



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2022-0160407
(43) 공개일자 2022년12월06일

(51) 국제특허분류(Int. Cl.)

G16H 50/50 (2018.01) G06N 20/00 (2019.01)
G06N 3/04 (2006.01) G16B 40/00 (2019.01)
G16H 10/60 (2018.01) G16H 50/70 (2018.01)
G16H 70/60 (2018.01)

(52) CPC특허분류

G16H 50/50 (2018.01)
G06N 20/00 (2021.08)

(21) 출원번호 10-2021-0068615

(22) 출원일자 2021년05월27일

심사청구일자 2021년05월27일

(71) 출원인

재단법인 아산사회복지재단
서울특별시 송파구 올림픽로43길 88 (풍납동)
울산대학교 산학협력단
울산광역시 남구 대학로 93(무거동)

(72) 발명자

김영학
서울특별시 강남구 남부순환로 2914, 3동 1102호
(대치동, 개포1차우성아파트)

전대준

서울특별시 송파구 풍성로6길 15, 101동 708호 (풍납동, 신동아아파트)
(뒷면에 계속)

(74) 대리인

특허법인 무한

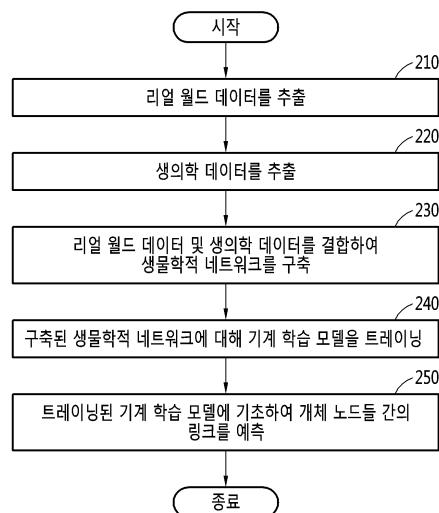
전체 청구항 수 : 총 14 항

(54) 발명의 명칭 생의학적 연관성 예측 방법 및 장치

(57) 요약

일 실시예에 따른 컴퓨팅 장치에 의해 수행되는 생의학적 연관성 예측 방법은, 리얼 월드 데이터베이스로부터 환자 정보와 연결된 진단 코드 및 처방 코드를 포함하는 리얼 월드 데이터를 추출하는 단계; 생의학 데이터베이스로부터 유전자 정보, 질병 식별자, 화학물질 식별자, 및 모구조 식별자를 포함하는 생의학 데이터를 추출하는 단계; 리얼 월드 데이터 및 생의학 데이터를 결합함으로써 새로운 생물학적 네트워크를 구축하는 단계; 구축된 새로운 생물학적 네트워크로부터 랜덤워크에 기반하여 생성된 랜덤워크 데이터를 이용하여, 기계 학습 모델을 트레이닝시키는 단계; 및 트레이닝된 기계 학습 모델에 기초하여 새로운 생물학적 네트워크의 개체 노드의 임베딩 벡터에 기초하여 개체 노드들 간의 링크(link)를 예측하는 단계를 포함할 수 있다.

대표도 - 도2



(52) CPC특허분류

G06N 3/0427 (2013.01)

G16B 40/00 (2019.02)

G16H 10/60 (2021.08)

G16H 50/70 (2018.01)

G16H 70/60 (2021.08)

(72) 발명자

김윤하

서울특별시 성동구 서울숲길 25, 102동 304호 (성수동1가, 현대아파트)

정연욱

부산광역시 금정구 금강로379번길 50, 102동 504호 (장전동, 금정산 쌍용예가 아파트)

안임진

경기도 의정부시 진등로19번길 20, 901동 1207호 (녹양동, 녹양동 신도9차 전원마을아파트)

권한슬

서울특별시 노원구 초안산로5길 22, 105동 202호 (월계동, 대우아파트)

강희준

경기도 성남시 중원구 둔촌대로 300, 1004동 1004호 (하대원동, 아튼빌아파트)

조승주

서울특별시 구로구 중앙로 121, 109동 302호 (고척동, 고척파크푸르지오)

이 발명을 지원한 국가연구개발사업

과제고유번호 9991006770

과제번호 202012B06-01

부처명 다부처

과제관리(전문)기관명 (재단)범부처전주기의료기기연구개발사업단

연구사업명 범부처전주기의료기기연구개발사업(R&D)

연구과제명 (주관)EMR기반 환자중심 진단치료 지원용 초지능 의료기기 소프트웨어 개발 및 실증

기 여 율 1/1

과제수행기관명 서울아산병원

연구기간 2020.09.01 ~ 2021.02.28

명세서

청구범위

청구항 1

컴퓨팅 장치에 의해 수행되는 생의학적 연관성(biomedical association) 예측 방법에 있어서,

리얼 월드 데이터베이스로부터 환자 정보와 연결된 진단 코드 및 처방 코드를 포함하는 리얼 월드 데이터를 추출하는 단계;

생의학 데이터베이스로부터 유전자 정보, 질병 식별자, 화학물질 식별자, 및 모구조 식별자를 포함하는 생의학 데이터를 추출하는 단계;

상기 리얼 월드 데이터 및 상기 생의학 데이터를 결합함으로써 새로운 생물학적 네트워크를 구축하는 단계;

상기 구축된 새로운 생물학적 네트워크로부터 기계 학습 모델을 트레이닝시키는 단계; 및

상기 트레이닝된 기계 학습 모델에 기초하여 상기 새로운 생물학적 네트워크의 개체 노드의 임베딩 벡터에 기초하여 개체 노드들 간의 링크(link)를 예측하는 단계

를 포함하는 생의학적 연관성 예측 방법.

청구항 2

제1항에 있어서,

상기 리얼 월드 데이터를 추출하는 단계는,

환자 정보를 지시하는 환자 노드를 상기 환자 정보에 대응하여 추출된 처방 코드를 지시하는 처방 노드와 진단 코드를 지시하는 진단 노드에 엣지를 통해 연결함으로써, 그래프 구조의 상기 리얼 월드 데이터를 생성하는 단계

를 포함하는 생의학적 연관성 예측 방법.

청구항 3

제1항에 있어서,

상기 생의학 데이터를 추출하는 단계는,

유전자 정보를 지시하는 유전자 노드를 상기 유전자 정보에 대응하는 질병 식별자는 지시하는 질병 노드에 엣지를 통해 연결하는 단계;

상기 질병 노드를 상기 질병 식별자에 대응하는 화학물질 식별자를 지시하는 화학물질 노드에 엣지를 통해 연결하는 단계;

상기 화학물질 노드를 상기 화학물질 식별자에 대응하는 모구조 식별자를 지시하는 모구조 노드에 엣지를 통해 연결하는 단계; 및

엣지를 통해 연결된 상기 유전자 노드, 상기 질병 노드, 상기 화학물질 노드, 및 상기 모구조 노드를 포함하는 그래프 구조의 생의학 데이터를 생성하는 단계

를 포함하는 생의학적 연관성 예측 방법.

청구항 4

제3항에 있어서,

상기 생의학 데이터에서 상기 질병 노드는 엣지를 통해 상기 화학물질 노드를 경유하여 상기 모구조 노드와 연결되고, 상기 화학물질 노드는 엣지를 통해 상기 질병 노드를 경유하여 상기 유전자 노드와 연결되는, 생의학적 연관성 예측 방법.

청구항 5

제1항에 있어서,

상기 새로운 생물학적 네트워크를 구축하는 단계는,

상기 리얼 월드 데이터의 진단 노드 및 상기 생의학 데이터의 질병 노드를 연결하고 상기 리얼 월드 데이터의 처방 노드 및 상기 생의학 데이터의 화학물질 노드를 연결함으로써 상기 새로운 생물학적 네트워크를 구축하는 단계

를 포함하는 생의학적 연관성 예측 방법.

청구항 6

제5항에 있어서,

상기 새로운 생물학적 네트워크를 구축하는 단계는,

생의학 용어(biomedical terminology)에 관한 공통 데이터 모델에 기초하여 동일한 고유 개념 식별자에 대응하는 상기 진단 노드 및 상기 질병 노드끼리 매핑하는 단계; 및

상기 공통 데이터 모델에 기초하여 동일한 고유 개념 식별자에 대응하는 상기 처방 노드 및 상기 화학물질 노드끼리 연결하는 단계

를 포함하는 생의학적 연관성 예측 방법.

청구항 7

제5항에 있어서,

상기 새로운 생물학적 네트워크를 구축하는 단계는,

상기 진단 노드에 의해 지시되는 진단 코드에 대응하는 제1 문자열을 추출하는 단계;

공통 데이터 모델로부터 상기 추출된 제1 문자열을 가지는 제1 개념 고유 식별자를 검색하는 단계;

상기 검색된 제1 개념 고유 식별자에 대응하는 질병 식별자를 지시하는 상기 질병 노드에 상기 진단 노드를 매핑하는 단계;

를 포함하는 생의학적 연관성 예측 방법.

청구항 8

제5항에 있어서,

상기 새로운 생물학적 네트워크를 구축하는 단계는,

상기 처방 노드에 의해 지시되는 처방 코드에 대응하는 제2 문자열을 추출하는 단계;

공통 데이터 모델로부터 상기 추출된 제2 문자열에 대응하는 제2 개념 고유 식별자를 검색하는 단계; 및

상기 검색된 제2 개념 고유 식별자에 대응하는 화학물질 식별자를 지시하는 상기 화학물질 노드에 상기 처방 노

드를 매핑하는 단계
를 포함하는 생의학적 연관성 예측 방법.

청구항 9

제1항에 있어서,
상기 기계 학습 모델을 트레이닝시키는 단계는,
상기 구축된 새로운 생물학적 네트워크에서 초기 개체 노드를 선정하는 단계;
상기 선정된 초기 개체 노드로부터 미리 결정된 랜덤워크 길이만큼 순차적으로 랜덤워크를 수행함으로써, 개별 랜덤워크에 따른 개체 노드를 선택하는 단계; 및
상기 초기 개체 노드 및 상기 랜덤워크에 기초하여 선택된 개체 노드를 포함하는 랜덤워크 셋트를 지시하는 랜덤워크 데이터를 생성하는 단계
를 포함하는 생의학적 연관성 예측 방법.

청구항 10

제9항에 있어서,
상기 기계 학습 모델을 트레이닝시키는 단계는,
상기 랜덤워크 데이터에 기초한 입력 데이터에 상기 기계 학습 모델을 적용함으로써 임시 임베딩 데이터를 산출하는 단계; 및
상기 산출된 임시 임베딩 데이터에 기초하여 상기 랜덤워크 셋트 내 개체 노드들 간의 연결 확률이 최대화되도록 기계 학습 모델의 파라미터를 업데이트하는 단계
를 더 포함하는 생의학적 연관성 예측 방법.

청구항 11

제1항에 있어서,
상기 개체 노드들 간의 링크(link)를 예측하는 단계는,
상기 새로운 생물학적 네트워크의 개체 노드들 중 제1 노드에 대응하는 제1 임베딩 벡터를 상기 기계 학습 모델로부터 추출하는 단계;
상기 새로운 생물학적 네트워크의 개체 노드들 중 제2 노드에 대응하는 제2 임베딩 벡터를 상기 기계 학습 모델로부터 추출하는 단계; 및
상기 추출된 제1 임베딩 벡터 및 상기 제2 임베딩 벡터에 기초하여, 상기 제1 노드 및 상기 제2 노드 간의 엣지 존재 여부를 결정하는 단계
를 포함하는 생의학적 연관성 예측 방법.

청구항 12

제11항에 있어서,
상기 엣지 존재 여부를 결정하는 단계는,
상기 추출된 제1 임베딩 벡터 및 상기 제2 임베딩 벡터 간의 벡터 거리 차이가 임계 거리 값 미만인 경우에 응

답하여, 상기 제1 노드 및 상기 제2 노드 간에 엣지가 존재한다고 결정하는 단계
를 포함하는 생의학적 연관성 예측 방법.

청구항 13

하드웨어와 결합되어 제1항 내지 제12항 중 어느 하나의 항의 방법을 실행시키기 위하여 컴퓨터 판독 가능한 기록매체에 저장된 컴퓨터 프로그램.

청구항 14

컴퓨팅 장치에 있어서,

생의학적 연관성 예측을 위한 기계 학습 모델을 저장한 메모리;

리얼 월드 데이터베이스로부터 환자 정보와 연결된 진단 코드 및 처방 코드를 포함하는 리얼 월드 데이터를 추출하고, 생의학 데이터베이스로부터 유전자 정보, 질병 식별자, 화학물질 식별자, 및 모구조 식별자를 포함하는 생의학 데이터를 추출하며, 상기 리얼 월드 데이터 및 상기 생의학 데이터를 결합함으로써 새로운 생물학적 네트워크를 구축하고, 상기 구축된 새로운 생물학적 네트워크로부터 기계 학습 모델을 트레이닝시키며, 상기 트레이닝된 기계 학습 모델에 기초하여 상기 새로운 생물학적 네트워크의 개체 노드의 임베딩 벡터에 기초하여 개체 노드들 간의 링크(link)를 예측하는 프로세서

를 포함하는 컴퓨팅 장치.

발명의 설명

기술 분야

[0001] 이하, 생의학적 연관성 예측 방법 및 장치에 관한 기술이 제공된다.

배경 기술

[0002] 전자 장치의 발전과 체계적인 전자 기록 수집이 보편 화되면서 환자 임상데이터가 전자 의무 기록(Electric Medical Record, EMR)의 형태로 축적되고 있다. EMR은 변수가 통제된 시험 환경에서의 임상 데이터가 아닌 다양한 변수들이 복합적으로 포함되는 실세계에서의 데이터로서, 리얼 월드 데이터(Real World Data, RWD)에 해당한다. RWD를 분석하여 얻을 수 있는 의약품 등의 사용 및 잠재적인 유익성과 위해성에 관한 임상적인 증거를 리얼 월드 에비던스(Real World Evidence, RWE)라고 한다. 최근의 임상 연구에서는 RWE를 근거로 하여 환자의 생물학적 현상이나 실제 임상에서의 사건을 예측하거나 설명하려는 시도들이 다각도로 연구되고 있다.

[0003] 질병의 복잡한 생명활동을 설명하고 질병과 약물의 관계성을 예측하려는 연구들이 진행되고 있다. 분자 생물학(Molecular biology), 유전체학(genomics) 등 다양한 필드에서도 각각의 관련 물질에 대한 오픈 생의학 데이터베이스들이 구축되었고 생명활동의 복잡한 현상을 이해하려는 연구에 사용된다. 생의학 데이터베이스 기반 관계성 연구는 유전자(gene), 단백질(protein), 장기(organs), 약물(drug) 및 질병(disease)과 같은 관련 요소(element)들의 상호작용과 이들이 이루는 하나의 큰 맥락(context)을 이해하는 새로운 접근법이 될 수 있다.

[0004] 위에서 설명한 배경기술은 발명자가 본원의 개시 내용을 도출하는 과정에서 보유하거나 습득한 것으로서, 반드시 본 출원 전에 일반 공중에 공개된 공지기술이라고 할 수는 없다.

발명의 내용

해결하려는 과제

[0006] 일 실시예에 따른 컴퓨팅 장치는 생의학 데이터베이스에 축적된 생물학적, 화학적 지식과 전자의무기록의 임상

지식을 결합할 수 있다.

- [0007] 일 실시예에 따른 컴퓨팅 장치는 그래프 구조로 결합 구축된 데이터베이스에서 그래프 임베딩 기술을 통한 생의학적 연관성을 탐색할 수 있다.
- [0008] 일 실시예에 따른 컴퓨팅 장치는 전자의무기록의 환자 및 테이블 추가에 따른 데이터베이스 구축을 할 수 있다.
- [0009] 일 실시예에 따른 컴퓨팅 장치는 생의학 데이터베이스 추가에 따른 데이터베이스를 구축할 수 있다.
- [0010] 다만, 기술적 과제는 상술한 기술적 과제들로 한정되는 것은 아니며, 또 다른 기술적 과제들이 존재할 수 있다.

과제의 해결 수단

- [0011] 일 실시예에 따른 컴퓨팅 장치에 의해 수행되는 생의학적 연관성 예측 방법은, 리얼 월드 데이터베이스로부터 환자 정보와 연결된 진단 코드 및 처방 코드를 포함하는 리얼 월드 데이터를 추출하는 단계; 생의학 데이터베이스로부터 유전자 정보, 질병 식별자, 화학물질 식별자, 및 모구조 식별자를 포함하는 생의학 데이터를 추출하는 단계; 상기 리얼 월드 데이터 및 상기 생의학 데이터를 결합함으로써 새로운 생물학적 네트워크를 구축하는 단계; 상기 구축된 새로운 생물학적 네트워크로부터 기계 학습 모델을 트레이닝시키는 단계; 및 상기 트레이닝된 기계 학습 모델에 기초하여 상기 새로운 생물학적 네트워크의 개체 노드의 임베딩 벡터에 기초하여 개체 노드들 간의 링크(link)를 예측하는 단계를 포함할 수 있다.
- [0012] 상기 리얼 월드 데이터를 추출하는 단계는, 환자 정보를 지시하는 환자 노드를 상기 환자 정보에 대응하여 추출된 처방 코드를 지시하는 처방 노드와 진단 코드를 지시하는 진단 노드에 엣지를 통해 연결함으로써, 그래프 구조의 상기 리얼 월드 데이터를 생성하는 단계를 포함할 수 있다.
- [0013] 상기 생의학 데이터를 추출하는 단계는, 유전자 정보를 지시하는 유전자 노드를 상기 유전자 정보에 대응하는 질병 식별자를 지시하는 질병 노드에 엣지를 통해 연결하는 단계; 상기 질병 노드를 상기 질병 식별자에 대응하는 화학물질 식별자를 지시하는 화학물질 노드에 엣지를 통해 연결하는 단계; 상기 화학물질 노드를 상기 화학물질 식별자에 대응하는 모구조 식별자를 지시하는 모구조 노드에 엣지를 통해 연결하는 단계; 및 엣지를 통해 연결된 상기 유전자 노드, 상기 질병 노드, 상기 화학물질 노드, 및 상기 모구조 노드를 포함하는 그래프 구조의 생의학 데이터를 생성하는 단계를 포함할 수 있다.
- [0014] 상기 생의학 데이터에서 상기 질병 노드는 엣지를 통해 상기 화학물질 노드를 경유하여 상기 모구조 노드와 연결되고, 상기 화학물질 노드는 엣지를 통해 상기 질병 노드를 경유하여 상기 유전자 노드와 연결될 수 있다.
- [0015] 상기 새로운 생물학적 네트워크를 구축하는 단계는, 상기 리얼 월드 데이터의 진단 코드 및 상기 생의학 데이터의 질병 코드를 연결하고 상기 리얼 월드 데이터의 처방 코드 및 상기 생의학 데이터의 화학물질 코드를 연결함으로써 상기 새로운 생물학적 네트워크를 구축하는 단계를 포함할 수 있다.
- [0016] 상기 새로운 생물학적 네트워크를 구축하는 단계는, 생의학 용어(biomedical terminology)에 관한 공통 데이터 모델에 기초하여 동일한 고유 개념 식별자에 대응하는 상기 진단 코드 및 상기 질병 코드끼리 매핑하는 단계; 및 상기 공통 데이터 모델에 기초하여 동일한 고유 개념 식별자에 대응하는 상기 처방 코드 및 상기 화학물질 코드끼리 연결하는 단계를 포함할 수 있다.
- [0017] 상기 새로운 생물학적 네트워크를 구축하는 단계는, 상기 진단 코드에 의해 지시되는 진단 코드에 대응하는 제1 문자열을 추출하는 단계; 공통 데이터 모델로부터 상기 추출된 제1 문자열을 가지는 제1 개념 고유 식별자를 검색하는 단계; 상기 검색된 제1 개념 고유 식별자에 대응하는 질병 식별자를 지시하는 상기 질병 코드에 상기 진단 코드를 매핑하는 단계를 포함할 수 있다.
- [0018] 상기 새로운 생물학적 네트워크를 구축하는 단계는, 상기 처방 코드에 의해 지시되는 처방 코드에 대응하는 제2 문자열을 추출하는 단계; 공통 데이터 모델로부터 상기 추출된 제2 문자열에 대응하는 제2 개념 고유 식별자를 검색하는 단계; 및 상기 검색된 제2 개념 고유 식별자에 대응하는 화학물질 식별자를 지시하는 상기 화학물질 코드에 상기 처방 코드를 매핑하는 단계를 포함할 수 있다.
- [0019] 상기 기계 학습 모델을 트레이닝시키는 단계는, 상기 구축된 새로운 생물학적 네트워크에서 초기 개체 노드를 선정하는 단계; 상기 선정된 초기 개체 노드로부터 미리 결정된 랜덤워크 길이만큼 순차적으로 랜덤워크를 수행함으로써, 개별 랜덤워크에 따른 개체 노드를 선택하는 단계; 및 상기 초기 개체 노드 및 상기 랜덤워크에 기초하여 선택된 개체 노드를 포함하는 랜덤워크 셋트를 지시하는 랜덤워크 데이터를 생성하는 단계를 포함할 수 있다.

다.

- [0020] 상기 기계 학습 모델을 트레이닝시키는 단계는, 상기 랜덤워크 데이터에 기초한 입력 데이터에 상기 기계 학습 모델을 적용함으로써 임시 임베딩 데이터를 산출하는 단계; 및 상기 산출된 임시 임베딩 데이터에 기초하여 상기 랜덤워크 셋트 내 개체 노드들 간의 연결 확률이 최대화되도록 기계 학습 모델의 파라미터를 업데이트하는 단계를 더 포함할 수 있다.
- [0021] 상기 개체 노드들 간의 링크(link)를 예측하는 단계는, 상기 새로운 생물학적 네트워크의 개체 노드들 중 제1 노드에 대응하는 제1 임베딩 벡터를 상기 기계 학습 모델로부터 추출하는 단계; 상기 새로운 생물학적 네트워크의 개체 노드들 중 제2 노드에 대응하는 제2 임베딩 벡터를 상기 기계 학습 모델로부터 추출하는 단계; 및 상기 추출된 제1 임베딩 벡터 및 상기 제2 임베딩 벡터에 기초하여, 상기 제1 노드 및 상기 제2 노드 간의 엣지 존재 여부를 결정하는 단계를 포함할 수 있다.
- [0022] 상기 엣지 존재 여부를 결정하는 단계는, 상기 추출된 제1 임베딩 벡터 및 상기 제2 임베딩 벡터 간의 벡터 거리 차이가 임계 거리 값 미만인 경우에 응답하여, 상기 제1 노드 및 상기 제2 노드 간에 엣지가 존재한다고 결정하는 단계를 포함할 수 있다.
- [0023] 일 실시예에 따른 컴퓨팅 장치는, 생의학적 연관성 예측을 위한 기계 학습 모델을 저장한 메모리; 및 리얼 월드 데이터베이스로부터 환자 정보와 연결된 진단 코드 및 처방 코드를 포함하는 리얼 월드 데이터를 추출하고, 생의학 데이터베이스로부터 유전자 정보, 질병 식별자, 화학물질 식별자, 및 모구조 식별자를 포함하는 생의학 데이터를 추출하며, 상기 리얼 월드 데이터 및 상기 생의학 데이터를 결합함으로써 새로운 생물학적 네트워크를 구축하고, 상기 구축된 새로운 생물학적 네트워크로부터 랜덤워크에 기반하여 생성된 랜덤워크 데이터를 이용하여, 기계 학습 모델을 트레이닝시키며, 상기 트레이닝된 기계 학습 모델에 기초하여 상기 새로운 생물학적 네트워크의 개체 노드의 임베딩 벡터에 기초하여 개체 노드들 간의 링크(link)를 예측하는 프로세서를 포함할 수 있다.

발명의 효과

- [0024] 일 실시예에 따른 컴퓨팅 장치는 생의학 데이터베이스에 축적된 생물학적, 화학적 지식과 전자의무기록의 임상 지식을 효과적으로 결합할 수 있다.
- [0025] 일 실시예에 따른 컴퓨팅 장치는 그래프 구조로 결합 구축된 데이터베이스에서 그래프 임베딩 기술을 통한 약물 적응증, 약물부작용, 약물치료효과 등의 생물학적 개체 간 연관성을 효과적으로 탐색할 수 있다.
- [0026] 일 실시예에 따른 컴퓨팅 장치는 전자의무기록의 환자 및 테이블 추가에 따른 확장성 있는 데이터베이스를 구축할 수 있다.
- [0027] 일 실시예에 따른 컴퓨팅 장치는 생의학 데이터베이스 추가에 따른 확장성 있는 데이터베이스를 구축할 수 있다.

도면의 간단한 설명

- [0029] 도 1은 일 실시예에 따른 생의학적 연관성 예측 장치의 블록도를 도시한다.
- 도 2는 일 실시예에 따른 생의학적 연관성 예측 방법을 도시한 흐름도이다.
- 도 3은 일 실시예에 따른 리얼 월드 데이터(real world data)를 나타낼 수 있다.
- 도 4는 일 실시예에 따른 생의학 데이터(biomedical data)를 나타낼 수 있다.
- 도 5 내지 도 7은 일 실시예에 따른 리얼 월드 데이터 및 생의학적 데이터의 매핑을 설명한다.
- 도 8은 일 실시예에 따른 리얼 월드 데이터 및 생의학적 데이터 간의 매핑 결과를 설명한다.
- 도 9 및 도 10은 일 실시예에 따라 생성된 새로운 병합된 네트워크(merged network)를 설명한다.
- 도 11은 일 실시예에 따른 기계 학습 모델의 트레이닝 및 트레이닝된 기계 학습 모델을 이용한 생의학적 연관성을 설명한다.
- 도 12는 일 실시예에 따라 잠재 공간(latent space)에 임베딩된 개별 개체 노드(entity node)에 대응하는 임베

당 표현 벡터를 예시적으로 도시한다.

발명을 실시하기 위한 구체적인 내용

- [0030] 실시예들에 대한 특정한 구조적 또는 기능적 설명들은 단지 예시를 위한 목적으로 개시된 것으로서, 다양한 형태로 변경되어 구현될 수 있다. 따라서, 실제 구현되는 형태는 개시된 특정 실시예로만 한정되는 것이 아니며, 본 명세서의 범위는 실시예들로 설명한 기술적 사상에 포함되는 변경, 균등물, 또는 대체물을 포함한다.
- [0031] 제1 또는 제2 등의 용어를 다양한 구성요소들을 설명하는데 사용될 수 있지만, 이런 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 해석되어야 한다. 예를 들어, 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소는 제1 구성요소로도 명명될 수 있다.
- [0032] 어떤 구성요소가 다른 구성요소에 "연결되어" 있다고 언급된 때에는, 그 다른 구성요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 할 것이다.
- [0033] 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "포함하다" 또는 "가지다" 등의 용어는 설명된 특징, 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것이 존재함으로 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.
- [0034] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 해당 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가진다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 의미를 갖는 것으로 해석되어야 하며, 본 명세서에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.
- [0035] 이하, 실시예들을 첨부된 도면들을 참조하여 상세하게 설명한다. 첨부 도면을 참조하여 설명함에 있어, 도면 부호에 관계없이 동일한 구성 요소는 동일한 참조 부호를 부여하고, 이에 대한 중복되는 설명은 생략하기로 한다.
- [0036] 최근, 질병의 복잡한 생명활동을 설명하고 질병과 관련 물질들의 관계성을 예측하려는 연구들이 진행되고 있다. 분자 생물학(molecular biology), 유전체학(genomics) 등 다양한 분야에서도 각각의 관련 물질에 대한 오픈소스 데이터베이스들이 구축되었고, 생명활동의 복잡한 현상을 이해하려는 연구에 사용되고 있다.
- [0037] 리얼 월드 데이터(Real World data, RWD)는 변수가 통제되는 실험적 설정에서와 달리, 일상적 프랙티스(routine practice)의 자연적 설정(natural setting)에서의 정보를 참조한다. 전체 생명(whole life)의 현상들의 작용이 완전히 밝혀지고 이해되지는 않기 때문에, 요소들(factors) 및 현상들을 측정하는 테스트들은 연구실에서 통제된 설정을 설계하기에 충분히 명확하지 않을 수 있다.
- [0038] 생의학(biomedical) 연구 보고에서 리얼 월드 데이터의 사용은 정보 시스템의 성장과 함께 활용될 수 있다. 최근 전자 장치들의 대중화와 함께 체계적인 데이터 수집의 보급과 함께, 생의학 데이터(biomedical data)가 전자 의무기록(Electrical Medical Record, EMR)의 형태로 누적될 수 있다. 이를 통해 합리적인 양의 데이터가 계산적으로(computationally) 수집될 수 있다. 임상 연구(clinical research)에서, 전자의무기록 데이터는 임상 사건(clinical event)의 예측, 의학적 예후(medical prognosis), 또는 생의학적 연관성 예측(예를 들어, 약물 재창출(drug repositioning))을 입증(demonstrate)하고, 임상 규제 감시 결정(clinical regulatory surveillance decisions)을 지원하기 위해 사용될 수 있다.
- [0039] 아래에서는 리얼 월드 데이터 및 생의학 데이터의 효과적인 결합 및 결합에 기초하여 새로 구축된 생물학적 네트워크를 활용한 생의학적 연관성을 설명한다. 생의학적 연관성은 생물학적 데이터베이스의 개체 간의 연관성으로서, 생물학적 특성들이 서로 관련되어 있는지 여부를 나타낼 수 있고, 개체 간의 연관성은 예시적으로 약물재창출, 약물 부작용 예측, 약물치료 효과 예측, 질환 발병 예측 등을 포함할 수 있다. 약물, 질병, 처방, 및/또는 진단 간의 이전에 알려지지 않았던 연관성을 예측하는 것을 약물 재창출(drug repositioning)이라고 나타낼 수 있다. 본 명세서에서는 생의학적 연관성 예측의 예시로서 주로 약물 재창출을 설명하지만, 이로 한정하는 것은 아니다.
- [0040] 도 1은 일 실시예에 따른 생의학적 연관성 예측 장치의 블록도를 도시한다.
- [0041] 일 실시예에 따른 컴퓨팅 장치(100)는 프로세서(110) 및 메모리(120)를 포함할 수 있다.

- [0042] 프로세서(110)는 리얼 월드 데이터베이스로부터 환자 정보와 연결된 진단 코드 및 처방 코드를 포함하는 리얼 월드 데이터를 추출할 수 있다. 프로세서(110)는 생의학 데이터베이스로부터 유전자 정보, 질병 식별자, 화학물질 식별자, 및 모구조 식별자를 포함하는 생의학 데이터를 추출할 수 있다. 프로세서(110)는 리얼 월드 데이터 및 생의학 데이터를 결합함으로써 새로운 생물학적 네트워크를 구축할 수 있다. 프로세서(110)는 구축된 새로운 생물학적 네트워크로부터 랜덤워크에 기반하여 생성된 랜덤워크 데이터를 이용하여, 기계 학습 모델을 트레이닝시키며, 트레이닝된 기계 학습 모델에 기초하여 새로운 생물학적 네트워크의 개체 노드의 임베딩 벡터에 기초하여 개체 노드들 간의 링크(link)를 예측할 수 있다. 프로세서(110)의 상세한 동작은 하기 도 2 내지 도 12에서 설명한다.
- [0043] 메모리(120)는 생의학적 연관성 예측을 위한 기계 학습 모델을 저장할 수 있다. 메모리(120)는 생의학적 연관성 및 기계 학습 모델의 트레이닝을 위해 요구되는 데이터를 임시적으로 또는 영구적으로 저장할 수 있다.
- [0044] 또한, 컴퓨팅 장치(100)는 출력부(예를 들어, 디스플레이)를 더 포함할 수 있다. 출력부는 잠재 공간으로 매핑된 임베딩 벡터 및/또는 개체 노드들 간의 링크 예측 결과를 시각적으로 출력할 수 있다.
- [0045] 일 실시예에 따른 컴퓨팅 장치(100) 전자의무기록과 생의학 데이터베이스에서 공통으로 사용되는 용어(Terminology)에 대한 매핑을 수행하고, 그래프 형태의 데이터베이스를 구축할 수 있다. 그래프 데이터베이스에서 약물의 잠재적인 유익성과 위해성을 탐색할 수 있는 효과적인 그래프 구조가 구축되며, 컴퓨팅 장치(100)는 그래프 임베딩 기술을 통해 약물 효과 탐색을 수행할 수 있다. 컴퓨팅 장치(100)는 전자의무기록과 생의학 데이터베이스를 동시에 활용하여 약물 적응증을 탐색할 수 있다.
- [0046] 컴퓨팅 장치(100)는 전자의무기록과 생의학 데이터베이스에서 공통으로 사용되는 용어(Terminology)에 대한 매핑을 수행할 수 있다. 컴퓨팅 장치(100)는 전자의무기록과 생의학 데이터베이스를 그래프 형태로 결합 구축하였을 때, 그래프의 잠재 표현(Latent Representation)을 효과적으로 임베딩할 수 있다.
- [0047] 본 명세서에서 생물학적 네트워크(biological network)는 생물학적 시스템에 관한 데이터에 적용되는 모든 네트워크로서, 데이터 간의 연결에 관한 수학적 표현을 제공할 수 있는 네트워크를 나타낼 수 있다. 생물학적 네트워크는 그래프 구조 데이터로 표현될 수 있다. 그래프는 정점(vertex)과 엣지(edge)로 구성된 자료 구조로서, 본 명세서에서 정점은 개체 노드(entity node)라고도 나타낼 수 있다. 예시적으로 생물학적 네트워크에 포함되는 생물학적 데이터(biological data)는 하나 이상의 개체 노드를 포함할 수 있다. 개체 노드는 생물학적 네트워크 데이터베이스에 포함된 개체에 따라 다양한 개체 타입으로 정의되어 해당 개체 타입에 해당하는 정보(예를 들어, 질병 타입은 질병의 종류를 지시하는 정보)를 지시하는 노드로서, 본 명세서에서 설명되는 예시적인 개체 타입은 리얼 월드 데이터베이스의 진단 타입, 처방 타입, 및 환자 타입, 그리고 생의학 데이터베이스의 유전자 타입, 질병 타입, 화학물질 타입, 및 모구조 타입이 설명된다. 각 개체 타입의 개체 노드는 하기 도 3 및 도 4에서 설명한다. 각 개체 노드는 하나 이상의 다른 개체 노드와 엣지(edge)를 통해 연결될 수 있다. 본 명세서에서는 방향 및 가중치를 가지지 않는 엣지를 주로 설명하나, 이로 한정하는 것은 아니다.
- [0048] 참고로, 본 명세서에서 생물학적 네트워크는 리얼 월드 데이터베이스(real world database)에 포함되는 리얼 월드 데이터 네트워크 및 미리 구축된(pre-built) 생의학 데이터베이스(biomedical database)에 포함되는 생의학 네트워크(biomedical network)를 포함할 수 있다. 리얼 월드 데이터 네트워크로부터 추출되는 데이터를 리얼 월드 데이터(RWD, real world data)로 나타낼 수 있고, 생의학 네트워크로부터 추출되는 데이터를 생의학 데이터(biomedical data)라고 나타낼 수 있다.
- [0049] 도 2는 일 실시예에 따른 생의학적 연관성 예측 방법을 도시한 흐름도이다.
- [0050] 우선, 단계(210)에서 컴퓨팅 장치는 리얼 월드 데이터베이스로부터 환자 정보와 연결된 진단 코드 및 처방 코드를 포함하는 리얼 월드 데이터를 추출할 수 있다. 예를 들어, 컴퓨팅 장치는 환자 정보를 지시하는 환자 노드를 환자 정보에 대응하여 추출된 처방 코드를 지시하는 처방 노드와 진단 코드를 지시하는 진단 노드에 엣지를 통해 연결함으로써, 그래프 구조의 리얼 월드 데이터를 생성할 수 있다.
- [0051] 그리고 단계(220)에서 컴퓨팅 장치는 생의학 데이터베이스로부터 유전자 정보, 질병 식별자, 화학물질 식별자, 및 모구조 식별자를 포함하는 생의학 데이터를 추출할 수 있다. 예를 들어, 컴퓨팅 장치는 유전자 정보를 지시하는 유전자 노드를 유전자 정보에 대응하는 질병 식별자는 지시하는 질병 노드에 엣지를 통해 연결할 수 있다. 컴퓨팅 장치는 질병 노드를 질병 식별자에 대응하는 화학물질 식별자를 지시하는 화학물질 노드에 엣지를 통해 연결할 수 있다. 컴퓨팅 장치는 화학물질 노드를 화학물질 식별자에 대응하는 모구조 식별자를 지시하는 모구조 노드에 엣지를 통해 연결할 수 있다. 컴퓨팅 장치는 엣지를 통해 연결된 유전자 노드, 질병 노드, 화학물질

노드, 및 모구조 노드를 포함하는 그래프 구조의 생의학 데이터를 생성할 수 있다.

- [0052] 참고로, 생의학 데이터에서 질병 노드는 엣지를 통해 화학물질 노드를 경유하여 모구조 노드와 연결되고, 화학물질 노드는 엣지를 통해 질병 노드를 경유하여 유전자 노드와 연결될 수 있다.
- [0053] 이어서 단계(230)에서 컴퓨팅 장치는 리얼 월드 데이터 및 생의학 데이터를 결합함으로써 새로운 생물학적 네트워크를 구축할 수 있다. 예를 들어, 컴퓨팅 장치는 리얼 월드 데이터의 진단 노드 및 생의학 데이터의 질병 노드를 연결하고 리얼 월드 데이터의 처방 노드 및 생의학 데이터의 화학물질 노드를 연결함으로써 새로운 생물학적 네트워크를 구축할 수 있다.
- [0054] 그리고 단계(240)에서 컴퓨팅 장치는 구축된 새로운 생물학적 네트워크로부터 랜덤워크에 기반하여 생성된 랜덤 워크 데이터를 이용하여, 기계 학습 모델을 트레이닝시킬 수 있다.
- [0055] 이어서 단계(250)에서 컴퓨팅 장치는 트레이닝된 기계 학습 모델에 기초하여 새로운 생물학적 네트워크의 개체 노드의 임베딩 벡터에 기초하여 개체 노드들 간의 링크(link)를 예측할 수 있다.
- [0056] 도 3은 일 실시예에 따른 리얼 월드 데이터(real world data)를 나타낼 수 있다.
- [0057] 컴퓨팅 장치는 리얼 월드 데이터 네트워크(real world data network)로부터 리얼 월드 데이터(300)를 추출할 수 있다. 리얼 월드 데이터 네트워크는 예시적으로 전자의무기록(EMR, Electric Medical Record)의 형태로 누적된(accumulated) 데이터로 구성되는 네트워크로서, 리얼 월드 데이터베이스라고도 나타낼 수 있다. 리얼 월드 데이터(300)는, 임의의 변수가 제한된 실험 세팅에서의 임상 데이터가 아닌, 다양한 변수들이 포함되는 실세계에서의 데이터를 나타낼 수 있다. 생명 현상에 영향을 주는 변수는 매우 다양한 바, 제한된 실험으로 얻어진 실험 결과는 리얼 월드에서 획득되는 결과와 다를 수 있다. 따라서, 리얼 월드 데이터(300)는 제한되지 않은 실제 다양한 변수를 내포하므로, 임상적인 근거를 가질 수 있다. 누적된 리얼 월드 데이터(300), 예를 들어, 전자의무기록 데이터는 새로운 접근을 통해 생물학적 현상이나 실제 임상에서의 사건을 더 잘 설명하기 위한 기회를 제공할 수 있다.
- [0058] 일 실시예에 따른 전자의무기록 데이터는 서울아산병원(Asan Medical Center, AMC)의 CardioNet 데이터베이스로부터 추출될 수 있다. 리얼 월드 데이터(300)의 구조화된 개체(entity)는 환자 인구통계 테이블(patient demographics table), 진단 테이블(310)(diagnosis table) 및 투약 테이블(320)(medication table)로부터 추출될 수 있다. 리얼 월드 데이터베이스의 환자는 익명으로 환자 식별자(Patient Identifier, PAID)에 의해 넘버링될 수 있다. 환자 인카운터 번호(patient encounter number)(EN_NO)는 다른 종류의 진단들을 위한 매 초진마다 주어질 수 있다. 환자 식별자(PAID) 및 환자 인카운터 번호(EN_NO)가 스트링 결합된 키(key)는 환자의 진단 테이블(310) 및 투약 테이블(320)을 조회하는데 사용될 수 있다. 환자 인카운터 번호(EN_NO)는 키(key)로 대체될 수도 있다. 환자 인카운터 정보는 환자 인카운터 번호(EN_NO) 및/또는 환자 식별자(PAID) 및 환자 인카운터 번호(EN_NO)가 스트링 결합된 키를 포함할 수 있다. 예를 들어, 환자 타입의 개체 노드(이하, '환자 노드')는 환자 정보로서, 환자 식별자(PAID)를 지시하는 환자 식별자 노드와 환자 인카운터 번호(EN_NO) 또는 키(key)를 지시하는 환자 인카운터 노드를 포함할 수 있다.
- [0059] 컴퓨팅 장치는 진단 테이블(310)(diagnosis table)로부터, 진단 코드(diagnosis code, DICD), 각 키의 인-데이트(in-date, INDT), 및 각 진단 코드(DICD)의 초진 날짜(The date of initial diagnosis, T_INDT)를 추출할 수 있다. 진단 테이블(310)은 한국질병분류(Korean Classification of Diseases, KCD)를 기반으로 부여된 진단 코드(DICD)를 포함할 수 있다. 진단 코드(DICD)는 한 환자에게 각 방문 시 발생한 질병들의 목록 및 증상(symptom)을 지시하는 코드일 수 있다. 진단 코드(DICD)(예를 들어, 서울아산병원 기관의 진단 코드)는 "질병 및 관련 건강 문제의 국제 질병 분류"(International Statistical Classification of Diseases and Related Health Problems, ICD)에 기초하여 생성된 한국질병분류(Korean Classification of Diseases, KCD)로부터 확장된 코드일 수 있다. 진단 코드(DICD)는 주로 소스 코드(source code)로부터 비롯되며(originated), 건강 보험 청구(health insurance claim) 또는 서비스 상세(service detail)(예를 들어, 응급 방문)와 같은 기관의 요구(institutional need)에 따른 코드를 포함할 수 있다. 예를 들어, 진단 타입의 개체 노드(이하, '진단 노드')는 진단 내용을 지시하는 진단 코드(DICD)를 지시하는 개체 노드일 수 있다.
- [0060] 컴퓨팅 장치는 투약 테이블(320)(medication table)로부터 처방 코드(prescription code)(ODCD)를 각 키에 대해 요소 컬럼(element column, ELEM)에 따라 추출할 수 있다. 투약 테이블(320)은 RxNorm와 RxNorm Extension을 기반으로 부여된 처방 코드(Order code, ODCD) 및 성분 명(ingredient name) 등을 포함할 수 있다. RxNorm 및 RxNorm Extension은 미국 시장에서 구할 수 있는 모든 약품을 포함하는 약품의 미국 특정 용어로서, 개인 건

강 기록 애플리케이션에도 사용할 수 있으며, RxNorm은 통합 의료 언어 시스템(Unified Medical Language System, UMLS) 용어의 일부이며 미국 국립 의학 도서관에서 관리될 수 있다. 처방 코드(ODCD)는 병원에 방문한 환자에게 투약된 약물에 관한 정보를 지시하는 코드일 수 있다. 예를 들어, 처방 타입의 개체 노드(이하, '처방 노드')는 처방 내용을 지시하는 처방 코드(ODCD)를 지시하는 개체 노드일 수 있다.

[0061] 전술한 바와 같이, 처방 코드(ODCD)와 진단 코드(DICD)는, 환자의 질병 이력과 함께 입원 시 환자의 상태 변화나 환자의 특성에 따라 투약된 약의 종류를 내포할 수 있다.

[0062] 일 실시예에 따르면 컴퓨팅 장치는 환자 정보(예를 들어, 전술한 환자 식별자(PAID) 및 환자 인카운터 번호(EN_NO))와 연결된 처방 코드(ODCD) 및 진단 코드(DICD)를 개체로서 포함하는 리얼 월드 데이터(300)를 추출할 수 있다. 컴퓨팅 장치는 환자 특성이 반영된 리얼 월드 데이터(300)를 추출할 수 있다. 예를 들어, 컴퓨팅 장치는 추출된 환자 정보를 지시하는 환자 노드를 해당 환자 노드에 대응하여 추출된 처방 코드(ODCD)를 지시하는 처방 노드와 진단 코드(DICD)를 지시하는 진단 노드에 엣지를 통해 연결함으로써 그래프 구조의 리얼 월드 데이터(300)를 생성할 수 있다. 또한, 컴퓨팅 장치는 정의된 결과(330)를 지시하는 아웃컴 노드(outcome node)를 환자 노드에 엣지를 통해 추가로 연결할 수 있다.

[0063] 또한, 컴퓨팅 장치는 구조(structure)의 특정 사용에 대해, 정의된 결과(defined outcome)(330)를 나타내는 컬럼 또는 테이블을 추출할 수 있다. 예를 들어, 도 3에서는 처방 코드(ODCD)에 대응하는 약물의 처방에 의해 환자의 상태가 긍정적(positive)으로 호전되었는지 여부 및/또는 부정적(negative)으로 악화되었는지 여부를 지시하는 결과를 포함할 수 있다. 다른 예를 들어, 컴퓨팅 장치는 항암제(anticancer drug)에 대한 환자 조정(patient adjustment)에 대한 정의된 결과로서, 환자 인구 통계 테이블로부터 각 환자 식별자(PAID)에 대해 암 등록 환자의 사망일(death date of a cancer registration patient)(CDTH)을 추출할 수 있다.

[0064] 도 4는 일 실시예에 따른 생의학 데이터(biomedical data)를 나타낼 수 있다.

[0065] 일 실시예에 따르면 컴퓨팅 장치는 미리 구축된 생의학 데이터베이스(pre-built biomedical database)로부터 생의학 데이터(400)를 추출할 수 있다. 생의학 데이터베이스는 전문가들에 의해 수동으로 큐레이팅된 관계(curated relationship)를 가지는 데이터를 포함할 수 있다. 예를 들어, 생의학 데이터베이스는 CTD(Comparative Toxicogenomic Database) 데이터베이스일 수 있으며,

[0066] CTD 데이터베이스는 독성 유전학적 관계(toxicogenomic relationships)의 상호작용들이 수동으로 큐레이팅된(curated) 비교 데이터베이스(comparative database)를 나타낼 수 있다. CTD 데이터베이스는 화학물질의 화학 메커니즘과 유전자의 연관 관계 및 약물과 질병의 관계에 대한 데이터를 제공한다. CTD 데이터베이스의 각 데이터셋에서 나타나는 개체 타입별 관계성은 전문가들에 의해 큐레이팅된 관계일 수 있다. CTD 데이터베이스는 환경 화학물질(environmental chemicals)의 화학적 독성(chemical toxicity) 뿐만 아니라 독성 물질(toxicogenomic)의 약리학적 활성(pharmacological activity)도 커버하도록 확장된 데이터베이스일 수 있다. 더 나아가, 유전자 관련 개체(gene related entities)에 관한 큐레이션(curation)은 표현형(phenotypes), 유전자 온톨로지(gene ontology), 및 경로로 확장(broaden)되어 유전자(genes), 단백질(protein), 대사 산물(metabolite) 및 그 관련 반응(reaction)을 포함하는 학제간 연구(interdisciplinary research)로 확장할 가능성을 줄 수 있다.

[0067] 질병 식별자(Disease_ID)는 미국 국립보건원 산하 국립의학도서관(National Library of Medicine, NLM)의 의학 주제 표제(Medical Subject Headings, MeSH)의 "질병" 브랜치에 따라 결정될 수 있다. 또한, 질병 식별자(Disease_ID)는 MeSH에 의해 커버되지 않는 일부 유전 질환(genetic disorders)에 대해서는 OMIM(Online Mendelian Inheritance in Man)에 따라 결정될 수 있다. 예를 들어, 질병 타입의 개체 노드(이하, '질병 노드')는 질병의 종류를 지시하는 질병 식별자(Disease_ID)를 지시하는 개체 노드일 수 있다.

[0068] 유전자 정보(Gene_Symbol)는 NCBI(National Center for Biotechnology Information)의 유전자 데이터베이스로부터 독성학적으로(toxicologically) 중요한 유전자 심볼 또는 단백질 심볼에 관한 정보를 포함할 수 있다. 예를 들어, 유전자 타입의 개체 노드(이하, '유전자 노드')는 독성학적으로(toxicologically) 중요한 유전자 심볼 및/또는 단백질 심볼의 종류를 지시하는 유전자 정보(Gene_Symbol)를 지시하는 개체 노드일 수 있다.

[0069] 화학물질 식별자(Chemical_ID)는, NLM MeSH의 "화학물질 및 약물" 브랜치에 기초하여 결정될 수 있다. 또한, 화학물질 식별자(Chemical_ID)는 분자 시약(molecular reagents), 환경 화학물질(environmental chemicals), 및 임상 약물(clinical drugs)에 따라서도 결정될 수 있다. 예를 들어, 화학물질 타입의 개체 노드(이하, '화학물질 노드')는 화학물질의 종류를 지시하는 화학물질 식별자(Chemical_ID)를 지시하는 개체 노드일 수 있다.

- [0070] 컴퓨팅 장치는 CTD 데이터베이스의 관계 데이터(relationship data)에 기초하여 생의학 데이터(400)를 생성 및 추출할 수 있다. 관계 데이터는, 예를 들어, 화학물질-질병(chemical-disease) 데이터셋(410), 질병-유전자(disease-gene) 데이터셋(420), 및 화학물질들(chemicals) 데이터셋(430)에 기초하여 생의학 데이터(400)를 생성 및 추출할 수 있다. 화학물질-질병 데이터셋(410)에서 화학물질 및 질병 간의 상호작용(예를 들어, 관계성), 질병-유전자 데이터셋(420)에서 질병 및 유전자 간의 상호작용, 및 화학물질들 데이터셋(430)에서 화학물질들 간 상호작용은 출판된 문헌에 기초하여 전문가에 의해 수동으로 큐레이팅된 데이터셋일 수 있다.
- [0071] 화학물질-질병 데이터셋(410)은 화학물질 및 질병 간의 상호작용에 관한 데이터셋으로서, 불필요한 환경 화학물질(environmental chemical)이 제거되고, 나머지 화학물질에 해당하는 약의 성분과 질병, 화학물질-질병 관계가 추출된 데이터일 수 있다. 참고로, 임상 약물의 성분(ingredients)은 화학물질 식별자(Chemical_ID) 및 전자의 무기록 데이터베이스의 처방 정보 간의 매핑에 의해 선별(sift)될 수 있다.
- [0072] 질병-유전자 데이터셋(420)은 타겟 단백질(protein)이나 타겟 유전자와의 반응(reaction)을 통해 작용하는 약의 기전을 반영하기 위한 질병-유전자 상호작용들 중 전자의무기록 데이터베이스에 있는 질병에 대응하여 큐레이팅된 질병-유전자 관계를 포함할 수 있다. 컴퓨팅 장치는 생의학 데이터베이스로부터 해당 질병에 관련된 복수의 유전자 정보 중 미리 결정된 개수의 유전자 정보를 추출할 수 있다. 컴퓨팅 장치는 생의학 데이터베이스에 의해 제공되는 추론 점수에 기초하여, 해당 질병에 대해 가장 높은 추론 점수를 갖는 유전자 정보부터 내림차순으로 미리 결정된 개수까지 해당 질병에 관련된 유전자 정보를 추출할 수 있다. 예를 들어, 컴퓨팅 장치는 추론 점수에 기초하여 상위 10개 유전자 정보를 추출할 수 있다. 노드 가중치 없는 그래프 구조의 네트워크에서, 모든 연관성들(예를 들어, 관계성)은 중요성에서의 차이에도 불구하고, 랜덤워크를 할 시, 동일한 가능성으로 평가되어야 하는 바, 생의학 데이터(400)에서 개체들 간의 연관성들(예를 들어, 개체 노드들 간의 관련성들)을 지시하는 엣지들의 개수가 정규화될 필요가 있다. 따라서, 한 질병에 관련된 복수의 유전자 정보 중 10개의 가장 중요한 연관성들로 연결된 유전자 정보가 중요성을 정규화하기 위해 추출될 수 있고, 중요한 유전자들에 대한 위킹의 가능성이 네트워크에 걸쳐 표준화될 수 있다.
- [0073] 추론 점수(Inference Score)는 PPI(Protein Protein Interaction)에서 단백질들 간 기능적 관계를 통계적으로 평가하는데 사용될 수 있다. 추론 점수는 화학물질-유전자 및 질병-유전자 연관성에서 각 질병 또는 화학물질에 연관된 상위 10개 유전자를 픽(pick)하는데 사용될 수 있다. 추론 점수는 생물학적 네트워크가 스케일 프리 랜덤 네트워크(scale free random network)를 형성하기 위해 내재적(innate) 방식의 개념으로부터 도출될 수 있다. 화학물질-유전자-질병 네트워크 및 비전형적인(atypical) 연결성의 정도를 평가하기 위한 유사한 스케일 프리 랜덤 네트워크 간 유사도를 나타낼 수 있다. 더 나아가, 추론 점수를 계산하기 위해 관련 질병/화학물질의 양과 유추된 화학물질/질병 자체의 연결성에 대한 화학물질/질병 및 유전자의 연결성과 같은 공통 이웃(Common Neighbors) 방식이 사용될 수 있다.
- [0074] 화학물질들 데이터셋(430)은 화학물질 어휘 데이터셋(Chemical vocabulary dataset)라고도 나타낼 수 있고, 약물이 모구조(parent structure)에서 작용기 차이로 달라지는 반응을 반영하기 위해 사용될 수 있다. 컴퓨팅 장치는 화학물질들 데이터셋(430)으로부터 화학물질-모구조(parent structure) 간의 관계, 예를 들어, 한 화학물질 식별자(Chemical_ID)에 연결되는 모구조 식별자(Parent_ID)를 추출할 수 있다. 화학물질-모구조 관계는, 모든 자손(descendant)에게 링크되는 부모 구조 정보(parental structure information)를 포함할 수 있다. 따라서, 컴퓨팅 장치는 화학물질의 모구조를 추출하여 작용기 차이를 가지는 화학물질들을 연결시킴으로써, 유사한 화학구조 식을 가지지만 작용기 차이로 인해 효과가 달라지는 약의 특성을 생의학 데이터(400)에 반영할 수 있다. 예를 들어, 모구조 타입의 개체 노드(이하, '모구조 노드')는 모구조의 종류를 지시하는 모구조 식별자(Parent_ID)를 지시하는 개체 노드일 수 있다.
- [0075] 또한, 독성 유전학적 관계(toxicogenomical relationships)를 결합하기 위해, 질병-유전자 데이터셋(420) 및 화학물질-유전자 데이터셋이 하에 사용될 수 있다.
- [0076] 도 4에 도시된 예시에서, 컴퓨팅 장치는 생의학 데이터베이스로부터, 유전자 정보(Gene_Symbol), 질병 식별자(Disease_ID), 화학물질 식별자(Chemical_ID), 및 모구조 식별자(Parent_ID)를 개체로서 포함하는 생의학 데이터(400)를 추출할 수 있다. 추출된 생의학 데이터(400)는 서로 대응하는 유전자 정보(Gene_Symbol) 및 질병 식별자(Disease_ID), 질병 식별자(Disease_ID) 및 화학물질 식별자(Chemical_ID), 화학물질 식별자(Chemical_ID) 및 모구조 식별자(Parent_ID)를 포함할 수 있다. 예를 들어, 컴퓨팅 장치는 추출된 유전자 정보(Gene_Symbol)를 지시하는 유전자 노드를 해당 유전자 정보(Gene_Symbol)에 대응하는 질병 식별자(Disease_ID)를 지시하는 질병 노드에 엣지를 통해 연결할 수 있다. 유사하게, 컴퓨팅 장치는, 엣지를 통해, 질병 노드에 해당 질병 식별

자(Disease_ID)에 대응하는 화학물질 식별자(Chemical_ID)를 지시하는 화학물질 노드를 연결하고, 화학물질 노드에 해당 화학물질 식별자(Chemical_ID)에 대응하는 모구조 식별자(Parent_ID)를 지시하는 모구조 노드를 연결할 수 있다. 따라서, 컴퓨팅 장치는 노드 간 엣지 연결을 통해 유전자 정보(Gene_Symbol), 질병 식별자(Disease_ID), 화학물질 식별자(Chemical_ID), 및 모구조 식별자(Parent_ID)를 포함하는 생의학 데이터(400)를 추출할 수 있다. 다시 말해, 컴퓨팅 장치는 엣지를 통해 연결된 유전자 노드, 질병 노드, 화학물질 노드, 및 모구조 노드를 포함하는 그래프 구조의 생의학 데이터를 생성할 수 있다. 예시적으로, 질병 노드는 엣지를 통해 화학물질 노드를 경유하여 모구조 노드와 연결될 수 있고, 화학물질 노드는 엣지를 통해 질병 노드를 경유하여 유전자 노드와 연결될 수 있다. 다만, 설명의 편의를 위하여 임의의 순서로 연결 과정을 설명하였으나, 연결 순서를 전술한 바로 한정하는 것은 아니다. 설계에 따라 전술한 노드 간 연결은 역순으로 수행되거나, 임의의 순서로 수행될 수도 있다.

[0077] 환자의 유전체 데이터는 EMR에 결여되기 때문에, 엣지가 만들어지지 않을 수 있다. 생의학 데이터베이스에 관련 정보가 추가되지 않을 수 있다. 그럼에도 불구하고, CTD 데이터베이스에서, 공통 이웃 방식의 통계적 평가로부터 도출되는 자체 통계적 추론 점수를 통해, 질병이 질병에 상관(corelate)되거나, 화학물질이 화학물질에 상관될 수 있다.

[0078] 도 5 내지 도 7은 일 실시예에 따른 리얼 월드 데이터 및 생의학 데이터의 매핑을 설명한다.

[0079] 일 실시예에 따르면 컴퓨팅 장치는 리얼 월드 데이터(510)의 진단 노드(512) 및 생의학 데이터(530)의 질병 노드를 연결하고 리얼 월드 데이터(510)의 처방 노드(511) 및 생의학 데이터(530)의 화학물질 노드를 연결함으로써 새로운 생물학적 네트워크를 구축할 수 있다. 예를 들어, 컴퓨팅 장치는 생의학 용어(biomedical terminology)에 관한 "공통 데이터 모델"(Common Data Model, CDM)에 기초하여 동일한 고유 개념 식별자에 대응하는 진단 노드(512) 및 질병 노드끼리 매핑함으로써, 리얼 월드 데이터(510) 및 생의학 데이터(530)를 결합할 수 있다. 컴퓨팅 장치는 공통 데이터 모델에 기초하여 동일한 고유 개념 식별자에 대응하는 처방 노드(511) 및 화학물질 노드끼리 연결할 수 있다. 리얼 월드 데이터베이스, 예를 들어, 임상 데이터 셋에서 병원 내 코드들(예를 들어, 전술한 진단 코드(DICD) 및 처방 코드(ODCD))은 각 병원의 고유한 코드 체계를 따라 생성되므로, 병원 내 코드를 다양한 소스에 적용하기 위해, 컴퓨팅 장치는 공통 데이터 모델에 기초한 매핑을 수행할 수 있다.

[0080] 공통 데이터 모델은 생의학 용어(biomedical terminology)의 시소러스(thesaurus), 온톨로지(ontology), 및 인포매틱스(informatics)에 관한 모델로서, 예를 들어, 통합 의료 언어 시스템(Unified Medical Language System, UMLS(520)), 몬도 질병 온톨로지(Mondo Disease Ontology, Mondo), 및 OHDSI(Observational Health Data Sciences and Informatics)를 포함할 수 있다.

[0081] UMLS(520)는 상호 운용 가능(interoperable)하게 어휘 소스(vocabulary source)를 통합하도록 만들어진 모든 생의학(biomedical) 도메인의 시소러스(thesaurus)일 수 있다. UMLS(520)는 생의학 개념의 용어 문자열(terminology string) 자체가 의미를 나타내며 동일한 의미에 대해 동의어(synonyms)를 가지는 것을 가정한다. UMLS(520)에서는 같은 의미를 갖는 다른 어휘들이 하나의 개념 고유 식별자(Concept Unique Identifier, CUI)로 그룹핑될 수 있다. 예를 들어, 같은 개념 내에서, 다른 동의어 고유 식별자(Lexical Unique Identifier, LUI)들에 각각 할당된 동의어들 및 어휘 변형(lexical variation)은 동일한 개념 고유 식별자로 그룹핑될 수 있다. 같은 동의어들 내에서도 문자열 변형들(예를 들어, 대문자 또는 소문자)에 대해, 다른 문자열 고유 식별자(String Unique Identifier, SUI)들이 주어지고, 해당 문자열 변형들도 동일한 개념 고유 식별자로 그룹핑될 수 있다. UMLS(520) 메타시소러스(Metathesaurus)는, 컬럼이 정의된 카테고리들 하에서 생의학 개념들을 갖는 명칭(name)으로 구성되는, 생의학 개념들의 대규모 시소러스(예를 들어, 동의어 사전)일 수 있다.

[0082] Mondo는 복수의 질병 용어 도메인을 병합한 온톨로지로서 어휘 소스들에 걸쳐 정확한 매핑들을 제공하고자 하는 온톨로지를 나타낼 수 있다. 개념 및 관련 유의어를 제공하는 UMLS(520)는 일대다 매핑이 가능할 수 있다. 한편, Mondo는 "K-평균 베이지안 온톨로지 추론"(K-means Bayesian-ontology inference, kBoom)을 이용하여 일대일 매핑을 제공할 수 있다. Mondo에서는 소스 제공된 매핑들로부터 확률적 온톨로지가 생성되고, kBoom에 의한 확률을 우선순위화(prioritize)함으로써 온톨로지로서 병합(merge)될 수 있다. 이 고품질 일대일 매핑은 랜덤워크 기반 그래프 임베딩을 위한 네트워크에게 활용될 수 있는데, 노드에서 링크의 개수는 임베딩의 퀄리티에 영향을 미칠 수 있다.

[0083] OHDSI는 글로벌 레벨(global level)에서 기관들(organizations) 간의 건강 데이터 분석을 가능케 하는 표준화된 관찰 연구(standardized observational research)를 제시할 수 있다. OMOP(Observational Medical Outcomes

Partnership) CDM은 각 의료기관들이 보유한 다른 구조의 의료 데이터에 적용 가능한 동일한 구조와 규격의 데이터 모델이며, 전세계 여러 기관 소스들(multi-institutional sources)로부터 데이터를 변환하는데 사용될 수 있다. 국가는 헬스케어 제품이 승인되는 자체 국가 건강 부서를 가지기 때문에, 임상 약물 제품(clinical drug product)은 국가마다 다를 수 있다. 글로벌 약물을 표준화하기 위하여, OMOP CDM은 RxNorm Extension이라고 불리는 약(medicine)에 대한 상호 운용 가능한 표준화된 어휘를 제공할 수 있다. 다시 말해, OHDSI는 개별 국가에서 처방되는 약을 글로벌 컨셉으로 분류하기 위해 사용될 수 있다. RxNorm Extension은, 미국 특정 표준 용어인, RxNorm으로부터 도출된 개념들의 확장을 갖는 어휘일 수 있다. 로컬 기관 용어 약물 데이터를 글로벌 약물 개념으로 OMOP CDM을 이용하여 전송하는 것이 가능할 수 있다.

[0084] 일 실시예에 따르면 리얼 월드 데이터(510)의 처방 노트(511)에 대응하는 개념 고유 식별자(CUI)는 스트링 매핑 및 요소 매핑(element mapping)에 기초하여 획득될 수 있다. 리얼 월드 데이터(510)의 진단 노트(512)에 대응하는 개념 고유 식별자(CUI)는 스트링 매핑 및 소스 매핑(source mapping)에 기초하여 획득될 수 있다.

[0085] 아래에서는 공통 데이터 모델로서 UMLS(520)를 사용하는 예시를 주로 설명하나, 이로 한정하는 것은 아니다.

[0086] 도 6은 예시적으로 UMLS를 이용한 진단 코드(DICD)를 지시하는 질병 노트 및 질병 식별자(Disease_ID)를 지시하는 질병 노트 간의 매핑 동작을 설명하는 도면이다.

[0087] 일 실시예에 따른 리얼 월드 데이터의 코드들(예를 들어, 진단 코드(DICD) 및 처방 코드(ODCD))은 공통 데이터 모델의 개념 명칭(concept name)을 기반으로 생성될 수 있다. 공통 데이터 모델에서 다른 문장 문자열(statement string)에 대응하는 식별자들(IDs, identifiers)(예를 들어, 동의어 고유 식별자(LUI) 및 문자열 고유 식별자(SUI) 등)이 같은 개념을 가지는 경우, 전술한 서로 다른 문장 문자열에 대응하는 식별자들은 동일한 개념 고유 식별자(CUI)로 그룹핑될 수 있다. 같은 개념 고유 식별자(CUI)의 개념 내에서는 여러가지 동의어(lexical)한 유의어(synonym)를 지시하는 문장 문자열들이 포함될 수 있다. 각 어휘 변형(lexical variation)에는 동의어 고유 식별자(LUI)가 부여될 수 있고, 각 어휘 변형의 문자열 변형(string variation)마다 문자열 고유 식별자(SUI)가 부여될 수 있다. 예를 들어, 도 6은 개념 고유 식별자(CUI)로서 "C0026919"를 가지는 다양한 문장 문자열들의 목록을 도시한다. 따라서, 컴퓨팅 장치는 공통 데이터 모델의 문자열 필드(string field)의 문자열 고유 식별자를 이용하여 개념 매핑(concept mapping)을 수행할 수 있다.

[0088] 컴퓨팅 장치는 리얼 월드 데이터의 개체 노트(예를 들어, 진단 노트)에 의해 지시되는 코드(예를 들어, 진단 코드(DICD) 및 처방 코드(ODCD))에 대응하는 문자열을 추출할 수 있다. 컴퓨팅 장치는 공통 데이터 모델(예를 들어, UMLS)로부터 리얼 월드 데이터의 코드에 대응하는 문자열과 일치하는 문자열을 가지는 개념 고유 식별자(CUI)를 추출할 수 있다. 예시적으로, 문자열은 UMLS의 가장 작은 단위이므로, 컴퓨팅 장치는 기본적으로 문자열 단위로 개념 고유 식별자(CUI)를 검색할 수 있으나, 이로 한정하는 것은 아니다. 컴퓨팅 장치는 리얼 월드 데이터의 코드에 일치하는 코드를 갖거나 코드에 대응하는 문자열에 일치하는 문자열을 갖는 고유 개념 식별자(CUI)를 검색할 수 있다. 컴퓨팅 장치는, 리얼 월드 데이터의 코드 및/또는 생의학 데이터의 식별자에 대응하는 문자열과 공통 데이터 모델의 문자열이 정확히 동일한 경우에만 서로 일치하는 것으로 판단할 수 있다. 도 6에서는 진단 코드(DICD) 및 질병 식별자(Disease_ID) 간의 매핑을 주로 설명한다.

[0089] 진단 코드(DICD)는 '코어' 분류로 구성되는 ICD(International Statistical Classification of Diseases and Related Health Problems) 10에 기초할 수 있다. ICD 10은 3개 문자 카테고리 및 소수점(decimal point) 후에 숫자(numeric character)에서 하위 구분(subdivision)의 목록일 수 있다. 유행성 질환과 중증 질환이 코어 분류의 제목으로 선택되었고 보고에 필수적이며, 최대 10개의 하위 구분이 소수점 이후 숫자로 나타날 수 있다. 하위 카테고리는 다양한 부위 또는 상태의 그룹의 상세 타임 및 다양성에 대해 사용될 수 있다. 예시적으로 장티푸스열에 대해 'A01.0', 림프절 폐스에 대해 'A20.0' 코드가 사용될 수 있다. 도 6에 도시된 'A31.9' 코드는 비전형적 마이코박테리아 감염(Atypical mycobacterial infection, NOS(Not Otherwise Specified))을 나타낼 수 있다. 진단 코드(DICD)는 의사(physicians)에 의해 질병 상세가 기록되지 않는 경우가 많아 3자리 숫자로 기록될 수 있는데, 소수점 이후 숫자는 기관(institution)(예를 들어, 병원) 용으로 확장된 의미를 계층적으로 나타낼 수 있다.

[0090] 일 실시예에 따르면 컴퓨팅 장치는 진단 노트에 의해 지시되는 진단 코드(DICD)에 대응하는 제1 문자열을 추출할 수 있다. 컴퓨팅 장치는 공통 데이터 모델로부터 추출된 제1 문자열을 가지는 제1 개념 고유 식별자(CUI)를 검색할 수 있다. 도 6에 도시된 예시에서, 컴퓨팅 장치는 진단 노트(610)의 진단 코드(DICD) 'A31.9'와 동일한 문자열 'Atypical mycobacterial infection NOS'을 가지는 UMLS의 컬럼(631)으로부터 개념 고유 식별자(CUI) 'C0026919'를 검색할 수 있다. 컴퓨팅 장치는 같은 개념 고유 식별자(CUI)를 갖는 다른 코드들 및 문자열들을

추출함으로써 진단 코드(DICD)의 다양한 표현을 획득할 수 있다. 예시적으로 컴퓨팅 장치는 같은 개념 고유 식별자(CUI)에 속하는 다른 문자열 고유 식별자(SUI)(예를 들어, 도 6에 도시된 'S0016794', 'S0016795' 등) 및 다른 동의어 고유 식별자(LUI)(예를 들어, 도 6에 도시된 'L0026919', 'L10114762', 및 'L6525836' 등)에 대응하는 정보(632)를 획득할 수 있다. 컴퓨팅 장치는 UMLS MeSH 코드에 기반하여 같은 개념 고유 식별자(CUI)에 속하는 문자열 고유 식별자(SUI) 및 동의어 고유 식별자(LUI)에 대응하는 질병 식별자(Disease_ID)를 결정할 수 있다. 컴퓨팅 장치는 검색된 제1 개념 고유 식별자(CUI)에 대응하는 질병 식별자(Disease_ID)를 지시하는 질병 노드에 진단 노드를 매핑할 수 있다. 따라서, 컴퓨팅 장치는 하나의 진단 노드에, 동의어 및 유의어에 대응하는 질병 노드들을 매핑할 수 있다. 또한, 처방된 약의 상표명이 컨셉 이름으로 쓰이는 경우, 약의 상표명에 대응하는 개념 고유 식별자(CUI)에 포함된 동의어 고유 식별자(LUI) 및 문자열 고유 식별자(SUI)에 대응하는 질병 식별자(Disease_ID)가 매핑될 수 있다.

[0091] MeSH 용어에 의해 진단 코드(DICD)의 61% 정도가 매핑되는데, 매핑되지 않은 나머지 진단 코드(DICD)는 소스 매핑에 기초하여 질병 식별자(Disease_ID)와 매핑될 수 있다. 예를 들어, 컴퓨팅 장치는 MeSH 코드 매핑 이후 남은 나머지 진단 코드(DICD)에 대해 질병 온톨로지(Disease Ontology) 및 Mondo에서 정의되는 필드(field)에 미리 매핑된 질병 식별자(Disease_ID)를 매핑할 수 있다. 예를 들어, Bell's palsy에 대해 스트링 매핑이 불충분할 수 있는데, 소스 매핑이 적용될 경우, 컴퓨팅 장치는 증상인 안면 마비(Facial paralysis) 및 가족유전(familial)의 개념을 가진 코드에 매핑할 수 있다.

[0092] 리얼 월드 데이터의 RxNorm에서 처방은 용량, 용법, 및 성분 등에 의해 분류되는 반면, 생의학 데이터의 MeSH에서 동일한 성분을 가진 약은 같은 MeSH 컨셉 코드를 가질 수 있다. 따라서 성분명 변환 필요하기 때문에, 컴퓨팅 장치는 OHDSI 데이터 소스를 이용하여 리얼 월드 데이터의 처방 코드(ODCD) 및/또는 생의학 데이터의 화학물질 식별자(Chemical_ID)에 대응하는 성분명 정보를 획득할 수 있다.

[0093] 일 실시예에 따른 컴퓨팅 장치는 리얼 월드 데이터의 처방 코드(ODCD) 및 생의학 데이터의 화학물질 식별자(Chemical_ID) 간의 매핑도 전술한 바와 유사하게 수행할 수 있다. 처방 코드(ODCD)는 RxNorm/RxNormExtension을 기반으로 생성될 수 있다. 처방 코드(ODCD)의 RxNorm의 어휘(vocabulary)는 용량, 용법, 성분 등에 의해 분류되기 때문에 복잡한 개념을 포함할 수 있다. 반면, MeSH에서는 동일한 성분을 가진 약이 같은 MeSH 컨셉 코드에 의해 지시되므로, 성분명 변환이 필요할 수 있다. 참고로, 컴퓨팅 장치는 도 7에 도시된 바와 같이, OHDSI의 'has_ingredient' 관계(700)를 이용한 요소 매핑(Element mapping)을 통해 RxNorm/RxNormExtension의 성분명을 추출할 수 있다.

[0094] 일 실시예에 따른 컴퓨팅 장치는 처방 노드에 의해 지시되는 처방 코드(ODCD)에 대응하는 제2 문자열을 추출할 수 있다. 예를 들어, 컴퓨팅 장치는 처방 코드(ODCD)에 따른 각 약의 성분명과 약명을 추출할 수 있다. 컴퓨팅 장치는 처방 코드(ODCD)에 대응하여 획득된 약의 성분명 및/또는 약명에 대해 전처리(예를 들어, 간단한 자연어 처리(Natural Language Processing, NLP), 대소문자 전환, 전치사의 위치 변경, 띄어쓰기 수정, 설명을 포함한 괄호 처리)를 수행하여 소문자로 전환된 제2 문자열을 추출할 수 있다.

[0095] 컴퓨팅 장치는 공통 데이터 모델로부터 추출된 제2 문자열에 대응하는 제2 개념 고유 식별자(CUI)를 검색할 수 있다. 컴퓨팅 장치는 예를 들어, 컴퓨팅 장치는 성분명에 대해서 OHDSI의 RxNormExtension을 성분명을 기준으로 RxNorm코드를 부여한 후 공통 데이터 모델인 UMLS MeSH 코드와 매핑할 수 있다. 다시 말해, 컴퓨팅 장치는 RxNorm과 용법 및 용량은 다르지만 동일한 성분을 가지는 RxNormExtension에 대해 동일한 성분을 가지는 RxNorm과 똑같은 MeSH 코드를 매핑할 수 있다. 컴퓨팅 장치는 처방 코드(ODCD)의 성분명을 UMLS에 스트링 매핑함으로써 각 성분명(예를 들어, 제2 문자열)에 대한 UMLS의 제2 개념 고유 식별자(CUI)를 획득할 수 있다.

[0096] 컴퓨팅 장치는 검색된 제2 개념 고유 식별자(CUI)에 대응하는 화학물질 식별자를 지시하는 화학물질 노드에 처방 노드를 매핑할 수 있다. 컴퓨팅 장치는 처방 코드(ODCD)에 대응하는 성분명의 다른 표현을 찾기 위해 UMLS에 있는 전체 어휘에 대해 같은 제2 개념 고유 식별자(CUI)를 가지는 개체 노드들을 추출할 수 있다. 컴퓨팅 장치는 전술한 바와 유사하게, 같은 제2 개념 고유 식별자(CUI)에 속하는 동의어 고유 식별자(LUI) 및 문자열 고유 식별자(SUI)에 대응하는 화학물질 노드를 처방 노드에 매핑함으로써 생의학 데이터 및 리얼 월드 데이터를 결합할 수 있다.

[0097] 도 8은 일 실시예에 따른 리얼 월드 데이터 및 생의학적 데이터 간의 매핑 결과를 설명한다.

[0098] 일 실시예에 따른 컴퓨팅 장치는 도 5 내지 도 7에서 전술한 매핑에 의해 결합된 리얼 월드 데이터(810) 및 생의학 데이터(820)를 포함하는 새로운 생의학적 데이터를 도시한다. 새로운 생의학적 데이터의 집합은 도 9 및

도 10에서 후술하는 바와 같이 새로운 생물학적 네트워크를 구성할 수 있다. 리얼 월드 데이터(810), 생의학 데이터(820), 새로운 생의학 데이터, 및 새로운 생물학적 네트워크는 그래프 구조로 구현될 수 있다.

[0099] 리얼 월드 데이터(810)는 전술한 바와 같이, 아웃컴 노드(815), 환자 식별자 노드(811), 환자 인카운터 노드(812), 처방 노드(813) 및 진단 노드(814)를 포함할 수 있다. 리얼 월드 데이터(810)는 아웃컴 노드(815) 및 환자 식별자 노드(811)를 연결하는 엣지, 환자 식별자 노드(811)와 환자 인카운터 노드(812)를 연결하는 엣지, 환자 인카운터 노드(812)와 처방 노드(813)를 연결하는 엣지, 및 환자 인카운터 노드(812)와 진단 노드(814)를 연결하는 엣지를 포함할 수 있다.

[0100] 생의학 데이터(820)는 전술한 바와 같이, 화학물질 노드(821), 모구조 노드(822), 질병 노드(823), 및 유전자 노드(824)를 포함할 수 있다. 생의학 데이터(820)는 화학물질 노드(821)와 모구조 노드(822)를 연결하는 엣지, 화학물질 노드(821)와 질병 노드(823)를 연결하는 엣지, 및 질병 노드(823)와 유전자 노드(824)를 연결하는 엣지를 포함할 수 있다.

[0101] 새로운 생의학 데이터는 전술한 리얼 월드 데이터(810) 및 생의학 데이터(820)의 개체 노드들 및 엣지들을 모두 포함할 수 있다. 또한, 새로운 생의학 데이터는 도 5 내지 도 7에서 전술한 매핑에 의해 새로 생성된 엣지들을 더 포함할 수 있다. 예를 들어, 새로운 생의학 데이터는 처방 노드(813)와 화학물질 노드(821)를 연결하는 엣지 및 진단 노드(815)와 질병 노드(823)를 연결하는 엣지를 포함할 수 있다. 처방 노드(813)와 화학물질 노드(821)를 연결하는 엣지 및 진단 노드(815)와 질병 노드(823)를 연결하는 엣지는 도 11에서 후술하는 랜덤워크 데이터를 생성할 시 사이클릭(cyclic)한 부분을 방지할 수 있다. 사이클은 랜덤워크의 출발 노드와 경유 노드가 동일한 경우를 의미하며, 랜덤워크에서 사이클이 자주 발생하는 경우 후술하는 벡터 임베딩 모델의 성능 하락이 발생할 수 있으므로 사이클릭을 방지하는 것이 필요하다.

[0102] 다만, 새로운 생물학적 데이터를 도 8에 도시된 바로 한정하는 것은 아니고, 환자 노드(예를 들어, 환자 식별자 노드(811) 및 환자 인카운터 노드(812))가 도 11에서 후술하는 벡터 임베딩 후 배제될 수 있다.

[0103] 도 9 및 도 10은 일 실시예에 따라 생성된 새로운 병합된 네트워크(merged network)를 설명한다.

[0104] 도 8에서는 이중 데이터의 결합에 의해 생성된 새로운 생의학 데이터를 설명했고, 도 9에서는 생의학 데이터 간의 연결을 설명한다. 도 9에 도시된 바와 같이, 제1 리얼 월드 데이터(911)는 제1 생의학 데이터(921)와 연결될 수 있다. 또한, 제1 생의학 데이터는 질병 식별자(Disease_ID) 및 화학물질 식별자(Chemical_ID)를 통해 다른 제2 리얼 월드 데이터(912) 및 모구조를 통해 제3 생의학 데이터(923)와도 연결될 수 있다. 제3 생의학 데이터(923)는 질병 식별자(Disease_ID) 및 화학물질 식별자(Chemical_ID)를 통해 제3 리얼 월드 데이터(913)와 연결될 수 있다. 도 9에 도시된 바와 같이 컴퓨팅 장치는 한 개체 노드에 복수의 개체 노드들을 엣지로 연결함으로써, 새로운 생물학적 네트워크를 구축할 수 있다.

[0105] 도 10은 새로 구축된 생물학적 네트워크(1000)의 예시를 설명한다. 예시적으로 새로 구축된 생물학적 네트워크(1000)의 개체 노드 및 엣지의 개수는 하기 표 1과 같이 나타낼 수 있다.

표 1

[0106]

Datasets	Entities and relationships				
EMR	Entities	PAID	KEY	DICD	ODCD
		360409	14967169	1025	4798
	Relationships (edge)	PAID - KEY	KEY- ODCD	KEY- DICD	KEY- CDTH
CTD	Entities	DiseaseID	ChemicalID	GeneSymbol	Parent ID
		7643	1805	1496	733
	Relationships (edge)	DiseaseID - ChemicalID	DiseaseID - GeneSymbol	ChemicalID - Parent ID	ChemicalID - GeneSymbol
Mapping	Relationships (edge)	ODCD - ChemicalID	DICD-DiseaseID		
		6633	7989		

[0107] 도 11은 일 실시예에 따른 기계 학습 모델의 트레이닝 및 트레이닝된 기계 학습 모델을 이용한 생의학적 연관성을 설명한다. 일 실시예에 따르면 컴퓨팅 장치는 구축된 새로운 생물학적 네트워크로부터 랜덤워크에 기반하여

생성된 랜덤워크 데이터(1120)를 이용하여, 기계 학습 모델을 트레이닝시킬 수 있다.

[0108] 컴퓨팅 장치는 도 10에서 구축된 새로운 생물학적 네트워크로부터 랜덤워크에 기반하여 랜덤워크 데이터(1120)를 생성할 수 있다. 본 명세서에서 리얼 월드 데이터(예를 들어, 전자의무기록 데이터)와 통합된 새로운 생물학적 네트워크의 그래프는 $G = (v, \varepsilon)$ 로 정의될 수 있다. 여기서, 노드 집합 v 는 새로운 생물학적 데이터로서, 리얼 월드 데이터의 5개 개체 타입의 개체 노드(1110) 및 생의학 데이터의 4개 개체 타입의 개체 노드(1110), 총 9개의 이종 개체 노드(1110)(heterogeneous entity node)의 셋트를 나타낼 수 있다. 예시적으로, 제1 개체 타입의 개체 노드 셋트 $v_1 = \{v_{11}, \dots, v_{1n}\}$, 내지 제9 개체 타입의 개체 노드 셋트 $v_9 = \{v_{91}, \dots, v_{9n}\}$ 로 표현될 수 있다. 노드 집합 v 는 총 V 개의 노드 개체들을 포함할 수 있고, 여기서, n 및 V 는 1이상의 정수일 수 있다. 개체 노드 셋트는 개체 간의 인접성을 나타내는 인접 행렬로부터 생성될 수 있다. 예를 들어, 인접 행렬의 각 행 및 각 열은 i 번째 개체 및 j 번째 개체를 나타내고, 인접 행렬의 i 번째 행의 j 번째 열의 원소 값은 i 번째 개체 및 j 번째 개체 간의 인접성(예를 들어, 인접 여부)을 나타낼 수 있다. 예시적으로 v_{ij} 는 제 i 개체 타입의 제 j 인스턴스(instance)의 개체 노드(1110)를 지시할 수 있다. 엣지 ε 는 개체 노드(1110) v_{ij} 간을 연결하는 엣지로서, 가중치 또는 방향 없이 데이터 셋트에서 연관성(association) 및/또는 관련성(relationship)을 나타낼 수 있다. 엣지는 $\varepsilon \subseteq V \times V$ 로 표현될 수 있다.

[0109] 랜덤워크 데이터(1120)는 이웃하는 개체 노드들 간의 랜덤워크에 기초한 랜덤 시퀀스를 나타내는 데이터일 수 있다. 그래프 구조의 데이터 네트워크에 대한 랜덤워크는, 데이터 네트워크를 구성하는 임의의 개체 노드(1110)로부터 해당 개체 노드(1110)에 엣지를 통해 연결된 하나 이상의 개체 노드(1110) 중 무작위로 선택된 개체 노드(1110)를 선택 및 추출하는 동작을 나타낼 수 있다. 랜덤워크 데이터(1120)의 길이는 랜덤워크 길이, 다시 말해 랜덤워크의 시행 횟수에 따라 결정될 수 있다. 랜덤워크의 방향 및 확률은 엣지의 방향 및 가중치에 따라 결정될 수 있는데, 전술한 바와 같이, 본 명세서에서는 설명의 편의를 위해 방향이 없고 가중치는 모두 동일한 것을 가정한다. 랜덤워크 데이터(1120)를 랜덤워크 입력(random walk input)이라고도 나타낼 수 있다.

[0110] 랜덤워크 입력은 그래프 네트워크의 위상 정보(topological information)를 캡처하기 위해 생성된 데이터로서, 2차 랜덤워크(2^{nd} order random walk) 데이터라고도 나타낼 수 있다. 새로운 생물학적 네트워크에 포함된 개체 노드들 간의 유사성(similarity)은 주어진 특징 표현(feature representation), 예를 들어, 임베딩 벡터에서 이웃하는 개체 노드(1110)의 발생 확률로서 정의될 수 있다. 이웃하는 개체 노드(1110)의 발생 확률은 이웃하는 개체 노드(1110) 간에 엣지가 존재할 확률이라고도 해석될 수 있다. 컴퓨팅 장치는 그래프 임베딩을 위해 랜덤워크 셋트 $N_s(u)$ 를 생성할 수 있다. 컴퓨팅 장치는 랜덤워크를 통해 초기 노드 u 로부터 시작하여, 랜덤워크 전략 S 를 이용하여 미리 결정된 랜덤워크 길이를 통해 이웃하는 노드로 워크를 생성할 수 있다. 여기서, u 는 $u \in v$ 일 수 있다. 후술하겠으나, 랜덤워크 입력으로 표현되는 랜덤워크 셋트 $N_s(u)$ 내에서, 이웃하는 노드들의 발생 확률이 임의로 출력된 임베딩 벡터에 기초하여 추정될 수 있고, 임베딩 공간의 유사도에 대해 최적 확률으로써, 후술하는 기계 학습 모델(1140)이 트레이닝될 수 있다.

[0111] 예를 들어, 컴퓨팅 장치는 구축된 새로운 생물학적 네트워크에서 초기 개체 노드 u 를 선정할 수 있다. 컴퓨팅 장치는 선정된 초기 개체 노드 u 로부터 미리 결정된 랜덤워크 길이만큼 순차적으로 랜덤워크를 수행함으로써, 개별 랜덤워크에 따른 개체 노드를 선택할 수 있다. 컴퓨팅 장치는 초기 개체 노드 u 및 랜덤워크에 기초하여 선택된 개체 노드를 포함하는 랜덤워크 셋트 $N_s(u)$ 를 지시하는 랜덤워크 데이터(1120)를 생성할 수 있다.

[0112] 랜덤워크 길이는 새로운 생물학적 네트워크를 구성하는 개체 노드들의 총 개체 타입 개수보다 큰 값을 가질 수 있다. 예를 들어, 도 11에서는 랜덤워크 길이가 10인 예를 설명한다. 예를 들어, 랜덤워크를 생성함에 있어서, 워크 길이는 10으로 설정될 수 있고, 개체 타입들의 380개 다른 조합들이 생성될 수 있다. 예를 들어, 랜덤워크 셋트 $N_s(u)$ 로서, $[v_{1n}, v_{2n}, v_{1n}, v_{2n}, v_{1n}, v_{2n}, v_{1n}, v_{2n}, v_{1n}, v_{2n}]$ 내지 $[v_{9n}, v_{8n}, v_{9n}, v_{8n}, v_{9n}, v_{8n}, v_{5n}, v_{7n}, v_{9n}, v_{8n}]$ 가 생성될 수 있다. 다른 예를 들어, 랜덤워크를 생성함에 있어서, 대략적으로 4개 개체 타입들 및 각 개체 노드(1110)에 대해 40개 조합이 생성될 수 있다.

[0113] 참고로, 리턴 파라미터 p 및 인아웃 파라미터 q 는 BFS(Breadth-first Sampling) 및 DFS(Depth-first Sampling)와 같은 샘플링 전략을 위한 서치를 가이드하기 위해 결정될 수 있다. 탐색 바이어스(search bias)를 엣지의 전이 확률에 부가하기 위해, 이전 노드 t 가 기억되고, 탐색 바이어스가 노드 t 에 비교하여 상대적 위치에 의존하여 계산될 수 있다. 다음 노드 x 에 대하여, 노드 t 및 노드 x 간의 가장 짧은 경로의 값들이 $d_{tx} = \{0, 1, 2\}$

로 고정되어, 엣지 (v, x) 상의 탐색 바이어스 $\alpha_{pq}(t, x)$ 가 계산될 수 있다. 탐색 바이어스 $\alpha_{pq}(t, x)$ 는 하기 수학적 식 1과 같이 결정될 수 있다.

수학적 식 1

$$\alpha_{pq}(t, x) = \begin{cases} \frac{1}{p} & \text{if } d_{tx} = 0 \\ 1 & \text{if } d_{tx} = 1 \\ \frac{1}{q} & \text{if } d_{tx} = 2 \end{cases}$$

[0114]

이전 노드에 더 가까워지면 p 의 역수, 이전 노드로부터 멀어지면 q 의 역수가 적용될 수 있다. 전술한 수학적 식 1에서, 예시적으로 새로운 생물학적 네트워크는 리턴 파라미터 $p=1$ 및 인아웃 파라미터 $q=1$ 을 갖는 엣지들을 가지므로, 전이 확률은 d_{tx} 와 무관하게, $\alpha_{pq}(t, x) = 1$ 로 설정될 수 있다

[0116]

컴퓨팅 장치는 전술한 랜덤워크 입력을 이용하여 기계 학습 모델(1140)을 트레이닝시킬 수 있다. 기계 학습 모델(1140)은 랜덤워크 입력에 의해 지시되는 각 개체 노드(1110)를 잠재 공간(latent space)의 임베딩 벡터로 투사(project)하도록 설계 및 트레이닝된 모델로서, 예를 들어, 뉴럴 네트워크 모델을 포함할 수 있다. 본 명세서에서는 하나의 히든 레이어를 포함하는 "얕은 뉴럴 네트워크"(Shallow Neural Network)를 주로 설명하나, 이로 한정하는 것은 아니다. 기계 학습 모델(1140)의 은닉층은 록업 테이블이라는 연산을 담당하며, 투사층(projection layer)라고도 나타낼 수 있다. 기계 학습 모델(1140)은 네거티브 샘플링을 이용한 스킵 그래프(skip-gram) 구조로 설계될 수 있다. 다만, 기계 학습 모델(1140)의 구조를 한정하는 것은 아니고, 컨볼루션 뉴럴 네트워크(CNN, convolutional neural network), 및 순환 뉴럴 네트워크(RNN, recurrent neural network) 등 다양한 구조로 구현될 수 있다. 또한, 기계 학습 모델(1140)은 그래프 구조의 데이터를 임베딩하는 그래프 임베딩 모델을 포함하고, 예를 들어, deep walk, node2vec, gcn 등을 포함할 수 있다.

[0117]

컴퓨팅 장치는 도 11에 도시된 바와 같이, 랜덤워크 입력의 각 개체 노드(1110)에 대응하는 입력 데이터(1130)를 기계 학습 모델(1140)에 입력할 수 있다. 컴퓨팅 장치는 랜덤워크 입력에 포함된 복수의 개체 노드들의 각각에 대응하는 원핫벡터(one-hot vector)를 입력 데이터(1130)로서 기계 학습 모델(1140)에 입력할 수 있다. 도 11에서는 랜덤워크 셋트 $N_S(u)$ 로서 $[v_{9n}, v_{8n}, v_{9n}, v_{8n}, v_{9n}, v_{8n}, v_{5n}, v_{7n}, v_{9n}, v_{8n}]$ 인 랜덤워크 입력을 예로 들어 설명한다. 예시적으로 기계 학습 모델(1140)에 입력되는 입력 데이터(1130)(예를 들어, 원핫 벡터)의 차원은 전술한 도 10에서 생성된 새로운 생물학적 네트워크(1000)의 개체 노드들의 개수와 같을 수 있다. 새로운 생물학적 네트워크(1000)를 구성하는 개체 노드들의 총 개수가 V 인 경우, 입력 데이터(1130)인 원핫 벡터의 차원도 V 일 수 있다. 여기서, V 는 2이상의 정수일 수 있다. 컴퓨팅 장치는 랜덤워크 입력에 기초하여, 랜덤워크 입력에 포함된 각 개체 노드(1110)를 지시하는 원핫 벡터를 생성할 수 있다. 한 개체 노드(1110)를 지시하는 원핫 벡터는 해당하는 개체 노드(1110)를 지시하는 원소 값이 '1'이고, 나머지 노드를 지시하는 원소 값이 '0'인 벡터를 나타낼 수 있다. 도 11에서, 컴퓨팅 장치는 랜덤워크 입력 중 랜덤워크 셋트 $N_S(u)$ 에 대해 입력 데이터(1130) 중 개체 노드 v_{9n} 에 대응하는 원소 값이 '1'이고 나머지 원소 값이 '0'인 원핫 벡터, 개체 노드 v_{8n} 에 대응하는 원소 값만 '1'인 원핫 벡터, 개체 노드 v_{5n} 에 대응하는 원소 값만 '1'인 원핫 벡터 등을 생성할 수 있다. 컴퓨팅 장치는 생성된 원핫 벡터를 이용하여 기계 학습 모델(1140)을 트레이닝시킬 수 있다.

[0118]

기계 학습 모델(1140)의 투사층 f 를 포함하는 경우, 파라미터, 예시적으로, 투사층 f 와 입력층(input layer) 사이의 가중치 행렬 W 는 $V \times d$ 차원의 행렬일 수 있다. 여기서, d 는 1이상의 정수일 수 있다. 다시 말해, 컴퓨팅 장치는 기계 학습 모델(1140)의 투사층 f 을 통해, V 차원의 원핫 벡터를 d 차원으로 투사함으로써, d 차원의 임베딩 벡터를 생성할 수 있다. 예를 들어, 컴퓨팅 장치는 랜덤워크 셋트 $N_S(u)$ 에 대응하는 원핫 벡터들로부터 기계 학습 모델(1140)을 이용하여, 출력 데이터들 $f(v_{9n})$, $f(v_{8n})$, 및 $f(v_{5n})$ 을 생성할 수 있다. 출력 데이터들 $f(v_{9n})$, $f(v_{8n})$, 및 $f(v_{5n})$ 은 전술한 바와 같이 M 차원의 임베딩 벡터로서, d 차원 잠재 공간에서 각 개체 노드(1110)의 벡터 좌표를 나타낼 수 있다.

[0119]

전술한 바와 같이 기계 학습 모델(1140)은 스킵 그래프 모델(skip-gram model)로서, 개체 노드들 간의 관계를 계산하기 위한 임베딩된 표현(embedded representation)을 각 개체 노드(1110)에 대해 출력하도록, 랜덤워크

(random walk)에 기초한 시퀀스들을 학습할 수 있다. 기계 학습 모델(1140)의 학습 및/또는 트레이닝은 후술하는 목적함수 값(1109)에 기초하여 수행될 수 있다. 도 10에서 전술한 생물학적 네트워크(1000)에 대응하는 그래프 구조 데이터 G 는 $G = (v, \varepsilon)$ 로 표현될 수 있다. 투사층 f 는 매핑함수로서, $f: v \rightarrow \mathbb{R}^d$ 로 표현될 수 있다. 여기서, $\mathbb{R}$ 은 노드 v 를 d 차원의 벡터 공간(예를 들어, 잠재 공간 및/또는 임베딩 공간)으로 매핑하는 $V \times d$ 차원의 가중치 행렬의 파라미터들을 나타낼 수 있다. 노드를 관찰하는 확률, 다시 말해, 두 노드를 연결하는 엣지가 존재할 확률은 하기 수학식 2과 같은 문제를 최대화하는 것으로 계산될 수 있다.

수학식 2

$$\max_f \sum_{u \in V} \log \Pr(N_S(u) | f(u))$$

[0120]

[0121] 전술한 수학식 2는 랜덤워크 셋트 $N_S(u)$ 에 포함된 개체 노드들 간의 연결 확률을 최대화하는 목적함수 값(1109)으로 해석될 수 있다. 전술한 수학식 2를 간소화하기 위해, 노드들의 각 발생 확률은 다른 노드들의 발생 확률에 독립적이며, 2개의 노드 페어에 대해 서로에 대한 효과는 대칭적으로 동일한 것으로 가정될 수 있다. 전술한 가정 및 네거티브 샘플링을 이용한 근사화(approximation) 하에서, 수학식 2의 목적함수 값(1109)은 특징들의 도트 곱으로 하기 수학식 3과 같이 간소화될 수 있다.

수학식 3

$$\max_f \sum_{u \in V} \left[-\log Z_u + \sum_{n_i \in N_S(u)} f(n_i) \cdot f(u) \right]$$

[0122]

[0123] 전술한 수학식 3은 임베딩 공간에서 노드들의 유사도 및/또는 노드들 간의 엣지 존재 확률을 나타내는 수학식 2를 최대화하는 확률적 경사 상승(stochastic gradient ascent)에 의해 f 의 파라미터를 최적화하는데 사용될 수 있다.

[0124] 전술한 바와 같이, 컴퓨팅 장치는 랜덤워크 데이터(1120)에 기초한 입력 데이터(1130)에 기계 학습 모델(1140)을 적용함으로써 임시 임베딩 데이터를 산출할 수 있다. 임시 임베딩 데이터는 트레이닝이 완료되기 전의 기계 학습 모델(1140), 다시 말해 임시 모델을 이용하여 입력 데이터(1130)(예를 들어, 임의의 개체 노드(1110)를 지시하는 원핫 벡터)로부터 추출된 임베딩 벡터를 나타낼 수 있다. 컴퓨팅 장치는 산출된 임시 임베딩 데이터에 기초하여 랜덤워크 셋트 $N_S(u)$ 내 개체 노드들 간의 연결 확률이 최대화되도록 기계 학습 모델(1140)의 파라미터를 업데이트할 수 있다. 따라서, 컴퓨팅 장치는 랜덤워크를 통해, 새로운 생물학적 네트워크(1000)에서 개체 노드들 간에 추출된 엣지 관계를 기계 학습 모델(1140)에 트레이닝시킴으로써, 서로 엣지를 통해 연결된 개체 노드들의 잠재 공간에서의 벡터 좌표가 서로 인접하게 출력되도록 기계 학습 모델(1140)의 파라미터를 업데이트할 수 있다. 여기서, 기계 학습 모델(1140)의 투사층 f 는 $V \times d$ 차원의 가중치 행렬일 수 있는데, 가중치 행렬의 각 행(row)에 대응하는 행 벡터(row vector)는 해당하는 개체 노드(1110)를 표현한 임베딩 벡터일 수 있다. 다시 말해, 가중치 행렬의 V 개의 복수의 행 벡터들 중 제 k 행 벡터는, 새로운 생물학적 네트워크(1000)의 V 개의 개체 노드들 중 제 k 개체 노드(1110)를 표현한 임베딩 벡터일 수 있다. 여기서, k 는 1이상 V 이하의 정수일 수 있다.

[0125] 컴퓨팅 장치는, 일정 비율(예를 들어, 10%)의 엣지가 무작위로 제거된 새로운 생물학적 네트워크(1000)에 대해 전술한 기계 학습 모델(1140)을 이용하여 링크를 예측함으로써, 기계 학습 모델(1140)의 성능을 검증할 수 있다.

[0126] 도 12는 일 실시예에 따라 잠재 공간(latent space)에 임베딩된 개별 개체 노드(entity node)에 대응하는 임베딩 표현 벡터를 예시적으로 도시한다.

- [0127] 약물의 새로운 적응증을 탐색하는 것을 약물 재창출 또는 신약재 창출이라고 나타낼 수 있다. 약물의 새로운 적응증이란 시판중인 약물의 기존 적응증과 다른 새로운 적응증(예를 들어, 질환)을 의미하거나, 임상시험에서 실패한 약물의 새로운 대상 환자군 선정을 의미할 수 있다. 신약 재창출은 기존 신약 개발 Process에 비해 적은 비용으로 실패율을 낮추고 보다 신속하게 출시할 수 있다는 장점이 있다.
- [0128] 전술한 도 10에서 구축된 새로운 생물학적 네트워크(1000)에서 관련성에 따라 개체 노드들이 엣지를 통해 연결될 수 있다. 특히, 질병 노드와 연결된 화학물질 노드는 해당 화학물질 노드에 대응하는 화학물질이 질병 노드에 대응하는 질병에 적응증이 있는 것으로 해석될 수 있다. 일 실시예에 따르면, 새로운 생물학적 네트워크(1000)가 화학물질 노드 및 질병 노드 간의 직접적인 엣지 이외에도, 모구조 노드, 처방 노드, 및 진단 노드 등 다른 개체 노드와의 직간접적인 연결을 나타내는 엣지를 포함하는 바, 도 11에서 전술한 기계 학습 모델(1140)은 새로운 생물학적 네트워크(1000)에서 밝혀지지 않은(undiscovered) 미성 엣지를 예측하도록 설계 및 트레이닝된 모델일 수 있다.
- [0129] 일 실시예에 따르면 새로운 생물학적 네트워크(1000)의 모든 개체 노드들의 각각에 대응하는 임베딩 벡터가, 잠재 공간(1200) 내 벡터 좌표로 매핑될 수 있다. 잠재 공간(1200)에서 서로 인접한 개체 노드들 사이에는 엣지가 존재하는 것으로 결정될 수 있다. 예를 들어, 도 12에서 컴퓨팅 장치는 잠재 공간(1200)에서 제1 개체 노드(1210) 및 제2 개체 노드(1220) 간의 거리가 임계 거리 미만인 경우, 제1 개체 노드(1210) 및 제2 개체 노드(1220) 간에 엣지(1290)가 존재하는 것으로 결정할 수 있다. 따라서, 컴퓨팅 장치는 새로운 생물학적 네트워크(1000)에서 미리 구축된 데이터베이스에 의해 식별된 명시적 엣지 외에도, 잠재 공간(1200) 내 개체 노드들에 대응하는 임베딩 벡터 간의 벡터 거리에 따라 잠재적인 엣지를 탐색할 수 있다.
- [0130] 예를 들어, 컴퓨팅 장치는 새로운 생물학적 네트워크의 개체 노드들 중 제1 노드(예를 들어, 대상 질병 노드)에 대응하는 제1 임베딩 벡터를 기계 학습 모델로부터 추출할 수 있다. 컴퓨팅 장치는 새로운 생물학적 네트워크의 개체 노드들 중 제2 노드(예를 들어, 대상 화학물질 노드)에 대응하는 제2 임베딩 벡터를 기계 학습 모델로부터 추출할 수 있다. 컴퓨팅 장치는 추출된 제1 임베딩 벡터 및 제2 임베딩 벡터에 기초하여, 제1 노드 및 제2 노드 간의 엣지 존재 여부를 결정할 수 있다. 임베딩 벡터들 간의 차이가 작아서, 임베딩 벡터들이 유사한 경우 두 임베딩 벡터 사이에는 잠재적인 엣지 및/또는 링크가 존재하는 것으로 예측될 수 있다. 컴퓨팅 장치는 추출된 제1 임베딩 벡터 및 제2 임베딩 벡터 간의 벡터 거리 차이가 임계 거리 값 미만인 경우에 응답하여, 제1 노드 및 제2 노드 간에 엣지가 존재한다고 결정할 수 있다. 여기서, 두 임베딩 벡터들 간의 벡터 거리는 L2 노름과 같은 유클리드 거리일 수 있다. 전술한 바와 같이, 컴퓨팅 장치는 미리 구축된 생물학적 네트워크에서 링크가 존재하지 않더라도, 잠재 공간 내 임베딩 벡터 간의 거리에 기초하여 잠재적인 링크를 예측함으로써, 생의학적 연관성을 예측할 수 있다.
- [0131] 참고로, 전술한 예시에서, 제1 노드 및 제2 노드가 각각 질병-약물이면 컴퓨팅 장치는 약물재창출을 수행하게 된다. 제1 노드 및 제2 노드가 약물-약물이 되면 컴퓨팅 장치는 약물-약물 상호작용 예측을 수행하게 된다. 제1 노드 및 제2 노드가 각각 질병-질병이 되면 컴퓨팅 장치는 질병-질병 연관성 탐색을 수행하게 된다. 제1 노드 및 제2 노드가 각각 환자-질병이 되면 컴퓨팅 장치는 대상 환자의 새로운 질병의 발병을 예측하게 된다.
- [0132] 예를 들어, 컴퓨팅 장치는 임의의 대상 질병을 지시하는 질병 노드에 대응하는 제1 임베딩 벡터를 기준으로 임계 거리 값 미만인 화학물질 노드를 검색할 수 있다. 따라서, 대상 질병에 대한 효과를 가지는 약물이 탐색될 수 있다. 다른 예를 들어, 컴퓨팅 장치는 임의의 대상 화학물질을 지시하는 화학물질 노드에 대응하는 제2 임베딩 벡터를 기준으로 임계 거리 값 미만인 질병 노드를 검색할 수 있다. 따라서, 대상 화학물질에 대해 새로운 적응증에 해당하는 질병이 탐색될 수 있다.
- [0133] 이상에서 설명된 실시예들은 하드웨어 구성요소, 소프트웨어 구성요소, 및/또는 하드웨어 구성요소 및 소프트웨어 구성요소의 조합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치, 방법 및 구성요소는, 예를 들어, 프로세서, 콘트롤러, ALU(arithmetic logic unit), 디지털 신호 프로세서(digital signal processor), 마이크로컴퓨터, FPGA(field programmable gate array), PLU(programmable logic unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(OS) 및 상기 운영 체제 상에서 수행되는 소프트웨어 애플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설명된 경우도 있지만, 해당 기술분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소(processing element) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알 수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하

나의 프로세서 및 하나의 컨트롤러를 포함할 수 있다. 또한, 병렬 프로세서(parallel processor)와 같은, 다른 처리 구성(processing configuration)도 가능하다.

[0134] 소프트웨어는 컴퓨터 프로그램(computer program), 코드(code), 명령(instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성요소(component), 물리적 장치, 가상 장치(virtual equipment), 컴퓨터 저장 매체 또는 장치, 또는 전송되는 신호 파(signal wave)에 영구적으로, 또는 일시적으로 구체화(embody)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

[0135] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있으며 매체에 기록되는 프로그램 명령은 실시예를 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic media), CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media), 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다.

[0136] 위에서 설명한 하드웨어 장치는 실시예의 동작을 수행하기 위해 하나 또는 복수의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

[0137] 이상과 같이 실시예들이 비록 한정된 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 이를 기초로 다양한 기술적 수정 및 변형을 적용할 수 있다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다.

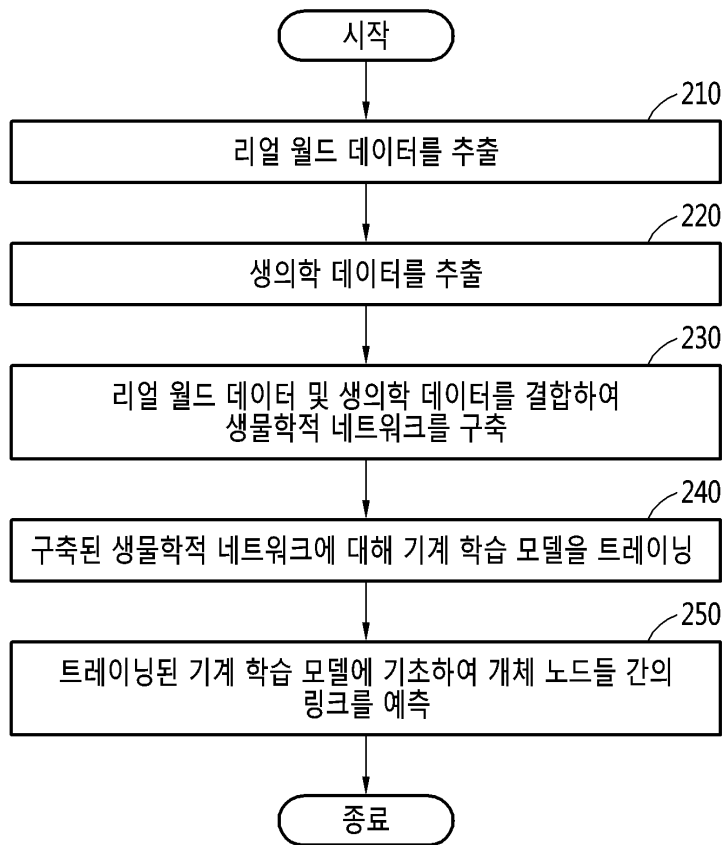
[0138] 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 특허청구범위의 범위에 속한다.

도면

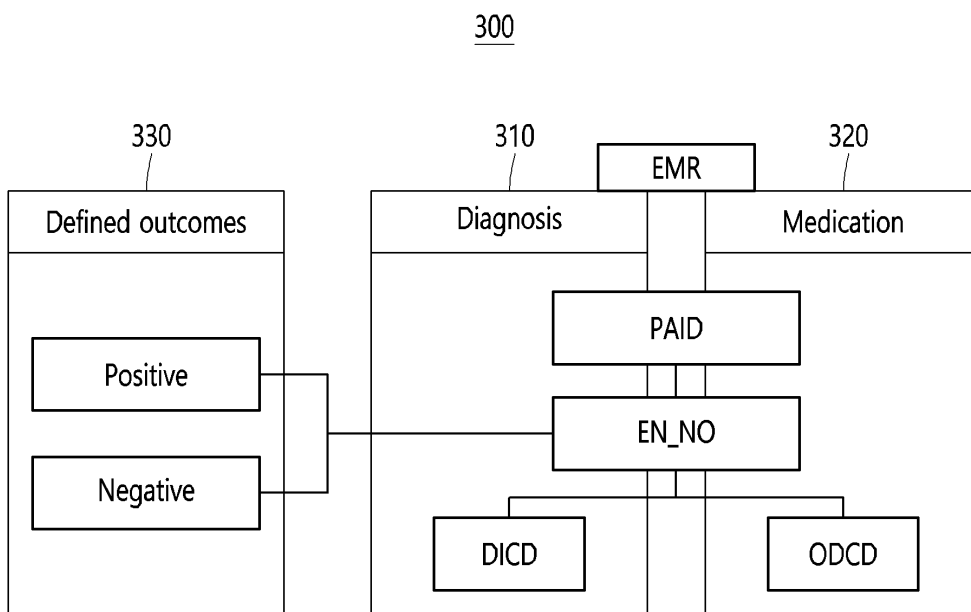
도면1



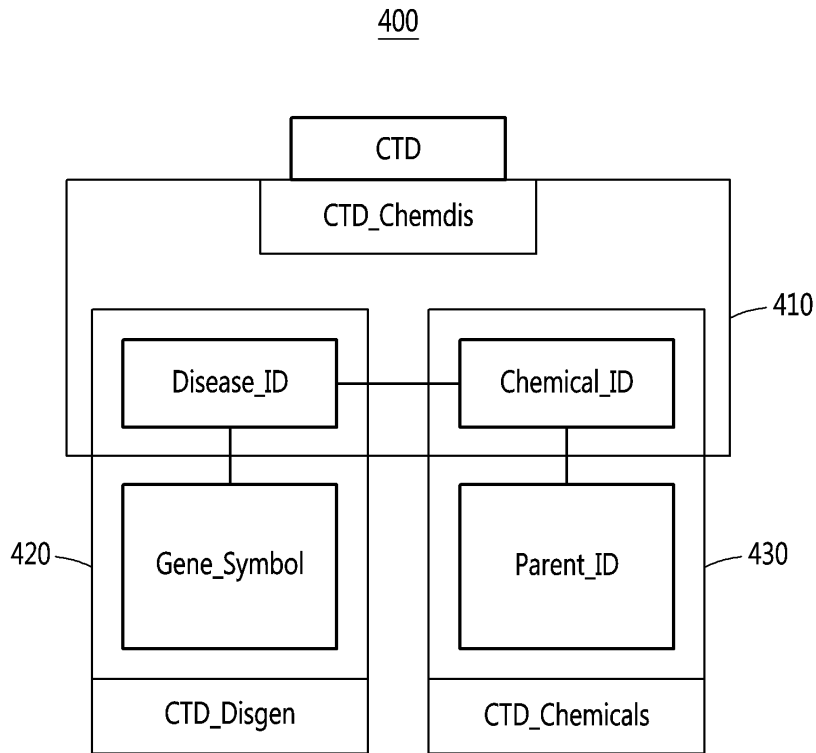
도면2



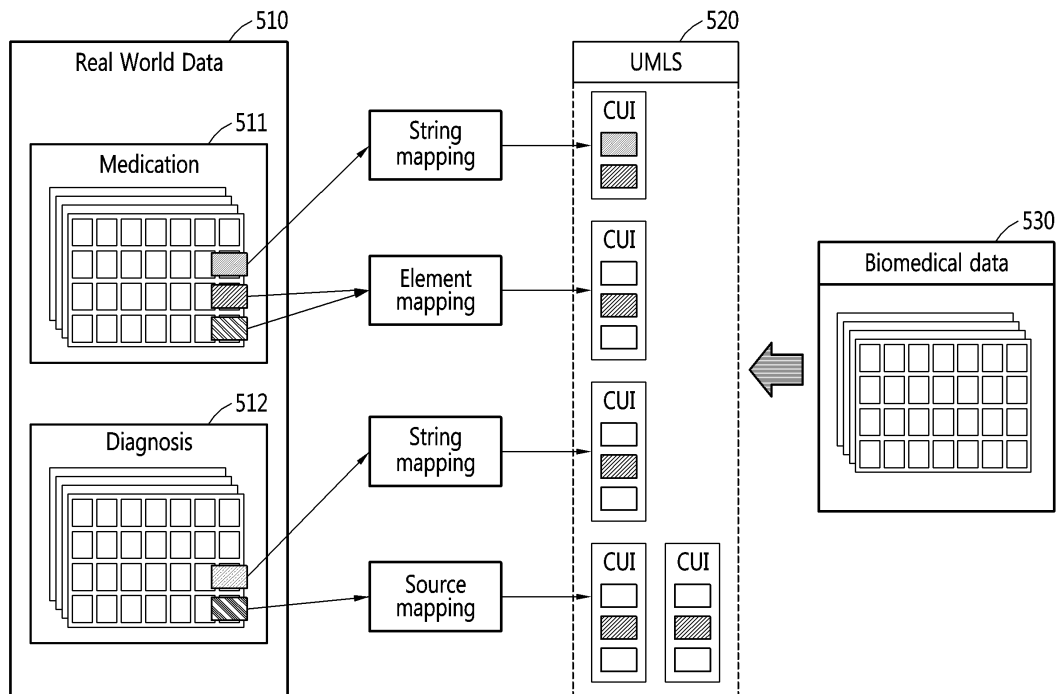
도면3



도면4



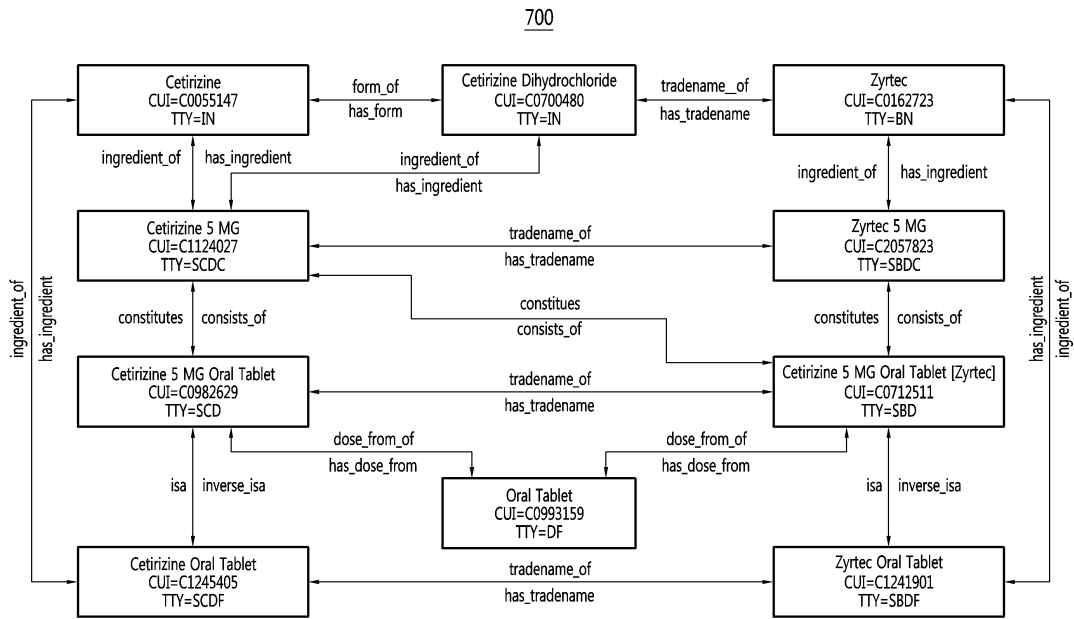
도면5



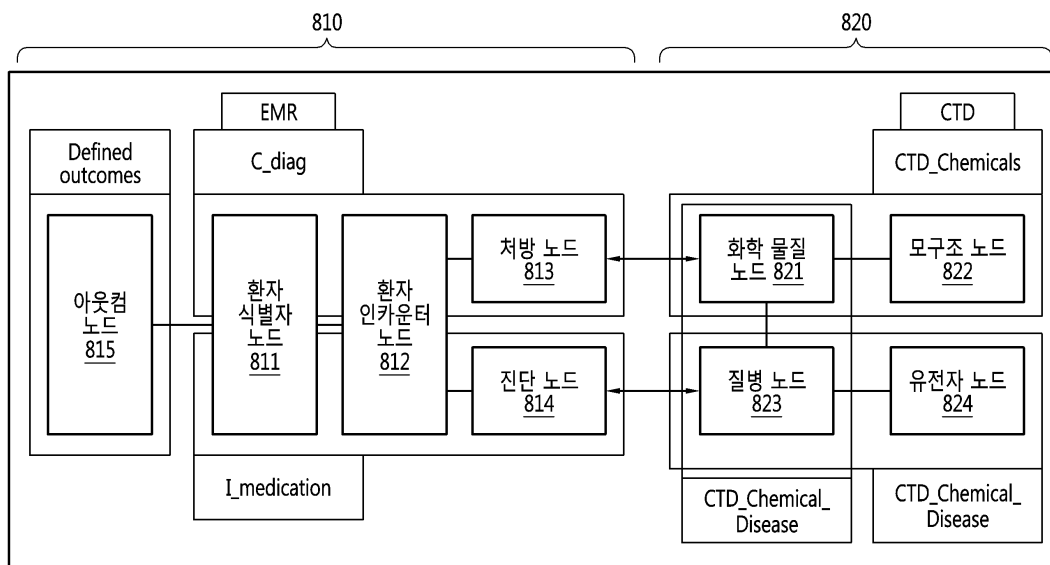
도면6

CUI	LUI	SUI	SAB	CODE	STR
C0026919	L0026909	S1623743	ICD10CM	A31.9	Atypical mycobacterial infection NOS
C0026919	L0026909	S1623743	MDR	10003753	Atypical mycobacterial infection NOS
C0026919	L0026919	S0016794	MSH	D009165	Atypical Mycobacterium Infection
C0026919	L0026919	S0016795	MSH	D009165	Atypical Mycobacterium Infections
C0026919	L0026919	S0052482	MSH	D009165	Infections, Atypical Mycobacterium
C0026919	L0026919	S0064433	MSH	D009165	Mycobacterium Infections, Atypical
C0026919	L0026919	S0388718	MSH	D009165	Mycobacterium Infection, Atypical
C0026919	L10114762	S12642904	MSH	D009165	Mycobacterium Infection, Nontuberculous
C0026919	L10114762	S12642905	MSH	D009165	Mycobacterium Infections, Nontuberculous
C0026919	L10114762	S12643537	MSH	D009165	Nontuberculous Mycobacterium Infections
C0026919	L10114762	S12643538	MSH	D009165	Nontuberculous Mycobacterium Infection
C0026919	L6525836	S7590273	MSH	D009165	ATYPICAL MYCOBACT INFECT
C0026919	L6525836	S7597239	MSH	D009165	INFECT ATYPICAL MYCOBACT
C0026919	L6525836	S7599515	MSH	D009165	MYCOBACT INFECT ATYPICAL

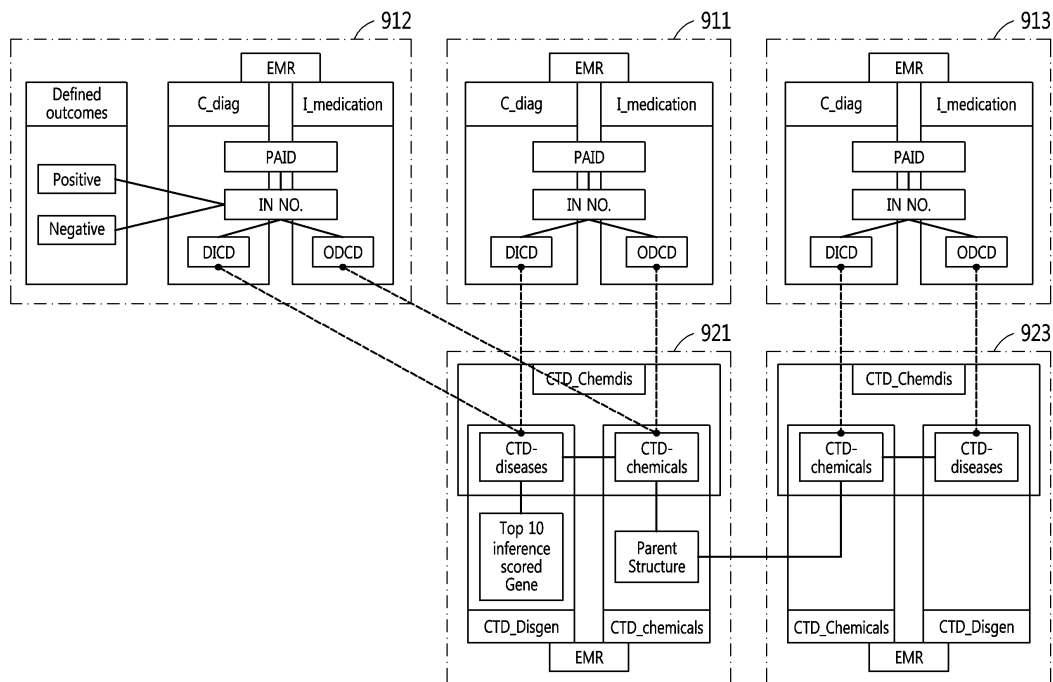
도면7



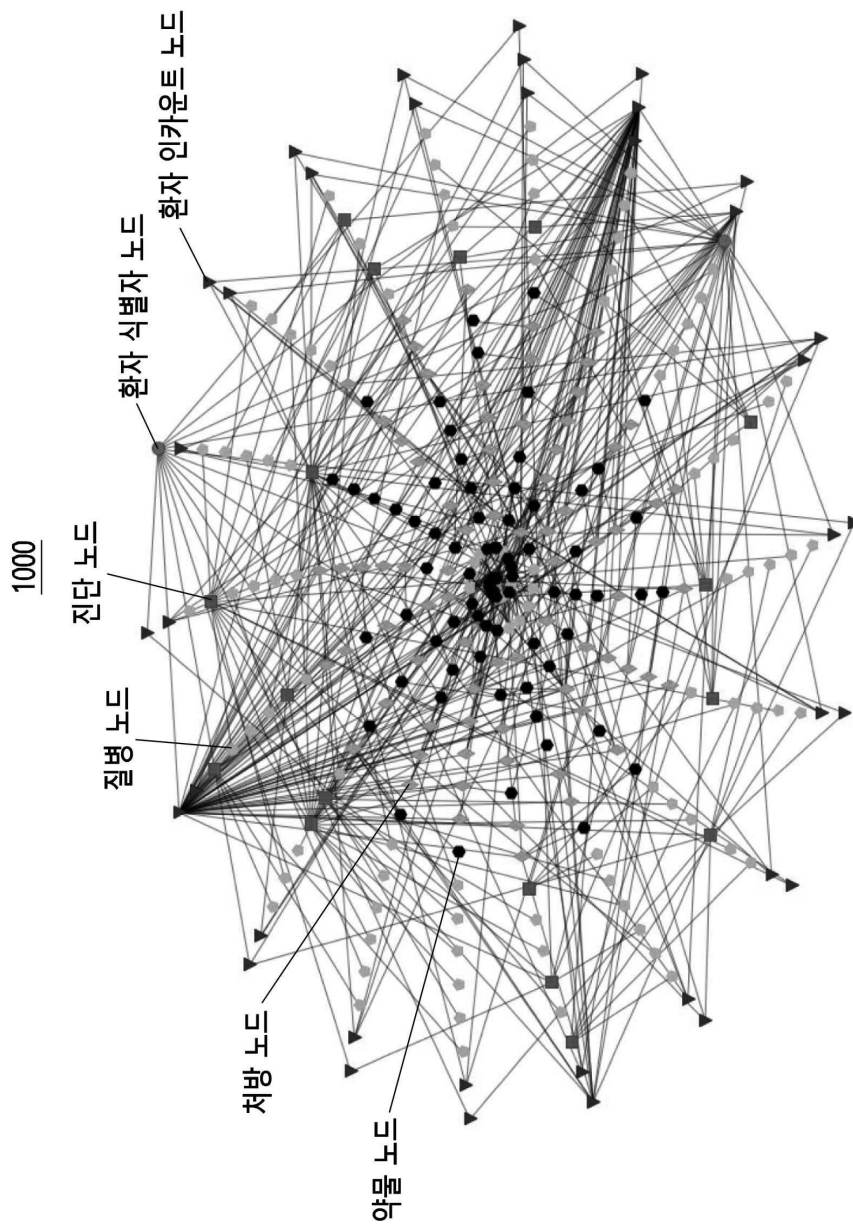
도면8



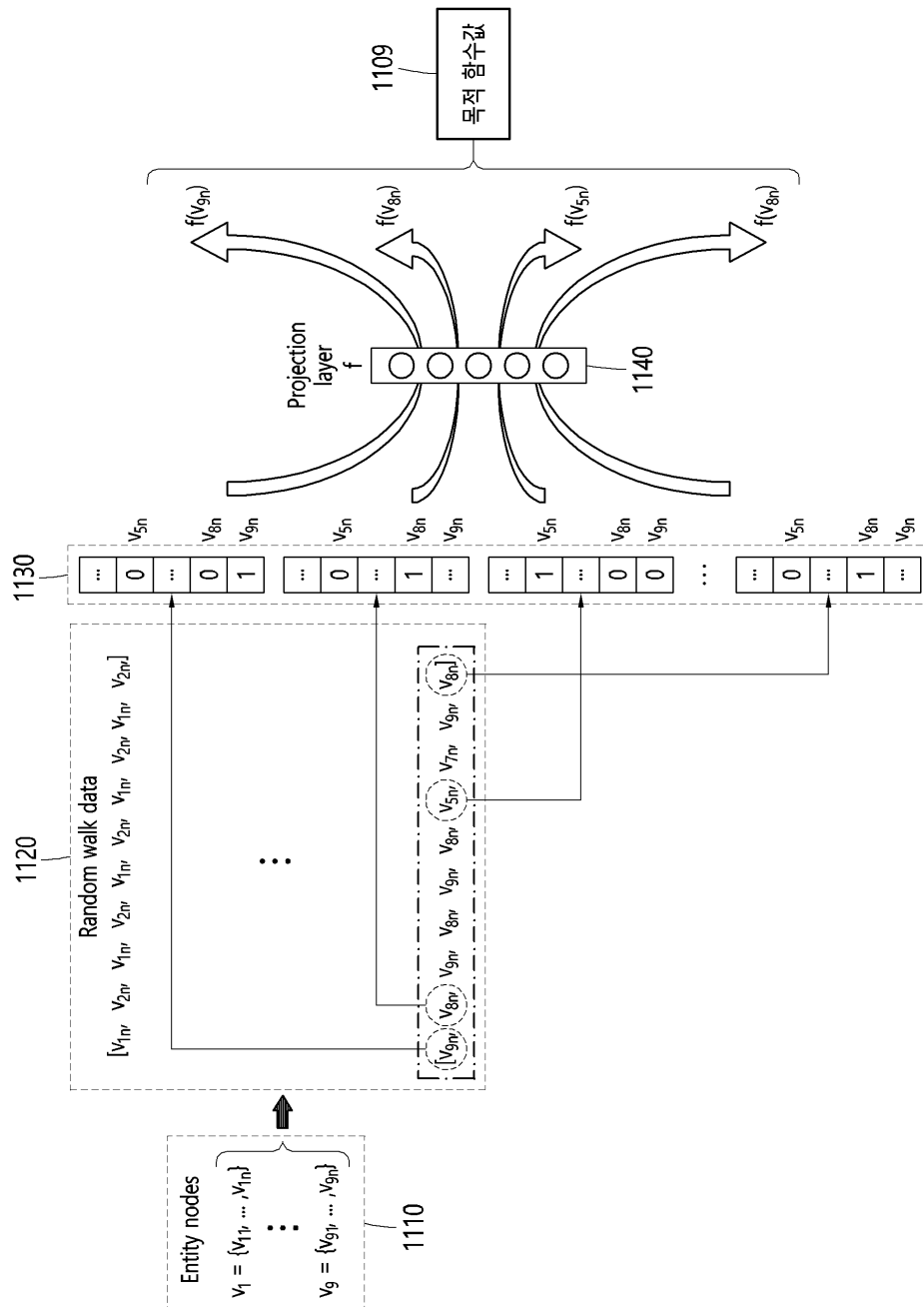
도면9



도면10



도면11



도면12

1200

