



(19) 대한민국특허청(KR)

(12) 등록특허공보(B1)

(45) 공고일자 2015년12월09일

(11) 등록번호 10-1575683

(24) 등록일자 2015년12월02일

(51) 국제특허분류(Int. Cl.)

G06F 17/27 (2006.01) G06F 17/21 (2006.01)

G06F 17/30 (2006.01)

(21) 출원번호 10-2014-0187026

(22) 출원일자 2014년12월23일

심사청구일자 2014년12월23일

(56) 선행기술조사문헌

US20140040281 A1

JP2011134230 A

JP2007011571 A

US20130275527 A1

(73) 특허권자

고려대학교 산학협력단

서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)

포항공과대학교 산학협력단

경상북도 포항시 남구 청암로 77 (지곡동)

(72) 발명자

이상근

서울 강남구 선릉로 221, 205동 1104호 (도곡동, 도곡렉슬아파트)

알람 히즈블

서울특별시 마포구 대흥로30길 40, 2층 (염리동)

류우종

서울 성북구 한천로 509, 106동 502호 (석관동, 코오롱아파트)

(74) 대리인

김동용, 김홍석, 이강녕

전체 청구항 수 : 총 6 항

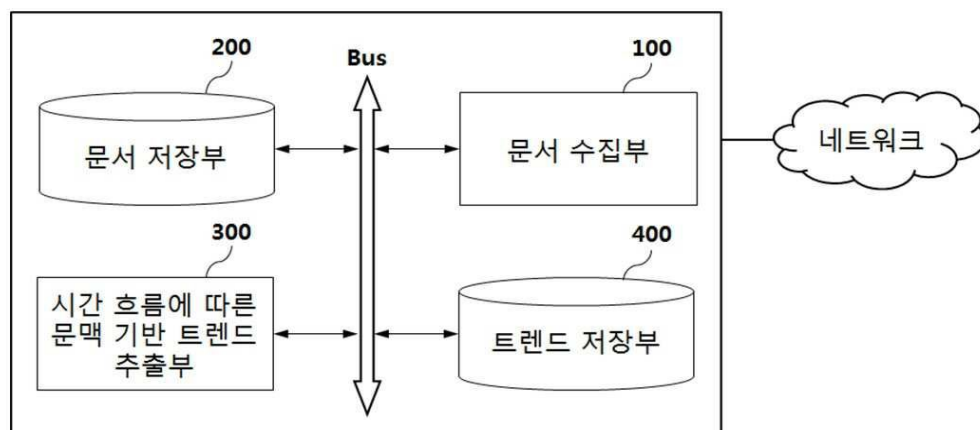
심사관 : 김병균

(54) 발명의 명칭 시간 흐름에 따른 문맥 기반 트렌드 분석 방법

(57) 요약

시간 흐름에 따른 문맥 기반 트렌드를 분석하는 방법이 개시된다. 상기 문맥 기반 트렌드를 분석하는 방법은 확률 분포를 계산할 수 있는 장치를 이용하여 문서 집합으로부터 시간 흐름에 따른 문맥 기반 트렌드를 분석하는 방법으로서, 각 토픽에 대해, 단어를 선택하기 위해 사용될 어휘 분포와 해시태그(hashtag)를 선택하기 위해 사용될 해시태그 분포를 추출하는 단계, 상기 문서 집합에 포함된 각 문서에 대해, 토픽 분포를 추출하는 단계, 상기 어휘 분포, 상기 해시태그 분포, 및 상기 토픽 분포에 대해 통계적 추론을 수행하는 단계, 상기 문서 집합에 포함된 각 문서의 각 단어에 대해, 상기 토픽 분포로부터 토픽을 추출하고, 시간에 대한 베타 분포(beta distribution)로부터 시간을 추출하는 단계, 및 상기 문서 집합에 포함된 각 문서의 각 단어에 대해, 단어 또는 해시태그를 추출하는 단계를 포함한다.

대표도 - 도1



이 발명을 지원한 국가연구개발사업

과제고유번호 2012M3C4A7033344

부처명 미래창조과학부

연구관리전문기관 한국연구재단

연구사업명 차세대정보컴퓨팅기술개발사업

연구과제명 비정형 빅 데이터 분석 및 개인화 기술 개발

기 여 율 1/1

주관기관 포항공과대학교

연구기간 2012.07.01 ~ 2017.06.30

명세서

청구범위

청구항 1

확률 분포를 계산할 수 있는 장치를 이용하여, 문서 집합으로부터 시간 흐름에 따른 문맥 기반 트렌드를 분석하는 방법에 있어서,

- (a) 각 토픽에 대해, 단어를 선택하기 위해 사용될 어휘 분포와 해시태그(hashtag)를 선택하기 위해 사용될 해시태그 분포를 추출하는 단계;
- (b) 상기 문서 집합에 포함된 각 문서에 대해, 토픽 분포를 추출하는 단계;
- (c) 상기 어휘 분포, 상기 해시태그 분포, 및 상기 토픽 분포에 대해 통계적 추론을 수행하는 단계;
- (d) 상기 문서 집합에 포함된 각 문서의 각 단어에 대해, 상기 토픽 분포로부터 토픽을 추출하고, 시간에 대한 베타 분포(beta distribution)로부터 시간을 추출하는 단계; 및
- (e) 상기 문서 집합에 포함된 각 문서의 각 단어에 대해, 단어 또는 해시태그를 추출하는 단계를 포함하는, 시간 흐름에 따른 문맥 기반 트렌드를 분석하는 방법.

청구항 2

제1항에 있어서,

상기 어휘 분포, 상기 해시태그 분포, 및 상기 토픽 분포는 디리클레 사전확률(Dirichlet prior)로부터 추출되는, 시간 흐름에 따른 문맥 기반 트렌드를 분석하는 방법.

청구항 3

제1항에 있어서,

상기 통계적 추론에는 깁스 샘플링(Gibbs sampling) 기법이 사용되는, 시간 흐름에 따른 문맥 기반 트렌드를 분석하는 방법.

청구항 4

제1항에 있어서,

상기 (e) 단계는,

상기 어휘 분포로부터 단어를 추출하는 단계; 및

추출된 단어가 해시태그인 경우, 상기 해시태그 분포로부터 해시태그를 추출하는 단계를 포함하는, 시간 흐름에 따른 문맥 기반 트렌드를 분석하는 방법.

청구항 5

제1항에 있어서,

상기 문서 집합은 복수의 트윗(tweet)을 포함하는, 시간 흐름에 따른 문맥 기반 트렌드를 분석하는 방법.

청구항 6

제1항에 있어서,

상기 (a) 단계 이전에 네트워크를 통하여 상기 문서 집합에 포함되는 적어도 하나의 문서를 수집하는 단계를 더 포함하는, 시간 흐름에 따른 문맥 기반 트렌드를 분석하는 방법.

발명의 설명

기술 분야

[0001] 본 발명의 개념에 따른 실시 예는 시간 흐름에 따른 문맥 기반 트렌드 분석 장치 및 방법에 관한 것으로, 특히 소셜 미디어 상에서 사용자가 기술한 문서 집합으로부터 시간 흐름에 따른 문맥 기반 트렌드를 분석하는 트렌드 분석 장치 및 방법에 관한 것이다.

배경 기술

[0002] 트위터, 페이스북, 마이크로 블로그와 같은 소셜 미디어의 사용자가 증가함에 따라, 사용자들은 소셜 미디어에 자신의 관심사, 사회적 이슈 등에 대한 의견을 기술하고 있다. 이러한 의견들 중 많은 사용자들이 기술하는 대상, 즉 많은 사용자들이 공통적으로 관심을 가지는 대상을 트렌드(trend)로써 정의한다. 따라서 소셜 미디어 상에 사용자들이 기술한 문서 집합을 이용해 트렌드를 분석하기 위한 많은 연구가 진행 중이다.

[0003] 상술한 종래기술에서는 트렌드를 분석할 때 시간적인 속성을 고려해 시간 흐름에 따라 트렌드 및 트렌드의 변화를 분석한다. 예를 들어, 소셜 미디어로부터 세월호 참사에 대한 트렌드를 추출한 경우, 사용자들의 관심이 언제, 어떻게 증가하는지 또는 감소하는지 등의 시간 흐름에 따른 트렌드의 변화를 분석할 수 있다. 하지만 트렌드를 분석함에 있어 상기 트렌드에 어떤 감정, 기관, 인물 등이 연관되어 있는지, 즉 트렌드의 문맥을 분석할 수 없다는 한계가 있다.

발명의 내용

해결하려는 과제

[0004] 본 발명이 이루고자 하는 기술적인 과제는 문서 집합으로부터 트렌드를 분석함에 있어 트렌드와 연관된 문맥을 추출하고, 추출된 문맥을 고려해 시간 흐름에 따른 트렌드의 변화를 분석할 수 있는 트렌드 분석 장치 및 방법을 제공하는 것이다.

과제의 해결 수단

[0005] 본 발명의 실시 예에 따른 문맥 기반 트렌드를 분석하는 방법은 확률 분포를 계산할 수 있는 장치를 이용하여 문서 집합으로부터 시간 흐름에 따른 문맥 기반 트렌드를 분석하는 방법으로서, 각 토픽에 대해, 단어를 선택하기 위해 사용될 어휘 분포와 해시태그를 선택하기 위해 사용될 해시태그 분포를 추출하는 단계, 상기 문서 집합에 포함된 각 문서에 대해, 토픽 분포를 추출하는 단계, 상기 어휘 분포, 상기 해시태그 분포, 및 상기 토픽 분포에 대해 통계적 추론을 수행하는 단계, 상기 문서 집합에 포함된 각 문서의 각 단어에 대해, 상기 토픽 분포로부터 토픽을 추출하고, 시간에 대한 베타 분포로부터 시간을 추출하는 단계, 및 상기 문서 집합에 포함된 각 문서의 각 단어에 대해, 단어 또는 해시태그를 추출하는 단계를 포함한다.

발명의 효과

[0006] 본 발명의 실시 예에 따른 트렌드 분석 장치 및 방법은 사용자들의 트렌드에 대한 관심이 시간에 따라 증가하는지 또는 감소하는지 여부를 모니터링할 수 있는 효과가 있다.

[0007] 또한, 트렌드와 연관된 감정, 기관, 인물 등과 같은 트렌드의 문맥을 분석할 수 있는 효과가 있다.

도면의 간단한 설명

[0008] 본 발명의 상세한 설명에서 인용되는 도면을 보다 충분히 이해하기 위하여 각 도면의 상세한 설명이 제공된다.

도 1은 본 발명의 일 실시 예에 의한 트렌드 분석 장치의 기능 블록도이다.

도 2는 도 1에 도시된 트렌드 분석 장치를 이용한 트렌드 분석 방법에서 사용되는 표기를 도시한다.

도 3은 도 1에 도시된 트렌드 분석 장치를 이용한 시간 흐름에 따른 문맥 기반 트렌드 분석 방법을 설명하기 위한 도면이다.

도 4는 도 1에 도시된 트렌드 추출 장치를 이용한 시간 흐름에 따른 문맥 기반 트렌드 추출 방법의 개념을 도시하고 있다.

도 5는 도 1에 도시된 트렌드 분석 장치를 이용한 시간 흐름에 따른 문맥 기반 트렌드 추출 방법을 설명하기 위

한 흐름도이다.

도 6은 도 1에 도시된 트렌드 분석 장치를 이용하여 분석된 트렌드의 예를 도시한다.

발명을 실시하기 위한 구체적인 내용

- [0009] 본 명세서에 개시되어 있는 본 발명의 개념에 따른 실시 예들에 대해서 특정한 구조적 또는 기능적 설명은 단지 본 발명의 개념에 따른 실시 예들을 설명하기 위한 목적으로 예시된 것으로서, 본 발명의 개념에 따른 실시 예들은 다양한 형태들로 실시될 수 있으며 본 명세서에 설명된 실시 예들에 한정되지 않는다.
- [0010] 본 발명의 개념에 따른 실시 예들은 다양한 변경들을 가할 수 있고 여러 가지 형태들을 가질 수 있으므로 실시 예들을 도면에 예시하고 본 명세서에서 상세하게 설명하고자 한다. 그러나, 이는 본 발명의 개념에 따른 실시 예들을 특정한 개시 형태들에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물, 또는 대체물을 포함한다.
- [0011] 제1 또는 제2 등의 용어는 다양한 구성 요소들을 설명하는데 사용될 수 있지만, 상기 구성 요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성 요소를 다른 구성 요소로부터 구별하는 목적으로만, 예컨대 본 발명의 개념에 따른 권리 범위로부터 벗어나지 않은 채, 제1 구성 요소는 제2 구성 요소로 명명될 수 있고 유사하게 제2 구성 요소는 제1 구성 요소로도 명명될 수 있다.
- [0012] 어떤 구성 요소가 다른 구성 요소에 "연결되어" 있다거나 "접속되어" 있다고 언급된 때에는, 그 다른 구성 요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성 요소가 존재할 수도 있다고 이해되어야 할 것이다. 반면에, 어떤 구성 요소가 다른 구성 요소에 "직접 연결되어" 있다거나 "직접 접속되어" 있다고 언급된 때에는 중간에 다른 구성 요소가 존재하지 않는 것으로 이해되어야 할 것이다. 구성 요소들 간의 관계를 설명하는 다른 표현들, 즉 "~사이에"와 "바로 ~사이에" 또는 "~에 이웃하는"과 "~에 직접 이웃하는" 등도 마찬가지로 해석되어야 한다.
- [0013] 본 명세서에서 사용한 용어는 단지 특정한 실시 예를 설명하기 위해 사용된 것으로서, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "포함하다" 또는 "가지다" 등의 용어는 본 명세서에 기재된 특징, 숫자, 단계, 동작, 구성 요소, 부분품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성 요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.
- [0014] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가진다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 의미를 갖는 것으로 해석되어야 하며, 본 명세서에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.
- [0015] 이하, 본 명세서에 첨부된 도면들을 참조하여 본 발명의 실시 예들을 상세히 설명한다.
- [0016] 도 1은 본 발명의 일 실시 예에 의한 트렌드 분석 장치의 기능 블록도이다. 도 1을 참조하면, 시간 흐름에 따른 문맥에 기반하여 트렌드를 분석할 수 있는 트렌드 분석 장치(10)는 문서 수집부(100), 문서 저장부(200), 트렌드 추출부(300), 및 트렌드 저장부(400)를 포함한다. 또한, 트렌드 분석 장치(10)는 확률 분포를 계산할 수 있는 장치로 명칭될 수도 있다.
- [0017] 문서 수집부(100)는 인터넷 등의 네트워크를 통해 소셜 미디어 또는 소셜 서비스를 제공하는 서버로부터 하나 이상의 문서를 수집할 수 있다. 문서 수집부(100)에 의해 수집된 문서 집합은 문서 저장부(200)에 저장될 수 있다. 이때, 문서 저장부(200)에 저장된 문서 집합 또는 문서 집합의 일부는 기존에 구축된 것일 수도 있다. 이때, 상기 문서 집합은 복수의 트윗(tweet)을 포함할 수 있다.
- [0018] 트렌드 추출부(300)는 문서 저장부(200)에 저장된 문서 집합에서 시간 흐름에 따른 문맥 기반 트렌드를 추출할 수 있다. 구체적으로, 트렌드 추출부(300)는 문서 저장부(200)에 저장된 문서 집합으로부터 시간 흐름에 따른 토픽 및 시간 흐름에 따른 문맥을 추출한다. 즉, 트렌드 추출부(300)는 토픽, 시간, 문맥을 동시에 고려함으로써 시간 흐름에 따른 문맥 기반 트렌드를 추출할 수 있다.
- [0019] 트렌드 추출부(300)는 시간 흐름에 따른 문맥 기반 트렌드 추출부로 명칭될 수도 있다. 트렌드 추출부(300)에

의해 추출된 트렌드는 트렌드 저장부(400)에 저장된다.

- [0020] 토픽 모델을 이용해 소셜 미디어 상에서의 많은 사용자들이 공통적으로 관심을 가지는 트렌드를 분석하기 위한 연구가 있었다. 또한, 고정된 시간에서의 트렌드를 분석하는데에서 나아가 시간이라는 속성을 고려해 트렌드를 분석할 뿐만 아니라 시간 흐름에 따른 트렌드의 변화를 모니터링하기 위한 연구가 진행되고 있다.
- [0021] 하지만 상술한 종래기술들은 단순히 트렌드를 분석하는 데에 초점이 맞춰져 있고, 분석한 트렌드에 대해 사용자들이 어떠한 감정 또는 생각을 가지고 있는지, 어떤 기관, 단체 또는 인물이 연관되어 있는지에 대한 문맥은 고려하지 않는다.
- [0022] 트렌드나 시간에 따른 트렌드의 변화를 분석함과 동시에 트렌드에 대한 문맥을 고려할 수 있다면 유용할 것이다. 사용자들이 트렌드에 대해 가지는 감정 또는 생각, 트렌드와 연관된 기관, 단체 또는 인물까지 분석할 수 있다면 트렌드를 분석하는 관점을 그 자체에 한정하지 않고, 트렌드를 다양한 시각에서 분석할 수 있다. 예를 들어, 트렌드 분석 업체 및 트렌드에 초점을 맞추는 마케팅 업체의 경우, 트렌드의 문맥까지 고려한다면 다양한 사용자들을 위한 다양한 마케팅 전략을 세울수 있을 것이다.
- [0023] 본 발명에 따른 시간 흐름에 따른 문맥 기반 트렌드 분석 방법 및 장치는 소셜 미디어 상에 사용자들이 기술한 하나 이상의 문서 집합으로부터 시간 흐름에 따른 트렌드 분석 뿐만 아니라, 트렌드의 문맥까지 자동으로 분석할 수 있다는 장점을 가지고 있다.
- [0024] 문맥(context)은 해당 개념을 사용하는 분야에 따라 다양한 의미로서 정의할 수 있다. 하지만 본 발명에서 고려하는 문맥은 트렌드와 함께 자주 나타나는 해시태그들의 군집(cluster of hashtags)으로 정의될 수 있다.
- [0025] 트위터의 트윗은 해시태그(hashtag), 멘션(mention), 이미지(image), 링크(link) 등과 같은 다양한 메타데이터를 포함하고 있다. 그 중 해시태그는 중요한 메타데이터로써 다양한 역할을 하고 있다.
- [0026] 트윗은 140자로 제한된 길이의 문자를 통해서만 의견의 기술이 가능하다. 따라서 사용자들은 이러한 제한을 극복하기 위해 다양한 의미 및 유용한 정보를 함축하고 있는 해시태그를 사용한다.
- [0027] 예를 들어, 세월호 참사와 관련된 해시태그 중, "#PrayForSouthKorea"는 세월호 참사의 희생자를 추모하기 위한 온라인 커뮤니티를, "#SadStory"는 세월호 참사에 대한 사용자들의 감정을, "#RestInPeace"는 세월호 참사의 희생자들에 대한 사용자들의 소망을, "#YellowRibbon"은 세월호 참사 실종자들의 구조를 기원하는 사용자들의 운동을 의미할 수 있다. 이와 같이 해시태그는 한 단어로써 표현되지만 140자로 제한된 길이의 문자 제한을 극복하기 위해 사용자들의 감정, 소망 또는 기관, 단체, 인물 등 의미있는 정보를 함축하고 있다.
- [0028] 또한, 해시태그는 의미적으로 유사한 트윗들을 연결시킨다. 예를 들어, 공통의 관심사를 가진 사용자들은 공통의 해시태그를 사용함으로써, 이 해시태그는 두 개 이상의 서로 다른 트윗이 의미적으로 유사하다는 것을 암시적으로 나타낸다.
- [0029] 따라서 본 발명에서 고려하는 문맥은 트렌드와 연관된 감정, 기관, 단체, 인물 등에 대한 정보를 함축하고 있는 해시태그들의 군집으로 정의될 수 있다.
- [0030] 트윗은 트윗이 작성된 시간 정보를 포함하고 있다. 본 발명의 생성 과정(generative process)에서 시간은 토픽에 의해 생성된다고 가정하지만, 한 트윗내의 모든 단어(또는 어휘) 및 해시태그에는 트윗이 작성된 시간을 부여한다.
- [0031] 시간은 연속적인 속성을 가진다. 따라서 연속적인 속성을 가진 시간을 샘플링하기 위해 0~1 사이의 구간에서 정의되는 연속 확률 분포인 베타 분포(beta distribution)를 따른다.
- [0032] 도 2는 도 1에 도시된 트렌드 분석 장치를 이용한 트렌드 분석 방법에서 사용되는 표기를 도시한다.
- [0033] 도 1과 도 2를 참조하면, 각 문서는 d, 토픽은 z, 단어(또는 어휘)는 w, 해시태그는 c, 시간은 t로 표기되고, 일반적으로 개수 또는 횟수는 n으로 표기된다. 토픽, 단어, 해시태그, 시간 등에 대한 확률 분포는 각각 그리스 문자로 표기하고 있으며, 각 확률 분포의 디리클레 사전확률(Dirichlet prior)에도 해당 그리스 문자를 할당한다.
- [0034] 또한, 이후 기술에서 Dir()은 괄호 안의 인수를 기초로 디리클레 분포(dirichlet distribution)를 생성함을 의미하고, B()는 괄호 안의 인수를 기초로 베타 분포를 생성함을 의미한다.
- [0035] 도 3은 도 1에 도시된 트렌드 분석 장치를 이용한 시간 흐름에 따른 문맥 기반 트렌드 분석 방법을 설명하기 위

한 도면이다.

- [0036] 도 1 내지 도 3을 참조하면, 각 토픽에 대해, 단어를 선택하기 위해 사용될 어휘 분포(또는 단어 분포), 해시태그를 선택하기 위해 사용될 해시태그 분포를 추출한다. 본 발명에서 단어, 해시태그, 및 시간은 토픽에 의해 생성된다고 가정한다. 따라서, 트윗을 모델링하기 위한 단어와 해시태그는 두 단계를 통해 생성된다.
- [0037] 먼저, 각 문서의 각 단어에 대해, 토픽 분포로부터 토픽을 선택하고, 토픽의 어휘 분포 및 시간에 대한 베타 분포로부터 단어와 시간을 선택한다.
- [0038] 다음으로, 각 문서의 각 해시태그에 대해, 토픽 분포로부터 토픽을 선택하고, 토픽의 해시태그 분포 및 시간에 대한 베타 분포로부터 해시태그와 시간을 선택한다.
- [0039] 도 4는 도 1에 도시된 트렌드 추출 장치를 이용한 시간 흐름에 따른 문맥 기반 트렌드 추출 방법의 개념을 도시하고 있다.
- [0040] 도 1 내지 도 4를 참조하면, 도 4는 토픽(z), 단어(w), 해시태그(c), 시간(t) 사이의 관계를 표기하고 있다. 또한 각 확률 분포 및 변수를 구하는 순서를 표기하고 있다.
- [0041] 도 4의 우측을 보면, 단어의 다항 분포, 즉 어휘 분포(word distribution)는 디리클레 사전확률로부터 산출되고, 최종적으로 단어를 샘플링하는데 사용된다. 도 4의 좌측을 보면, 해시태그의 다항 분포, 즉 해시태그 분포(hashtag distribution)는 디리클레 사전확률로부터 산출되고, 최종적으로 해시태그를 샘플링하는데 사용된다.
- [0042] 도 5는 도 1에 도시된 트렌드 분석 장치를 이용한 시간 흐름에 따른 문맥 기반 트렌드 추출 방법을 설명하기 위한 흐름도이다. 도 5에 도시된 각 단계는 트렌드 분석 장치(10), 구체적으로는 트렌드 추출부(300)에 의해 수행될 수 있다.
- [0043] 도 1 내지 도 5를 참조하면, 단계 S110에서, 단어를 선택하기 위한 어휘 분포와 해시태그를 선택하기 위한 해시태그 분포가 추출된다. 즉, 토픽을 구성하는 어휘와 해시태그에 대해 디리클레 사전확률을 기초로 확률 분포가 구축된다.
- [0044] 단계 S120에서, 각 문서에 대해, 토픽의 다항분포, 즉 토픽 분포가 추출된다. 상기 토픽 분포는 토픽 분포에 대한 디리클레 사전확률로부터 추출될 수 있다. 이때, 단계 S120 이전 또는 단계 S110 이전에 네트워크를 통하여 상기 문서 집합에 포함되는 적어도 하나의 문서를 수집하는 단계가 더 포함될 수 있고, 이 단계는 문서 수집부(100)에 의해 수행될 수 있다.
- [0045] 단계 S130에서, 추출된 각 분포, 즉 어휘 분포, 해시태그 분포, 및 토픽 분포에 대해 통계적 추론을 수행한다. 이때, 통계적 추론에는 깁스 샘플링(Gibbs sampling) 기법이 사용될 수 있다. 아래의 수학적 식 1과 수학적 식 2는 각각 해시태그에 대한 샘플링 분포와 단어에 대한 샘플링 분포를 나타내고 있다.

수학적 식 1

$$P(z_{d,i} = z | \mathbf{z}', \mathbf{w}, \mathbf{t}, \mathbf{c}) \propto \frac{n_{w_{d,i}}^z + \beta}{n^z + V\beta} \times \frac{n_c^{zh} + \gamma}{n^{zh} + C\gamma} \times \frac{n_z^d + \alpha}{n^d + M\alpha} \times \frac{(1 - t_d)^{\lambda_{z1}-1} t_d^{\lambda_{z2}-1}}{B(\lambda_{z1}, \lambda_{z2})}$$

[0046]

수학적 식 2

$$P(z_{d,i} = z | \mathbf{z}', \mathbf{w}, \mathbf{t}) \propto \frac{n_{w_{d,i}}^z + \beta}{n^z + V\beta} \times \frac{n_z^d + \alpha}{n^d + M\alpha} \times \frac{(1 - t_d)^{\lambda_{z1}-1} t_d^{\lambda_{z2}-1}}{B(\lambda_{z1}, \lambda_{z2})}$$

[0047]

- [0048] 수학적 식 1은 해시태그, 시간, 및 단어가 주어졌을 때, 토픽이 z일 확률을 의미하고, 수학적 식 2는 시간과 단어가 주어졌을 때 토픽이 z일 확률을 의미한다. $\mathbf{z}'$ 은 토픽에 대한 할당 벡터(assignment vector)로, 문서 d의 i번째 단어와 해시태그를 제외한 모든 단어 및 해시태그에 대한 것이다.

[0049] 아래의 수학적식 3과 수학적식 4는 토픽 z 에 대한 어휘 분포와 해시태그 분포를 구하기 위한 수식이다.

수학적식 3

$$\varphi = \frac{n_w^z + \beta}{n^z + V\beta}$$

[0050]

수학적식 4

$$\phi = \frac{n_c^z h + \gamma}{n^z h + C\gamma}$$

[0051]

[0052] 수학적식 3을 이용하면 각 토픽을 구성하는 어휘 분포를 추출할 수 있고, 수학적식 4를 이용하면 각 문맥을 구성하는 해시태그 분포를 추출할 수 있다.

[0053] 단계 S140에서, 각 문서의 각 단어에 대해, 토픽 분포로부터 토픽을 추출(또는 선택)하고, 시간에 대한 베타 분포로부터 시간을 추출(또는 선택)한다.

[0054] 단계 S150에서, 어휘 분포로부터 단어를 선택(또는 추출)한다.

[0055] 단계 S160에서, 단어가 해시태그인지 여부를 판단한다. 단어가 해시태그인 경우 단계 S170에서 해시태그 분포로부터 해시태그를 선택(또는 추출)한다.

[0056] 이와 같이, 토픽, 시간, 및 단어를 선택하거나 토픽, 시간, 단어, 및 해시태그를 선택함에 따라, 시간 흐름에 따른 문맥 기반 트렌드 추출이 완료된다.

[0057] 도 6은 도 1에 도시된 트렌드 분석 장치를 이용하여 분석된 트렌드의 예를 도시한다.

[0058] 분석을 위한 자료는 2014년 4월 16일부터 약 한달간 "ferry"와 "prayforsouthkorea" 두 개의 키워드를 이용해 트위터로부터 "세월호 참사"와 관련된 트윗 데이터를 수집했고, 도 6은 트렌드 분석 장치(10)를 통해 추출된 시간 흐름에 따른 문맥 기반 트렌드 중 3개의 예를 도시하고 있다.

[0059] 도 6에서 각 트렌드의 첫째 줄은 트렌드의 이름이고, 둘째 줄은 시간 흐름에 따른 트렌드의 변화를 나타내는 히스토그램이며, 셋째 줄은 트렌드에 대한 문맥 및 트렌드에 대한 단어의 집합이다.

[0060] "Brother saved sister" 트렌드를 살펴보면, 히스토그램을 통해 트렌드에 대한 사용자들의 관심이 어떻게 변화하는지 관찰할 수 있다. 또한, 문맥을 통해, 이 트렌드에 대해 사용자들은 충격을 받았고, 슬픈 감정을 가진다는 것을 알 수 있다. 또한, 이 트렌드는 뉴스 속보로 방송되었고, 청와대와 오바마 대통령과도 연관되어 있음을 추측할 수 있다.

[0061] 본 발명은 도면에 도시된 실시 예를 참고로 설명되었으나 이는 예시적인 것에 불과하며, 본 기술 분야의 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 타 실시 예가 가능하다는 점을 이해할 것이다. 따라서, 본 발명의 진정한 기술적 보호 범위는 첨부된 등록청구범위의 기술적 사상에 의해 정해져야 할 것이다.

부호의 설명

[0062] 10 : 트렌드 분석 장치

100 : 문서 수집부

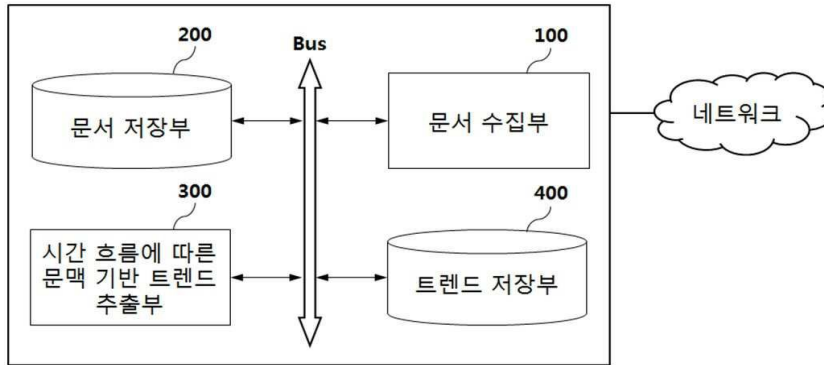
200 : 문서 저장부

300 : 트렌드 추출부

400 : 트렌드 저장부

도면

도면1



10

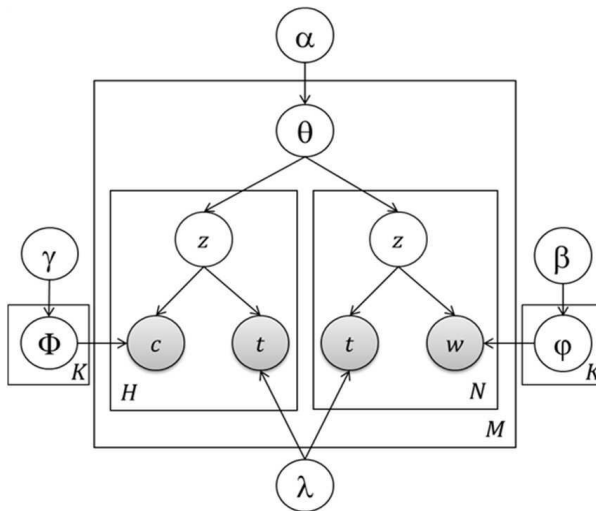
도면2

M	문서(document)의 개수
V	문서 집합 내 유일한(unique) 단어의 개수
N	문서 내 단어(word)의 개수
H	문서 내 해시태그(hashtag)의 개수
K	토픽(topic)의 개수
z, d	각각 토픽 및 문서
w, c, t	각각 단어, 해시태그 및 시간(time)
θ	토픽의 다항 분포(multinomial distribution)
φ	단어의 다항 분포
ϕ	해시태그의 다항 분포
λ	시간에 대한 베타 분포(beta distribution)
α	토픽 분포에 대한 디리클레 사전확률(dirichlet prior)
β	어휘 분포에 대한 디리클레 사전확률
γ	해시태그 분포에 대한 디리클레 사전확률
n_w^z	토픽 z 에 할당된 단어 w 의 횟수
n^z	토픽 z 에 할당된 단어의 개수
n_c^{zh}	토픽 z 에 할당된 해시태그 c 의 횟수
n^{zh}	토픽 z 에 할당된 해시태그의 개수
n_z^d	문서 d 에서 토픽 z 에 할당된 단어의 개수
t_d	문서 d 의 시간
λ_z	토픽 z 의 시간에 대한 베타 분포

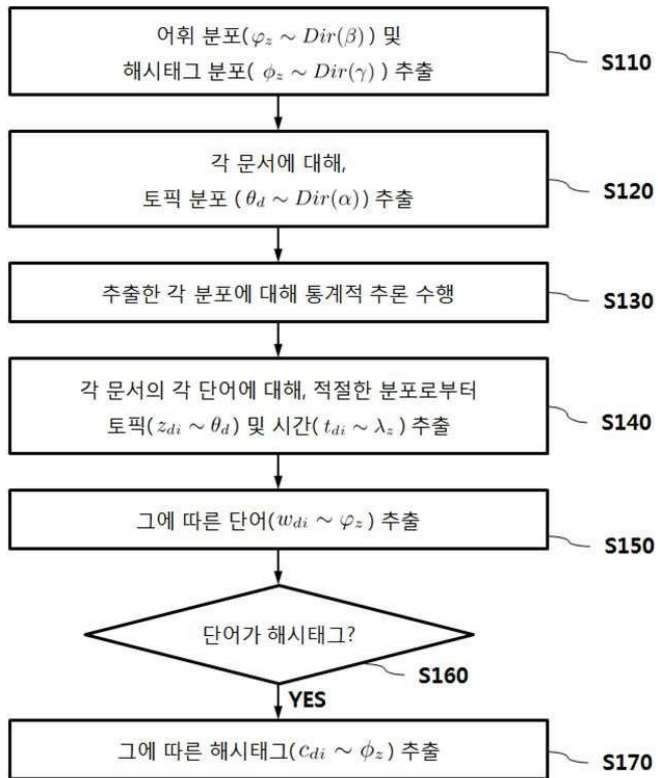
도면3

1. (a) Draw K word distributions $\varphi_z \sim \text{Dir}(\beta)$.
 (b) Draw K hashtag distributions $\Phi_z \sim \text{Dir}(\gamma)$.
2. For each document d ,
 - (a) Draw a distribution of topics $\theta_d \sim \text{Dir}(\alpha)$.
 - (b) For each word w_i ,
 - Draw a topic $z_{d,i} \sim \theta_d$.
 - Draw a word $w_{d,i} \sim \varphi_{z_{d,i}}$.
 - Draw a timestamp $t_{d,i} \sim \text{Beta } \lambda_z$.
 - (c) For each hashtag c_i ,
 - Draw a topic $z_{d,i} \sim \theta_d$.
 - Draw a hashtag $c_{d,i} \sim \Phi_{z_{d,i}}$.
 - Draw a timestamp $t_{d,i} \sim \text{Beta } \lambda_z$.

도면4



도면5



도면6

Brother saved sister	Danwon classroom picture, death toll	Samsung sent crane to raise ferry
#heartbreak #sad #airangnews #samsung #hyundai #ansan #heartbreak #bluehouse #seoul #breakingnew #vixx #obama #pray4southkorea	disaster sewol miss confirm search #heartbreak save dead away float sister sunken 120m korean boy stay 104 strong #sad little	#sewol #together #death toll #hope #danwon #tragedy #restinpeace #live #yellowribbon #staystrongsouthkor ea #school #cnn #northkorea
#sewol student tragedy body kst death toll latest diver victim #rip dead school life recover high sunken vest discover #together	#kb #dna #neverlosehope #video #ferrytragedy #sadstory #worldnews #sunkenship #prayforssouthkorea #komchadluek	lift crane send samsung ceo lee kunhe thank billion thankyou you spent big did damn meant sir #kb