(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2023년10월23일
(11) 등록번호 10-2592079
(24) 등록일자 2023년10월17일

(51) 국제특허분류(Int. Cl.)
G06F 16/35 (2019.01) *G06F 16/2458* (2019.01)
G06F 16/33 (2019.01) *G06F 16/34* (2019.01)
G06F 17/16 (2006.01) *G06F 17/18* (2006.01)
G06F 18/10 (2023.01) *G06F 18/2321* (2023.01)
G06F 18/2413 (2023.01)
(52) CPC특허분류
G06F 16/355 (2019.01)
G06F 16/2462 (2019.01)
(21) 출원번호 10-2023-0036825
(22) 출원일자 2023년03월21일
심사청구일자 2023년03월21일
(56) 선행기술조사문헌
"Latent Dirichlet Allocation(LDA) 모델 기반의 인공지능(AI) 기술 관련 연구 활동 및 동향 분석", Journal of the Korea Industrial Information Systems Research Vol. 23 No. 3(pp. 87-95), 2018.*
"Topic Model Analysis of Research Themes and Trends in the Journal of Economic and Environmental Geology", Econ. Environ. Geol. 2021 54(3)(pp. 353-364), 2021.*
"Topic Modeling with Document Relative Similarities", Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence. 2015.*
KR1020200053334 A
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
부경대학교 산학협력단
부산광역시 남구 용소로 45, 부경대학교내 (대연동)
주식회사 와이블
부산광역시 남구 신선로 365, 7공학관동 219호(용당동, 부경대학교용당캠퍼스)
(72) 발명자
양민준
부산광역시 남구 용소로 45 (대연동, 부경대학교 대연캠퍼스)
김태용
부산광역시 남구 용소로21번길 120, 701호(대연동, 행복빌리지)
박혜민
부산광역시 북구 덕천로234번길 47, 503동 405호(만덕동, 그린코아아파트)
(74) 대리인
김정수

전체 청구항 수 : 총 3 항 심사관 : 박승철

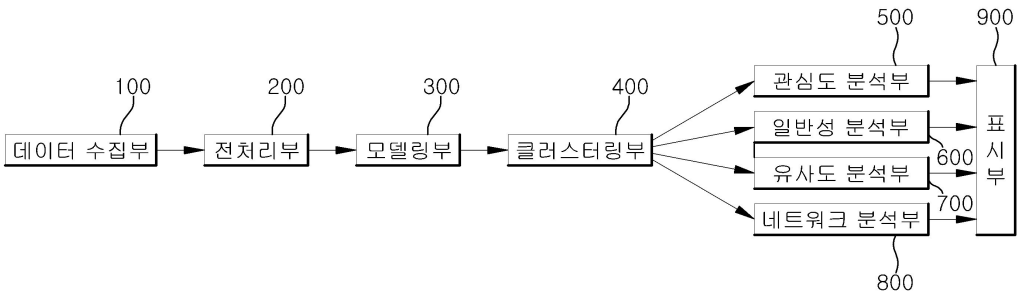
(54) 발명의 명칭 **잠재 디리클레 할당을 이용한 문헌 검토 장치 및 방법**

(57) 요 약

본 발명은 잠재 디리클레 할당(Latent Dirichlet Allocation; LDA)을 이용한 문헌 검토 장치 및 방법에 관한 것이다. 본 발명의 잠재 디리클레 할당을 이용한 문헌 검토 장치는 수집된 대용량의 문헌 데이터를 텍스트 전처리 기법을 이용하여 처리하여서 모델링에 적합한 텍스트 데이터를 추출하도록 구성된 전처리부; 추출된 상기 텍스트

(뒷면에 계속)

대 표 도



데이터에 잠재 디리클레 할당(Latent Dirichlet Allocation)을 적용하여 설정된 개수(K)만큼 주제를 자동으로 분류하고 문헌별 주제의 확률분포 및 주제별 단어의 확률분포를 생성하도록 구성된 모델링부; 생성된 상기 문헌별 주제의 확률분포 및 주제별 단어의 확률분포를 확인하여 주제의 개수(K)를 업데이트하도록 구성된 클러스터링부; 생성된 상기 문헌별 주제의 확률분포와 해당 문헌의 발행연도를 기초로 시간에 따른 주제별 동향을 확인하도록 구성된 관심도 분석부; 생성된 상기 문헌별 주제의 확률분포를 사용하여 일반성을 계산함으로써 특정 주제의 연구범위를 정량적으로 확인하도록 구성된 일반성 분석부; 생성된 상기 주제별 단어의 확률분포를 기초로 주제 간 유사도를 정량적으로 확인하도록 구성된 유사도 분석부; 생성된 상기 문헌별 주제의 확률분포를 기초로 주제 간 연관성을 정량적으로 확인하도록 구성된 네트워크 분석부; 및 확인된 시간에 따른 주제별 동향, 정략적인 특정 주제의 연구범위, 정량적인 주제 간 유사도, 및 정량적인 주제 간 연관성을 디스플레이하도록 구성된 표시부;를 포함한다.

(52) CPC특허분류

G06F 16/2477 (2019.01)
G06F 16/3334 (2019.01)
G06F 16/34 (2019.01)
G06F 17/16 (2013.01)
G06F 17/18 (2013.01)
G06F 18/10 (2023.01)
G06F 18/2321 (2023.01)
G06F 18/2413 (2023.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711135245
과제번호	2021-0-01533-001
부처명	과학기술정보통신부
과제관리(전문)기관명	부경대학교산학협력단
연구사업명	정보통신방송기술개발사업
연구과제명	지능형정보기술 기반 환경오염 특성 및 인체 위해성 평가
기 여 율	1/1
과제수행기관명	지질환경연구소
연구기간	2021.05.01 ~ 2022.04.30

공지예외적용 : 있음

명 세 서

청구범위

청구항 1

대용량의 문헌 데이터를 수집하도록 구성된 데이터 수집부;

수집된 상기 대용량의 문헌 데이터를 텍스트 전처리 기법을 이용하여 처리하여서 모델링에 적합한 텍스트 데이터를 추출하도록 구성된 전처리부;

추출된 상기 텍스트 데이터에 잠재 디리클레 할당(Latent Dirichlet Allocation)을 적용하여 설정된 개수(K)만큼 주제를 자동으로 분류하고 문헌별 주제의 확률분포 및 주제별 단어의 확률분포를 생성하도록 구성된 모델링부;

생성된 상기 문헌별 주제의 확률분포 및 주제별 단어의 확률분포를 확인하여 주제의 개수(K)를 업데이트하도록 구성된 클러스터링부;

생성된 상기 문헌별 주제의 확률분포와 해당 문헌의 발행연도를 기초로 시간에 따른 주제별 동향을 확인하도록 구성된 관심도 분석부;

생성된 상기 문헌별 주제의 확률분포를 사용하여 일반성을 계산함으로써 특정 주제의 연구범위를 정량적으로 확인하도록 구성된 일반성 분석부;

생성된 상기 주제별 단어의 확률분포를 기초로 주제 간 유사도를 정량적으로 확인하도록 구성된 유사도 분석부;

생성된 상기 문헌별 주제의 확률분포를 기초로 주제 간 연관성을 정량적으로 확인하도록 구성된 네트워크 분석부; 및

상기 관심도 분석부, 일반성 분석부, 유사도 분석부 및 네트워크 분석부에서 확인된 시간에 따른 주제별 동향, 정략적인 특정 주제의 연구범위, 정량적인 주제 간 유사도, 및 정량적인 주제 간 연관성을 디스플레이하도록 구성된 표시부;를 포함하며,

상기 클러스터링부는 상기 모델링부에서 소정의 분류 결과가 도출될 때까지 주제의 개수 업데이트를 반복하며,

상기 관심도 분석부는

생성된 상기 문헌별 주제의 확률분포와 해당 문헌의 발행연도를 사용하여 단순선형회귀분석을 실시함으로써 시간에 따른 주제별 동향을 확인하도록 더욱 구성되며, 여기서 선형회귀계수는 다음의 [수학식 1]을 통해 계산되며,

상기 유사도 분석부는

상기 주제별 단어의 확률분포를 이용하여 주제 간 유클리디언 거리를 계산함으로써 주제 간 유사도를 정량적으로 확인하도록 더욱 구성되며, 상기 주제 간 유클리디언 거리는 다음의 [수학식 4]에 의해 계산되는, 잠재 디리클레 할당을 이용한 문헌 검토 장치.

[수학식 1]

$$\theta_k^t = \text{slope}_k(t) + b$$

[여기서, θ_k^t는 연도(t)일 때 주제(k)에 따른 비중을 나타내며, b는 회귀식의 절편을 나타내며, $slope_k$는 주제(k)에 대한 선형회귀계수를 나타냄]

[수학식 4]

$$E_d(\text{topic}_i, \text{topic}_j) = \sqrt{\sum_{n=1}^{N}(w_{i,v} - w_{j,v})^2}$$

[여기서, N은 전체 문헌 내에 존재하는 단어의 총 개수를 나타내며, $w_{i,v}$와 $w_{j,v}$는 주제(i,j)에서 단어 (v)에 대한 가중치를 나타내며, topic_i는 주제 i를 나타내고, topic_j는 주제 j를 나타내며, $E_d(\text{topic}_i, \text{topic}_j)$는 주제 i, j간 유클리디언 거리를 나타냄]

청구항 2

삭제

청구항 3

삭제

청구항 4

제1 항에 있어서,

상기 네트워크 분석부는

상기 문헌별 주제의 확률분포와 해당 행렬의 전치행렬 간 내적을 통해 주제-주제 동시출현행렬을 계산함으로써 주제 간 연관성을 정량적으로 확인하도록 더욱 구성된, 잠재 디리클레 할당을 이용한 문헌 검토 장치.

청구항 5

잠재 디리클레 할당을 이용한 문헌 검토 장치를 이용한 문헌 검토 방법으로서,

데이터 수집부에 의해 대용량의 문헌 데이터가 수집되는 단계;

전처리부가 수집된 상기 대용량의 문헌 데이터를 텍스트 전처리 기법을 이용하여 처리하여서 모델링에 적합한 텍스트 데이터를 추출하는 단계;

모델링부가 추출된 상기 텍스트 데이터에 잠재 디리클레 할당(Latent Dirichlet Allocation)을 적용하여 설정된 개수(K)만큼 주제를 자동으로 분류하고 문헌별 주제의 확률분포 및 주제별 단어의 확률분포를 생성하는 단계;

클러스터링부가 생성된 상기 문헌별 주제의 확률분포 및 주제별 단어의 확률분포를 확인하여 주제의 개수(K)를 업데이트하는 단계;

관심도 분석부가 생성된 상기 문헌별 주제의 확률분포와 해당 문헌의 발행연도를 기초로 시간에 따른 주제별 동향을 확인하는 단계;

일반성 분석부가 생성된 상기 문헌별 주제의 확률분포를 사용하여 일반성을 계산함으로써 특정 주제의 연구범위를 정량적으로 확인하는 단계;

유사도 분석부가 생성된 상기 주제별 단어의 확률분포를 기초로 주제 간 유사도를 정량적으로 확인하는 단계;

네트워크 분석부가 생성된 상기 문헌별 주제의 확률분포를 기초로 주제 간 연관성을 정량적으로 확인하는 단계; 및

표시부를 통해 상기 관심도 분석부, 일반성 분석부, 유사도 분석부 및 네트워크 분석부에서 확인된 시간에 따른

주제별 동향, 정략적인 특정 주제의 연구범위, 정량적인 주제 간 유사도, 및 정량적인 주제 간 연관성이 디스플레이되는 단계;를 포함하며,

상기 클러스터링부는 상기 모델링부에서 소정의 분류 결과가 도출될 때까지 주제의 개수 업데이트를 반복하며,

상기 관심도 분석부는

생성된 상기 문헌별 주제의 확률분포와 해당 문헌의 발행연도를 사용하여 단순선형회귀분석을 실시함으로써 시간에 따른 주제별 동향을 확인하도록 더욱 구성되며, 여기서 선형회귀계수는 다음의 [수학식 1]을 통해 계산되며,

상기 유사도 분석부는

상기 주제별 단어의 확률분포를 이용하여 주제 간 유클리디언 거리를 계산함으로써 주제 간 유사도를 정량적으로 확인하도록 더욱 구성되며, 상기 주제 간 유클리디언 거리는 다음의 [수학식 4]에 의해 계산되는 문헌 검토 방법.

[수학식 1]

$$\theta_k^t = \text{slope}_k(t) + b$$

[여기서, θ_k^t는 연도(t)일 때 주제(k)에 따른 비중을 나타내며, b는 회귀식의 절편을 나타내며, $slope_k$는 주제(k)에 대한 선형회귀계수를 나타냄]

[수학식 4]

$$E_d(\text{topic}_i, \text{topic}_j) = \sqrt{\sum_{n=1}^{N}(w_{i,v} - w_{j,v})^2}$$

[여기서, N은 전체 문헌 내에 존재하는 단어의 총 개수를 나타내며, $w_{i,v}$와 $w_{j,v}$는 주제(i,j)에서 단어(v)에 대한 가중치를 나타내며, topic_i는 주제 i를 나타내고, topic_j는 주제 j를 나타내며, $E_d(\text{topic}_i, \text{topic}_j)$는 주제 i, j간 유클리디언 거리를 나

타냄]

발명의 설명

기 술 분 야

[0001] 본 발명은 잠재 디리클레 할당(Latent Dirichlet Allocation; LDA)을 이용한 문헌 검토 장치 및 방법에 관한 것으로, 특히 대용량 문헌 데이터로부터 추출된 텍스트 데이터에 잠재 디리클레 할당을 적용하여 주제를 자동으로 분류하고 문헌별 주제의 확률분포 및 주제별 단어의 확률분포를 생성하고, 생성된 확률분포를 기초로 시간에 따른 주제별 동향, 특정 주제의 연구범위에 대한 정량적 확인, 주제 간 유사도의 정량적 확인, 및 주제 간 연관성 확인을 수행하여 그 결과를 디스플레이하는, 잠재 디리클레 할당을 이용한 문헌 검토 장치 및 방법에 관한 것이다.

배 경 기 술

[0002] 최근, 인터넷의 발달로 웹 문서 양이 급격하게 증가함에 따라 인터넷에서 생성되는 수많은 대용량의 문서를 토픽별로 분류하는 토픽 분석 기술이 주목을 받고 있으며, 연구가 활발히 진행되고 있다.

[0003] 특히, 토픽 분석 기술 중에서 잠재 디리클레 할당이 가장 널리 이용되고 있다. 잠재 디리클레 할당은 관찰된 단

어를 바탕으로 주제에 따른 단어의 분포를 추정하고 문헌별 주제의 분포를 추정할 수 있는 토픽 모델중 하나이다.

[0004] 그러나, 종래의 잠재 디리클레 할당 모델은 이질적인 문서의 텍스트 데이터를 동질적인 텍스트 데이터로 취급하여 이질적인 문서의 토픽을 분석하는데 한계를 보이는 문제점이 있었다.

[0005] 이를 해결하기 위해, 국내 등록 특허 공보 10-2181744호(이하, 종래 기술이라 함)에는 하이퍼파라미터 값을 자동으로 찾을 수 있고 이질적인 문서의 토픽을 분석할 수 있는 향상된 잠재 디리클레 할당 모델을 이용한 토픽 분석 장치 및 방법이 개시되어 있다.

[0006] 그러나, 종래 기술은 대용량 문헌을 검토함에 있어서 시간 및 비용적인 측면이 고려되지 않았으며, 주제의 동향 및 주제 간 연관성에 대한 수치가 제시되어 있지 않다는 문제점이 있었다.

선행기술문헌

특허문헌

[0007] (특허문헌 0001) 국내 등록 특허 공보 10-2181744호(발명의 명칭: 향상된 잠재 디리클레 할당 모델을 이용한 토픽 분석 장치 및 방법)

발명의 내용

해결하려는 과제

[0008] 따라서 본 발명은 상기와 같은 점을 고려하여 이루어진 것으로서, 본 발명의 목적은 대용량 문헌을 검토함에 있어서 시간 및 비용적인 측면에서 효율적이며, 주제별 동향, 주제의 연구범위, 주제 간 유사도, 및 주제 간 연관성을 사용자에게 정량적으로 제공할 수 있는, 잠재 디리클레 할당을 이용한 문헌 검토 장치 및 방법을 제공하는 데에 있다.

과제의 해결 수단

[0009] 상기 목적을 달성하기 위해, 본 발명의 일실시형태에 의한 잠재 디리클레 할당을 이용한 문헌 검토 장치는 대용량의 문헌 데이터를 수집하도록 구성된 데이터 수집부; 수집된 상기 대용량의 문헌 데이터를 텍스트 전처리 기법을 이용하여 처리하여서 모델링에 적합한 텍스트 데이터를 추출하도록 구성된 전처리부; 추출된 상기 텍스트 데이터에 잠재 디리클레 할당(Latent Dirichlet Allocation)을 적용하여 설정된 개수(K)만큼 주제를 자동으로 분류하고 문헌별 주제의 확률분포 및 주제별 단어의 확률분포를 생성하도록 구성된 모델링부; 생성된 상기 문헌별 주제의 확률분포 및 주제별 단어의 확률분포를 확인하여 주제의 개수(K)를 업데이트하도록 구성된 클러스터링부; 생성된 상기 문헌별 주제의 확률분포와 해당 문헌의 발행연도를 기초로 시간에 따른 주제별 동향을 확인하도록 구성된 관심도 분석부; 생성된 상기 문헌별 주제의 확률분포를 사용하여 일반성을 계산함으로써 특정 주제의 연구범위를 정량적으로 확인하도록 구성된 일반성 분석부; 생성된 상기 주제별 단어의 확률분포를 기초로 주제 간 유사도를 정량적으로 확인하도록 구성된 유사도 분석부; 생성된 상기 문헌별 주제의 확률분포를 기초로 주제 간 연관성을 정량적으로 확인하도록 구성된 네트워크 분석부; 및 상기 관심도 분석부, 일반성 분석부, 유사도 분석부 및 네트워크 분석부에서 확인된 시간에 따른 주제별 동향, 정략적인 특정 주제의 연구범위, 정량적인 주제 간 유사도, 및 정량적인 주제 간 연관성을 디스플레이하도록 구성된 표시부;를 포함하는 것을 특징으로 한다.

[0010] 상기 일실시형태에 의한 잠재 디리클레 할당을 이용한 문헌 검토 장치에 있어서, 상기 관심도 분석부는 생성된 상기 문헌별 주제의 확률분포와 해당 문헌의 발행연도를 사용하여 단순선형회귀분석을 실시함으로써 시간에 따른 주제별 동향을 확인하도록 더욱 구성될 수 있다.

[0011] 상기 일실시형태에 의한 잠재 디리클레 할당을 이용한 문헌 검토 장치에 있어서, 상기 유사도 분석부는 상기 주제별 단어의 확률분포를 이용하여 주제 간 유클리디언 거리를 계산함으로써 주제 간 유사도를 정량적으로 확인하도록 더욱 구성될 수 있다.

[0012] 상기 일실시형태에 의한 잠재 디리클레 할당을 이용한 문헌 검토 장치에 있어서, 상기 네트워크 분석부는 상기 문헌별 주제의 확률분포와 해당 행렬의 전치행렬 간 내적을 통해 주제-주제 동시출현행렬을 계산함으로써 주제

간 연관성을 정량적으로 확인하도록 더욱 구성될 수 있다.

[0013] 상기 목적을 달성하기 위해, 본 발명의 다른 실시형태에 의한 잠재 디리클레 할당을 이용한 문헌 검토 장치를 이용한 문헌 검토 방법은 데이터 수집부에 의해 대용량의 문헌 데이터가 수집되는 단계; 전처리부가 수집된 상기 대용량의 문헌 데이터를 텍스트 전처리 기법을 이용하여 처리하여서 모델링에 적합한 텍스트 데이터를 추출하는 단계; 모델링부가 추출된 상기 텍스트 데이터에 잠재 디리클레 할당(Latent Dirichlet Allocation)을 적용하여 설정된 개수(K)만큼 주제를 자동으로 분류하고 문헌별 주제의 확률분포 및 주제별 단어의 확률분포를 생성하는 단계; 클러스터링부가 생성된 상기 문헌별 주제의 확률분포 및 주제별 단어의 확률분포를 확인하여 주제의 개수(K)를 업데이트하는 단계; 관심도 분석부가 생성된 상기 문헌별 주제의 확률분포와 해당 문헌의 발행연도를 기초로 시간에 따른 주제별 동향을 확인하는 단계; 일반성 분석부가 생성된 상기 문헌별 주제의 확률분포를 사용하여 일반성을 계산함으로써 특정 주제의 연구범위를 정량적으로 확인하는 단계; 유사도 분석부가 생성된 상기 주제별 단어의 확률분포를 기초로 주제 간 유사도를 정량적으로 확인하는 단계; 네트워크 분석부가 생성된 상기 문헌별 주제의 확률분포를 기초로 주제 간 연관성을 정량적으로 확인하는 단계; 및 표시부를 통해 상기 관심도 분석부, 일반성 분석부, 유사도 분석부 및 네트워크 분석부에서 확인된 시간에 따른 주제별 동향, 정략적인 특정 주제의 연구범위, 정량적인 주제 간 유사도, 및 정량적인 주제 간 연관성이 디스플레이되는 단계;를 포함하는 것을 특징으로 한다.

발명의 효과

[0014] 본 발명의 실시형태에 의한 잠재 디리클레 할당을 이용한 문헌 검토 장치 및 방법에 의하면, 대용량의 문헌 데이터가 수집되고, 수집된 상기 대용량의 문헌 데이터를 텍스트 전처리 기법을 이용하여 처리하여서 모델링에 적합한 텍스트 데이터를 추출하고, 추출된 상기 텍스트 데이터에 잠재 디리클레 할당(Latent Dirichlet Allocation)을 적용하여 설정된 개수(K)만큼 주제를 자동으로 분류하고 문헌별 주제의 확률분포 및 주제별 단어의 확률분포를 생성하고, 생성된 상기 문헌별 주제의 확률분포 및 주제별 단어의 확률분포를 확인하여 주제의 개수(K)를 업데이트하고, 생성된 상기 문헌별 주제의 확률분포와 해당 문헌의 발행연도를 기초로 시간에 따른 주제별 동향을 확인하고, 생성된 상기 문헌별 주제의 확률분포를 사용하여 일반성을 계산함으로써 특정 주제의 연구범위를 정량적으로 확인하고, 생성된 상기 주제별 단어의 확률분포를 기초로 주제 간 유사도를 정량적으로 확인하고, 생성된 상기 문헌별 주제의 확률분포를 기초로 주제 간 연관성을 정량적으로 확인하고, 확인된 상기 시간에 따른 주제별 동향, 정략적인 특정 주제의 연구범위, 정량적인 주제 간 유사도, 및 정량적인 주제 간 연관성이 디스플레이되도록 구성됨으로써, 대용량 문헌을 검토함에 있어서 시간 및 비용적인 측면에서 효율적이며, 주제별 동향, 주제의 연구범위, 주제 간 유사도, 및 주제 간 연관성을 사용자에게 정량적으로 제공할 수 있다는 뛰어난 효과가 있다.

도면의 간단한 설명

[0015] 도 1은 본 발명의 실시예에 의한 잠재 디리클레 할당을 이용한 문헌 검토 장치의 블록 구성도이다.

도 2는 도 1의 잠재 디리클레 할당을 이용한 문헌 검토 장치를 이용한 문헌 검토 방법을 설명하기 위한 플로우챠트이다.

도 3은 도 1의 전처리부에 의해 수행되는 텍스트 전처리 기법의 상세 플로우챠트이다.

도 4는 도 1의 모델링부에서 사용하는 잠재 디리클레 할당 모델의 개략도이다.

발명을 실시하기 위한 구체적인 내용

[0016] 본 발명의 실시예를 설명함에 있어서, 본 발명과 관련된 공지기술에 대한 구체적인 설명이 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명을 생략하기로 한다. 그리고 후술되는 용어들은 본 발명에서의 기능을 고려하여 정의된 용어들로서 이는 사용자, 운용자의 의도 또는 관례 등에 따라 달라질 수 있다. 그러므로 그 정의는 본 명세서 전반에 걸친 내용을 토대로 내려져야 할 것이다. 상세한 설명에서 사용되는 용어는 단지 본 발명의 실시예를 기술하기 위한 것이며, 결코 제한적으로 해석되어서는 안 된다. 명확하게 달리 사용되지 않는 한, 단수 형태의 표현은 복수 형태의 의미를 포함한다. 본 설명에서, "포함" 또는 "구비"와 같은 표현은 어떤 특성들, 숫자들, 단계들, 동작들, 요소들, 이들의 일부 또는 조합을 가리키기 위한 것이며, 기술된 것 이외에 하나 또는 그 이상의 다른 특성, 숫자, 단계, 동작, 요소, 이들의 일부 또는 조합의 존재 또는 가능성을 배제하는 것으로 해석되어서는 안 된다.

[0017] 도면에서 도시된 각 시스템에서, 몇몇 경우에서의 요소는 각각 동일한 참조 번호 또는 상이한 참조 번호를 가져서 표현된 요소가 상이하거나 유사할 수가 있음을 시사할 수 있다. 그러나 요소는 상이한 구현을 가지고 본 명세서에서 보여지거나 기술된 시스템 중 몇몇 또는 전부와 작동할 수 있다. 도면에서 도시된 다양한 요소는 동일하거나 상이할 수 있다. 어느 것이 제1 요소로 지칭되는지 및 어느 것이 제2 요소로 불리는지는 임의적이다.

[0018] 본 명세서에서 어느 하나의 구성요소가 다른 구성요소로 데이터 또는 신호를 '전송', '전달' 또는 '제공'한다 함은 어느 한 구성요소가 다른 구성요소로 직접 데이터 또는 신호를 전송하는 것은 물론, 적어도 하나의 또 다른 구성요소를 통하여 데이터 또는 신호를 다른 구성요소로 전송하는 것을 포함한다.

[0019] 이하, 본 발명의 실시예를 도면을 참조하여 상세히 설명하기로 한다.

[0020] 도 1은 본 발명의 실시예에 의한 잠재 디리클레 할당을 이용한 문헌 검토 장치의 블록 구성도이다.

[0021] 본 발명의 실시예에 의한 잠재 디리클레 할당을 이용한 문헌 검토 장치는, 도 1에 도시된 바와 같이, 데이터 수집부(100), 전처리부(200), 모델링부(300), 클러스터링부(400), 관심도 분석부(500), 일반성 분석부(600), 유사도 분석부(700), 네트워크 분석부(800) 및 표시부(900)를 포함한다.

[0022] 데이터 수집부(100)는 대용량의 문헌 데이터를 수집하는 역할을 한다. 예컨대, 해외 문헌분석의 경우, 학술 데이터베이스인 Web of science (https://www.webofscience.com/wos/woscc/basic-search) 또는 Scopus (https://www.scopus.com/search/form.uri?display=basic#basic)를 활용하여 title, abstract, keywords, Journal, published year 등 문헌과 관련된 데이터를 일괄적으로 추출하여 모델링에 알맞은 데이터를 준비할 수 있다. 한편, 국내 문헌분석의 경우, RISS (http://www.riss.kr/index.do)를 사용하여 데이터를 수집할 수 있다.

[0023] 데이터 수집부(100)는 문헌분석의 목적에 따라 데이터 수집 방법이 달라질 수 있다. 즉, 특정 저널을 대상으로 문헌분석을 수행할 경우, 해당 저널과 관련된 모든 데이터를 검색하여 수집할 수 있다. 한편, 특정 분야를 대상으로 수집할 경우, 해당 분야에서 주로 사용되는 키워드들을 검색하여 데이터를 수집할 수 있다.

[0024] 전처리부(200)는 데이터 수집부(100)에 의해 수집된 대용량의 문헌 데이터를 텍스트 전처리 기법을 이용하여 처리하여서 모델링에 적합한 텍스트 데이터를 추출하는 역할을 한다.

[0025] 전처리부(200)가 텍스트 전처리 기법을 이용하여 텍스트 데이터를 추출하는 과정에 대해, 도 3을 참조하여 설명하기로 한다.

[0026] 먼저, 수집된 대용량의 문헌 데이터인 원본 데이터(raw data) 중 국문 데이터를 제외한 영문데이터를 추출한 후 대문자를 소문자로 변환한다.

[0027] 다음, 변환된 소문자 데이터에서 구두점 및 띄어쓰기에 따라 단어로 분리해주는 토큰화를 수행한다.

[0028] 다음, 토큰화된 데이터로부터 불용어 사전(stop words)을 사전에 정의하여 관사(예컨대, a, an, the, that) 또는 문헌분석에 불필요한 단어(예컨대, study, method, research)를 제거한다.

[0029] 다음, 공통된 단어(unique)의 개수를 줄이고 해당 단어의 빈도수를 높이기 위해 관사 또는 문헌분석에 불필요한 단어가 제거된 토큰화된 데이터에 어간 추출법(stemming)을 적용하여 단어의 원형 데이터(텍스트 데이터)를 추출한다.

[0030] 모델링부(300)는 전처리부(200)에 의해 추출된 텍스트 데이터에 잠재 디리클레 할당을 적용하여 설정된 개수(K)만큼 주제를 자동으로 분류하고 문헌별 주제의 확률분포 및 주제별 단어의 확률분포를 생성하는 역할을 한다.

[0031] 잠재 디리클레 할당은 확률적 생성 모델로 문헌 내에 존재하는 단어를 기반으로 주제에 따른 단어의 분포를 추정하고, 문헌에 따른 주제의 분포를 추정할 수 있어, 다양한 분야에서 문헌 고찰을 위해 연구 방법으로 채용하고 있다[도 4 참조].

[0032] 클러스터링부(400)는 모델링부(300)에 의해 생성된 문헌별 주제의 확률분포 및 주제별 단어의 확률분포를 확인하여 주제의 개수(K)를 업데이트하는 역할을 한다. 클러스터링부(400)는 기존의 설정되었던 주제의 개수를 바탕으로 분류된 문헌별 주제의 확률분포와 주제별 단어의 확률분포를 확인하고 업데이트한다. 업데이트 과정은 적절한 분류 결과가 도출되기까지 반복 수행된다.

[0033] 관심도 분석부(500)는 모델링부(300)에 의해 생성된 문헌별 주제의 확률분포와 해당 문헌의 발행연도를 기초로

단순선형회귀분석을 실시함으로써 시간에 따른 주제별 동향을 확인하는 역할을 한다.

[0034] 선형회귀계수는 다음의 [수학식 1]을 통해 계산된다.

[0035] 선형회귀계수는 양측 t-검정을 통해 통계적 유의성($p = 0.05$)을 확인하며, 선형회귀계수에 따라 3가지 유형(hot, cold, and no trend)으로 분류한다.

[0036] [수학식 1]

[0037] $$\theta_k^t = \text{slope}_k(t) + b$$

[0038] [여기서, θ_k^t는 연도(t)일 때 주제(k)에 따른 비중을 나타내며, b는 회귀식의 절편을 나타내며, $slope$는 주제(k)에 대한 선형회귀계수를 나타냄]

[0039] 일반성 분석부(600)는 모델링부(300)에 의해 생성된 문헌별 주제의 확률분포를 사용하여 일반성을 계산함으로써 특정 주제의 연구범위를 정량적으로 확인하는 역할을 한다.

[0040] 일반성 분석부(600)는 각 주제의 연구범위를 정량적으로 계산하기 위해 문헌별 주제의 확률분포를 사용하여 일반성을 계산하고, 그 결과를 차원 축소 방법인 주성분 분석(Principal Component Analysis, PCA)을 사용하여 1-D plot으로 시각화한다. 일반성 분석부(600)는 특정 주제($k-i$)를 선택한 후, 주제(k)의 가중치가 가장 높은 문헌들만 추출하여 평균가중치(θ_t)를 계산한다. 다음, 주제(k)에 대해 선택되지 않은 주변 평균 가중치($\theta_{\text{not } k}$)를 계산하였다.

[0041] 이와 같은 과정은 다음의 [수학식 2]와 [수학식 3]을 통해 표현되며, 광범위한 연구범위를 나타내는 주제(generic)와 특정한 분야, 물질, 방법 등을 다루는 주제(specific)로 분류하였다.

[0042] [수학식 2]

[0043] $$\lambda_k = \frac{\theta_k}{\theta_{\text{not } k}}$$

[0044] [여기서, λ_k는 주제(k)의 연구범위를 나타내며, $\lambda_k > 1$이면 광범위한 주제(generic topic)로 분류하였으며, $\lambda_k < 1$이면 특정한 주제(specific)로 분류하였음]

[0045] [수학식 3]

[0046] $$\theta_{\text{not } k} = \sum_{k=1, k \neq i}^{K} \theta_k$$

[0047] 유사도 분석부(700)는 모델링부(300)에 의해 생성된 주제별 단어의 확률분포를 이용하여 다음의 [수학식 4]와 같이 주제 간 유클리디언 거리를 계산함으로써 주제 간 유사도를 정량적으로 확인하는 역할을 한다. 계산된 주제 간 유사도는 덴드로그램(dendrogram)으로 시각화된다.

[0048] [수학식 4]

[0049] $$E_d(\text{topic}_i, \text{topic}_j) = \sqrt{\sum_{n=1}^{N} (w_{i,v} - w_{j,v})^2}$$

[0050] [여기서, N은 전체 문헌 내에 존재하는 단어의 총 개수를 나타내며, $w_{i,v}$와 $w_{j,v}$는 주제(i,j)에서 단어(v)에

대한 가중치를 나타내며, topic_i는 주제 i를 나타내고, topic_j는 주제 j를 나타내며, $E_d(\text{topic}_i, \text{topic}_j)$는 주제 i, j간 유클리디언 거리를 나타냄]

[0051] 네트워크 분석부(800)는 모델링부(300)에 의해 생성된 문헌별 주제의 확률분포(행렬)와 해당 행렬의 전치행렬 간 내적을 통해 주제-주제 동시출현행렬을 다음의 [수학식 5]와 같이 계산함으로써 주제 간 연관성을 정량적으로 확인하는 역할을 한다. 계산된 주제-주제 동시출현행렬을 이용하여 K개 노드(node)와 K^2개 엣지(edge)를 가진 무방향 가중치 네트워크를 생성한다. 여기서, 노드는 각 주제를 의미하며, 엣지는 주제 간 가중치를 의미한다. 높은 가중치는 해당 주제 간 연관성이 높은 것을 의미한다.

[0052] [수학식 5]

[0053] $$M_k = \theta_d \cdot \theta_d^T$$

[0054] [여기서, θ_d는 문헌별 주제의 확률분포를 나타내고, θ_d^T는 θ_d의 전치행렬을 나타내며, M_k는 주제-주제 동시출현행렬을 나타냄]

[0055] 표시부(900)는 관심도 분석부(500), 일반성 분석부(600), 유사도 분석부(700) 및 네트워크 분석부(800)에서 확인된 시간에 따른 주제별 동향, 정략적인 특정 주제의 연구범위, 정량적인 주제 간 유사도, 및 정량적인 주제 간 연관성을 디스플레이하는 역할을 한다.

[0056] 이하, 상기한 바와 같이 구성된 본 발명의 실시예에 의한 잠재 디리클레 할당을 이용한 문헌 검토 장치를 이용한 문헌 검토 방법에 대해서 설명하기로 한다.

[0057] 도 2는 도 1의 잠재 디리클레 할당을 이용한 문헌 검토 장치를 이용한 문헌 검토 방법을 설명하기 위한 플로우 챠트로서, 여기서 S는 스텝(step)을 의미한다.

[0058] 먼저, 데이터 수집부(100)에 의해 대용량의 문헌 데이터가 수집된다(S100).

[0059] 다음, 전처리부(200)가 스텝(S100)에서 수집된 대용량의 문헌 데이터를 텍스트 전처리 기법을 이용하여 처리하여서 모델링에 적합한 텍스트 데이터를 추출한다(S200).

[0060] 다음, 모델링부(300)가 스텝(S200)에서 추출된 텍스트 데이터에 잠재 디리클레 할당(Latent Dirichlet Allocation)을 적용하여 설정된 개수(K)만큼 주제를 자동으로 분류하고, 문헌별 주제의 확률분포 및 주제별 단어의 확률분포를 생성한다(S300).

[0061] 다음, 클러스터링부(400)가 스텝(S300)에서 생성된 문헌별 주제의 확률분포 및 주제별 단어의 확률분포를 확인하여 주제의 개수(K)를 업데이트한다(S400).

[0062] 다음, 관심도 분석부(S500)가 스텝(S300)에서 생성된 문헌별 주제의 확률분포와 해당 문헌의 발행연도를 기초로 시간에 따른 주제별 동향을 확인한다(S500).

[0063] 한편, 스텝(S400) 이후, 일반성 분석부(600)가 스텝(S300)에서 생성된 문헌별 주제의 확률분포를 사용하여 일반성을 계산함으로써 특정 주제의 연구범위를 정량적으로 확인한다(S600).

[0064] 한편, 스텝(S400) 이후, 유사도 분석부(700)가 스텝(S300)에서 생성된 주제별 단어의 확률분포를 기초로 주제 간 유사도를 정량적으로 확인한다(S700).

[0065] 한편, 스텝(S400) 이후, 네트워크 분석부(800)가 생성된 스텝(S300)에서 생성된 문헌별 주제의 확률분포를 기초로 주제 간 연관성을 정량적으로 확인한다(S800).

[0066] 다음, 표시부(900)를 통해 상기 스텝(S500, S600, S700, S800)에서 확인된 시간에 따른 주제별 동향, 정략적인 특정 주제의 연구범위, 정량적인 주제 간 유사도, 및 정량적인 주제 간 연관성이 디스플레이된다(S900).

[0067] 본 발명의 실시형태에 의한 잠재 디리클레 할당을 이용한 문헌 검토 장치 및 방법에 의하면, 대용량의 문헌 데이터가 수집되고, 수집된 상기 대용량의 문헌 데이터를 텍스트 전처리 기법을 이용하여 처리하여서 모델링에 적합한 텍스트 데이터를 추출하고, 추출된 상기 텍스트 데이터에 잠재 디리클레 할당(Latent Dirichlet Allocation)을 적용하여 설정된 개수(K)만큼 주제를 자동으로 분류하고 문헌별 주제의 확률분포 및 주제별 단어의 확률분포를 생성하고, 생성된 상기 문헌별 주제의 확률분포 및 주제별 단어의 확률분포를 확인하여 주제의

개수(K)를 업데이트하고, 생성된 상기 문헌별 주제의 확률분포와 해당 문헌의 발행연도를 기초로 시간에 따른 주제별 동향을 확인하고, 생성된 상기 문헌별 주제의 확률분포를 사용하여 일반성을 계산함으로써 특정 주제의 연구범위를 정량적으로 확인하고, 생성된 상기 주제별 단어의 확률분포를 기초로 주제 간 유사도를 정량적으로 확인하고, 생성된 상기 문헌별 주제의 확률분포를 기초로 주제 간 연관성을 정량적으로 확인하고, 확인된 상기 시간에 따른 주제별 동향, 정략적인 특정 주제의 연구범위, 정량적인 주제 간 유사도, 및 정량적인 주제 간 연관성이 디스플레이되도록 구성됨으로써, 대용량 문헌을 검토함에 있어서 시간 및 비용적인 측면에서 효율적이며, 주제별 동향, 주제의 연구범위, 주제 간 유사도, 및 주제 간 연관성을 사용자에게 정량적으로 제공할 수 있다.

[0068] 도면과 명세서에는 최적의 실시예가 개시되었으며, 특정한 용어들이 사용되었으나 이는 단지 본 발명의 실시형태를 설명하기 위한 목적으로 사용된 것이지 의미를 한정하거나 특허청구범위에 기재된 본 발명의 범위를 제한하기 위하여 사용된 것은 아니다. 그러므로 본 기술 분야의 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 타 실시예가 가능하다는 점을 이해할 수 있을 것이다. 따라서 본 발명의 진정한 기술적 보호범위는 첨부된 특허청구범위의 기술적 사상에 의해 정해져야 할 것이다.

부호의 설명

[0069] 100: 데이터 수집부

200: 전처리부

300: 모델링부

400: 클러스터링부

500: 관심도 분석부

600: 일반성 분석부

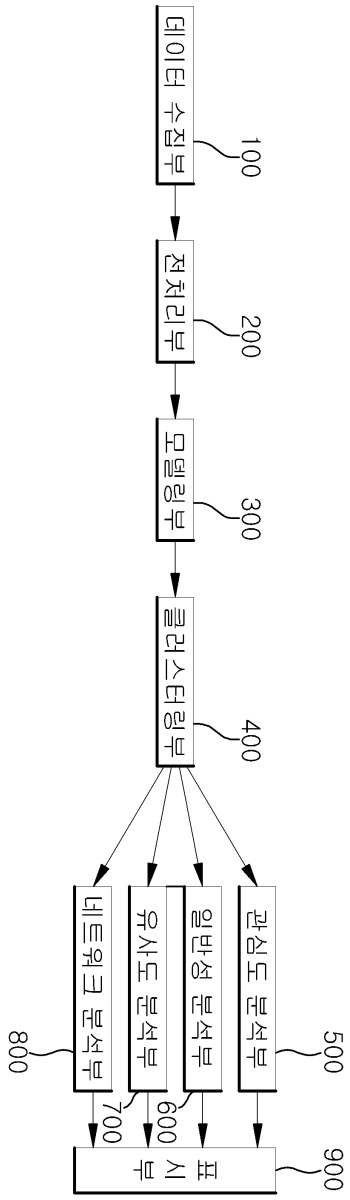
700: 유사도 분석부

800: 네트워크 분석부

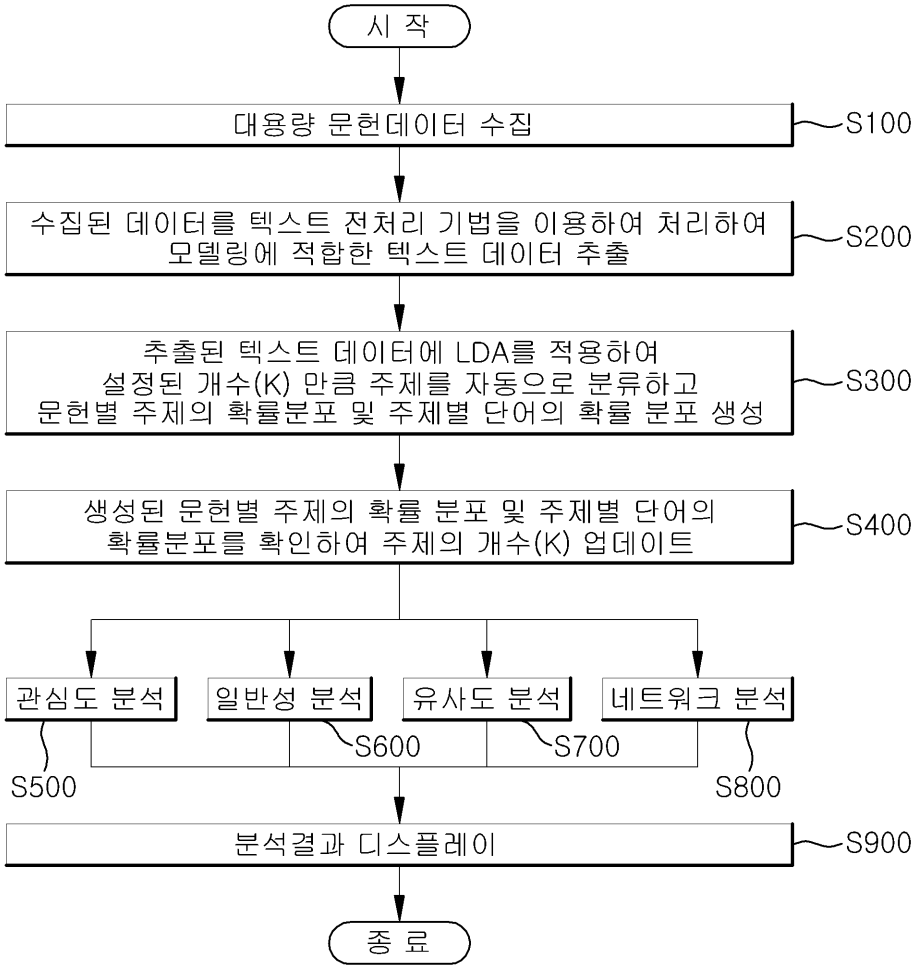
900: 표시부

도면

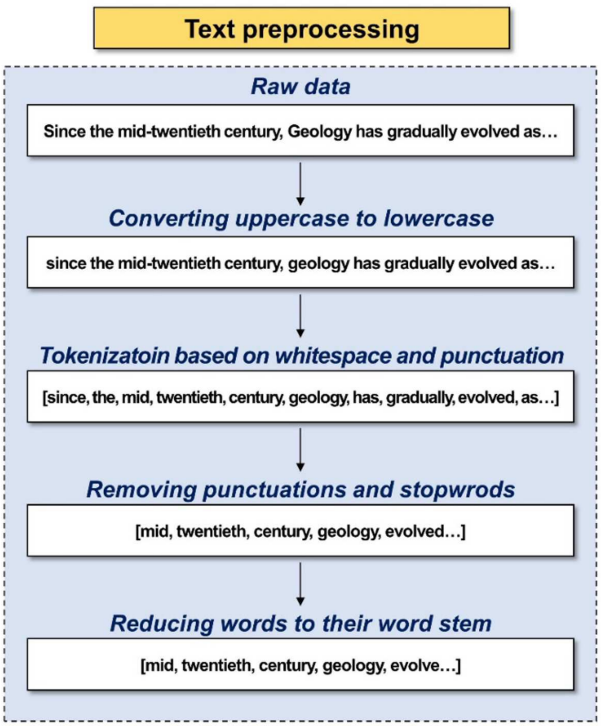
도면1


데이터 수집부
100
전처리부
200
모델링부
300
클러스터링부
400
관심도 분석부
500
일반성 분석부
600
유사도 분석부
700
네트워크 분석부
800
표시부
900

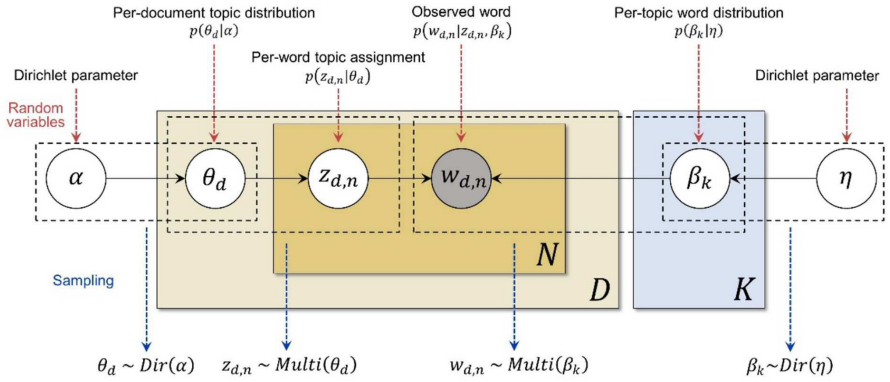
도면2


시 작
대용량 문헌데이터 수집
S100
수집된 데이터를 텍스트 전처리 기법을 이용하여 처리하여
모델링에 적합한 텍스트 데이터 추출
S200
추출된 텍스트 데이터에 LDA를 적용하여
설정된 개수(K) 만큼 주제를 자동으로 분류하고
문헌별 주제의 확률분포 및 주제별 단어의 확률 분포 생성
S300
생성된 문헌별 주제의 확률 분포 및 주제별 단어의
확률분포를 확인하여 주제의 개수(K) 업데이트
S400
관심도 분석
일반성 분석
유사도 분석
네트워크 분석
S500
S600
S700
S800
분석결과 디스플레이
S900
종 료

도면3


Text preprocessing
Raw data
Since the mid-twentieth century, Geology has gradually evolved as…
Converting uppercase to lowercase
since the mid-twentieth century, geology has gradually evolved as…
Tokenizatoin based on whitespace and punctuation
[since, the, mid, twentieth, century, geology, has, gradually, evolved, as…]
Removing punctuations and stopwrods
[mid, twentieth, century, geology, evolved…]
Reducing words to their word stem
[mid, twentieth, century, geology, evolve…]

도면4



【심사관 직권보정사항】

【직권보정 1】

【보정항목】 청구범위

【보정세부항목】 청구항 1

【변경전】

대용량의 문헌 데이터를 수집하도록 구성된 데이터 수집부;

수집된 상기 대용량의 문헌 데이터를 텍스트 전처리 기법을 이용하여 처리하여서 모델링에 적합한 텍스트 데이터를 추출하도록 구성된 전처리부;

추출된 상기 텍스트 데이터에 잠재 디리클레 할당(Latent Dirichlet Allocation)을 적용하여 설정된 개수(K)만큼 주제를 자동으로 분류하고 문헌별 주제의 확률분포 및 주제별 단어의 확률분포를 생성하도록 구성된 모델링부;

생성된 상기 문헌별 주제의 확률분포 및 주제별 단어의 확률분포를 확인하여 주제의 개수(K)를 업데이트하도록 구성된 클러스터링부;

생성된 상기 문헌별 주제의 확률분포와 해당 문헌의 발행연도를 기초로 시간에 따른 주제별 동향을 확인하도록 구성된 관심도 분석부;

생성된 상기 문헌별 주제의 확률분포를 사용하여 일반성을 계산함으로써 특정 주제의 연구범위를 정량적으로 확인하도록 구성된 일반성 분석부;

생성된 상기 주제별 단어의 확률분포를 기초로 주제 간 유사도를 정량적으로 확인하도록 구성된 유사도 분석부;

생성된 상기 문헌별 주제의 확률분포를 기초로 주제 간 연관성을 정량적으로 확인하도록 구성된 네트워크 분석부; 및

상기 관심도 분석부, 일반성 분석부, 유사도 분석부 및 네트워크 분석부에서 확인된 시간에 따른 주제별 동향, 정략적인 특정 주제의 연구범위, 정량적인 주제 간 유사도, 및 정량적인 주제 간 연관성을 디스플레이하도록 구성된 표시부;를 포함하며,

상기 클러스팅부는 상기 모델링부에서 소정의 분류 결과가 도출될 때까지 주제의 개수 업데이트를 반복하며,

상기 관심도 분석부는

생성된 상기 문헌별 주제의 확률분포와 해당 문헌의 발행연도를 사용하여 단순선형회귀분석을 실시함으로써 시간에 따른 주제별 동향을 확인하도록 더욱 구성되며, 여기서 선형회귀계수는 다음의 [수학식 1]을 통해 계산되며,

상기 유사도 분석부는

상기 주제별 단어의 확률분포를 이용하여 주제 간 유클리디언 거리를 계산함으로써 주제 간 유사도를 정량적으로 확인하도록 더욱 구성되며, 상기 주제 간 유클리디언 거리는 다음의 [수학식 4]에 의해 계산되는, 잠재 디리

클레 할당을 이용한 문헌 검토 장치.

[수학식 1]

$$\theta_k^t = \text{slope}_k(t) + b$$

[여기서, θ_k^t

는 연도(t)일 때 주제(k)에 따른 비중을 나타내며, b는 회귀식의 절편을 나타내며, $slop\ e_k$

는 주제(k)에 대한 선형회귀계수를 나타냄]

[수학식 4]

$$E_d(\text{topic}_i, \text{topic}_j) = \sqrt{\sum_{n=1}^{N}(w_{i,v} - w_{j,v})^2}$$

[여기서, N은 전체 문헌 내에 존재하는 단어의 총 개수를 나타내며, $w_{i,v}$

와 $w_{j,v}$

는 주제(i,j)에서 단어(v

)에 대한 가중치를 나타내며, topic_i

는 주제 i를 나타내고, topic_j

는 주제 j를 나타내며, $E_d(\text{topic}_i, \text{topic}_j)$

는 주제 i, j간 유클리디언 거리를 나타냄]

【변경후】

대용량의 문헌 데이터를 수집하도록 구성된 데이터 수집부;

수집된 상기 대용량의 문헌 데이터를 텍스트 전처리 기법을 이용하여 처리하여서 모델링에 적합한 텍스트 데이터를 추출하도록 구성된 전처리부;

추출된 상기 텍스트 데이터에 잠재 디리클레 할당(Latent Dirichlet Allocation)을 적용하여 설정된 개수(K)만큼 주제를 자동으로 분류하고 문헌별 주제의 확률분포 및 주제별 단어의 확률분포를 생성하도록 구성된 모델링부;

생성된 상기 문헌별 주제의 확률분포 및 주제별 단어의 확률분포를 확인하여 주제의 개수(K)를 업데이트하도록 구성된 클러스터링부;

생성된 상기 문헌별 주제의 확률분포와 해당 문헌의 발행연도를 기초로 시간에 따른 주제별 동향을 확인하도록 구성된 관심도 분석부;

생성된 상기 문헌별 주제의 확률분포를 사용하여 일반성을 계산함으로써 특정 주제의 연구범위를 정량적으로 확인하도록 구성된 일반성 분석부;

생성된 상기 주제별 단어의 확률분포를 기초로 주제 간 유사도를 정량적으로 확인하도록 구성된 유사도 분석부;

생성된 상기 문헌별 주제의 확률분포를 기초로 주제 간 연관성을 정량적으로 확인하도록 구성된 네트워크 분석

부; 및

상기 관심도 분석부, 일반성 분석부, 유사도 분석부 및 네트워크 분석부에서 확인된 시간에 따른 주제별 동향, 정략적인 특정 주제의 연구범위, 정량적인 주제 간 유사도, 및 정량적인 주제 간 연관성을 디스플레이하도록 구성된 표시부;를 포함하며,

상기 클러스터링부는 상기 모델링부에서 소정의 분류 결과가 도출될 때까지 주제의 개수 업데이트를 반복하며,

상기 관심도 분석부는

생성된 상기 문헌별 주제의 확률분포와 해당 문헌의 발행연도를 사용하여 단순선형회귀분석을 실시함으로써 시간에 따른 주제별 동향을 확인하도록 더욱 구성되며, 여기서 선형회귀계수는 다음의 [수학식 1]을 통해 계산되며,

상기 유사도 분석부는

상기 주제별 단어의 확률분포를 이용하여 주제 간 유클리디언 거리를 계산함으로써 주제 간 유사도를 정량적으로 확인하도록 더욱 구성되며, 상기 주제 간 유클리디언 거리는 다음의 [수학식 4]에 의해 계산되는, 잠재 디리클레 할당을 이용한 문헌 검토 장치.

[수학식 1]

$$\theta_k^t = \text{slope}_k(t) + b$$

[여기서, θ_k^t

는 연도(t)일 때 주제(k)에 따른 비중을 나타내며, b는 회귀식의 절편을 나타내며, $slop\ e_k$

는 주제(k)에 대한 선형회귀계수를 나타냄]

[수학식 4]

$$E_d(\text{topic}_i, \text{topic}_j) = \sqrt{\sum_{n=1}^{N}(w_{i,v} - w_{j,v})^2}$$

[여기서, N은 전체 문헌 내에 존재하는 단어의 총 개수를 나타내며, $w_{i,v}$

와 $w_{j,v}$

는 주제(i,j)에서 단어(v

)에 대한 가중치를 나타내며, topic_i

는 주제 i를 나타내고, topic_j

는 주제 j를 나타내며, $E_d(\text{topic}_i, \text{topic}_j)$

는 주제 i, j간 유클리디언 거리를 나타냄]

【직권보정 2】

【보정항목】 청구범위

【보정세부항목】청구항 5

【변경전】

잠재 디리클레 할당을 이용한 문헌 검토 장치를 이용한 문헌 검토 방법으로서,

데이터 수집부에 의해 대용량의 문헌 데이터가 수집되는 단계;

전처리부가 수집된 상기 대용량의 문헌 데이터를 텍스트 전처리 기법을 이용하여 처리하여서 모델링에 적합한 텍스트 데이터를 추출하는 단계;

모델링부가 추출된 상기 텍스트 데이터에 잠재 디리클레 할당(Latent Dirichlet Allocation)을 적용하여 설정된 개수(K)만큼 주제를 자동으로 분류하고 문헌별 주제의 확률분포 및 주제별 단어의 확률분포를 생성하는 단계;

클러스터링부가 생성된 상기 문헌별 주제의 확률분포 및 주제별 단어의 확률분포를 확인하여 주제의 개수(K)를 업데이트하는 단계;

관심도 분석부가 생성된 상기 문헌별 주제의 확률분포와 해당 문헌의 발행연도를 기초로 시간에 따른 주제별 동향을 확인하는 단계;

일반성 분석부가 생성된 상기 문헌별 주제의 확률분포를 사용하여 일반성을 계산함으로써 특정 주제의 연구범위를 정량적으로 확인하는 단계;

유사도 분석부가 생성된 상기 주제별 단어의 확률분포를 기초로 주제 간 유사도를 정량적으로 확인하는 단계;

네트워크 분석부가 생성된 상기 문헌별 주제의 확률분포를 기초로 주제 간 연관성을 정량적으로 확인하는 단계; 및

표시부를 통해 상기 관심도 분석부, 일반성 분석부, 유사도 분석부 및 네트워크 분석부에서 확인된 시간에 따른 주제별 동향, 정략적인 특정 주제의 연구범위, 정량적인 주제 간 유사도, 및 정량적인 주제 간 연관성이 디스플레이되는 단계;를 포함하며,

상기 클러스팅부는 상기 모델링부에서 소정의 분류 결과가 도출될 때까지 주제의 개수 업데이트를 반복하며,

상기 관심도 분석부는

생성된 상기 문헌별 주제의 확률분포와 해당 문헌의 발행연도를 사용하여 단순선형회귀분석을 실시함으로써 시간에 따른 주제별 동향을 확인하도록 더욱 구성되며, 여기서 선형회귀계수는 다음의 [수학식 1]을 통해 계산되며,

상기 유사도 분석부는

상기 주제별 단어의 확률분포를 이용하여 주제 간 유클리디언 거리를 계산함으로써 주제 간 유사도를 정량적으로 확인하도록 더욱 구성되며, 상기 주제 간 유클리디언 거리는 다음의 [수학식 4]에 의해 계산되는 문헌 검토 방법.

[수학식 1]

$$\theta_k^t = \text{slope}_k(t) + b$$

[여기서, θ_k^t 는 연도(t)일 때 주제(k)에 따른 비중을 나타내며, b는 회귀식의 절편을 나타내며, $slope_k$ 는 주제(k)에 대한 선형회귀계수를 나타냄]

[수학식 4]

$$E_d(\text{topic}_i, \text{topic}_j) = \sqrt{\sum_{n=1}^{N}(w_{i,v} - w_{j,v})^2}$$

[여기서, N은 전체 문헌 내에 존재하는 단어의 총 개수를 나타내며, $w_{i,v}$ 와 $w_{j,v}$ 는 주제(i,j)에서 단어(v)에 대한 가중치를 나타내며, topic_i 는 주제 i를 나타내고, topic_j 는 주제 j를 나타내며, $E_d(\text{topic}_i, \text{topic}_j)$ 는 주제 i, j간 유클리디언 거리를 나타냄]

【변경후】

잠재 디리클레 할당을 이용한 문헌 검토 장치를 이용한 문헌 검토 방법으로서,

데이터 수집부에 의해 대용량의 문헌 데이터가 수집되는 단계;

전처리부가 수집된 상기 대용량의 문헌 데이터를 텍스트 전처리 기법을 이용하여 처리하여서 모델링에 적합한 텍스트 데이터를 추출하는 단계;

모델링부가 추출된 상기 텍스트 데이터에 잠재 디리클레 할당(Latent Dirichlet Allocation)을 적용하여 설정된 개수(K)만큼 주제를 자동으로 분류하고 문헌별 주제의 확률분포 및 주제별 단어의 확률분포를 생성하는 단계;

클러스터링부가 생성된 상기 문헌별 주제의 확률분포 및 주제별 단어의 확률분포를 확인하여 주제의 개수(K)를 업데이트하는 단계;

관심도 분석부가 생성된 상기 문헌별 주제의 확률분포와 해당 문헌의 발행연도를 기초로 시간에 따른 주제별 동향을 확인하는 단계;

일반성 분석부가 생성된 상기 문헌별 주제의 확률분포를 사용하여 일반성을 계산함으로써 특정 주제의 연구범위를 정량적으로 확인하는 단계;

유사도 분석부가 생성된 상기 주제별 단어의 확률분포를 기초로 주제 간 유사도를 정량적으로 확인하는 단계;

네트워크 분석부가 생성된 상기 문헌별 주제의 확률분포를 기초로 주제 간 연관성을 정량적으로 확인하는 단계; 및

표시부를 통해 상기 관심도 분석부, 일반성 분석부, 유사도 분석부 및 네트워크 분석부에서 확인된 시간에 따른 주제별 동향, 정략적인 특정 주제의 연구범위, 정량적인 주제 간 유사도, 및 정량적인 주제 간 연관성이 디스플레이되는 단계;를 포함하며,

상기 클러스터링부는 상기 모델링부에서 소정의 분류 결과가 도출될 때까지 주제의 개수 업데이트를 반복하며,

상기 관심도 분석부는

생성된 상기 문헌별 주제의 확률분포와 해당 문헌의 발행연도를 사용하여 단순선형회귀분석을 실시함으로써 시간에 따른 주제별 동향을 확인하도록 더욱 구성되며, 여기서 선형회귀계수는 다음의 [수학식 1]을 통해 계산되며,

상기 유사도 분석부는

상기 주제별 단어의 확률분포를 이용하여 주제 간 유클리디언 거리를 계산함으로써 주제 간 유사도를 정량적으로 확인하도록 더욱 구성되며, 상기 주제 간 유클리디언 거리는 다음의 [수학식 4]에 의해 계산되는 문헌 검토 방법.

[수학식 1]

$$\theta_k^t = \text{slope}_k(t) + b$$

[여기서, θ_k^t

는 연도(t)일 때 주제(k)에 따른 비중을 나타내며, b는 회귀식의 절편을 나타내며, $slope_k$

는 주제(k)에 대한 선형회귀계수를 나타냄]

[수학식 4]

$$E_d(\text{topic}_i, \text{topic}_j) = \sqrt{\sum_{n=1}^{N}(w_{i,v} - w_{j,v})^2}$$

[여기서, N은 전체 문헌 내에 존재하는 단어의 총 개수를 나타내며, $w_{i,v}$

와 $w_{j,v}$

는 주제(i,j)에서 단어(v

)에 대한 가중치를 나타내며, topic_i

는 주제 i를 나타내고, topic_j

는 주제 j를 나타내며, $E_d(\text{topic}_i, \text{topic}_j)$

는 주제 i, j간 유클리디언 거리를 나

타냄]