



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0091271
(43) 공개일자 2023년06월23일

(51) 국제특허분류(Int. Cl.)

G06F 16/35 (2019.01) G06F 17/16 (2006.01)
G06F 40/216 (2020.01) G06F 40/289 (2020.01)
G06N 3/08 (2023.01)

(52) CPC특허분류

G06F 16/35 (2019.01)
G06F 17/16 (2013.01)

(21) 출원번호 10-2021-0180220

(22) 출원일자 2021년12월16일

심사청구일자 2021년12월16일

(71) 출원인

인하대학교 산학협력단

인천광역시 미추홀구 인하로 100(용현동, 인하대학교)

(72) 발명자

강성우

경기도 성남시 분당구 정자일로 248, 611동 3103호 (정자동, 파크뷰)

(74) 대리인

김수호

전체 청구항 수 : 총 6 항

(54) 발명의 명칭 신성장분야에 최적화된 딥러닝을 이용한 특허 분류시스템 및 분류방법

(57) 요약

본 발명은 신성장분야에 최적화된 딥러닝을 이용한 특허 분류방법에 관한 것으로서, 보다 구체적으로는 텍스트마이닝을 이용한 신성장분야에 최적화된 딥러닝을 이용한 특허 분류시스템 및 분류방법에 관한 발명으로써, 본 발명에 따른 신성장분야에 최적화된 딥러닝을 이용한 특허 분류시스템 및 분류방법에 의하면, 문서가 보유하고 있는 자체 정보를 클러스터링에 이용하여 양질의 클러스터링 결과를 제공할 수 있다는 효과가 얻어진다. 또한, 클러스터링 결과 의 적절성을 시각적으로 판단할 수 있어 클러스터들을 일일이 검토할 필요가 없다는 효과도 얻어진다.

대표도 - 도1



(52) CPC특허분류

G06F 40/216 (2020.01)

G06F 40/289 (2020.01)

G06N 3/08 (2023.01)

명세서

청구범위

청구항 1

다수의 문서를 저장하는 데이터베이스를 구비한 서버와 네트워크를 통하여 연결되는 컴퓨터 시스템을 이용하는 문서 그룹화 시스템에 있어서,

상기 서버에 그룹화할 문서를 요청하여 다수의 문서를 전송받는 요청수단;

전송받은 상기 다수의 문서를 구문분석하여 다차원 벡터를 생성하는 벡터수단;

생성된 상기 다차원 벡터와 상기 문서가 갖고 있는 다수의 문서정보를 이용하여 상기 다수의 문서를 클러스터링하는 클러스터링수단;

클러스터링된 다수의 클러스터와 상기 클러스터에 속한 상기 다수의 문서를 시각화하기 위한 정보를 산출하는 시각화수단;

산출된 시각화 정보를 이용하여 상기 클러스터와 상기 다수의 문서를 출력하는 출력수단을 포함하는 것을 특징으로 하는 신성장분야에 최적화된 딥러닝을 이용한 특허 분류시스템

청구항 2

제 1항에 있어서,

상기 클러스터링수단은

상기 다수의 문서에 포함된 각각의 단어를 수집하되, 단어가 수집된 횟수를 기준으로 정규분포로 변환을 진행하며,

상기 정규분포 변환이 진행된 단어 중 z 값이 2이상에 해당하는 단어의 경우 키워드로 설정하며,

각각의 카테고리과 상기 키워드 간에 유사도는,

$\text{sim}(A) = \{\text{simcos}(a, A), \text{simcos}(b, A), \text{simcos}(c, A), \dots\}$ 의 실행에 의해 연산되며,

여기서, A 는 분류 카테고리이고, $a, b, c, \dots$ 는 상기 수집된 키워드이며, $\text{simcos}(a, A)$ 는 a 키워드와 카테고리 A 사이의 코사인유사도이며, $\text{sim}(A)$ 는 각각의 키워드와 A 카테고리 사이의 유사도 행렬인 신성장분야에 최적화된 딥러닝을 이용한 특허 분류시스템

청구항 3

제2항에 있어서,

상기 클러스터링수단은,

$\text{REF}(AA) = \{\text{val}(a, AA), \text{val}(b, AA), \text{val}(c, AA), \dots\}$

여기서, AA 는 분류가 필요한 문서이고, $\text{var}(a, AA)$ 는 a 키워드가 문서 AA 에 포함되었는지 여부이며, $\text{REF}(AA)$ 는 각각의 키워드가 AA 문서에 포함되었는지 여부를 행렬로 나타낸 포함행렬이며,

$\text{REF}(AA)$ 와 $\text{sim}(A)$ 의 내적을 통해 AA 문서의 카테고리를 결정하는 신성장분야에 최적화된 딥러닝을 이용한 특허 분류시스템

청구항 4

다수의 문서를 저장하는 데이터베이스를 구비한 서버와 네트워크를 통하여 연결되는 컴퓨터 시스템을 이용하는 문서 그룹화 방법에 있어서,

상기 서버에 그룹화할 문서를 요청하여 다수의 문서를 전송받는 단계;

전송받은 상기 다수의 문서를 구문분석하여 다차원 벡터를 생성하는 단계;

생성된 상기 다차원 벡터와 상기 문서가 갖고 있는 다수의 문서정보를 이용하여 상기 다수의 문서를 클러스터링하는 단계;

클러스터링된 다수의 클러스터와 상기 클러스터에 속한 상기 다수의 문서를 시각화하기 위한 정보를 산출하는 단계; 및

산출된 시각화 정보를 이용하여 상기 클러스터와 상기 다수의 문서를 출력하는 단계를 포함하는 것을 특징으로 하는 신성장분야에 최적화된 딥러닝을 이용한 특허 분류방법

청구항 5

제 4항에 있어서,

상기 클러스터링하는 단계는,

상기 다수의 문서에 포함된 각각의 단어를 수집하되, 단어가 수집된 횟수를 기준으로 정규분포로 변환을 진행하는 단계;

상기 정규분포 변환이 진행된 단어 중 z 값이 2이상에 해당하는 단어의 경우 키워드로 설정하는 단계 및

상기 두 문서의 유사도를 식

$\text{sim}(A) = \{\text{simcos}(a, A), \text{simcos}(b, A), \text{simcos}(c, A), \dots\}$ 의 실행에 의해 연산하는 단계를 포함하되,

여기서, A 는 분류 카테고리이고, $a, b, c, \dots$ 는 상기 수집된 키워드이며, $\text{simcos}(a, A)$ 는 a 키워드와 카테고리 A 사이의 코사인유사도이며, $\text{sim}(A)$ 는 각각의 키워드와 A 카테고리 사이의 유사도 행렬인 신성장분야에 최적화된 딥러닝을 이용한 특허 분류방법

청구항 6

제5항에 있어서,

상기 클러스터링하는 단계는,

상기 키워드가 문서에 포함되었는지 여부를 아래식을 통해 행렬로 표현하며,

$\text{REF}(AA) = \{\text{val}(a, AA), \text{val}(b, AA), \text{val}(c, AA), \dots\}$

여기서, AA 는 분류가 필요한 문서이고, $\text{val}(a, AA)$ 는 a 키워드가 문서 AA 에 포함되었는지 여부이며, $\text{REF}(AA)$ 는 각각의 키워드가 AA 문서에 포함되었는지 여부를 행렬로 나타낸 포함행렬이며,

$\text{REF}(AA)$ 와 $\text{sim}(A)$ 의 내적을 통해 AA 문서의 카테고리를 결정하는 신성장분야에 최적화된 딥러닝을 이용한 특허 분류방법

발명의 설명

기술 분야

[0001] 본 발명은 신성장분야에 최적화된 딥러닝을 이용한 특허 분류방법에 관한 것으로서, 보다 구체적으로는 텍스트 마이닝을 이용한 신성장분야에 최적화된 딥러닝을 이용한 특허 분류시스템 및 분류방법에 관한 것이다.

배경 기술

[0002] 오늘날 특허정보는 연구개발뿐만 아니라, 과학기술에 대한 국가 정책에 영향을 줄 만큼 그 활용도에 있어서 다양하게 활용되고 있다. 또한 특허정보가 가지는 장점인 잘 정리된 형태의 문서를 이용하여 원하는 정보를 찾아내는 것은 연구개발 시 기존 연구 개발된 결과를 파악하는 것뿐만 아니라, 심판이나 소송에서의 증거자료를 찾는 목적에서 매우 중요한 일이 아닐 수 없다. 따라서 기존에 구축되어 있는 특허정보를 잘 활용하는 것은 기업에 있어서 불필요한 연구개발비를 예방할 수 있고, 새로운 연구개발 방향을 설정하는 측면에서도 매우 중요한 일이라고 할 수 있다.

- [0003] 조사 및 검색된 특허 정보들을 구체적으로 기술개발에 활용하기 위해서는 특허 정보를 체계적으로 구분하여 분류를 해 놓아야 하는데 분류를 하는 과정을 살펴보면 관련 특허를 사람이 일일이 확인하여 분류를 해야 하므로 상당한 인적, 물적, 시간이 소요될 뿐만 아니라 단순 작업에 의한 오류, 사람마다의 이해 차이에 의해 통일된 기준이 없어 체계적인 분류가 되지 못하는 경우가 많았다.
- [0004] 등록특허 10-0718745 "텍스트 마이닝을 이용한 특허 검색 시스템 및 그 방법"은 기존의 텍스트 마이닝을 통한 주제어 도출 및 이를 이용한 자동 특허검색 기술에 관한 것으로서 하나의 특허에 대하여 관련 특허를 검색하는 기술이기 때문에 많은 특허 데이터가 모여져 있는 특허 데이터 베이스를 분석하는 데에는 그 효용성이 감소된다.
- [0005] 또한 여기에서 이용되고 있는 빈도수는 각각의 특허 문헌의 분량을 고려하지 않고 일괄적으로 단어들이 나타나는 횟수만을 기준으로 하였다는 점에서 그 문제점이 있다.
- [0006] 등록특허 10-0809452 "컴퓨팅 장치를 이용한 특허분류 방법 및 시스템"은 컴퓨팅 장치에 의해 기술분류(IPC, UPC, ECLA, FI, F-term, 산업기술분류, 미국산업분류, 사내기술분류 등의 분류표)의 계층구조에 따라 특허 데이터를 분류하는 방법에 관한 것으로 기존에 존재하는 IPC분류 등을 이용하여 특허를 단순 나열식으로 분류하는 것이기 때문에 그 안에서 어떤 새로운 기술을 발견하거나 기술 사이에 어떤 연관 관계를 파악하고 새로운 구성의 기술분류를 하는 데에는 미흡한 문제점이 있다.
- [0007] 또한 IPC 국제 특허분류는 그 분류내용에 있어서 현재의 신기술들을 잘 표현하지 못하고 있으며 상기 분류방법은 이미 작성된 특허분류체계에 맞추어 특허 데이터 베이스에 있는 특허문헌들을 단순 배열하는 것에 불과하다 할 것이다.
- [0008] 따라서 일일이 특허 데이터 베이스에 있는 특허들의 내용을 확인하지 않고도 보다 편리하게 특허 데이터 베이스에 있는 기술 내용을 파악할 수 있도록 새로운 특허 분류를 할 수 있는 방법의 필요성이 대두 되고 있다 할 것이다
- [0009] 따라서 상술된 문제점을 해결하기 위한 기술이 필요하게 되었다.
- [0010] 한편 전술한 배경기술은 발명자가 본 발명의 도출을 위해 보유하고 있었거나, 본 발명의 도출 과정에서 습득한 기술 정보로서, 반드시 본 발명의 출원 전에 일반 공중에게 공개된 공지기술이라 할 수는 없다.

발명의 내용

해결하려는 과제

- [0011] 본 발명의 일 실시예는 신성장분야에 최적화된 딥러닝을 이용한 특허 분류방법을 제공하는 데에 목적이 있다.

과제의 해결 수단

- [0012] 본 발명은 상술한 문제점을 해결하기 위한 것으로서, 다수의 문서를 저장하는 데이터베이스를 구비한 서버와 네트워크를 통하여 연결되는 컴퓨터 시스템을 이용하는 문서 그룹화 시스템에 있어서, 상기 서버에 그룹화할 문서를 요청하여 다수의 문서를 전송받는 요청수단; 전송받은 상기 다수의 문서를 구문분석하여 다차원 벡터를 생성하는 벡터수단; 생성된 상기 다차원 벡터와 상기 문서가 갖고 있는 다수의 문서정보를 이용하여 상기 다수의 문서를 클러스터링하는 클러스터링수단; 클러스터링된 다수의 클러스터와 상기 클러스터에 속한 상기 다수의 문서를 시각화하기 위한 정보를 산출하는 시각화수단; 산출된 시각화 정보를 이용하여 상기 클러스터와 상기 다수의 문서를 출력하는 출력수단을 포함하는 것을 특징으로 한다.
- [0013] 본 발명은 상술한 문제점을 해결하기 위한 것으로서, 다수의 문서를 저장하는 데이터베이스를 구비한 서버와 네트워크를 통하여 연결되는 컴퓨터 시스템을 이용하는 문서 그룹화 방법에 있어서, 상기 서버에 그룹화할 문서를 요청하여 다수의 문서를 전송받는 단계; 전송받은 상기 다수의 문서를 구문분석하여 다차원 벡터를 생성하는 단계; 생성된 상기 다차원 벡터와 상기 문서가 갖고 있는 다수의 문서정보를 이용하여 상기 다수의 문서를 클러스터링하는 단계; 클러스터링된 다수의 클러스터와 상기 클러스터에 속한 상기 다수의 문서를 시각화하기 위한 정보를 산출하는 단계; 및 산출된 시각화 정보를 이용하여 상기 클러스터와 상기 다수의 문서를 출력하는 단계를 포함할 수 있다.

발명의 효과

- [0014] 상술한 바와 같이, 본 발명에 따른 신성장분야에 최적화된 딥러닝을 이용한 특허 분류시스템 및 분류방법에 의하면, 문서가 보유하고 있는 자체 정보를 클러스터링에 이용하여 양질의 클러스터링 결과를 제공할 수 있다는 효과가 얻어진다. 또한, 클러스터링 결과의 적절성을 시각적으로 판단할 수 있어 클러스터들을 일일이 검토할 필요가 없다는 효과도 얻어진다.
- [0015] 또한, 본 발명의 일실시예에 따르면 각각의 카테고리별로 문서들이 자동으로 분류되어 문서분류에 인력이 필요하지 않아 인건비 등이 절약된다.
- [0016] 또한, 본 발명의 일실시예에 따르면, 신성장 기술분야는 전문가가 많지 않지만 본 발명의 일실시예에 따라 각각의 문서를 카테고리별로 분류해야 하는 문제가 해결될 수 있다.
- [0017] 본 발명에서 얻을 수 있는 효과는 이상에서 언급한 효과들로 제한되지 않으며, 언급하지 않은 또 다른 효과들은 아래의 기재로부터 본 발명이 속하는 분야에서 통상의 지식을 가진 자에게 명확하게 이해될 수 있을 것이다.

도면의 간단한 설명

- [0018] 도 1은 본 발명의 일 실시예에 따른 특허 분류시스템 및 분류방법에 있어서, 카테고리과 키워드 사이에 유사도 행렬을 구성하는 예시를 나타낸 도면이다.
- 도 2는 본 발명의 일 실시예에 따른 특허 분류시스템 및 분류방법에 있어서, 카테고리과 문서 사이에 유사도 행렬을 구성하는 예시를 나타낸 도면이다.
- 도 3은 본 발명의 일 실시예에 따른 특허 분류시스템 및 분류방법에 있어서, 각각의 문서가 특정 카테고리로 분류되는 모습의 예시를 나타낸 도면이다.

발명을 실시하기 위한 구체적인 내용

- [0019] 아래에서는 첨부한 도면을 참조하여 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 본 발명의 실시 예를 상세히 설명한다. 그러나 본 발명은 여러 가지 상이한 형태로 구현될 수 있으며 여기에서 설명하는 실시 예에 한정되지 않는다. 그리고 도면에서 본 발명을 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략하였으며, 명세서 전체를 통하여 유사한 부분에 대해서는 유사한 도면 부호를 붙였다.
- [0020] 명세서 전체에서, 어떤 부분이 다른 부분과 “연결”되어 있다고 할 때, 이는 “직접적으로 연결”되어 있는 경우뿐만 아니라 그 중간에 다른 소자를 두고 “전기적으로 연결”되어 있는 경우도 포함한다. 또한, 어떤 부분이 어떤 구성요소를 “포함”한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있는 것을 의미한다.
- [0021] 본 발명의 실시예를 설명하기 위한 도면에 개시된 형상, 크기, 비율, 각도, 개수 등은 예시적인 것이므로 본 발명이 도시된 사항에 한정되는 것은 아니다. 명세서 전체에 걸쳐 동일 참조 부호는 동일 구성 요소를 지칭한다. 또한, 본 발명을 설명함에 있어서, 관련된 공지 기술에 대한 구체적인 설명이 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우 그 상세한 설명은 생략한다.
- [0022] 본 명세서에서 언급된 '포함한다', '갖는다', '이루어진다' 등이 사용되는 경우 '~만'이 사용되지 않는 이상 다른 부분이 추가될 수 있다. 구성 요소를 단수로 표현한 경우에 특별히 명시적인 기재 사항이 없는 한 복수를 포함하는 경우를 포함한다.
- [0023] 구성 요소를 해석함에 있어서, 별도의 명시적 기재가 없더라도 오차 범위를 포함하는 것으로 해석한다.
- [0024] 위치 관계에 대한 설명일 경우, 예를 들어, '~상에', '~상부에', '~하부에', '~옆에' 등으로 두 부분의 위치 관계가 설명되는 경우, '바로' 또는 '직접'이 사용되지 않는 이상 두 부분 사이에 하나 이상의 다른 부분이 위치할 수도 있다.
- [0025] 시간 관계에 대한 설명일 경우, 예를 들어, '~후에', '~에 이어서', '~다음에', '~전에' 등으로 시간적 선후 관계가 설명되는 경우, '바로' 또는 '직접'이 사용되지 않는 이상 연속적이지 않은 경우도 포함할 수 있다.
- [0026] 제1, 제2 등이 다양한 구성요소들을 서술하기 위해서 사용되나, 이들 구성요소들은 이들 용어에 의해 제한되지 않는다. 이들 용어들은 단지 하나의 구성요소를 다른 구성요소와 구별하기 위하여 사용하는 것이다. 따라서, 이하에서 언급되는 제1 구성요소는 본 발명의 기술적 사상 내에서 제2 구성요소일 수도 있다.
- [0027] "적어도 하나"의 용어는 하나 이상의 관련 항목으로부터 제시 가능한 모든 조합을 포함하는 것으로 이해되어야

한다. 예를 들어, "제1 항목, 제2 항목 및 제 3 항목 중에서 적어도 하나"의 의미는 제1 항목, 제2 항목 또는 제3 항목 각각 뿐만 아니라 제1 항목, 제2 항목 및 제3 항목 중에서 2개 이상으로부터 제시될 수 있는 모든 항목의 조합을 의미할 수 있다.

[0028] 본 발명의 여러 실시예들의 각각 특징들이 부분적으로 또는 전체적으로 서로 결합 또는 조합 가능하고, 기술적으로 다양한 연동 및 구동이 가능하며, 각 실시예들이 서로에 대하여 독립적으로 실시 가능할 수도 있고 연관 관계로 함께 실시할 수도 있다.

[0029] 이하에서는 본 발명의 일 실시예인 신성장분야에 최적화된 딥러닝을 이용한 특허 분류시스템에 대해서 구체적으로 설명하기로 한다.

[0030] 도 1은 본 발명의 일 실시예에 따른 특허 분류시스템 및 분류방법에 있어서, 카테고리과 키워드 사이에 유사도 행렬을 구성하는 예시를 나타낸 도면이고, 도 2는 본 발명의 일 실시예에 따른 특허 분류시스템 및 분류방법에 있어서, 카테고리과 문서 사이에 유사도행렬을 구성하는 예시를 나타낸 도면이며, 도 3은 본 발명의 일 실시예에 따른 특허 분류시스템 및 분류방법에 있어서, 각각의 문서가 특정 카테고리로 분류되는 모습의 예시를 나타낸 도면이다.

[0031] 우선 본 발명의 일 실시예는, 신성장분야에 최적화된 딥러닝을 이용한 특허 분류시스템에 관한 발명으로써, 서버에 그룹화할 문서를 요청하여 다수의 문서를 전송받는 요청수단; 전송받은 상기 다수의 문서를 구문분석하여 다차원 벡터를 생성하는 벡터수단; 생성된 상기 다차원 벡터와 상기 문서가 갖고 있는 다수의 문서정보를 이용하여 상기 다수의 문서를 클러스터링하는 클러스터링수단; 클러스터링된 다수의 클러스터와 상기 클러스터에 속한 상기 다수의 문서를 시각화하기 위한 정보를 산출하는 시각화수단; 산출된 시각화 정보를 이용하여 상기 클러스터와 상기 다수의 문서를 출력하는 출력수단을 포함할 수 있다.

[0032] 또한, 본 발명에 적용되는 문서는 특허문서로써, 특허문서라 함은 대한민국 특허 또는 실용신안만 한정하는 것은 아니고, 다른 나라의 특허 또는 실용신안도 포함될 수 있다.

[0033] 본 발명의 일 실시예인 벡터수단은 우선 불용어처리를 진행할 수 있다. 불용어처리는 특허 서류의 분석에 필요하지 않은 단어들을 모두 제거하고 필요한 단어만을 남기는 과정을 의미한다. 예를 들어 수소와 관련된 특허를 분류하는 경우, hydrogen, tank와 같은 단어들은 남겨둘 수 있으나, I, can과 같이 특허 분류에 상관 없는 단어는 제거할 수 있으며, 불용어처리는 공개되어 있는 기술로써 자세한 설명은 생략하기로 하겠다.

[0034] 한편, 클러스터링수단은 상기 다수의 문서에 포함된 각각의 단어를 수집하되, 단어가 수집된 횟수를 기준으로 정규분포로 변환을 진행하며, 상기 정규분포 변환이 진행된 단어 중 z값이 2이상에 해당하는 단어의 경우 키워드로 설정할 수 있다.

[0035] 구체적으로 예를 들어 설명하면, 클러스터링 수단은 다수의 문서에 포함된, 불용어처리가 끝난 단어들을 수집할 수 있다. 예를 들어, 다수의 문서에 hydrogen이 200회, I가 400회, tank가 100회, can이 50회가 포함되어 있다고 하면, 불용어처리가 끝난 hydrogen 200회, tank 100회를 수집, 저장할 수 있다.

[0036] 클러스터링 수단은 수집한 단어들을 수집된 횟수에 기초하여 정규분포로 변환할 수 있는데, 아래 식에 의해 정규분포 변환이 진행될 수 있다.

$$z = \frac{x - \varphi}{\sigma}$$

[0037] φ

φ 는 수집된 단어들의 수집평균횟수, σ 는 표준편차이다.

[0038] z는 표준변환 값, x는 각 단어의 수집횟수,

[0039] 예를 들어, 불용어처리가 끝난 단어들이 평균 10회 정도 수집된다고 가정하면, tank는 0보다 큰 수가 될 것이며, hydrogen의 경우 tank보다 큰 수가 될 것이다. 특히 정규분포로 변환되었을 때, 값이 크면 클수록 특허 문서에 해당 단어가 많이 포함되었다는 것을 알 수 있다.

[0040] 각각의 카테고리과 상기 키워드 간에 유사도는, 도 1에 도시된 바와 같이 $\text{sim}(A) = \{\text{simcos}(a, A), \text{simcos}(b, A), \text{simcos}(c, A), \dots\}$ 의 실행에 의해 연산되며, 여기서, A는 분류 카테고리이고, a,b,c,...는 상기 수집된 키워드이며, $\text{simcos}(a, A)$ 는 a키워드와 카테고리 A사이의 코사인유사도이며, $\text{sim}(A)$ 는 각각의 키워드와 A카테고리 사이의 유사도 행렬일 수 있다.

[0041] 예를 들어, 키워드와 각각의 카테고리 사이에 유사도를 판단할 수 있다. 이때 코사인 유사도를 기준으로 판단할 수 있는데, 코사인 유사도(simcos)를 판단하는 공식은 다음과 같다.

$$\frac{V_a \cdot V_A}{|V_a| |V_A|}$$

[0042] $\text{simcos}(a, A) =$

$$\frac{V_a \cdot V_A}{|V_a| |V_A|}$$

[0043] 와 는 각각 키워드 a와 카테고리A의 단어의 통계정보를 갖는 다차원 벡터이다

[0044] 키워드와 벡터는 다차원 벡터로 이루어져있으며, 코사인 유사도 공식에 따라 그 크기는 -1에서 1사이의 값으로 결정된다. 즉, 임베딩이 끝난 각각의 키워드 벡터와와 카테고리 벡터를 내적으로 통해 코사인유사도를 결정하면, 서로 반대되는 의미의 키워드와 카테고리는 -1에 가까운 코사인 유사도를, 서로 대응되는 의미의 키워드와 카테고리는 1에 가까운 코사인 유사도를 나타낼 수 있다.

[0045] 만약 키워드 a,b,c,d,e,f와 카테고리 A가 각각 0.652, 0.152, -0.0342, 0.751, 0.022, 0.096 이라고 가정한다면, $\text{sim}(A)$ 는 {0.652, 0.152, -0.034, 0.751, 0.022, 0.096}가 될 수 있다.

[0046] 한편, 클러스터링수단은, 각각의 문서에 키워드들이 포함되어 있는지 여부를 판단하며, 나아가 각 문서의 카테고리를 결정할 수 있다.

[0047] 구체적으로 클러스터링 수단은 $\text{REF}(AA) = \{\text{val}(a, AA), \text{val}(b, AA), \text{val}(c, AA), \dots\}$ 식에 기초하여 각각의 문서에 키워드 들이 포함되어 있는지 여부를 여기서, AA는 분류가 필요한 문서이고, $\text{val}(a, AA)$ 는 a키워드가 문서AA에 포함되었는지 여부이며, $\text{REF}(AA)$ 는 각각의 키워드가 AA문서에 포함되었는지 여부를 행렬로 나타낸 포함행렬이며, $\text{REF}(AA)$ 와 $\text{sim}(A)$ 의 내적을 통해 AA문서의 카테고리를 결정할 수 있다.

[0048] 더욱 구체적으로 설명하면, 만약 키워드 a가 AA라는 문서에 포함되어 있다면, $\text{val}(a, AA)$ 는 1이 될 것이며, 만약 포함되지 않는다면 $\text{val}(a, AA)$ 는 0이 될 것이다. 만약 AA에 a,c,d가 있고, b,e,f가 없다면 $\text{REF}(AA)$ 의 경우 {1,0,1,1,0,0}이라는 벡터 형태가 될 것이다. 그리고, 최종적으로 AA라는 문서는 A라는 카테고리와의 유사도는 $\text{sim}(A)$ 벡터와 $\text{REF}(AA)$ 벡터를 내적하여 판단할 수 있다. 본 예시의 경우 {0.652, 0.152, -0.034, 0.751, 0.022, 0.096}와 {1,0,1,1,0,0}의 내적값인 1.367이 될 수 있다. AA문서를 각각의 카테고리 A,B,C 등의 카테고리와 유사도를 검토하고, 최종적으로 내적값이 가장 높은 카테고리에 해당한다고 분류할 수 있다.

[0049] 이하에서는, 신성장분야에 최적화된 딥러닝을 이용한 특허 분류방법에 대해서 구체적으로 설명하기로 한다.

[0050] 본 발명의 일실시예에 따르면, 신성장분야에 최적화된 딥러닝을 이용한 특허 분류방법에 있어서, 상기 서버에 그룹화할 문서를 요청하여 다수의 문서를 전송받는 단계; 전송받은 상기 다수의 문서를 구문분석하여 다차원 벡터를 생성하는 단계; 생성된 상기 다차원 벡터와 상기 문서가 갖고 있는 다수의 문서정보를 이용하여 상기 다수의 문서를 클러스터링하는 단계; 클러스터링된 다수의 클러스터와 상기 클러스터에 속한 상기 다수의 문서를 시각화하기 위한 정보를 산출하는 단계; 및 산출된 시각화 정보를 이용하여 상기 클러스터와 상기 다수의 문서를 출력하는 단계를 포함할 수 있다.

[0051] 본 발명의 일실시예는 우선 불용어처리를 진행할 수 있다. 불용어처리는 특허 서류의 분석에 필요하지 않은 단어들을 모두 제거하고 필요한 단어만을 남기는 과정을 의미한다. 예를 들어 수소와 관련된 특허를 분류하는 경우, hydrogen, tank와 같은 단어들은 남겨둘 수 있으나, I, can과 같이 특허 분류에 상관 없는 단어는 제거할 수 있으며, 불용어처리는 공개되어 있는 기술로써 자세한 설명은 생략하기로 하겠다.

[0052] 한편, 본 발명의 일실시예는 상기 다수의 문서에 포함된 각각의 단어를 수집하되, 단어가 수집된 횟수를 기준으로 정규분포로 변환을 진행하며, 상기 정규분포 변환이 진행된 단어 중 z값이 2이상에 해당하는 단어의 경우 키워드로 설정할 수 있다.

[0053] 구체적으로 예를 들어 설명하면, 본 발명의 일실시예는 다수의 문서에 포함된, 불용어처리가 끝난 단어들을 수집할 수 있다. 예를 들어, 다수의 문서에 hydrogen이 200회, I가 400회, tank가 100회, can이 50회가 포함되어 있다고 하면, 불용어처리가 끝난 hydrogen 200회, tank 100회를 수집, 저장할 수 있다.

[0054] 한편, 본 발명의 일실시예에 포함된 클러스터링하는 단계는, 다수의 문서에 포함된 각각의 단어를 수집하되, 단어가 수집된 횟수를 기준으로 정규분포로 변환을 진행하는 단계; 상기 정규분포 변환이 진행된 단어 중 z값이 2 이상에 해당하는 단어의 경우 키워드로 설정하는 단계 및 상기 두 문서의 유사도를 식 $\text{sim}(A) = \{\text{simcos}(a, A),$

$\text{simcos}(b, A), \text{simcos}(c, A), \dots$ 의 실행에 의해 연산하는 단계를 포함할 수 있다.

[0055] 예를 들어, 키워드와 각각의 카테고리 사이에 유사도를 판단할 수 있다. 이때 코사인 유사도를 기준으로 판단할 수 있는데, 코사인 유사도(simcos)를 판단하는 공식은 다음과 같다.

$$\frac{V_a \cdot V_A}{|V_a| |V_A|}$$

[0056] $\text{simcos}(a, A) =$

$$\frac{V_a \cdot V_A}{|V_a| |V_A|}$$

[0057] 와 는 각각 키워드 a와 카테고리A의 단어의 통계정보를 갖는 다차원 벡터이다

[0058] 키워드와 벡터는 다차원 벡터로 이루어져있으며, 코사인 유사도 공식에 따라 그 크기는 -1에서 1사이의 값으로 결정된다. 즉, 임베딩이 끝난 각각의 키워드 벡터와와 카테고리 벡터를 내적으로 통해 코사인유사도를 결정하면, 서로 반대되는 의미의 키워드와 카테고리는 -1에 가까운 코사인 유사도를, 서로 대응되는 의미의 키워드와 카테고리는 1에 가까운 코사인 유사도를 나타낼 수 있다.

[0059] 만약 키워드 a,b,c,d,e,f와 카테고리 A가 각각 0.652, 0.152, -0.0342, 0.751, 0.022, 0.096 이라고 가정한다면, $\text{sim}(A)$ 는 {0.652, 0.152, -0.034, 0.751, 0.022, 0.096}가 될 수 있다.

[0060] 한편, 클러스터링수단은, 각각의 문서에 키워드들이 포함되어 있는지 여부를 판단하며, 나아가 각 문서의 카테고리들을 결정할 수 있다.

[0061] 여기서, A는 분류 카테고리이고, a,b,c,...는 상기 수집된 키워드이며, $\text{simcos}(a, A)$ 는 a키워드와 카테고리 A사이의 코사인유사도이며, $\text{sim}(A)$ 는 각각의 키워드와 A카테고리 사이의 유사도 행렬일 수 있다.

[0062] 한편, 클러스터링하는 단계는, 키워드가 문서에 포함되었는지 여부를 아래식을 통해 행렬로 표현하며, $\text{REF}(AA) = \{\text{val}(a, AA), \text{val}(b, AA), \text{val}(c, AA), \dots$ 여기서, AA는 분류가 필요한 문서이고, $\text{var}(a, AA)$ 는 a키워드가 문서 AA에 포함되었는지 여부이며, $\text{REF}(AA)$ 는 각각의 키워드가 AA문서에 포함되었는지 여부를 행렬로 나타낸 포함행렬이며, $\text{REF}(AA)$ 와 $\text{sim}(A)$ 의 내적을 통해 AA문서의 카테고리를 결정할 수 있다.

[0063] 구체적으로 클러스터링 수단은 $\text{REF}(AA) = \{\text{val}(a, AA), \text{val}(b, AA), \text{val}(c, AA), \dots$ 식에 기초하여 각각의 문서에 키워드 들이 포함되어 있는지 여부를 여기서, AA는 분류가 필요한 문서이고, $\text{val}(a, AA)$ 는 a키워드가 문서AA에 포함되었는지 여부이며, $\text{REF}(AA)$ 는 각각의 키워드가 AA문서에 포함되었는지 여부를 행렬로 나타낸 포함행렬이며, $\text{REF}(AA)$ 와 $\text{sim}(A)$ 의 내적을 통해 AA문서의 카테고리를 결정할 수 있다.

[0064] 더욱 구체적으로 설명하면, 만약 키워드 a가 AA라는 문서에 포함되어 있다면, $\text{val}(a, AA)$ 는 1이 될 것이며, 만약 포함되지 않는다면 $\text{val}(a, AA)$ 는 0이 될 것이다. 만약 AA에 a,c,d가 있고, b,e,f가 없다면 $\text{REF}(AA)$ 의 경우 {1,0,1,1,0,0}이라는 벡터 형태가 될 것이다. 그리고, 최종적으로 AA라는 문서는 A라는 카테고리 and 유사도는 $\text{sim}(A)$ 벡터와 $\text{REF}(AA)$ 벡터를 내적하여 판단할 수 있다. 본 예시의 경우 {0.652, 0.152, -0.034, 0.751, 0.022, 0.096}와 {1,0,1,1,0,0}의 내적값인 1.367이 될 수 있다. AA문서를 각각의 카테고리 A,B,C 등의 카테고리와 유사도를 검토하고, 최종적으로 내적값이 가장 높은 카테고리에 해당한다고 분류할 수 있다.

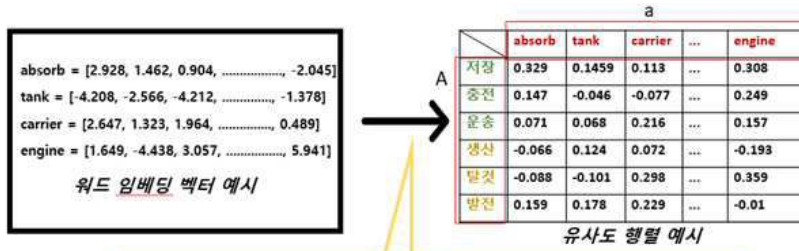
[0065] 전술한 본 발명의 설명은 예시를 위한 것이며, 본 발명이 속하는 기술분야의 통상의 지식을 가진 자는 본 발명의 기술적 사상이나 필수적인 특징을 변경하지 않고서 다른 구체적인 형태로 쉽게 변형이 가능하다는 것을 이해할 수 있을 것이다. 그러므로 이상에서 기술한 실시예들은 모든 면에서 예시적인 것이며 한정적이 아닌 것으로 이해해야만 한다. 예를 들어 단일형으로 설명되어 있는 각 구성 요소는 분산되어 실시될 수도 있으며, 마찬가지로 분산된 것으로 설명되어 있는 구성요소들도 결합된 형태로 실시될 수 있다.

[0066] 본 발명의 범위는 상기 상세한 설명보다는 후술하는 특허청구범위에 의하여 나타내어지며, 특허청구범위의 의미 및 범위 그리고 그 균등 개념으로부터 도출되는 모든 변경 또는 변형된 형태가 본 발명의 범위에 포함되는 것으로 해석되어야 한다.

부호의 설명

도면

도면1



코사인 유사도 측정

- 명시된 기술 분야와 특허 문서 단어들의 코사인 유사도 측정

도면2



도면3

	Doc(1)	Doc(2)	Doc(3)	Doc(4)	...	Doc(m)
저장	1.148 ✓	0.613	0.667	1.944 ✓	...	0.296
충전	-0.259	-0.299	-0.576	-0.336	...	-2.256
운송	0.619	0.044	0.59	1.279	...	3.921 ✓
생산	0.306	1.241 ✓	1.646 ✓	0.372	...	0.426
탈것	-0.091	-1.294	-1.214	0.985	...	-4.201
발전	-0.253	0.576	0.484	0.084	...	0.897
대분류	저장	생산	생산	저장	...	운송