



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2022-0097170
(43) 공개일자 2022년07월07일

(51) 국제특허분류(Int. Cl.)
G06F 16/2458 (2019.01) G06F 16/2457 (2019.01)
G06F 16/26 (2019.01) G06F 16/34 (2019.01)
(52) CPC특허분류
G06F 16/2465 (2019.01)
G06F 16/24578 (2019.01)
(21) 출원번호 10-2021-0116462
(22) 출원일자 2021년09월01일
심사청구일자 2021년09월01일
(30) 우선권주장
1020200187216 2020년12월30일 대한민국(KR)

(71) 출원인
계명대학교 산학협력단
대구광역시 달서구 달구벌대로 1095, 계명대학교 산학협력관 201호(신당동)
(72) 발명자
이수경
대구광역시 달성군 다사읍 서재로12길 38(에코폴리스 동화 아이위시)
정상원
광주광역시 남구 서문대로663번안길 12, 104동 1202호(진월동, 금광하늘연가아파트)
(뒷면에 계속)
(74) 대리인
특허법인(유한)아이시스

전체 청구항 수 : 총 5 항

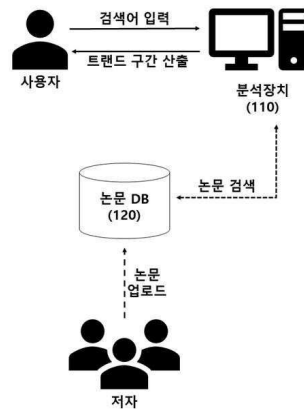
(54) 발명의 명칭 텍스트랭크 기반 헬스 케어 빅데이터 분석 방법 및 장치

(57) 요약

개시된 기술은 텍스트랭크 기반 헬스 케어 빅데이터 분석 방법 및 장치에 관한 것으로, 데이터베이스에 저장된 복수의 헬스 케어 빅데이터들의 초록을 추출하는 단계, 상기 초록에서 복수개의 키워드를 추출하는 단계, 상기 복수의 키워드 각각에 대한 빈도수를 계산하여 일부의 키워드끼리 클러스터링하는 단계 및 클러스터링된 키워드의 빈도수 및 유사도에 따라 트렌드 구간을 산출하는 단계를 포함한다.

대표도 - 도1

100



(52) CPC특허분류

G06F 16/26 (2019.01)

G06F 16/34 (2019.01)

(72) 발명자

김홍기

서울특별시 관악구 관악로 1, 86동 403호(신림동,
서울대학교)

남세진

대전광역시 서구 문정로 131, 9동 1102호(탄방동,
공작한양아파트)

김신실

서울특별시 노원구 동일로207길 186, 109동 1002
호(하계동, 학여울청구아파트)

이 발명을 지원한 국가연구개발사업

과제고유번호 1345315082

과제번호 2018R1D1A3B07050652

부처명 교육부

과제관리(전문)기관명 한국연구재단

연구사업명 지역대학우수과학자지원사업

연구과제명 사회연결망 분석과 기계학습 기법을 활용한 헬스 빅데이터 지능형 분석시스템 개발

및 적용

기 여 율 1/1

과제수행기관명 계명대학교

연구기간 2020.03.01 ~ 2021.02.28

명세서

청구범위

청구항 1

분석장치가 데이터베이스에 저장된 복수의 헬스 케어 빅데이터들의 초록(Abstract)을 추출하는 단계;

상기 분석장치가 상기 초록에서 복수개의 키워드를 추출하는 단계;

상기 분석장치가 상기 복수의 키워드 각각에 대한 빈도수를 계산하여 일부의 키워드끼리 클러스터링하는 단계; 및

상기 분석장치가 클러스터링 된 키워드의 빈도수 및 유사도에 따라 트렌드 구간을 산출하는 단계;를 포함하는 텍스트랭크 기반 헬스 케어 빅데이터 분석 방법.

청구항 2

제 1 항에 있어서,

상기 분석장치는 텍스트랭크(TextRank) 알고리즘을 이용하여 상기 복수개의 키워드를 추출하는 텍스트랭크 기반 헬스 케어 빅데이터 분석 방법.

청구항 3

제 1 항에 있어서,

상기 분석장치는 추출된 복수개의 키워드들 중 특정 키워드의 빈도수와 상기 특정 키워드와 유사한 적어도 하나의 키워드의 빈도수를 계산하여 클러스터링하는 텍스트랭크 기반 헬스 케어 빅데이터 분석 방법.

청구항 4

제 1 항에 있어서,

상기 분석장치는 클러스터링 된 키워드의 유사도 변화를 이하의 수학식 1에 따라 계산하는 텍스트랭크 기반 헬스 케어 빅데이터 분석 방법.

[수학식 1]

$$J(S, T) = \frac{\sum_k \min(S_k, T_k)}{\sum_k \max(S_k, T_k)}$$

(여기에서 S 와 T 는 키워드 빈도수를 의미하고, S_k 와 T_k 는 각각 S 와 T 의 k 번째 값을 의미함.)

청구항 5

복수의 헬스 케어 빅데이터들을 저장하는 저장장치; 및

상기 헬스 케어 빅데이터들의 초록을 추출하고 텍스트랭크 알고리즘을 이용하여 상기 초록에 포함된 복수개의 키워드를 추출하고 상기 복수의 키워드 각각에 대한 빈도수를 계산하여 일부의 키워드를 클러스터링하고 키워드의 빈도수 및 유사도에 따라 트렌드 구간을 산출하는 프로세서;를 포함하는 텍스트랭크 기반 헬스 케어 빅데이터 분석 장치.

발명의 설명

기술 분야

[0001]

개시된 기술은 데이터베이스에 저장된 헬스 케어 빅데이터를 텍스트랭크 알고리즘으로 분석하여 주요 키워드와 트렌드 구간을 추출하는 방법 및 장치에 관한 것이다.

배경 기술

- [0002] 사회 연결망 분석이란 사회 연결망 데이터를 활용하여 사회 구조를 분석하는 하나의 방식이다. 기술의 발전에 더불어 사회의 객체들 사이의 관계는 더욱 복잡해지고, 정교하게 변화하였다. 이러한 관계들을 통해 특정 네트워크의 구조나 개인 간의 상호관계 등을 파악하는 것을 목표로 한다. 주로 수치화, 통계화, 그래프 기반 시각화 등을 이용하여 일련의 과정을 통해 발견된 어떤 형태나 패턴을 분석함으로써 사회구조, 연결망, 관계성, 그룹 또는 개인의 속성을 파악하는 것이 가능하다.
- [0003] 한편, 다수의 논문을 저장하는 데이터베이스를 분석하는 과정에서도 이러한 사회 연결망 분석 방식이 이용되고 있다. 논문에 포함된 주요 키워드, 요약 등을 추출하고 특정 키워드의 빈도수나 유사도, 다수 키워드 사이의 관계를 분석하여 논문에 나타내는 핵심 키워드나 논문 간의 관계 등을 파악할 수 있다.

선행기술문헌

특허문헌

- [0004] (특허문헌 0001) 한국 등록특허 제10-1663454호

발명의 내용

해결하려는 과제

- [0005] 개시된 기술은 데이터베이스에 저장된 헬스 케어 빅데이터를 텍스트랭크 알고리즘으로 분석하여 주요 키워드와 트렌드 구간을 추출하는 방법 및 장치를 제공하는데 있다.

과제의 해결 수단

- [0006] 상기의 기술적 과제를 이루기 위하여 개시된 기술의 제 1 측면은 데이터베이스에 저장된 복수의 헬스 케어 빅데이터들의 초록을 추출하는 단계, 상기 초록에서 복수개의 키워드를 추출하는 단계, 상기 복수의 키워드 각각에 대한 빈도수를 계산하여 일부의 키워드끼리 클러스터링하는 단계 및 클러스터링된 키워드의 빈도수 및 유사도에 따라 트렌드 구간을 산출하는 단계를 포함하는 텍스트랭크 기반 헬스 케어 빅데이터 분석 방법을 제공하는데 있다.
- [0007] 상기의 기술적 과제를 이루기 위하여 개시된 기술의 제 2 측면은 복수의 헬스 케어 빅데이터들을 저장하는 저장장치 및 상기 헬스 케어 빅데이터들의 초록을 추출하고 상기 초록에 포함된 복수개의 키워드를 추출하고 상기 복수의 키워드 각각에 대한 빈도수를 계산하여 일부의 키워드를 클러스터링하고 키워드의 빈도수 및 유사도에 따라 트렌드 구간을 산출하는 프로세서를 포함하는 텍스트랭크 기반 헬스 케어 빅데이터 분석 장치를 제공하는데 있다.

발명의 효과

- [0008] 개시된 기술의 실시 예들은 다음의 장점들을 포함하는 효과를 가질 수 있다. 다만, 개시된 기술의 실시 예들이 이를 전부 포함하여야 한다는 의미는 아니므로, 개시된 기술의 권리범위는 이에 의하여 제한되는 것으로 이해되어서는 아니 될 것이다.
- [0009] 개시된 기술의 일 실시예에 따르면 텍스트랭크 기반 사회 연결망 분석 방법 및 장치는 헬스 케어 빅데이터 간의 유사도를 정량적으로 파악하는 효과가 있다.
- [0010] 또한, 헬스 케어 빅데이터에 포함된 키워드에 대한 트렌드 구간을 설정하여 구간 별로 키워드의 유사도 및 키워드 간의 관계를 분석하는 효과가 있다.

도면의 간단한 설명

- [0011] 도 1은 개시된 기술의 일 실시예에 따른 사회 연결망 분석 시스템을 나타낸 도면이다.
- 도 2는 개시된 기술의 일 실시예에 따라 키워드를 클러스터링하는 것을 나타낸 도면이다.

발명을 실시하기 위한 구체적인 내용

- [0012] 본 발명은 다양한 변경을 가할 수 있고 여러 가지 실시예를 가질 수 있는 바, 특정 실시예들을 도면에 예시하고 상세한 설명에 상세하게 설명하고자 한다. 그러나, 이는 본 발명을 특정한 실시 형태에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다.
- [0013] 제 1, 제 2, A, B 등의 용어는 다양한 구성요소들을 설명하는데 사용될 수 있지만, 해당 구성요소들은 상기 용어들에 의해 한정되지는 않으며, 단지 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 사용된다. 예를 들어, 본 발명의 권리 범위를 벗어나지 않으면서 제 1 구성요소는 제 2 구성요소로 명명될 수 있고, 유사하게 제 2 구성요소도 제 1 구성요소로 명명될 수 있다. 및/또는 이라는 용어는 복수의 관련된 기재된 항목들의 조합 또는 복수의 관련된 기재된 항목들 중의 어느 항목을 포함한다.
- [0014] 본 명세서에서 사용되는 용어에서 단수의 표현은 문맥상 명백하게 다르게 해석되지 않는 한 복수의 표현을 포함하는 것으로 이해되어야 한다. 그리고 "포함한다" 등의 용어는 실시된 특징, 개수, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것이 존재함을 의미하는 것이지, 하나 또는 그 이상의 다른 특징들이나 개수, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 배제하지 않는 것으로 이해되어야 한다.
- [0015] 도면에 대한 상세한 설명을 하기에 앞서, 본 명세서에서의 구성부들에 대한 구분은 각 구성부가 담당하는 주기능 별로 구분한 것에 불과함을 명확히 하고자 한다. 즉, 이하에서 설명할 2개 이상의 구성부가 하나의 구성부로 합쳐지거나 또는 하나의 구성부가 보다 세분화된 기능별로 2개 이상으로 분화되어 구비될 수도 있다.
- [0016] 그리고 이하에서 설명할 구성부 각각은 자신이 담당하는 주기능 이외에도 다른 구성부가 담당하는 기능 중 일부 또는 전부의 기능을 추가적으로 수행할 수도 있으며, 구성부 각각이 담당하는 주기능 중 일부 기능이 다른 구성부에 의해 전담되어 수행될 수도 있음은 물론이다. 따라서, 본 명세서를 통해 설명되는 각 구성부들의 존재 여부는 기능적으로 해석되어야 할 것이다.
- [0017] 도 1은 개시된 기술의 일 실시예에 따른 사회 연결망 분석 시스템을 나타낸 도면이다. 도 1을 참조하면 시스템(100)은 분석장치(110) 및 데이터베이스(120)로 구성되며 사용자가 분석장치(110)를 이용하여 데이터베이스(120)에 저장된 헬스케어 빅데이터들 중 일부를 검색할 수 있다. 헬스케어 빅데이터는 헬스케어에 관련된 논문이나 뉴스, 기사(Article)를 포함한다. 분석장치(110)는 사용자가 이용하는 PC나 노트북과 같은 단말기일 수 있으며 사용자의 검색어에 매칭되는 복수의 헬스케어 빅데이터들을 검색할 수 있다. 사용자가 입력하는 검색어는 키워드 중 하나로 이용할 수도 있고 단순히 검색단계에서만 이용할 수도 있다.
- [0018] 데이터베이스(120)는 분석장치 내부에 탑재된 저장공간일 수도 있고 별도로 마련된 저장장치일 수도 있다. 데이터베이스(120)는 다수의 논문 저자들이 각자의 논문을 업로드할 수 있다. 뉴스나 기사 또한 마찬가지로 업로드할 수 있다. 데이터베이스(120)는 업로드되는 헬스케어 빅데이터들을 추후 검색이 용이하도록 업로드 날짜나 주요 키워드 등의 메타데이터를 생성하여 함께 저장할 수 있다. 데이터베이스는 자체적으로 헬스케어 빅데이터의 내용을 검색하여 키워드를 메타데이터로 입력할 수도 있고 헬스케어 빅데이터의 저자로부터 입력되는 정보에 따라 메타데이터를 저장할 수도 있다.
- [0019] 분석장치는 사용자의 검색어에 매칭되는 복수의 헬스케어 빅데이터들을 검색하면 초록을 추출할 수 있다. 초록영역을 추출하는 것은 기계적인 방법으로 처리될 수 있다. 예컨대, 작성된 문서의 양식에서 앞부분만 추출하거나 뒷부분만 추출할 수 있다. 또는 'Abstract' 라고 명시된 문단만 추출할 수도 있다.
- [0020] 한편, 분석장치는 추출된 초록에서 키워드를 추출한다. 분석장치는 초록 내에서 자주 사용되는 단어나 서로 유사한 의미를 갖는 용어를 추출할 수 있다. 키워드는 문서 저장 시 데이터베이스가 생성한 메타데이터를 참고하여 키워드를 추출할 수 있다. 또는 헬스케어 빅데이터에 포함된 모든 단어를 계산한 뒤 가장 많이 사용된 단어를 키워드로 추출할 수도 있다.
- [0021] 한편, 키워드를 추출함에 있어서 분석장치는 텍스트랭크(TextRank) 알고리즘을 각 문서마다 적용하여 키워드를 보다 효율적으로 추출할 수 있다. 텍스트랭크 알고리즘은 유사도를 비교하여 문장과 단어 사이에 가중치를 매긴 뒤 하나의 문서에서 많은 빈도수를 갖는 단어나 문장 등을 추출하는 알고리즘이다. 주로 문서 집합을 요약하는 용도로 이용되고 있으나 문서의 요약 보다는 문서를 대표하는 키워드나 핵심 용어 등을 파악하는데 더 직관적이다. 따라서, 검색된 헬스케어 빅데이터의 초록에 대하여 텍스트랭크 알고리즘을 적용하는 것으로 최소 빈도수 이상의 키워드들을 단시간에 추출할 수 있다.

[0022] 이와 같이 키워드를 추출하면 분석장치는 각 키워드에 대한 빈도수를 계산하여 일부의 키워드끼리 클러스터링할 수 있다. 키워드의 빈도수는 초록 내에서 키워드의 등장 횟수일 수 있다. 그리고 일부 키워드를 클러스터링하는 것은 서로 유사한 의미를 갖거나 특정 용어를 대체하는 용도로 다른 용어가 사용되는 경우 서로 유사한 것으로 판단하여 클러스터링할 수 있다. 일 실시예로, ‘약물’이나 ‘약품’과 같은 키워드의 경우 서로 유사한 의미를 갖는 키워드이므로 같은 그룹으로 클러스터링할 수 있다. 반면, ‘인삼’이나 ‘별침’과 같은 키워드 간에는 서로 대체하는 용도의 키워드가 아니고 유사도 또한 없으므로 각각 별개로 클러스터링할 수 있다.

[0023] 한편, 키워드를 유사도에 따라 클러스터링하면 분석장치는 키워드의 빈도수 및 유사도에 따라 트렌드 구간을 산출할 수 있다. 트렌드 구간은 시간대 별로 나뉠 수 있으며 특정 키워드가 시간대 별로 점진적으로 등장횟수가 증가하거나 감소하는 등의 추세를 파악할 수 있다. 따라서 사용자는 자신이 검색한 논문들에 사용된 키워드가 주로 어느 구간에서 자주 사용되고 어떻게 변화하는지 또는 시간적 흐름에 따라 연구 동향이 어떻게 변화하는지 등을 파악할 수 있다.

[0024] 도 2는 개시된 기술의 일 실시예에 따라 키워드를 클러스터링하는 것을 나타낸 도면이다. 도 2와 같이 시간대 별로 복수개의 트렌드 구간이 설정될 수 있다. 여기에서 각 트렌드 구간에 포함되는 년도의 개수는 정해진 것이 아니라 해당 트렌드 구간에 포함된 키워드의 빈도수와 유사도에 따라 결정될 수 있다. 가령, 트렌드 구간 1(Phase 1)의 경우 1964년부터 1978년까지 14년을 포함하지만 트렌드 구간4(Phase 4)는 1996년부터 2002년까지 6년만 포함될 수 있다. 즉, 모든 트렌드 구간이 동일한 수의 년도를 포함하는 것은 아니다.

[0025] 한편, 구간별 빈도를 기준으로 키워드의 유사도에 대한 변화는 아래의 수학식 1에 따라 계산할 수 있다.

수학식 1

$$J(S, T) = \frac{\sum_k \min(S_k, T_k)}{\sum_k \max(S_k, T_k)}$$

[0026]

[0028] 여기에서 S 와 T 는 키워드 빈도수를 의미하고, S_k 와 T_k 는 각각 S와 T의 K번째 값을 의미한다. 가령 A년도의 S값이 {17, 21, 9, 3, 0, ...}이고, T값이 {61, 101, 59, 22, 4, ...}라면 유사도 $J(S, T)$ 는 $(17+21+9+3+0+...)/(61+101+59+22+4+...)$ 이 될 수 있다. 계산 결과, 키워드 사이의 유사도가 완벽하게 일치하는 경우에는 1.0의 값으로 나타날 수 있다. 그리고 키워드가 1년 이내에 비슷한 빈도로 나타나는 경우에는 유사성 값이 높게 나타날 수 있다.

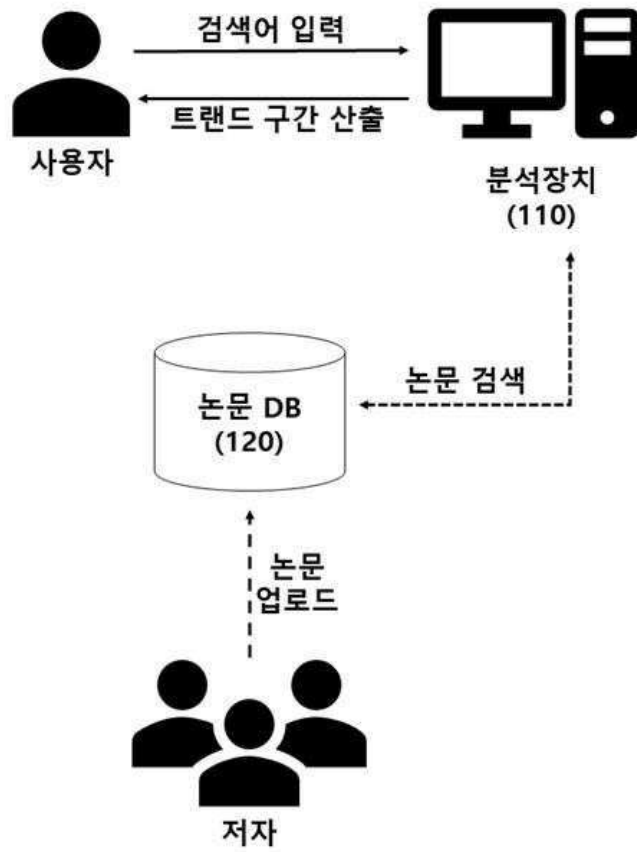
[0029] 한편, 각 트렌드 구간마다 키워드의 빈도수 및 유사도의 증가세 또는 감소세가 서로 다르게 나타날 수 있다. 예컨대, 트렌드 구간1에서는 키워드의 유사도에 대한 증가 폭이 크게 변화하지는 않았으나 트렌드 구간5에서는 유의미한 수준의 증가세를 나타낼 수 있다. 이는 초기에는 특정 키워드가 논문 상에서 자주 언급되지 않았거나 언급되더라도 정해진 몇 가지의 키워드로만 언급되는 수준이었으나 연구 횟수가 증가하고 연구자들 사이에서 보다 폭넓은 연구가 진행됨에 따라 키워드에 대한 언급이 증가하고, 키워드와 관련한 용어들의 사용 또한 증가하였기 때문이다. 사용자는 이와 같이 키워드가 사용되는 시간대를 트렌드 구간으로 나누어 파악함으로써 현재의 헬스케어 연구 동향이나 앞으로의 헬스케어 연구 방향이나 체계 등을 파악할 수 있다.

[0030] 개시된 기술의 일 실시예에 따른 텍스트랭크 기반 헬스케어 빅데이터 분석 방법 및 장치는 이해를 돕기 위하여 도면에 도시된 실시예를 참고로 설명되었으나, 이는 예시적인 것에 불과하며, 당해 분야에서 통상적 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 타 실시예가 가능하다는 점을 이해할 것이다. 따라서, 개시된 기술의 진정한 기술적 보호범위는 첨부된 특허청구범위에 의해 정해져야 할 것이다.

도면

도면1

100



도면2

200

