(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2022-0003704
(43) 공개일자 2022년01월11일

(51) 국제특허분류(Int. Cl.)
G16H 50/70 (2018.01) *G06F 16/35* (2019.01)
G06F 16/36 (2019.01) *G16H 70/00* (2018.01)
(52) CPC특허분류
G16H 50/70 (2018.01)
G06F 16/35 (2019.01)
(21) 출원번호 10-2020-0081236
(22) 출원일자 2020년07월02일
심사청구일자 2020년07월02일

(71) 출원인
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
주형준
서울특별시 강남구 대치동 503 개포1,2차 우성아파트, 10동 1405호
김종호
서울특별시 성북구 화랑로 76, 102동 3504호 (하월곡동,코업스타클래스)
(뒷면에 계속)
(74) 대리인
특허법인태백

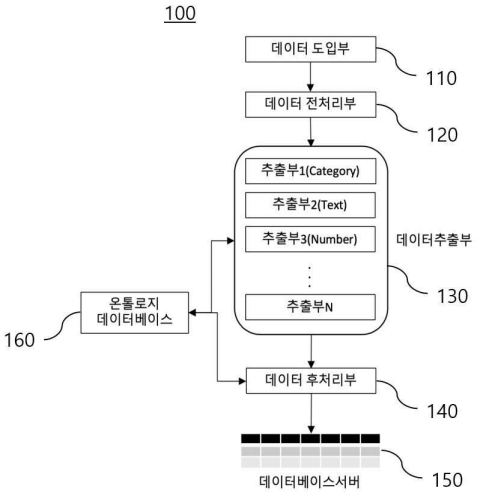
전체 청구항 수 : 총 12 항

(54) 발명의 명칭 **온톨로지 데이터베이스 기반의 룰 기반 알고리즘을 이용한 병원진단검사결과 정제 시스템**

(57) 요 약

본 발명의 일 실시예에 따른 온톨로지 데이터베이스 기반의 룰 기반 알고리즘을 이용한 병원진단검사결과 정제 시스템은, 병원진단검사결과의 정제를 위해 기 구축된 온톨로지 데이터베이스; 입력된 병원진단검사결과를 기 정해진 형태로 변환하여 데이터를 생성하는 데이터 도입부; 상기 데이터 도입부에 의해 생성된 데이터를 전달 받아서 룰 기반으로 데이터 형식의 변환 및 검증, 데이터 삭제 및 데이터 분리 중 적어도 하나 이상을 수행하는 데이터 전처리부; 상기 데이터 전처리부에 의해 전처리된 데이터를 전달 받고 상기 온톨로지 데이터베이스를 기반으로 기 정의된 기준에 따라 룰 기반으로 데이터를 추출하는 데이터 추출부; 및 상기 데이터 추출부에 의해 추출된 데이터를 원본 데이터를 기준으로 분석하여 데이터 전환의 오류 가능성을 평가하는 데이터 후처리부를 포함할 수 있다.

대 표 도 - 도1



(52) CPC특허분류
G06F 16/367 (2019.01)
G16H 70/00 (2021.08)

(72) 발명자
박수완
경기도 안산시 단원구 광덕2로 241, 806동 1403호
(고잔동,그린빌주공아파트)

장준호
서울특별시 관악구 봉천로19길 18(봉천동, 삼성주택), 303호

이 발명을 지원한 국가연구개발사업

과제고유번호	HI19C0832
과제번호	HI19C0832
부처명	보건복지부
과제관리(전문)기관명	한국보건산업진흥원
연구사업명	CDM 기반 정밀의료 데이터 통합 플랫폼 기술개발
연구과제명	클라우드 기반 공통데이터모델 보안관리 및 운영 네트워크 개발
기 여 율	1/2
과제수행기관명	고려대학교 산학협력단
연구기간	2019.07.25 ~ 2021.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호	1415163624
과제번호	20004861
부처명	산업통상자원부
과제관리(전문)기관명	한국산업기술평가관리원
연구사업명	산업기술혁신사업
연구과제명	CDM 기반의 지능형 진료 가이드 알고리즘 개발과 확산을 위한 CDSS 플랫폼 개발
기 여 율	1/2
과제수행기관명	고려대학교의료원
연구기간	2019.04.01 ~ 2021.12.31

명 세 서

청구범위

청구항 1

병원진단검사결과의 정제를 위해 기 구축된 온톨로지 데이터베이스;

입력된 병원진단검사결과를 기 정해진 형태로 변환하여 데이터를 생성하는 데이터 도입부;

상기 데이터 도입부에 의해 생성된 데이터를 전달 받아서 룰 기반으로 데이터 형식의 변환 및 검증, 데이터 삭제 및 데이터 분리 중 적어도 하나 이상을 수행하는 데이터 전처리부;

상기 데이터 전처리부에 의해 전처리된 데이터를 전달 받고 상기 온톨로지 데이터베이스를 기반으로 기 정의된 기준에 따라 룰 기반으로 데이터를 추출하는 데이터 추출부; 및

상기 데이터 추출부에 의해 추출된 데이터를 원본 데이터를 기준으로 분석하여 데이터 전환의 오류 가능성을 평가하는 데이터 후처리부를 포함하는 온톨로지 데이터베이스 기반의 룰 기반 알고리즘을 이용한 병원진단검사결과 정제 시스템.

청구항 2

제 1 항에 있어서,

상기 병원진단검사결과는 의료기관에서 수집 및 저장된 것으로, 인위적으로 작성되거나 의료 장비에서 추출된 것을 특징으로 하는 온톨로지 데이터베이스 기반의 룰 기반 알고리즘을 이용한 병원진단검사결과 정제 시스템.

청구항 3

제 1 항에 있어서,

상기 병원진단검사결과는 단일 또는 복수의 결과를 포함하며, 문자, 숫자 및 기호 중 적어도 하나의 조합으로 이루어지는 것을 특징으로 하는 온톨로지 데이터베이스 기반의 룰 기반 알고리즘을 이용한 병원진단검사결과 정제 시스템.

청구항 4

제 1 항에 있어서,

상기 데이터 도입부는 상기 병원진단검사결과가 저장된 데이터베이스에 연결하여 데이터를 불러오는 것을 특징으로 하는 온톨로지 데이터베이스 기반의 룰 기반 알고리즘을 이용한 병원진단검사결과 정제 시스템.

청구항 5

제 1 항에 있어서,

상기 데이터 전처리부는 입력된 데이터를 분석하여 문자 형식으로 변환하거나, 기 정의된 데이터 형식 검증 룰 기반으로 데이터 형식을 검증하여 불필요한 데이터를 삭제하고, 복수 개의 정보를 포함하는 데이터를 분리하는 것을 특징으로 하는 온톨로지 데이터베이스 기반의 룰 기반 알고리즘을 이용한 병원진단검사결과 정제 시스템.

청구항 6

제 1 항에 있어서,

상기 데이터 추출부는 기 정의된 종류 별로 데이터를 추출하는 것을 특징으로 하는 온톨로지 데이터베이스 기반의 룰 기반 알고리즘을 이용한 병원진단검사결과 정제 시스템.

청구항 7

제 6 항에 있어서,

상기 기 정의된 종류는 오퍼레이터(Operator) 데이터, 단위(Unit) 데이터, 카테고리(Category) 데이터, 비율(Ratio) 데이터, 강도(Grade) 데이터, 타입(Type) 데이터, 텍스트(Text) 데이터 및 수치(Number) 데이터 중 적어도 하나 이상을 포함하는 것을 특징으로 하는 온톨로지 데이터베이스 기반의 룰 기반 알고리즘을 이용한 병원진단검사결과 정제 시스템.

청구항 8

제 6 항에 있어서,

상기 데이터 추출부는 상기 기 정의된 종류의 데이터에 해당하는지 여부를 기 정의된 순서에 따라 순차로 확인하고, 확인된 값을 해당 종류의 데이터로 추출하는 것을 특징으로 하는 온톨로지 데이터베이스 기반의 룰 기반 알고리즘을 이용한 병원진단검사결과 정제 시스템.

청구항 9

제 1 항에 있어서,

상기 데이터 후처리부는 상기 데이터 추출부에 의해 추출되지 않은 후처리 데이터에 대해 기 정의된 추가 추출 룰을 기반으로 데이터 추출을 진행하여 상기 온톨로지 데이터베이스에 대한 업데이트를 진행하는 것을 특징으로 하는 온톨로지 데이터베이스 기반의 룰 기반 알고리즘을 이용한 병원진단검사결과 정제 시스템.

청구항 10

제 1 항에 있어서,

상기 데이터 후처리부는 추출된 데이터에 대해 기 정의된 용어매핑 룰 및 데이터 정리 룰을 기반으로 표준 용어 및 표준 코드로 전환을 수행하는 것을 특징으로 하는 온톨로지 데이터베이스 기반의 룰 기반 알고리즘을 이용한 병원진단검사결과 정제 시스템.

청구항 11

제 1 항에 있어서,

상기 데이터 후처리부에 의해 출력된 데이터를 저장 및 제공하는 데이터베이스 서버를 더 포함하는 것을 특징으로 하는 온톨로지 데이터베이스 기반의 룰 기반 알고리즘을 이용한 병원진단검사결과 정제 시스템.

청구항 12

제 1 항에 있어서,

상기 온톨로지 데이터베이스는 국제 표준인 LOINC(Logical Observation Identifiers Names and Codes) 및 의료기관에서 수집한 데이터를 기반으로 병원진단검사에서 사용되는 기호 및 용어를 정리하여 구축한 데이터베이스

인 것을 특징으로 하는 온톨로지 데이터베이스 기반의 룰 기반 알고리즘을 이용한 병원진단검사결과 정제 시스템.

발명의 설명

기 술 분 야

[0001] 본 출원은 온톨로지 데이터베이스 기반의 룰 기반 알고리즘을 이용한 병원진단검사결과 정제 시스템에 관한 것이다.

배 경 기 술

[0002] 의료 분야의 빅데이터를 구축하기 위해서는 데이터의 정제가 필수적이다. 그러나, 의료기관 내의 대다수의 데이터가 코드화되지 않은 비정형 텍스트 혹은 이미지로 구성되어 있으며, 특히 병원검사결과들은 전문적인 용어 및 도메인 별로 특이한 구성으로 이루어져 있어서 일반적인 데이터 정제 과정을 통해 가치 있는 데이터로 정제하기 어렵다.

[0003] 실제로 의료기관의 모든 임상데이터를 의학 분야별로 맞추어 데이터베이스로 구축하기 위한 기술은 전무한 실정이며, 모든 의학분야를 아우르는 정제 시스템을 개발하는 것은 현실적으로 불가능하다.

[0004] 따라서, 고품질의 빅데이터를 구축하기 위해서는 각 검사 종류별로 적합한 자동화 알고리즘을 개발하는 것이 필요하다.

[0005] 한편, 의료기관에서 시행되는 진단검사결과는 환자의 치료법을 결정하는데 중요할 뿐만 데이터 양적으로도 가장 많은 분량을 차지하고 있다. 이러한 진단검사결과는 숫자 형태의 데이터로 이루어진 경우도 있으나 문자 또는 기호가 혼재되어 있는 경우가 많아서 해당 결과를 문자, 즉 텍스트 형태의 데이터로 저장하고 있다. 이러한 데이터를 적절하게 사용 및/또는 분석하기 위해서는 텍스트 형태의 데이터에서 다른 형태에 해당하는 부분을 추출하여 변환할 필요가 있다.

[0006] 특히, 병원진단검사결과의 경우, 최근 자동화된 시스템을 통해 병원정보 시스템 등의 서버에 자동으로 입력되는 경우도 있으나, 시스템 오류로 인해 수기로 보정하는 경우가 있으며 검사결과에 따라서 관련 담당자가 추가적인 코멘트나 주석을 달기도 한다. 이러한 과정에서 데이터 형식의 일관성이 손상되고, 이러한 이유로 데이터를 비정형 텍스트 형태로 저장하고 있는 실정이다.

발명의 내용

해결하려는 과제

[0007] 따라서, 당해 기술분야에서는 보다 정확하고 효율적으로 병원진단검사결과 데이터를 정제하기 위한 방안이 요구되고 있다.

[0008] 특히, 병원진단검사결과를 자동으로 단시간 내에 타입별 데이터베이스를 구축하고, 필요한 경우 사용자가 원하는 용어 및 코드로 전환하기 위한 방안이 요구되고 있다.

과제의 해결 수단

[0009] 상기 과제를 해결하기 위해서, 본 발명의 일 실시예는 온톨로지 데이터베이스 기반의 룰 기반 알고리즘을 이용한 병원진단검사결과 정제 시스템을 제공한다.

[0010] 상기 온톨로지 데이터베이스 기반의 룰 기반 알고리즘을 이용한 병원진단검사결과 정제 시스템은, 병원진단검사결과의 정제를 위해 기 구축된 온톨로지 데이터베이스; 입력된 병원진단검사결과를 기 정해진 형태로 변환하여 데이터를 생성하는 데이터 도입부; 상기 데이터 도입부에 의해 생성된 데이터를 전달 받아서 룰 기반으로 데이터 형식의 변환 및 검증, 데이터 삭제 및 데이터 분리 중 적어도 하나 이상을 수행하는 데이터 전처리부; 상기 데이터 전처리부에 의해 전처리된 데이터를 전달 받고 상기 온톨로지 데이터베이스를 기반으로 기 정의된 기준에 따라 룰 기반으로 데이터를 추출하는 데이터 추출부; 및 상기 데이터 추출부에 의해 추출된 데이터를 원본 데이터를 기준으로 분석하여 데이터 전환의 오류 가능성을 평가하는 데이터 후처리부를 포함할 수 있다.

[0012] 덧붙여 상기한 과제의 해결수단은, 본 발명의 특징을 모두 열거한 것이 아니다. 본 발명의 다양한 특징과 그에 따른 장점과 효과는 아래의 구체적인 실시형태를 참조하여 보다 상세하게 이해될 수 있을 것이다.

발명의 효과

[0013] 본 발명의 일 실시예에 따르면, 의료기관에서 활용도가 높은 병원진단검사결과를 대상으로 온톨로지 데이터베이스 기반의 룰 기반 알고리즘을 이용하여 정제함으로써 고품질의 데이터베이스를 구축할 수 있다.

[0014] 이를 통해, 데이터베이스의 다기관 상호운용성을 높이고 데이터 분석을 용이하게 할 수 있다.

도면의 간단한 설명

[0015] 도 1은 본 발명의 일 실시예에 따른 온톨로지 데이터베이스 기반의 룰 기반 알고리즘을 이용한 병원진단검사결과 정제 시스템의 구성도이다.

도 2는 도 1에 도시된 데이터 도입부의 일 구현예를 도시하는 도면이다.

도 3은 도 1에 도시된 데이터 전처리부에 의한 처리 흐름을 도시하는 도면이다.

도 4는 도 1에 도시된 데이터 전처리부의 일 구현예를 도시하는 도면이다.

도 5는 도 1에 도시된 데이터 추출부에 의한 처리 흐름을 도시하는 도면이다.

도 6은 도 1에 도시된 데이터 추출부의 일 구현예를 도시하는 도면이다.

도 7은 도 1에 도시된 데이터 후처리부에 의한 처리 흐름을 도시하는 도면이다.

도 8은 도 1에 도시된 데이터 후처리부의 일 구현예를 도시하는 도면이다.

도 9는 본 발명의 일 실시예에 따른 온톨로지 데이터베이스 기반의 룰 기반 알고리즘을 이용한 병원진단검사결과 정제 시스템에 의한 병원진단검사결과 처리 결과의 일 예를 도시하는 도면이다.

발명을 실시하기 위한 구체적인 내용

[0016] 이하, 첨부된 도면을 참조하여 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자가 본 발명을 용이하게 실시할 수 있도록 바람직한 실시예를 상세히 설명한다. 다만, 본 발명의 바람직한 실시예를 상세하게 설명함에 있어, 관련된 공지 기능 또는 구성에 대한 구체적인 설명이 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명을 생략한다. 또한, 유사한 기능 및 작용을 하는 부분에 대해서는 도면 전체에 걸쳐 동일한 부호를 사용한다.

[0017] 덧붙여, 명세서 전체에서, 어떤 부분이 다른 부분과 '연결'되어 있다고 할 때, 이는 '직접적으로 연결'되어 있는 경우뿐만 아니라, 그 중간에 다른 소자를 사이에 두고 '간접적으로 연결'되어 있는 경우도 포함한다. 또한, 어떤 구성요소를 '포함'한다는 것은, 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있다는 것을 의미한다.

[0019] 도 1은 본 발명의 일 실시예에 따른 온톨로지 데이터베이스 기반의 룰 기반 알고리즘을 이용한 병원진단검사결과 정제 시스템의 구성도이다.

[0020] 도 1을 참조하면, 본 발명의 일 실시예에 따른 온톨로지 데이터베이스 기반의 룰 기반 알고리즘을 이용한 병원진단검사결과 정제 시스템(100)은 데이터 도입부(110), 데이터 전처리부(120), 데이터 추출부(130), 데이터 후처리부(140), 데이터베이스 서버(150) 및 온톨로지 데이터베이스(160)를 포함하여 구성될 수 있다.

[0022] 데이터 도입부(110)는 병원정보시스템 등과 같은 의료기관의 서버에 저장된 데이터, 즉, 병원진단검사결과를 입력 받고 이를 후술하는 구성에 의해 처리할 수 있는 기 정해진 형태로 변환하여 데이터를 생성할 수 있다. 여기서, 병원진단검사결과는 의료기관에서 수집 및 저장된 것으로, 인위적으로 작성되거나 의료 장비에서 추출된 것일 수 있으며, 다양한 형태로 저장된 것일 수 있다. 또한, 병원진단검사결과는 단일 결과뿐만 아니라 복수의 결과를 포함할 수도 있다. 또한, 병원진단검사결과는 문자, 숫자 및 기호 중 적어도 하나의 조합으로 이루어질 수

있다.

[0023] 일 실시예에 따르면, 데이터 도입부(110)은 사용자에 의해 직접 입력된 텍스트를 입력 받을 수도 있고, 예를 들어 CSV(comma separated value) 파일, 혹은 EXCEL 파일, 텍스트 파일 등과 같은 임의의 파일 형식으로 데이터를 입력 받을 수도 있다. 그러나, 상술한 파일 형식은 예시에 불과한 것으로, 특정한 파일로 한정되지 않고 어떠한 파일 형식으로도 데이터를 입력 받을 수 있다.

[0024] 다른 실시예에 따르면, 데이터 도입부(110)는 예를 들어 MS-SQL 등과 같은 DBMS(database management system)를 통해 병원진단검사결과가 저장된 데이터베이스에 접속하여 데이터를 입력 받을 수도 있다.

[0026] 도 2는 도 1에 도시된 데이터 도입부의 일 구현예를 도시하는 도면이다.

[0027] 일 실시예에 따르면, 데이터 도입부(110)는 도 2의 (a)에 도시된 알고리즘에 따라 병원진단검사결과가 저장된 데이터베이스에 연결하여 데이터를 불러와서 기 정해진 형태로 변환된 데이터를 생성할 수 있다.

[0028] 도 2의 (b)는 데이터베이스의 연결 정보 및 상태값을 도시하는 것이고, 도 2의 (c)는 데이터 도입부(110)에 의해 데이터베이스로부터 데이터를 불러오는데 소요된 시간을 나타내는 것으로, 일 실시예에 따라 780 만개의 데이터를 도입하는데 약 9.56분이 소요됨을 확인하였다.

[0030] 데이터 전처리부(120)는 데이터 도입부(110)에 의해 생성된 데이터를 전달 받아서 룰 기반으로 데이터 형식의 변환 및 검증, 데이터 삭제 및 데이터 분리 중 적어도 하나 이상을 수행할 수 있다.

[0032] 도 3은 도 1에 도시된 데이터 전처리부에 의한 처리 흐름을 도시하는 도면이다.

[0033] 도 3을 참조하면, 데이터 전처리부(120)는 입력된 원본 데이터를 분석하여 문자 형식으로 변환하거나, 데이터 형식을 검증하여 불필요한 데이터를 삭제하고, 복수 개의 정보를 포함하는 데이터를 분리할 수 있다.

[0034] 구체적으로, 데이터 전처리부(120)는 원본 데이터를 입력 받으면(S310), 해당 데이터의 데이터 타입이 문자 형식인지 확인하여(S320), 문자 형식인 경우 문자 형식으로 변환하여(S330) 임시 데이터(temp_df)를 생성할 수 있다(S340).

[0035] 이후, 데이터 전처리부(120)는 기 정의된 데이터 형식 검증 룰 기반으로 검증을 수행하여(S350), 진단 검사 결과 형식에 합당한지 여부를 판단하고(S360), 합당하지 않으면 일부 혹은 전체 데이터를 삭제하여(S370), 전처리가 완료된 데이터(전처리_df)을 생성할 수 있다(S380).

[0037] 도 4는 도 1에 도시된 데이터 전처리부의 일 구현예를 도시하는 도면으로, 도 4에 도시된 바와 같이, 데이터 형식 검증 및 분리/삭제의 경우 코드, 날짜 및 검사결과 종류에 대한 검증 및 분리를 수행하고, 대문자를 소문자로 변환하는 단계 등을 수행할 수 있다.

[0039] 데이터 추출부(130)는 전처리된 데이터를 전달 받고 온톨로지 데이터베이스(160)를 기반으로 기 정의된 기준에 따라 룰 기반으로 데이터를 추출할 수 있다.

[0040] 일 실시예에 따르면, 데이터 추출부(130)는 기 정의된 종류(예를 들어, N개 종류의 데이터)로 데이터를 추출할 수 있다.

[0042] 도 5는 도 1에 도시된 데이터 추출부에 의한 처리 흐름을 도시하는 도면이다.

[0043] 도 5를 참조하면, 데이터 추출부(130)는 전처리된 데이터(전처리_df)에서 오퍼레이터(Operator) (또는 기호(Sign)) 데이터(예를 들어, 이상, 이하, 초과 및 미만을 표현하는 데이터), 단위(Unit) 데이터, 카테고리(Category) 데이터, 비율(Ratio) 데이터, 강도(Grade) 데이터, 타입(Type) 데이터 및 수치(Number) 데이터 등을 포함하는 기 정의된 종류의 데이터에 해당하는지 여부를 순차로 확인하고(S510, S520, S530, S540, S550,

S560, S570), 확인된 값을 해당 종류의 데이터로 추출함으로써(S515, S525, S535, S545, S555, S565, S575), 추출 데이터(Extract_df)를 생성할 수 있다.

[0044] 한편, 상술한 과정을 거쳐서 추출되지 않은 데이터는 후술하는 데이터 후처리부(140)에 의한 후처리를 위해 후처리 데이터(후처리_df)를 생성할 수 있다.

[0046] 도 6은 도 1에 도시된 데이터 추출부의 일 구현예를 도시하는 도면으로, 도 6을 참조하면, 데이터 추출부(130)는 전달된 데이터에서 필요로 하는 의료 정보를 추출하기 위해 기호, 단위, 1차 텍스트, 1차 카테고리, 2차 텍스트, 숫자 및 2차 카테고리를 추출함으로써 순도 높은 데이터를 추출할 수 있다.

[0047] 데이터 추출부(130)에 의한 처리 흐름 및 추출 데이터 종류가 반드시 이로 제한되는 것은 아니며, 입력되는 데이터의 구조에 따라 변경 가능하다. 즉, 의료기관별로 병원진단검사결과 데이터의 구조가 상이할 수 있으므로, 필요에 따라 데이터의 추출 순서 및 추출 데이터의 종류를 변경함으로써 보다 순도 높은 데이터를 추출하도록 할 수 있다.

[0049] 데이터 후처리부(140)는 데이터 추출부(130)에 의해 추출된 데이터(Extract_df)를 원본 데이터를 기준으로 분석하여 데이터 전환의 오류 가능성을 평가하여 최종 결과를 출력함과 더불어, 데이터 추출부(130)에 의해 추출되지 않은 후처리 데이터(후처리_df)에 대해 기 정의된 추가 추출 룰을 기반으로 데이터 추출을 진행하여 온톨로지 데이터베이스(160)에 대한 업데이트를 진행할 수 있다.

[0050] 또한, 데이터 후처리부(140)는 추출된 데이터에 대해 온톨로지 데이터베이스(160)에 저장된 사용자가 원하는 표준 용어 및 표준 코드로 전환을 수행할 수도 있다.

[0052] 도 7은 도 1에 도시된 데이터 후처리부에 의한 처리 흐름을 도시하는 도면이다.

[0053] 도 7을 참조하면, 데이터 후처리부(140)는 후처리 데이터(후처리_df)에 남은 데이터가 존재하는 경우(S710, S715), 기 정의된 추가 추출 룰(추가 추출 Rule)을 기반으로 추출을 수행하여 추출 데이터(Extract_df)에 추가할 수 있다(S720).

[0054] 또한, 추가 추출을 진행한 후에도 남은 데이터가 존재하는 경우에는(S725), 온톨로지 데이터베이스(160)에 저장된 온톨로지를 분석하여 온톨로지 데이터베이스(160)에 대한 업데이트를 진행할 수 있다(S730, S735).

[0055] 한편, 추출 데이터(Extract_df)는 추가 추출 Rule에 의해 추가적으로 추출된 데이터와 합쳐져서 확장된 추출 데이터(Extract_Ext)가 생성될 수 있다(S720, S740, S745). 확장된 추출 데이터(Extract_Ext)는 사용자의 필요에 따라 용어매핑 여부를 결정하여(S750), 용어매핑이 필요한 경우에는 사전에 정의된 메타데이터 등을 활용한 용어매핑 룰을 기반으로 표준용어 및 표준코드로 전환을 수행할 수 있으며(S755), 용어매핑이 필요 없는 경우에는 표준용어 및 표준코드로 전환하는 단계를 생략할 수 있다. 이후, 기 정의된 데이터 정리 Rule에 의해 데이터 형식을 점검 및 수정하여 최종 결과(Final_df)을 출력할 수 있다(S765).

[0057] 도 8은 도 1에 도시된 데이터 후처리부의 일 구현예를 도시하는 도면으로, 도 8의 (a)에 도시된 바와 같이 원본 데이터를 기준으로 추추된 데이터를 분석하여 데이터 전환의 오류를 검증할 수 있으며, 온톨로지 데이터베이스(160)를 기반으로 컬럼을 정리할 수 있다.

[0058] 도 8의 (b)는 도 8의 (a)의 알고리즘을 수행하는데 소요된 시간을 나타내는 것으로, 일 실시예에 따라 2억여건을 처리 및 검증하는데 약 35분 정도 소요됨을 확인하였다.

[0060] 데이터베이스 서버(150)는 데이터 후처리부(140)에 의해 출력된 데이터를 저장 및 제공할 수 있다.

[0062] 온톨로지 데이터베이스(160)는 병원진단검사에서 사용되는 기호 및 용어 (예를 들어, 단위, 반복적으로 사용하는 구문, 약자, 표준용어, 표준코드 등)를 정리하여 구축한 데이터베이스로서, 국제 표준인 LOINC(Logical

Observation Identifiers Names and Codes) 및 의료기관에서 수집한 데이터를 기반으로 구축된 것일 수 있다.

[0064] 도 9는 본 발명의 일 실시예에 따른 온톨로지 데이터베이스 기반의 룰 기반 알고리즘을 이용한 병원진단검사결과 정제 시스템에 의한 병원진단검사결과 처리 결과의 일 예를 도시하는 도면이다.

[0065] 도 9의 (a)에 도시된 바와 같이 입력된 병원진단검사결과는 도 1 내지 도 8을 참조하여 상술한 바와 같은 과정에 따라 처리되어 도 9의 (b)에 도시된 바와 같이 데이터 타입별로 추출되어 정리된 결과 데이터를 획득할 수 있다.

[0067] 상술한 본 발명의 실시예를 종합병원의 진단검사결과 데이터베이스에 저장된 약 2억 건의 샘플데이터를 적용한 결과, 데이터 타입별로 데이터를 추출하여 보다 정교한 데이터베이스를 구축하였으며, 그 중 일부 데이터는 의료 표준용어 및 코드로 전환하여 구축할 수 있었다.

[0068] 또한, 7,840,263건의 데이터의 질검정 수행결과, 미처리 데이터 종류는 43건, 미처리 데이터 건수는 143건(0.0018%)에 불과하여, 고품질의 데이터베이스를 구축 가능함을 확인하였다.

[0070] 본 발명은 전술한 실시예 및 첨부된 도면에 의해 한정되는 것이 아니다. 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에게 있어, 본 발명의 기술적 사상을 벗어나지 않는 범위 내에서 본 발명에 따른 구성요소를 치환, 변형 및 변경할 수 있다는 것이 명백할 것이다.

부호의 설명

[0071] 100: 병원진단검사결과 정제 시스템

110: 데이터 도입부

120: 데이터 전처리부

130: 데이터 추출부

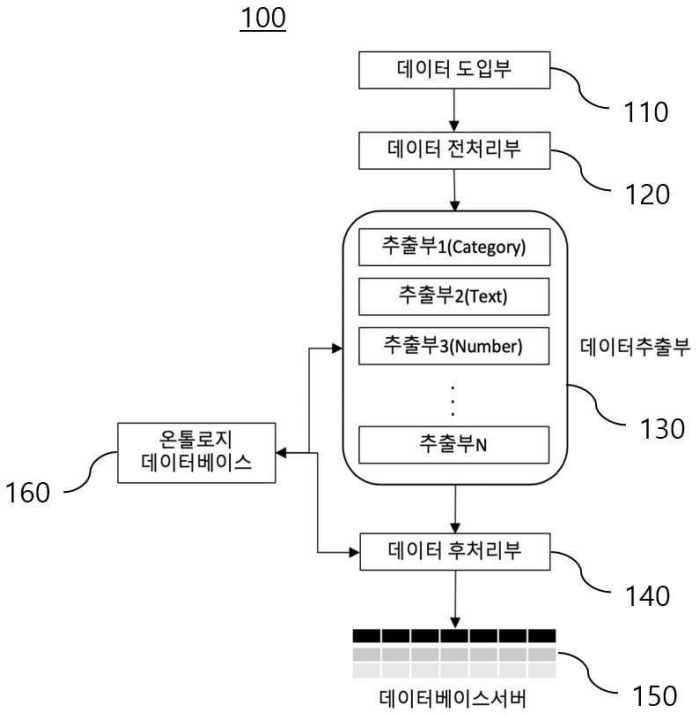
140: 데이터 후처리부

150: 데이터베이스 서버

160: 온톨로지 데이터베이스

도면

도면1



도면2

(a)

```
1  # DB connection
2  dbms <- '[redacted]'
3  user <- '[redacted]'
4  pw <- '[redacted]'
5  host <- '[redacted]'
6
7  details <- DatabaseConnector::createConnectionDetails(
8    dbms = dbms,
9    user = user,
10   password = pw,
11   server = host
12 )
13
14 con <- DatabaseConnector::connect(details)
15
16 start_time <- Sys.time()
17 query <- "SELECT *  FROM [TEMP_DB].[dbo][redacted]
18 data <- DatabaseConnector::dbGetQuery(conn = con, statement = query)
19 end_time <- Sys.time()
20 end_time - start_time
```

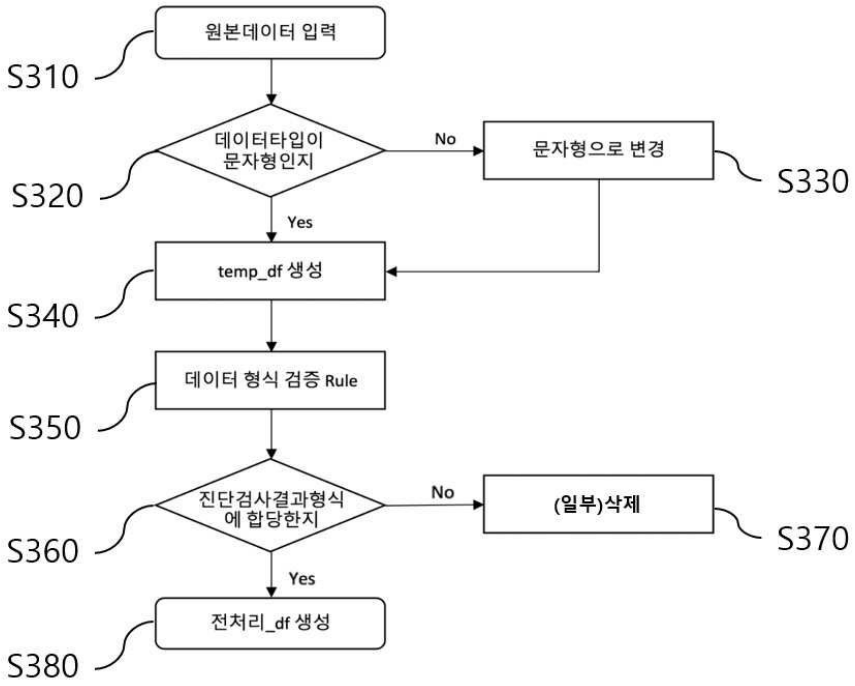
(b)

Data	
con	Formal class DatabaseConnectorJdbcConnection
dat[redacted]	7840263 obs. of [redacted] variables

(c)

```
> end_time <- Sys.time()
> end_time - start_time
Time difference of 9.56581 mins
```

도면3



도면4

```
##  << Pre-processing >>
## 1.문자형식으로 변환
data$result <- as.character(data$result)

## 2. 데이터형식 검증 및 분리/삭제
## (1) 코드로 생각되는 검사결과
data$result <- ifelse(!is.na(str_extract(data$result, "^[:digit:]+[:alpha:]+[:digit:]+")),NA, data$result)
## [코드로 생각되는 검사결과] 이하 코드 생략

## (2) 날짜로 생각되는 검사결과
data$result <- ifelse(!is.na(str_extract(data$result, "^20[:digit:][:digit:]//-")),NA, data$result)
## [날짜로 생각되는 검사결과] 이하 코드 생략

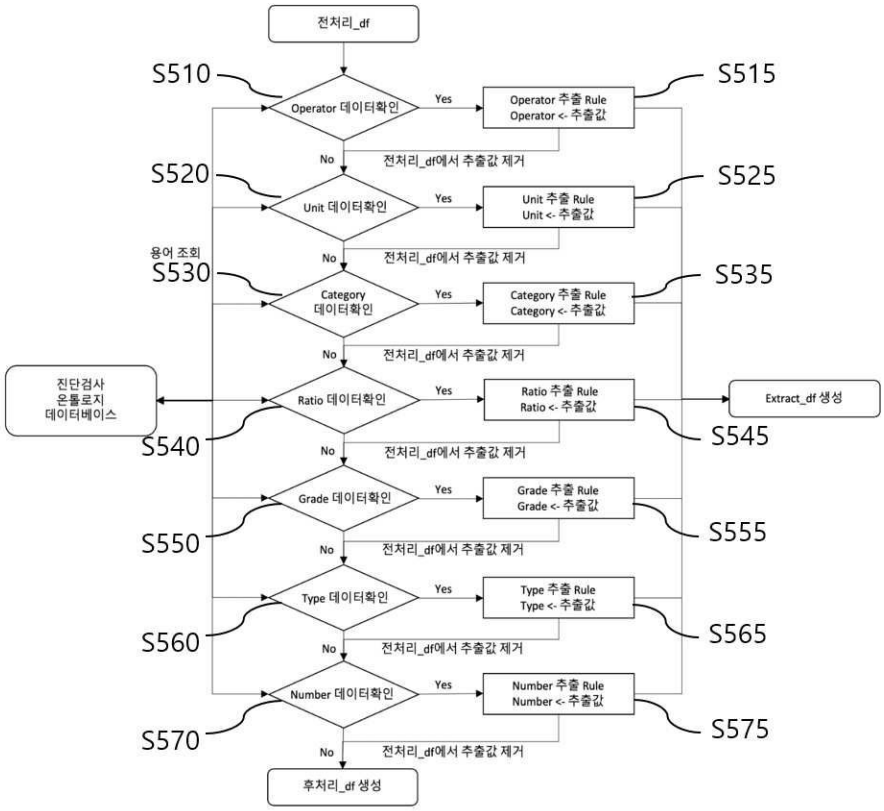
## (3) 결과가 없는 검사결과
data$result <- ifelse(data$result ==".", NA, data$result)
## [결과가 없는 검사결과] 이하 코드 생략

## (4)진단검사결과가 아닌 검사결과 (이름 등)
data$result <- ifelse(!is.na(str_extract(data$result, "^[가-힣][가-힣][가-힣]+")), NA, data$result)
## [진단검사결과가 아닌 검사결과 (이름 등)] 이하 코드 생략

## (5) 그밖에 특수한 검사결과 (글자수 50개 이상 등)
data$nchar <- nchar(data$result)
data$result_1 <- ifelse (nchar(data$result) >= 50, data$result, NA)
data$result <- ifelse(nchar(data$result) >= 50, NA, data$result)
data$nchar <- NULL
## [그밖에 특수한 검사결과 (글자수 50개 이상 등)] 이하 코드 생략

## (6) 소문자로 변경
data$result <- tolower(data$result)
```

도면5



도면6

```r
## << Data Extraction & Transformation >>
## 1. sign 추출 (column name: sign.final)
data$sign1 <- str_extract(data$result, "[<>=≤≥]+")
## 이하 코드 생략
data$result <- str_replace_all(data$result, "[<>=≤≥]+", "")
## 이하 코드 생략
data$sign.final <- paste(data$sign1, data$sign2, data$sign3, data$sign4, data$sign5, data$sign6)
## 이하 코드 생략
data$sign.final <- str_replace_all(data$sign.final, "NA", "")
## 이하 코드 생략
data$sign.final <- ifelse(data$sign.final=="", NA, data$sign.final)
data$sign1 <- NULL; data$sign2 <- NULL; data$sign3 <- NULL; data$sign4 <- NULL; data$sign5 <- NULL; data$sign6 <- NULL

## 2. 단위 추출  (column name: unit)
data$unit1 <- ifelse(str_extract(data$result, "iu\\/ml") == "iu/ml", "IU/ml", NA)
## 이하 코드 생략
data$unit <- paste(data$unit1, data$unit2, data$unit3, data$unit4, data$unit5, data$unit6, data$unit7, data$unit8, data$unit9)
## 이하 코드 생략
data$unit <- str_replace_all(data$unit, "\\s{2,}", "")
## 이하 코드 생략
data$unit <- ifelse(data$unit=="", NA, data$unit)
data$unit1 <- NULL; data$unit2 <- NULL; data$unit3 <- NULL; data$unit4 <- NULL; data$unit5 <- NULL; data$unit6 <- NULL; data$unit7 <- NULL;


## 3. comment text 추출 (검체종류 포함) (column name: comment1)
data$com1.1 <- ifelse(!is.na(str_extract(data$result, "[:digit:]+분")), str_extract(data$result, "[:digit:]+분"), NA)
## 이하 코드 생략

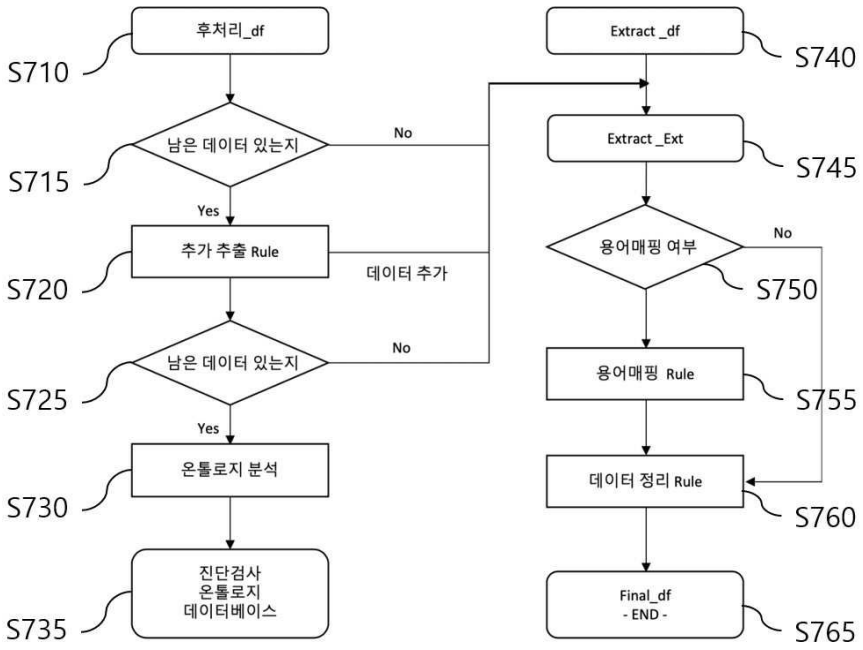
## 4. Category 추출 (general, ratio, grade, type)
## cate2 (general)
## positive
data$cate2.1.1 <- ifelse(!is.na(str_extract(data$result, "w\\s\\+")), "weak_positive",0)
## 이하 코드 생략

## 5. Text 추출  (2자, column name: text2)
data$text2.1 <- ifelse(!is.na(str_extract(data$result, "dar[:print:]+")),str_extract(data$result, "dar[:print:]+"), NA)        # dark...
## 이하 코드 생략

## 6. Number  추출 (column name: num1)
data$num1 <- ifelse(!is.na(str_extract(data$result, "[:digit:]+\\.[:digit:]+e\\+[:digit:]+")),str_extract(data$result, "[:digit:]+\\.[:digit
## 이하 코드 생략

## 7. Category (grade) 추출
data$cate3.2.101 <- ifelse(data$result == "+", "+", NA)
## 이하 코드 생략
```

도면7



도면8

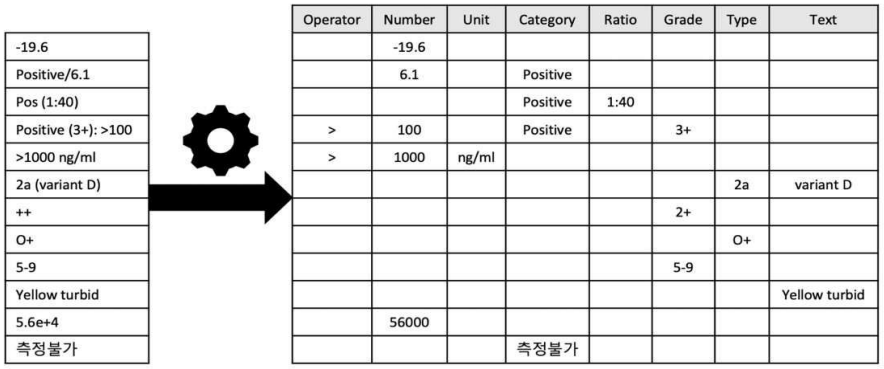
(a)

```
666 ## << Post-processing >>
667 ## 마지막 남은 데이터 정리 (for data mining validation)
668 ## 불필요한 텍스트 삭제
669 data$result <- str_replace_all(data$result, "y", "")
670 data$result <- ifelse(data$result == "'", NA, data$result)
671 ## [불필요한 텍스트 삭제] 이하 코드 생략
672
673 ## Column 정리 (Ontology DB transfer column 등)
674 data$Index <- data$X
675 data$Data <- data$result
676 data$Operator <- data$sign.final; data$sign.final <- NULL
677 data$Number <- data$number.final; data$number.final <- NULL
678 data$Unit <- data$unit; data$unit <- NULL
679 data$Category <- data$cate2
680 data$Ratio <- data$cate3.1
681 data$Grade <- data$cate3.2
682 data$Type <- data$cate3.3
683 data$cate2 <- NULL;data$cate3.1 <- NULL;data$cate3.2.final <- NULL;data$cate3.3 <- NULL
684 data$Text <- ifelse(!is.na(data$text1), data$text1, data$text2)
685 data$text1 <- NULL; data$text2 <- NULL
686 data$Comment <- data$comment1; data$comment1 <- NULL
687 data$result <- NULL
688 data$X <- NULL
689 ## [Column 정리 (Ontology DB transfer column 등)] 이하 코드 생략
690
691 end_time <- Sys.time()
692 end_time - start_time
```

(b)

```
> end_time <- Sys.time()
> end_time - start_time
Time difference of 35.24835 mins
```

도면9



-19.6
Positive/6.1
Pos (1:40)
Positive (3+): >100
>1000 ng/ml
2a (variant D)
++
O+
5-9
Yellow turbid
5.6e+4
측정불가

(a)

Operator	Number	Unit	Category	Ratio	Grade	Type	Text
	-19.6						
	6.1		Positive				
			Positive	1:40			
>	100		Positive		3+		
>	1000	ng/ml					
						2a	variant D
					2+		
						O+	
					5-9		
							Yellow turbid
	56000						
			측정불가				

(b)