(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2016-0124742
(43) 공개일자 2016년10월28일

(51) 국제특허분류(Int. Cl.)
*G06F 17/30* (2006.01) *G06F 17/27* (2006.01)
*G06F 19/00* (2011.01) *G06F 19/24* (2011.01)
(52) CPC특허분류
*G06F 17/30705* (2013.01)
*G06F 17/2765* (2013.01)
(21) 출원번호 10-2016-7017515
(22) 출원일자(국제) 2014년12월01일
심사청구일자 없음
(85) 번역문제출일자 2016년06월30일
(86) 국제출원번호 PCT/US2014/067918
(87) 국제공개번호 WO 2015/084724
국제공개일자 2015년06월11일
(30) 우선권주장
61/910,739 2013년12월02일 미국(US)

(71) 출원인
**큐베이스 엘엘씨**
미국 버지니아 20191 레스턴 스위트 300 선라이즈 밸리 드라이브 12018
(72) 발명자
**라이트너 스캇**
미국 버지니아 20175 리즈버그 레드힐 매너 코트 22596
**베케서 프란츠**
미국 오하이오 45370 스프링 밸리 센터빌 로드 3942 이
*(뒷면에 계속)*
(74) 대리인
**특허법인신성**

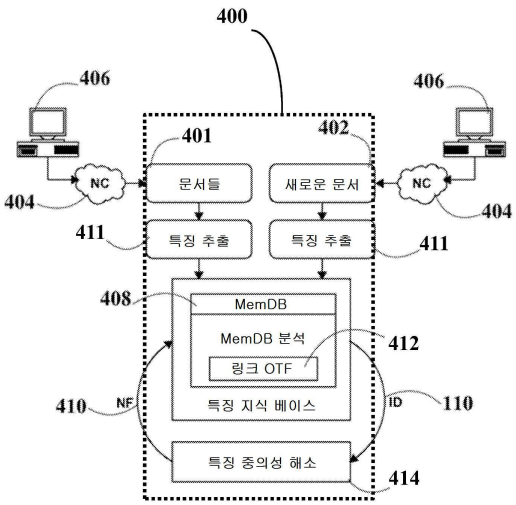
전체 청구항 수 : 총 20 항

(54) 발명의 명칭 **비정형 텍스트내의 특징들의 중의성을 해소하는 방법**

## (57) 요 약

비정형 텍스트에 있어서의 특징들의 중의성을 해소하는 방법이 제공된다. 개시된 방법은 기존의 링크가 존재할 것을 요구하지 않는다. 비정형 텍스트에 있어서의 특징들의 중의성을 해소하는 방법은 소스 문선 및 대형 문서 코퍼스로부터 도출되는 동시 발생 특징들을 이용할 수 있다. 개시된 방법은 소스 문서로부터 도출된 특징들을 기존 지식베이스의 동시 발생 특징에 링크 접속시키는 링그 접속 모듈을 포함하는, 다수의 모듈을 포함할 수 있다. 특징들의 중의성을 해소하는 개시된 방법은 고유한 동시 발생 특징들의 세트를 가진 엔티티들을 포함하는 지식베이스로부터 고유 엔티티들을 식별할 수 있게 하여, 지식 발견 및 탐색 결과에 있이서 정밀성의 증가가 가능하게 하며, 대형 코퍼스를 통해 진보된 분석 방법을 채용하고, 엔티티들, 동시 발생 엔티티들, 토픽 ID 및 다른 도출된 특징들의 조합을 채용한다.

*대 표 도* - 도4



(52) CPC특허분류
*G06F 17/30684* (2013.01)
*G06F 19/24* (2013.01)
*G06F 19/707* (2013.01)
(72) 발명자
**보두 산제이**
미국 오하이오 45440 데이턴 서머셋 패스 4408
**데이브 라케시**
미국 오하이오 45440 데이턴 앰허스트 벤드 376
**플랙 로버트**
미국 메인 04101 포틀랜드 아파트1 이스턴 프롬 6

# *명 세 서*

## *청구범위*

### 청구항 1

인-메모리 데이터베이스(in-memory database)를 호스팅(hosting)하는 시스템의 노드(node)가, 후보 레코드들(candidate records)의 세트를 탐색하여 하나 이상의 추출 특징(extracted feature)들과 매칭되는 하나 이상의 후보를 식별하고 - 후보와 매칭되는 추출 특징이 1차 특징(primary feature)임 - ;

상기 노드가, 상기 추출 특징들 각자를 하나 이상의 기계 생성 토픽 식별자들(machine-generated topic identifiers(토픽 ID))과 연계시키고;

상기 노드가, 토픽 ID들의 관련성에 기초하여 상기 1차 특징들 각자의 중의성을 서로 간에 해소하고;

상기 노드가, 토픽 ID들의 관련성에 기초하여 상기 1차 특징들 각자와 연계된 2차 특징들의 세트를 식별하고;

상기 노드가, 토픽 ID들의 관련성에 기초하여 상기 연계된 2차 특징들의 세트내의 2차 특징들 각자로부터 상기 1차 특징들 각자의 중의성을 해소하고;

상기 노드가, 상기 연계된 2차 특징들의 세트에 상기 1차 특징들의 각자를 링크 접속시켜 새로운 클러스터를 형성하고;

상기 노드가, 상기 새로운 클러스터가 기존의 지식베이스 클러스터와 매칭되는지 판정하되, 매칭이 있으면, 인-메모리 데이터베이스의 서버 컴퓨터의 중의성 해소 모듈이, 지식베이스 클러스터 내의 각각의 매칭되는 1차 특징에 대응하는 기존의 고유 식별자("고유 ID")를 판정하고, 새로운 클러스터를 포함하도록 지식베이스 클러스터를 갱신하고, 매칭이 없으면, 상기 노드가, 새로운 지식베이스 클러스터를 생성하고 새로운 지식베이스 클러스터의 1차 특징에 새로운 고유 ID를 할당하고 ;

상기 노드가, 상기 1차 특징에 대한 기존의 고유 ID와 새로운 고유 ID 중 하나를 전송하는 것을 구비하는

방법.

### 청구항 2

제 1 항에 있어서,

상기 노드가, 추출 특징과 매칭되는 후보 레코드들의 각각을 비교하고;

상기 노드가, 상기 비교에 기초하여 추출 특징들의 각각에 가중 매치 스코어 결과(weighted match score result)를 할당하는 것을 더 구비하는

방법.

### 청구항 3

제 2 항에 있어서,

상기 노드가, 상기 추출 특징들의 각각을 가중 특징 속성들(weighted feature attributes)의 세트와 연계시키는 것을 더 구비하는

방법.

### 청구항 4

제 3 항에 있어서,

상기 노드가, 하나 이상의 가중 특징 속성들에 기초하여, 추출 특징들의 각각의 관련성을 판정하는 것을 더 구비하는

방법.

**청구항 5**

제 1 항에 있어서,

상기 노드의 추출 모듈이 하나 이상의 추출 특징들을 인식하고 추출하고 - 하나 이상의 1차 특징들이 하나 이상의 추출 특징들에서 식별됨 -;

상기 노드의 추출 모듈이 추출 특징들의 각각을 데이터베이스내에 저장하는 것을 더 구비하는

방법.

**청구항 6**

제 5 항에 있어서,

상기 노드의 추출 모듈이, 상기 특징들의 각각에 추출 확실성 스코어를 할당하는 것을 더 구비하는

방법.

**청구항 7**

제 1 항에 있어서,

각 1차 특징은 하나 이상의 특징 속성들의 세트와 연계되는

방법.

**청구항 8**

제 7 항에 있어서,

토픽 ID, 문서 식별자(문서 ID), 특징 유형, 특징 이름, 신뢰 스코어 및 특징 위치를 포함하는 그룹으로부터 특징 속성이 선택되는

방법.

**청구항 9**

제 1 항에 있어서,

각 연계된 특징은 사전 정의된 클러스터 계층에 따라 낮은 순서의 특징(lower-ordinal feature)들의 세트와 연계되는

방법.

**청구항 10**

제 1 항에 있어서,

상기 노드가, 후보 레코드들의 세트의 퍼지 키 탐색(fuzzy key search)을 실행하는 것을 더 구비하는

방법.

**청구항 11**

제 7 항에 있어서,

상기 노드의 링크 온더플라이 모듈(link-on-the-fly module)이 하나 이상의 특징 속성들 및 관련된 토픽 ID들의 동시 발생에 기초하여 둘 이상의 데이터 소스들을 링크 접속시키는 것을 더 구비하는

방법.

**청구항 12**

제 1 항에 있어서,

상기 노드가, 제 2 데이터 소스내의 특징과 추출 특징을 비교함에 의해, 데이터 소스내의 추출 특징이 제 2 데이터 소스에서 동시에 발생하는지를 판정하고;

상기 노드가, 상기 비교에 기초하여 데이터 소스들의 각각을 링크 접속시키는 것을 더 구비하는

방법.

**청구항 13**

제 1 항에 있어서,

상기 노드가, 추출 특징의 중의성 해소의 정확성을 개선하기 위해, 다른 데이터 소스들로부터 추출 특징의 동시 발생을 분석하는 것을 더 구비하는

방법.

**청구항 14**

제 1 항에 있어서,

상기 노드가, 하나 이상의 새로운 데이터 소스를 계속적으로 수신하고;

상기 노드가, 하나 이상의 추출 특징들을 계속적으로 추출하고;

상기 노드가, 하나 이상의 추출 특징들에 대해 후보 탐색을 계속적으로 실행하고;

상기 노드가, 추출 특징들의 중의성을 계속적으로 해소하고;

상기 노드가, 추출 특징들을 하나 이상의 새로운 클러스터내로 계속적으로 링크 접속시키는 것을 더 구비하는

방법.

**청구항 15**

컴퓨터 실행가능 명령어를 저장한 비일시적 컴퓨터 독출가능 매체로서,

인-메모리 데이터베이스(in-memory database)를 호스팅(hosting)하는 시스템의 노드(node)가, 후보 레코드들(candidate records)의 세트를 탐색하여 하나 이상의 추출 특징(extracted feature)들과 매칭되는 하나 이상의 후보를 식별하고 - 후보와 매칭되는 추출 특징은 1차 특징(primary feature)임 - ;

상기 노드가, 추출 특징들 각자를 하나 이상의 기계 생성 토픽 식별자들(machine-generated topic

identifiers(토픽 ID))과 연계시키고;

상기 노드가, 토픽 ID들의 관련성에 기초하여 1차 특징들 각자의 중의성을 서로 간에 해소하고;

상기 노드가, 토픽 ID들의 관련성에 기초하여 1차 특징들 각자와 연계된 2차 특징들의 세트를 식별하고;

상기 노드가, 토픽 ID들의 관련성에 기초하여 상기 연계된 2차 특징들의 세트내의 2차 특징들 각자로부터 1차 특징들 각자의 중의성을 해소하고;

상기 노드가, 상기 연계된 2차 특징들의 세트에 1차 특징들의 각자를 링크 접속시켜 새로운 클러스터를 형성하고;

상기 노드가, 새로운 클러스터가 기존의 지식베이스 클러스터와 매칭되는지 판정하되; 매칭이 있으면, 상기 노드가, 지식베이스 클러스터 내의 각각의 매칭되는 1차 특징에 대응하는 기존의 고유 식별자("고유 ID")를 판정하여, 새로운 클러스터를 포함하도록 지식베이스 클러스터를 갱신하고, 매칭이 없으면 새로운 지식베이스 클러스터를 생성하여 새로운 지식베이스 클러스터의 1차 특징에 새로운 고유 ID를 할당하고;

상기 노드가, 1차 특징에 대한 기존 고유 ID와 새로운 고유 ID 중 하나를 전송하는 것을 구비하는

컴퓨터 독출 가능 매체.

**청구항 16**

제 15 항에 있어서,

상기 명령어는,

상기 노드가, 추출 특징과 매칭되는 후보 레코드들의 각각을 비교하고;

상기 비교에 기초하여 추출 특징들의 각각에 가중 매치 스코어 결과(weighted match score result)를 할당하는 것을 더 구비하는

컴퓨터 독출 가능 매체.

**청구항 17**

제 16 항에 있어서,

상기 명령어는,

상기 노드가, 상기 추출 특징들의 각각을 가중 특징 속성들(weighted feature attributes)의 세트와 연계시키는 것을 더 구비하는

컴퓨터 독출 가능 매체.

**청구항 18**

제 17 항에 있어서,

상기 명령어는,

상기 노드가, 하나 이상의 가중 특징 속성들에 기초하여, 추출 특징들의 각각의 관련성을 판정하는 것을 더 구비하는

컴퓨터 독출 가능 매체.

**청구항 19**

제 15 항에 있어서,

상기 명령어는,

상기 노드의 추출 모듈이 하나 이상의 추출 특징들을 인식하고 추출하고 - 하나 이상의 1차 특징들이 하나 이상의 추출 특징들에서 식별됨 -;

상기 노드의 추출 모듈이 추출 특징들의 각각을 데이터베이스내에 저장하는 것을 더 구비하는

컴퓨터 독출 가능 매체.

**청구항 20**

제 19 항에 있어서,

상기 명령어는,

상기 노드의 추출 모듈이, 상기 특징들의 각각에 추출 확실성 스코어를 할당하는 것을 더 구비하는

컴퓨터 독출 가능 매체.

## *발명의 설명*

### *기 술 분 야*

[0001] 본 개시는 통상적으로 데이터 관리에 관한 것으로, 보다 구체적으로는, 네트워크를 통해 수신된 소스 아이템들(source items)로부터 자료(material)를 추출하여 저장하는 방법 및 데이터 관리 시스템에 관한 것이다.

### *배 경 기 술*

[0002] 네트워크와 같은 소스를 포함하는, 대형 문서 콜렉션(large document collection)에서 엔티티들(즉, 사람, 위치, 조직)에 대한 정보를 탐색하는 것은 불명확할 수 있어서, 부정확한 텍스트 프로세싱, 지식 추출(knowledge extract) 동안의 특징들의 부정확한 연계성 및 그에 따른 부정확한 데이터 분석으로 귀결될 수 있다.

[0003] 최신 시스템들은 페이지랭크(PageRank) 및 HITS(Hyperlink-Induced Topic Search) 알고리즘과 같은 여러 알고리즘에 링크 기반 클러스터링 및 랭킹(linkage based clustering and ranking)을 이용한다. 이러한 관련 방식들 이면의 기본적인 개념은 기존의 링크들이 전형적으로 관련 페이지(page)들 또는 콘셉트(concept)들간에 존재한다는 것이다. 클러스터링 기반 기술의 한계는 엔티티들의 중의성을 해소하는데 필요한 문맥 정보(contextual information)가 콘텍스트(context)내에 존재하지 않아, 잘못된 중의성 해소 결과(incorrectly disambiguated result)를 이끄는 경우가 있다는 것이다. 유사하게, 동일하거나 피상적으로 유사한 콘텍스트내의 다른 엔티티들에 대한 문서들이 부정확하게 함께 클러스터링될 수도 있다.

[0004] 다른 시스템들은 엔티티들의 하나 이상의 외부 사전(또는 지식베이스(knowledgebase))를 참조하여 엔티티들의 중의성을 해소하고자 한다. 그러한 시스템에서는, 엔티티의 콘텍스트가 그 사전에 있는 가능성있는 매칭 엔티티들과 비교되고, 가장 근접한 매칭이 리턴(return)된다. 현재의 사전 기반 기술과 연계된 제약은, 엔티티들의 개수가 언제라도 증가할 수 있고, 그에 따라 전 세계의 모든 엔티티들의 표현을 포함하는 사전은 없다는 사실로부터 기인하다. 따라서, 문서의 콘텍스트가 사전내의 엔티티와 매칭되면, 그 기술은 그 사전내의 가장 유사한 엔티티만을 식별한 것일 뿐, 사전 외부에 있을 수 있는 정확한 엔티티를 필수적으로 식별한 것은 아닐 수 있다.

[0005] 대부분의 방법들은 단지 중의성 해소 프로세스에 엔티티들과 키 구문(key phrase)만을 이용한다. 그러므로, 정밀한 데이터 분석을 허용하는 정확한 엔티티 중의성 해소 기술이 여전히 필요하다.

## *발명의 내용*

### *해결하려는 과제*

[0006] 일부 실시 예들은 특징들의 중의성을 해소하는 방법을 설명한다. 그 방법은 하나 이상의 특징 추출 모듈, 하나 이상의 중의성 해소 모듈, 하나 이상의 스코어링 모듈(scoring modules) 및 하나 이상의 링크 접속 모듈

(linking modules)과 같은 다수의 모듈을 포함할 수 있다.

## 과제의 해결 수단

[0007] 특징들의 중의성 해소는, MC-LDA(multi-component extension of Latent Dirichlet Allocation) 토픽 모델을 채용하여, 특징의 주변 문서로부터 토픽(topic)들을 추출함에 의해 부분적으로 지원된다. 여기에서, 각 구성 요소(component)는 기존 지식 베이스내에 저장된 각각의 2차적 특징을 기준으로 모델링되거나 수신 문서(incoming document)상에서 추출된다. 또한, 링크 접속 또는 중의성 해소 프로세스는 MC-LDA로부터의 토픽 추론(topic inference)으로서 모델링되어, MC-LDA 트레이닝(training) 동안 자동 가중 추정(automated weight estimation)을 제공하고 추론 동안에 실질적으로 그들을 적용한다.

[0008] 예시적인 방법은 문서 링크 무접속(no document linking)을 고려함에 의해 달성될 수 있는 것 이외에, 엔티티 중의성 해소의 정확성을 개선할 수 있다. 문서 링크를 고려할 경우 링크에 의해 암시되는 문서 및 엔티티 관련성을 고려함에 의해 보다 나은 중의성 해소가 가능하게 된다.

[0009] 일 실시예에 있어서, 방법은, 인-메모리 데이터베이스(in-memory database)를 호스팅(hosting)하는 시스템의 노드(node)가, 후보 레코드들(candidate records)의 세트를 탐색하여 하나 이상의 추출 특징(extracted feature)들과 매칭되는 하나 이상의 후보를 식별하고 - 후보와 매칭되는 추출 특징은 1차 특징(primary feature)임 - ; 노드가, 추출 특징들 각자를 하나 이상의 기계 생성 토픽 식별자(machine-generated topic identifiers(토픽 ID))와 연계시키고; 노드가 토픽 ID들의 관련성에 기초하여 1차 특징들 각자의 중의성을 서로 간에 해소하고; 노드가, 토픽 ID들의 연관성에 기초하여 1차 특징들 각자와 연계된 2차 특징들의 세트를 식별하고; 노드가 토픽 ID들의 연관성에 기초하여 2차 특징들의 연계 세트내의 2차 특징들 각자로부터 1차 특징들 각자의 중의성을 해소하고; 노드가 2차 특징들의 연계 세트에 1차 특징들의 각자를 링크 접속시켜 새로운 클러스터를 형성하고; 노드가, 새로운 클러스터가 기존의 지식베이스 클러스터와 매칭되는지 판정하고 - 매치가 있으면, 인-메모리 데이터베이스 서버 컴퓨터의 중의성 해소 모듈이, 지식베이스 클러스터 내의 각각의 매칭되는 1차 특징에 대응하는 기존의 고유 식별자("고유 ID")를 판정하여, 새로운 클러스터를 포함하도록 지식베이스 클러스터를 갱신하고, 매치가 없으면 노드가 새로운 지식베이스 클러스터를 생성하여 새로운 지식베이스 클러스터의 1차 특징에 새로운 고유 ID를 할당함 -; 노드가 1차 특징에 대한 기존 고유 ID와 새로운 고유 ID 중 하나를 전송하는 것을 구비한다.

[0010] 다른 실시 예에 있어서, 컴퓨터 실행가능 명령어를 저장한 비일시적 컴퓨터 독출가능 매체는 인-메모리 데이터베이스(in-memory database)를 호스팅(hosting)하는 시스템의 노드(node)가, 후보 레코드들(candidate records)의 세트를 탐색하여 하나 이상의 추출 특징(extracted feature)들과 매칭되는 하나 이상의 후보를 식별하고 - 후보와 매칭되는 추출 특징은 1차 특징(primary feature)임 - ; 노드가, 추출 특징들 각자를 하나 이상의 기계 생성 토픽 식별자(machine-generated topic identifiers(토픽 ID))와 연계시키고; 노드가 토픽 ID들의 관련성에 기초하여 1차 특징들 각자의 중의성을 서로 간에 해소하고; 노드가, 토픽 ID들의 관련성에 기초하여 1차 특징들 각자와 연계된 2차 특징들의 세트를 식별하고; 노드가 토픽 ID들의 관련성에 기초하여 2차 특징들의 연계 세트내의 2차 특징들 각자로부터 1차 특징들 각자의 중의성을 해소하고; 노드가 2차 특징들의 연계 세트에 1차 특징들의 각자를 링크 접속시켜 새로운 클러스터를 형성하고; 노드가, 새로운 클러스터가 기존의 지식베이스 클러스터와 매칭되는지 판정하고 - 매치가 있으면, 노드가, 지식베이스 클러스터 내의 각각의 매칭되는 1차 특징에 대응하는 기존의 고유 식별자("고유 ID")를 판정하여, 새로운 클러스터를 포함하도록 지식베이스 클러스터를 갱신하고, 매치가 없으면 새로운 지식베이스 클러스터를 생성하여 새로운 지식베이스 클러스터의 1차 특징에 새로운 고유 ID를 할당함 -; 노드가 1차 특징에 대한 기존 고유 ID와 새로운 고유 ID 중 하나를 전송하는 것을 구비한다.

[0011] 실시 예의 추가적인 특징 및 장점은 이하의 상세한 설명에서 설명될 것이며, 부분적으로 그 설명으로부터 명확해질 것이다. 본 발명의 목적 및 다른 장점은 살세한 설명, 그의 청구범위 및 첨부 도면에서의 예시적인 실시예에서 특정하게 지정된 구조에 의해 실현 및 이루어질 수 있을 것이다.

[0012] 상술한 전반적인 설명 및 이하의 상세한 설명은 예시적이고 설명을 위한 것이며, 청구된 바와 같은 본 발명의 추가적인 설명을 제공하기 위한 것이다.

## 도면의 간단한 설명

[0013] 본 개시는 이하의 도면을 참조하면 더욱 잘 이해될 수 있을 것이다. 첨부된 도면은 본 명세서의 일부를 구성하

며, 본 발명의 실시 예를 도시한 것으로, 명세서와 함께 본 발명을 설명한다. 도면에서의 구성 요소는 반드시 축척으로 도시된 것은 아니며, 대신에 본 개시의 원리의 설명이 강조되어 있다. 도면에 있어서, 참조 번호들은 다른 도면들간에 대응하는 부분을 나타낸다.

도 1은 예시적인 실시 예에 따라 비정형 텍스트에 있어서의 특징들의 중의성을 해소하는 방법의 흐름도이다.

도 2는 예시적인 실시 예에 따라, 특징들의 중의성을 해소하는 방법에 채용되는 중의성 해소 모듈에 의해 실행되는 단계들의 흐름도이다.

도 3은 예시적인 실시 예에 따라, 특징들의 중의성을 해소하는 방법에 채용되는 링크 온더플라이 모듈(link on-the-fly module)에 의해 실행되는 단계들의 흐름도이다.

도 4는 예시적인 실시 예에 따라, 특징들의 중의성을 해소하는 방법을 구현하는데 채용된 시스템의 도면이다.

도 5는 예시적인 실시 예에 따라, MC-LDA(multi-component, conditionally-independent Latent Dirichlet Allocation) 토픽 모델의 그래픽도이다.

도 6은 예시적인 실시 예에 따라, MC-LDA 토픽 모델에 대한 Gibbs 샘플링 수학식의 실시 예를 나타낸 도면이다.

도 7은 예시적인 실시 예에 따라, MC-LDA 토픽 모델에 있어서의 트레이닝 및 추론을 위한 확률론적 변분 추론 알고리즘(stochastic variational inference algorithm)의 구현의 실시 예를 나타낸 도면이다.

도 8은 예시적인 실시 예에 따라, MC-LDA 토픽 모델에 대한 샘플 토픽을 도시한 테이블이다.

## *발명을 실시하기 위한 구체적인 내용*

[0014] **정의**

[0015] 본 명세서에 이용된 이하의 용어들은 다음과 같은 정의를 가질 수 있다.

[0016] "문서"는 시작 및 종료를 가진 정보의 이산 전자 표현을 지칭한다.

[0017] "멀티-문서"는 개별적인 "백오프서피스폼(bag-of-surface-forms)" 구성 요소들로 조직된 토큰, 서로 다른 유형들의 지명 엔티티들(named entities) 및 키 구문들을 가진 문서를 지칭한다.

[0018] "데이터베이스"는 하나 이상의 콜렉션들을 저장하는데 적합하고 하나 이상의 조회(quiry)를 처리하는데 적합한 모듈들 및 클러스터들의 임의 조합을 포함하는 임의 시스템을 지칭한다.

[0019] "코퍼스(corpus)"는 하나 이상의 문서들의 콜렉션을 지칭한다.

[0020] "라이브 코퍼스(live corpus)" 또는 "문서 스트림(document stream)"은 새로운 문서들이 네트워크내에 업로드됨에 따라 일정하게 피딩(feeding)되는 코퍼스를 지칭한다.

[0021] "특징(feature)"은 문서로부터 적어도 부분적으로 도출되는 임의 정보를 지칭한다.

[0022] "특징 속성(feature attribute)"은 다른 것들 중에서 문서내의 특징들의 위치, 신뢰 스코어와 같은 특징과 연계된 메타데이터를 지칭한다.

[0023] "클러스터"는 특징들의 콜렉션을 지칭한다.

[0024] "엔티티 지식베이스"는 특징들/엔티티들을 포함하는 베이스를 지칭한다.

[0025] "링크 온더플라이 모듈(link on-the-fly module)" 또는 "링크 OTF"는 라이브 코퍼스가 갱신됨에 따라 데이터를 갱신하는 임의 링크 접속 모듈을 지칭한다.

[0026] "메모리"는 충분히 높은 속도로 정보를 저장하고 정보를 탐색하는데 적합한 임의 하드웨어 구성 요소를 지칭한다.

[0027] "모듈"은 하나 이상의 정의된 태스크(task)를 실행하는데 적합한 컴퓨터 소프트웨어 구성 요소를 지칭한다.

[0028] "센티먼트(sentiment)"는 문서, 문서의 일부 또는 특징과 연계된 주관적 평가를 지칭한다.

[0029] "토픽"은 코퍼스로부터 적어도 부분적으로 도출되는 테마 정보(thematic information) 세트를 지칭한다.

[0030] "토픽 식별자" 또는 "토픽 ID"는 토픽의 특정 인스턴스(specific instance)라고 지칭되는 식별자를 지칭한다.

[0031] "토픽 콜렉션"은 각 토픽이 고유 식별자("고유 ID")를 가진, 코퍼스로부터 도출되는 토픽들의 특정 세트를 지칭한다.

[0032] "토픽 분류"는 문서의 특징으로서 특정 토픽 식별자의 할당을 지칭한다.

[0033] "조회"는 하나 이상의 적당한 데이터베이스로부터 정보를 검색하기 위한 요청을 지칭한다.

[0034] 바람직한 실시 예에 대한 참조가 세부적으로 이루어질 것이며, 그의 예시들이 첨부 도면에 도시된다. 본 명세서에서 설명한 실시 예들은 예시적인 것이다. 당업자라면 수많은 대안적인 구성 요소들 및 실시 예들이 본 명세서에서 설명하는 특정 예시를 대신할 수 있고, 본 발명의 범주내에 있음을 알 것이다.

[0035] 본 개시는 비정형 텍스트에 있어서의 특징들의 중의성을 해소하는 방법을 설명한다. 예시적인 실시 예가 본 개시에 따른 특징들의 중의성을 해소하는 실시를 설명하지만, 본 명세서에서 설명하는 시스템 및 방법은 본 개시의 범주내에서 임의의 적당한 이용을 위해 구성될 수 있다.

[0036] 기존의 지식 베이스는 불명확하지 않은 특징 및 그들의 관련 특징들을 포함함으로써, 낮은 신뢰성의 텍스트 분석으로 귀결될 수 있다. 본 개시의 측면은 특징 및 엔티티 중의성 해소에 있어서 정확성이 증가되고 텍스트 분석에 있어서 정확성이 증가될 수 있다.

[0037] 실시 예에 따르면, 특징들의 중의성을 해소하는 개시된 방법은 데이터의 초기 코퍼스에 채용되어, 초기 코퍼스에 포함된 각 문서에 대한 토픽 분류 및 다른 텍스트 분석을 가능하게 하는 문서 수용(documnent ingestion) 및 특징 추출을 실행한다. 각 특징은 다른 것들 중에서도 문서내의 이름, 유형 및 위치 정보와, 신뢰 스코어로서 식별되고 기록될 수 있다.

[0038] 도 1은 비정형 텍스트내의 특징들의 중의성을 해소하는 다수의 단계들을 도시한 방법(100)의 흐름도이다. 실시 예에 따르면, 특징들의 중의성을 해소하는 방법(100)은, 기존 지식 베이스에서 새로운 문서 입력 단계(102)가 이루어짐에 따라 개시된다. 후속하여, 그 문서에 대해 특징 추출 단계(104)가 실행된다. 실시 예에 따르면, 특징은, 다른 것들 중에서도, 예를 들어, 토픽 식별자(토픽 ID), 문서 식별자(문서 ID), 특징 유형, 특징 이름, 신뢰 스코어 및 특징 위치와 같은 서로 다른 특징 속성과 관련될 수 있다.

[0039] 여러 실시 예에 따르면, 단계 102에서 입력된 문서는 (인터넷 또는 네트워크 접속 코퍼스와 같은) 대형 코퍼스 또는 라이브 코퍼스로부터 피딩되고, 그 다음  매순간마다 피딩될 수 있다.

[0040] 다른 실시 예들에 따르면, 특징 추출 단계(104) 동안에 하나 이상의 특징 인식 및 추출 알고리즘이 채용되어 문서 입력 단계(102)의 비정형 텍스트를 분석할 수 있다. 각 추출 특징에 스코어가 할당될 수 있다. 그 스코어는 정확한 속성을 가진채 정확하게 추출되는 특징의 확실성 레벨을 나타낼 수 있다.

[0041] 추가적으로, 특징 추출 단계(104) 동안, 단계(102)에서 입력된 문서로부터 하나 이상의 1차 특징들이 식별될 수 있다. 각 1차 특징은 특징 속성들의 세트 및 하나 이상의 2차 특징과 연계되어 있을 수 있다. 각 2차 특징은 특징 속성들의 세트와 연계될 수 있다. 일부 실시 예에 있어서, 하나 이상의 2차 특징들은 각각이 그 자신의 특징 속성 세트들 가지고 있는 하나 이상의 3차 특징들을 가질 수 있다.

[0042] 특징 속성들을 고려하여, 단계(102)에서 입력된 문서내의 특징들 각각의 상대적 가중치(relative weight) 또는 관련성이 판정될 수 있다. 추가적으로, 가중 스코어링 모델(weighted scoring model)을 이용하여 특징들간의 연계 관련성이 판정될 수 있다.

[0043] 특징 추출 단계(104)에 이어서, 특징들의 중의성 해소 요청 단계(108)의 일부로서, MemDB(in-memory database)내에 특징들을 포함시키는 단계 106 동안, 단계(102)에서 입력된 문서로부터 추출된 특징들과 그와 관련된 정보 모두가 MemDB에 로딩될 수 있다.

[0044] 실시 예에 있어서, MemDB는 도 1 내지 도 8과 관련하여 설명되는 단계들을 실행하는 하나 이상의 프로세서를 가진 중의성 해소 컴퓨터 서버 환경의 일부를 형성한다. 일 실시 예에 있어서, MemDB는 하나 이상의 탐색 제어기, 다수의 탐색 모듈들, 압축 데이터의 콜렉션 및 중의성 해소 서브 모듈을 포함할 수 있는 컴퓨터 모듈이다. 하나의 탐색 제어기는 하나 이상의 탐색 노드에 선택적으로 연계될 수 있다. 각 탐색 노드는 압축 데이터의 콜렉션을 통해 퍼지 키 탐색(fuzzy key search)을 독자적으로 실행하고 그와 연계된 탐색 제어기에 스코어 결과 세트를 리턴할 수 있다.

[0045] 특징 중의성 해소 단계(108)는 MemDB내의 중의성 해소 모듈에 의해 실행될 수 있다. 특징 중의성 해소(108) 프로세스는 특징들, 문서들 또는 코퍼스들을 분류하기 위해 채용될 수 있는 기계 생성 토픽 ID를 포함할 수 있다.

개별 특징들과 특정 토픽 ID들의 관련성은 중의성 해소 알고리즘을 이용하여 판정될 수 있다. 일부 문서들에서는, 문서내의 특징의 서로 다른 발생 상황에 따라 동일한 특징이 하나 이상의 토픽 ID들과 관련될 수 있다.

[0046] 하나의 문서로부터 추출된 (토픽, 근접 용어 및 엔티티, 키 구문, 이벤트 및 팩트(fact)와 같은) 특징 세트가 다른 문서로부터의 특징 세트와 비교될 수 있는데, 이것은 서로 다른 문서들에 걸쳐 있는 둘 이상의 특징들이 단일 특징인지, 그들이 별개의 특징인지를 일정 레벨의 정확성으로 정의하도록 중의성 해소 알고리즘을 이용한다. 일부 예시들에서는, 데이터베이스내의 문서들의 콜렉션에 걸쳐있는 둘 이상의 특징들의 동시 발생(co-occurrence)이 분석되어 특징 중의성 해소 프로세스(108)의 정확성을 개선할 수 있다. 일부 실시 예에서는, 글로벌 스코어링 알고리즘(global scoring algorithm)이 특징들의 확률이 동일함을 판정하는데 이용될 수 있다.

[0047] 일부 실시 예들에 있어서, 특징 중의성 해소 프로세스(108)의 일부로서, 지식 베이스가 MemDB내에 생성될 수 있다. 이러한 지식 베이스는 관련된 중의성 해소된 1차 특징과 그들의 관련 2차 특징들의 클러스터들을 일시 저장하는데 이용될 수 있다. 새로운 문서들이 MemDB내에 로딩되면, 특징들의 새로운 중의성 해소 세트가 기존 지식 베이스와 비교되어, 특징들간의 연관성을 판정하고, 새로운 특징들과 이미 추출된 특징들간에 매칭이 있는지를 판정할 수 있다.

[0048] 비교된 특징들이 매칭되면, 지식 베이스는 갱신될 수 있고, 매칭 특징들의 특징 ID는 사용자 및/또는 요청 애플리케이션이나 프로세스로 리턴될 수 있고, 추가로 매칭들의 빈도수에 기초하여, 프로미넌스(prominence) 측정치에 특징 ID가 부착되어, 주어진 코퍼스에 있어서의 그의 대중성 인덱스(popularity index)를 포착한다. 비교된 특징이 이미 추출된 특징들과 매칭되지 않으면, 고유 특징 ID가 중의선 해소 엔티티 또는 특징에 할당되고, 고유 특징 ID가 특징을 정의하는 클러스터와 연계되어 MemDB의 지식 베이스내에 저장된다. 후속적으로, 단계 110에서, 중의성 해소된 특징의 특징 ID가 시스템 인터페이스를 통해 소스로 리턴될 수 있다. 일부 실시 예에 있어서, 중의성 해소된 특징의 특징 ID는 2차 특징, 특징의 클러스터, 관련 특징 속성 또는 다른 요청 데이터를 포함할 수 있다. 특징 중의성 해소 단계(108)를 위해 채용된 중의성 해소 서브 모듈에 대해서는 이하의 도 2에서 보다 상세하게 설명하겠다.

[0049] 중의성 해소 서브 모듈

[0050] 도 2는, 실시 예에 따라, 방법(100)(도 1)의 특징 중의성 해소 단계(108)를 위한 비정형 텍스트에 채용된 중의성 해소 서브 모듈에 의해 실행되는 프로세스(200)의 흐름도이다. 중의성 해소 프로세스(200)는 도 1의 단계(106)에서의 MemDB내에 특징들의 포함 후에 시작될 수 있다. 단계(202)에서 제공된 추출 특징은 단계(204)에서 후보 탐색을 실행하는데 이용될 수 있으며, 단계(204)에서는, 동시 발생 특징을 포함하는 모든 후보 레코드에 대해 실행될 수 있다.

[0051] 여러 실시 예에 따르면, 후보들은 특징 중의성 해소 프로세스(108)에 이용될 수 있는 연계된 2차 특징들의 세트를 가진 1차 특징들일 수 있다.

[0052] 중의성 해소 결과는 토픽 ID들의 동시 발생 및 토픽 ID들간의 관련성에 의해 개선될 수 있다. 서로 다른 토픽 모델들간의 관련성을 포함하여 토픽 ID 들의 관련성은, 토픽 ID들이 할당되었던 대형 코퍼스로부터 발견될 수 있다. 관련된 토픽 ID들은, 레코드 링크 단계(206) 동안에, 정확한 토픽 ID를 포함하는 것이 아니라 하나 이상의 관련된 토픽 ID를 포함할 수 있는 문서에 링크를 제공하는데 이용될 수 있다. 이러한 방식은 레코드 링크 단계(206)에 포함될 관련 특징의 리콜(recall)을 개선하고, 특정 경우에는 중의성 해소 결과를 개선할 수 있다.

[0053] 일단, 잠재적으로 관련된 문서들의 세트가 식별되었고, 이들 문서내의 관련된 1차 및 2차 특징들의 세트가 추출되었으면, 레코드 링크 프로세스(206) 동안에, 특징 속성, 동일 문서의 특징들간의 관련성(의미있는 콘텍스트), 특징들의 상대적 가중치 및 다른 변수는, 문서들에 걸쳐 있는 1차 및 2차 특징들의 중의성을 해소하는데 이용될 수 있다. 그 다음, 레코드들의 각각은 중의성 해소된 특징들 및 그들의 관련 2차 특징들의 클러스터들을 판정하기 위해 다른 레코드들에 링크 접속될 수 있다. 레코드 링크(206)에 이용되는 알고리즘은 철자법 오류 또는 음역(transliterations) 및 마이닝 비정형 데이터 세트(mining unstructured data sets)의 다른 챌린지(challenge)를 극복할 수 있다.

[0054] 클러스터 비교 단계(208)는 중의성 해소된 특징들의 클러스터에 관련 매칭 스코어를 할당하는 것을 포함하는데, 다른 애플리케이션에 대해서는 다른 승인 임계치가 정의될 수 있다. 정의된 승인 레벨들은 어느 스코어가 포지티브 매치 탐색(positive match search)으로 고려될 수 있고, 어느 스코어가 네거티브 매치 탐색(negative match search)으로 고려될 수 있는지를 판정할 수 있다 (단계 210). 새로운 클러스터의 각각은 고유 ID를 제공

받으며, 지식 베이스에 일시적으로 저장될 수 있다. 각각의 새로운 클러스터는 중의성 해소된 새로운 1차 특징과, 그의 2차 특징 세트를 포함할 수 있다. 새로운 클러스터가 지식 베이스에 이미 저장된 클러스터와 매칭되면, 시스템은 단계(212)에서 지식 베이스를 갱신하고 단계(214)에서, 사용자 및/또는 요청 애플리케이션 또는 프로세스에 매칭된 특징 ID를 리턴시킨다. 지식 베이스의 갱신(212)은 1차 또는 2차 특징들과 이전에 연계되지 않았던 특징 속성들의 추가 또는 하나의 1차 특징에 2차 특징들의 추가의 연계를 암시할 수 있다.

[0055] 평가중인 클러스터가 포지티브 매치 탐색의 임계 미만의 스코어를 할당받으면(단계 210), 시스템은 단계 216에서 클러스터의 1차 특징에 고유 ID 할당을 실행하고, 지식 베이스를 갱신한다(단계 212). 이 후, 시스템은 매칭된 ID 프로세스의 리턴을 실행한다(단계 214). 레코드 링크 단계(단계 216)는 도 3에서 상세하게 추가로 설명하겠다.

[0056] 링크 온더플라이 서브 모듈

[0057] 도 3은 실시 예에 따라, 특징의 중의성을 해소하는 방법(100)에 채용되는 링크 온더플라이(링크 OTF) 서브 모듈에 의해 실행되는 프로세스(300)의 흐름도이다. 링크 OTF 프로세스(300)는 정보의 피드를 일정하게 평가하고, 스코어링하고, 링크 접속하고 클러스터링할 수 있다. 링크 OTF 서브 모듈은 다수의 알고리즘들을 이용하여 레코드 링크(206)를 실행한다. 단계 204의 후보 탐색 결과는 링크 OTF 모듈로 일정하게 피딩된다. 데이터의 입력 후, 매치 스코어링 알고리즘 애플리케이션(단계 302)이 실행되는데, 그 애플리케이션에서는 하나 이상의 매치 스코어링 알고리즘이 다른 것들 중에서도 스트링 편집 거리(string edit distance), 음성학(phonetics) 및 센티먼트와 같은 다수의 특징 속성을 고려하여, 관련 결과들을 평가 및 스코어링하기 위한 퍼지 키 탐색을 실행하면서 MemDB의 다수의 탐색 노드들에 동시에 적용될 수 있다.

[0058] 이후, 매치 스코어링 알고리즘 애플리케이션 단계(단계 302) 동안에 식별된 모든 후보 레코드들을 서로 비교하기 위해 링크 접속 알고리즘 애플리케이션 단계(단계 304)가 추가될 수 있다. 링크 접속 알고리즘 애플리케이션(단계 304)은 MemDE의 다수의 탐색 노드들 내부에서 실행되는 퍼지 키 탐색의 스코어링된 결과를 필터링하고 평가할 수 있는 하나 이상의 분석 링크 접속 알고리즘의 이용을 포함할 수 있다. 일부 예시에 있어서, MemDB에 있는 식별된 후보 레코드의 콜렉션에 걸쳐있는 2 이상의 특징들의 동시 발생이 분석되어, 프로세스의 정확성을 개선한다. 서로 다른 특징 속성들과 연계된 다른 가중 모델과 신뢰 스코어가 링크 접속 알고리즘 애플리케이션(단계 304)을 위해 고려될 수 있다.

[0059] 링크 접속 알고리즘 애플리케이션 단계(단계 304)후, 링크 접속 결과는 단계 306에서 관련 특징들의 클러스터에 배열되고, 링크 접속된 레코드 클러스터의 리턴의 일부로서 리턴된다.

[0060] 도 4는 도 1과 관련하여 상술한 바와 같이 비정형 텍스터에 있어서의 특징들의 중의성을 해소하는 시스템(400)의 실시 예의 예시적 도면이다. 시스템(400)은 인-메모리 데이터베이스를 호스팅하며, 하나 이상의 노드를 구비한다.

[0061] 실시 예에 따르면, 시스템(400)은 하나 이상의 문서들내의 특징들의 중의성을 해소하기 위하여, 다수의 전용 컴퓨터 모듈(401,402,411,412 및 414: 이하에서 설명할 것임)에 대한 컴퓨터 명령을 실행하는 하나 이상의 프로세서를 포함한다. 도 4에 도시된 바와 같이, 문서 입력 모듈(401,402)은 인터넥 기반 소스 및/또는 문서의 라이브 코퍼스로부터 문서를 수신한다. 상당량의 새로운 문서들이 네트워크 접속(404)을 통해 문서 입력 모듈(402)로 업로딩될 수 있다. 그러므로, 소스는 사용자 워크스테이션(406)에 의해 갱신된 새로운 지식을 일정하게 획득하고 있을 수 있으며, 거기에서는 그러한 새로운 지식이 정적 방식(static way)으로 사전 링크 접속되지 않는다. 따라서, 평가될 문서들의 개수가 무한정 증가하게 된다.

[0062] 이러한 평가는 MemDB 컴퓨터(408)를 통해 달성될 수 있다. MemDB(408)는 보다 빠른 중의성 해소 프로세스를 도모하고, 중의성 해소 프로세스 온더플라이를 도모하여, MemDB(408)에 기고될 예정인 가장 최근의 정보의 수신을 도모한다. 특징들을 링크 결합시키는 여러 방법이 채용될 수 있는데, 이들은 어느 엔티티 유형들이 가장 중요한지 및 어느 것이 더 높은 가중치를 가지는지를 판정하는 가중 모델을 필수적으로 이용하고, 신뢰 스코어에 기초하여 정확한 특징들의 추출 및 중의성 해소가 어느 정도의 신뢰도로 실행되었는지를 판정하며, 정확한 특징이 특징들의 결과 클러스터로 들어갈 수 있다. 도 4에 도시된 바와 같이, 보다 많은 시스템 노드들이 병렬로 작업함에 따라 그 프로세스는 보다 효율적으로 된다.

[0063] 여러 실시 예에 따르면, 네트워크 접속(404)을 통해 문서 입력 모듈(401,402)을 거쳐 새로운 문서가 시스템(400)내로 도착하면, 추출 모듈(411)을 통해 특징 추출이 실행되고, 그 다음, MemDB(408)의 특징 중의성 해소 서브 모듈(414)를 통해 새로운 문서에 대해 특징 중의성 해소가 실행될 수 있다. 일 실시 예에 있어서, 새로운

문서의 특징 중의성 해소가 실행된 후, 추출된 새로운 특징(410)은 링크 OTF 서브 모듈(412)을 통과하도록 MemDB에 포함될 수 있으며, 거기에서는 특징들이 비교되고 링크 접속되며, 중의성 해소된 특징의 특징 ID(110)가 조회로부터의 결과로서 사용자에게 리턴될 수 있다. 특징 ID에 추가하여, 중의성 해소된 특징을 정의하는 결과하는 특징 클러스터가 선택적으로 리턴될 수 있다.

[0064] MemDB 컴퓨터(408)는, "디스크" 메모리에 데이터를 저장하는 종래의 데이터베이스 관리 시스템(database management system: DBMS)(도시되지 않음) 모듈들 및 데이터 베이스와는 반대로, 디바이스의 주 메모리에 데이터 레코드를 저장하도록 구성된 DBMS에 의해 제어되는, 레코드에 데이터를 저장하는 데이터베이스일 수 있다. 종래의 디스크 저장 장치는 프로세서(CPU들)가 디바이스의 하드 디스크에 독출 및 기록 명령들을 실행하도록 요청하고, 그에 따라 CPU는 데이터에 대한 메모리 위치를 위치 결정하고(탐지(seek)), 검색하기 위한 명령을 실행하도록 요청한다. 인-메모리 데이터베이스 시스템은 주 메모리내에 자리하여 어드레싱된 데이터를 액세스하며, 그에 의해 CPU에 의해 실행되는 명령어들의 개수를 경감시키고, 하드 디스크상의 데이터를 탐지하는 CPU와 연계된 탐지 시간을 제거한다.

[0065] 인-메모리 데이터베이스는 노드들의 각 리소스(예를 들어, 메모리, 디스크, 프로세서)들을 결집시키도록 구성된 하나 이상의 노드들을 구비하는 컴퓨팅 시스템일 수 있는 분산형 컴퓨팅 아키텍쳐에 구현될 수 있다. 본 명세서에서 개시된 바와 같이, 인-메모리 데이터베이스를 호스팅하는 컴퓨팅 시스템의 실시 예는 하나 이상의 노드들 간에 데이터베이스의 데이터 레코드들을 분산 저장할 수 있다. 일부 실시 예들에 있어서, 이들 노드들은 노드들의 "클러스터들"로 된다. 일부 실시 예들에 있어서, 이들 노드들의 클러스터들은 데이터베이스 정보의 일부 또는 "콜렉션"을 저장한다.

[0066] 여러 실시 예들은, 동시 발생 토픽들, 키 구문들, 근접 용어들, 이벤트, 팩트 및 동향 대중성 인덱스(trending popularity index)와 같은 2차 특징들을 저장하도록 구성되는 진화 및 효율적 링크 접속 가능 특징 지식 베이스(evolving and efficiently linkable feature knowledge base)를 채용하는 컴퓨터 실행 특징 중의성 해소 기술을 제공한다. 개시된 실시 예는 주어진 추출 특징을 지식 베이스내의 저장된 특징으로 변형시키는데 도움을 주는 수반된 2차 특징들의 디멘션(dimensions)에 기초하여 단순한 개념적 거리 측정에서 정교한 그래픽 클러스터링 방식으로 가변할 수 있는 아주 다양한 링크 접속 알고리즘을 통해 실행될 수 있다. 추가적으로, 실시 예들은 기존의 특징 엔트리의 2차 특징을 갱신하고, 지식 베이스에 추가될 수 있는 새로운 특징을 발견함에 의해 그것을 확장하는 기능(capability)에 의해 기존의 특징 지식 베이스를 진화시키기 위한 방식을 도입할 수 있다.

[0067] 중의성 해소 방식의 실시 예는 토픽 인터페이스로서 모델링되는 (모든 2차 특징들에 걸쳐) 자동 가중 링크 접속 프로세스(automated weighted linking process: 중의성 해소)를 제공하기 위해 토픽 모델링 방식을 채용할 수 있다. 자동 가중 링크 접속 프로세스를 지원하기 위해, 실시 예들은 종래의 LDA 토픽 모델링을 확장하여, 조건부 독립으로서 임의 개수의 구성 요소(2차 특징들)를 지원할 수 있는 멀티-구성 요소 LDA(MC-LDA)라고 지칭되는 신규한 토픽 모델링 방식을 구축한다. 모델링 방식의 실시 예들은, 또한, 트레이닝 동안 구성 요소들의 가중치들을 자동으로 학습하고, 중의성 해소와 관련된 추론(링크 접속)을 위해 그들을 채용한다. 중의성 해소를 위해 도입된 MC-LDA 방식은 중의성 해소를 증가시키기 위해 도입될 수 있었던 임의 추가적인 개수의 2차 특징들을 스케일링(scaling)할 수 있다.

[0068] 도 5는 상술한 도 4의 시스템(400)에 의해 채용된 멀티-구성요소 조건부 독립 LDA(MC-LDA) 토픽 컴퓨터 모델링 방식의 실시 예의 그래픽도이다. 도시된 실시 예에 있어서, 각 구성 요소 블럭은, 예시적으로, 도 5에 설명한 파라메타들로 개시되는 도 4의 MemEB(408)를 통해 실행되는 지식 베이스에 걸쳐 있는 각 2차 특징의 모델링을 나타낸다.

[0069] 도 6은 상술한 도 5에 채용된 MC-LDA 토픽 모델링에 대한 Gibbs 샘플링 수학식의 실시 예를 도시한 도면이다. 이 샘플링 방식의 실시 예는 자동화 방식 및 효율적인 방식으로 개별 구성 요소(2차 특징) 가중치를 트레이닝시키는데 있어서 도 4의 시스템(400)을 원조한다.

[0070] 도 7은, 예를 들어, 도 7에 설명된 파라메타들로 개시되는 도 4의 시스템(400)의 MemDB(408)를 통해 실행되는 도 5 및 6의 MC-LDA 토픽 모델에 있어서의 트레이닝 및 추론을 위한 확률론적 변분 추론 알고리즘(stochastic variational inference algorithm)의 컴퓨터 실행 구현의 실시 예를 나타낸 도면이다. 이 추론 방식의 실시 예는 입력으로서 (관심 문서로부터 추출된) 모든 2차 특징들을 취하고, 출력으로서 가중 토픽을 제공함에 의해, 토픽 추론으로서 링크 접속/중의성 해소 프로세스를 모델링하는데 쉽게 적용된다. 이러한 가중 토픽은 저장된 특징 지식 베이스 엔트리들에 대한 유사성 스코어를 계산하는데 이용될 수 있다.

[0071] 도 8은 MC-LDA 토픽 모델에 대한 샘플 토픽을 도시한 테이블이다. 도 8은, 실시 예에 따라, 예를 들어, 도 4의 시스템(400)의 MemDB(408)를 통해 실행되는, 모델의 각 구성 요소에 대한 토픽 스코어링 표면 형태(topic scoring surface forms)를 디스플레이한다.

[0072] 예시 #1은, 비정형 텍스트에 있어서의 특징의 중의성을 해소하는 방법(100)의 애플리케이션으로서, 관심 특징(1차 특징)은 풋볼 선수인 John Doe이며, 사용자는 John Doe를 인용하는 뉴스를 모니터링하기 원한다. 일 실시 예에 따라, John Doe를 언급하는 문서 입력(102)이 네트워크내로 업로드될 수 있다. 문서 입력(102)의 특징이 추출되어 MemDB(408)내로 포함되어, 중의성 해소되고 1차 특징(John Doe)에 연계된 2차 특징들의 클러스터에 링크 접속되고 유사한 특징들의 기존의 클러스터와 비교된다. 방법(100)은, 예를 들어, John Doe, 엔지니어; John Doe, 선생; 및 John Doe, 풋볼 선수와 같은 John Doe에 대한 모든 관련된 2차 특징을 포함하는 서로 다른 특징 ID 및 특징 ID의 연계 클러스터를 출력할 수 있다. 예를 들어, 별명 및 약식 이름과 같은 유사한 2차 특징들을 가진 다른 1차 특징들이 고려될 수 있다. 그 다음, John Doe 풋볼 선수와 동일 팀으로부터 동일한 나이 및 경력을 가진 "JD" 풋볼 선수가 동일한 1차 특징으로 고려될 수 있다. 그러므로, John Doe, 풋볼 선수와 관련된 모든 문서가 쉽게 액세스될 수 있다.

[0073] 예시 #2는 비정형 문서에 있어서의 특징들의 중의성을 해소하는 방법(100)의 애플리케이션으로서, 1차 특징이 화상일 수 있다. 일 실시 예에 따르면, 방법(100)은 특징이 다른 것들 중에서도 에지 및 형상과 같은 일반적인 속성일 수 있거나, 다른 것들 중에서도 탱크, 사람, 시계와 같은 특정 속성일 수 있는 특징 추출(104)을 포함할 수 있다. 예를 들어, 화상이 (예를 들어, 정사각형, 사람 또는 차량과 같은) 특정 형상과 같은 2차 특징을 가질 수 있고, 2차 특징이 추출되어 유사한 2차 특징을 가지는 모든 다른 화상들 중에서 매칭이 발견될 수 있는 MemDB(408)에 포함될 수 있는 새로운 화상이 입력될 수 있다. 본 실시 예에 따르면, 단지 텍스트와 같은 화상을 포함하는 특징이 특징으로서 포함되지 않을 수 있다.

[0074] 예시 # 3은, 비정형 문서에 있어서의 특징들의 중의성을 해소하는 방법(100)의 애플리케이션으로서, 1차 특징이 이벤트일 수 있다. 일 실시 예에 따르면, 조회가 이루어지면, 방법(100)은 다른 것들 중에서 지진, 화재 또는 전염병 발발과 같은 이벤트와 연계된 결과를 사용자가 수신할 수 있게 한다. 방법(100)은 특징 추출(104)과 특징들의 특징 중의성 해소(108)를 실행하여 이벤트와 연계된 특징을 발견하고 중의성 해소된 특징들의 특징 ID를 제공할 수 있다(110).

[0075] 예시 #4 는 발생할 수 있는 하나 이상의 이벤트의 예측이 이루어질 수 있는 방법(100)의 실시 예이다. 일 실시 예에 따르면, 사용자는 동작전에 관심 이벤트 및 특징을 사전에 나타낼 수 있고, 그러므로 관심 이벤트와 연계된 서로 다른 특징들간의 링크가 사전에 수립될 수 있다. 연계된 특징들이 높은 발생 수로 네트워크에 나타나고 있으면, 방법(100)은 연계된 특징들의 증가된 발생 수에 기초하여 관심 이벤트의 발생을 예측할 수 있다. 임박한 이벤트가 검출되면, 사용자에게 경고가 전송될 수 있다. 예를 들어, 태국으로부터의 위생국(health department)에 근무하는 사용자는 댕기열 전염병 발발에 대한 경고를 수신하도록 선택할 수 있다. 예를 들어, 소셜 네트워크로부터의 다른 사용자들(406)들이 댕기열의 증상 또는 입원(inclusion into a hospital)을 포함하는 코멘트를 업로딩함에 따라, 방법(100)은 소셜 네트워크로부터 모든 관련된 코멘트들의 중의성을 해소할 수 있고, 관련된 정보를 포함하는 사용자(406)들의 수를 고려하여, 뎅기열 전염병 발발이 일어날 수 있음을 예측하여 위생국에 경고할 수 있다. 그러므로, 위생국 근무자는 추가적인 증거를 가질 수 있고, 전염병이 확산되지 않도록 감염된 커뮤니티에 대해 추가적인 행동이 취해질 수 있다.

[0076] 예시 #5는 비정형 텍스트에 있어서의 특징들의 중의성을 해소하는 방법(100)의 애플리케이션으로서, 1차 특징이 지리적 소재명일 수 있다. 실시 예에 따르면, 방법(100)은 도시 명의 중의성을 해소하는데 채용될 수 있으며, 거기에서는 다른 스코어링 가중치들이 중의성 해소 서브 모듈내의 2차 특징들과 연계될 수 있다. 예를 들어, 방법(100)은 프랑스, 파리로부터 텍사스, 파리의 중의성을 해소하는데 채용될 수 있다.

[0077] 예시 #6은 비정형 텍스트에 있어서의 특징들의 중의성을 해소하는 방법(100)의 애플리케이션으로서, 1차 특징이 다른 것들 중에서 사람, 이벤트 또는 회사와 연계된 센티먼트일 수 있고, 센티먼트들이 다른 것들 중에서도 소셜 네트워크를 포함한 임의의 적당한 소스로부터 피딩될 수 있는 사람, 이벤트 또는 회사에 대한 포지티브 또는 네거티브 코멘트일 수 있다. 일 실시 예에 따르면, 방법(100)은 대중들이 받아들일 수 있는 것을 회사가 알아채는데 채용될 수 있다.

[0078] 예시 #7은 방법(100)의 실시 예로서, 방법(100)은 특징의 신뢰 스코어를 증가시키기 위해 인간의 확인을 포함할 수 있다. 일 실시 예에 따르면, 링크 OTF 프로세스(300)(도 4)는 사용자에 의해 보조될 수 있으며, 사용자는 중의성 해소된 특징이 올바르게 중의성 해소되었는지를 나타낼 수 있고, 사용자는 2개의 다른 클러스터가 한가지

인지를 나타낼 수 있다(이것은, (모든 특징 및 토픽 동시 발생 정보를 고려하여) 사용자가 알고 있는 2개의 다른 1차 특징들으로서 방법(100)이 나타내고 있는 것이 동일할 수 있음을 의미함). 그러므로, 그 클러스터와 연계된 신뢰 스코어는 보다 높을 수 있으며, 따라서, 특징이 정확하게 중의성 해소될 확률이 보다 높아질 수 있다.

[0079] 예시 #8은 중의성 해소 프로세스(200)와 링크 OTF 프로세스(300)를 이용한 방법(100)의 실시 예이다. 본 예시에 있어서, 링크 접속 알고리즘 애플리케이션(304)에 이용되는 링크 접속 알고리즘은 1000ms의 기간내에 0.85보다 높은 신뢰 스코어를 제공하도록 구성된다.

[0080] 예시 #9는 중의성 해소 프로세스(200)와 링크 OTF 프로세스(300)를 이용한 방법(100)의 실시 예이다. 본 예시에 있어서, 링크 접속 알고리즘 애플리케이션(304)에 이용되는 링크 접속 알고리즘은 300ms를 초과하지 않는 기간내에 0.80보다 높은 신뢰 스코어를 제공하도록 구성된다. 본 예시에서 이용되는 알고리즘은 예시 #8에서 이용된 알고리즘에 비해 짧은 시 기간에 응답을 제공하지만, 통상적으로 보다 낮은 신뢰 스코어를 리턴한다.

[0081] 예시 #10은 중의성 해소 프로세스(200)와 링크 OTF 프로세스(300)를 이용한 방법(100)의 실시 예이다. 본 예시에 있어서, 링크 접속 알고리즘 애플리케이션(304)에 이용되는 링크 접속 알고리즘은 통상적으로 3000ms를 초과하는 기간내에 0.90보다 높은 신뢰 스코어를 제공하도록 구성된다. 본 예시에서 이용되는 알고리즘은 예시 #8에서 이용된 알고리즘에 의해 리턴된 것 보다 통상적으로 더 큰 신뢰 스코어를 가진 응답을 제공하지만, 통상적으로 아주 긴 시 기간을 요구한다.

[0082] 예시 #11은 다수의 소스로부터 문서의 대형 코퍼스에 대해 이-디스커버리(e-discovery)를 실행하기 위해 비정형 문서에 있어서의 특징들의 중의성을 해소하는 방법(100)의 예시이다. 다수의 소스로부터 문서의 대형 코퍼스가 주어질 경우, 이들 문서에 있어서의 모든 특징들의 중의성을 해소하는데 방법(100)을 적용하게 되면, 코퍼스내의 모든 특징들의 발견이 가능하게 된다. 발견된 특징들의 콜렉션은 관련된 특징들의 발견 및 특징과 관련된 모든 문서들을 발견하는데 추가로 이용될 수 있다.

[0083] 상술한 방법 설명 및 프로세스 흐름도는 단지 예시적으로 제공된 것으로 여러 실시 예의 단계들이 안출된 순서로 실행되어야만 함을 요구하거나 암시하는 것은 아니다. 당업자라면 알겠지만, 상술한 실시 예들에 있어서의 단계들은 임의 순서로 실행될 수 있다. "그 다음", "다음"들과 같은 단어는 단계들의 순서를 제한하고자 하는게 아니며, 이들 단어들은 단지 그 방법들의 설명을 독자에게 안내하는데 이용된다. 프로세스 흐름도가 동작들을 순차적인 프로세스로서 설명하였지만, 그 동작들 중 많은 동작들이 나란히 또는 동시에 실행될 수 있다. 또한, 동작들의 순서는 재배열될 수 있다. 프로세스는 방법, 기능, 절차, 서브루틴, 서브프로그램 등에 대응할 수 있다. 프로세스가 소정 기능에 대응할 경우, 그의 종료는 콜링 기능(calling function) 또는 주 기능에 대한 그 기능의 리턴에 대응할 수 있다.

[0084] 본 명세서에 개시된 실시 예들과 관련하여 설명된 여러 예시적인 논리 블럭, 모듈, 회로 및 알고리즘 단계들은 전자 하드웨어, 컴퓨터 소프트웨어 또는 그들의 결합으로서 구현될 수 있다. 하드웨어와 소프트웨어의 이러한 호환성을 명확하게 설명하기 위해, 여러 예시적인 구성 요소, 블럭들, 모듈들, 회로들 및 단계들이 그들의 기능성 견지에서 전반적으로 상기에 설명되었다. 그들의 기능성이 하드웨어 또는 소프트웨어로서 구현되는지는 전체 시스템에 대해 부과된 고안 제약 및 특정 애플리케이션에 좌우된다. 당업자라면 특정 애플리케이션마다 여러 방식으로 설명된 기능성들을 구현할 수 있겠지만, 그러한 구현 결정이 본 발명의 범주를 벗어나는 것으로 해석되어서는 안될 것이다.

[0085] 컴퓨터 소프트웨어로 구현된 실시 예는 소프트웨어, 펌웨어, 미들웨어, 마이크로코드, 하드웨어 기술 언어 또는 그들의 임의 조합으로 구현될 수 있다. 코드 세그먼트 또는 기계-실행 가능 명령어는 절차, 기능, 서브프로그램, 프로그램, 루틴, 서브루틴, 모듈, 소프트웨어 패키지, 클래스(class), 또는 명령어, 데이터 구조 또는 프로그램 진술의 임의 조합을 나타낼 수 있다. 코드 세그먼트는 정보, 데이터, 인수(argument), 파라메타 또는 메모리 콘텐츠를 통과 및/또는 수신함에 의해 또 다른 코드 세그먼트 또는 하드웨어 회로에 결합될 수 있다. 정보, 인수, 파라메타, 데이터 등은 메모리 공유, 메시지 패싱(passing), 토큰 패싱, 네트워크 전송들을 포함하는 임의 적절한 수단을 통해 패싱되거나, 전달되거나, 전송될 수 있다.

[0086] 이들 시스템 및 방법들을 구현하는데 이용되는 실제 소프트웨어 코드 또는 전용 제어 하드웨어는 본 발명을 제한하는 것이 아니다. 따라서, 시스템 및 방법의 동작 및 작용은, 소프트웨어 및 제어 하드웨어가 본 명세서에서의 설명에 기초하여 시스템 및 방법을 구현하도록 고안될 수 있다고 알고 있는 특정 소프트웨어 코드에 대한 참조없이 설명되었다.

[0087] 소프트웨어로 구현되었을 경우, 그 기능들은 비-일시적 컴퓨터 독출 가능 또는 프로세서-독출 가능 저장 매체상에 하나 이상의 명령 또는 코드로서 저장될 수 있다. 본 명세서에 개시된 방법 또는 알고리즘의 단계들은 컴퓨터 독출 가능 또는 프로세서 독출 가능 저장 매체상에 상주할 수 있는 프로세서-실행 가능 소프트웨어 모듈로 구현될 수 있다. 비-일시적 컴퓨터 독출 가능 또는 프로세서-독출 가능 매체는 컴퓨터 프로그램을 이곳 저곳으로 전달할 수 있는 컴퓨터 저장 매체 또는 유형(tangible) 저장 매체를 포함한다. 비-일시적 프로세서-독출 가능 저장 매체는 컴퓨터에 의해 액세스될 수 있는 임의 이용 가능 매체일 수 있다. 예를 들어, 제한을 위한 것은 아니지만, 그러한 비-일시적 프로세서-독출 가능 매체는 RAM, ROM, EEPROM, CD-ROM 또는 다른 광학 디스크 저장, 자기 디스크 저장 또는 다른 자기 저장 디바이스, 또는 명령어 또는 원하는 프로그램 코드를 데이터 구조 형태로 저장하는데 이용되고 컴퓨터 또는 프로세서에 의해 액세스될 수 있는 임의 다른 유형의 저장 매체를 구비할 수 있다. 본 명세서에서 이용된 디스크(disk and disc)는 컴팩트 디스크(CD), 레이저 디스크, 광학 디스크, 디지털 다기능 디스크(DVD), 플로피 디스크 및 블루-레이(Blu-ray) 디스크(디스크(disk)들은 통상 자기적으로 데이터를 재생하지만, 디스크(disc)는 레이저로 광학적으로 데이터를 재생한다)를 포함한다. 상술한 조합들은 컴퓨터-독출 가능 매체의 범주내에 포함되어야 한다. 또한, 방법 또는 알고리즘의 동작들은 컴퓨터 프로그램 제품내에 합체될 수 있는 비-일시적 프로세서-독출 가능 매체 및/또는 컴퓨터-독출 가능 매체상에 코드들 및/또는 명령어들의 하나 또는 임의 조합 또는 세트로서 상주할 수 있다.

[0088] 본 기술의 여러 구성 요소들은 분산형 네트워크 또는 인터넷의 원거리 부분 또는 전용 보안, 보안 해제 및/또는 암호화 시스템내에 배치될 수 있음을 알 것이다. 따라서, 시스템의 구성 요소들은 하나 이상의 디바이스내에 조합될 수 있거나, 원거리 통신 네트워크와 같은 분산형 네트워크의 특정 노드상에 함께 배치될 수 있음을 알 것이다. 본 명세서의 설명으로부터 및 계산적 효율 때문에 알겠지만, 시스템의 구성 요소들은 시스템의 동작에 영향을 주지 않고 분산형 네트워크내의 임의 위치에 배열될 수 있다. 더욱이, 그 구성 요소들은 전용 기계내에 내장될 수 있다.

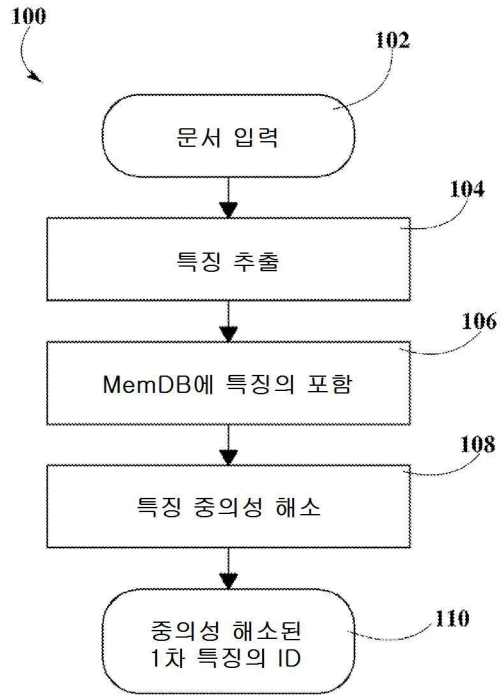
[0089] 또한, 요소들을 접속시키는 여러 링크들은 유선 또는 무선 링크이거나, 그들의 임의 조합 또는 접속된 요소들에/들로부터 데이터를 공급 및/또는 통신할 수 있는 임의의 다른 알려지거나 나중에 개발된 요소(들)일 수 있음을 알 수 있을 것이다. 본 명세서에서 이용된 용어인 모듈은 임의의 알려지거나 나중에 개발된 하드웨어, 소프트웨어, 펌웨어 또는 그 요소와 연계된 기능성을 실행할 수 있는 그들의 임의 조합을 지칭할 수 있다. 본 명세서에서 이용된 용어, 판정, 계산 및 비교와, 그들의 변형은 교호적으로 이용되고, 임의 유형의 방법론, 프로세스, 수학적 작용 또는 기술을 포함한다.

[0090] 개시된 실시 예의 상술한 설명은 당업자가 본 발명을 제조 및 이용할 수 있도록 제공된다. 이들 실시 예들의 여러 수정은 당업자에게는 아주 명백한 것이며, 본 명세서에서 정의된 포괄적인 원리는 본 발명의 사상 및 범주를 벗어나지 않고서 다른 실시 예에 적용될 수 있다. 따라서, 본 발명은 본 명세서에서 설명된 실시 예에 국한되는 것이 아니라, 이하의 청구범위 및 본 명세서에서 개시된 원리 및 신규한 특징과 일치하는 가장 넓은 범주를 부여받는다.

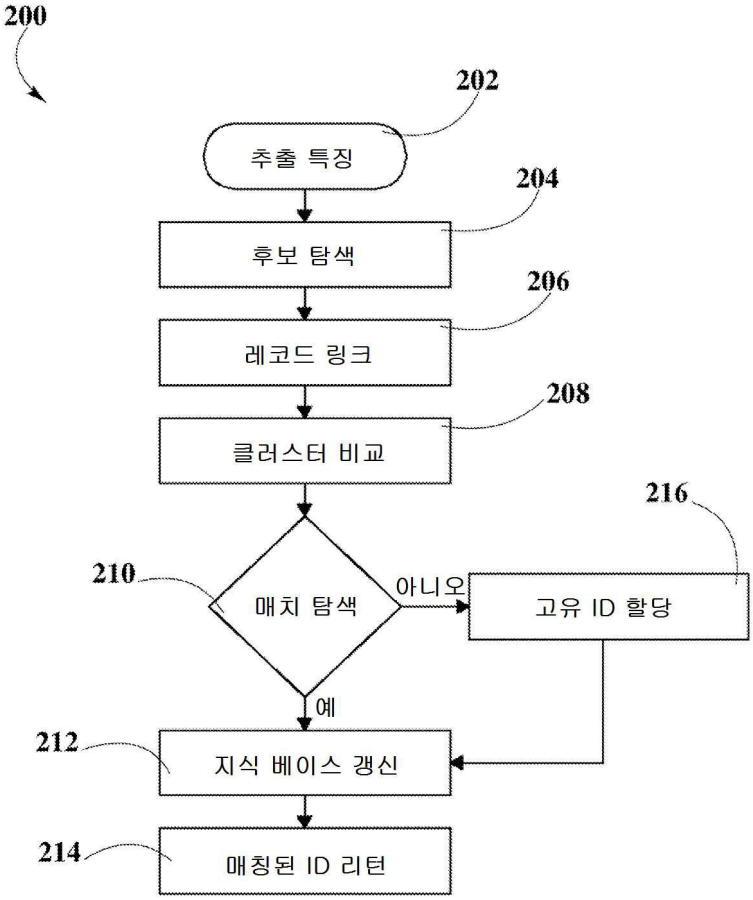
[0091] 상술한 실시 예들은 예시적인 것이다. 당업자라면 수많은 대안적인 구성 요소 및 실시 예들이 본 명세서에서 설명된 특정 예시를 대신하고 본 발명의 범주내에 속할 수 있음을 알 것이다.

## 도면

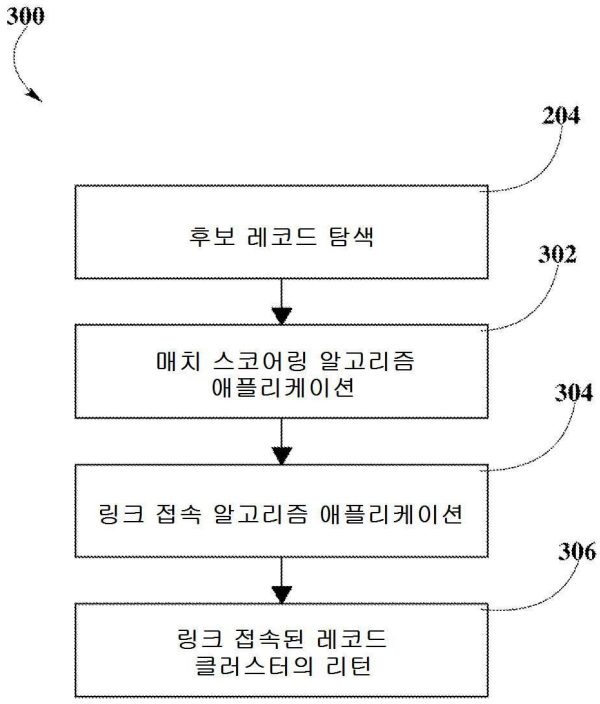
### 도면1


100
102
문서 입력
104
특징 추출
106
MemDB에 특징의 포함
108
특징 중의성 해소
110
중의성 해소된
1차 특징의 ID

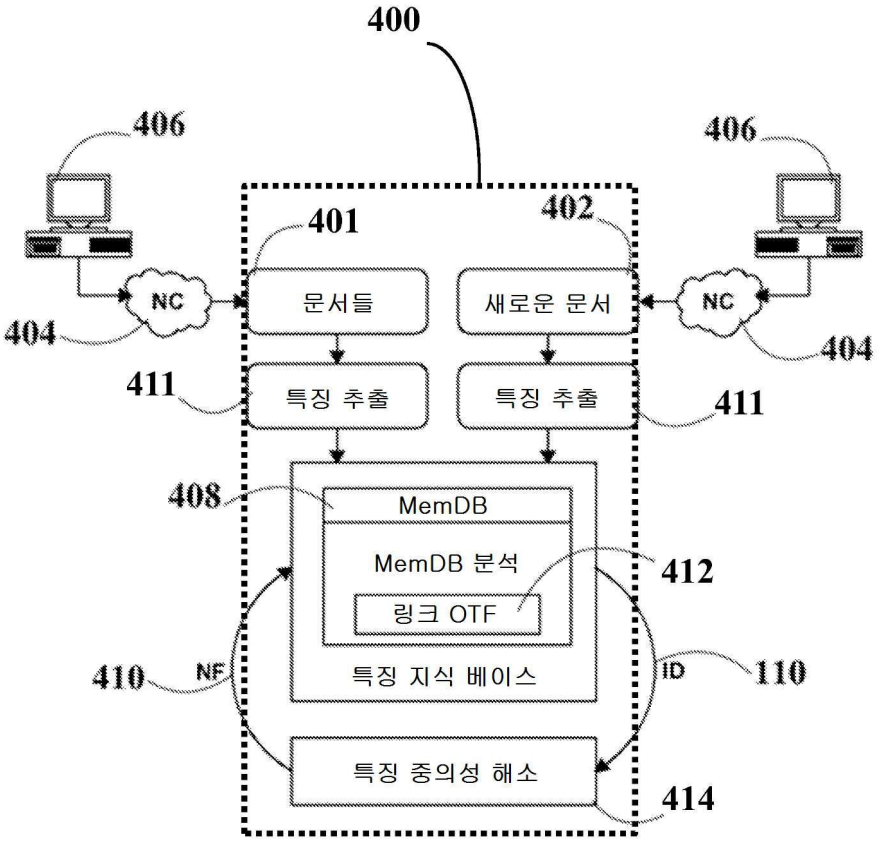
도면2


200
202
추출 특징
204
후보 탐색
206
레코드 링크
208
클러스터 비교
216
210
매치 탐색
아니오
고유 ID 할당
예
212
지식 베이스 갱신
214
매칭된 ID 리턴

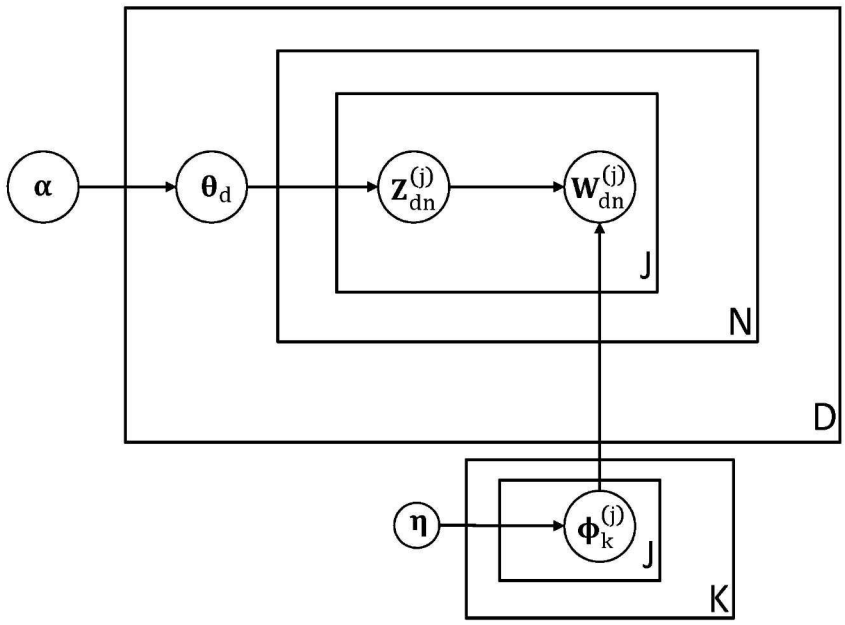
도면3


300
204
후보 레코드 탐색
302
매치 스코어링 알고리즘
애플리케이션
304
링크 접속 알고리즘 애플리케이션
306
링크 접속된 레코드
클러스터의 리턴

도면4


400
406
406
401
402
NC
문서들
새로운 문서
NC
404
404
411
특징 추출
특징 추출
411
408
MemDB
MemDB 분석
412
링크 OTF
410
NF
특징 지식 베이스
ID
110
특징 중의성 해소
414

도면5




$J$: 멀티-문서 구성 요소들의 수
$V^{(j)}$: j번째 구성 요소의 어휘에 있어서의 용어의 개수
$D$: 문서들의 수
$N$: 문서의 구성 요소에 있어서의 토큰의 수(실제적으로는 j와 d에 의존함)
$K$: 토픽들의 수
$\alpha$: 배합비(mixing proportion)에 대한 하이퍼파라메타 (대칭일 경우 K-벡터 또는 스칼라)
$\eta^{(j)}$: 혼합 구성 요소들에 대한 하이퍼파라메타 (대칭일 경우 K-벡터 또는 스칼라)
$\theta_d$: 문서 d에 대한 토픽 혼합비(topic mixture proportion)
$\phi_k^{(j)}$: k번째 토픽의 j번째 구성 요소에 대한 혼합 구성 요소
$z_{d,n}^{(j)}$: 문서 d의 j번째 구성 요소에 있어서의 n번째 단어에 대한 토픽을 선택하는 혼합 표시자
$w_{d,n}^{(j)}$: 문서 d의 j번째 구성 요소에 있어서의 n번째 단어에 대한 용어 표시자

도면6

$$p(z_i^{(j)} = k|w_i^{(j)} = t, z_{\neg i}^{(j)}, z^{(-j)}, w_{\neg i}^{(j)}, w^{(-j)}, \alpha, \eta) \propto \frac{n_{d,\neg i}^{(j,k)} + \alpha}{\sum_{k=1}^{K}(n_{d,\neg i}^{(j,k)} + \alpha)} \cdot \frac{n_{k,\neg i}^{(j,t)} + \beta^{(j)}}{\sum_{t'=1}^{V_j}(n_{k,\neg i}^{(j,t')} + \beta^{(j)})}$$

$$\mathrm{E}(\theta_{d,k}|z^{(j)}, w^{(j)}, \beta) = \frac{n_k^{(j,t)} + \beta^{(j)}}{\sum_{t'=1}^{V_j}(n_k^{(j,t')} + \beta^{(j)})}$$

$$\mathrm{E}(\theta_{d,k}|z^{(j)}, \ldots, z^{(J)}, \alpha) = \frac{\sum_{j=1}^{J}(n_d^{(j,k)} + \alpha)}{\sum_{k=1}^{K}\sum_{j=1}^{J}(n_d^{(j,k)} + \alpha)}$$

$J$: 멀티-문서 구성 요소들의 수

$V^{(j)}$: j번째 구성 요소의 어휘에 있어서의 용어의 개수

$K$: 토픽들의 수

$\alpha$: 배합비(mixing proportion)에 대한 하이퍼파라메타 (대칭일 경우 K-벡터 또는 스칼라)

$\eta$: 혼합 구성 요소들에 대한 하이퍼파라메타 (실제적으로, 각각 길이 V(j)의 J 벡터들)

$\theta_d$: 문서 d에 대한 토픽 혼합비(topic mixture proportion)

$\phi_k^{(j)}$: k번째 토픽의 j번째 구성 요소에 대한 혼합 구성 요소

$z_{d,n}^{(j)}$: 문서 d의 j번째 구성 요소에 있어서의 n번째 단어에 대한 토픽을 선택하는 혼합 표시자

$w_{d,n}^{(j)}$: 문서 d의 j번째 구성 요소에 있어서의 n번째 단어에 대한 용어 표시자

$n_{d,(\cdot)}^{(j,k)}$: 토픽 k에 할당된 문서 d의 구성 요소 j에 있어서의 용어의 수

$n_{k,(\cdot)}^{(j,t)}$: 코퍼스의 구성 요소 j에 있어서의 용어 t가 토픽 k에 할당되었던 횟수

도면7

알고리즘 1 MC-TM에 대한 확률론적 변분 추론

1: Set $\epsilon = 0.0001$
2: Define $\rho_t := (\tau_0 + t)^{-\kappa}$
3: Initialize $\boldsymbol{\lambda}$ randomly.
4: **for** $t = 0$ to $\infty$ **do**
*E-Step:*
5: Initialize $\gamma_{tk} = 1$. (The constant 1 is arbitrary.)
6: **repeat**
7: Set $\phi_{tmk}^{(j)} \propto e^{\mathbf{E}_q[\log \theta_{tk}] + \mathbf{E}_q[\log \beta_{kw}^{(j)}]}$
8: Set $\gamma_{tk} = \alpha + \sum_j \sum_{w=1}^{V} w_{tm}^{(j)} \phi_{tmk}^{(j)} n_{tw}^{(j)}$
9: **until** $\frac{1}{K} \sum_k |\Delta \gamma_{tk}| < \epsilon$
*M-Step:*
10: Compute $\lambda_{kw}^{(j)} = \eta^{(j)} + D \cdot n_{tw}^{(j)} \cdot \phi_{tmk}^{(j)}$
11: Set $\lambda = (1 - \rho_t)\lambda + \rho_t \hat{\lambda}$
12: **end-for**

$\alpha$: 문서-토픽 배합비(mixing proportion)에 대한 하이퍼파라메타
$\eta$: 토픽-용어 혼합비율에 대한 하이퍼파라메타
$\phi_{d,w}$: 문서 d에 있어서 토픽 할당의 다항 분산에 대한 파라메타
$\gamma_d$: 문서 d에 대한 문서-토픽 비례(document-topic proportion)에 대한 디리클레(Dirichlet) 분포에 대한 파라메타
$\lambda_k$: 토픽 k에 대한 토픽-용어 비율(topic-term proportion)에 대한 디지클레 분포에 대한 파라메타

도면8

| 토큰 | album, music, song, released, songs, single, albums, love, tour, first ... |
|---|---|
| 키 구문 | album, song, music, rock, debut_album, studio_album, solo_album, edit, music_video, guitar ... |
| 조직 | mtv, beatles, rolling_stone, youtube, emi, warner_bros, rca, official_charts_company, rovi_corporation, atlantic_records ... |
| 사람 | britney_spears, bruce_springsteen, robin_thicke, bob_dylan, mariah_carey, john_lennon, jennifer_lopez, stevie_nicks, melissa_etheridge, dylan ... |
| 위치 | us, united_states, uk, united_kingdom, united_states_of_america, u_s, australia, glastonbury ... |