(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2022-0000661
(43) 공개일자 2022년01월04일

(51) 국제특허분류(Int. Cl.)
G16B 50/10 (2019.01) *G06F 16/33* (2019.01)
G06F 16/332 (2019.01) *G06F 16/335* (2019.01)
G16H 70/00 (2018.01)
(52) CPC특허분류
G16B 50/10 (2019.02)
G06F 16/3329 (2019.01)
(21) 출원번호 10-2020-0078557
(22) 출원일자 2020년06월26일
심사청구일자 2020년06월26일

(71) 출원인
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
손장욱
서울특별시 강남구 영동대로72길 10(대치동, 대치동3차풍림아이원아파트), 301동 1403호
한성원
서울특별시 강남구 압구정로 151(압구정동, 현대아파트), 102동 1105호
(뒷면에 계속)
(74) 대리인
특허법인태백

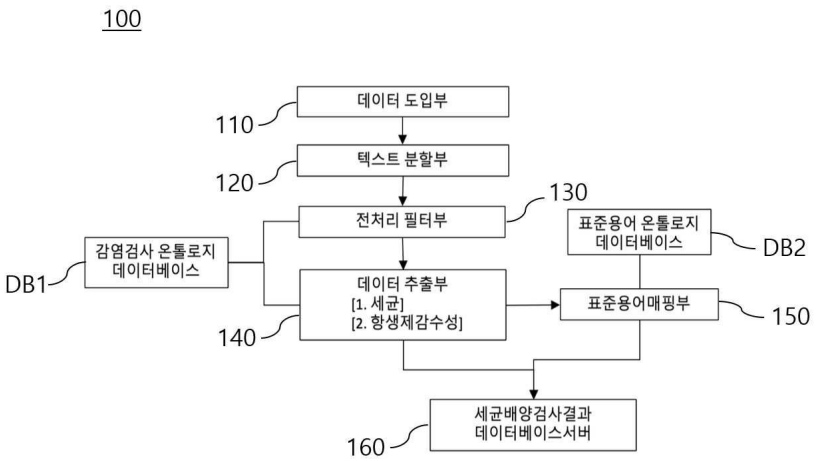
전체 청구항 수 : 총 13 항

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

(57) 요 약

본 발명의 일 실시예에 따른 온톨로지 데이터베이스 기반의 룰 기반 알고리즘을 이용한 세균배양검사결과보고서 정제 시스템은, 세균배양검사결과보고서의 정제를 위해 기 구축된 감염검사 온톨로지 데이터베이스; 복수의 표준용어를 바탕으로 기 구축된 표준용어 온톨로지 데이터베이스; 입력된 데이터를 상기 감염검사 온톨로지 데이터베이스에 저장된 데이터와 비교 분석하여 기 정해진 룰 기반으로 필요한 부분만을 필터링하는 전처리 필터부; 상기 전처리 필터부에 의해 출력된 데이터에서 상기 감염검사 온톨로지 데이터베이스와 기 정해진 룰 기반으로 세균에 해당하는 데이터 및 항생제 감수성 결과 데이터를 추출하는 데이터 추출부; 및 상기 데이터 추출부에 의해 추출된 데이터에서 상기 표준용어 온톨로지 데이터베이스 및 기 정해진 룰 기반으로 최종 매핑된 표준용어 및 코드를 출력하는 표준용어 매핑부를 포함할 수 있다.

대 표 도 - 도1



(52) CPC특허분류
G06F 16/3335 (2019.01)
G06F 16/334 (2019.01)
G06F 16/335 (2019.01)
G16H 70/00 (2021.08)

(72) 발명자
주형준
서울특별시 강남구 대치동 503 개포1,2차 우성아파트, 10동 1405호
김태형
경기도 성남시 분당구 분당로263번길 24 장안타운 건영아파트, 119동 501호
김종호
서울특별시 성북구 하월곡동 46-73 코업스타클래스, 102동 3504호
강민지
서울특별시 중랑구 동일로 752 중화한신아파트, 103동 2302호
장준호
서울특별시 관악구 봉천로19길 18(봉천동, 삼성주택), 303호
김종훈
서울특별시 성북구 안암로 145 고려대학교안암캠퍼스

이 발명을 지원한 국가연구개발사업

과제고유번호	HI19C0201
과제번호	HI19C0201
부처명	보건복지부
과제관리(전문)기관명	한국보건산업진흥원
연구사업명	CDM 기반 정밀의료 데이터 통합 플랫폼 기술개발
연구과제명	감염성질환 데이터의 CDM 확장모델 개발 및 적용
기 여 율	1/1
과제수행기관명	고려대학교 안암병원
연구기간	2019.04.17 ~ 2021.12.31

명 세 서

청구범위

청구항 1

세균배양검사결과보고서의 정제를 위해 기 구축된 감염검사 온톨로지 데이터베이스;

복수의 표준용어를 바탕으로 기 구축된 표준용어 온톨로지 데이터베이스;

입력된 데이터를 상기 감염검사 온톨로지 데이터베이스에 저장된 데이터와 비교 분석하여 기 정해진 룰 기반으로 필요한 부분만을 필터링하는 전처리 필터부;

상기 전처리 필터부에 의해 출력된 데이터에서 상기 감염검사 온톨로지 데이터베이스와 기 정해진 룰 기반으로 세균에 해당하는 데이터 및 항생제 감수성 결과 데이터를 추출하는 데이터 추출부; 및

상기 데이터 추출부에 의해 추출된 데이터에서 상기 표준용어 온톨로지 데이터베이스 및 기 정해진 룰 기반으로 최종 매핑된 표준용어 및 코드를 출력하는 표준용어 매핑부를 포함하는 온톨로지 데이터베이스 기반의 룰 기반 알고리즘을 이용한 세균배양검사결과보고서 정제 시스템.

청구항 2

제 1 항에 있어서,

상기 세균배양검사결과보고서는 의료기관에서 수집 및 저장된 것으로, 인위적으로 작성되거나 의료 장비에서 추출된 것이며, 텍스트 및 특수문자 중 적어도 하나에 의해 기술된 것을 특징으로 하는 온톨로지 데이터베이스 기반의 룰 기반 알고리즘을 이용한 세균배양검사결과보고서 정제 시스템.

청구항 3

제 1 항에 있어서,

상기 세균배양검사결과보고서는 단일 또는 복수의 결과보고서를 포함하는 것을 특징으로 하는 온톨로지 데이터베이스 기반의 룰 기반 알고리즘을 이용한 세균배양검사결과보고서 정제 시스템.

청구항 4

제 1 항에 있어서,

입력된 세균배양검사결과보고서를 기 정해진 형태로 변환하여 데이터를 생성하는 데이터 도입부; 및

상기 데이터 도입부에 의해 생성된 데이터를 전달받아 단일 분석단위로 분할하는 텍스트 분할부 중 적어도 하나를 더 포함하며,

상기 데이터 도입부 및 상기 텍스트 분할부 중 적어도 하나의 의해 처리된 데이터를 상기 전처리 필터부에 전달하는 것을 특징으로 하는 온톨로지 데이터베이스 기반의 룰 기반 알고리즘을 이용한 세균배양검사결과보고서 정제 시스템.

청구항 5

제 1 항에 있어서,

상기 전처리 필터부는, 입력된 데이터에서 세균명이 있는 검사 결과만을 추출하고, 항생제 감수성 결과 중에서 항생제 감수성 여부 및 기 정의된 항생제 감수성에 대한 특징에 해당하는 결과를 필터링하는 것을 특징으로 하

는 온톨로지 데이터베이스 기반의 룰 기반 알고리즘을 이용한 세균배양검사결과보고서 정제 시스템.

청구항 6

제 1 항에 있어서,

상기 전처리 필터부는 상기 세균배양검사결과보고서에 포함된 복수의 결과보고서를 단일 검사로 분리하는 것을 특징으로 하는 온톨로지 데이터베이스 기반의 룰 기반 알고리즘을 이용한 세균배양검사결과보고서 정제 시스템.

청구항 7

제 1 항에 있어서,

상기 데이터 추출부는, 입력된 데이터에서 세균명 및 배양정보를 확인하여 모두 존재하는 경우 세균에 해당하는 데이터로 추출하며, 상기 세균명이 존재하면 예외 처리와 오타 검사를 수행하는 것을 특징으로 하는 온톨로지 데이터베이스 기반의 룰 기반 알고리즘을 이용한 세균배양검사결과보고서 정제 시스템.

청구항 8

제 1 항에 있어서,

상기 데이터 추출부는, 입력된 데이터에서 항생제 감수성에 대한 정보를 순차적으로 확인하여 항생제명과 상기 항생제명에 대한 결과값을 추출하며, 예외 처리와 오타 검사를 수행하는 것을 특징으로 하는 온톨로지 데이터베이스 기반의 룰 기반 알고리즘을 이용한 세균배양검사결과보고서 정제 시스템.

청구항 9

제 1 항에 있어서,

상기 데이터 추출부는, 입력된 세균배양검사결과보고서에 대한 데이터 추출 시에 상기 감염검사 온톨로지 데이터베이스의 응답에 기반하여 세균 및 항생제 감수성에 대한 결과를 생성 및 제공하는 것을 특징으로 하는 온톨로지 데이터베이스 기반의 룰 기반 알고리즘을 이용한 세균배양검사결과보고서 정제 시스템.

청구항 10

제 1 항에 있어서,

상기 표준용어 매핑부는, 상기 추출된 데이터에서 기 정해진 라벨 룰에 따라 라벨을 생성하며, 라벨이 없으면 상기 세균명을 라벨로 생성하고, 라벨이 복수개의 경우 기 정해진 계층을 고려하여 순서를 재정렬한 후 표준용어 및 코드 매핑을 수행하는 것을 특징으로 하는 온톨로지 데이터베이스 기반의 룰 기반 알고리즘을 이용한 세균배양검사결과보고서 정제 시스템.

청구항 11

제 1 항에 있어서,

상기 표준용어 매핑부는 입력된 데이터의 정제 실행 시에 상기 표준용어 온톨로지 데이터베이스의 응답에 기반하여 매핑된 표준용어 및 코드를 생성 및 제공하는 것을 특징으로 하는 온톨로지 데이터베이스 기반의 룰 기반 알고리즘을 이용한 세균배양검사결과보고서 정제 시스템.

청구항 12

제 1 항에 있어서,

상기 감염검사 온톨로지 데이터베이스는 세균배양검사에서 사용하는 세균, 항생제 및 용어 또는 약자가 정리된 것을 특징으로 하는 온톨로지 데이터베이스 기반의 룰 기반 알고리즘을 이용한 세균배양검사결과보고서 정제 시스템.

청구항 13

제 1 항에 있어서,

상기 표준용어 온톨로지 데이터베이스는 표준화된 코드를 매핑하기 위해 SNOMED CT(SNOMED Clinical Terms), OMOP(Observational Medical Outcomes Partnership) 및 LOINC(Logical Observation Identifiers Names and Codes)를 포함하는 표준용어를 바탕으로 구축된 것을 특징으로 하는 온톨로지 데이터베이스 기반의 룰 기반 알고리즘을 이용한 세균배양검사결과보고서 정제 시스템.

발명의 설명

기 술 분 야

[0001] 본 출원은 온톨로지 데이터베이스 기반의 룰 기반 알고리즘을 이용한 세균배양검사결과보고서 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] 이에, 데이터의 분석을 위해서 사람이 직접 보고서를 분석 및 정리하여 데이터베이스로 구축하고 있는 실정이다. 그러나, 이 과정에서 데이터베이스의 일관성이 손상되고, 기입 과정에서 오타가 발생하는 등의 문제가 발생한다.

[0004] 한편, 감염 질환과 관련된 데이터 중 가장 대표적인 데이터로 세균배양검사 데이터가 있다. 이러한 감염균에 대한 데이터는 개별 환자의 치료법을 결정하는데 중요할 뿐만 아니라, 보건학적, 병원 경영적 측면에서도 중요하다.

[0005] 그러나, 현재 대다수의 병원에서는 이러한 세균배양검사 데이터를 이미지로 변환하여 보관하거나, 비정형 텍스트 형식으로 보관하고 있는 실정이다.

발명의 내용

해결하려는 과제

[0006] 따라서, 당해 기술분야에서는 보다 정확하고 효율적으로 세균배양검사 데이터를 정제하기 위한 방안이 요구되고 있다.

과제의 해결 수단

[0007] 상기 과제를 해결하기 위해서, 본 발명의 일 실시예는 온톨로지 데이터베이스 기반의 룰 기반 알고리즘을 이용한 세균배양검사결과보고서 정제 시스템을 제공한다.

[0008] 상기 온톨로지 데이터베이스 기반의 룰 기반 알고리즘을 이용한 세균배양검사결과보고서 정제 시스템은, 세균배양검사결과보고서의 정제를 위해 기 구축된 감염검사 온톨로지 데이터베이스; 복수의 표준용어를 바탕으로 기 구축된 표준용어 온톨로지 데이터베이스; 입력된 데이터를 상기 감염검사 온톨로지 데이터베이스에 저장된 데이

터와 비교 분석하여 기 정해진 룰 기반으로 필요한 부분만을 필터링하는 전처리 필터부; 상기 전처리 필터부에 의해 출력된 데이터에서 상기 감염검사 온톨로지 데이터베이스와 기 정해진 룰 기반으로 세균에 해당하는 데이터 및 항생제 감수성 결과 데이터를 추출하는 데이터 추출부; 및 상기 데이터 추출부에 의해 추출된 데이터에서 상기 표준용어 온톨로지 데이터베이스 및 기 정해진 룰 기반으로 최종 매핑된 표준용어 및 코드를 출력하는 표준용어 매핑부를 포함할 수 있다.

[0010] 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

[0011] 본 발명의 일 실시예에 따르면, 의료기관에서 활용도가 높은 병리검사결과보고서를 대상으로 온톨로지 데이터베이스 기반의 룰 기반 알고리즘을 이용하여 정제함으로써 고품질의 데이터베이스를 구축할 수 있다.

[0012] 이를 통해, 데이터베이스의 다기관 상호운용성을 높이고 데이터 분석을 용이하게 할 수 있다.

도면의 간단한 설명

[0013] 도 1은 본 발명의 일 실시예에 따른 온톨로지 데이터베이스 기반의 룰 기반 알고리즘을 이용한 세균배양검사결과보고서 정제 시스템의 구성도이다.

도 2는 도 1에 도시된 전처리 필터부에 의한 처리 흐름을 도시하는 도면이다.

도 3은 도 2에 도시된 바에 따라 데이터가 전처리되는 일 예를 도시하는 도면이다.

도 4는 도 1에 도시된 데이터 추출부에 의한 처리 흐름을 도시하는 도면이다.

도 5는 도 4에 도시된 바에 따라 추출된 데이터의 일 예를 도시하는 도면이다.

도 6은 도 1에 도시된 데이터 추출부에 의한 처리 흐름을 도시하는 도면이다.

도 7은 도 6에 도시된 바에 따라 추출된 데이터의 일 예를 도시하는 도면이다.

도 8은 도 1에 도시된 표준용어 온톨로지 데이터베이스의 일 예를 도시하는 도면이다.

도 9는 도 1에 도시된 표준용어 매핑부에 의한 처리 흐름을 도시하는 도면이다.

도 10은 도 9에 도시된 바에 따라 매핑된 데이터의 일 예를 도시하는 도면이다.

도 11 및 도 12는 본 발명의 일 실시예에 따라 온톨로지 데이터베이스를 기반으로 룰 기반 알고리즘을 이용한 경우와 그렇지 않은 경우의 성능을 비교한 결과를 도시하는 도면이다.

발명을 실시하기 위한 구체적인 내용

[0014] 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.

[0015] 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.

[0017] 도 1은 본 발명의 일 실시예에 따른 온톨로지 데이터베이스 기반의 룰 기반 알고리즘을 이용한 세균배양검사결과보고서 정제 시스템의 구성도이다.

[0018] 도 1을 참조하면, 본 발명의 일 실시예에 따른 온톨로지 데이터베이스 기반의 룰 기반 알고리즘을 이용한 세균

배양검사결과보고서 정제 시스템(100)은 데이터 도입부(110), 텍스트 분할부(120), 전처리 필터부(130), 데이터 추출부(140), 표준용어 매핑부(150), 감염검사 온톨로지 데이터베이스(DB1) 및 표준용어 온톨로지 데이터베이스(DB2)를 포함하여 구성될 수 있다.

[0020] 데이터 도입부(110)는 병원정보시스템 등과 같은 의료기관의 서버에 저장된 데이터, 즉, 세균배양검사결과보고서를 입력 받고 이를 후술하는 구성에 의해 처리할 수 있는 기 정해진 형태로 변환하여 데이터를 생성할 수 있다. 여기서, 세균배양검사결과보고서는 의료기관에서 수집 및 저장된 것으로, 인위적으로 작성되거나 의료 장비에서 추출된 것일 수 있으며, 텍스트 및 특수문자 중 적어도 하나에 의해 기술된 것일 수 있다. 또한, 세균배양검사결과보고서는 단일 결과보고서뿐만 아니라 복수의 결과보고서를 포함할 수도 있다.

[0021] 일 실시예에 따르면, 데이터 도입부(110)은 사용자에 의해 직접 입력된 텍스트를 입력 받을 수도 있고, 예를 들어 CSV(comma separated value) 파일, 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.

[0022] 다른 실시예에 따르면, 데이터 도입부(110)는 예를 들어 MS-SQL 등과 같은 DBMS(database management system)를 통해 세균배양검사결과보고서가 저장된 데이터베이스에 접속하여 데이터를 입력 받을 수도 있다.

[0024] 텍스트 분할부(120)는 데이터 도입부(110)에 의해 생성된 하나 이상의 세균배양검사결과 데이터를 전달 받아서 단일 분석단위로 분할할 수 있다.

[0026] 본 발명의 일 실시예에 따른 세균배양검사결과보고서 정제 시스템(100)은 필요에 따라 상술한 데이터 도입부(110) 및 텍스트 분할부(120) 중 적어도 하나 이상을 포함할 수 있으며, 데이터 도입부(110) 및 텍스트 분할부(120) 중 적어도 하나의 의해 처리된 데이터가 후술하는 전처리 필터부(130)에 전달될 수 있다.

[0028] 전처리 필터부(130)는 텍스트 분할부(120)에 의해 단일 분석단위로 분할된 데이터를 감염검사 온톨로지 데이터베이스(DB1)에 저장된 데이터와 비교 분석하여 도 2에 도시된 알고리즘에 따라 필요한 부분만을 필터링할 수 있다.

[0029] 도 2는 도 1에 도시된 전처리 필터부에 의한 처리 흐름을 도시하는 도면이고, 도 3은 도 2에 도시된 바에 따라 데이터가 전처리되는 일 예를 도시하는 도면이다.

[0030] 도 2를 참조하면, 전처리 필터부(130)는 입력된 미가공 데이터(input raw data)(S21)에서 소문자로 통일시킨 후, 세균명(Organism명)이 있는지 확인하여(S22) Organism 명이 없는 검사 결과는 삭제하고(S23) Organism 명이 있는 검사 결과만을 추출하여 organ_df로 생성할 수 있다(S24).

[0031] 이후, 전처리 필터부(130)는 organ_df에서 필요한 검사 결과만을 필터링하기 위해 예를 들어 항생제 감수성 결과 중에서 항생제 감수성 여부(예를 들어, 양성, 음성, 판정불가, sensitive, resistant, intermediate, equivocal 혹은 이에 해당하는 약자인 s, i, r이 있는 결과)를 필요한 검사 결과로 필터링하여 organ_result로 생성할 수 있다(S25~S27). 한편, 기 정의된 항생제 감수성에 대한 특징에 해당하는 결과(예를 들어, esbl의 결과, 즉 esbl 양성여부)는 상이한 형식으로 표기되므로, 별도의 알고리즘에 따라 필터링할 수 있다.

[0032] 또한, 전처리 필터부(130)는 입력된 세균배양검사결과보고서에 복수의 결과보고서가 포함된 경우 이를 단일 검사결과보고서로 분리할 수도 있다.

[0034] 데이터 추출부(140)는 전처리 필터부(130)에 의해 출력된 데이터(즉, organ_result)에서 감염검사 온톨로지 데이터베이스(DB1)와 도 4 및 도 6에 도시된 알고리즘에 따라 세균에 해당하는 데이터 및 항생제 감수성 결과 데이터를 각각 추출할 수 있다.

[0035] 도 4는 도 1에 도시된 데이터 추출부에 의한 처리 흐름을 도시하는 도면이고, 도 5는 도 4에 도시된 바에 따라 추출된 데이터의 일 예를 도시하는 도면으로, 세균에 해당하는 데이터를 추출하기 위한 처리 흐름 및 예를 도시

한다.

[0036] 도 4를 참조하면, 데이터 추출부(140)는 입력된 데이터인 organ_result(S41)에서 세균명(organism명) 및 배양 정보(amount값 및 colony_count값)을 확인하여(예를 들어, organism명, amount값 및 colony_count값이 존재하는지 순차적으로 확인하여)(S42, S45, S47), 해당 사항이 없는 경우 None을 출력하고(S43, S46, S48), 해당 값이 모두 존재하는 경우 해당 데이터를 세균에 해당하는 데이터인 temp_df로 생성하여 추출할 수 있다(S49). 또한, organism명이 존재하는 경우에는 organ_check 함수를 통해 예외 처리와 오타 검사를 추가로 수행할 수 있다(S44).

[0037] 또한, 도 6은 도 1에 도시된 데이터 추출부에 의한 처리 흐름을 도시하는 도면이고, 도 7은 도 6에 도시된 바에 따라 추출된 데이터의 일 예를 도시하는 도면으로, 항생제 감수성 결과 데이터를 추출하기 위한 처리 흐름 및 예를 도시한다.

[0038] 도 6을 참조하면, 데이터 추출부(140)는 입력된 데이터인 temp_df의 organ_result(S61)에서 항생제 감수성에 대한 정보(예를 들어, Esbl의 결과, high 및 양성, 음성, 판정불가, sensitive, resistant, intermediate, equivocal 혹은 이에 해당하는 약자인 s, i, r이 있는지 여부)를 순차적으로 확인하여(S62, S64, S66), s, i, r을 기준으로 앞 부분을 항생제명(anti_name)으로 추출하고, 뒷 부분을 해당 항생제명에 대한 결과값(anti_value)으로 추출할 수 있다(S63, S65, S67). 이 경우, anti_check 함수를 통해 예외 처리와 오타 검사를 추가로 수행할 수 있다.

[0039] 또한, 데이터 추출부(140)는 입력된 세균배양검사결과보고서에 대한 데이터 추출 시에 감염검사 온톨로지 데이터베이스(DB1)의 응답에 기반하여 세균 및 항생제 감수성에 대한 결과를 생성 및 제공할 수 있다.

[0041] 표준용어 매핑부(150)는 데이터 추출부(140)에 의해 추출된 데이터에서 표준용어 온톨로지 데이터베이스(DB2) 및 도 9에 도시된 표준용어 라벨링 및 재정렬 알고리즘에 따라 최종 매핑된 표준용어 및 코드를 출력할 수 있다.

[0042] 도 9는 도 1에 도시된 표준용어 매핑부에 의한 처리 흐름을 도시하는 도면이고, 도 10은 도 9에 도시된 바에 따라 매핑된 데이터의 일 예를 도시하는 도면이다.

[0043] 도 9를 참조하면, 추출된 데이터, 즉 df에서 기 정해진 룰인 resistance_label Rule에 따라 라벨을 생성할 수 있다(S91, S92, S94). 이 경우, 해당하는 라벨이 없으면, 세균명(organism명)을 라벨로 생성할 수 있다(S93).

[0044] 또한, 해당하는 라벨이 복수개인 경우(S95), 기 정해진 계층(hierarchy)을 고려하여 순서를 재정렬하고(S97) 표준용어 온톨로지 데이터베이스(DB2)에 해당하는 표준코드(concept_id)로 매핑할 수 있다(S98).

[0045] 또한, 표준용어 매핑부(150)는 입력된 데이터의 정제 실행 시에 표준용어 온톨로지 데이터베이스(DB2)의 응답에 기반하여 매핑된 표준용어 및 코드를 생성 및 제공할 수 있다.

[0047] 감염검사 온톨로지 데이터베이스(DB1)는 세균배양검사에서 사용하는 세균, 항생제 및 기타 용어 또는 약자 등을 정리하여 구축한 데이터베이스이다.

[0048] 또한, 표준용어 온톨로지 데이터베이스(DB2)는 전 세계적으로 표준화된 코드를 매핑하기 위하여 예를 들어 SNOMED CT(SNOMED Clinical Terms), OMOP(Observational Medical Outcomes Partnership), LOINC(Logical Observation Identifiers Names and Codes) 등을 포함하는 표준용어를 바탕으로 구축한 데이터베이스이다. 여기서, SNOMED CT는 의학 임상 문서와 보고서에 사용되는 체계적으로 구조화된 의학용어의 집합으로 세계 최대의 다중언어를 지원하는 임상 용어체계이고, OMOP은 의학 분야 용어의 표준안을 만들어 제공하는 국제적 협의체인 OHDSI(Observational Health Data Sciences and Informatics) 기반의 표준용어이며, LOINC는 임상 및 의료 실험실 관찰을 식별하기 위해 사용되는 국제 표준용어이다. 이와 같은 3가지의 표준용어를 바탕으로 표준용어 온톨로지 데이터베이스(DB2)를 구축하면 약 6만 2천 개의 표준용어를 매핑할 수 있다. 도 8은 도 1에 도시된 표준용어 온톨로지 데이터베이스의 일 예를 도시하는 도면이다.

[0050] 상술한 바와 같이, 본 발명의 실시예에 따르면, 감염검사 온톨로지 데이터베이스(DB1) 및 표준용어 온톨로지 데

이터베이스(DB2)를 기반으로 하여 기 정해진 룰을 이용하여 세균배양검사 데이터를 정제함으로써 정확도를 크게 향상시킬 수 있다.

[0052] 도 11 및 도 12는 본 발명의 일 실시예에 따라 온톨로지 데이터베이스를 기반으로 룰 기반 알고리즘을 이용한 경우와 그렇지 않은 경우의 성능을 비교한 결과를 도시하는 도면이다.

[0053] 도 11 및 도 12를 참조하면, 본 발명의 일 실시예에 따라 온톨로지 데이터베이스를 사용하는 경우(도 12의 상단 박스) 온톨로지 데이터베이스를 사용하지 않는 경우(도 12의 하단 박스)에 비해 높은 추출 값을 얻을 수 있으며, 매핑 정확도에 있어서도 20% 이상 향상됨을 확인하였다. 즉, 본 발명에서는 세균배양검사결과보고서를 정제함에 있어서 온톨로지를 기반으로 룰 기반 알고리즘을 적용함으로써 보다 정확하고 효율적으로 정제를 수행할 수 있다.

[0055] 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

[0056] 100: 세균배양검사결과보고서 정제 시스템

110: 데이터 도입부

120: 텍스트 분할부

130: 전처리 필터부

140: 데이터 추출부

150: 표준용어 매핑부

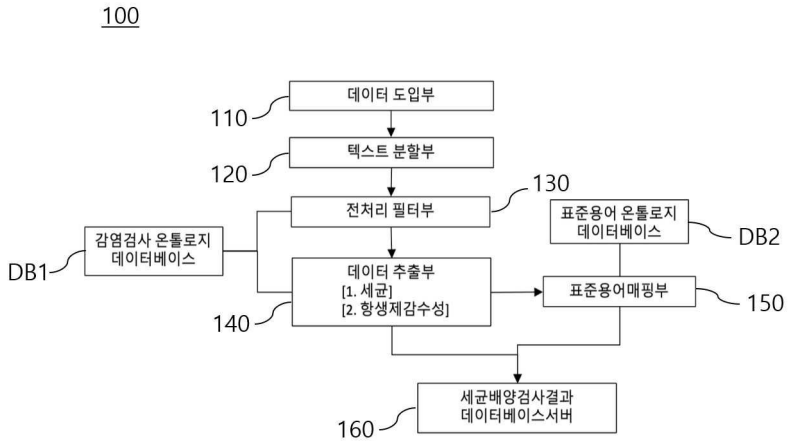
160: 세균배양검사결과 데이터베이스 서버

DB1: 감염검사 온톨로지 데이터베이스

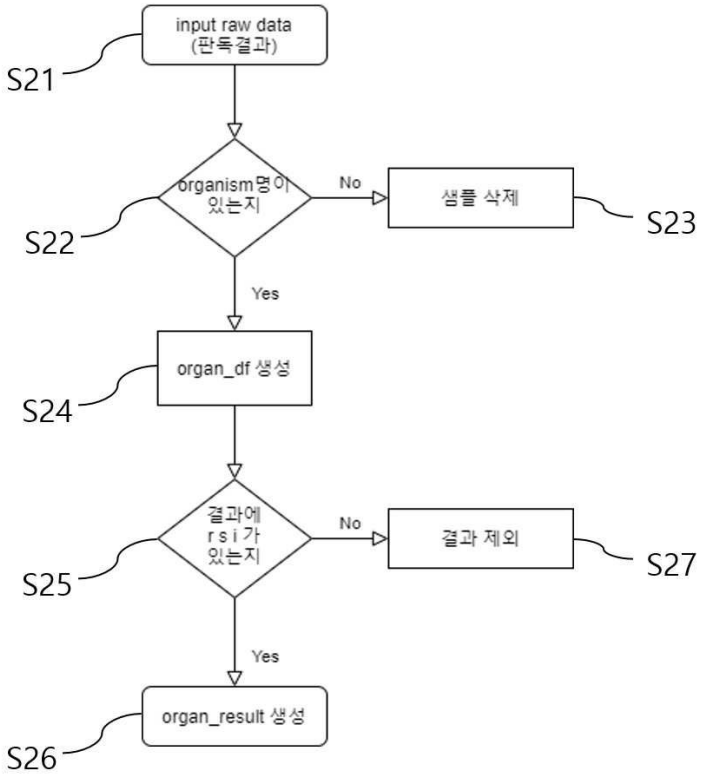
DB2: 표준용어 온톨로지 데이터베이스

도면

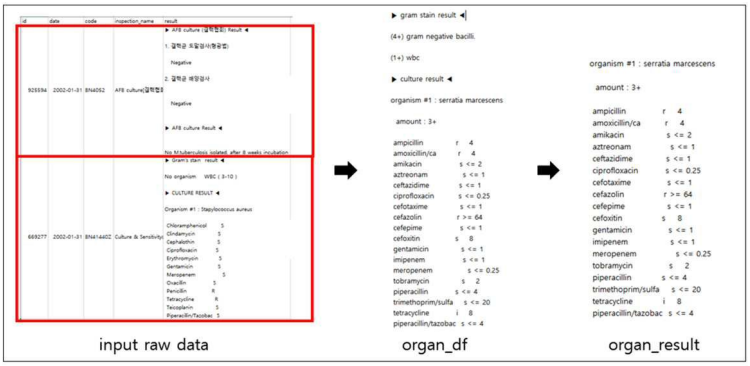
도면1



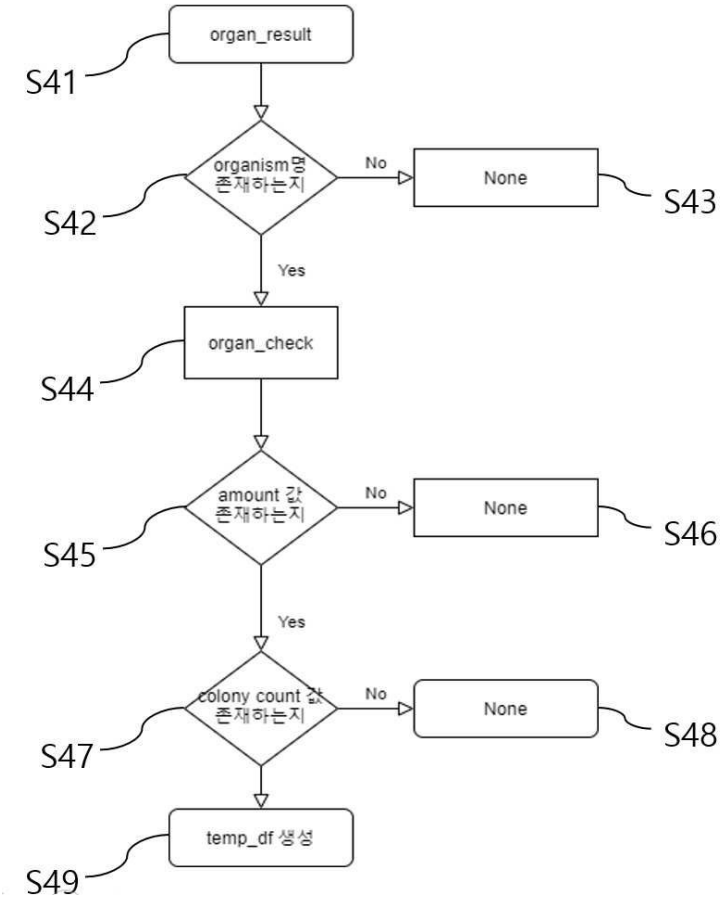
도면2


input raw data
(판독결과)
S21
organism명이
있는지
No
샘플 삭제
S23
S22
Yes
organ_df 생성
S24
결과에
r s i 가
있는지
No
결과 제외
S27
S25
Yes
organ_result 생성
S26

도면3


input raw data
organ_df
organism #1 : serratia marcescens
amount : 3+
ampicillin r 4
amoxicillin/ca r 4
amikacin s <= 2
aztreonam s <= 1
ceftazidime s <= 1
ciprofloxacin s <= 0.25
cefotaxime s <= 1
cefazolin r >= 64
cefepime s <= 1
cefoxitin s 8
gentamicin s <= 1
imipenem s <= 1
meropenem s <= 0.25
tobramycin s 2
piperacillin s <= 4
trimethoprim/sulfa s <= 20
tetracycline i 8
piperacillin/tazobac s <= 4
organ_result

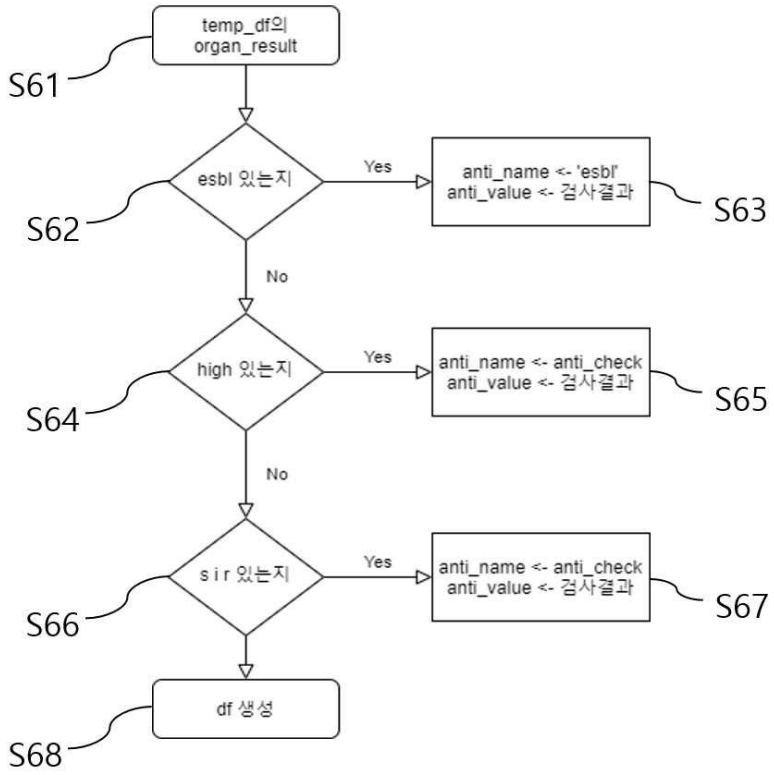
도면4



도면5

result	organism	amount	colony_count
organism #1 : corynebacterium minutissimum amount : 2+ ampicillin r ciprofloxacin r cefotaxime r penicillin r vancomycin s imipenem r tigecycline s	corynebacterium minutissimum	2+	None

도면6



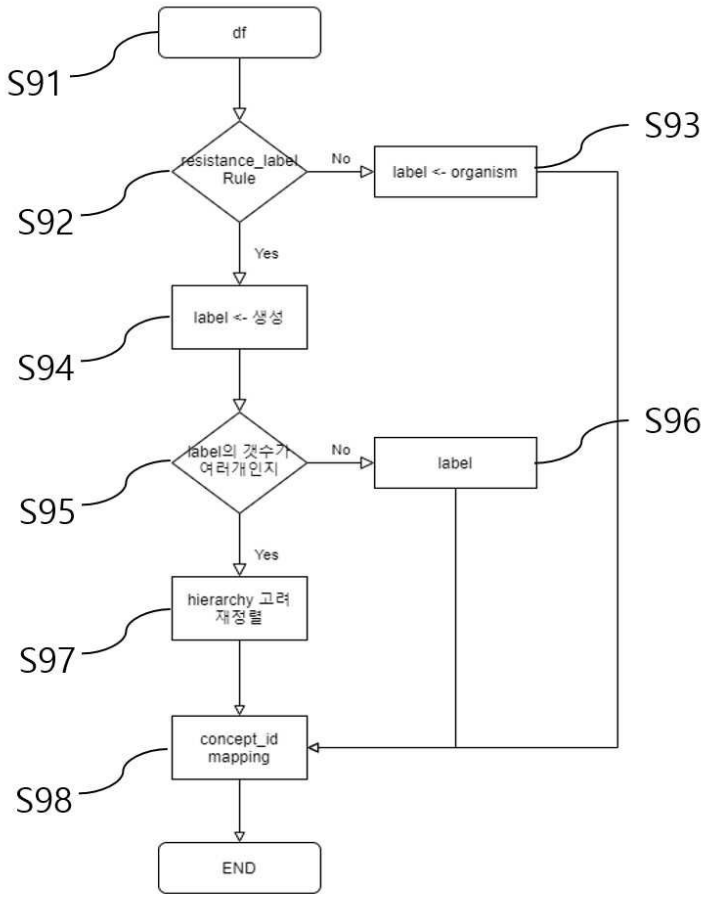
도면7

result	organism	amount	colony_count	high k	trimeth	ampicillin	norfloi	cefepime	cephal	cefaclc	piperacillin	minoc	etham	cefoxitin	voricol
organism #1 : serratia marcescens amount : 3+ ampicillin r 4 amoxicillin/ca r 4 amikacin s <= 2 aztreonam s <= 1 ceftazidime s <= 1 ciprofloxacin s <= 0.25 cefotaxime s <= 1 cefazolin r >= 64 cefepime s <= 1 cefoxitin s 8 gentamicin s <= 1 imipenem s <= 1 meropenem s <= 0.25 tobramycin s 2 piperacillin s <= 4 trimethoprim/sulfa s <= 20 tetracycline i 8 piperacillin/tazobac s <= 4	serratia marcescens	3+	None			['r', '4']		['s', '<=', '1']			['s', '<=', '4']			['s', '8']	

도면8

concept_id	concept_name
761937	Influenza A virus subtype H1 2009 pandemic strain
762597	Influenza A virus subtype H3N2 variant
762609	Aeromonas caviae complex
762610	Aeromonas veronii complex
762611	Aeromonas hydrophila complex
763143	Hippurate-negative Campylobacter jejuni/Campylobacter coli
763389	Prevotella oralis group
763395	Nocardia-like organism
763516	Coagulase-negative Staphylococcus, not Staphylococcus saprophyticus
763517	Lactobacillus-like gram positive bacillus
763523	Bacillus species not Bacillus anthracis or Bacillus cereus
763524	Streptococcus species not group A, B or D
763525	Mycobacterium franklinii
763686	Non-shiga toxin producing Escherichia coli
763860	Multidrug resistant Acinetobacter baumannii
763931	Shiga toxin-producing Escherichia coli serogroup O132
763932	Shiga toxin-producing Escherichia coli serogroup O78
765319	Influenza A virus subtype H1 swine-like
765611	Coccidioides immitis/Coccidioides posadasii
765850	Beta-hemolytic Streptococcus, non-group B

도면9



도면10

resistance_label	hierarchy_key	Key1	Key2	Key3	Key4	concept_id_fina
['multidrug resistant acinetobacter baumannii', 'fluoroquinolone resistant acinetobacter baumannii']	['multidrug resistant acinetobacter baumannii', 'fluoroquinolone resistant acinetobacter baumannii']	multidrug resistant acinetobacter baumannii	fluoroquinolone resistant acinetobacter baumannii			763860

도면11

	성능평가 (대상 27,542건)	
	추출 값	매핑 정확도 Pathology
온톨로지 데이터베이스 기반	27,542	99.92%
온톨로지 데이터베이스 기반 하지 않은	21,750	78.97%

도면12

```
In [5]: test=df[["concept_id","concept_id_final"]]
        test.head()
```

Out[5]:

	concept_id	concept_id_final
22	[4149419]	4149419
23	[4321152]	4321152
24	[4149419]	4149419
25	[4149419]	4149419
26	[4149419]	4149419

```
In [6]: len(test)
Out[6]: 27542
```

```
In [16]: len(test[test["concept_id_final"]=="no label"])
Out[16]: 21

In [8]: (test["concept_id"]==test["concept_id_final"].apply(lambda x : [x])).value_counts()
Out[8]: True     21771
        False     5771
        dtype: int64
```