대한 설명은 도1, 도2, 도7, 및 도8에서 전술한 바 있다. 네트워크 생성모듈(250)은 추출된 연관된 지식과 생성된 워드벡터들을 결합하여 이기종 정보 네트워크를 생성하는 단계(S1305)를 수행할 수 있다. 이에 대한 설명은 도1, 도2, 도7, 및 도8에서 전술한 바 있다. 확장단어 생성모듈(260)은 입력단어에 대응하여 생성된 네트워크 내의 타겟 텍스트 컬렉션으로부터 워드벡터를 활용하여 확장단어를 생성하는 단계(S1306)를 수행할 수 있다. 이에 대한 설명은 도1, 도2, 도9, 및 도10에서 전술한 바 있다. 문서추천모듈(270)은 확장단어를 기반으로 타겟 컬렉션으로부터 문서를 추천하는 단계(S1307)를 수행할 수 있다. 이에 대한 설명은 도1, 도2, 및 도13에서 전술한 바 있다.

## *부호의 설명*

[0048] 200 : 수집모듈

210 : 클러스터링모듈

220 : 탐지모듈

230 : 추출모듈

240 : 워드벡터 생성모듈

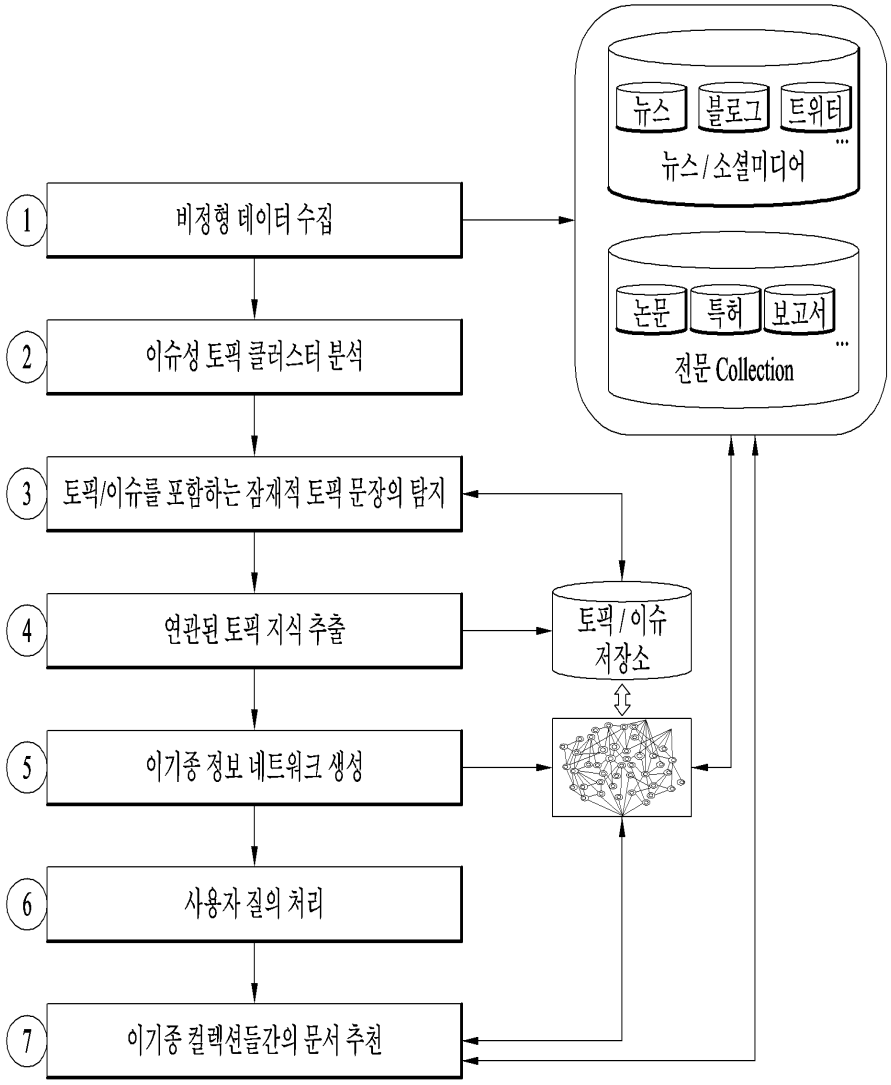
250 : 네트워크 생성모듈

260 : 확장단어 생성모듈

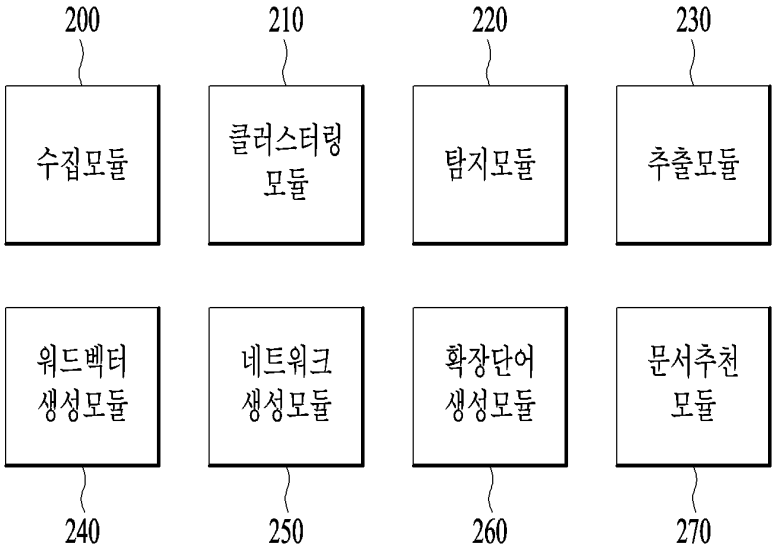
270 : 문서추천모듈

## 도면

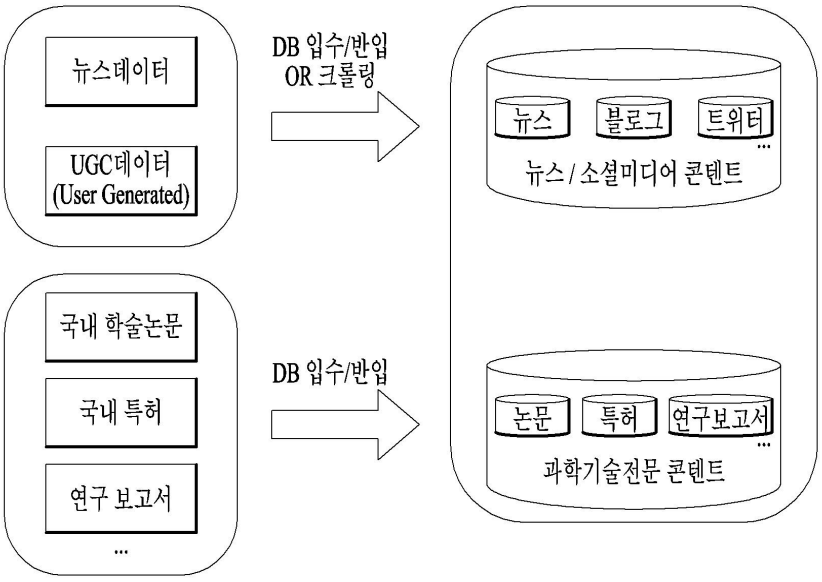
### 도면1


뉴스
블로그
트위터
...
뉴스/소셜미디어
논문
특허
보고서
...
전문 Collection
1
비정형 데이터 수집
2
이슈성 토픽 클러스터 분석
3
토픽/이슈를 포함하는 잠재적 토픽 문장의 탐지
4
연관된 토픽 지식 추출
토픽/이슈 저장소
5
이기종 정보 네트워크 생성
6
사용자 질의 처리
7
이기종 컬렉션들간의 문서 추천

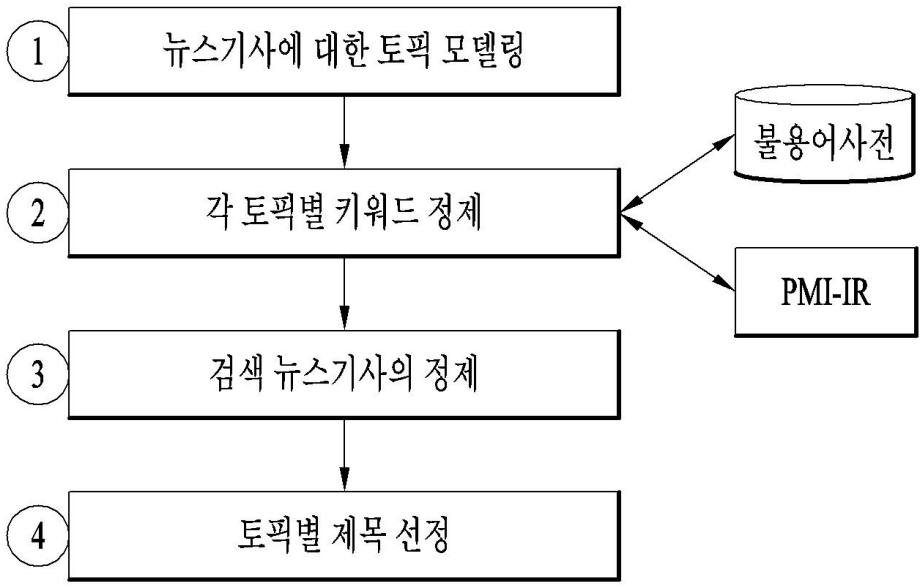
도면2


200
210
220
230
수집모듈
클러스터링 모듈
탐지모듈
추출모듈
워드벡터 생성모듈
네트워크 생성모듈
확장단어 생성모듈
문서추천 모듈
240
250
260
270

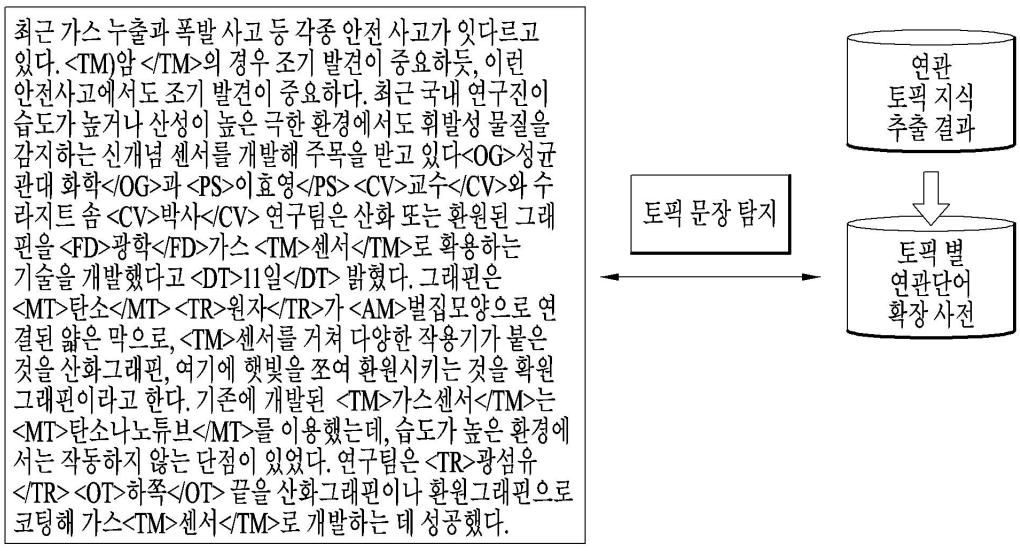
도면3


뉴스데이터
UGC데이터 (User Generated)
DB 입수/반입 OR 크롤링
뉴스
블로그
트위터
뉴스/소셜미디어 콘텐트
국내 학술논문
국내 특허
연구 보고서
DB 입수/반입
논문
특허
연구보고서
과학기술전문 콘텐트

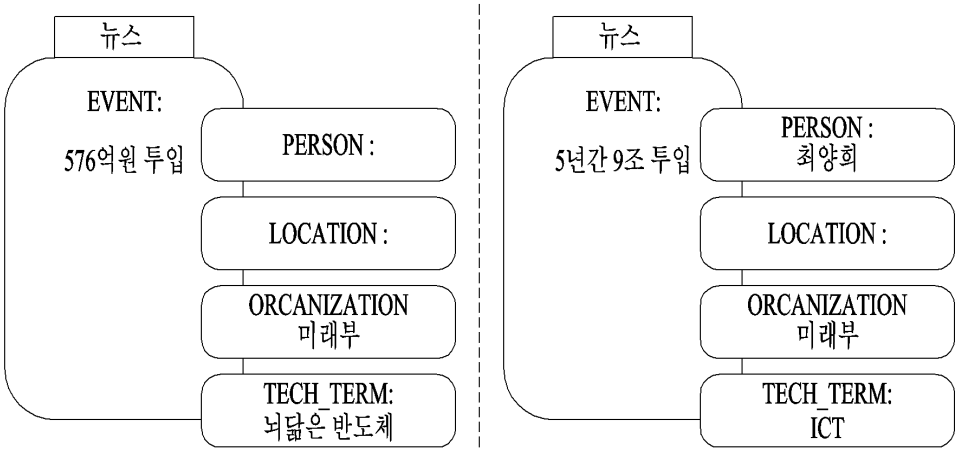
도면4


1
뉴스기사에 대한 토픽 모델링
2
각 토픽별 키워드 정제
3
검색 뉴스기사의 정제
4
토픽별 제목 선정
불용어사전
PMI-IR

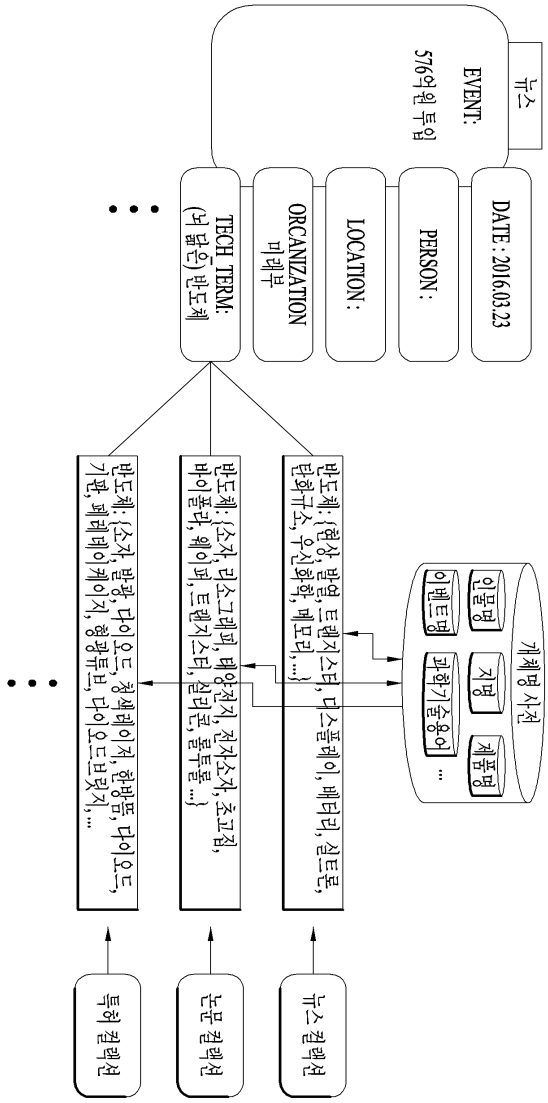
도면5


최근 가스 누출과 폭발 사고 등 각종 안전 사고가 잇다르고 있다. <TM)암 </TM>의 경우 조기 발견이 중요하듯, 이런 안전사고에서도 조기 발견이 중요하다. 최근 국내 연구진이 습도가 높거나 산성이 높은 극한 환경에서도 휘발성 물질을 감지하는 신개념 센서를 개발해 주목을 받고 있다<OG>성균관대 화학</OG>과 <PS>이효영</PS> <CV>교수</CV>와 수라지트 솜 <CV>박사</CV> 연구팀은 산화 또는 환원된 그래핀을 <FD>광학</FD>가스 <TM>센서</TM>로 확용하는 기술을 개발했다고 <DT>11일</DT> 밝혔다. 그래핀은 <MT>탄소</MT> <TR>원자</TR>가 <AM>벌집모양으로 연결된 얇은 막으로, <TM>센서를 거쳐 다양한 작용기가 붙은 것을 산화그래핀, 여기에 햇빛을 쪼여 환원시키는 것을 확원 그래핀이라고 한다. 기존에 개발된 <TM>가스센서</TM>는 <MT>탄소나노튜브</MT>를 이용했는데, 습도가 높은 환경에서는 작동하지 않는 단점이 있었다. 연구팀은 <TR>광섬유 </TR> <OT>하쪽</OT> 끝을 산화그래핀이나 환원그래핀으로 코팅해 가스<TM>센서</TM>로 개발하는 데 성공했다.
토픽 문장 탐지
연관 토픽 지식 추출 결과
토픽 별 연관단어 확장 사전

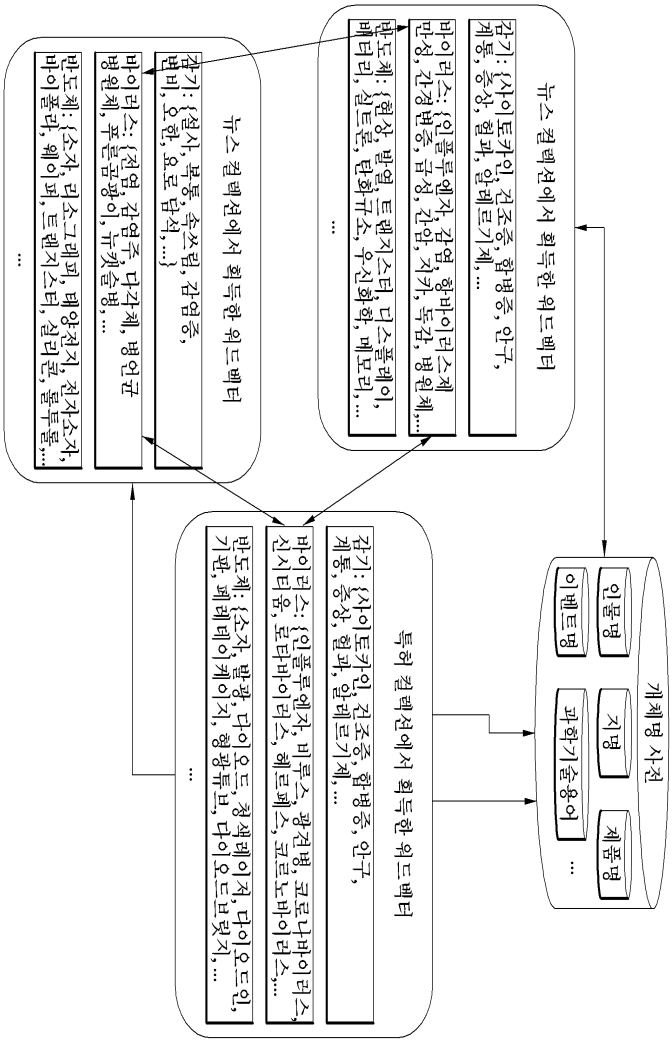
도면6


뉴스
EVENT:
576억원 투입
PERSON :
LOCATION :
ORCANIZATION
미래부
TECH_TERM:
뇌닮은 반도체
뉴스
EVENT:
5년간 9조 투입
PERSON :
최양희
LOCATION :
ORCANIZATION
미래부
TECH_TERM:
ICT

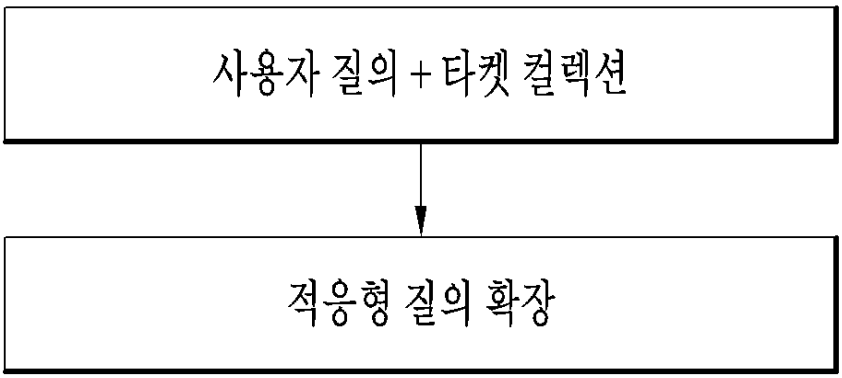
도면7


뉴스
EVENT:
576억원 투입
DATE : 2016.03.23
PERSON :
LOCATION :
ORCANIZATION
미래부
TECH TERM:
(뇌닮은) 반도체
개체명 사전
인물명
지명
제품명
이벤트명
과학기술용어
...
반도체: {현상, 발열, 트랜지스터, 디스플레이, 메터리, 실트론, 탄화규소, 우신화학, 메모리, ...}
반도체: {소자, 리소그래피, 태양전지, 전자소자, 초고집, 바이폴라, 웨이퍼, 트랜지스터, 실리콘, 롤투롤 ...}
반도체: {소자, 발광 다이오드 청색레이저, 한방품 다이오드, 기판, 패키데이케이지, 형광튜브 다이오드낫지, ...
뉴스 컬렉션
논문 컬렉션
특허 컬렉션

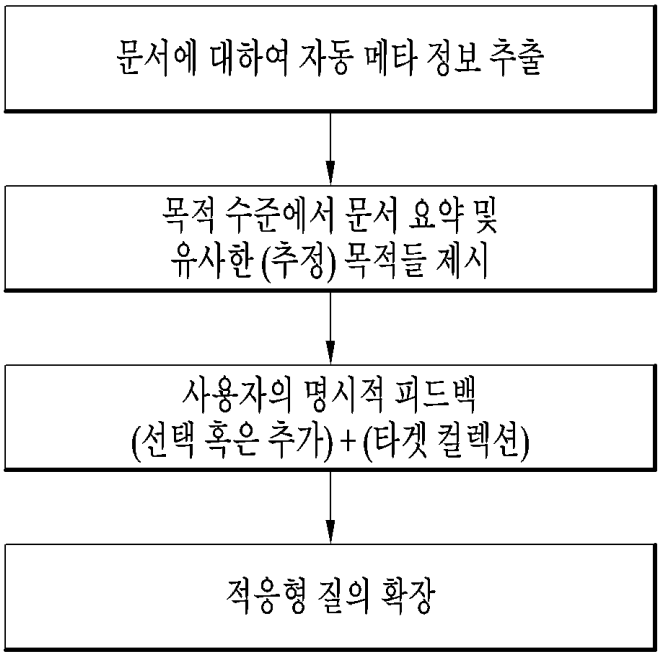
도면8


개체명 사전
인물명
이벤트명
지명
과학기술용어
제품명
...
뉴스 컬렉션에서 획득한 워드벡터
감기: {사이토카인, 건조증, 합병증, 안구, 계통, 증상, 혈과, 알레르기제, ...
바이러스: {인플루엔자, 감염, 항바이러스제 만성, 간경변증, 급성, 간암, 지카, 독감, 병원체,...
반도체: {웨상, 발열, 트랜지스터, 디스플레이, 배터리, 실트로, 탄화규소, 우신화학, 메모리, ...
...
특허 컬렉션에서 획득한 워드벡터
감기: {사이토카인, 건조증, 합병증, 안구, 계통, 증상, 혈과, 알레르기제, ...
바이러스: {인플루엔자, 비루스, 광견병, 코로나바이러스, 신시티움, 로타바이러스, 헤르페스, 코르노바이러스,...
반도체: {소자, 발광, 다이오드, 청색레이저, 다이오드인, 기판, 페레데이케이지, 형광튜브, 다이오드브릿지, ...
...
뉴스 컬렉션에서 획득한 워드벡터
감기: {설사, 복통, 손쓰림, 감염증, 변비, 오한, 요로담석, ...}
바이러스: {전염, 감염주, 다각체, 병언균 병원체, 푸른곰팡이, 뉴캣슬병, ...
반도체: {소자, 리소그래피, 태양전지, 전자소자, 바이폴라, 웨이퍼, 트랜지스터, 실리콘, 롤투롤, ...
...

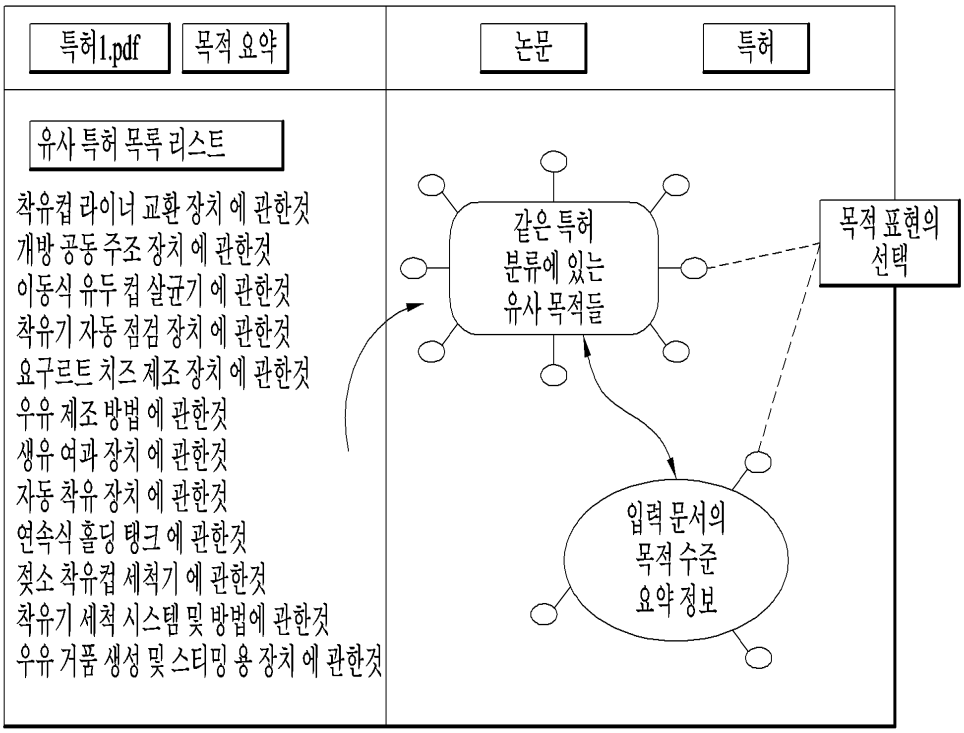
도면9


사용자 질의 + 타켓 컬렉션
적응형 질의 확장

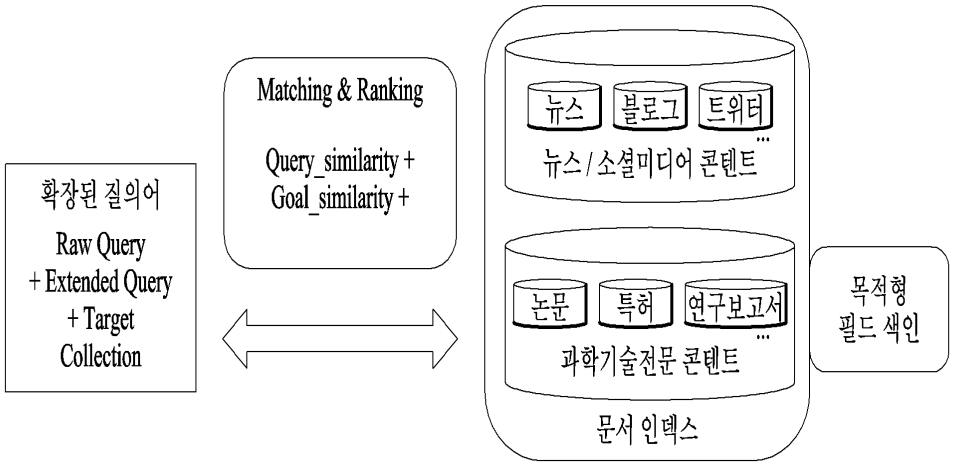
도면10


문서에 대하여 자동 메타 정보 추출
목적 수준에서 문서 요약 및
유사한 (추정) 목적들 제시
사용자의 명시적 피드백
(선택 혹은 추가) + (타겟 컬렉션)
적응형 질의 확장

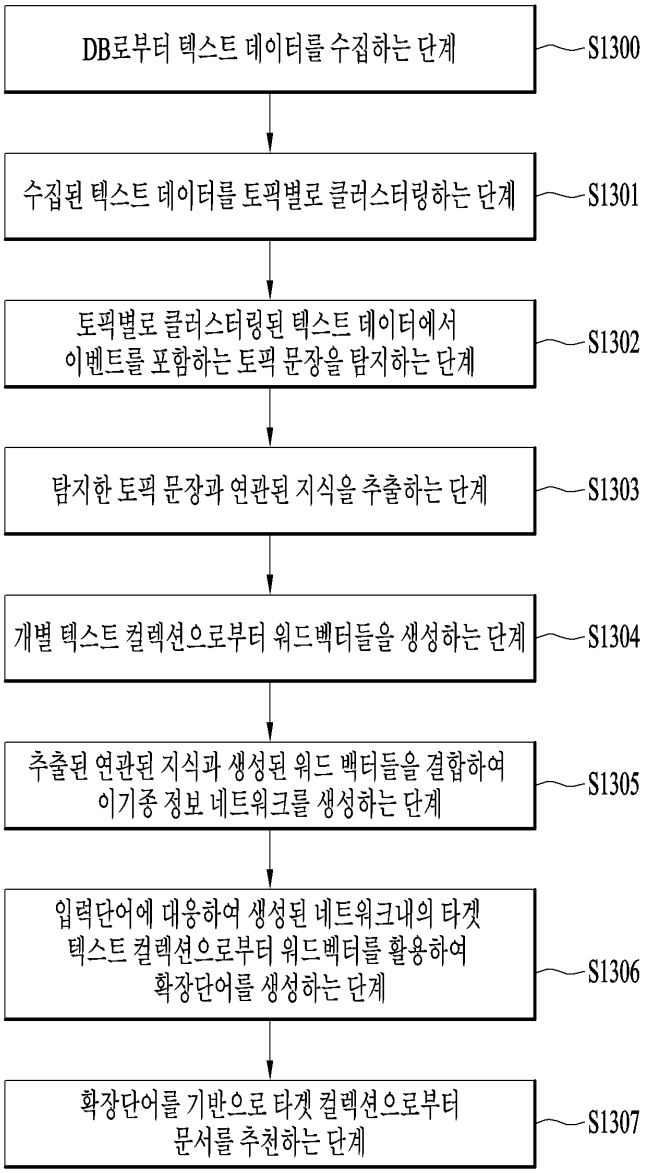
도면11


특허1.pdf
목적 요약
논문
특허
유사 특허 목록 리스트
착유컵 라이너 교환 장치 에 관한것
개방 공동 주조 장치 에 관한것
이동식 유두 컵 살균기 에 관한것
착유기 자동 점검 장치 에 관한것
요구르트 치즈 제조 장치 에 관한것
우유 제조 방법 에 관한것
생유 여과 장치 에 관한것
자동 착유 장치 에 관한것
연속식 홀딩 탱크 에 관한것
젖소 착유컵 세척기 에 관한것
착유기 세척 시스템 및 방법에 관한것
우유 거품 생성 및 스티밍 용 장치 에 관한것
같은 특허
분류에 있는
유사 목적들
목적 표현의
선택
입력 문서의
목적 수준
요약 정보

도면12


확장된 질의어
Raw Query
+ Extended Query
+ Target
Collection
Matching & Ranking
Query_similarity +
Goal_similarity +
뉴스
블로그
트위터
...
뉴스 / 소셜미디어 콘텐트
논문
특허
연구보고서
...
과학기술전문 콘텐트
문서 인덱스
목적형
필드 색인

도면13


DB로부터 텍스트 데이터를 수집하는 단계
S1300
수집된 텍스트 데이터를 토픽별로 클러스터링하는 단계
S1301
토픽별로 클러스터링된 텍스트 데이터에서
이벤트를 포함하는 토픽 문장을 탐지하는 단계
S1302
탐지한 토픽 문장과 연관된 지식을 추출하는 단계
S1303
개별 텍스트 컬렉션으로부터 워드벡터들을 생성하는 단계
S1304
추출된 연관된 지식과 생성된 워드 백터들을 결합하여
이기종 정보 네트워크를 생성하는 단계
S1305
입력단어에 대응하여 생성된 네트워크내의 타겟
텍스트 컬렉션으로부터 워드벡터를 활용하여
확장단어를 생성하는 단계
S1306
확장단어를 기반으로 타겟 컬렉션으로부터
문서를 추천하는 단계
S1307