(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 118394954 B
(45)授权公告日 2024.10.22

(21)申请号 202410595015.0

(22)申请日 2024.05.14

(65)同一申请的已公布的文献号
申请公布号 CN 118394954 A

(43)申请公布日 2024.07.26

(73)专利权人 中国医学科学院医学信息研究所
地址 100020 北京市朝阳区雅宝路3号

(72)发明人 吴思竹 胡拯涌 修晓蕾 王安然

(74)专利代理机构 北京睿智保诚专利代理事务所(普通合伙) 11732
专利代理师 刘刚

(51)Int.Cl.
*G06F 16/36*(2019.01)
*G06F 16/35*(2019.01)
*G16H 50/70*(2018.01)
*G06F 40/284*(2020.01)
*G06F 40/30*(2020.01)
*G06F 40/247*(2020.01)
*G06F 18/22*(2023.01)

(56)对比文件
CN 112199511 A,2021.01.08
CN 112542223 A,2021.03.23

审查员 姚子琪

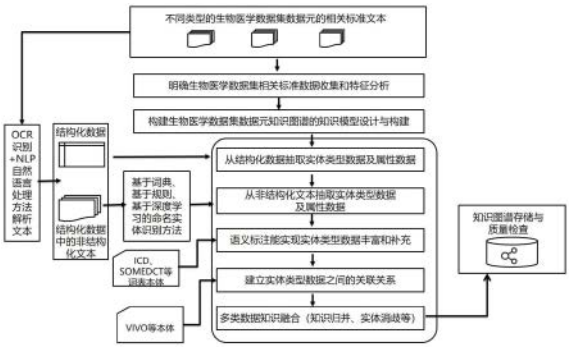
权利要求书4页 说明书15页 附图3页

(54)发明名称

一种生物医学数据集标准数据元的知识图谱构建方法及系统

(57)摘要

本发明公开了一种生物医学数据集标准数据元的知识图谱构建方法及系统,涉及医学数据处理技术领域,收集不同类型的生物医学数据集数据元的相关标准文本和生物医学数据集相关标准的数据;并进行分析和归纳;构建生物医学数据集标准数据元知识图谱的知识模型;从结构化数据和结构化数据中的非结构化文本抽取实体类型数据及属性数据;根据建立的实体类型之间的语义关联关系类型,进行多类数据的知识融合,得到生物医学数据集标准数据元知识图谱。本发明不仅增强领域数据集元数据和数据元、分类、值域标准的可用性和利用率,而且还实现数据元的统一和数据集创建的规范性以及提高机器可读性和语义互操作性。




CN 118394954 B

1.一种生物医学数据集标准数据元的知识图谱构建方法,其特征在于,包括:

收集不同类型的生物医学数据集数据元的相关标准文本和生物医学数据集相关标准的数据;

通过对收集数据元的相关标准文本和生物医学数据集相关标准的数据进行分析和归纳,用于支持构建生物医学数据集标准数据元知识图谱的知识模型和进行数据的解析和细粒度内容抽取;

构建生物医学数据集标准数据元知识图谱的知识模型,定义实体类型并同时建立各实体类型的属性和实体类型之间的语义关联关系类型;

从结构化数据和结构化数据中的非结构化文本抽取实体类型数据及属性数据;

根据建立的实体类型之间的语义关联关系类型,进行多类数据的知识融合,得到生物医学数据集标准数据元知识图谱;

实体类型之间的语义关联关系具体包括:数据标准之间的关系、数据元集合和数据元之间的关系、数据元与数据元概念之间的关系、数据元之间的关系、数据元与值域之间的关系、数据集标准与医学量表/问卷的关系、数据元与医学量表/问卷的关系;其中数据标准层面的关系是多元的;数据标准与数据元集合是包含关系,数据元集合和数据是包含关系,数据元集合下包含多个数据元;数据元之间的关系包括3类:同义关系、相关关系、无关关系;数据元值域根据值域来源和使用方式划分为枚举引它型、枚举自引型、枚举定义型和非枚举型四种类型;数据集标准中使用了医学量表,量表名称和信息从文本中提取,通过补足量表资源建立连接;数据元为医学量表规范化的数据库存储名称,建立数据元和特定医学量表之间的关联;

数据元之间的关系判断方法:

识别完数据元概念后,进行数据元同义关系识别,如果在任何同一医学领域主题词表中,数据元的概念相同,则两个数据元为同义关系,相似度标记为1;

如果非同义关系,则对两个标准编码和数据元标识完全不同的数据元进行相似度计算,计算方法采用了Jaccard相似度,集合的交集和并集的比值,计算公式如下:

$$Sim\_ele(E1,E2)=\frac{|A\cap B|}{|A|+|B|-|A\cap B|}$$

其中E1,E2分别表示两个数据元,每个数据元的文本被进行分词处理,E为该数据元的数据元名称和数据元定义组成的分词文本,Sim_ele_name()表示数据元相似度,A表示E1的分词文本,B表示E2的分词文本,最终相似度结果控制在[0,1]范围;

如果两个数据元非同义,则根据计算公式计算第一数据元和第二数据元的相似度值;如果两个数据元的相似度大于数据元同义阈值,二者为候选同义关系;

如果两个数据元的相似度大于数据元相关阈值,小于数据元同义阈值,二者为候选相关关系;

如果相似度小于数据元相关阈值,仅记录二者相似度值,则标记二者关系为无关;

判断数据元和值域的类型与关系方法如下:

a,数据元和对应值域,判断数据元的允许值是否包含标准号或值域代码表编号或名称,通过编码规则库进行判断,如果包括则为枚举引用;如果不包括,则执行步骤b;

b,如果为枚举引用,进一步判断是否当前引用值域的数据集标准编码或值域代码表编

码是当前数据元的标准编号或包含的值域代码表编码,不同则为枚举引它,如果为相同为枚举自引;

c,如果允许值域不满足a且值包含“;”分割的数字项则为枚举定义;

d,如果不属于c则为非枚举型。

2.根据权利要求1所述的一种生物医学数据集标准数据元的知识图谱构建方法,其特征在于,通过对不同类型的生物医学数据集数据元的相关标准文本,进行OCR识别加NLP自然语言处理方法解析文本,得到结构化数据和结构化数据中的非结构化文本。

3.根据权利要求1所述的一种生物医学数据集标准数据元的知识图谱构建方法,其特征在于,还包括知识图谱的存储与质量检查;存储,建立多张实体属性表和实体三元组关系表,批量转换,三元组导入转换为utf-8,用Neo4j图数据库来存储知识图谱;检查,将所有三元组数据导入neo4j之后,进行数据抽查,核对三元组数据的正确性,保证实体类型和关联关系的正确性。

4.根据权利要求1所述的一种生物医学数据集标准数据元的知识图谱构建方法,其特征在于,所述从结构化数据抽取实体类型数据及属性数据的具体过程为:

通过人机结合的方式对数据元的相关标准文本内容进行识别和提取;提取后的内容需进行数据清洗、数据审核和数据质控,标识类数据结合明确规定的编码规则要求编写正则表达式,对不同编码进行拼写检查和质控,对于有问题的标识进行修正,并对标识进行统一;提取的内容中存在识别错误、无用空格和换行、乱码和遗漏的情况,由人工进行补充和修改,完成所有文本内容的提取和整理,形成初步的结构化数据。

5.根据权利要求1所述的一种生物医学数据集标准数据元的知识图谱构建方法,其特征在于,所述从结构化数据中的非结构化文本抽取实体类型数据及属性数据的具体过程为:

从结构化数据中的非结构化文本中借助领域词表或机器学习方法识别抽取及标注,对实体类型进行人工标注和审核质控。

6.根据权利要求1所述的一种生物医学数据集标准数据元的知识图谱构建方法,其特征在于,所述多类数据的知识融合具体包括:

(1)利用已有唯一编码进行消歧,包括对跨级别编号进行处理;

(2)名称规范,通过《WS/T306卫生信息数据集分类与编码规则》、《WS370-2012卫生信息基本数据集编制规范制定规则》规则标准、机构规范库和领域词表、相似度计算和人工核查质控实现命名和编码的归一;其中术语、缩略语也通过领域主题词表、通用主题词表进行语义归并;

(3)数据元名称通过数据元间的相似度计算、数据元概念归并和人工判别实现归并;

(4)数据值域表名称归并,数据集标准文本中值域表和数据元允许值中均涉及值域表相关名称,包括表号、表编码和表名称,需要结构化处理这三个部分、进行数据纠错、组合归并,并且融合标准号,实现值域表的归并和消除歧义。

7.一种生物医学数据集标准数据元的知识图谱构建系统,应用如权利要求1-6任一所述的一种生物医学数据集标准数据元的知识图谱构建方法,其特征在于,包括以下模块:

数据收集模块,收集不同类型的生物医学数据集数据元的相关标准文本和生物医学数据集相关标准的数据;

数据分析模块,通过对收集数据元的相关标准文本和生物医学数据集相关标准的数据进行分析和归纳,用于支持构建生物医学数据集标准数据元知识图谱的知识模型和进行数据的解析和细粒度内容抽取;

知识模型构建模块,构建生物医学数据集标准数据元知识图谱的知识模型,定义实体类型并同时建立各实体类的属性和实体类型之间的语义关联关系类型;

实体类型抽取模块,从结构化数据和结构化数据中的非结构化文本抽取实体类型数据及属性数据;

知识图谱获取模块,根据建立的实体类型之间的语义关联关系类型,进行多类数据的知识融合,得到生物医学数据集标准数据元知识图谱;

实体类型之间的语义关联关系具体包括:数据标准之间的关系、数据元集合和数据元之间的关系、数据元与数据元概念之间的关系、数据元之间的关系、数据元与值域之间的关系、数据集标准与医学量表/问卷的关系、数据元与医学量表/问卷的关系;其中数据标准层面的关系是多元的;数据标准与数据元集合是包含关系,数据元集合和数据是包含关系,数据元集合下包含多个数据元;数据元之间的关系包括3类:同义关系、相关关系、无关关系;数据元值域根据值域来源和使用方式划分为枚举引它型、枚举自引型、枚举定义型和非枚举型四种类型;数据集标准中使用了医学量表,量表名称和信息从文本中提取,通过补足量表资源建立连接;数据元为医学量表规范化的数据库存储名称,建立数据元和特定医学量表之间的关联;

数据元之间的关系判断方法:

识别完数据元概念后,进行数据元同义关系识别,如果在任何同一医学领域主题词表中,数据元的概念相同,则两个数据元为同义关系,相似度标记为1;

如果非同义关系,则对两个标准编码和数据元标识完全不同的数据元进行相似度计算,计算方法采用了Jaccard相似度,集合的交集和并集的比值,计算公式如下:

$$Sim\_ele(E1,E2)=\frac{|A\cap B|}{|A|+|B|-|A\cap B|}$$

其中E1,E2分别表示两个数据元,每个数据元的文本被进行分词处理,E为该数据元的数据元名称和数据元定义组成的分词文本,Sim_ele_name()表示数据元相似度,A表示E1的分词文本,B表示E2的分词文本,最终相似度结果控制在[0,1]范围;

如果两个数据元非同义,则根据计算公式计算第一数据元和第二数据元的相似度值;如果两个数据元的相似度大于数据元同义阈值,二者为候选同义关系;

如果两个数据元的相似度大于数据元相关阈值,小于数据元同义阈值,二者为候选相关关系;

如果相似度小于数据元相关阈值,仅记录二者相似度值,则标记二者关系为无关;

判断数据元和值域的类型与关系方法如下:

a,数据元和对应值域,判断数据元的允许值是否包含标准号或值域代码表编号或名称,通过编码规则库进行判断,如果包括则为枚举引用;如果不包括,则执行步骤b;

b,如果为枚举引用,进一步判断是否当前引用值域的数据集标准编码或值域代码表编码是当前数据元的标准编号或包含的值域代码表编码,不同则为枚举引它,如果为相同为枚举自引;

c,如果允许值域不满足a且值包含“;”分割的数字项则为枚举定义;

d,如果不属于c则为非枚举型。

# 一种生物医学数据集标准数据元的知识图谱构建方法及系统

## 技术领域

[0001] 本发明涉及医学数据处理技术领域,更具体的说是涉及一种生物医学数据集标准数据元的知识图谱构建方法及系统。

## 背景技术

[0002] 目前,生物医学数据共享可提高医学研究效率,增强医学研究透明性,学术领域对研究复现和数据的公开也提出了硬性要求,越来越多的医学研究人员选择将原始生物医学数据公开乃至共享,但生物医学数据有着高复杂性语义,容易出现同义、歧义等情况,而共享的生物医学数据缺乏在数据字段或值域层面的统一标准和规范,导致数据语义模糊、不同数据集间无法比对和联合分析,例如,数据集中字段或变量“性别”的英文名称可以用gender或sex表示,值域上可以直接用文字表示为男性、女性,也可以用数值0和1的表示,0表示男性、1表示女性。如果没有统一的数据元名称和值域规范,对于不同数据集的同一语义的字段或变量就没有办法进行集成整合或者联合分析,研究者也难以理解数据语义和进行分析利用,极大地阻碍了数据共享。由此,数据集的元数据和数据元标准非常重要,能够规范和统一数据结构及语义表达。但当前的数据标准多以标准规范形式发布为PDF等非结构化形式,很多临床专业领域的数据集标准中涉及的数据元达200-300多个,而且不同数据元可能定义或使用了不同的值域,现仅能提供文本查找阅读和理解,而在数据元数据创建时很难有效利用、机器可读、可处理性差,这也是标准难以被应用和实施的原因。

[0003] 因此,如何在增强领域数据集元数据和数据元、分类、值域标准的可用性和利用率的基础上,提高机器可读性和语义互操作性是本领域技术人员亟需解决的问题。

## 发明内容

[0004] 有鉴于此,本发明提供了一种生物医学数据集标准数据元的知识图谱构建方法及系统,收集生物医学科学数据领域的数据集标准和分类、值域标准,进行碎片化和规范化处理,并通过词性、语义计算等进行数据元语义归并建立有效关联。而后设计生物医学数据集数据元知识模式和构建知识图谱,用于支持领域数据字段/变量的标准化和其值域标准化。本发明以生物医学数据集标准数据元为例,方法可推广到其他领域数据集的数据元知识图谱的设计和实现。以此一方面可以增强领域数据集元数据和数据元、分类、值域标准的可用性和利用率,另一方面有助于实现数据元的统一和数据集创建的规范性、细化和丰富跨数据集标准、数据元集合、数据元、数据元概念、数据值域之间的关联,以及提高机器可读性和语义互操作性。

[0005] 为了实现上述目的,本发明采用如下技术方案:

[0006] 一种生物医学数据集标准数据元的知识图谱构建方法,包括:

[0007] 收集不同类型的生物医学数据集数据元的相关标准文本和生物医学数据集相关标准的数据;

[0008] 通过对收集数据元的相关标准文本和生物医学数据集相关标准的数据进行分析

和归纳,用于支持构建生物医学数据集标准数据元知识图谱的知识模型和进行数据的解析和细粒度内容抽取;

[0009] 构建生物医学数据集标准数据元知识图谱的知识模型,定义实体类型并同时建立各实体类的属性和实体类型之间的语义关联关系类型;

[0010] 从结构化数据和结构化数据中的非结构化文本抽取实体类型数据及属性数据;

[0011] 根据建立的实体类型之间的语义关联关系类型,进行多类数据的知识融合,得到生物医学数据集标准数据元知识图谱。

[0012] 可选的,通过对不同类型的生物医学数据集数据元的相关标准文本,进行OCR识别+NLP自然语言处理方法解析文本,得到结构化数据和结构化数据中的非结构化文本。

[0013] 可选的,还包括知识图谱的存储与质量检查;存储,建立多张实体属性表和实体三元组关系表,批量转换,三元组导入转换为utf-8,用Neo4j图数据库来存储知识图谱;检查,将所有三元组数据导入neo4j之后,进行数据抽查,核对三元组数据的正确性,保证实体类型和关联关系的正确性。

[0014] 可选的,所述从结构化数据抽取实体类型数据及属性数据的具体过程为:

[0015] 通过人机结合的方式进行文本内容的识别和提取;提取后的内容需进行数据清洗、数据审核和数据质控,标识类数据结合明确规定的编码规则要求编写正则表达式,对不同编码进行拼写检查和质控,对于有问题的标识进行修正,并对标识进行统一;提取的内容中存在识别错误、无用空格和换行、乱码和遗漏的情况,由人工进行补充和修改,完成所有文本内容的提取和整理,形成初步的结构化数据。

[0016] 可选的,所述从结构化数据中的非结构化文本抽取实体类型数据及属性数据的具体过程为:

[0017] 从结构化数据中的非结构化文本中借助领域词表或机器学习方法识别抽取及标注,对实体类型进行人工标注和审核质控,用于丰富和增强数据集标准和数据元的领域特征和应用场景特征,进而实现更细粒度和更多维度内容的揭示。

[0018] 可选的,实体类型之间的关联关系具体包括:数据标准之间的关系、数据元集和数据元之间的关系、数据元与数据元概念之间的关系、数据元之间的关系、数据元与值域之间的关系、数据集标准与医学量表/问卷的关系、数据元与医学量表/问卷的关系;其中数据标准层面的关系是多元的;数据标准与数据元集合是包含关系,数据元集合和数据元是包含关系,数据元集合下包含多个数据元;数据元之间的关系包括3类:同义关系、相关关系、无关关系;数据元值域根据值域来源和使用方式划分为枚举引它型、枚举自引型、枚举定义型和非枚举型四种类型;数据集标准中使用了医学量表,量表名称和信息从文本中提取,通过补足量表资源建立连接;数据元为医学量表规范化的数据库存储名称,建立数据元和特定医学量表之间的关联。

[0019] 可选的,数据元之间的关系判断方法:

[0020] 识别完数据元概念后,进行数据元同义关系识别,如果在任何同一医学领域主题词表中,数据元的概念相同,则两个数据元为同义关系,相似度标记为1;

[0021] 如果非同义关系,则进入数据元相似度计算程序,两个标准编码和数据元标识完全不同的数据元进行相似度计算,计算方法采用了Jaccard相似度,集合的交集和并集的比值,计算公式如下:

[0022] $$Sim\_ele(E1,E2)=\frac{|A\cap B|}{|A|+|B|-|A\cap B|}$$

[0023] 其中E1,E2分别表示两个数据元,每个数据元的文本被进行分词处理,E为该数据元的数据元名称和数据元定义组成的分词文本,Sim_ele_name()表示数据元相似度,A表示E1的分词文本,B表示E2的分词文本,最终相似度结果控制在[0,1]范围;

[0024] 如果两个数据元非同义,则根据计算公式计算第一数据元和第二数据元的相似度值;如果两个数据元的相似度大于数据元同义阈值,二者为候选同义关系;

[0025] 如果两个数据元的相似度大于数据元相关阈值,小于数据元同义阈值,二者为候选相关关系;

[0026] 如果相似度小于数据元相关阈值,仅记录二者相似度值,则标记二者关系为无关。

[0027] 可选的,判断数据元和值域的类型与关系方法如下:

[0028] a,数据元和对应值域,判断数据元的允许值是否包含标准号或值域代码表编号或名称,通过编码规则库进行判断,如果包括则为枚举引用;如果没有跳转进入下一条件判断;

[0029] b,如果为枚举引用,进一步判断是否当前引用值域的数据集标准编码或值域代码表编码是当前数据元的标准编号或包含的值域代码表编码,不同则为枚举引它,如果为相同为枚举自引;

[0030] c,如果允许值域不满足a且值包含“;”分割的数字项则为枚举定义;

[0031] d,如果不属于c则为非枚举型。

[0032] 可选的,所述多类数据的知识融合具体包括:

[0033] (1)利用已有唯一编码进行消歧,但跨级别编号还是需要进一步处理;

[0034] (2)名称规范,通过《WS/T306卫生信息数据集分类与编码规则》、

[0035] 《WS370-2012卫生信息基本数据集编制规范制定规则》规则标准、机构规范库和领域词表、相似度计算和人工核查质控实现命名和编码的归一;其中术语、缩略语也通过领域主题词表、通用主题词表进行语义归并;

[0036] (3)数据元名称通过数据元间的相似度计算、数据元概念归并和人工判别实现归并;

[0037] (4)数据值域表名称归并,数据集标准文本中值域表和数据元允许值中均涉及值域表相关名称,包括表号、表编码和表名称,需要结构化处理这三个部分、进行数据纠错、组合归并,并且融合标准号,实现值域表的归并和消除歧义。

[0038] 另一方面,提供一种生物医学数据集标准数据元的知识图谱构建系统,包括以下模块:

[0039] 数据收集模块,收集不同类型的生物医学数据集数据元的相关标准文本和生物医学数据集相关标准的数据;

[0040] 数据分析模块,通过对收集数据元的相关标准文本和生物医学数据集相关标准的数据进行分析和归纳,用于支持构建生物医学数据集标准数据元知识图谱的知识模型和进行数据的解析和细粒度内容抽取;

[0041] 知识模型构建模块,构建生物医学数据集标准数据元知识图谱的知识模型,定义实体类型并同时建立各实体类的属性和实体类型之间的语义关联关系类型;

[0042] 实体类型抽取模块,从结构化数据和结构化数据中的非结构化文本抽取实体类型数据及属性数据;

[0043] 知识图谱获取模块,根据建立的实体类型之间的语义关联关系类型,进行多类数据的知识融合,得到生物医学数据集标准数据元知识图谱。

[0044] 经由上述的技术方案可知,与现有技术相比,本发明公开提供了一种生物医学数据集标准数据元的知识图谱构建方法及系统,收集不同类型的生物医学数据集数据元的相关标准文本和生物医学数据集相关标准的数据;通过对收集数据元的相关标准文本和生物医学数据集相关标准的数据进行分析和归纳,用于支持构建生物医学数据集标准数据元知识图谱的知识模型和进行数据的解析和细粒度内容抽取;构建生物医学数据集标准数据元知识图谱的知识模型,定义实体类型并同时建立各实体类的属性和实体类型之间的语义关联关系类型;从结构化数据和结构化数据中的非结构化文本抽取实体类型数据及属性数据;根据建立的实体类型之间的语义关联关系类型,进行多类数据的知识融合,得到生物医学数据集标准数据元知识图谱。本发明不仅可以增强领域数据集元数据和数据元、分类、值域标准的可用性和利用率,而且还有助于实现数据元的统一和数据集创建的规范性、细化和丰富跨数据集标准、数据元集合、数据元、数据元概念、数据值域之间的关联,以及提高机器可读性和语义互操作性。

## 附图说明

[0045] 为了更清楚地说明本发明实施例或现有技术中的技术方案,下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍,显而易见地,下面描述中的附图仅仅是本发明的实施例,对于本领域普通技术人员来讲,在不付出创造性劳动的前提下,还可以根据提供的附图获得其他的附图。

[0046] 图1附图为本发明提供的生物医学科学数据集标准数据元知识图谱构建框架图;

[0047] 图2附图为本发明提供的生物医学数据集标准数据元知识图谱知识模型示意图;

[0048] 图3附图为本发明提供的建立的部分知识图谱数据实例图;

[0049] 图4附图为本发明提供的数据集标准、数据元集合、数据元及值域代码间的关系图。

## 具体实施方式

[0050] 下面将结合本发明实施例中的附图,对本发明实施例中的技术方案进行清楚、完整地描述,显然,所描述的实施例仅仅是本发明一部分实施例,而不是全部的实施例。基于本发明中的实施例,本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例,都属于本发明保护的范围。

[0051] 本发明实施例公开了一种生物医学数据集标准数据元的知识图谱构建方法,如图1所示,包括:

[0052] 收集不同类型的生物医学数据集数据元的相关标准文本和生物医学数据集相关标准的数据;

[0053] 通过对收集数据元的相关标准文本和生物医学数据集相关标准的数据进行分析和归纳,用于支持构建生物医学数据集标准数据元知识图谱的知识模型和进行数据的解析

和细粒度内容抽取；

[0054] 构建生物医学数据集标准数据元知识图谱的知识模型，定义实体类型并同时建立各实体类的属性和实体类型之间的语义关联关系类型；

[0055] 从结构化数据和结构化数据中的非结构化文本抽取实体类型数据及属性数据；

[0056] 根据建立的实体类型之间的语义关联关系类型，进行多类数据的知识融合，得到生物医学数据集标准数据元知识图谱。

[0057] 其中，收集的领域数据集相关标准包括但不限于数据集标准、分类和编码标准以及值域代码标准，同时扩展涉及的相关外部资源包括科学文献、医学词表(ICD、UMLS等)等，涉及的标准级别包括国家标准、行业标准、地方标准和团体标准等。其中，数据集标准既包括卫生信息数据元目录、卫生信息数据元值域代码、疾病控制基本数据集、基本信息数据集、医疗服务基本数据集、电子病历基本数据集等多类通用数据集，也包括骨伤科、中医药、高血压等电子病历基本数据集等专病数据集标准。本发明通过对这些不同级别、类型的领域标准的结构和重要数据集标准数据元等特征进行分析和归纳，用于支持构建生物医学数据集标准数据元知识图谱的知识模型和进行数据的解析和细粒度内容抽取。

[0058] 具体的，医学科学数据集标准数据元知识图谱构建的核心在于面向特定需求的图谱知识模型的设计和建设。虽然现有研究中有少部分聚焦于通用标准文本的知识图谱构建，但普遍存在知识粒度过粗、标准化程度和关联程度低等问题，缺乏针对特定领域和应用的精细化框架建模和知识抽取及关联关系构建，机器可读的数据集标准构建、数据元复用和数据值域复用。本发明在数据处理和图谱建设中主要参考ISO/IEC 11179《元数据注册系统》标准、《卫生健康信息基本数据集编制标准》等，面向生物医学领域数据集标准建设、数据元管理、集成、使用、复用、创建、对比等业务需求发展目标，进行设计了生物医学数据集标准数据元知识图谱的知识模型。实现对生物医学数据集标准的细粒度拆分和语义丰富，知识模型共包括但不限于21种实体类型和30种关系类型，其中这些实体和关系可以结合需求进一步扩展，建立细粒度的不同类型标准、内容单元和不同类型资源的细粒度关联和计算特定实体间的关联程度。

[0059] 生物医学数据集标准数据元知识图谱的知识模式中的实体类型包括但不限于标准、术语、缩略语、规定内容、适用范围、前言、引言、数据元集合、数据元、数据元概念、值域代码、疾病、领域、科室、出版物、归口单位、提出单位、起草单位等21类实体。同时建立各实体类型的属性和实体类型之间的语义关联关系类型。知识模型如图2所示。

[0060] 在一个具体的实施例中，通过对不同类型的生物医学数据集数据元的相关标准文本，进行OCR识别+NLP自然语言处理方法解析文本，得到结构化数据和结构化数据中的非结构化文本。

[0061] 在一个具体的实施例中，还包括知识图谱的存储与质量检查；存储，建立多张实体属性表和实体三元组关系表，批量转换，三元组导入转换为utf-8，用Neo4j图数据库来存储知识图谱；检查，将所有三元组数据导入neo4j之后，进行数据抽查，核对三元组数据的正确性，保证实体类型和关联关系的正确性。

[0062] 在一个具体的实施例中，从结构化数据抽取实体类型数据及属性数据的具体过程为：

[0063] 本发明涉及的标准文档，无论是数据集标准、学科代码标准还是代码值域标准，都

有不同的文本结构,参考《WS/T 370—2022卫生健康信息基本数据集编制标准》、《T/CHIA6-2018专科电子病历数据集编制规范》等指导文件,并结合实际文本和国家标准、行业标准、地方标准和团体标准的差异,针对每类文本结构进行文本解析和内容单元识别。对于多类标准的共性内容单元进行合并和共性特征提取,特有单元进行单独提取。针对不同文本结构设计数据库用于存储提取的结构化对象。

[0064] 由于文本类型属于非结构化文本,多以.pdf和.doc格式为主,因此,通过人机结合的方式进行文本内容的识别和提取。机器方式主要通过OCR图像识别和PDF内容抽取技术,如前言、引言、规定内容、适用范围、引用文件、术语、缩略语、参考文献等。提取后的内容需进行数据清洗、数据审核和数据质控,例如对于标准号、内部标识符等标识类数据结合明确规定的编码规则要求编写正则表达式,对不同编码进行拼写检查和质控,对于有问题的标识进行修正,并对标识进行统一,便于归一化和统计,例如其中标准号是唯一的,可以被直接用于图谱构建,而数据元标识符在不同标准中可能重复,不可以直接用于标识,需要重新定义唯一编码。此外,提取的内容中会存在识别错误、无用空格和换行、乱码和遗漏等情况,需由人工进行补充和修改,完成所有文本内容的提取和整理,形成初步的结构化数据。

[0065] 在一个具体的实施例中,从结构化数据中的非结构化文本抽取实体类型数据及属性数据的具体过程为:

[0066] 并不是所有实体类型均来自结构化数据,一些表征生物医学领域标准特征的数据需要从结构化数据中的非结构化描述如标题、摘要等中借助领域词表或机器学习方法识别抽取及标注,实体类型疾病、科室、主题词等进行人工标注和审核质控,用于丰富和增强数据集标准和数据元的领域特征和应用场景特征,进而实现从生物医学领域标准到数据元集合、数据元到值域等实现更细粒度和更多维度内容的揭示。

[0067] 数据元的概念识别,在本发明收集的数据集标准中,少部分数据集标准例如广东省医院协会发布的团体标准,涉及的专病领域包含慢性疾病、高血压病、冠心病、脑梗死等,标准参考ISO/IEC 11179《元数据注册系统》标准,如《T/GDPHA 031—2021脑血管疾病研究通用标准数据集》等团标中已实现数据元和CDISC、SNOMED CT、LOINC、NIH CDE等词表或通用数据元仓储中的数据元概念间的映射,标注了概念英文名称或概念ID编码。因此,可以从这类数据集标准中提取数据元和数据元概念间的关系。这类数据元概念的提取是基于英文医学领域词表/本体获得的。但多数数据集标准中的数据元是没有定义数据元概念的,并且均为中文表达。因此,本发明中利用中文/英文医学领域词表/本体获得数据元的概念,例如医学主题词表包括主题词、入口词,具有概念树层次结构,一个主题词下包含多个具有同义关系的入口词。通过数据元和主题词及该主题词下入口词的匹配,以获得数据元的概念。

[0068] 此外,如涉及到具体资源如参考论文、引用的政策、引用的标准等资源需要通过数据抓取文本或补充外源性可获得链接信息,保证数据关联和资源的可访问性。

[0069] 在一个具体的实施例中,实体类型之间的关联关系构建:

[0070] 下面重点阐述需要建立的几类重要的实体之间的关系的定义和处理过程。

[0071] (1)数据标准之间的关系,标准层面的关系是多元的,比如数据集标准引用了其他标准,新的标准替代了废止的标准、标准遵循了其他标准等。此外,通常被忽略的一种关系是标准之间的组成关系。一个数据集标准可能是由多个标准组成的,如高血压专科电子病历数据集,包括等14个部分。这些标准之间共同构成数据集的标准,这些标准之间是有同属

于一个数据集的组成部分的关系。值域标准也类似,如《WS 364卫生信息数据元值域代码》,包括人口学及社会经济学特征,健康史,健康危险因素等17个部分,其中除第1部分和第2部分为编制规则外,有15个部分为可用代码表,它们共同构成卫生信息数据元值域。标准之间的关系如下表:

[0072] 表1数据标准之间的关系

[0073]

| 实体类型1 | 关系类型 | 实体类型2 |
|---|---|---|
| 数据集标准 | 属于 | 数据集标准 |

[0074]

| | | |
|---|---|---|
| 数据集标准 | 引用 | 标准 |
| 数据集标准(变化和修改) | 替代 | 数据集标准 |
| 数据集标准 | 引用 | 出版物 |
| 数据元集合 | 补充 | 数据元 |
| 数据元集合 | 复用 | 数据元 |
| 数据元集合 | 包含 | 数据元 |
| 数据集标准 | 参考 | 出版物 |
| 数据集标准 | 包含 | 值域 |
| 数据集标准 | 涉及 | 医学量表/问卷 |

[0075] (2)数据元集合和数据元之间的关系。数据元集合在生物医学数据集标准中具体体现为特定命名的数据元专用属性集合。每个数据元专用属性集合中一般包含很多数据元。现有研究和应用中忽略了数据元的集合划分。数据集标准中数据元的专用属性下是对数据元的分类,例如胃癌临床科学研究通用数据元标准,包含7个数据元专用属性集合,分别是通用数据元、受试者人口学基本信息、受试者门(急)诊病历、受试者检查信息、受试者检验信息、胃受试者入院出院信息、受试者不良事件信息。因此,数据标准与数据元集合是包含关系,数据元集合和数据是包含关系,数据元集合下包含多个数据元。

[0076] (3)数据元与数据元概念间的关系。数据元主要来自数据元专用属性,其中中文生物医学数据集标准中的数据元多数没有按照ISO/IEC 11179《元数据注册系统》标准中的要求提供数据元概念和对象的信息。这部分需要借助医学领域主题词表等进行数据元的概念补全。

[0077] (4)数据元之间的关联关系。数据元之间的关系包括3类:同义关系、相关关系、无关关系。具体通过以下步骤实现数据元之间关系的确定。

[0078] 1)识别完数据元概念后,进行数据元同义关系识别,如果在任何同一医学领域主题词表中,数据元的概念相同,则两个数据元为同义关系,相似度标记为1。

[0079] 2)如果非同义关系,则进入数据元相似度计算程序,两个标准编码和数据元标识完全不同的数据元相似度计算,计算方法采用了Jaccard相似度,集合的交集和并集的比值,见公式1:

[0080] $$Sim\_ele(E1,E2)=\frac{|A\cap B|}{|A|+|B|-|A\cap B|}$$ (公式1)

[0081] 其中E1,E2分别表示两个数据元,每个数据元的文本被进行分词处理,E为该数据元的数据元名称和数据元定义组成的分词文本,Sim_ele_name()表示数据元相似度,A表示E1的分词文本,B表示E2的分词文本,最终相似度结果控制在[0,1]范围。

[0082] 3)如果两个数据元非同义,则根据公式1计算第一数据元和第二数据元的相似度值;如果两个数据元的相似度大于数据元同义阈值,二者为候选同义关系;

[0083] 4)如果两个数据元的相似度大于数据元相关阈值,小于数据元同义阈值,二者为候选相关关系;

[0084] 5)如果相似度小于数据元相关阈值,仅记录二者相似度值,则标记二者关系为无关。

[0085] 6)每对数据元的候选关系不能仅通过相似度计算获得,还需通过人工核查和调整后确定准确的关系,以确保关系的准确性。由此建立数据元间的多维细粒度的关联性和关联程度,为后续数据元创建和复用提供智能推荐。

[0086] (5)数据元与值域之间的关系。本发明细化了数据元和值域之间的关系,将数据元使用值域的方式进行了细粒度的划分。数据元值域根据值域来源和使用方式划分为枚举引它型、枚举自引型、枚举定义型和非枚举型四种类型。

[0087] 枚举引它型指引用其它标准(非数据元和值域所在本标准内的)值域表,有明确值域标准或值域表名称。

[0088] 枚举自引型指引用数据元和值域所在标准内定义的值域表,允许值条目大于4项,且有明确的表名和表编码。

[0089] 枚举定义型指数据元和值域所在标准内在数据元部分定义了允许值,但没有使用值域表形式,一般定义的允许值条目少于4项。

[0090] 非枚举型指没有按允许值条目列出的值域,一般采用文字进行描述允许值或自由填写方式。

[0091] 基于上述定义和方法,判断数据元和值域的类型与关系方法如下。

[0092] 1)数据元和对应值域,判断数据元的允许值是否包含标准号或值域代码表编号或名称,通过编码规则库进行判断,如果包括则为枚举引用;如果没有跳转进入下一条件判断。

[0093] 2)如果为枚举引用,进一步判断是否当前引用值域的数据集标准编码或值域代码表编码是当前数据元的标准编号或包含的值域代码表编码,不同则为枚举引它,如果为相同为枚举自引。

[0094] 3)如果允许值域不满足1)且值包含“;”分割的数字项则为枚举定义;

[0095] 4)如果不属于3)则为非枚举型。

[0096] (6)数据集标准与医学量表/问卷的关系。数据集标准中使用了医学量表,量表名称和信息从文本中提取,通过补足量表资源建立连接。

[0097] (7)数据元与医学量表/问卷的关系。数据元为医学量表规范化的数据库存储名称,建立数据元和特定医学量表之间的关联。

[0098] 具体的,数据集标准、数据元集合、数据元及值域代码间的关系如图4所示。

[0099] 知识融合:

[0100] 通过知识合并、实体消歧、共指消解等方法实现知识融合。不同类实体类型的实例

数据需要进行去重和消歧,针对不同实体类型的特点进行针对性的处理。

[0101] (1)利用已有唯一编码进行消歧,但跨级别编号还是需要进一步处理。例如数据标准通过对标准名称和标准号进行处理,虽然标准号是唯一的,但是,数据集标准文档中可能在不同的位置出现不同的标准描述,如标准名称、标准号、名称缩写等,这些会导致同一对象无法被认定。类似的还有值域代码表名称、词表名称存在差异,如CV03.00.107、WS364.5CV03.00.107饮食习惯代码表、WS364.5卫生信息数据元值域代码第5部分,其实都是对应一个值域代码表。标准内部编码和外部编码也存在重复问题,因为目前没有提供细粒度中文数据元查询系统,因此会导致编码重复。

[0102] (2)名称规范,机构名称也有不同的表达,如卫生部统计信息中心、中华人民共和国卫生部统计信息中心、卫生部卫生统计信息中心也是共指一家单位需要进行名称规范和归并。因此通过《WS/T306卫生信息数据集分类与编码规则》、《WS370-2012卫生信息基本数据集编制规范制定规则》等规则标准、机构规范库和领域词表、相似度计算和人工核查质控实现命名和编码的归一。术语、缩略语等也通过领域主题词表、通用主题词表等进行语义归并。

[0103] (3)数据元名称通过数据元间的相似度计算、数据元概念归并和人工判别实现归并。

[0104] (4)数据值域表名称归并,数据集标准文本中值域表和数据元允许值中均涉及值域表相关名称,包括表号、表编码和表名称,需要结构化处理这三个部分、进行数据纠错、组合归并等,并且融合标准号实现值域表的归并和消除歧义。

[0105] 数据存储:

[0106] 建立了多张实体属性表和实体三元组关系表,批量转换,三元组(主语、谓词、宾语)导入转换应为utf-8避免乱码。选择用Neo4j图数据库来存储知识图谱。对于Neo4j数据库的数据导入可使用Neo4j-import工具导入整理好的结构化三元组知识数据形成最终的知识图谱,并通过Cyber语句可查询和可视化全部数据,用于支持生物医学数据集标准数据元知识图谱实体和关系的查询。

[0107] 数据更新:

[0108] 随着生物医学数据集标准的新标准制定、原有标准修订,内容会发生变化。持续进行数据集标准、数据元相关数据的收集和加工,对于变化的内容对应实体类型进行实例数据的更新和补充。进行数据元和标准、机构的归并,并将新生成的类和数据填充到生物医学数据集标准数据元知识图谱中。

[0109] 具体的,引入一个具体的实施例来进一步解释本发明。

[0110] (1)设计知识模型的实体类型和实体间关系,如表2、表3所示。

[0111] 表2实体类型示例

[0112]

| 序号 | 实体类型 | 说明 |
|---|---|---|
| 1 | 数据标准 | 规范数据集、值域代码、分类编码等的标准 |
| 2 | 术语 | 标准中使用的术语 |
| 3 | 缩略语 | 标准中使用的缩略语 |
| 4 | 规定内容 | 标准内容的主要技术内容 |
| 5 | 适用范围 | 标准的适用的范围、场景等方面 |
| 6 | 前言 | 标准文本中对应的前言章节 |
| 7 | 引言 | 标准文本中对应的引言章节，不是所有标准文本均包含 |
| 8 | 领域 | 对应学科分类的二级分类 |
| 9 | 出版物 | 标准文本参考的参考文献 |
| 10 | 科室 | 根据标准内容，如某种疾病可能对应的医院科室 |
| 11 | 疾病 | 根据标准内容，可能关于某种疾病，可取ICD10编码 |
| 12 | 数据元 | 标准文本中的数据元 |
| 13 | 数据元概念 | 数据元的规范化名称和来源 |
| 14 | 归口机构 | 标准的归口机构名称 |
| 15 | 提出机构 | 标准的提出机构名称 |
| 16 | 起草机构 | 标准的起草机构名称 |
| 17 | 注册机构 | 标准的注册机构名称 |
| 18 | 数据元集合 | 多个数据元的分类集合，是根据数据元特征进行分类 |
| 19 | 值域代码 | 数据元的值的范围、来源等。 |
| 20 | 起草人 | 标准的起草人的名称 |
| 21 | 医学量表 | 医学量表是一种由多个项目构成的标准化测量工具，旨在揭示那些不适宜用直接方法测量的理论变量的水平。 |

[0113] 表3实体类型关系示例

[0114]

| 序号 | 实体类型1 | 关系属性 | 实体类型2 |
|---|---|---|---|
| 1 | 标准 | 引用 | 标准 |

[0115]

| | | | |
|---|---|---|---|
| 2 | 标准 | 属于 | 标准 |
| 3 | 标准 | 采用 | 标准 |
| 4 | 标准 | 替代 | 标准 |
| 5 | 标准 | 前言 | 前言 |
| 6 | 标准 | 引言 | 引言 |
| 7 | 标准 | 包含术语 | 术语 |
| 8 | 标准 | 包含缩略语 | 缩略语 |
| 9 | 标准 | 适用范围 | 适用范围 |
| 10 | 标准 | 规定内容 | 规定内容 |
| 11 | 标准 | 参考 | 出版物 |
| 12 | 标准 | 有关科室 | 科室 |
| 13 | 标准 | 有关疾病 | 疾病 |
| 14 | 标准 | 有关领域 | 领域 |
| 15 | 标准 | 起草 | 起草机构 |
| 16 | 标准 | 发布 | 发布机构 |
| 17 | 标准 | 归口 | 归口机构 |
| 18 | 标准 | 提出 | 提出机构 |
| 19 | 标准 | 起草 | 起草人 |
| 20 | 标准 | 包含数据元集合 | 数据元集合 |
| 21 | 数据元集合 | 包含 | 数据元 |
| 22 | 数据元 | 属于 | 数据元概念 |
| 23 | 数据元 | 枚举引它 | 值域代码 |
| 24 | 数据元 | 枚举自引 | 值域代码 |
| 25 | 数据元 | 枚举定义 | 值域代码 |
| 26 | 数据元 | 非枚举 | 值域代码 |
| 27 | 数据元 | 复用 | 数据元 |
| 28 | 数据元 | 补充 | 数据元 |
| 29 | 标准 | 包含 | 值域代码 |

[0116] (2)从结构化数据和非结构化数据中提取结构化实体类型实例,如表4、表5、表6所示。

[0117] 表4提取的结构化实体类型及属性——标准实例

[0118]

| | |
|---|---|
| 标准号 | WS 375.20-2016 |
| 国际标准分类号 | 11.020 |
| 中国标准分类号 | C 07 |
| 标准中文名称 | 疾病控制基本数据集 第20部分：脑卒中登记报告 |
| 标准类型 | 数据集标准 |
| 标准英文名称 | Basic dataset of discase control-Part 20: Registration and report of stroke |
| 类目 | 行业标准 |
| 发布者 | 中华人民共和国国家卫生和计划生育委员会 |
| 前言 | WS 375《疾病控制基本数据集》现分为以下部分：<br>——第1部分：艾滋病综合防治；…<br>本部分为WS 375的第20部分。本部分按照GB/T 1.1—2009给出的规则起草。<br>本部分主要起草单位：上海市疾病预防控制中心、上海市预防医学研究院、上海市卫生和计划生育委员会、上海市卫生和计划生育委员会信息中心。本部分主要起草人：吴凡、缪隼、李新建、仲伟鉴、夏天、姜智海、蔡淳、谢桦。 |
| 起草机构 | 上海市疾病预防控制中心、上海市预防医学研究院、上海市卫生和计划生育委员会、上海市卫生和计划生育委员会信息中心 |
| 规定内容 | WS 375本部分规定了脑卒中登记报告基本数据集的数据集元数据属性和数据元属性。 |
| 适用范围 | 本部分适用于疾病预防控制机构、提供相关服务的医疗机构及卫生行政部门进行相关业务数据采集、传输、存储等工作。 |

[0119]

| | |
|---|---|
| 疾病 | 脑卒中（ICD10/163.9） |
| 科室 | 神经内科 |
| 引用标准 | GB/T 2260中华人民共和国行政区划代码<br>GB/T 2261.1个人基本信息分类与代码 第1部分：人的性别代码<br>GB/T 2261.2个人基本信息分类与代码 第2部分：婚姻状况代码GB/T 3304 中国各民族名称的罗马字母拼写法和代码GB/T 4658 学历代码GB/T 6565职业分类与代码<br>WS 364.3 卫生信息数据元值域代码 第3部分：人口学及社会经济学特征<br>WS 364.5卫生信息数据元值域代码 第5部分：健康危险因素<br>WS 364.11卫生信息数据元值域代码 第11部分：医学评估WS 370 卫生信息基本数据集编制规范<br>ICD-10 国际疾病分类标准编码 |

[0120] 表5提取的结构化实体类型及属性——数据元实例

[0121]

| 外部标识符 | 内部标识符 | 数据元中文名称 | 数据元定义 | 数据元标识格式 | 数据元数据类型 | 数据元允许值 |
|---|---|---|---|---|---|---|
| HDSD00.C6. | DE05.01.50 | 中医诊法 | 中医临床诊断方法在特定编码体系中的代码 | S3 | N1 | T/CIATCM003-2019 中医药信息数据元值域代码 05.01.501 中医诊 |

[0122]

| | | | | | | |
|---|---|---|---|---|---|---|
| 003 | 1.00 | 代码 | | | | 法代码表 |
| HDSD00.C6.004 | DE05.01.502.00 | 中医望诊代码 | 中医临床医生用视觉观察病人的神、色、形、态、舌象、排泄物、小儿指纹等的异常变化，以了解病情的诊断内容在特定编码体系中的代码 | S3 | N1 | T/CIATCM003-2019 中医药信息数据元值域代码 中医望诊代码表 |

[0123] 表6结构化数据元及数据元概念

[0124]

| 标准号 | 内部标识符 | 数据元标识符 | 数据元名称 | CDISC数据元概念 | SNOMEDCT数据元概念 |
|---|---|---|---|---|---|
| T/GDPHA014—2021 | LV010101.001.04.020 | DE02.01.004.01 | 淋巴结转移 | \ | C96660 |
| T/GDPHA013—2021 | COPD010101.001.04.030 | DE02.01.004.11 | 用力肺活量 | Forcevitalcapacitymeasurement | C111361 |
| T/GDPHA013—2021 | COPD010101.001.04.031 | DE02.01.004.12 | 用力肺活量预计值 | FVCVolRespiratoryPredicted | |

[0125] (3)根据知识图谱构建的实体类型间关联关系,生成三元组,如表7所示、表8所示。

[0126] 表7实体类型三元组数据

[0127]

| SetID:ID | name | :LABEL |
|---|---|---|
| SetID1 | 慢性疾病临床研究项目信息通用数据元专用属性 | 数据元集合 |
| SetID2 | 慢性疾病临床研究受试者人口学基本信息子集数据元专用属性 | 数据元集合 |
| SetID3 | 慢性疾病临床研究受试者门（急）诊病历数据元专用属性 | 数据元集合 |

[0128]

| | | |
|---|---|---|
| SetID4 | 慢性疾病临床研究受试者检查信息数据元专用属性 | 数据元集合 |
| SetID5 | 慢性疾病临床研究受试者检验信息数据元专用属性 | 数据元集合 |

[0129] 表8实体类型标准引用关系数据

[0130]

| StID:START_ID | StID:END_ID | :TYPE |
|---|---|---|
| T/GDPHA002—2021 | GB/T2261.1—2003 | 引用 |

| | | |
|---|---|---|
| T/GDPHA002—2021 | GB/T2261.2—2003 | 引用 |
| T/GDPHA002—2021 | GB/T2261.4—2003 | 引用 |
| T/GDPHA002—2021 | GB/T3304—1991 | 引用 |

[0131] (4)进行知识图谱融合构建,利用规则和词典实现实体归并。

[0132] 合并如将CV03.00.107、WS364.5CV03.00.107饮食习惯代码表、WS364.5卫生信息数据元值域代码第5部分,统一归并为WS364.5CV03.00.107饮食习惯代码表。

[0133] 卫生部统计信息中心、中华人民共和国卫生部统计信息中心、卫生部卫生统计信息中心归并为中华人民共和国卫生部统计信息中心。

[0134] (5)数据存储和质量检查。

[0135] 将所有三元组数据导入neo4j之后,进行数据抽查,核对三元组数据的正确性,保证实体类型和关联关系的正确性。

[0136] 最后建立的部分知识图谱数据实例如图3所示。

[0137] 本发明实施例提供一种生物医学数据集标准数据元的知识图谱构建系统,包括以下模块:

[0138] 数据收集模块,收集不同类型的生物医学数据集数据元的相关标准文本和生物医学数据集相关标准的数据;

[0139] 数据分析模块,通过对收集数据元的相关标准文本和生物医学数据集相关标准的数据进行分析和归纳,用于支持构建生物医学数据集标准数据元知识图谱的知识模型和进行数据的解析和细粒度内容抽取;

[0140] 知识模型构建模块,构建生物医学数据集标准数据元知识图谱的知识模型,定义实体类型并同时建立各实体类的属性和实体类型之间的语义关联关系类型;

[0141] 实体类型抽取模块,从结构化数据和结构化数据中的非结构化文本抽取实体类型数据及属性数据;

[0142] 知识图谱获取模块,根据建立的实体类型之间的语义关联关系类型,进行多类数据的知识融合,得到生物医学数据集标准数据元知识图谱。

[0143] 本说明书中各个实施例采用递进的方式描述,每个实施例重点说明的都是与其他实施例的不同之处,各个实施例之间相同相似部分互相参见即可。对于实施例公开的装置而言,由于其与实施例公开的方法相对应,所以描述的比较简单,相关之处参见方法部分说明即可。

[0144] 对所公开的实施例的上述说明,使本领域专业技术人员能够实现或使用本发明。对这些实施例的多种修改对本领域的专业技术人员来说将是显而易见的,本文中所定义的一般原理可以在不脱离本发明的精神或范围的情况下,在其它实施例中实现。因此,本发明将不会被限制于本文所示的这些实施例,而是要符合与本文所公开的原理和新颖特点相一致的最宽的范围。

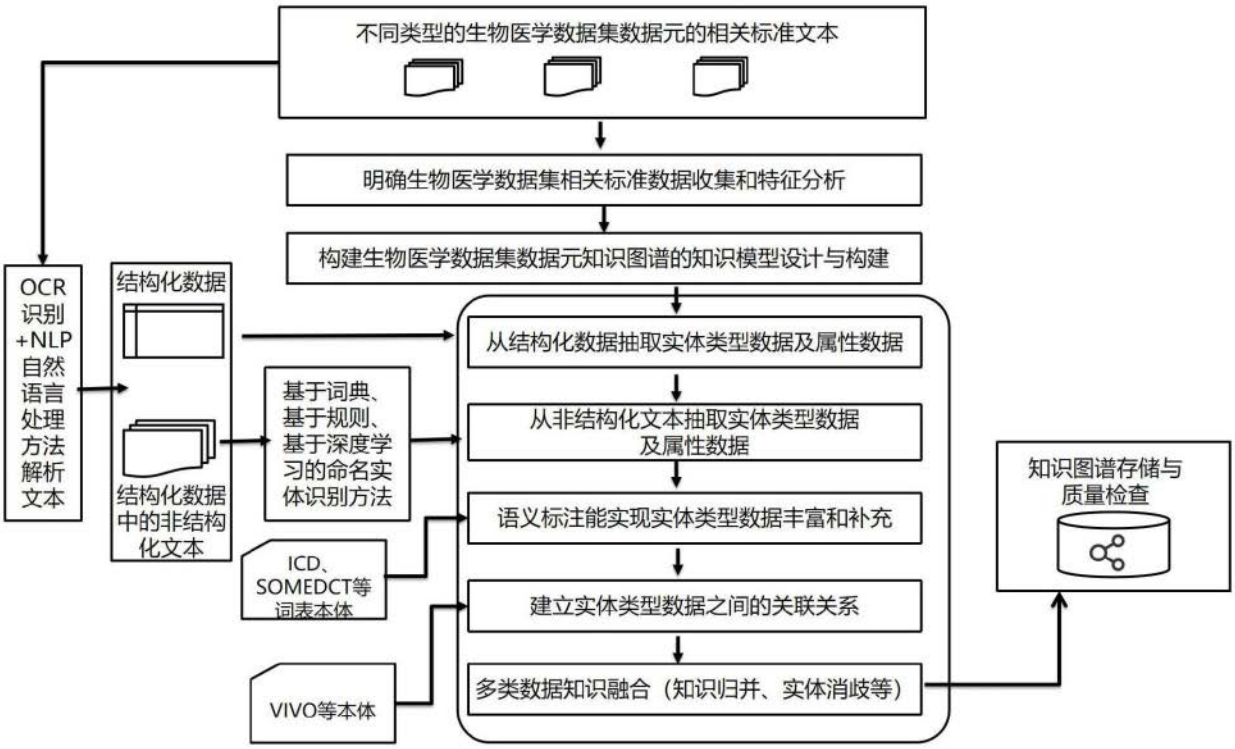

不同类型的生物医学数据集数据元的相关标准文本
明确生物医学数据集相关标准数据收集和特征分析
构建生物医学数据集数据元知识图谱的知识模型设计与构建
OCR识别+NLP自然语言处理方法解析文本
结构化数据
结构化数据中的非结构化文本
基于词典、基于规则、基于深度学习的命名实体识别方法
ICD、SOMEDCT等词表本体
VIVO等本体
从结构化数据抽取实体类型数据及属性数据
从非结构化文本抽取实体类型数据及属性数据
语义标注能实现实体类型数据丰富和补充
建立实体类型数据之间的关联关系
多类数据知识融合（知识归并、实体消歧等）
知识图谱存储与质量检查

图1

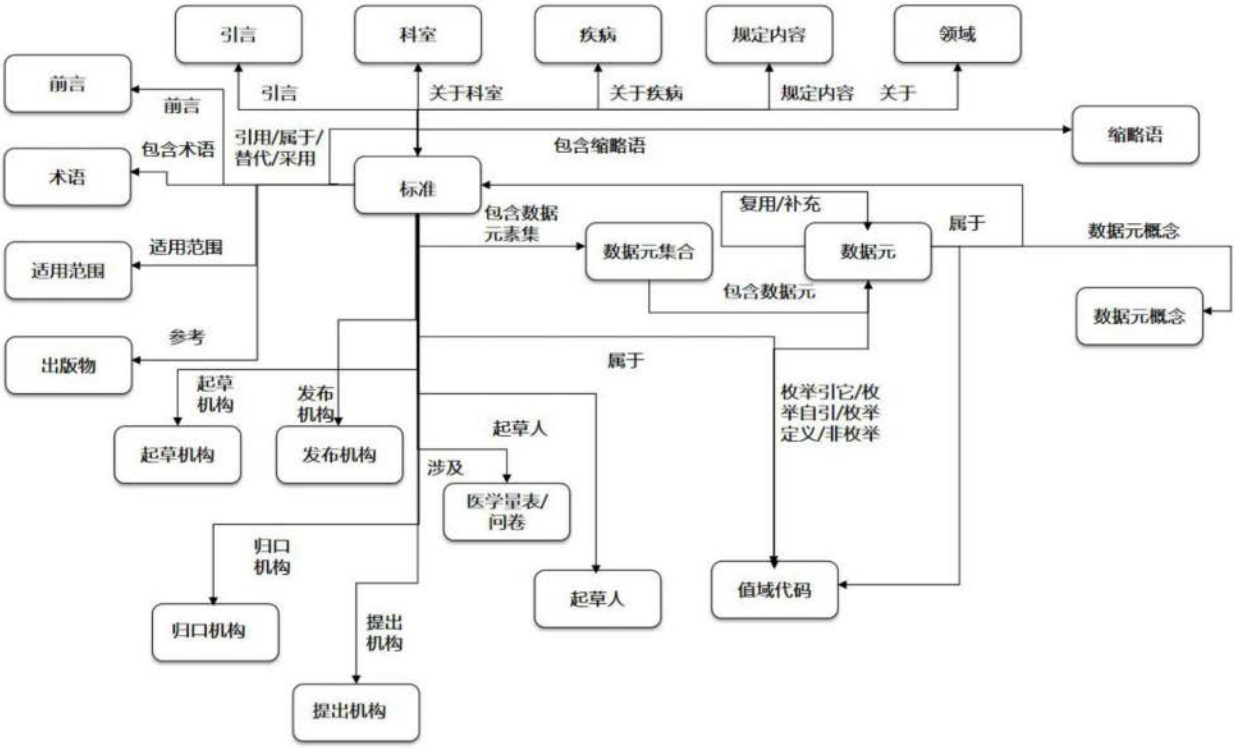

引言
科室
疾病
规定内容
领域
前言
术语
适用范围
出版物
缩略语
标准
数据元集合
数据元
数据元概念
起草机构
发布机构
医学量表/问卷
归口机构
提出机构
起草人
值域代码
引言
关于科室
关于疾病
规定内容
关于
前言
引用/属于/替代/采用
包含术语
包含缩略语
适用范围
参考
包含数据元素集
复用/补充
属于
包含数据元
起草机构
发布机构
起草人
涉及
归口机构
提出机构
枚举引它/枚举自引/枚举定义/非枚举

图2

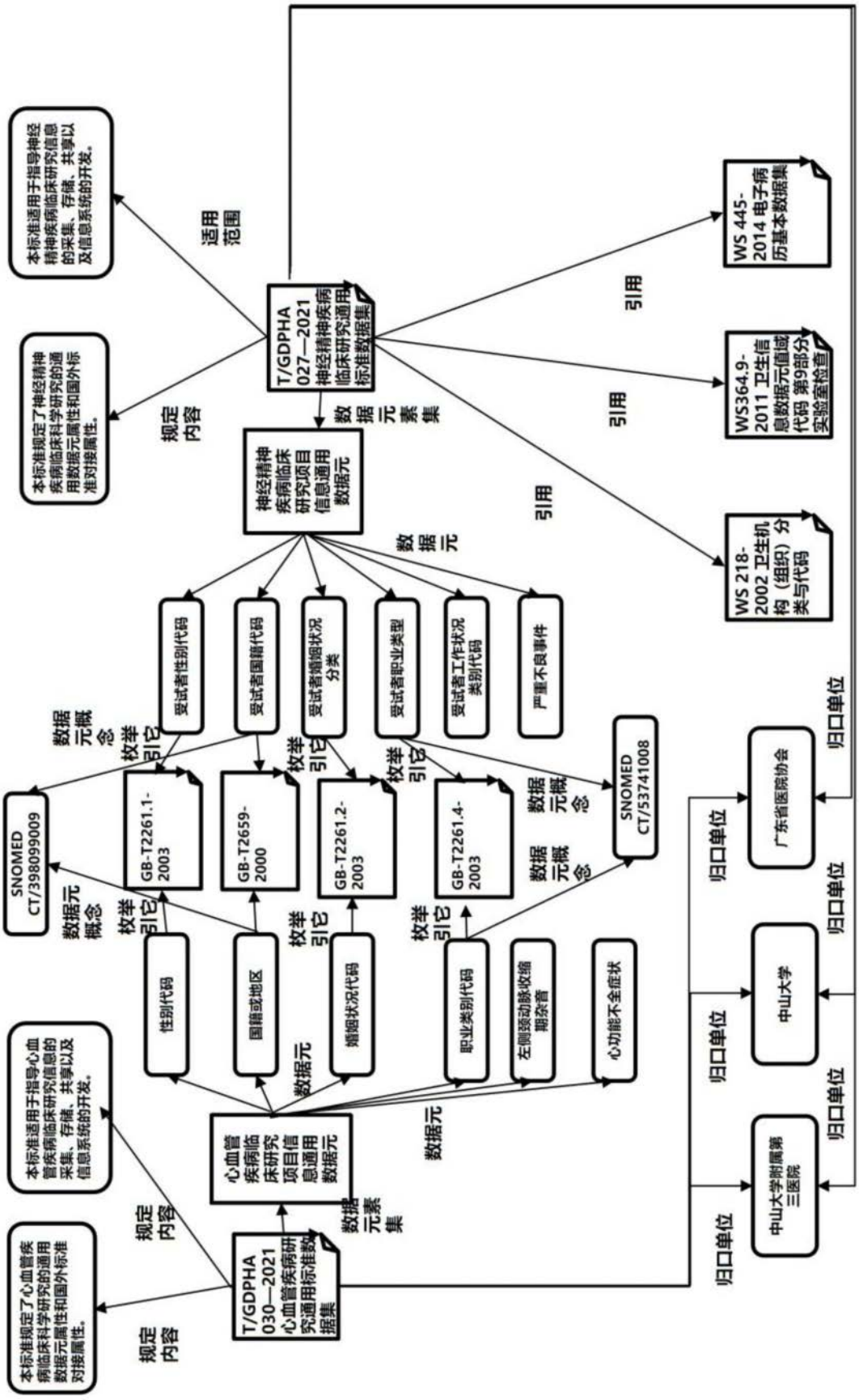

本标准适用于指导神经精神疾病临床研究信息的采集、存储、共享以及信息系统的开发。
适用范围
本标准规定了神经精神疾病临床科学研究的通用数据元属性和国外标准对接属性。
规定内容
T/GDPHA 027—2021 神经精神疾病临床研究通用标准数据集
数据元素集
神经精神疾病临床研究项目信息通用数据元
数据元
引用
WS 445-2014 电子病历基本数据集
WS364.9-2011 卫生信息数据元值域代码 第9部分 实验室检查
WS 218-2002 卫生机构（组织）分类与代码
受试者性别代码
受试者国籍代码
受试者婚姻状况分类
受试者职业类型
受试者工作状况类别代码
严重不良事件
数据元概念
枚举
引它
GB-T2261.1-2003
GB-T2659-2000
GB-T2261.2-2003
GB-T2261.4-2003
SNOMED CT/398099009
SNOMED CT/53741008
性别代码
国籍或地区
婚姻状况代码
职业类别代码
左侧颈动脉收缩期杂音
心功能不全症状
本标准适用于指导心血管疾病临床研究信息的采集、存储、共享以及信息系统的开发。
本标准规定了心血管疾病临床科学研究的通用数据元属性和国外标准对接属性。
T/GDPHA 030—2021 心血管疾病研究通用标准数据集
心血管疾病临床研究项目信息通用数据元
归口单位
广东省医院协会
中山大学
中山大学附属第三医院

图3

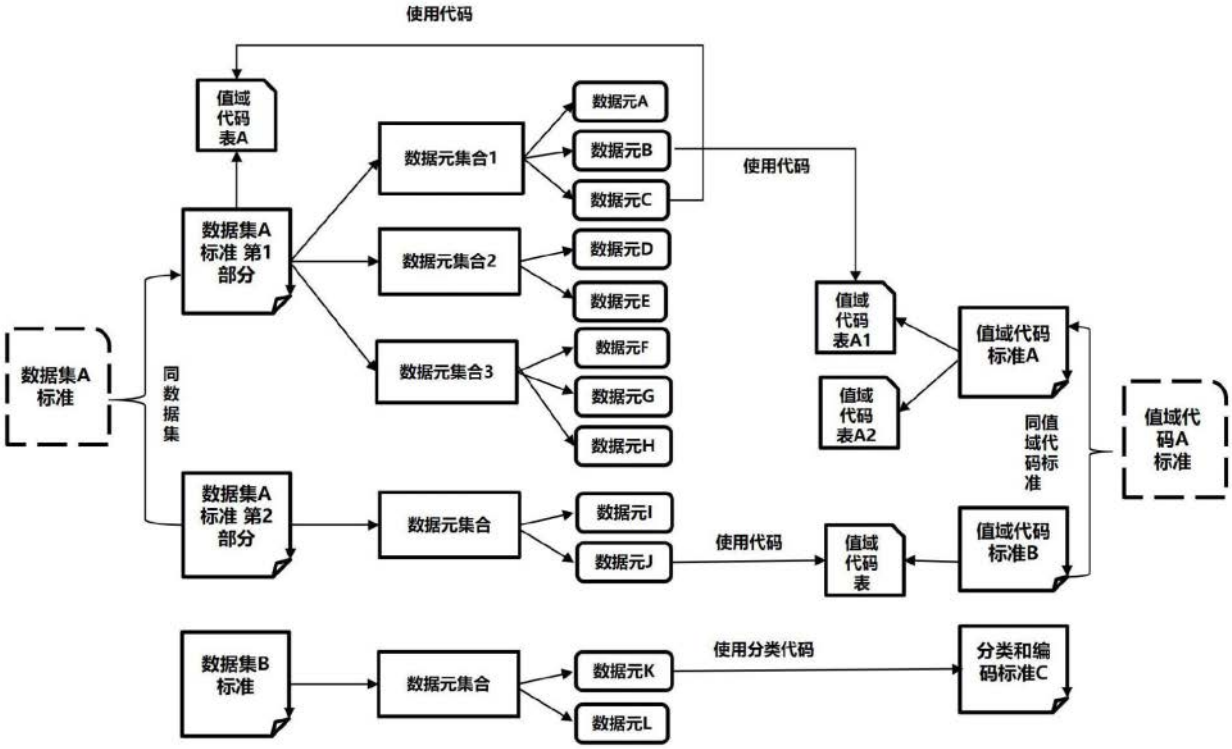

使用代码
值域代码表A
数据集A标准 第1部分
数据元集合1
数据元集合2
数据元集合3
数据元A
数据元B
数据元C
数据元D
数据元E
数据元F
数据元G
数据元H
使用代码
值域代码表A1
值域代码表A2
值域代码标准A
同值域代码标准
值域代码A标准
数据集A标准
同数据集
数据集A标准 第2部分
数据元集合
数据元I
数据元J
使用代码
值域代码表
值域代码标准B
数据集B标准
数据元集合
数据元K
数据元L
使用分类代码
分类和编码标准C

图4