기술명 : 기계학습 기반의 특허 분쟁 예측 방법 및 장치

IPC : G06Q 50/18|G06F 40/258|G06F 40/242

발명자 : 청주대학교 박상성

요 약
본 명세서의 일 실시예에 따른 특허 분쟁 예측 방법은 특허 정량 정보 및 텍스트 정보를 포함하는 복수의 특허 데이터 각각에 대하여 상기 텍스트 정보를 전처리하여 문서-단어 행렬(DTM)을 생성하는 단계, 상기 문서-단어 행 렬 및 벡터 공간에서 특정 단어가 위치한 위치 정보를 포함하는 사전 학습 행렬을 이용하여 특허 임베딩 행렬을 산출하는 단계, 특허 임베딩 행렬에 기반하여 상기 복수의 특허 데이터를 복수의 기술 군집으로 군집화하고, 상 기 기술 군집별로 세부 기술 레이블을 부여하는 단계 및 세부 기술 레이블 및 상기 특허 정량 정보를 입력으로 학습된 특허 분쟁 예측 모델을 통해 상기 복수의 특허 데이터 중 하나인 분석 대상 특허의 특허 분쟁 위험 가능 성을 산출하는 단계를 포함한다. - 도1

청구범위
청구항 1
특허 정량 정보 및 텍스트 정보를 포함하는 복수의 특허 데이터 각각에 대하여 상기 텍스트 정보를 전처리하여 문서-단어 행렬(DTM)을 생성하는 단계; 상기 문서-단어 행렬 및 벡터 공간에서 특정 단어가 위치한 위치 정보를 포함하는 사전 학습 행렬을 이용하여 특허 임베딩 행렬을 산출하는 단계; 상기 특허 임베딩 행렬에 기반하여 상기 복수의 특허 데이터를 복수의 기술 군집으로 군집화하고, 상기 기술 군 집별로 세부 기술 레이블을 부여하는 단계; 및 상기 세부 기술 레이블 및 상기 특허 정량 정보를 입력으로 학습된 특허 분쟁 예측 모델을 통해 상기 복수의 특 허 데이터 중 하나인 분석 대상 특허의 특허 분쟁 위험 가능성을 산출하는 단계를 포함하는 특허 분쟁 예측 방법

청구항 2
제1항에 있어서, 상기 텍스트 정보를 전처리하여 문서-단어 행렬을 생성하는 단계는 상기 텍스트 정보로부터 말뭉치를 추출하는 단계; 상기 말뭉치에서 불용어, 구두점 및 숫자를 제거하여 복수의 표제어를 추출하는 단계; 및 상기 추출한 복수의 표제어에 기초하여 상기 문서-단어 행렬을 생성하는 단계를 포함하는 특허 분쟁 예측 방법.

청구항 3
제1항에 있어서, 상기 특허 임베딩 행렬을 산출하는 단계는 상기 사전 학습 행렬에서 상기 문서-단어 행렬에 포함된 단어와 동일한 단어를 추출하여 상기 룩업 테이블 행렬 (Look-up table)을 생성하는 단계; 및 상기 문서-단어 행렬과 상기 룩업 테이블 행렬의 행렬곱을 수행하여 상기 특허 임베딩 행렬을 산출하는 단계를 포함하는 특허 분쟁 예측 방법.

청구항 4
제1항에 있어서, 상기 특허 정량 정보는 상기 복수의 특허 데이터 각각에 대한 IPC 코드 수,

인용 수, 피인용 수, 패밀리 특허 수, 패밀리 국가 수, 청 구항 수, 발명자 수, 기술이전 여부 및 출원부터 등록까지 소요된 일수 중 적어도 하나를 포함하는 특허 분쟁 예측 방법.

청구항 5
제1항에 있어서, 상기 사전 학습 행렬은 Word2vec 모델을 이용하여 미리 학습된 행렬인 특허 분쟁 예측 방법.

청구항 6
제1항에 있어서, 상기 기술 군집별로 세부 기술 레이블을 부여하는 단계는 상기 복수의 특허 데이터에 포함된 단어 중 등장 횟수 상위 N개의 단어를 이용하여 상기 기술 군집별로 레이블 단어 리스트를 선정하는 단계; 및 상기 레이블 단어 리스트에 기초하여 상기 기술 군집별로 상기 세부 기술 레이블을 부여하는 단계를 포함하는 특허 분쟁 예측 방법.

청구항 7
제1항에 있어서, 상기 군집화는 K 평균 군집화(K-means clustering)를 이용하여 수행되는 특허 분쟁 예측 방법.

청구항 8
제1항에 있어서, 상기 특허 분쟁 예측 모델은 Random survival forest 모델인 특허 분쟁 예측 방법.

청구항 9
특허 정량 정보 및 텍스트 정보를 포함하는 복수의 특허 데이터 각각에 대하여 상기 텍스트 정보를 전처리하여 문서-단어 행렬(DTM)을 생성하는 DTM 생성부; 상기 문서-단어 행렬 및 벡터 공간에서 특정 단어가 위치한 위치 정보를 포함하는 사전 학습 행렬을 이용하여 특허 임베딩 행렬을 산출하는 임베딩 행렬 산출부; 상기 특허 임베딩 행렬에 기반하여 상기 복수의 특허 데이터를 복수의 기술 군집으로 군집화하고, 상기 기술 군 집별로 세부 기술 레이블을 부여하는 레이블링부; 및 상기 세부 기술 레이블 및 상기 특허 정량 정보를 입력으로 학습된 특허 분쟁 예측 모델을 통해 상기 복수의 특 허 데이터 중 하나인 분석 대상 특허의 특허 분쟁 위험 가능성을 산출하는 위험예측부를 포함하는 특허 분쟁 예측 장치. 청구항 10 제9항에 있어서, 상기 DTM생성부는 상기 텍스트 정보로부터 말뭉치를 추출하고, 상기 말뭉치에서 불용어, 구두점 및 숫자를 제거하여 복수의 표제 어를 추출하고, 상기 추출한 복수의 표제어에 기초하여 상기 문서-단어 행렬을 생성하는 특허 분쟁 예측 장치. 청구항 11 제9항에 있어서, 상기 임베딩 행렬 산출부는 상기 사전 학습 행렬에서 상기 문서-단어 행렬에 포함된 단어와 동일한 단어를 추출하여 상기 룩업 테이블 행렬 (Look-up table)을 생성하고, 상기 문서-단어 행렬과 상기 룩업 테이블 행렬의 행렬곱을 수행하여 상기 특허 임 베딩 행렬을 산출하는 특허 분쟁 예측 장치. 청구항 12 제9항에 있어서, 상기 특허 정량 정보는 상기 복수의 특허 데이터 각각에 대한 IPC 코드 수, 인용 수, 피인용 수, 패밀리 특허 수, 패밀리 국가 수, 청구항 수, 발명자 수, 기술이전 여부 및 출원부터 등록까지 소요된 일수 중 적어도 하나를 포함하는 특허 분쟁 예측 장치. 청구항 13 제9항에 있어서, 상기 사전 학습 행렬은 Word2vec 모델을 이용하여 미리 학습된 행렬인 특허 분쟁 예측 장치. 청구항 14 제9항에 있어서, 상기

레이블링부는 상기 복수의 특허 데이터에 포함된 단어 중 등장 횟수 상위 N개의 단어를 이용하여 상기 기술 군집별로 레이블 단어 리스트를 선정하고, 상기 레이블 단어 리스트에 기초하여 상기 기술 군집별로 상기 세부 기술 레이블을 부 여하는 특허 분쟁 예측 장치. 청구항 15 제9항에 있어서, 상기 특허 분쟁 예측 모델은 Random survival forest 모델인 특허 분쟁 예측 장치.

기 술 분 야

본 명세서는 특허 분쟁 예측 방법 및 장치에 관한 것으로, 보다 상세하게는 기계학습에 기반한 특허 분쟁 예측 방법 및 장치에 관한 것이다.

배 경 기 술

특허권이 발생하면 특허권자에게 발명을 독점, 배타적으로 실시할 수 있는 권한이 부여된다. 따라서, 특허 소송 은 무형의 지적 재산인 특허권자의 특허 발명을 정당 권원이 없는 자가 실시하는 경우 등 기업의 제품 또는 서 비스가 다른 기업의 특허 권리 범위를 침해하였을 때 발생한다. 특허 소송이 발생하면, 기업의 제품 및 서비스 판매에 차질이 생겨 경영활동의 지속 가능성을 저해한다. 따라서, 이러한 특허 분쟁을 대비하고 분쟁을 사전에 차단하기 위해서 특허 분쟁을 예측하는 기술의 필요성이 증가하고 있다. 한편, 특허 발명은 수치, 도면, 텍스트 등 개발된 기술의 정보를 다양한 형태로 표현한다. 따라서, 이를 적합한 기법으로 처리하고 활용하는 것이 매우 중요하다. 그 중 특허 빅데이터 기반으로 통계 및 기계학습을 활용한 방법들은 데이터 수집 시점을 기준으로 특허 정보 및 소송 여부를 반영하여 분석을 수행한다. 그러나, 특허의 가치 및 소송 위험은 시간의 흐름에 따라 변화하는데 반해 종래의 방법들은 데이터 수집 시점 이후에 발생할 수 있는 소송 위험을 고려하지 못한다. 따라서, 시간의 흐름에 따른 특허 소송 위험을 객관적이면서도 정량적으로 예측하기 위한 기술의 필요성이 대두되고 있다

해결하려는 과제

본 명세서의 목적은 문서-단어 행렬 및 사전 학습 행렬에 기초한 특허 임베딩 행렬을 이용하여 학습 데이터의 세부 기술을 객관적이면서도 정밀하게 분류할 수 있는 특허 분쟁 예측 방법 및 장치를 제공하는 것이다. 또한, 본 명세서의 목적은 분류된 세부 기술과 특허 정량 지표를 통해 분쟁 예측 모델인 Random Survival Forest(RSF)를 학습시킴으로써 시간 변화에 따른 특허 분쟁 위험을 정량적으로 예측할 수 있는 특허 분쟁 예측 방법 및 장치를 제공하는 것이다. 본 명세서의 목적들은 이상에서 언급한 목적으로 제한되지 않으며, 언급되지 않은 본 명세서의 다른 목적 및 장 점들은 하기의 설명에 의해서 이해될 수 있고, 본 명세서의 실시예에 의해 보다 분명하게 이해될 것이다. 또한, 본 명세서의 목적 및 장점들은 특허 청구 범위에 나타낸 수단 및 그 조합에 의해 실현될 수 있음을 쉽게 알 수 있을 것이다.

과제의 해결 수단

본 명세서의 일 실시예에 따른 특허 분쟁 예측 방법은 특허 정량 정보 및 텍스트 정보를 포함하는 복수의 특허 데이터 각각에 대하여 상기 텍스트 정보를 전처리하여 문서-단어 행렬(DTM)을 생성하는 단계, 상기 문서-단어 행렬 및 벡터 공간에서 특정 단어가 위치한 위치 정보를 포함하는 사전 학습 행렬을 이용하여 특허 임베딩 행렬 을 산출하는 단계, 특허 임베딩

행렬에 기반하여 상기 복수의 특허 데이터를 복수의 기술 군집으로 군집화하고, 상기 기술 군집별로 세부 기술 레이블을 부여하는 단계 및 세부 기술 레이블 및 상기 특허 정량 정보를 입력으 로 학습된 특허 분쟁 예측 모델을 통해 상기 복수의 특허 데이터 중 하나인 분석 대상 특허의 특허 분쟁 위험 가능성을 산출하는 단계를 포함한다. 본 명세서의 일 실시예에서 텍스트 정보를 전처리하여 문서-단어 행렬을 생성하는 단계는 텍스트 정보로부터 말 뭉치를 추출하는 단계, 말뭉치에서 불용어, 구두점 및 숫자를 제거하여 복수의 표제어를 추출하는 단계 및 추출 한 복수의 표제어에 기초하여 상기 문서-단어 행렬을 생성하는 단계를 포함한다. 본 명세서의 일 실시예에서 특허 임베딩 행렬을 산출하는 단계는 사전 학습 행렬에서 상기 문서-단어 행렬에 포 함된 단어와 동일한 단어를 추출하여 상기 룩업 테이블 행렬 (Look-up table)을 생성하는 단계 및 문서-단어 행 렬과 상기 룩업 테이블 행렬의 행렬곱을 수행하여 상기 특허 임베딩 행렬을 산출하는 단계를 포함한다. 본 명세서의 일 실시예에서 상기 특허 정량 정보는 복수의 특허 데이터 각각에 대한 IPC 코드 수, 인용 수, 피 인용 수, 패밀리 특허 수, 패밀리 국가 수, 청구항 수, 발명자 수, 기술이전 여부 및 출원부터 등록까지 소요된 일수 중 적어도 하나를 포함한다. 본 명세서의 일 실시예에서 사전 학습 행렬은 Word2vec 모델을 이용하여 미리 학습된 행렬이다. 본 명세서의 일 실시예에서 기술 군집별로 세부 기술 레이블을 부여하는 단계는 복수의 특허 데이터에 포함된 단어 중 등장 횟수 상위 N개의 단어를 이용하여 기술 군집별로 레이블 단어 리스트를 선정하는 단계 및 레이블 단어 리스트에 기초하여 상기 기술 군집별로 상기 세부 기술 레이블을 부여하는 단계를 포함한다. 본 명세서의 일 실시예에서 군집화는 K 평균 군집화(K-means clustering)를 이용하여 수행된다. 본 명세서의 일 실시예에서 특허 분쟁 예측 모델은 Random survival forest 모델이다. 본 명세서의 일 실시예에 따른 특허 분쟁 예측 장치는 특허 정량 정보 및 텍스트 정보를 포함하는 복수의 특허 데이터 각각에 대하여 상기 텍스트 정보를 전처리하여 문서-단어 행렬(DTM)을 생성하는 DTM 생성부, 문서-단어 행렬 및 벡터 공간에서 특정 단어가 위치한 위치 정보를 포함하는 사전 학습 행렬을 이용하여 특허 임베딩 행렬 을 산출하는 임베딩 행렬 산출부, 특허 임베딩 행렬에 기반하여 상기 복수의 특허 데이터를 복수의 기술 군집으 로 군집화하고, 상기 기술 군집별로 세부 기술 레이블을 부여하는 레이블링부 및 세부 기술 레이블 및 상기 특 허 정량 정보를 입력으로 학습된 특허 분쟁 예측 모델을 통해 상기 복수의 특허 데이터 중 하나인 분석 대상 특 허의 특허 분쟁 위험 가능성을 산출하는 위험예측부를 포함한다. 본 명세서의 일 실시예에서 DTM생성부는 텍스트 정보로부터 말뭉치를 추출하고, 상기 말뭉치에서 불용어, 구두 점 및 숫자를 제거하여 복수의 표제어를 추출하고, 상기 추출한 복수의 표제어에 기초하여 상기 문서-단어 행렬 을 생성한다. 본 명세서의 일 실시예에서 임베딩 행렬 산출부는 사전 학습 행렬에서 상기 문서-단어 행렬에 포함된 단어와 동 일한 단어를 추출하여 상기 룩업 테이블 행렬 (Look-up table)을 생성하고, 상기 문서-단어 행렬과 상기 룩업 테이블 행렬의 행렬곱을 수행하여 상기 특허 임베딩 행렬을 산출한다. 본 명세서의 일 실시예에서 특허 정량 정보는 복수의 특허 데이터 각각에 대한 IPC 코드 수, 인용 수, 피인용 수, 패밀리 특허 수, 패밀리 국가 수, 청구항 수, 발명자 수, 기술이전 여부 및 출원부터 등록까지 소요된 일수 중 적어도 하나를 포함한다. 본 명세서의 일 실시예에서 사전 학습 행렬은 Word2vec 모델을 이용하여 미리 학습된 행렬이다. 본 명세서의 일 실시예에서 레이블링부는 복수의 특허 데이터에 포함된 단어 중 등장 횟수 상위 N개의 단어를 이용하여 상기 기술 군집별로 레이블 단어 리스트를 선정하고, 상기 레이블 단어 리스트에 기초하여 상기 기술 군집별로 상기 세부 기술 레이블을 부여한다. 본 명세서의 일 실시예에서 특허 분쟁 예측 모델은 Random survival

forest 모델이다.

발명의 효과

본 명세서의 일 실시예에 따른 특허 분쟁 예측 방법 및 장치는 문서-단어 행렬 및 사전 학습 행렬에 기초한 특 허 임베딩 행렬을 이용하여 학습 데이터의 세부 기술을 객관적이면서도 정밀하게 분류할 수 있다. 또한, 본 명세서의 일 실시예에 따른 특허 분쟁 예측 방법 및 장치는 분류된 세부 기술과 특허 정량 지표를 통 해 분쟁 예측 모델인 Random Survival Forest(RSF)를 학습시킴으로써 시간 변화에 따른 특허 분쟁 위험을 정량 적으로 예측할 수 있다.

도면의 간단한 설명

도 1은 본 명세서의 일 실시예에 따른 특허 분쟁 예측 장치의 구성도이다. 도 2는 본 명세서의 일 실시예에서 텍스트 정보를 전처리하여 문서-단어 행렬을 생성하는 과정을 나타낸 순서도 이다 도 3은 본 명세서의 일 실시예에서 특허 데이터로부터 문서-단어 행렬을 생성하는 과정을 나타낸 도면이다. 도 4는 본 명세서의 일실시예에서 특허 임베딩 행렬을 산출하는 과정을 나타낸 순서도이다. 도 5는 본 명세서의 일실시예에서 사전 학습 행렬 및 문서-단어 행렬을 이용하여 특허 임베딩 행렬을 산출하는 과정을 나타낸 도면이다. 도 6은 본 명세서의 일 실시예에서 특허 정량 정보를 나타낸 표이다. 도 7은 본 명세서의 일 실시예에서 학습된 특허 분쟁 예측 모델로부터 특허 정량 정보가 특허 분쟁 위험에 영향 을 미치는 정도를 수치화하여 나타낸 표이다. 도 8은 본 명세서의 일 실시예에 따른 특허 분쟁 예측 모델과 다른 모델과의 성능 평가표이다. 도 9는 본 명세서의 일 실시예에 따른 특허 분쟁 예측 방법의 순서도이다.

발명을 실시하기 위한 구체적인 내용

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 이해되어 야 한다. 각 도면을 설명하면서 유사한 참조부호를 유사한 구성요소에 대해 사용하였다. 제1, 제2, A, B 등의 용어는 다양한 구성요소들을 설명하는데 사용될 수 있지만, 상기 구성요소들은 상기 용어 들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 사용된다. 예를 들어, 본 발명의 권리 범위를 벗어나지 않으면서 제1 구성요소는 제2 구성요소로 명명될 수 있 고, 유사하게 제2 구성요소도 제1 구성요소로 명명될 수 있다. 및/또는 이라는 용어는 복수의 관련된 기재된 항 목들의 조합 또는 복수의 관련된 기재된 항목들 중의 어느 항목을 포함한다. 어떤 구성요소가 다른 구성요소에 "연결되어" 있다거나 "접속되어" 있다고 언급된 때에는, 그 다른 구성요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이 해되어야 할 것이다. 반면에, 어떤 구성요소가 다른 구성요소에 "직접 연결되어" 있다거나 "직접 접속되어" 있 다고 언급된 때에는, 중간에 다른 구성요소가 존재하지 않는 것으로 이해되어야 할 것이다. 본 출원에서 사용한 용어는 단지 특정한 실시예를 설명하기 위해 사용된 것으로, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 출원에서, "포함 하다" 또는 "가지다" 등의 용어는 명세서상에 기재된 특징, 숫자, 단계, 동작, 구성

요소, 부품 또는 이들을 조 합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부 품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다. 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 가지고 있다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥 상 가지는 의미와 일치하는 의 미를 가지는 것으로 해석되어야 하며, 본 출원에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적 인 의미로 해석되지 않는다. 이하, 첨부된 도면을 참조하여 본 발명의 바람직한 실시예를 상세하게 설명한다. 도 1은 본 명세서의 일 실시예에 따른 특허 분쟁 예측 장치의 구성도이다. 도면을 참조하면, 특허 분쟁 예측 장치(100)는 특허 분쟁 예측 모델을 통해 분석 대상 특허의 특허 분쟁 위험을 예측하는 장치로써, DTM 생성부(110), 임베딩 행렬 산출부(130), 레이블링부(150) 및 위험예측부(170)를 포함한 다. DTM 생성부(110)는 복수의 특허 데이터로부터 문서-단어 행렬(DTM)을 생성한다. 구체적으로, DTM 생성부(110)는 특허 정량 정보 및 텍스트 정보를 포함하는 복수의 특허 데이터 각각에 대하여 텍스트 정보를 전처리하여 문서- 단어 행렬(DTM)을 생성한다. 여기서, 텍스트 정보는 복수의 특허 데이터에 포함된 텍스트에 관한 정보이다. 일반적으로, 특허 데이터는 텍스 트(문자)뿐만 아니라 도면, 표 및 기호를 포함하여 다양한 정보로 구성되므로, 텍스트 정보는 다양한 정보 중 하나인 텍스트와 관련되어 사람이 읽을 수 있는 정보만을 의미한다. DTM 생성부(110)는 이러한 텍스트 정보를 전처리함으로써 문서-단어 행렬을 생성한다. 구체적으로 DTM 생성부 (110)는 텍스트 정보의 전처리 과정을 통해 텍스트 정보로부터 복수의 표제어를 추출하고, 추출된 복수의 표제 어 각각이 등장하는 빈도를 행렬로 표현하여 문서-단어 행렬을 생성한다. DTM 생성부(110)가 문서-단어 행렬을 생성하는 구체적인 방법에 대해서는 도 2 및 도 3을 참조하여 상세히 설명하도록 한다. 임베딩 행렬 산출부(130)는 문서-단어 행렬 및 벡터 공간에서 특정 단어가 위치한 위치 정보를 포함하는 사전 학습 행렬을 이용하여 특허 임베딩 행렬을 산출한다. 여기서, 사전 학습 행렬이란 벡터 공간에서 특정 단어가 위치한 위치 정보를 포함하는 행렬이다. 즉, 사전 학습 행렬은 인터넷 서버에 저장되어 있는 뉴스, 기사 등의 온라인 데이터에 포함되어 있는 다수의 말뭉치를 전처리 하여 단어를 추출하고, 추출된 단어를 벡터 공간에 임베딩(word embedding)한 행렬을 의미한다. 따라서, 사전 학습 행렬을 이용하면 온라인 데이터에 포함된 다수의 단어가 벡터화되어 벡터 공간에 표시되므로 각 단어가 벡터 공간에서 어느 영역에 위치해 있는지 알 수 있다. 또한, 벡터 공간에서 단어 사이의 거리가 멀 지 않고, 각 단어가 비슷한 영역에 위치한 경우 단어간의 유사성이 높음을 쉽게 파악할 수 있다. 사전 학습 행 렬은 예를 들어, Word2vec 모델을 이용하여 미리 학습된 행렬일 수 있다. 이와 같이, 특허 분쟁 예측 장치(100)는 문서-단어 행렬뿐만 아니라 사전 학습 행렬을 사용함으로써 방대한 온 라인 데이터에 포함된 다수의 단어를 이용하고, 단어간의 관계를 정확하게 파악할 수 있다. 또한, 특허 분쟁 예측 장치(100)는 문서-단어 행렬 및 사전 학습 행렬을 이용하여 특허 임베딩 행렬을 산출함으 로써, 특허 데이터에 포함된 단어가 벡터 공간상에서 어디에 위치하였는지 알 수 있어 정량적인 분석을 할 수 있다. 레이블링부(150)는 특허 임베딩 행렬에 기반하여 복수의 특허 데이터를 복수의 기술 군집으로 군집화하고, 기술 군집별로 세부 기술 레이블을 부여한다. 특허 임베딩 행렬은 복수의 특허 데이터에 포함된 단어의 위치 정보를 포함하므로 레이블링부(150)는 벡터 공간 에서 복수의 특허 데이터를 복수의 기술 군집으로 군집화할 수 있다. 이에 따라 복수의 특허 데이터는 특허 데 이터가 속한 기술 분야에 따라 같

거나 서로 다른 기술 군집으로 분류될 수 있다. 여기서, 군집화는 K 평균 군집 화(K-means clustering)를 이용할 수 있다. 이후, 레이블링부(150)는 군집화된 기술 군집에 대하여 기술 군집별로 세부 기술 레이블을 부여한다. 이때, 세 부 기술 레이블은 각각의 기술 군집에 포함된 복수의 특허 데이터의 단어에 기초하여 부여될 수 있다. 따라서, 세부 기술 레이블은 기술 군집을 대표하는 문장 또는 단어로 구성될 수 있고, 기술 군집에 포함된 복수의 특허 데이터들이 어떠한 분야의 기술에 대한 특허인지 여부를 나타낼 수 있다. 구체적인 세부 기술 레이블 부여 방법 은 후술하여 자세히 설명하도록 한다. 위험예측부(170)는 기계 학습을 통해 분석 대상 특허의 특허 분쟁 위험 가능성을 산출한다. 구체적으로, 위험예측부(170)는 세부 기술 레이블 및 특허 정량 정보를 입력으로 학습된 특허 분쟁 예측 모델을 통해 복수의 특허 데이터 중 하나인 분석 대상 특허의 특허 분쟁 위험 가능성을 산출한다. 여기서, 특허 정량 정보는 복수의 특허 데이터 각각에 대한 정량적인 지표인 IPC 코드 수, 인용 수, 피인용 수, 패밀리 특허 수, 패밀리 국가 수, 청구항 수, 발명자 수, 기술이전 여부 및 출원부터 등록까지 소요된 일수 중 적어도 하나를 포함할 수 있다. 위험예측부(170)는 특허 정량 정보와 세부 기술 레이블을 입력으로 하여 특허 분쟁 예측 모델을 학습시킬 수 있 다. 위험예측부(170)는 학습된 특허 분쟁 예측 모델을 통해 분석하고자 하는 특허 데이터인 분석 대상 특허의 특허 분쟁 위험 가능성을 산출한다. 여기서, 분석 대상 특허는 복수의 특허 데이터 중 하나일 수 있고, 특허 분쟁 예측 모델은 Random Survival Forest(RSF) 모델일 수 있다. RSF 모델을 사용하는 경우 특허 데이터의 수집 시점 이전 뿐만 아니라 특허 데이 터의 수집 시점 이후에 분쟁 가능성이 있는 특허 데이터를 고려할 수 있어 시간의 흐름에 따라 변동하는 특허 분쟁 위험 가능성을 산출할 수 있다. 이하, 특허 분쟁 예측 모델은 RSF 모델임을 전제로 하여 설명하도록 한다. 도 2는 본 명세서의 일 실시예에서 텍스트 정보를 전처리하여 문서-단어 행렬을 생성하는 과정을 나타낸 순서도 이고, 도 3은 본 명세서의 일 실시예에서 특허 데이터로부터 문서-단어 행렬을 생성하는 과정을 나타낸 도면이 다. 이하, 도 2 및 도 3을 참조하여 설명하도록 한다. 도 2 및 도 3을 참조하면, DTM 생성부(110)는 복수의 특허 데이터(10)에 포함된 텍스트 정보로부터 말뭉치를 추 출한다(S112). 말뭉치(Corpus)란 컴퓨터가 텍스트를 가공,처리, 분석할 수 있도록 텍스트 정보를 모아 놓은 형태 로써, DTM 생성부(110)는 말뭉치에서 불용어, 구두점, 숫자를 제거하여 복수의 표제어(단어)를 추출한다(S114). 이후, DTM 생성부(110)는 복수의 표제어를 이용하여 문서-단어 행렬을 생성한다(S116). 문서-단어 행렬이란 복 수의 표제어 각각이 등장하는 빈도를 표현한 행렬이다. 복수의 표제어는 추출된 표제어의 수 만큼 문서-단어 행렬의 행 그룹(124)으로 표현되고, 복수의 특허 데이터 (10) 각각은 문서-단어 행렬의 열 그룹(122)으로 표현된다. 도 3을 참조하면, 특허 데이터 P1에는 표제어 sensor가 1번 등장하였고, 특허 데이터 P2에는 표제어 ai가 1번 등장하였음을 알 수 있다. 이와 같이, 문서-단어 행렬을 이용하면 각각의 특허 데이터에서 어느 단어가 많이 등 장했는지 파악할 수 있어 해당 특허 데이터의 세부 기술 파악이 용이할 수 있다. 또한, 다수의 특허 데이터를 이용하는 경우 문서-단어 행렬은 더 많은 특허 데이터 및 표제어를 포함할 수 있어 문서-단어 행렬의 사이즈는 더욱 커질 수 있고, 특허 데이터의 세부 기술을 세밀하게 파악할 수 있다. 도 4는 본 명세서의 일실시예에서 특허 임베딩 행렬을 산출하는 과정을 나타낸 순서도이고, 도 5는 본 명세서의 일실시예에서 사전 학습 행렬 및 문서-단어 행렬을 이용하여 특허 임베딩 행렬을 산출하는 과정을 나타낸 도면 이다. 도 4 및 도 5를 참조하면, 임베딩 행렬 산출부(130)는 미리 학습된 사전 학습 행렬(137)을 호출하고(S131), 사 전 학습 행렬로부터 룩업 테이블(Look-up table) 행렬(139)을 추출한다(S133). 사전 학습 행렬(136)은 상술한 바와 같이,

인터넷 서버에 저장되어 있는 뉴스, 기사 등의 온라인 데이터에 포함 되어 있는 다수의 말뭉치를 전처리하여 단어를 추출하고, 추출된 단어를 벡터 공간에 임베딩(word embedding)한 행렬을 의미한다. 도 5를 참조하면, 추출된 단어는 사전 학습 행렬의 열 그룹(138)으로 표현되고, 벡터 공간상의 차원은 사전 학 습 행렬의 행 그룹(137)으로 표현된다. 추출된 단어 각각은 특정 차원의 벡터 공간에서의 위치값을 갖는다. 예를 들어, 1차원 벡터 공간에서 단어 W1은 1482의 위치값을 갖고, 2차원 벡터 공간에서 단어 W1은 1623의 위치 값을 갖는다. 즉, 동일한 단어라도 벡터 공간의 차원에 따라 서로 다른 값을 가질 수 있다. 임베딩 행렬 산출부(130)는 사전 학습 행렬(136)에서 문서-단어 행렬에 포함된 표제어와 동일한 단어를 추출하여 룩업 테이블 행렬(139)을 생성한다. 예를 들어, 임베딩 행렬 산출부(130)는 문서-단어 행렬에 포함된 'sensor, ai, ..., chip'과 같은 표제어(단어)를 사전 학습 행렬로부터 추출하여 룩업 테이블 행렬을 구성할 수 있다. 따라서, 룩업 테이블 행렬의 열로 표현되는 단어는 문서-단어 행렬의 행으로 표현되는 표제어와 일치할 수 있다. 이후, 임베딩 행렬 산출부(130)는 룩업 테이블 행렬(139)과 문서-단어 행렬(120)의 행렬곱을 수행하여 특허 임 베딩 행렬(140)을 산출한다(S135). 특허 임베딩 행렬은 추출된 단어의 위치값을 포함하는 룩업 테이블을 이용하여 산출되므로 문서-단어 행렬에 포함된 복수의 특허 데이터의 벡터 공간에서의 위치 정보를 포함한다. 예를 들 어, 특허 임베딩 행렬(140)에서 특허 데이터 P1은 1차원 벡터 공간에서 1482의 위치값을 갖고, 2차원 벡터 공간 에서 3246의 위치값을 갖는다. 이와 같이, 특허 임베딩 행렬(140)은 복수의 특허 데이터 각각에 대한 위치 정보를 포함하므로 특허 데이터가 벡터 공간에서 어느 영역에 위치하는지 알 수 있어 복수의 특허 데이터 각각에 대한 정량적 분석이 가능하다. 또한, 본 명세서의 특허 분쟁 예측 장치는 문서-단어 행렬만을 이용하는 것이 아니라 문서-단어 행렬과 함께 방 대한 온라인 데이터에 기초한 사전 학습 행렬을 이용하여 특허 분쟁 예측 모델을 학습시키므로 정확한 분석이 가능하다. 레이블링부(150)는 특허 임베딩 행렬에 기반하여 복수의 특허 데이터를 복수의 기술 군집으로 군집화하고, 기술 군집별로 세부 기술 레이블을 부여한다. 특허 임베딩 행렬은 복수의 특허 데이터에 포함된 단어의 위치 정보를 포함하므로 레이블링부(150)는 벡터 공간 에서 복수의 특허 데이터를 복수의 기술 군집으로 군집화할 수 있다. 이에 따라 복수의 특허 데이터는 특허 데 이터가 속한 기술 분야에 따라 같거나 서로 다른 기술 군집으로 분류될 수 있다. 여기서, 군집화는 K 평균 군집 화(K-means clustering)를 이용할 수 있고, K 평균 군집화를 이용하는 경우 복수의 특허 데이터를 K개의 기술 군집으로 분류할 수 있다. 군집의 개수 K는 하기의 식 1에 의해 산출될 수 있다. 여기서, i는 기술 군집에 포함된 특허 데이터, a(i)는 같은 기술 군집에 포함된 특허 데이터들 간의 평균 거리, b(i)는 다른 기술 군집에 포함된 특허 데이터들 간의 평균 거리를 의미한다. 이후, 레이블링부(150)는 군집화된 기술 군집에 대하여 기술 군집별로 세부 기술 레이블을 부여한다. 이때, 세 부 기술 레이블은 각각의 기술 군집에 포함된 복수의 특허 데이터의 단어에 기초하여 부여될 수 있다. 따라서, 세부 기술 레이블은 기술 군집을 대표하는 문장 또는 단어로 구성될 수 있고, 기술 군집에 포함된 복수의 특허 데이터들이 어떠한 분야의 기술에 대한 특허인지 여부를 나타낼 수 있다. 보다 상세하게, 세부 기술 레이블은 복수의 특허 데이터에 포함된 단어 중 등장 횟수 상위 N개의 단어를 이용하 여 기술 군집별로 레이블 단어 리스트를 선정하고, 레이블 단어 리스트에 기초하여 기술 군집별로 세부 기술 레 이블을 부여할 수 있다. 예를 들어, 등장 횟수 상위 3개의 단어를 이용하여 레이블 단어 리스트를 선정하는 경우, 특정 기술 군집에 포 함된 복수의 특허 데이터의 등장 횟수 상위 3개의 단어가 'signal, control, detect'이면, 세부 기술 레이블은 'Object recognition technololgy'일 수

있고, 특정 기술 군집에 포함된 복수의 특허 데이터의 등장 횟수 상위 3개의 단어가 'Communic, receiv, network'이면, 세부 기술 레이블은 'signal communication technology'일 수 있다. 즉, 세부 기술 레이블은 레이블 단어 리스트에 포함된 단어의 상위 개념이거나, 단어 리스트에 포함된 단어들간 에 밀접한 연관성 있는 기술 명칭일 수 있다. 도 6은 본 명세서의 일 실시예에서 특허 정량 정보를 나타낸 표이고, 도 7은 본 명세서의 일 실시예에서 학습된 특허 분쟁 예측 모델로부터 특허 정량 정보가 특허 분쟁 위험에 영향을 미치는 정도를 수치화하여 나타낸 표이 다. 위험예측부(170)는 세부 기술 레이블과 특허 정량 정보(Quantitative Information)를 입력으로 하여 특허 분쟁 예측 모델인 RSF 모델을 학습시킨다. 여기서, 특허 정량 정보는 도 6에 도시된 바와 같이 IPC 코드 수 (all_IPC_count), 인용수(citation_count), 피인용 수(forward_citation_count), 패밀리 국가 수 (family_nation_count), 패밀리 특허 수(family_doc_count), 청구항 수(all_claim_count), 발명자 수 (inventor_count), 기술이전 여부(transfer_yn) 및 출원부터 등록까지 소요된 일수 (app_to_regi) 중 적어도 하 나를 포함할 수 있다. IPC 코드 수는 해당 특허와 연관되는 기술 분야 코드를 부여한 것으로 부여된 IPC 코드가 많을 수록 다양한 기술 에 확장 가능함 (Technology scalability)을 의미한다. 인용 수 및 피인용 수는 해당 특허의 인용 정도를 나타내 므로 해당 특허의 기술영향력(Technology impact)을 측정할 수 있다. 패밀리 국가 수 및 패밀리 특허 수는 해외 국가에 출원한 수로, 해당 특허의 시장 영향력(Market impact)을 측정할 수 있다. 청구항 수는 해당 특허의 권 리범위(Rights)에 대한 측정이 가능하며, 발명자 수는 해당 특허에 대한 지속가능한 발전(Sustainable development)여부를 판단할 수 있다. 또한, 기술이전 여부는 해당 특허의 타인에 대한 매도 여부(Utility value)를 나타내며 기술이전이 많을 수록 분석 대상 특허의 활용 가치는 높다는 것을 의미한다. 반면, 출원부터 등록까지 소요된 일수는 작을 수록 해당 특허의 활용 가치가 높다는 것을 의미한다. 한편, 본 명세서에서 학습된 특허 분쟁 예측 모델에서 특허 정량 정보 각각은 서로 다른 특허 분쟁 위험도를 갖 는다. 즉, 특허 정량 정보가 무엇이냐에 따라 특허 분쟁 위험에 영향을 주는 정도가 다르다. 도 7을 참조하면, 인용 수(citation_count)가 0.0317로 가장 높은 특허 분쟁 위험도를 갖고, 발명자 수(inventor_count)가 0.0002로 가장 낮은 특허 분쟁 위험도를 갖는다. 따라서, 복수의 특허 데이터 중 하나인 분석 대상 특허의 특허 분쟁 위험 가능성을 산출할 때, 위험예측부(17 0)는 각각의 특허 정량 정보가 갖는 특허 분쟁 위험도를 반영함으로써 보다 정확한 특허 분쟁 위험 가능성을 산 출할 수 있다. 도 8은 본 명세서의 일 실시예에 따른 특허 분쟁 예측 모델과 다른 모델과의 성능 평가표이다. 특허 분쟁 예측 모델의 특허 분쟁 예측 성능 평가를 위해 예측 오차(Prediction error)와 C-index(concordance index)의 두가지 지표가 사용될 수 있다. 도면을 참조하면, 성능 평가 결과, 본 명세서에서 RSF 모델을 사용한 특허 분쟁 예측 장치(100)는 예측 오차의 평균과 표준편차 SD(Standard Deviation)는 0.11±0.07이며, C-index의 평균과 표준편차 SD는 0.81±0.14로, 다른 모델인 KM, Cox 등의 모델과 비교할 때 낮은 예측 오차와 높은 C-index 값을 가지며 좋은 예측 성능을 보 였다. 따라서, 본 명세서에 따른 특허 분쟁 예측 장치를 이용하면 분석 대상 특허의 특허 분쟁 위험 가능성을 정확하게 산출할 수 있다. 또한, 본 명세서의 일 실시예에 따르면 RSF 모델을 특허 분쟁 예측에 사용함으로써 시간의 흐름에 따라 증감 변동하는 특허 분쟁 위험을 실시간으로 파악할 수 있다. 도 9는 본 명세서의 일 실시예에 따른 특허 분쟁 예측 방법의 순서도이다. 도면을 참조하면, 특허 분쟁 예측 장치(100)는 특허 정량 정보 및 텍스트 정보를 포함하는 복수의 특허 데이터 각각에 대하여 텍스트 정보를 전처리하여 문서-단어 행렬(DTM)을 생성한다(S200).

또한, 특허 분쟁 예측 장치(100)는 문서-단어 행렬 및 벡터 공간에서 특정 단어가 위치한 위치 정보를 포함하는 사전 학습 행렬을 이용하여 특허 임베딩 행렬을 산출한다(S210). 이후, 특허 분쟁 예측 장치(100)는 특허 임베딩 행렬에 기반하여 복수의 특허 데이터를 복수의 기술 군집으로 군집화하고, 기술 군집별로 세부 기술 레이블을 부여한다(S220). 마지막으로, 특허 분쟁 예측 장치(100)는 세부 기술 레이블 및 특허 정량 정보를 입력으로 학습된 특허 분쟁 예측 모델을 통해 복수의 특허 데이터 중 하나인 분석 대상 특허의 특허 분쟁 위험 가능성을 산출하여 특허 분쟁 을 예측한다(S230). 이상과 같이 본 발명에 대해서 예시한 도면을 참조로 하여 설명하였으나, 본 명세서에 개시된 실시 예와 도면에 의해 본 발명이 한정되는 것은 아니며, 본 발명의 기술사상의 범위 내에서 통상의 기술자에 의해 다양한 변형이 이루어질 수 있음은 자명하다. 아울러 앞서 본 발명의 실시 예를 설명하면서 본 발명의 구성에 따른 작용 효과 를 명시적으로 기재하여 설명하지 않았을지라도, 해당 구성에 의해 예측 가능한 효과 또한 인정되어야 함은 당 연하다.