(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 119862531 A
(43)申请公布日 2025.04.22

(21)申请号 202411939614.6

(22)申请日 2024.12.26

(71)申请人 广州汇通国信科技有限公司
地址 510000 广东省广州市黄埔区开源大道11号B9栋601室自编6310房

(72)发明人 李保平 谢超 杨建荣 戴思敏 欧再辉 龙荣豪

(74)专利代理机构 广州科捷知识产权代理事务所(普通合伙) 44560
专利代理师 钟慧增

(51)Int.Cl.
G06F 18/25(2023.01)
G06F 18/22(2023.01)
G06F 18/213(2023.01)
G06F 18/214(2023.01)
G06F 40/30(2020.01)

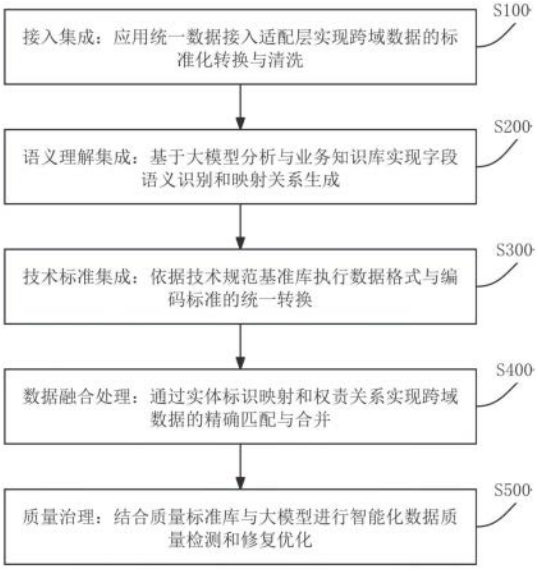
权利要求书2页 说明书9页 附图4页

(54)发明名称

基于大模型的跨域数据集成融合方法、终端及存储介质

(57)摘要

本发明提供一种基于大模型的跨域数据集成融合方法、终端及存储介质,所述方法包括:通过统一数据接入适配层对跨域原始数据进行格式转换和清洗,输出标准化的数据集;利用大模型进行语义理解集成,解析字段命名并结合业务术语知识库生成字段语义标签,通过计算相似度判定业务关联性,生成域间字段映射关系;基于所述映射关系执行技术标准转换和数据融合处理,并采用质量治理标准库与大模型相结合的方式进行智能化质量治理。本发明通过引入大模型辅助数据理解和治理,有效解决了现有跨域数据集成中语义理解不足、标准不统一等技术问题,显著提升了数据集成的准确性和自动化水平。



CN 119862531 A

1.一种基于大模型的跨域数据集成与融合治理方法,其特征在于,所述方法包括:

接入集成,接收跨域原始数据,通过统一数据接入适配层进行格式转换和清洗,输出标准化的数据集,其中所述标准化的数据集采用统一的命名格式,数据类型映射为系统支持的类型,数据内容完整且无重复值,缺失值标记统一;

语义理解集成,基于大模型分析各数据域内标准化数据集的特征,提取各域内字段数据特征,通过解析字段命名并结合所述业务术语知识库识别各域内字段的业务含义并生成字段语义标签,通过计算字段名称相似度并结合字段语义标签判定业务关联性,生成域间字段映射关系;

技术标准集成,基于所述字段语义标签,对标准化数据集执行统一的技术标准转换,包括统一数据格式规范和编码标准,输出符合统一技术规范的标准化数据;

数据融合处理,基于所述统一技术规范的标准化数据和所述域间字段映射关系,识别跨域同一实体的数据记录并进行匹配,对匹配实体的属性按照数据域权责关系进行合并处理,生成融合后的实体数据;

质量治理,基于质量治理标准库对所述融合后实体数据进行问题特征匹配和修复,对匹配成功的质量问题执行标准修复规则,对于未匹配问题采用所述大模型进行分析,识别实体属性值一致性及实体关系正确性的质量问题,生成并执行修复规则,将质量问题及处理方法更新至质量治理标准库,输出经治理的高质量数据;

其中,所述大模型为经过海量数据预训练并具备跨领域知识理解能力的语言模型,能够对数据类型、格式规范和业务规则进行智能分析。

2.根据权利要求1所述的一种基于大模型的跨域数据集成与融合治理方法,其特征在于,所述语义理解集成步骤中所述大模型处理具体包括:

提取各域内字段数据特征,通过对字段数据的分析将其分类为数值型、字符型、日期时间型、布尔型,并根据统计分布确定各类型字段的有效范围;

对各域内字段名称进行解析和语义识别,结合所述业务术语知识库生成包含数据域、业务类型、字段属性的字段语义标签;

生成域间字段映射关系,基于编辑距离算法计算不同数据域间字段名称相似度,基于所述字段语义标签判断不同数据域间字段之间的业务关联关系,将所述域间字段名称相似度和业务关联程度转换为特征向量,通过计算待判定字段映射特征向量与已验证字段映射集合中标准映射样本特征向量之间的欧氏距离,确定域间字段映射关系。

3.根据权利要求1所述的一种基于大模型的跨域数据集成与融合治理方法,其特征在于,所述方法在执行前需完成以下基础配置:

构建业务术语知识库,包含标准术语定义、字段命名规则及业务映射关系;

建立已验证的字段映射集合,用于指导新的字段映射;

配置技术规范转换基准库,包含数值规范、文本规范、时间规范和分类规范;设定数据域间的优先级顺序,用于数据属性合并时的冲突处理;

初始化质量治理标准库,包含问题特征模式库、标准修复规则库及问题-规则映射关系表,其中所述问题特征模式库用于识别数据一致性、完整性、准确性及关联性问题,所述标准修复规则库包含与问题特征对应的处理规则,所述问题-规则映射关系表用于实现质量问题与处理方法的快速匹配。

4.根据权利要求1所述的一种基于大模型的跨域数据集成与融合治理方法,其特征在于,所述技术标准集成步骤具体包括:

初始化转换环境,读取所述技术规范转换基准库中的数值规范、文本规范、时间规范和分类规范;

基于所述字段语义标签,从所述技术规范转换基准库中检索并确定各字段的目标转换规范;

依据确定的目标规范执行数据格式转换操作。

5.根据权利要求1所述的一种基于大模型的跨域数据集成与融合治理方法,其特征在于,所述数据融合处理步骤具体包括:

基于所述字段语义标签分析各数据域中的业务属性,识别具有唯一标识特征的字段,其中唯一标识特征指字段值不重复的约束特征;

通过所述域间字段映射关系关联唯一标识字段,生成实体标识字段映射集,并基于该映射集执行跨域数据记录的匹配;

基于所述域间字段映射关系和数据域权责关系,对同一实体的不同数据记录进行属性值合并,当权责数据域中属性值缺失时,按照数据域优先级采用其他域的有效值;

将合并后的属性值整合形成融合后的实体数据。

6.根据权利要求1所述的一种基于大模型的跨域数据集成与融合治理方法,其特征在于,所述质量治理步骤具体包括:

对所述融合后实体数据进行问题特征匹配,判定是否存在与质量治理标准库中记录的问题特征相符的质量问题,对于特征匹配的质量问题,执行标准库中的修复规则;

对于标准库中未涵盖的质量问题,采用所述大模型分析所述融合后实体数据,基于所述技术规范转换基准库识别属性值在数值精度、字符编码、时间格式及分类代码方面的一致性,验证实体间的关系规则,输出质量问题清单;

采用所述大模型基于所述字段语义标签和所述域间字段映射关系,生成包含问题定位条件和处理操作的修复规则并执行;

将所述质量问题及其处理方法更新至质量治理标准库,用于指导后续的质量治理工作。

7.一种基于大模型的跨域数据集成与融合治理装置,其特征在于,包括:

存储器,所述存储器用于存储包含计算机程序代码的程序指令,所述程序指令用于实现权利要求1-6任一项所述的基于大模型的跨域数据集成与融合治理方法;

处理器,所述处理器与所述存储器通过系统总线连接,用于调用并执行所述程序指令;

计算机可读存储介质,所述计算机可读存储介质与所述处理器连接,用于存储所述跨域数据集成与融合治理方法的执行结果以及所述方法执行过程中所需的数据信息。

基于大模型的跨域数据集成融合方法、终端及存储介质

技术领域

[0001] 本发明涉及数据处理技术领域,尤其涉及一种基于大模型的跨域数据集成融合方法、终端及存储介质。

背景技术

[0002] 随着数字化转型的深入推进,企业面临着跨部门、跨系统、跨平台的数据集成与融合需求日益增长。跨域数据集成是指将分布在不同数据域(如不同业务系统、不同部门、不同组织)的异构数据进行统一管理和融合利用的技术。近年来,大模型技术的快速发展为解决跨域数据集成中的复杂问题提供了新的技术手段,有望突破传统方法的局限性。

[0003] 然而,现有的跨域数据集成方案存在以下几个关键技术难点:

[0004] 第一,数据语义理解能力不足。现有技术主要依赖预定义的映射规则和人工经验进行字段匹配,这种方法难以准确理解和处理不同数据域之间的语义差异。在实际应用中,由于业务术语的多样性和复杂性,基于规则的映射方法往往需要大量人工干预,既降低了效率,又难以保证映射的准确性。

[0005] 第二,技术标准不统一的问题突出。不同数据域往往采用不同的数据格式、编码标准和技术规范,这种异构性导致数据集成过程中的一致性难以保证。在数据转换和融合过程中,由于缺乏统一的标准规范,容易出现数据失真、格式错误等问题,影响了数据集成的质量。

[0006] 第三,数据质量治理缺乏智能化。传统的数据治理方法主要依赖静态的规则库进行质量检查和修复,这种方法无法有效应对新出现的数据质量问题。随着数据规模和复杂度的增加,质量问题的类型也在不断演变,固定的治理规则难以满足动态变化的治理需求。

[0007] 基于上述分析,亟需一种能够提供智能化语义理解、统一技术标准转换、自适应质量治理的跨域数据集成解决方案。本发明正是针对这些技术难点,提出了基于大模型的创新性技术方案。

发明内容

[0008] 本发明的目的在于提供一种基于大模型的跨域数据集成融合方法,以解决现有技术中存在的语义理解不足、技术标准不统一、质量治理缺乏智能化等技术问题,提高跨域数据集成的准确性和自动化水平。

[0009] 为实现上述目的,本发明采用的技术方案如下:

[0010] 一种基于大模型的跨域数据集成融合方法,包括如下步骤:

[0011] 接入集成,接收跨域原始数据,通过统一数据接入适配层进行格式转换和清洗,输出标准化的数据集,其中所述标准化的数据集采用统一的命名格式,数据类型映射为系统支持的类型,数据内容完整且无重复值,缺失值标记统一;

[0012] 语义理解集成,基于大模型分析各数据域内标准化数据集的特征,提取各域内字段数据特征,通过解析字段命名并结合所述业务术语知识库识别各域内字段的业务含义并

生成字段语义标签,通过计算字段名称相似度并结合字段语义标签判定业务关联性,生成域间字段映射关系;

[0013] 技术标准集成,基于所述字段语义标签,对标准化数据集执行统一的技术标准转换,包括统一数据格式规范和编码标准,输出符合统一技术规范的标准化数据;

[0014] 数据融合处理,基于所述统一技术规范的标准化数据和所述域间字段映射关系,识别跨域同一实体的数据记录并进行匹配,对匹配实体的属性按照数据域权责关系进行合并处理,生成融合后的实体数据;

[0015] 质量治理,基于质量治理标准库对所述融合后实体数据进行问题特征匹配和修复,对匹配成功的质量问题执行标准修复规则,对于未匹配问题采用所述大模型进行分析,识别实体属性值一致性及实体关系正确性的质量问题,生成并执行修复规则,将质量问题及处理方法更新至质量治理标准库,输出经治理的高质量数据。

[0016] 其中,所述大模型为经过海量数据预训练并具备跨领域知识理解能力的语言模型,能够对数据类型、格式规范和业务规则进行智能分析;

[0017] 进一步的技术方案在于:语义理解集成步骤中所述大模型处理具体包括:

[0018] 提取各域内字段数据特征,通过对字段数据的分析将其分类为数值型、字符型、日期时间型、布尔型,并根据统计分布确定各类型字段的有效范围;

[0019] 对各域内字段名称进行解析和语义识别,结合所述业务术语知识库生成包含数据域、业务类型、字段属性的字段语义标签;

[0020] 生成域间字段映射关系,基于编辑距离算法计算不同数据域间字段名称相似度,基于所述字段语义标签判断不同数据域间字段之间的业务关联关系,将所述域间字段名称相似度和业务关联程度转换为特征向量,通过计算待判定字段映射特征向量与已验证字段映射集合中标准映射样本特征向量之间的欧氏距离,确定域间字段映射关系。

[0021] 进一步的技术方案在于:方法在执行前需完成以下基础配置:

[0022] 构建业务术语知识库,包含标准术语定义、字段命名规则及业务映射关系;

[0023] 建立已验证的字段映射集合,用于指导新的字段映射;

[0024] 配置技术规范转换基准库,包含数值规范、文本规范、时间规范和分类规范;

[0025] 设定数据域间的优先级顺序,用于数据属性合并时的冲突处理;

[0026] 初始化质量治理标准库,包含问题特征模式库、标准修复规则库及问题-规则映射关系表,其中所述问题特征模式库用于识别数据一致性、完整性、准确性及关联性问题,所述标准修复规则库包含与问题特征对应的处理规则,所述问题-规则映射关系表用于实现质量问题与处理方法的快速匹配。

[0027] 进一步的技术方案在于:技术标准集成步骤具体包括:

[0028] 初始化转换环境,读取所述技术规范转换基准库中的数值规范、文本规范、时间规范和分类规范;

[0029] 基于所述字段语义标签,从所述技术规范转换基准库中检索并确定各字段的目标转换规范;

[0030] 依据确定的目标规范执行数据格式转换操作。

[0031] 进一步的技术方案在于:数据融合处理步骤具体包括:

[0032] 基于所述字段语义标签分析各数据域中的业务属性,识别具有唯一标识征的字

段,其中唯一标识特征指字段值不重复的约束特征;

[0033] 通过所述域间字段映射关系关联唯一标识字段,生成实体标识字段映射集,并基于该映射集执行跨域数据记录的匹配;

[0034] 基于所述域间字段映射关系和数据域权责关系,对同一实体的不同数据记录进行属性值合并,当权责数据域中属性值缺失时,按照数据域优先级采用其他域的有效值;

[0035] 将合并后的属性值整合形成融合后的实体数据。

[0036] 进一步的技术方案在于:质量治理步骤具体包括:

[0037] 对所述融合后实体数据进行问题特征匹配,判定是否存在与质量治理标准库中记录的问题特征相符的质量问题,对于特征匹配的质量问题,执行标准库中的修复规则;

[0038] 对于标准库中未涵盖的质量问题,采用所述大模型分析所述融合后实体数据,基于所述技术规范转换基准库识别属性值在数值精度、字符编码、时间格式及分类代码方面的一致性,验证实体间的关系规则,输出质量问题清单;

[0039] 采用所述大模型基于所述字段语义标签和所述域间字段映射关系,生成包含问题定位条件和处理操作的修复规则并执行;

[0040] 将所述质量问题及其处理方法更新至质量治理标准库,用于指导后续的质量治理工作。

[0041] 本发明还提供了一种实现上述方法的终端装置。具体地,所述终端装置包括:存储器,用于存储包含程序指令的计算机程序代码,所述程序指令用于实现所述基于大模型的跨域数据集成与融合治理方法;处理器,所述处理器与所述存储器通过系统总线连接,用于调用并执行所述程序指令;以及计算机可读存储介质,所述计算机可读存储介质与所述处理器连接,用于存储所述方法的执行结果及其执行过程中所需的中间数据。当所述程序指令被所述处理器执行时,所述处理器执行基于大模型的跨域数据集成与融合治理方法的各个步骤。

[0042] 本发明的有益效果如下:

[0043] (1)本发明基于大模型的语义理解集成机制,通过智能解析字段命名并结合业务术语知识库生成字段语义标签,实现了跨域数据字段的智能映射,解决了传统方法中字段映射依赖人工经验、效率低下的问题,显著提高了数据集成的准确性和效率。

[0044] (2)本发明设计的技术标准集成方案,通过统一的数据接入适配层和技术规范转换基准库,实现了跨域异构数据的标准化处理,解决了不同数据域之间标准不统一的问题,有效保证了数据集成过程中的一致性和规范性。

[0045] (3)本发明创新性地将质量治理标准库与大模型相结合,构建了自适应的质量治理机制,能够自动识别和修复数据质量问题,并持续更新治理规则,解决了传统质量治理方法缺乏智能化、可扩展性差的问题,实现了数据治理的持续优化。

附图说明

[0046] 图1为基于大模型的跨域数据集成与融合治理方法实施例流程图。

[0047] 图2为语义理解集成流程图。

[0048] 图3为技术标准集成流程图。

[0049] 图4为数据融合处理流程图。

[0050] 图5为质量治理闭环过程示意图。

具体实施方式

[0051] 为使本发明的目的、技术方案及优点更加清楚、明确,以下参照附图并举实施例对本发明进一步详细说明。应当理解,此处所描述的具体实施例仅用以解释本发明,并不用于限定本发明。

[0052] 作为本发明一种可能的实施例,在执行跨域数据集成与融合治理方法的核心步骤前,需完成以下预置配置:

[0053] SP:预置配置步骤,用于初始化系统运行所需的基础环境。

[0054] 在本发明中,进一步的,SP步骤具体包括:

[0055] SP01:构建业务术语知识库。具体的,该库需包含:标准业务术语定义,包括各业务域的规范术语及其定义说明;字段命名规则,规定字段名称的格式标准,包括命名方法(如驼峰式或下划线式)、命名要素及其顺序;业务映射关系,定义不同业务域间的关联规则及数据流转约束。

[0056] SP02:建立已验证的字段映射集合。具体的,该集合用于存储历史验证通过的字段映射关系,包括源字段和目标字段的对应关系,作为字段映射验证的参考样本,用于评估新建映射的准确性。

[0057] SP03:配置技术规范转换基准库。具体的,该库包含:数值规范,规定数值类型的精度要求、度量单位及有效范围等;文本规范,定义字符集编码、文本长度限制等规则;时间规范,统一日期时间格式及时区处理方式;分类规范,规定状态代码、枚举值等分类数据的标准。

[0058] SP04:设定数据域优先级顺序。具体的,该配置用于明确各数据域的数据管理职责和权限,规定数据冲突时的取值优先顺序,确定跨域数据同步的流向和规则。

[0059] SP05:初始化质量治理标准库。具体的,在系统运行前需要建立完整的质量治理标准体系,该标准库作为质量问题识别和处理的基础支撑。首先,构建问题特征模式库,该库覆盖数据一致性问题特征(用于识别属性值在格式、精度、编码等方面的规范性问题)、数据完整性问题特征(用于识别必填字段缺失、关键属性空值等问题)、数据准确性问题特征(用于识别异常值、错误值等数据偏差问题)以及数据关联性问题特征(用于识别实体关系冲突、引用完整性等问题)。

[0060] 其次,构建标准修复规则库,该库包含与问题特征相对应的处理规则,涵盖数据转换规则(用于处理格式不规范、编码不统一等问题)、数据补全规则(用于处理数据缺失、空值等完整性问题)、数据校正规则(用于处理异常值、错误值等准确性问题)以及关系修复规则(用于处理实体关系冲突等关联性问题)。

[0061] 最后,建立问题-规则映射关系表,该表将问题特征与对应的修复规则进行关联,实现质量问题与处理方法的快速匹配,为后续的质量治理工作提供规则支撑。通过建立完整的质量治理标准库,确保了质量治理过程的规范性和可控性。

[0062] 本实施例中,通过完善的预置配置,为后续的数据集成与融合治理提供了基础支撑。这些配置既规范了数据处理标准,也为确保数据质量提供了基础保障。同时,各项配置可根据业务需求进行动态调整,保证了系统运行的灵活性和适应性。

[0063] 作为本发明一种可能的实施例,如图1所示,提供了一种基于大模型的跨域数据集成与融合治理方法,包括如下步骤:

[0064] S100:接入集成步骤,用于接收跨域原始数据,并通过统一的数据接入适配层进行格式转换和清洗,输出标准化的数据集。

[0065] 在本发明中,进一步的,S100步骤具体包括:

[0066] S101:接收来自不同数据域的原始数据。具体的,原始数据可能来自多个不同的业务系统或数据源,其数据格式、命名规则和编码标准均可能不同。举例说明,不同业务系统可能采用不同的数据库类型,如Oracle、MySQL等,或者采用不同的文件格式,如CSV、XML等。

[0067] S102:通过统一数据接入适配层进行数据格式转换。具体的,根据原始数据的格式特征,调用相应的适配器组件进行格式转换。举例说明,对于关系型数据库中的数据,通过数据库连接适配器读取并转换;对于文件类数据,通过文件格式适配器进行解析和转换。

[0068] S103:对转换后的数据进行清洗处理。具体的,清洗处理包括以下操作:统一命名格式,将字段名称转换为系统规定的统一格式,如采用驼峰命名法;数据类型映射,将原始数据类型映射为系统支持的标准类型;数据完整性检查,检查并处理重复值,对缺失值进行统一标记。

[0069] S104:输出标准化的数据集。具体的,将经过格式转换和清洗处理的数据,按照预定义的标准格式进行存储,形成标准化的数据集。该数据集具有统一的命名格式、标准的数据类型、完整且无重复的数据内容,以及统一的缺失值标记。

[0070] 本实施例中,通过设置统一的数据接入适配层,实现了对不同来源、不同格式的原始数据的标准化处理,为后续的语义理解和数据融合奠定了基础。同时,通过规范的数据清洗流程,保证了数据的质量和一致性,提高了数据的可用性。

[0071] 作为本发明一种可能的实施例,参照图2,接收到标准化的数据集后,执行语义理解集成步骤,具体如下:

[0072] S200:语义理解集成步骤,用于基于大模型分析各数据域内标准化数据集的特征,识别域内字段的业务含义并生成字段间的映射关系。

[0073] 在本发明中,进一步的,S200步骤中所述大模型处理具体包括:

[0074] S201:提取各域内字段数据特征。具体的,将各域内字段数据类型分为数值型、字符型、日期时间型、布尔型。通过对字段数据进行统计分析,设定其有效范围:对于数值型字段,根据数据分布设定上下限阈值;对于字符型字段,确定其长度范围;对于日期时间型字段,设定有效时间区间。这种基于统计分布的特征提取方法,可以有效识别异常值并确保数据的有效性。

[0075] S202:执行字段名称解析和语义识别。具体的,首先按照预置配置的命名规则对字段名称进行解析。然后将解析结果与业务术语知识库中的标准术语进行匹配,通过语义识别生成字段语义标签。该标签包含三个关键维度:数据域(标识字段所属的业务领域)、业务类型(说明字段的业务用途)、字段属性(描述字段的特征属性)。这种多维度的语义标注方式为后续的字段映射提供了可靠的语义基础。

[0076] S203:生成域间字段映射关系。具体的,该步骤包括以下过程:

[0077] 采用编辑距离算法计算不同数据域间字段名称的相似度;

[0078] 基于前述生成的字段语义标签判断字段间的业务关联关系;

[0079] 将字段名称相似度和业务关联程度转换为向量空间中的特征点,构建特征向量。该特征向量包含字段名称相似度分量和业务语义相关度分量,每个分量经过归一化处理后的取值范围为[0,1]。

[0080] S204:验证字段映射关系。具体的,基于SP02中已验证字段映射集合的标准映射样本,计算待判定字段映射对应的特征向量与标准映射样本特征向量之间的欧氏距离。当计算所得欧氏距离小于预设阈值时,即可确认存在域间字段映射关系。

[0081] 举例说明,在某数据集成场景中,预设的字段映射判定阈值为0.1。当计算得到待判定字段映射特征向量<0.85,0.75>与已验证映射样本特征向量<0.82,0.78>之间的欧氏距离为0.058时,由于0.058小于预设阈值0.1,因此可以确认这两个字段之间存在有效的映射关系。反之,若计算得到的欧氏距离为0.15,则超出预设阈值,表明两个字段之间不存在可靠的映射关系。

[0082] 本实施例中,通过采用大模型进行智能分析,结合特征提取、语义识别和映射验证等步骤,实现了对字段业务含义的准确理解和映射关系的可靠建立。该方法不仅提高了数据集成的准确性,也为后续的数据融合处理奠定了坚实的基础。

[0083] 作为本发明一种可能的实施例,参照图3,在完成语义理解集成后,执行技术标准集成步骤,具体如下:

[0084] S300:技术标准集成步骤,用于基于字段语义标签,对标准化数据集执行统一的技术标准转换,确保数据格式规范和编码标准的一致性。

[0085] 在本发明中,进一步的,S300步骤具体包括:

[0086] S301:初始化转换环境。具体的,通过配置文件加载SP03中技术规范转换基准库的规范定义,将数值规范、文本规范、时间规范和分类规范加载至存储器,建立规范索引表,初始化转换参数。该步骤通过缓存机制提高后续规范检索的效率,并确保转换过程的稳定性。

[0087] S302:确定目标转换规范。具体的,基于字段语义标签中的数据域、业务类型和字段属性信息,在技术规范转换基准库中确定目标转换规范。

[0088] 首先,根据字段语义标签构建检索条件,依次匹配数据域、业务类型和字段属性,在规范转换基准库中精确定位相关的规范定义。通过数据域匹配确保在正确的业务范围内检索,基于业务类型定位相应规范集合,再利用字段属性信息缩小规范查找范围。

[0089] 其次,对检索到的候选规范进行适用性判断。通过检查数据类型匹配度,确保源数据类型可以无损转换为目标类型;验证数据取值范围的适配性,保证转换后数据满足目标规范的取值要求;同时核实数据格式的兼容性,确认源数据格式能够按规范要求进行转换。

[0090] 最后,完成目标规范的确定。当存在唯一符合条件的规范时,直接确定为目标转换规范;当出现多个候选规范时,按照数据域优先级顺序选择最优规范。确定后的字段规范映射关系将记录到转换配置表中,作为后续转换操作的依据。这种结构化的规范确定流程,确保了每个字段都能找到最合适的转换规范。

[0091] S303:执行数据格式转换。具体的,根据确定的目标规范,采用以下转换处理方法:

[0092] 对于数值型数据,调用数值处理函数执行精度调整、单位换算和区间归一化;对于文本型数据,使用字符串处理函数进行字符集转换、编码标准化和格式规范化;对于时间型数据,通过日期时间处理函数统一格式,处理时区转换,标准化表示;对于分类型数据,基于映射表执行代码转换、枚举值标准化处理。

[0093] 本实施例中,通过规范化的技术标准集成步骤,实现了数据在技术层面的标准统一,为后续的数据融合和质量治理提供了可靠的基础。同时,通过细致的规范匹配和转换处理,保证了数据转换的准确性和一致性。

[0094] 作为本发明一种可能的实施例,如图4所示,在完成技术标准集成后,执行数据融合处理步骤,具体如下:

[0095] S400:数据融合处理步骤,用于基于统一技术规范的标准化数据和域间字段映射关系,识别和匹配跨域同一实体的数据记录,并进行属性合并处理。

[0096] 在本发明中,进一步的,S400步骤具体包括:

[0097] S401:识别唯一标识字段。具体的,基于字段语义标签分析各数据域中的业务属性,识别具有唯一标识特征的字段。唯一标识特征是指在该数据域内字段值不重复的约束特征。通过分析字段的取值分布特征和业务规则约束,确定各数据域中的唯一标识字段集合。

[0098] S402:生成实体标识字段映射集。具体的,该步骤包括以下处理过程:

[0099] 首先,基于S200步骤生成的域间字段映射关系,筛选出涉及唯一标识字段的映射关系;其次,对筛选出的映射关系进行验证,确保映射字段在跨域场景下的一致性。验证过程包括:检查字段值的格式一致性,验证字段值的对应关系,确认字段值的时效性;最后,将验证通过的映射关系整合形成实体标识字段映射集,用于后续的数据记录匹配。该映射集包含源域标识字段、目标域标识字段、映射规则和有效期等信息。

[0100] S403:执行跨域数据记录匹配。具体的,基于实体标识字段映射集,采用以下匹配策略:

[0101] 对于具有直接映射关系的标识字段,采用精确匹配方式,即字段值完全相同的记录被判定为同一实体;对于存在格式差异的标识字段,先按照映射规则进行格式标准化,再执行匹配。

[0102] S404:执行属性值合并。具体的,对匹配确认的同一实体数据记录,基于域间字段映射关系和数据域权责关系,进行属性值合并处理:

[0103] 首先,根据SP04步骤设定的数据域优先级顺序,确定各属性的主管数据域。

[0104] 其次,按照“以主管域为准”的原则处理属性值。当主管数据域中的属性值存在且有效时,采用主管域的值;当主管域中属性值缺失时,按照SP04步骤中设定的数据域优先级顺序,采用其他域中的有效值。

[0105] 最后,将合并后的属性值整合形成融合后的实体数据。该数据包含统一的实体标识、来自各域的有效属性值、属性值的来源域标识等信息。

[0106] 本实施例中,通过严格的唯一标识字段识别、实体记录匹配和属性值合并处理,实现了跨域数据的精确融合。同时,通过引入数据域权责关系和优先级机制,确保了数据融合结果的准确性和权威性。该方法可有效处理跨域数据集成中的实体识别和属性合并问题,为后续的质量治理提供了可靠的数据基础。

[0107] 作为本发明一种可能的实施例,参照图5,在完成数据融合处理后,执行质量治理步骤,具体如下:

[0108] S500:质量治理步骤,用于基于质量治理标准库和大模型相结合的方式,对融合后实体数据进行智能化的质量检查和修复。

[0109] 在本发明中,进一步的,S500步骤具体包括:

[0110] S501:执行问题特征匹配。具体的,基于SP05步骤中初始化的质量治理标准库内的问题特征模式,对融合后实体数据进行特征检测。通过将数据特征与标准库中的问题特征模式进行匹配,识别数据中存在的一致性、完整性、准确性及关联性等质量问题。对于与标准库中特征模式相匹配的质量问题,系统自动调用SP05步骤中预设的标准修复规则进行处理。

[0111] S502:大模型分析未匹配问题。具体的,对于在SP05步骤初始化的质量治理标准库中未涵盖的质量问题,系统采用大模型进行智能分析。首先,基于SP03步骤中配置的技术规范转换基准库,系统识别实体属性值在数值精度、字符编码、时间格式及分类代码等方面的一致性问题。其次,验证实体间的关系规则,检查实体之间的从属关系、互斥关系等业务约束是否满足。通过大模型的分析,系统将识别出的所有质量问题整理形成问题清单,该清单包含问题类型、涉及字段、违规程度等详细信息。

[0112] S503:生成修复规则。具体的,系统采用大模型基于字段语义标签和域间字段映射关系,针对质量问题清单中的各类问题生成修复规则。在规则生成过程中,首先分析问题出现的数据特征和业务场景,构建精确的问题定位条件。其次,根据技术规范和业务规则,设计相应的问题处理操作。最后,将问题定位条件和处理操作组装形成完整的修复规则。

[0113] S504:执行修复处理。具体的,系统按照生成的修复规则对质量问题进行处理。在修复过程中,首先根据问题定位条件筛选出待处理的数据记录,然后按照处理操作对问题数据执行修复。系统同步记录修复过程中的关键信息,包括修复前后的数据状态、应用的规则等,确保修复过程的可追溯性。

[0114] S505:更新质量治理标准库。具体的,系统将本次质量治理过程中发现的问题特征和处理方法更新至SP05步骤中初始化的质量治理标准库。在更新过程中,首先将新发现的质量问题特征及其识别方法添加到问题特征模式库,其次将新生成的修复规则添加到标准修复规则库,最后根据修复效果对规则进行评分和优化,实现质量治理能力的持续提升。

[0115] 本实施例中,通过SP05步骤初始化的质量治理标准库与大模型相结合的方式,实现了对融合数据的智能化质量治理。该方法不仅能够处理已知的质量问题,还能够通过大模型分析发现和解决新出现的质量问题,同时通过持续更新质量治理标准库实现了治理能力的动态优化,为数据质量的持续提升提供了有效保障。

[0116] 作为本发明一种可能的实施例,提供一种实现上述跨域数据集成与融合治理方法的具体装置,其技术实现如下:

[0117] 所述装置包括存储器、处理器及计算机可读存储介质。其中,存储器和处理器通过系统总线连接,形成数据交互通路。

[0118] 在本发明中,存储器可以包括但不限于高速RAM存储器、非易失性存储器(如固态硬盘、机械硬盘等)。所述存储器用于存储本发明的计算机程序,该程序包含实现上述跨域数据集成与融合治理全部步骤的程序代码。具体而言,存储器中设置有以下关键程序模块:

[0119] (1)数据接入适配模块:包含各类数据源的接入适配器程序,用于实现S100步骤中的数据格式转换和清洗功能;

[0120] (2)语义理解集成模块:包含调用大模型接口的程序代码,实现S200步骤中的字段特征提取、语义识别和映射关系生成;

[0121] (3)技术标准转换模块:包含数据规范化处理的程序代码,实现S300步骤中的技术标准统一转换;

[0122] (4)数据融合处理模块:包含实体识别匹配和属性合并的程序代码,实现S400步骤的数据融合功能;

[0123] (5)质量治理模块:包含质量检查和修复处理的程序代码,实现S500步骤中的数据质量治理。

[0124] 处理器可以是通用处理器(如Intel、AMD的CPU),也可以是专用的数据处理芯片。所述处理器通过执行存储器中的程序代码,实现上述各个步骤的具体功能。在程序执行过程中,处理器可以:

[0125] 调用数据接入适配模块,完成跨域原始数据的格式转换和标准化处理;

[0126] 加载语义理解集成模块,通过大模型分析实现字段语义理解和映射;

[0127] 运行技术标准转换模块,执行数据规范的统一转换;

[0128] 启动数据融合处理模块,完成实体匹配和属性合并;

[0129] 触发质量治理模块,实现数据质量的智能检查和修复。

[0130] 计算机可读存储介质可以是磁盘、光盘、固态存储器等多种形式,用于存储上述计算机程序。该程序被处理器加载和执行时,将实现如前所述的跨域数据集成与融合治理方法的全部步骤。程序在执行过程中,可以访问存储器中的各类配置信息,包括业务术语知识库、技术规范转换基准库、质量治理标准库等,确保方法的正常运行。

[0131] 通过上述硬件环境和软件组件的有机结合,本发明提供的装置能够高效实现跨域数据的智能集成与融合治理,为企业数据资产的统一管理提供可靠的技术支撑。

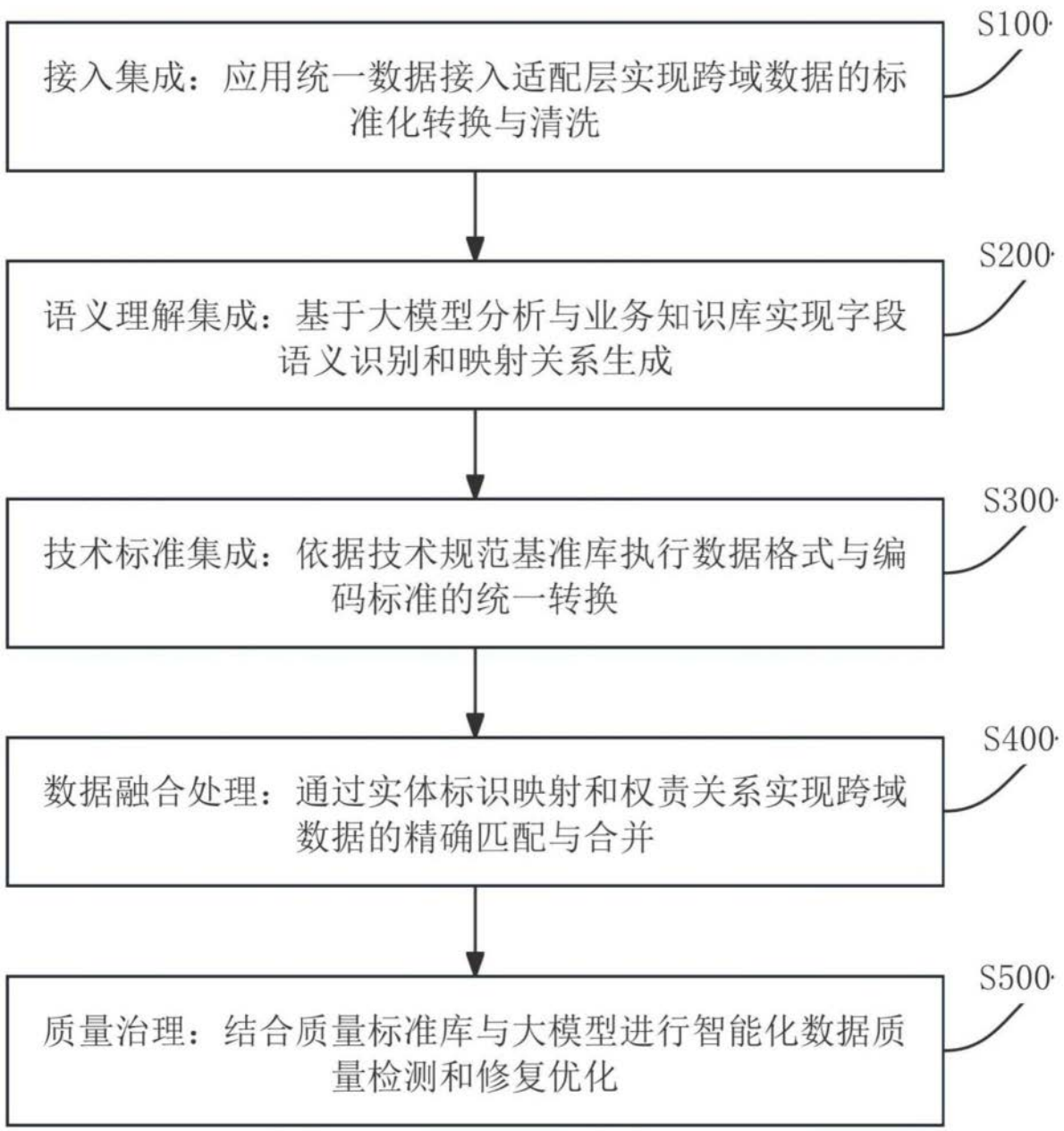

S100
接入集成：应用统一数据接入适配层实现跨域数据的标准化转换与清洗
S200
语义理解集成：基于大模型分析与业务知识库实现字段语义识别和映射关系生成
S300
技术标准集成：依据技术规范基准库执行数据格式与编码标准的统一转换
S400
数据融合处理：通过实体标识映射和权责关系实现跨域数据的精确匹配与合并
S500
质量治理：结合质量标准库与大模型进行智能化数据质量检测和修复优化

图1

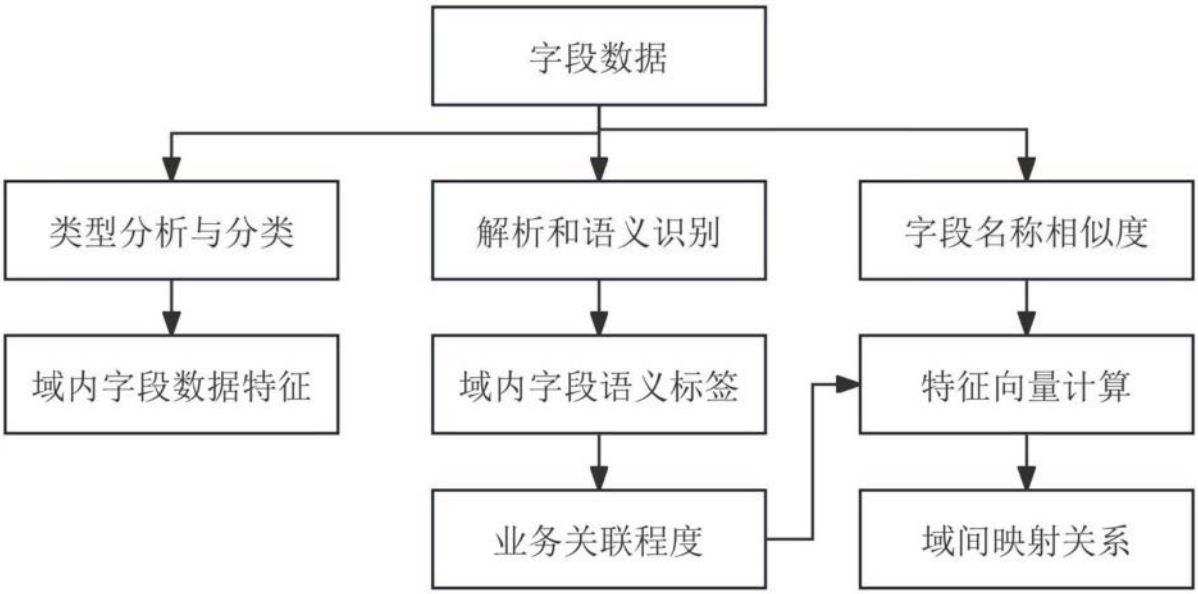

字段数据
类型分析与分类
解析和语义识别
字段名称相似度
域内字段数据特征
域内字段语义标签
特征向量计算
业务关联程度
域间映射关系

图2

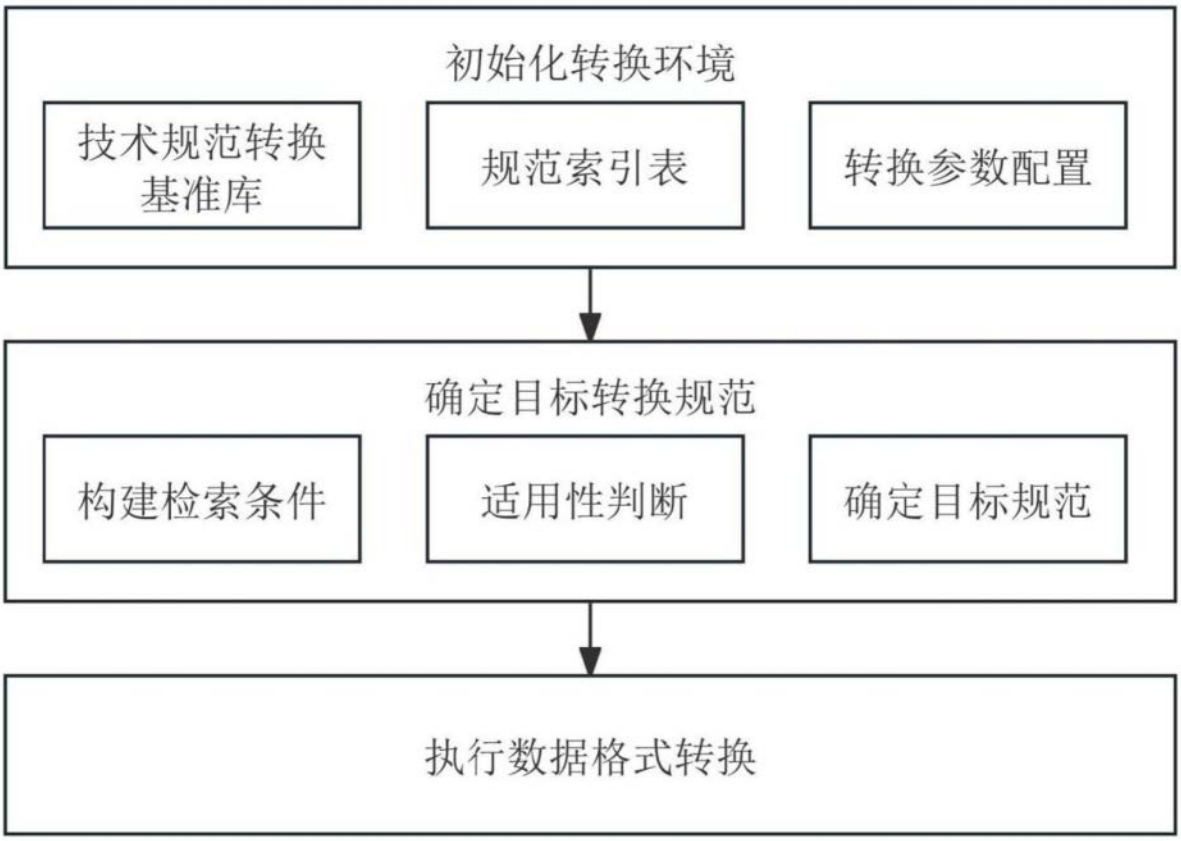

初始化转换环境
技术规范转换基准库
规范索引表
转换参数配置
确定目标转换规范
构建检索条件
适用性判断
确定目标规范
执行数据格式转换

图3

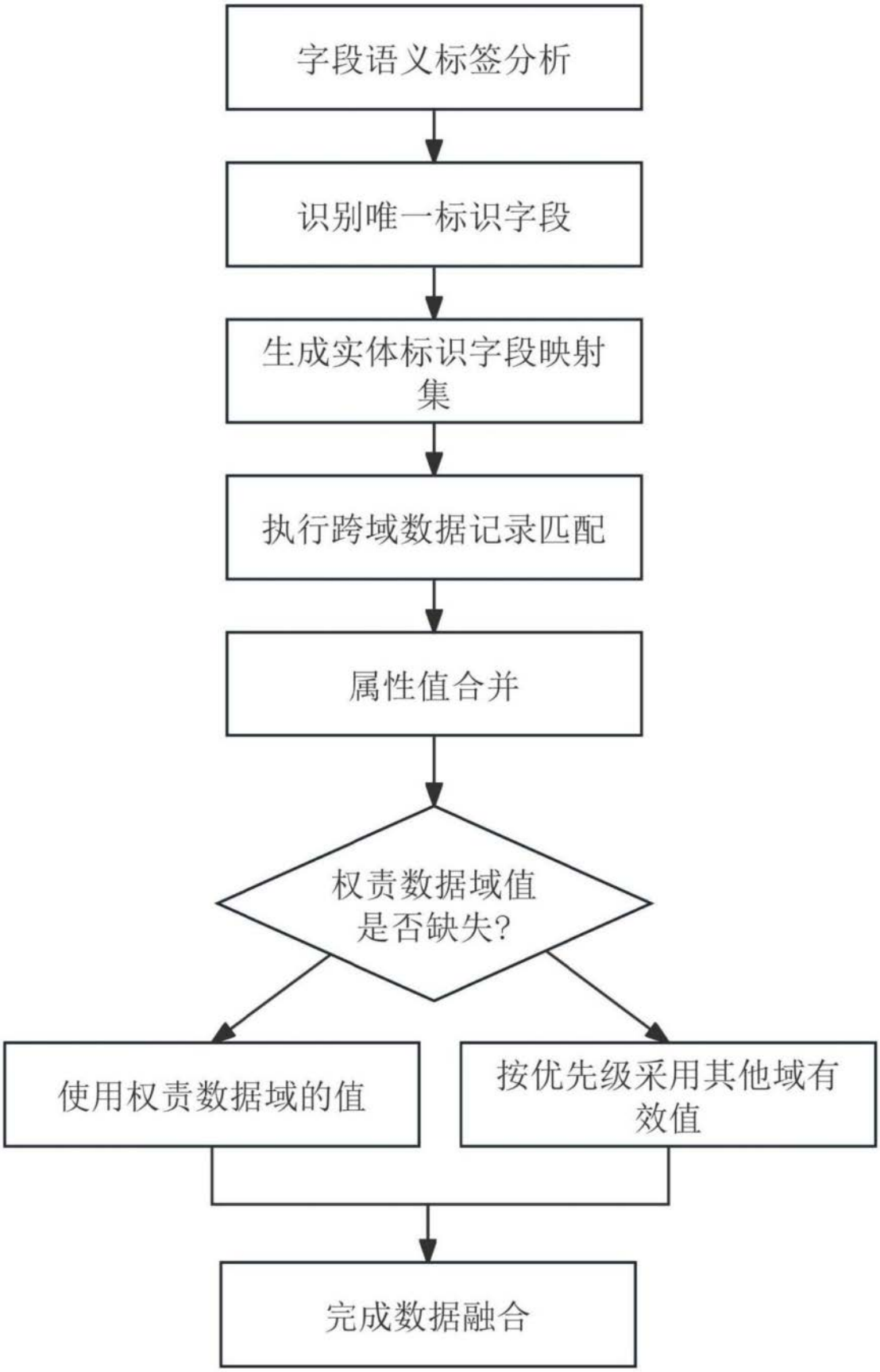

字段语义标签分析
识别唯一标识字段
生成实体标识字段映射集
执行跨域数据记录匹配
属性值合并
权责数据域值是否缺失?
使用权责数据域的值
按优先级采用其他域有效值
完成数据融合

图4

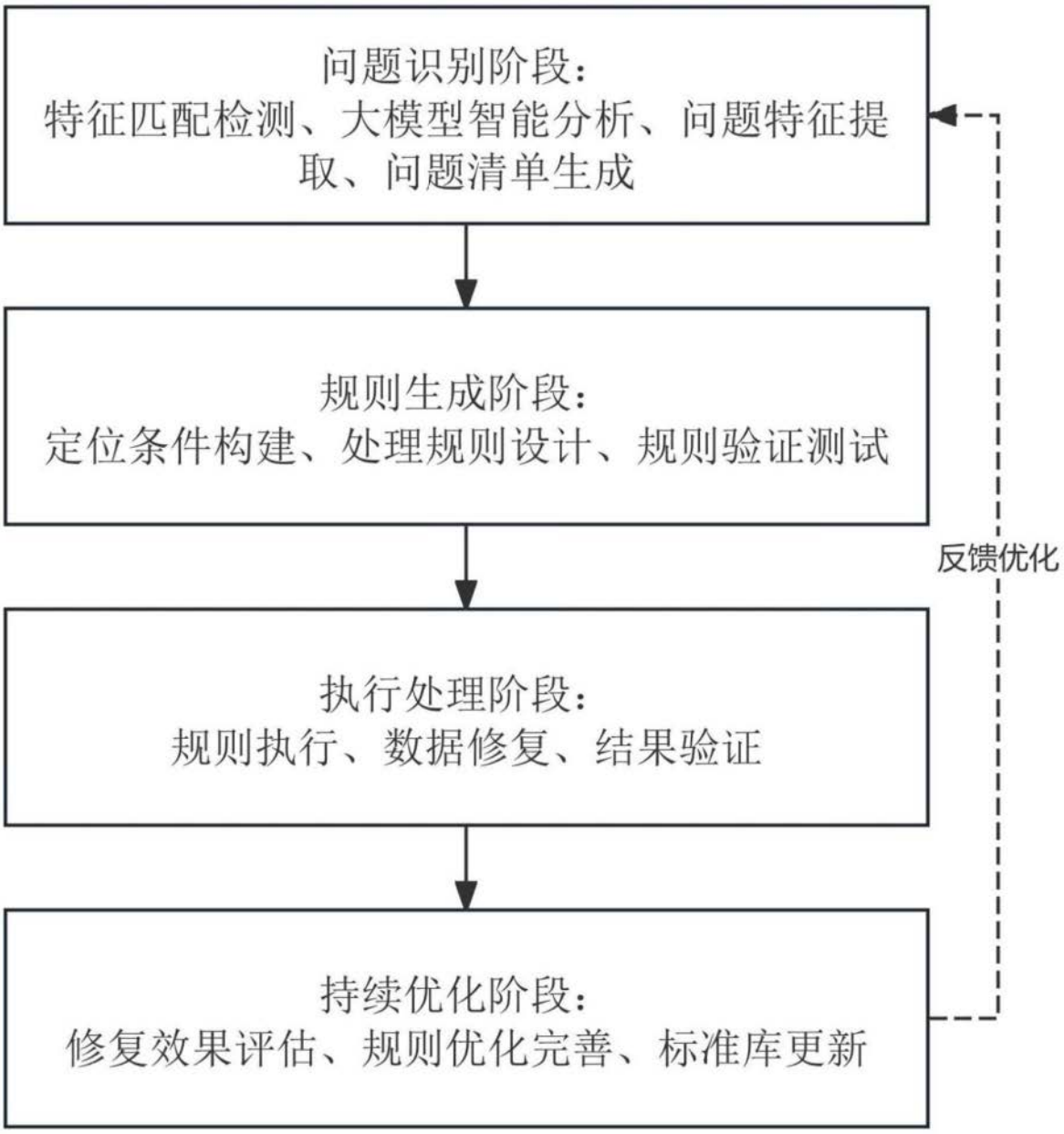

问题识别阶段：
特征匹配检测、大模型智能分析、问题特征提取、问题清单生成
规则生成阶段：
定位条件构建、处理规则设计、规则验证测试
执行处理阶段：
规则执行、数据修复、结果验证
持续优化阶段：
修复效果评估、规则优化完善、标准库更新
反馈优化

图5