(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0109296
(43) 공개일자 2023년07월20일

(51) 국제특허분류(Int. Cl.)
G06F 16/35 (2019.01) *G06F 16/26* (2019.01)
G06F 16/34 (2019.01) *G06N 3/08* (2023.01)
(52) CPC특허분류
G06F 16/35 (2019.01)
G06F 16/26 (2019.01)
(21) 출원번호 10-2022-0005088
(22) 출원일자 2022년01월13일
심사청구일자 없음

(71) 출원인
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
장동식
서울특별시 성북구 안암로9길 30, 2동 1102호
이주현
서울특별시 성북구 개운사2길 13-39, 205호
(뒷면에 계속)
(74) 대리인
김준석, 박민욱

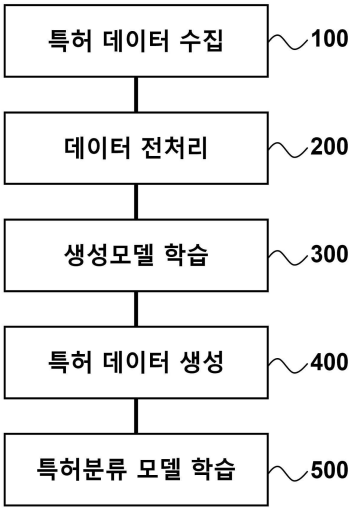
전체 청구항 수 : 총 6 항

(54) 발명의 명칭 **생성모델을 활용한 특허 분류 시스템**

(57) 요 약

본 발명은 기술이전 데이터에서 발생하는 레이블 불균형 문제 등을 해결하기 위해, 생성모델을 통해 범주가 불균형한 경우에 적용 가능한 생성모델인 conditional generative model을 활용한 특허분류 시스템에 관한 것이다. 본 발명에 따른 특허의 정량정보와 서지정보를 통해 특허분류 시스템을 생성하는데 있어서, 수집한 특허 데이터를 전처리하는 단계; 수집한 특허의 서지정보를 정형화하는 단계; 정량정보와 정형화된 서지정보를 통해 생성모델을 학습하는 단계; 생성모델을 통해 특정 범주에 속하는 특허를 생성하는 단계; 및 생성된 데이터를 추가하여 특허분류 시스템을 구축하는 단계를 포함한다.

대 표 도 - 도1



(52) CPC특허분류
G06F 16/34 (2019.01)
G06N 3/08 (2023.01)
Y10S 707/93 (2013.01)
(72) 발명자
강지호
서울특별시 용산구 효창원로37가길 7, 203호

김영호
서울특별시 성북구 동소문로 211, A동 602호

이 발명을 지원한 국가연구개발사업

과제고유번호	1415177481
과제번호	P0017311
부처명	산업통상자원부
과제관리(전문)기관명	한국산업기술진흥원
연구사업명	산업혁신인재성장지원(R&D)
연구과제명	로봇산업 혁신을 위한 글로벌인재 성장지원 프로젝트
기 여 율	100/100
과제수행기관명	고려대학교 산학협력단
연구기간	2021.03.01 ~ 2021.12.31

명 세 서

청구범위

청구항 1

특허분류 시스템을 생성하는 단계에 있어서,

특허분류 시스템을 생성하는데 필요한 특허 데이터를 수집하는 단계;

상기 특허 데이터를 특허 데이터베이스로부터 수집하고, 수집된 특허 데이터를 정량정보와 서지정보로 구분하여 서지정보를 구조화되고 분석 가능한 형태로 전처리하는 단계;

수집된 특허를 통해 생성모델을 학습하는 단계;

생성모델로부터 특허 데이터를 생성하는 단계;

상기 수집된 데이터와 생성된 데이터를 병합하여 특허분류 모델을 구축하는 단계;

를 포함하는 것을 특징으로 하는 생성모델을 활용한 특허분류 시스템.

청구항 2

제 1항에 있어서,

특허 정량정보를 활용하기 위해서, 특허의 청구항 수, 피인용 수, 패밀리 국가 수, 패밀리 특허 수, IPC 코드 수 외 특허를 정량적으로 평가하기 위한 정보를 포함하는 특허분류 시스템.

청구항 3

제 1항에 있어서,

상기 도출한 특허 정량정보와 구조화된 서지정보를 이용한 특허분류 시스템을 생성하는 단계는, 정량정보와 서지정보를 변수로 하는 생성모델을 생성하며, 생성모델은 AutoEncoder, Variational AutoEncoder, Conditional-Variational AutoEncoder, Generative Adversarial Networks를 포함하는 알고리즘에 기반한 방법.

청구항 4

제 1항에 있어서,

상기 생성모델을 통해 생성한 데이터를 활용해 특허를 분류하는 모델은 로지스틱 회귀모델, 의사결정 나무모델, 랜덤포레스트모델을 포함하는 알고리즘에 기반한 방법.

청구항 5

제 4항에 있어서,

상기 모델은 데이터의 수가 적은 범주를 결정하고, 이를 생성하기 위한 모델을 학습하고, 학습된 모델을 통해 데이터를 생성하고 방법.

청구항 6

제 4항에 있어서,

상기 모델은 생성된 데이터와 수집한 데이터를 병합하고, 병합된 데이터를 통해 특허를 분류하는 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 특허분류를 위한 시스템에 관한 것이며, 보다 상세하게는 특허를 분류하는 범주의 수가 불균형할 때

발생하는 문제점을 해결하기 위해 생성모델을 활용하는 시스템에 관한 것이다.

배 경 기 술

[0002] 특허는 연구자가 개발한 기술을 권리화하기 위한 제도이다. 기술 시장에 진입하기 위한 기업이나 연구소는 다른 특허의 권리를 침해하지 않도록 기존에 등록된 특허를 검토한다. 그러나 이미 등록된 특허의 기술성이 우수한 경우, 이를 회피하는 것은 많은 시간과 비용이 소모된다. 기술이전은 이러한 경우에 고려할 수 있는 전략이다. 기술이전은 권리자가 보유한 특허를 양도하는 것을 의미한다.

[0003] 최근, 기술이전의 중요성이 증가함에 따라, 특허 빅데이터를 기반으로 기술이전을 예측하는 연구가 활발히 진행중이며 이를 위해, 기술이전된 특허를 선별하거나, 가치가 높게 평가된 특허만을 선택하여 우선적으로 검토하는 방법이 활용되고 있다.

[0004] 특허분류는 특허의 특징을 반영하는 범주에 할당하는 것이 목적이다. 예를 들어, 국제특허분류 코드는 특허의 기술적인 특징에 따라 다수의 범주로 분류한다. 또한, 기술이전 여부에 따라 특허를 분류할 수 있다.

[0005] 최근에는 기계학습 모델을 통한 특허분류의 연구가 주로 수행되고 있다. 특허출원 제10-2019-0078907호의 경우, 기술이전된 특허와 그렇지 않은 특허의 분류를 위한 기계학습 모델을 제안했다. 그러나 범주의 수가 불균형한 경우, 예측의 결과가 단일 범주에 편향될 수 있는 문제가 있다.

[0006] 그러나 이러한 특허의 분류는 특정한 범주의 수가 매우 적은 상황이 발생할 수 있다. 예를 들어, 기술이전된 특허는 그렇지 않은 데이터보다 그 양이 매우 적다. 또한, 미래 가치가 높게 평가된 특허는 그렇지 않은 특허보다 그 양이 매우 적다. 이 경우, 범주의 불균형 문제가 발생한다. 범주가 불균형한 경우, 분류모델은 특정 범주에 편향되어 학습될 수 있다.

발명의 내용

해결하려는 과제

[0007] 본 발명은 기술이전 데이터에서 발생하는 레이블 불균형 문제 등을 해결하기 위해, 생성모델을 통해 범주가 불균형한 경우에 적용 가능한 생성모델인 conditional generative model을 활용한 특허분류 시스템을 제공하고자 한다.

[0008] 본 발명은 선행문헌과 다르게 생성모델을 활용하여 범주의 수가 적은 특허를 생성모델로 생성하여 예측성능의 향상을 도모하는 것에 그 차이가 있다.

과제의 해결 수단

[0009] 상기 목적을 달성하기 위한 본 발명에 따른 시스템은, 특허의 정량정보와 서지정보를 통해 특허분류 시스템을 생성하는데 있어서, 수집한 특허 데이터를 전처리하는 단계; 수집한 특허의 서지정보를 정형화하는 단계; 정량정보와 정형화된 서지정보를 통해 생성모델을 학습하는 단계; 생성모델을 통해 특정 범주에 속하는 특허를 생성하는 단계; 및 생성된 데이터를 추가하여 특허분류 시스템을 구축하는 단계를 포함한다.

[0010] 또한, 본 발명에서는 비정형화된 특허의 서지정보를 구조화된 행렬 형태로 표현하는 방법에 있어서, 특허 문서별 단어의 빈도나 tf-idf 또는 임베딩된 값을 통해 구조화하는 단계를 포함한다.

발명의 효과

[0011] 본 발명은 생성모델을 통해 범주의 수가 적은 특허를 생성할 수 있으며, 이를 활용하여 기계학습 모델의 학습을 위해 제공할 수 있는 데이터의 수를 증가할 수 있다.

[0012] 본 발명은 기술이전 여부를 분류하는 효과를 얻을 수 있다. 기술이전된 특허와 그렇지 않은 특허를 수집하고, 수집된 데이터를 생성모델에 적용하고, 생성모델을 통해 상대적으로 수가 적은 기술이전된 특허의 특징을 갖는 데이터를 생성하고, 이를 기계학습 모델의 학습에 적용하면 특정 범주에 편향되지 않은 예측의 효과를 얻을 수 있다.

[0013] 본 발명은 기술가치를 평가하는 효과를 얻을 수 있다. 기술의 가치가 높게 평가된 특허와 그렇지 않은 특허를 수집하고, 수집된 데이터를 생성모델에 적용하고, 생성모델을 통해 상대적으로 수가 적은 가치가 높은 특허의 특징을 갖는 데이터를 생성하고, 이를 기계학습 모델의 학습에 적용하면 특정 범주에 편향되지 않은 예측의 효

과를 얻을 수 있다.

도면의 간단한 설명

[0014] 도 1은 본 발명인 특허분류 시스템의 흐름도를 나타낸 것이다.

도 2는 본 발명에서 로지스틱 회귀모델로 test set에 대해 기술이전 여부를 예측한 결과를 나타낸 것이다.

도 3은 본 발명의 특허분류 시스템에 사용된 방법의 흐름도를 나타낸 것이다.

도 4는 본 발명에서 기술이전된 조건에 대해, C-VAE의 Decoder에 표준 정규분포로부터 추출한 난수를 입력하여 디코딩된 결과를 나타낸 것이다.

발명을 실시하기 위한 구체적인 내용

[0015] 본 발명은 conditional generative model을 통해 레이블이 불균형한 텍스트 데이터를 생성하는 방법을 제안한다. 기존의 생성모델은 레이블에 관계없이 랜덤한 데이터를 생성하는 한계가 있다. 그러나 conditional generative model은 특정 조건에 부합한 데이터를 선별하여 생성할 수 있는 장점이 있다.

[0016] 본 발명은 특허의 기술이전 데이터에서 발생하는 레이블 불균형 문제를 해결하기 위해 conditional generative model을 활용한다. 제안된 방법은 기술이전된 경우의 데이터만을 생성할 수 있다. 실험은 실제 적용 가능성을 검토하기 위해 바이오 로봇 관련 9,977건을 사용하여 진행했다. 실험 결과, conditional generative model로 생성된 데이터를 추가한 경우, 기술이전 여부를 분류하는 모델의 성능이 향상된 것을 확인할 수 있었다.

[0017] Kingma “Auto-encoding variational bayes”(2013)는 입력 데이터로부터 산출된 평균과 표준 편차를 모수로 갖는 정규분포에서 잠재변수를 추출하는 variational auto-encoder (VAE)를 제안했다. VAE는 디코더로 데이터의 종속변수와 관계없이 잠재변수를 입력 데이터의 공간으로 디코딩한다. 그러면 디코딩된 잠재변

[0018] 수는 입력 데이터와 유사한 확률분포를 따른다.

[0019] Conditional-VAE (C-VAE)는 VAE를 기반으로 분석자가 지정한 조건에 맞는 데이터를 생성하는 방법이다. 예를 들어, 0과 1을 손글씨로 작성한 이미지가 있을 때, 이미지들의 레이블은 0과 1이다. 이 경우, VAE는 잠재변수에 따라 랜덤하게 이미지를 생성하는 반면, C-VAE는 특정 레이블에 대한 이미지만 선택해서 생성할 수 있다.

[0020] 도1은 본 발명인 특허분류 시스템의 실시예 흐름도이다. 도면에 도시된 바와 같이, 본 발명에 따른 특허분류 시스템은, 특허 데이터를 수집하는 단계(100)와, 상기 수집한 특허 데이터(100)를 구조화되고 분석 가능한 형태로 가공하는 전처리단계(200)로 구성되며 생성모델을 학습하는 단계(300)와 이를 통해 특허 데이터를 생성(400)하고 이를 수집한 특허 데이터(100)과 결합하여 특허분류 모델(500)을 학습 및 구축하는 단계로 구성된다.

[0021] 본 발명은 C-VAE 모델을 기반으로 기술이전된 특허를 생성한다. 도 3은 본 발명의 특허분류 시스템에 사용된 방법의 흐름도이다. 제안된 방법은 특허 Database로부터 기술이전 이력이 있는 데이터를 수집한다. 그리고 수집된 데이터의 텍스트는 pre-trained language model로 context 공간에 임베딩된다.

[0022] 임베딩된 텍스트는 벡터로 변환되므로 C-VAE에 입력으로 사용할 수 있다. 그리고 기술이전된 경우를 조건으로 설정하여 C-VAE로부터 데이터를 생성할 수 있다. 생성된 데이터는 기존의 데이터와 합쳐져 기술이전 여부를 예측하는 모델의 학습에 사용된다.

[0023] [실시예]

[0024] 실시예는 바이오 로봇 특허 9,977건으로 기술이전 여부를 포함한 데이터에 관한 본 발명의 적용 사례를 포함한다. 실시예를 위해 수집된 데이터는 training, validation, test set으로 분할했다. 실시예에서는 pre-trained language model로 특허의 텍스트를 384차원으로 임베딩했다. 임베딩된 training과 validation set은 조건부 생성모델 (Conditional-Variational AutoEncoder, C-VAE)의 학습에 사용된다. 생성모델은 특허의 기술이전 여부에 따라 데이터를 생성한다. 실시예는 기술이전된 조건에 대해, C-VAE의 Decoder에 표준 정규분포로부터 추출한 난수를 입력하여 데이터를 생성한다.

[0025] 도 4는 기술이전된 조건에 대해, C-VAE의 Decoder에 표준 정규분포로부터 추출한 난수를 입력하여 디코딩된 결과를 보여준다. 생성된 데이터의 개수는 validation set 중 기술이전된 관측치의 절반인 24개이다. 도 4는 validation set과 생성된 데이터를 차원축소한 결과이다.

[0026] 도 2는 로지스틱 회귀모델로 test set에 대해 기술이전 여부를 예측한 결과이다. 로지스틱 회귀모델에 적용한 데이터는 두 개이다. 첫 번째는 원래 수집된 데이터이다. 그리고 두 번째는 첫 번째 데이터에 C-VAE로부터 기술이전된 조건하에 생성된 관측치를 더한 결과이다. 로지스틱 회귀모델의 성능은 Accuracy, Precision, Recall, F1-score로 측정했다. 실시예에서는 원래 수집된 데이터만을 학습한 모델보다, 기술이전된 조건하에 생성된 데이터를 추가한 경우의 성능이 향상되었다.

[0027] 본 발명은 특허의 기술이전 여부를 분류하는 모델을 구축할 때 발생하는 레이블 불균형 문제를 개선하기 위해 C-VAE를 활용한 방법을 제안했다. 제안된 방법은 특허의 텍스트를 pre-trained language model로 임베딩한 뒤, context 벡터를 활용하여 기술이전된 조건하의 데이터를 생성한다. 실험을 통해 생성모델로 표본의 크기를 늘린 경우, 레이블 불균형 문제가 해소되는 것을 확인할 수 있었다.

[0028] 이상 첨부된 도면을 참조하여 본 발명의 실시예들을 설명하였으나 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자는 본 발명이 그 기술적 사상이나 필수적인 특징을 변경하지 않고서 다른 구체적인 형태로 실시될 수 있다는 것을 이해할 수 있을 것이다. 그러므로 이상에서 기술한 실시예들은 모든 면에서 예시적인 것이며 한정적이 아닌 것으로 이해해야만 한다.

부호의 설명

[0029] 100 : 특허 데이터 수집부

200 : 특허 데이터의 전처리부

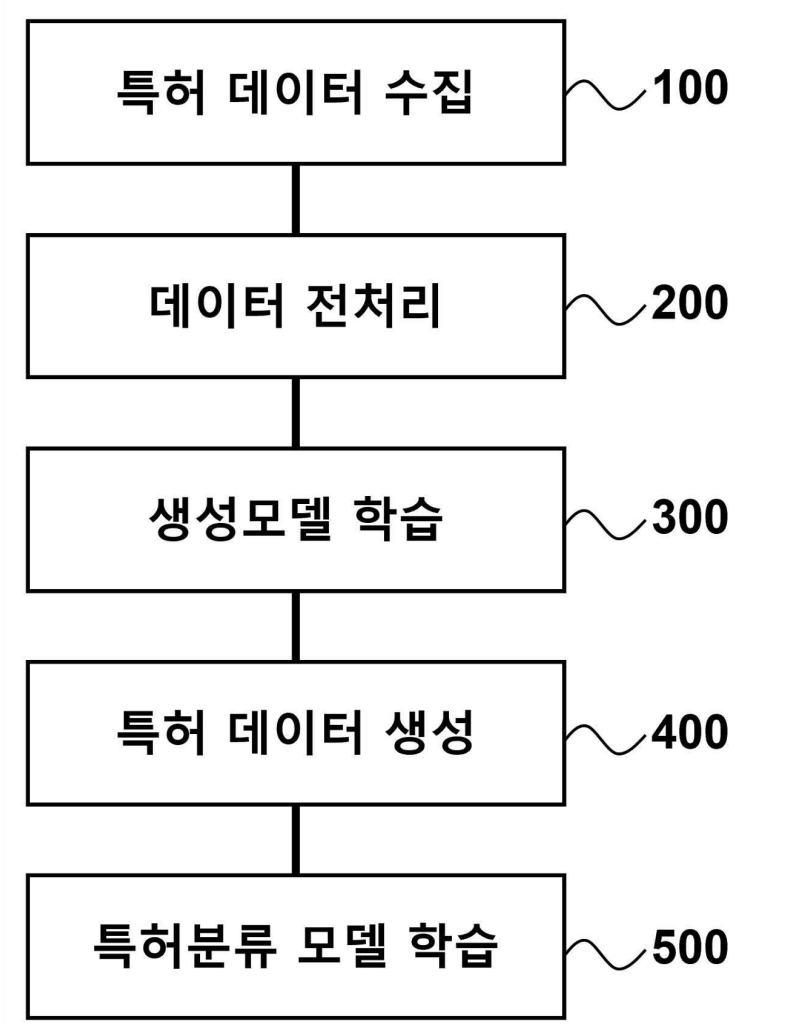
300 : 생성모델을 학습하는 산출부

400 : 생성모델을 통해 데이터를 생성하는 생성부

500 : 특허분류 모델을 학습하는 산출부

도면

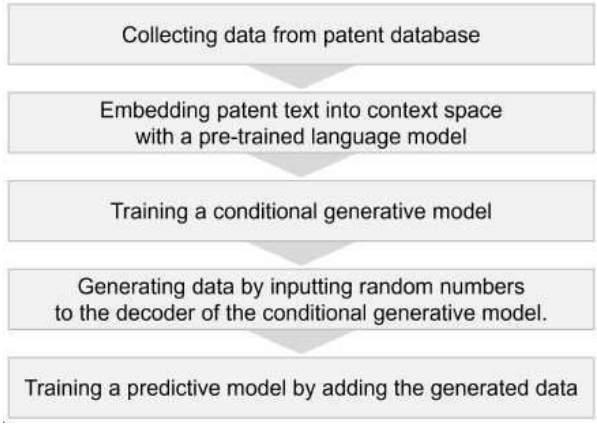
도면1



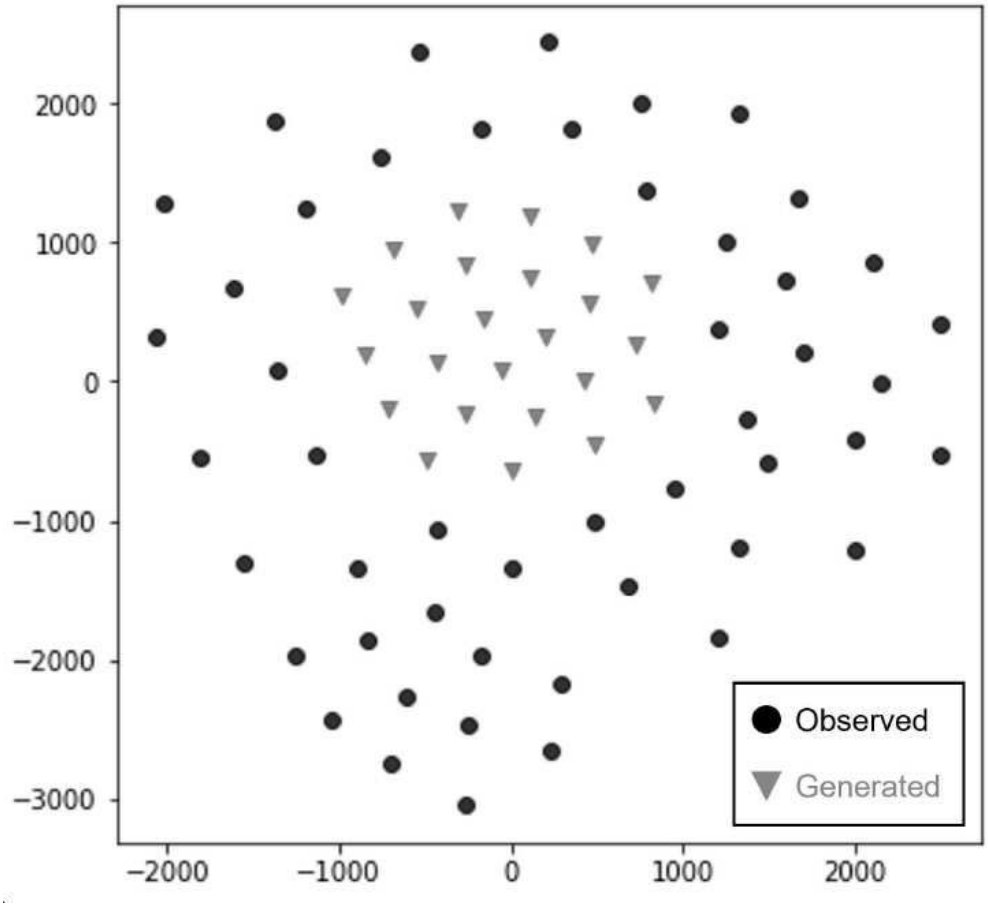
도면2

Parameter	Raw data	Raw data + Generated data
Accuracy	0.96	0.96
Precision	0.42	0.46
Recall	0.21	0.30
F1-score	0.28	0.36

도면3


Collecting data from patent database
Embedding patent text into context space with a pre-trained language model
Training a conditional generative model
Generating data by inputting random numbers to the decoder of the conditional generative model.
Training a predictive model by adding the generated data

도면4


2000
1000
0
−1000
−2000
−3000
−2000
−1000
0
1000
2000
Observed
Generated