(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 115391315 A
(43)申请公布日 2022.11.25

(21)申请号 202210837118.4

(22)申请日 2022.07.15

(71)申请人 生命奇点(北京)科技有限公司
地址 100080 北京市海淀区苏州街3号6层603

(72)发明人 安西平 徐辉

(74)专利代理机构 北京工信联合知识产权代理有限公司 11266
专利代理师 王舰

(51)Int.Cl.
G06F 16/215(2019.01)
G06F 16/2453(2019.01)

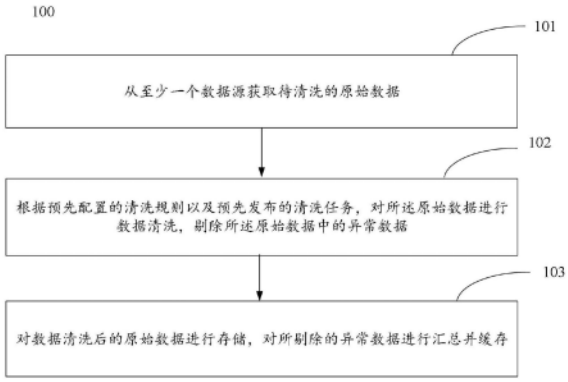
权利要求书2页 说明书10页 附图3页

(54)发明名称

一种数据清洗方法及装置

(57)摘要

本发明公开了一种数据清洗方法及装置。其中,方法包括:从至少一个数据源获取待清洗的原始数据;根据预先配置的清洗规则以及预先发布的清洗任务,对所述原始数据进行数据清洗,剔除所述原始数据中的异常数据;对数据清洗后的原始数据进行存储,对所剔除的异常数据进行汇总并缓存。



CN 115391315 A

1.一种数据清洗方法,其特征在于,包括:

从至少一个数据源获取待清洗的原始数据;

根据预先配置的清洗规则以及预先发布的清洗任务,对所述原始数据进行数据清洗,剔除所述原始数据中的异常数据;

对数据清洗后的原始数据进行存储,对所剔除的异常数据进行汇总并缓存。

2.根据权利要求1所述的方法,其特征在于,对所述原始数据进行数据清洗之前,还包括通过以下步骤配置所述清洗规则:

配置所述清洗规则中的字段清洗规则;

配置所述清洗规则中的正则表达式清洗规则;

配置所述清洗规则中的复杂逻辑清洗规则。

3.根据权利要求2所述的方法,其特征在于,配置所述清洗规则之前,还包括:推送预先设置的候选清洗规则,并且

配置所述清洗规则,包括:

根据推送的候选清洗规则,配置所述清洗规则中的字段清洗规则、正则表达式清洗规则以及复杂逻辑清洗规则。

4.根据权利要求1所述的方法,其特征在于,对所述原始数据进行数据清洗之前,还包括发布所述清洗任务,其中所述清洗任务包括提取不符合要求的数据、提取出的数据是否直接过滤掉以及是否由业务单位修正之后再进行数据的抽取。

5.根据权利要求1所述的方法,其特征在于,对所述原始数据进行数据清洗,包括以下至少一个操作步骤:

对所述原始数据进行一致性检查;

对所述原始数据不符合目标类型的进行类型转换和有效值提取;

对所述原始数据中的无效值和缺失值进行处理;

对所述原始数据中的残缺数据进行过滤;

对所述原始数据中的错误数据进行处理,其中所述错误数据为非标准格式的数据;

对所述原始数据中的重复数据进行提取。

6.根据权利要求5所述的方法,其特征在于,对所述原始数据进行一致性检查,包括:

根据每个变量的正常取值范围和相互关系,检查所述原始数据是否合乎要求,提取出超出正常取值范围、逻辑上不合理或者相互矛盾的数据。

7.根据权利要求1所述的方法,其特征在于,从至少一个数据源获取待清洗的原始数据之前,还包括:对数据源进行配置。

8.一种数据清洗装置,其特征在于,包括:

数据获取模块,用于从至少一个数据源获取待清洗的原始数据;

数据清洗模块,用于根据预先配置的清洗规则以及预先发布的清洗任务,对所述原始数据进行数据清洗,剔除所述原始数据中的异常数据;

数据存储模块,用于对数据清洗后的原始数据进行存储,对所剔除的异常数据进行汇总并缓存。

9.一种计算机可读存储介质,其特征在于,所述存储介质存储有计算机程序,所述计算机程序用于执行上述权利要求1-7任一所述的方法。

10.一种电子设备,其特征在于,所述电子设备包括:

处理器;

用于存储所述处理器可执行指令的存储器;

所述处理器,用于从所述存储器中读取所述可执行指令,并执行所述指令以实现上述权利要求1-7任一所述的方法。

一种数据清洗方法及装置

技术领域

[0001] 本发明涉及数据处理技术领域，并且更具体地，涉及一种数据清洗方法及装置。

背景技术

[0002] 数据网关，是指一类设备实现多个业务数据系统互连，实现各个系统之间数据的集成、共享和管理。智能医疗健康数据网关(Intelligent Clinical&Health Data GateWay，以下简称IDCHGW)是指一个设备，提供面向医疗数据的数据治理、再生产和数据服务，提供可视化、智能化、可交互、可编程的操作方式，以及提供安全的数据交换能力。

[0003] 数据清洗是整个智能医疗健康数据网关建设过程中不可缺少的一个环节，其结果质量直接关系到后续所有相关研究的模型效果和最终结论。目前，市面上的数据清洗软件大都是针对自家的业务系统做简单的清洗，对于其他的不同数据来源的业务数据系统，由于数据不一致，因而无法对异构异源海量离散的数据进行有效的数据清洗。

[0004] 另一方面，医院的临床数据有其特殊的特征以及要求。主要涉及临床数据中体征、检查、检验指标的单位、量纲和有效值；用药、麻醉、手术等各种医嘱的剂量、日期等；以及由于信息系统用户使用造成的错误；系统升级造成的数据断裂等。

[0005] 因此，如何提供一种有效的方案，以便对异构异源海量离散的数据进行有效的数据清洗，以及针对医疗数据量身定做的清洗系统，已成为现有技术中一亟待解决的难题。

发明内容

[0006] 针对现有技术的不足，本发明提供一种数据清洗方法及装置。

[0007] 根据本发明的一个方面，提供了一种数据清洗方法，包括：

[0008] 从至少一个数据源获取待清洗的原始数据；

[0009] 根据预先配置的清洗规则以及预先发布的清洗任务，对所述原始数据进行数据清洗，剔除所述原始数据中的异常数据；

[0010] 对数据清洗后的原始数据进行存储，对所剔除的异常数据进行汇总并缓存。

[0011] 可选地，对所述原始数据进行数据清洗之前，该方法还包括通过以下步骤配置所述清洗规则：

[0012] 配置所述清洗规则中的字段清洗规则；

[0013] 配置所述清洗规则中的正则表达式清洗规则；

[0014] 配置所述清洗规则中的复杂逻辑清洗规则。

[0015] 可选地，配置所述清洗规则之前，该方法还包括：推送预先设置的候选清洗规则，并且

[0016] 配置所述清洗规则，包括：

[0017] 根据推送的候选清洗规则，配置所述清洗规则中的字段清洗规则、正则表达式清洗规则以及复杂逻辑清洗规则。

[0018] 可选地，对所述原始数据进行数据清洗之前，该方法还包括发布所述清洗任务，其

中所述清洗任务包括提取不符合要求的数据、提取出的数据是否直接过滤掉以及是否由业务单位修正之后再进行数据的抽取。

[0019] 可选地，对所述原始数据进行数据清洗，包括以下至少一个操作步骤：

[0020] 对所述原始数据进行一致性检查；

[0021] 对所述原始数据不符合目标类型的进行类型转换和有效值提取；

[0022] 对所述原始数据中的无效值和缺失值进行处理；

[0023] 对所述原始数据中的残缺数据进行过滤；

[0024] 对所述原始数据中的错误数据进行处理，其中所述错误数据为非标准格式的数据；

[0025] 对所述原始数据中的重复数据进行提取。

[0026] 可选地，对所述原始数据进行一致性检查，包括：

[0027] 根据每个变量的正常取值范围和相互关系，检查所述原始数据是否合乎要求，提取出超出正常取值范围、逻辑上不合理或者相互矛盾的数据。

[0028] 可选地，从至少一个数据源获取待清洗的原始数据之前，该方法还包括：对数据源进行配置。

[0029] 根据本发明的另一个方面，提供了一种数据清洗装置，包括：

[0030] 数据获取模块，用于从至少一个数据源获取待清洗的原始数据；

[0031] 数据清洗模块，用于根据预先配置的清洗规则以及预先发布的清洗任务，对所述原始数据进行数据清洗，剔除所述原始数据中的异常数据；

[0032] 数据存储模块，用于对数据清洗后的原始数据进行存储，对所剔除的异常数据进行汇总并缓存。

[0033] 根据本发明的又一个方面，提供了一种计算机可读存储介质，所述存储介质存储有计算机程序，所述计算机程序用于执行本发明上述任一方面所述的方法。

[0034] 根据本发明的又一个方面，提供了一种电子设备，所述电子设备包括：处理器；用于存储所述处理器可执行指令的存储器；所述处理器，用于从所述存储器中读取所述可执行指令，并执行所述指令以实现本发明上述任一方面所述的方法。

[0035] 本发明首先从至少一个数据源获取待清洗的原始数据，然后根据预先配置的清洗规则以及预先发布的清洗任务，对所述原始数据进行数据清洗，剔除所述原始数据中的异常数据，最后对数据清洗后的原始数据进行存储，对所剔除的异常数据进行汇总并缓存。本发明通过规范清洗规则和清洗任务，对异构异源海量离散的原始数据进行数据清洗，生成易于分析利用的、可共享的数据。通过对数据清洗后的原始数据进行存储，对所剔除的异常数据进行汇总并缓存，以促使业务单位尽快的修正错误，同时也可以作为将来验证数据的依据。

附图说明

[0036] 通过参考下面的附图，可以更为完整地理解本发明的示例性实施方式：

[0037] 图1是本发明一示例性实施例提供的数据清洗方法的流程示意图；

[0038] 图2是本发明一示例性实施例提供的应用有数据清洗方法的数据清洗服务系统的框架图；

[0039] 图3是本发明一示例性实施例提供的将病历中的原始术语映射成统一的标准术语的示意图；

[0040] 图4是本发明一示例性实施例提供的对病历中的原始术语进行结构化和标准化处理的示意图；

[0041] 图5是本发明一示例性实施例提供的数据清洗装置的结构示意图。

具体实施方式

[0042] 下面，将参考附图详细地描述根据本发明的示例实施例。显然，所描述的实施例仅仅是本发明的一部分实施例，而不是本发明的全部实施例，应理解，本发明不受这里描述的示例实施例的限制。

[0043] 应注意到：除非另外具体说明，否则在这些实施例中阐述的部件和步骤的相对布置、数字表达式和数值不限制本发明的范围。

[0044] 本领域技术人员可以理解，本发明实施例中的“第一”、“第二”等术语仅用于区别不同步骤、设备或模块等，既不代表任何特定技术含义，也不表示它们之间的必然逻辑顺序。

[0045] 还应理解，在本发明实施例中，“多个”可以指两个或两个以上，“至少一个”可以指一个、两个或两个以上。

[0046] 还应理解，对于本发明实施例中提及的任一部件、数据或结构，在没有明确限定或者在前后文给出相反启示的情况下，一般可以理解为一个或多个。

[0047] 另外，本发明中术语“和/或”，仅仅是一种描述关联对象的关联关系，表示可以存在三种关系，例如，A和/或B，可以表示：单独存在A，同时存在A和B，单独存在B这三种情况。另外，本发明中字符“/”，一般表示前后关联对象是一种“或”的关系。

[0048] 还应理解，本发明对各个实施例的描述着重强调各个实施例之间的不同之处，其相同或相似之处可以相互参考，为了简洁，不再一一赘述。

[0049] 同时，应当明白，为了便于描述，附图中所示出的各个部分的尺寸并不是按照实际的比例关系绘制的。

[0050] 以下对至少一个示例性实施例的描述实际上仅仅是说明性的，决不作为对本发明及其应用或使用的任何限制。

[0051] 对于相关领域普通技术人员已知的技术、方法和设备可能不作详细讨论，但在适当情况下，技术、方法和设备应当被视为说明书的一部分。

[0052] 应注意到：相似的标号和字母在下面的附图中表示类似项，因此，一旦某一项在一个附图中被定义，则在随后的附图中不需要对其进行进一步讨论。

[0053] 本发明实施例可以应用于终端设备、计算机系统、服务器等电子设备，其可与众多其它通用或专用计算系统环境或配置一起操作。适于与终端设备、计算机系统、服务器等电子设备一起使用的众所周知的终端设备、计算系统、环境和/或配置的例子包括但不限于：个人计算机系统、服务器计算机系统、瘦客户机、厚客户机、手持或膝上设备、基于微处理器的系统、机顶盒、可编程消费电子产品、网络个人电脑、小型计算机系统、大型计算机系统和包括上述任何系统的分布式云计算技术环境，等等。

[0054] 终端设备、计算机系统、服务器等电子设备可以在由计算机系统执行的计算机系

统可执行指令(诸如程序模块)的一般语境下描述。通常,程序模块可以包括例程、程序、目标程序、组件、逻辑、数据结构等等,它们执行特定的任务或者实现特定的抽象数据类型。计算机系统/服务器可以在分布式云计算环境中实施,分布式云计算环境中,任务是由通过通信网络链接的远程处理设备执行的。在分布式云计算环境中,程序模块可以位于包括存储设备的本地或远程计算系统存储介质上。

[0055] 示例性方法

[0056] 图1是本发明一示例性实施例提供的数据清洗方法的流程示意图。本实施例可应用在电子设备上,例如但不限于应用在数据清洗服务系统上。

[0057] 如图1所示,数据清洗方法100包括以下步骤:

[0058] 步骤101,从至少一个数据源获取待清洗的原始数据。

[0059] 可选地,从至少一个数据源获取待清洗的原始数据之前,该方法还包括:对数据源进行配置。

[0060] 在本发明实施例中,待清洗的原始数据例如但不限于为医院数据。该医院数据可以从多个业务数据系统获取,每一个业务数据系统对应于一个数据源。因此,如图2所示,在获取待清洗的原始数据之前,业务单位可以通过数据源配置管理模块,根据实际需求对数据源进行配置,即确定对接哪几个数据源。完成配置之后,根据配置信息,从对应的数据源中采集待清洗的原始数据,以便后续进行清洗加工处理,并做标准化整理。

[0061] 在本发明实施例中,基于知识图谱对原始数据的术语进行标准化处理。知识图谱的schema的内容主要由实体类型定义和以及实体间的关系类型定义构成。主要包含疾病、手术、药品、检查、检验、人体组织器官等多类医学概念及关系。其中,知识图谱的实体类型例如但不限于包括下1表所示的内容,实体间的关系类型例如但不限于包括下2表所示的内容。

[0062] 表1

[0063]

事物	物质	病理过程	关系运算符
临床所见	生物	精度类型	用药目的
疾病	组织机构	给药途径	治疗
操作	观察对象	评价结果	药物治疗方案
观测操作	限定词	量纲类型	操作治疗方案
药品	严重程度	频次	人
事件	入路	身体状态	人群
人体形态与结构	剂型	传播途径	地点
异常形态结构	单位	优先级别	文献资料
基因	发病和发病过程	意图	同义词
基因突变	时间范围	分期状态	化疗方案
标本	技术	体位	免疫治疗方案
物理实体	操作方法	麻醉方式	亲属
医疗器械	方位	饮食类型	
物理能量	生命周期	药物处方分类	

[0064] 表2

[0065]

子类	有效成分	入路	症状
高发地区	规格成分	使用的能量	常见症状
传染源	相加作用	目标物质	少见症状
传播途径	协同作用	辅助性物质	体征
相关人群	拮抗作用	操作前检查	常见体征
易感人群	配伍禁忌	麻醉方式	少见体征
高危人群	化学配伍禁忌	麻醉用药	常伴发
适用人群		操作后饮食类型	发展为
禁忌人群	物理配伍禁忌		分型
高峰期		观察目标	危险因素
发生部位	剂量单位	上限运算符	一线治疗
形态学改变	用药频次	下限运算符	预防用药
临床过程	用药时间	危急值运算符	治疗相关检查
严重程度	用药间隔	受检标本	体格检查
遗传基因	用药疗程	受检成分	实验室检查

[0066]

致病原因	用药目的	精度类型	辅助检查
临床表现	剂型	受检时长	随访复查
后发于	分子规格单位	指标	病原学检查
相关检查	分母规格单位	使用技术	血清学检查
评价对象	生产厂家	单位	病情监测
诊疗依据	药品分类标签	结果提示	护理操作
诊断标准	适应证	方位	护理方案
鉴别诊断	禁忌证	组成部分	后遗症
治疗目标	不良反应	并发症	禁忌药物
诊疗操作	常见不良反应	治疗方式	同义词
治疗器械	少见不良反应	治疗前检查	治疗方案
治疗药物	过敏反应	治疗前必需检查	免疫治疗
一线用药	给药途径	治疗前备选检查	抗病毒治疗
二线用药	优先级别	治疗后复查	一般治疗
三线用药	操作重心	治疗后必需检查	康复治疗
预防措施	操作意图	治疗后备选检查	化疗方案
一级预防	操作分期	分支	免疫治疗方案
二级预防	最佳操作时间	诊断相关检查	包含关系_知识树
科室	操作方法	术中操作	知识库药品层级关系
发生于某期间	操作部位	术中用药	一级亲属
评价结果	操作形态学改变	术后用药	二级亲属
病理过程	操作条件	术前用药	三级亲属
药物成分	使用的器械	选择用药	其他亲属
	使用的通路器械	用药相关检查	
	植介入器械		

[0067] 进一步地，知识图谱的应用接口如下表3所示，术语标准化主要使用其中的“术语标准化”接口。

[0068] 表3

[0069]

接口名称	功能说明	备注
路径推断	即路径查询，查询满足预定条件的路径，返回结果中包含路径中的所有节点已经节点之间的关	

[0070]

接口名称	功能说明	备注
	系。	
节点推断	即节点查询，根据设定的条件在图谱中搜索需要的节点。	
自定义查询	支持用户根据数据存储结构设计图谱查询语句，接口负责执行用户的查询语句并返回查询结果。	仅支持查询功能
术语类别体系查询	获取一个在知识图谱中所属的上下层类别。	
术语标准化	根据图谱的 schema 调整输入术语的描述形式，将同类术语以统一的格式表示。	
实体链接	以图谱中的节点名称为标准词，将输入的术语映射到对应的标准词上。	
schema 查询	获取知识图谱的简要 schema 说明，仅包含数据存储结构和简要的属性说明。	

[0071] 术语标准化接口说明如下：

[0072] 1、任务：将病历中的原始术语映射成统一的标准术语。

[0073] 术语标准化结果可以认为是树状结构。根节点是原始术语，叶子节点是最细粒度的实体，每一级节点都有在图谱中对应的实体。

[0074] 如：原始术语：左室瘘->标准术语：左侧心室瘘。如图3所示

[0075] 2、方法，分成两步：

[0076] 1)结构化，即识别节点名称以及节点标签。通过命名实体识别算法，将原始术语分层级结构化成实体向量以及标签。这里的实体标签和知识图谱中的标签体系一致。得到树形表示。

[0077] 2)标准化：将树的叶子节点调用“术语标准化”接口，查到在图谱中的节点，得到对应节点的标准名称；然后向上合并，得到根节点的标准名称。例如图4所示。

[0078] 3、术语标准化算法构造，可以分为两个子任务，交互式迭代优化：

[0079] 第一步、使用词典及规则识别：

[0080] 1)识别节点名称及标签：通过事件生产结果、规则及参考资料等挖掘叶子节点标签词典，通过词典识别各节点名称及标签，并自底向上、从左向右构建标准化树；

[0081] 2)标准词映射表整理：通过词频统计、相似等方法整理叶子节点的标准词映射表，最终由医生确定标准词。

[0082] 第二步，使用第一步得到的标注数据做模块优化。

[0083] 1)节点名称及标签识别任务使用嵌套NER识别算法。

[0084] 2)准词映射任务可通过相似方法升级，但前期只考虑词典映射。

[0085] 步骤102，根据预先配置的清洗规则以及预先发布的清洗任务，对所述原始数据进行数据清洗，剔除所述原始数据中的异常数据。

[0086] 可选地，对所述原始数据进行数据清洗之前，该方法还包括通过以下步骤配置所述清洗规则：配置所述清洗规则中的字段清洗规则；配置所述清洗规则中的正则表达式清洗规则；配置所述清洗规则中的复杂逻辑清洗规则。

[0087] 在本发明实施例中，清洗规则配置主要包括配置字段清洗规则、正则表达式清洗规则以及复杂逻辑清洗规则。

[0088] 可选地，配置所述清洗规则之前，该方法还包括：推送预先设置的候选清洗规则，并且配置所述清洗规则，包括：根据推送的候选清洗规则，配置所述清洗规则中的字段清洗规则、正则表达式清洗规则以及复杂逻辑清洗规则。

[0089] 可选地，对所述原始数据进行数据清洗之前，该方法还包括发布所述清洗任务，其中所述清洗任务包括提取不符合要求的数据、提取出的数据是否直接过滤掉以及是否由业务单位修正之后再进行数据的抽取。

[0090] 在本发明实施例中，数据清洗的任务是过滤那些不符合要求的数据，确认是否过滤掉或者由业务单位修正之后再进行抽取。其中，不符合要求的数据主要是不完整的数据、错误的数据、重复的数据、与目标数据类型不一致的数据四大类。

[0091] 可选地，对所述原始数据进行数据清洗，包括以下至少一个操作步骤：对所述原始数据进行一致性检查；对所述原始数据中的无效值和缺失值进行处理；对所述原始数据中的残缺数据进行过滤；对所述原始数据中的错误数据进行处理，其中所述错误数据为非标准格式的数据；对所述原始数据中的重复数据进行提取。

[0092] 可选地，对所述原始数据进行一致性检查，包括：根据每个变量的正常取值范围和相互关系，检查所述原始数据是否合乎要求，提取出超出正常取值范围、逻辑上不合理或者相互矛盾的数据。

[0093] 在本发明实施例中，数据清洗要进行一致性检查。一致性检查是根据每个变量的合理取值范围和相互关系，检查数据是否合乎要求，发现超出正常范围、逻辑上不合理或者相互矛盾的数据。

[0094] 在本发明实施例中，数据清洗要进行对原始数据不符合目标类型的进行类型转换和有效值提取。原始数据不符合目标类型，目标数据类型根据使用场景确认，原始数据中有存在不符合的数据类型，清洗过程中需根据目标类型进行转换和提取有效值。

[0095] 在本发明实施例中，数据清洗要进行无效值和缺失值的处理。由于调查、编码和录入误差，数据中可能存在一些无效值和缺失值，需要给予适当的处理。常用的处理方法有：估算、整例删除、变量删除和成对删除。

[0096] 在本发明实施例中，数据清洗要进行残缺数据的过滤。残缺数据主要是一些应该有的信息缺失，对于这一类数据过滤出来，按缺失的内容向客户提交，要求在规定的时间内补全或者选择删除。补全后才写入数据仓库。

[0097] 在本发明实施例中，数据清洗要进行错误数据的修正。错误数据产生的原因是业务系统不够健全，在接收输入后没有进行判断直接写入后台数据库造成的，比如数值数据输成全角数字字符、字符串数据后面有一个回车操作、日期格式不正确、日期越界等。通过数据清洗把格式统一成标准格式。

[0098] 在本发明实施例中，数据清洗要进行重复数据的提取。对于这一类重复数据，特别是维表中会出现这种情况，需要将重复数据记录的所有字段导出来，让客户确认并整理。

[0099] 步骤103，对数据清洗后的原始数据进行存储，对所剔除的异常数据进行汇总并缓存。

[0100] 在本发明实施例中，数据清洗是一个反复的过程，不可能在短时间内完成，只有不断的发现问题，解决问题。对于是否过滤，是否修正一般要求客户确认，对于过滤掉的数据，写入Excel文件或者将过滤数据写入数据表并进行缓存，促使业务单位能够尽快地修正错误，同时也可以作为将来验证数据的依据。

[0101] 在本发明实施例中，数据清洗包括对数据的完整性、一致性、合法性、正确性等的清洗，并且按照一定规则转化成统一标准。比如：当数据包含不同量纲的多种变量时，数值间的差别可能很大。归一化将数据按比例缩放，使之落入一个小的特定区间。去除数据的单位限制，将其转化为无量纲的纯数值，便于不同单位或量级的指标能够进行比较和加权。经过清洗后的数据，才可以用于后续的统计分析。

[0102] 在本发明实施例中，数据清洗对采集汇聚的数据进行清洗加工处理，并做标准化整理。主要包括制定数据清洗流程、清洗流程控制、清洗质量控制、清洗过程管理等。通过规范流程和规则库，基于流程引擎构建统一的、可配置的数据转换、清洗、比对、关联、融合等加工处理过程，对异构异源海量离散的数据资源加工生产，生成易于分析利用的、可共享的数据。

[0103] 从而，本发明首先从至少一个数据源获取待清洗的原始数据，然后根据预先配置的清洗规则以及预先发布的清洗任务，对所述原始数据进行数据清洗，剔除所述原始数据中的异常数据，最后对数据清洗后的原始数据进行存储，对所剔除的异常数据进行汇总并缓存。本发明通过规范清洗规则和清洗任务，对异构异源海量离散的原始数据进行数据清洗，生成易于分析利用的、可共享的数据。通过对数据清洗后的原始数据进行存储，对所剔除的异常数据进行汇总并缓存，以促使业务单位尽快的修正错误，同时也可以作为将来验证数据的依据。

[0104] 示例性装置

[0105] 图5是本发明一示例性实施例提供的数据清洗装置的结构示意图。如图5所示，数据清洗装置500包括：

[0106] 数据获取模块510，用于从至少一个数据源获取待清洗的原始数据；

[0107] 数据清洗模块520，用于根据预先配置的清洗规则以及预先发布的清洗任务，对所述原始数据进行数据清洗，剔除所述原始数据中的异常数据；

[0108] 数据存储模块530，用于对数据清洗后的原始数据进行存储，对所剔除的异常数据进行汇总并缓存。

[0109] 可选地，数据清洗装置500还包括清洗规则配置管理模块，用于通过以下步骤配置所述清洗规则：

[0110] 配置所述清洗规则中的字段清洗规则；

[0111] 配置所述清洗规则中的正则表达式清洗规则；

[0112] 配置所述清洗规则中的复杂逻辑清洗规则。

[0113] 可选地，数据清洗装置500还包括清洗规则推荐模块，用于推送预先设置的候选清

洗规则，并且

[0114] 清洗规则配置管理模块，具体用于：根据推送的候选清洗规则，配置所述清洗规则中的字段清洗规则、正则表达式清洗规则以及复杂逻辑清洗规则。

[0115] 可选地，数据清洗装置500还包括清洗任务配置管理模块，用于发布所述清洗任务，其中所述清洗任务包括提取不符合要求的数据、提取出的数据是否直接过滤掉以及是否由业务单位修正之后再进行数据的抽取。

[0116] 可选地，数据清洗模块520，具体用于执行以下至少一个操作步骤：

[0117] 对所述原始数据进行一致性检查；

[0118] 对所述原始数据不符合目标类型的进行类型转换和有效值提取；

[0119] 对所述原始数据中的无效值和缺失值进行处理；

[0120] 对所述原始数据中的残缺数据进行过滤；

[0121] 对所述原始数据中的错误数据进行处理，其中所述错误数据为非标准格式的数据；

[0122] 对所述原始数据中的重复数据进行提取。

[0123] 可选地，数据清洗模块520，具体用于：

[0124] 根据每个变量的正常取值范围和相互关系，检查所述原始数据是否合乎要求，提取出超出正常取值范围、逻辑上不合理或者相互矛盾的数据。

[0125] 可选地，数据清洗装置500还包括数据源配置管理模块，用于对数据源进行配置。

[0126] 本发明的实施例的数据清洗装置500与本发明的另一个实施例的数据清洗方法100相对应，在此不再赘述。

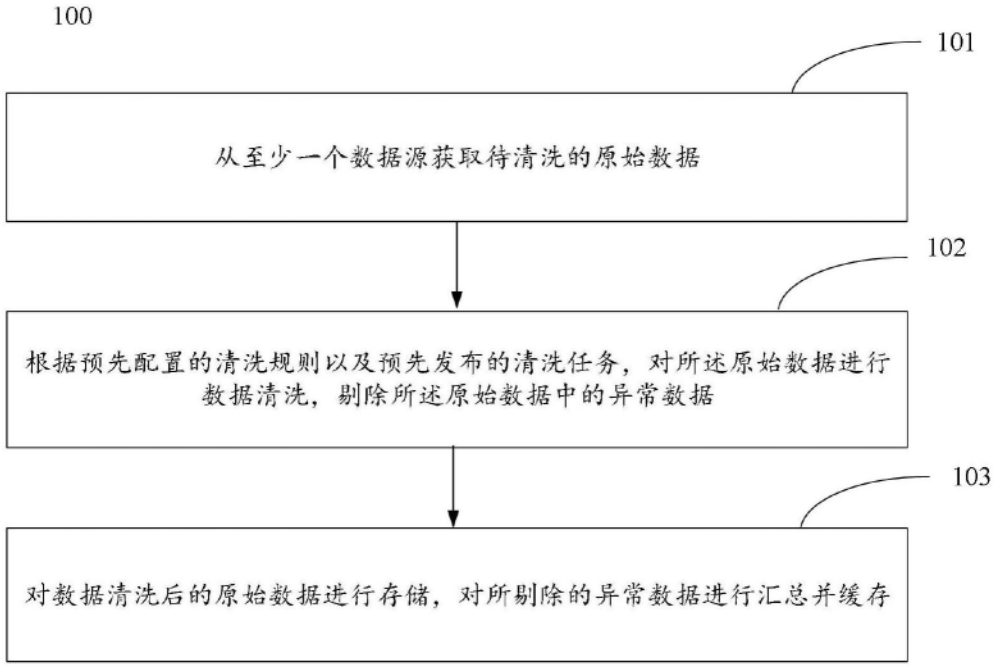

100
101
从至少一个数据源获取待清洗的原始数据
102
根据预先配置的清洗规则以及预先发布的清洗任务，对所述原始数据进行数据清洗，剔除所述原始数据中的异常数据
103
对数据清洗后的原始数据进行存储，对所剔除的异常数据进行汇总并缓存

图1

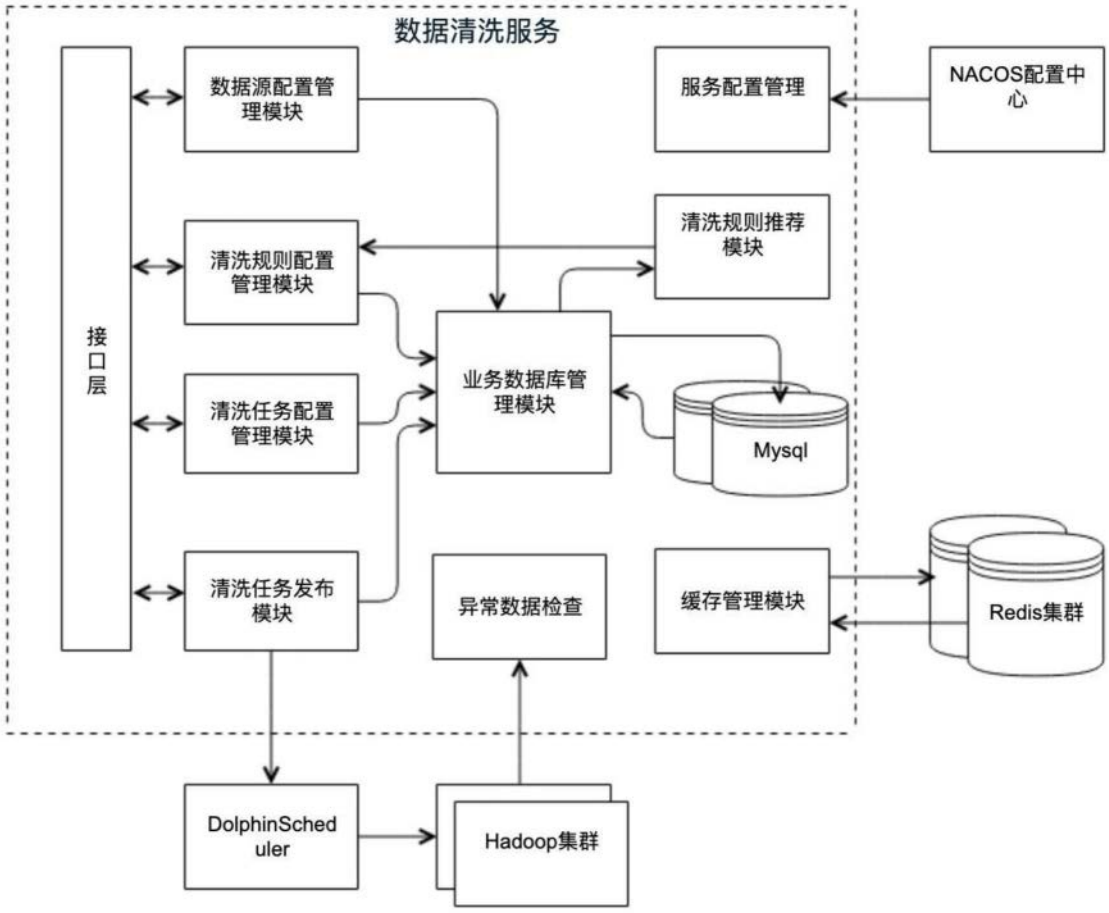

数据清洗服务
接口层
数据源配置管理模块
服务配置管理
NACOS配置中心
清洗规则配置管理模块
清洗规则推荐模块
业务数据库管理模块
Mysql
清洗任务配置管理模块
清洗任务发布模块
异常数据检查
缓存管理模块
Redis集群
DolphinScheduler
Hadoop集群

图2

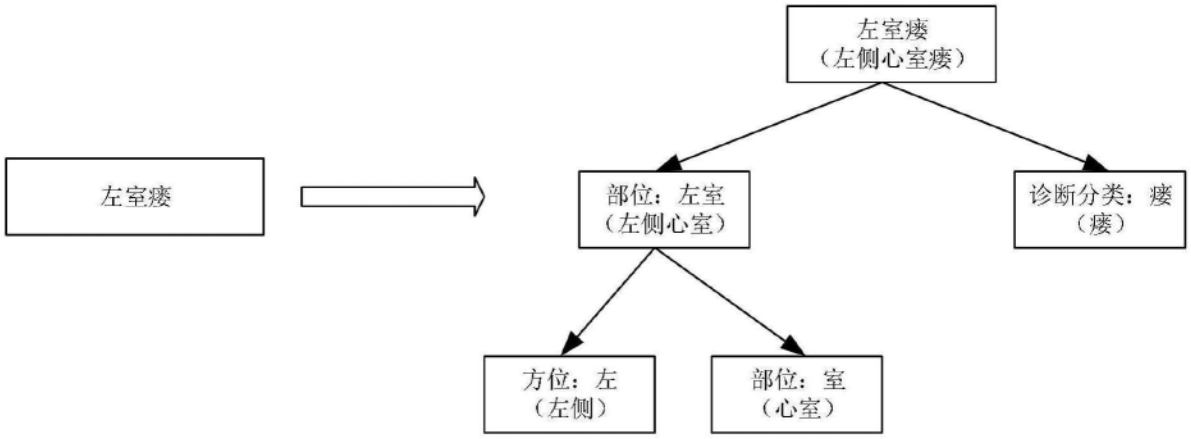

左室瘘
左室瘘
（左侧心室瘘）
部位：左室
（左侧心室）
诊断分类：瘘
（瘘）
方位：左
（左侧）
部位：室
（心室）

图3

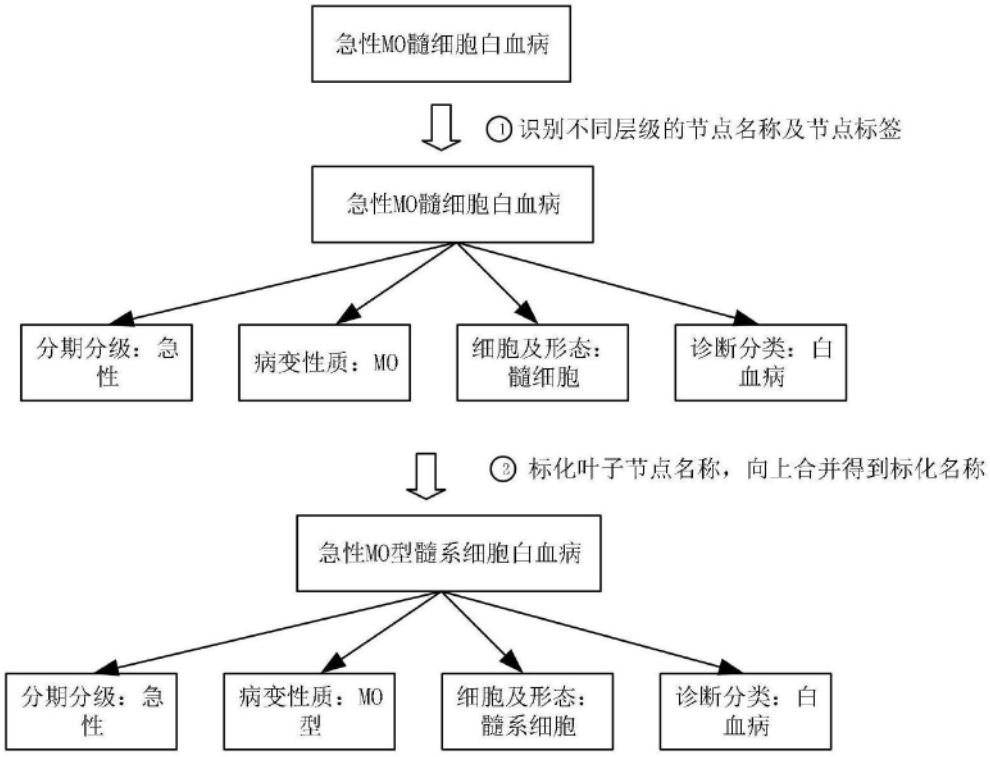

急性MO髓细胞白血病
① 识别不同层级的节点名称及节点标签
急性MO髓细胞白血病
分期分级：急性
病变性质：MO
细胞及形态：髓细胞
诊断分类：白血病
② 标化叶子节点名称，向上合并得到标化名称
急性MO型髓系细胞白血病
分期分级：急性
病变性质：MO型
细胞及形态：髓系细胞
诊断分类：白血病

图4

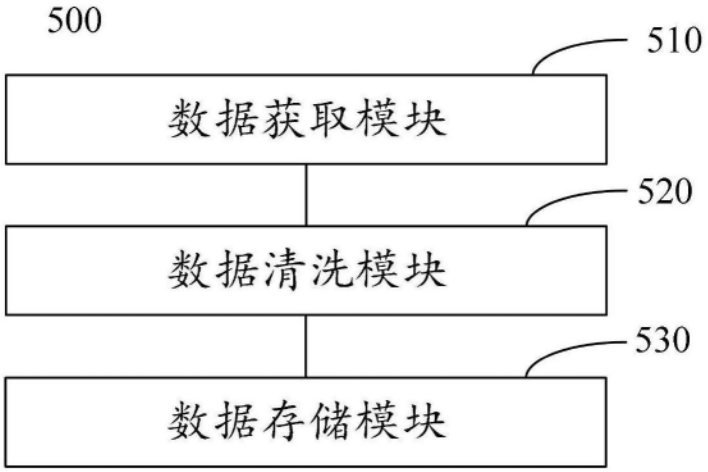


图5