(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 112363996 A
(43)申请公布日 2021.02.12

(21)申请号 202011197189.X

(22)申请日 2020.10.30

(71)申请人 国家电网有限公司大数据中心
地址 100052 北京市西城区骡马市大街18号中再中心

(72)发明人 沈亮 杨帅 朱广新 廖小琦 王春梅 宜东海 吴桂栋 吴一郝保聪

(74)专利代理机构 北京汇知杰知识产权代理有限公司 11587
代理人 李洁 董江虹

(51)Int.Cl.
G06F 16/21(2019.01)
G06Q 50/06(2012.01)
G06F 16/36(2019.01)

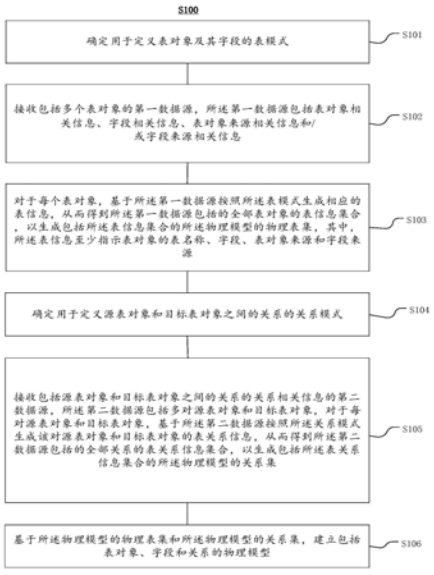
权利要求书2页 说明书12页 附图4页

(54)发明名称

用于建立电网知识图谱的物理模型的方法及系统和介质

(57)摘要

本发明提出一种用于建立用于电网知识图谱的物理模型的方法、系统及介质。方法包括：确定用于定义表对象及其字段的表模式；基于第一数据源按照表模式生成所有表对象的表信息以生成物理模型的物理表集；确定用于定义源表对象和目标表对象之间的关系的关系模式；对于第二数据源中的经过去重复处理的每对源表对象和目标表对象，基于第二数据源按照关系模式生成相应的表关系信息以生成物理模型的关系集；基于物理表集和关系集建立包括表对象、字段和关系的物理模型。利用本发明的方案，可以对不同数据源进行知识抽取，对现有模型进行查漏补缺以弥补现有模型的设计短板，给用户提供更合理的管控模型，并支持统一数据模型的信息匹配。



CN 112363996 A

1.一种用于建立用于电网知识图谱的物理模型的方法,包括:

确定用于定义表对象及其字段的表模式;

接收包括多个表对象的第一数据源,所述第一数据源包括表对象相关信息、字段相关信息、表对象来源相关信息和/或字段来源相关信息;

对于每个表对象,基于所述第一数据源按照所述表模式生成相应的表信息,从而得到所述第一数据源包括的全部表对象的表信息集合,以生成包括所述表信息集合的所述物理模型的物理表集,其中,所述表信息至少指示表对象的表名称、字段、表对象来源和字段来源;

确定用于定义源表对象和目标表对象之间的关系的关系模式;

接收包括源表对象和目标表对象之间的关系的关系相关信息的第二数据源,所述第二数据源包括多对源表对象和目标表对象,对于每对源表对象和目标表对象,基于所述第二数据源按照所述关系模式生成该对源表对象和目标表对象的表关系信息,从而得到所述第二数据源包括的全部关系的表关系信息集合,以生成包括所述表关系信息集合的所述物理模型的关系集;

基于所述物理模型的物理表集和所述物理模型的关系集,建立包括表对象、字段和关系的物理模型。

2.根据权利要求1所述的方法,所述字段基于所述第一数据源中的字段相关信息和字段来源相关信息按照预定义的字段模式确定,所述字段模式包括字段的字段名称、字段数据类型、字段描述、标准代码、数据存储格式、哈希列、责任部门、数据来源系统的名称、数据来源系统的表名称、数据来源系统的字段名称和数据来源系统的字段类型。

3.根据权利要求1所述的方法,所述表模式包括表对象的表名称、主题域、二级主题域、表类型、表描述、责任部门、数据来源系统的名称、数据来源系统的表名称和字段列表。

4.根据权利要求1所述的方法,所述关系模式包括源表对象的表名称、目标表对象的表名称、源表对象与目标表对象之间的关联关系、源表对象与目标表对象之间的关联字段、主题域和二级主题域。

5.根据权利要求4所述的方法,基于所述第二数据源按照所述关系模式生成一对源表对象和目标表对象的表关系信息包括:对于所述第二数据源中的一对源表对象和目标表对象以及另一对源表对象和目标表对象,如果其各自的源表对象的表名称、目标表对象的表名称、源表对象与目标表对象之间的关联关系和源表对象与目标表对象之间的关联字段都相同,则判定该一对源表对象和目标表对象之间的关系与该另一对源表对象和目标表对象的关系相同,对于相同的关系,只对其中的一个关系进行规范化处理并按照所述关系模式生成相应的源表对象和目标表对象的表关系信息。

6.一种用于建立用于电网知识图谱的物理模型的系统,包括:物理表集生成单元、关系集生成单元和处理单元,

其中,所述物理表集生成单元被配置为:

确定用于定义表对象及其字段的表模式;

接收包括多个表对象的第一数据源,所述第一数据源包括表对象相关信息、字段相关信息、表对象来源相关信息和/或字段来源相关信息;

对于每个表对象,基于所述第一数据源按照所述表模式生成相应的表信息,从而得到

所述第一数据源包括的全部表对象的表信息集合，以生成包括所述表信息集合的所述物理模型的物理表集，其中，所述表信息至少指示表对象的表名称、字段、表对象来源和字段来源；

其中，所述关系集生成单元被配置为：

确定用于定义源表对象和目标表对象之间的关系的关系模式；

接收包括源表对象和目标表对象之间的关系的关系相关信息的第二数据源，所述第二数据源包括多对源表对象和目标表对象，对于每对源表对象和目标表对象，基于所述第二数据源按照所述关系模式生成该对源表对象和目标表对象的表关系信息，从而得到所述第二数据源包括的全部关系的表关系信息集合，以生成包括所述表关系信息集合的所述物理模型的关系集；

其中，所述处理单元被配置为：

基于所述物理模型的物理表集和所述物理模型的关系集，建立包括表对象、字段和关系的物理模型。

7.根据权利要求6所述的系统，所述字段基于所述第一数据源中的字段相关信息和字段来源相关信息按照预定义的字段模式确定，所述字段模式包括字段的字段名称、字段数据类型、字段描述、标准代码、数据存储格式、哈希列、责任部门、数据来源系统的名称、数据来源系统的表名称、数据来源系统的字段名称和数据来源系统的字段类型。

8.根据权利要求6所述的系统，所述表模式包括表对象的表名称、主题域、二级主题域、表类型、表描述、责任部门、数据来源系统的名称、数据来源系统的表名称和字段列表，所述关系模式包括源表对象的表名称、目标表对象的表名称、源表对象与目标表对象之间的关联关系、源表对象与目标表对象之间的关联字段、主题域和二级主题域。

9.根据权利要求8所述的系统，基于所述第二数据源按照所述关系模式生成一对源表对象和目标表对象的表关系信息包括：对于所述第二数据源中的一对源表对象和目标表对象以及另一对源表对象和目标表对象，如果其各自的源表对象的表名称、目标表对象的表名称、源表对象与目标表对象之间的关联关系、源表对象与目标表对象之间的关联字段都相同，则判定该一对源表对象和目标表对象之间的关系与该另一对源表对象和目标表对象的关系相同，对于相同的关系，只对其中的一个关系进行规范化处理并按照所述关系模式生成相应的源表对象和目标表对象的表关系信息。

10.一种计算机可读存储介质，其特征在于，其上存储有计算机程序，所述计算机程序在被处理器执行时实现权利要求1至5任一项所述的方法。

用于建立电网知识图谱的物理模型的方法及系统和介质

技术领域

[0001] 本发明涉及知识图谱技术,更具体而言,涉及一种用于建立用于电网知识图谱的物理模型的方法及相应的系统和计算机可读存储介质。

背景技术

[0002] 随着知识图谱技术的进一步发展,知识图谱以其强大的语义处理能力和知识组织能力为大规模知识库组织和智能化应用奠定了基础。知识图谱由大量实体和实体关联构成。通过知识图谱,可以检索地标、人名、城市、运动队、建筑物、地理特征、电影、天体、艺术作品等实体,并获取与这些实体相关的信息。这是构建智能应用的关键,它融入了网络的集体智慧,并且能更像人去理解世界。在具体的应用场合,需要基于特定领域本体库建设领域知识图谱,支撑面向特定领域的信息智能检索和领域智能应用建设。面向特定领域的知识图谱建设不仅需要通用知识,更侧重结合领域专业知识。领域知识图谱的建设需要支撑实际工程应用,相比通用知识图谱的建设在识别率、准确性等相关指标方面有更高的要求。为了满足面向领域的大规模知识库及智能应用建设,需要研究适应领域特征的信息抽取技术及领域知识图谱的构建方法。

[0003] 近些年,国内推出了大量以中文为主语言的知识图谱,它们主要都是基于百度百科和维基百科的结构化信息构建起来的,旨在利用社区力量维护开放域知识图谱的Schema标准。知识图谱的构建方式包括人工编辑和自动抽取,但自动抽取方法主要是基于在线百科中的结构化信息而忽略了非结构化文本,而互联网中大部分的信息恰恰是以非结构化的自由文本形式呈现。在链接数据发展的同期,很多基于信息抽取技术的知识获取方法被提出,用以构建基于自由文本的开放域知识图谱。2007年,华盛顿大学Banko等人率先提出开放域信息抽取(OIE),直接从大规模自由文本中直接抽取实体关系三元组,即头实体、关系指示词以及尾实体三部分。在OIE提出之前,也有很多面向自由文本的信息抽取被提出,但这些方法主要的思路都是为每个目标关系训练相应的抽取器。这类传统的信息抽取方法在面对互联网文本中海量的关系类别时无法高效地工作,即为每个目标关系训练抽取器是不现实的,更为严重的是,很多情况下面对海量的网络文本我们无法事先明确关系的类型。

[0004] 此外,当前基于企业级数据模型的知识资源分类、智能搜索、以及对于跨域的知识融合和表示尚处于起步阶段,缺乏面向相关管理人员、业务人员的直观通俗的模型界面,同时数据模型的逻辑链路搜索能力及静态语义分析评估能力也受到严重限制。诸如国家电网公司企业公共数据模型(SG-CIM)的数据模型作为公司企业级电网、资产、财务等方面数据的全面抽象,不仅数量庞大,而且涉及专业门类极多,使得在模型成果、应用和支撑三个方面仍存在以下问题:(1)模型设计质量仍需完善,即在目前模型设计成果中,仍存在部分数据对象抽象程度不一致、实体关系不准确、数据对象及属性不完整、去重不彻底、数据溯源不完整、标准编码与源端业务系统编码不对应等实际问题;(2)模型映射率不高,即各单位基于不同版本的物理模型进行映射比对,导致平均映射率较低;(3)缺乏工具支撑,即目前数据模型管控多采用线下方式,流程复杂、沟通效率低,且模型设计成果较为抽象,造成各

级人员对模型难以理解，应用能力不足，模型应用与迭代完善质量无法保证。

[0005] 因此，需要提供一种改进的技术方案，以克服现有数据模型中存在的缺陷。

发明内容

[0006] 本发明的目的在于提供一种方案，以解决上述技术问题。

[0007] 具体地，根据本发明的第一方面，提供一种用于建立用于电网知识图谱的物理模型的方法，包括：

[0008] 确定用于定义表对象及其字段的表模式；

[0009] 接收包括多个表对象的第一数据源，所述第一数据源包括表对象相关信息、字段相关信息、表对象来源相关信息和/或字段来源相关信息；

[0010] 对于每个表对象，基于所述第一数据源按照所述表模式生成相应的表信息，从而得到所述第一数据源包括的全部表对象的表信息集合，以生成包括所述表信息集合的所述物理模型的物理表集，其中，所述表信息至少指示表对象的表名称、字段、表对象来源和字段来源；

[0011] 确定用于定义源表对象和目标表对象之间的关系的关系模式；

[0012] 接收包括源表对象和目标表对象之间的关系的关系相关信息的第二数据源，所述第二数据源包括多对源表对象和目标表对象，对于每对源表对象和目标表对象，基于所述第二数据源按照所述关系模式生成该对源表对象和目标表对象的表关系信息，从而得到所述第二数据源包括的全部关系的表关系信息集合，以生成包括所述表关系信息集合的所述物理模型的关系集；

[0013] 基于所述物理模型的物理表集和所述物理模型的关系集，建立包括表对象、字段和关系的物理模型。

[0014] 在一个实施例中，所述字段基于所述第一数据源中的字段相关信息和字段来源相关信息按照预定义的字段模式确定，所述字段模式包括字段的字段名称、字段数据类型、字段描述、标准代码、数据存储格式、哈希列、责任部门、数据来源系统的名称、数据来源系统的表名称、数据来源系统的字段名称和数据来源系统的字段类型。

[0015] 在一个实施例中，所述表模式包括表对象的表名称、主题域、二级主题域、表类型、表描述、责任部门、数据来源系统的名称、数据来源系统的表名称和字段列表。

[0016] 在一个实施例中，所述关系模式包括源表对象的表名称、目标表对象的表名称、源表对象与目标表对象之间的关联关系、源表对象与目标表对象之间的关联字段、主题域和二级主题域。

[0017] 在一个实施例中，基于所述第二数据源按照所述关系模式生成一对源表对象和目标表对象的表关系信息包括：对于所述第二数据源中的一对源表对象和目标表对象以及另一对源表对象和目标表对象，如果其各自的源表对象的表名称、目标表对象的表名称、源表对象与目标表对象之间的关联关系和源表对象与目标表对象之间的关联字段都相同，则判定该一对源表对象和目标表对象之间的关系与该另一对源表对象和目标表对象的关系相同，对于相同的关系，只对其中的一个关系进行规范化处理并按照所述关系模式生成相应的源表对象和目标表对象的表关系信息。

[0018] 在一个实施例中，设有表对象及其字段的表模式的库，从所述表模式的库确定用

于定义表对象及其字段的表模式。

[0019] 在一个实施例中，设有表示表对象之间的关系的关系模式的库，从所述关系模式的库确定用于定义源表对象和目标表对象之间的关系的关系模式。

[0020] 在一个实施例中，设有表对象、其字段、表对象之间的关系的别名集库，所述别名集库包括既往记录的别名及其出现频次，将所述第一数据源和所述第二数据源中出现的表对象、其字段、表对象之间的关系记录到所述别名库中，并将出现的频次累加；显示的表对象、其字段、表对象之间的关系为出现频次最大的表对象、其字段、表对象之间的关系。

[0021] 根据本发明的第二方面，提供一种用于建立用于电网知识图谱的物理模型的系统，包括：物理表集生成单元、关系集生成单元和处理单元，

[0022] 其中，所述物理表集生成单元被配置为：

[0023] 确定用于定义表对象及其字段的表模式；

[0024] 接收包括多个表对象的第一数据源，所述第一数据源包括表对象相关信息、字段相关信息、表对象来源相关信息和/或字段来源相关信息；

[0025] 对于每个表对象，基于所述第一数据源按照所述表模式生成相应的表信息，从而得到所述第一数据源包括的全部表对象的表信息集合，以生成包括所述表信息集合的所述物理模型的物理表集，其中，所述表信息至少指示表对象的表名称、字段、表对象来源和字段来源；

[0026] 其中，所述关系集生成单元被配置为：

[0027] 确定用于定义源表对象和目标表对象之间的关系的关系模式；

[0028] 接收包括源表对象和目标表对象之间的关系的关系相关信息的第二数据源，所述第二数据源包括多对源表对象和目标表对象，对于每对源表对象和目标表对象，基于所述第二数据源按照所述关系模式生成该对源表对象和目标表对象的表关系信息，从而得到所述第二数据源包括的全部关系的表关系信息集合，以生成包括所述表关系信息集合的所述物理模型的关系集；

[0029] 其中，所述处理单元被配置为：

[0030] 基于所述物理模型的物理表集和所述物理模型的关系集，建立包括表对象、字段和关系的物理模型。

[0031] 在一个实施例中，所述字段基于所述第一数据源中的字段相关信息和字段来源相关信息按照预定义的字段模式确定，所述字段模式包括字段的字段名称、字段数据类型、字段描述、标准代码、数据存储格式、哈希列、责任部门、数据来源系统的名称、数据来源系统的表名称、数据来源系统的字段名称和数据来源系统的字段类型。

[0032] 在一个实施例中，所述表模式包括表对象的表名称、主题域、二级主题域、表类型、表描述、责任部门、数据来源系统的名称、数据来源系统的表名称和字段列表。

[0033] 在一个实施例中，所述关系模式包括源表对象的表名称、目标表对象的表名称、源表对象与目标表对象之间的关联关系、源表对象与目标表对象之间的关联字段、主题域和二级主题域。

[0034] 在一个实施例中，基于所述第二数据源按照所述关系模式生成一对源表对象和目标表对象的表关系信息包括：对于所述第二数据源中的一对源表对象和目标表对象以及另一对源表对象和目标表对象，如果其各自的源表对象的表名称、目标表对象的表名称、源表

对象与目标表对象之间的关联关系、源表对象与目标表对象之间的关联字段都相同，则判定该一对源表对象和目标表对象之间的关系与该另一对源表对象和目标表对象的关系相同，对于相同的关系，只对其中的一个关系进行规范化处理并按照所述关系模式生成相应的源表对象和目标表对象的表关系信息。

[0035] 在一个实施例中，设有表对象及其字段的表模式的库，从所述表模式的库确定用于定义表对象及其字段的表模式。

[0036] 在一个实施例中，设有表示表对象之间的关系的关系模式的库，从所述关系模式的库确定用于定义源表对象和目标表对象之间的关系的关系模式。

[0037] 在一个实施例中，设有表对象、其字段、表对象之间的关系的别名集库，所述别名集库包括既往记录的别名及其出现频次，将所述第一数据源和所述第二数据源中出现的表对象、其字段、表对象之间的关系记录到所述别名库中，并将出现的频次累加；显示的表对象、其字段、表对象之间的关系为出现频次最大的表对象、其字段、表对象之间的关系。

[0038] 根据本发明的第三方面，提供一种计算机可读存储介质，其上存储有计算机程序，所述计算机程序在由处理器执行时导致上述的用于建立用于电网知识图谱的物理模型的方法被执行。

[0039] 根据本发明的方案，通过从多个数据源获取表对象、字段和关系的数据，对这些数据进行规范化处理并按照预定义的表模式、字段模式和关系模式建立用于电网知识图谱的统一的、完整的数据模型。利用本发明，可以对不同数据源进行知识抽取，对现有模型进行查缺补漏并弥补现有模型设计短板，同时可以给管理和业务人员提供更加合理的管控模型，支持公司统一数据模型的信息匹配、共享。另外，本发明可以基于现有数据模型进一步推进模型标准实施和构建完整的体系，为进一步推进数据质量管理打下坚实基础，同时支持数据中台与业务中台的建设，在实际应用中获得直接或间接效益。

附图说明

[0040] 以示例的方式参考以下附图描述本发明的非限制性且非穷举性实施例，其中：

[0041] 图1是示意性示出根据本发明一个实施例的用于建立用于电网知识图谱的物理模型的方法的流程图；

[0042] 图2是示意性示出根据本发明一个实施例的建立物理模型的物理表集的流程图；

[0043] 图3是示意性示出根据本发明一个实施例的建立物理模型的关系集的流程图；以及

[0044] 图4是示出根据本发明一个实施例的用于建立用于电网知识图谱的物理模型的系统的示意图。

具体实施方式

[0045] 为了使本发明的上述以及其他特征和优点更加清楚，下面结合附图进一步描述本发明。应当理解，本文给出的具体实施例是出于向本领域技术人员解释的目的，仅是示例性的，而非限制性的。

[0046] 作为本发明第一方面，提供一种用于建立用于电网知识图谱的物理模型的方法。图1示意性示出根据本发明一个实施例的用于建立用于电网知识图谱的物理模型的方法

S100。如图1所示，S100可以包括步骤S101、步骤S102、步骤S103、步骤S104、步骤S105和步骤S106。

[0047] 在步骤S101中，确定用于定义表对象及其字段的表模式。本文中表模式也可以被称为表定义，其用于限定表对象的组成成员，例如可以包括各种合适的用于将一个表对象与其他表对象区别开的组成成员。

[0048] 在一个实施例中，所述表模式可以包括表对象的表名称、主题域、二级主题域、表类型、表描述、责任部门、数据来源系统的名称、数据来源系统的表名称和字段列表。表名称可以包括表对象的表英文名称和表中文名称中的至少一个。例如，可以json格式确定表模式，如下：

[0049] {

[0050] 'name':[表名(英文),表名(中文)],

[0051] 'area':主题域,

[0052] 'secondary area':二级主题域,

[0053] 'type':表类型,

[0054] 'description':表描述,

[0055] 'department':责任部门,

[0056] 'source system':数据来源系统名,

[0057] 'source table':[数据来源系统表名(英文),数据来源系统表名(中文)],

[0058] 'more':备注,

[0059] 'fields':[字段列表]

[0060] }

[0061] 在一个实施例中，表模式中的字段列表基于所述第一数据源中的字段相关信息和字段来源相关信息按照预定义的字段模式确定，所述字段模式可以包括字段的字段名称、字段数据类型、字段描述、标准代码、数据存储格式、哈希列、责任部门、数据来源系统的名称、数据来源系统的表名称、数据来源系统的字段名称和数据来源系统的字段类型。本文中字段模式也可以被称为字段定义，其用于限定字段的组成成员。例如，可以json格式确定字段模式，如下：

[0062] {

[0063] 'name':[字段(英文),字段(中文)],

[0064] 'datatype':字段数据类型,

[0065] 'description':字段描述,

[0066] 'standard code':标准代码,

[0067] 'storage format':数据存储格式,

[0068] 'hash column':哈希列,

[0069] 'department':责任部门,

[0070] 'source system':[数据来源系统英文名,数据来源系统中文名],

[0071] 'source table':[数据来源系统表名(英文),数据来源系统表名(中文)],

[0072] 'source field':[数据来源系统字段(英文),数据来源系统字段(中文)],

[0073] 'source datatype':数据来源系统字段类型,

[0074] 'more':备注

[0075] }

[0076] 在步骤S102中,接收包括多个表对象的第一数据源,所述第一数据源包括表对象相关信息、字段相关信息、表对象来源相关信息和/或字段来源相关信息。这里,第一数据源可以被广义地理解为囊括各种可能形式的数据源,包括结构化、半结构化和非结构化形式的数据源,例如关系型数据库、数仓、非关系型数据库、文档库、各类报表等。优选地,本发明的第一数据源包括excel文档形式的数据源。

[0077] 在步骤S103中,对于每个表对象,基于所述第一数据源按照所述表模式生成相应的表信息,从而得到所述第一数据源包括的全部表对象的表信息集合,以生成包括所述表信息集合的所述物理模型的物理表集,其中,所述表信息至少指示表对象的表名称、字段、表对象来源和字段来源。应理解,对于每个表对象,可以将表名称作为表对象的识别标准,例如,将“表英文名称+表中文名称”作为表对象的识别标准、将“表英文名称”作为表对象的识别标准、或将“表中文名称”作为表对象的识别标准。根据需要,物理模型的物理表集可以各种合适的文件形式存储,例如json存储文件形式。在一个实施例中,物理模型的物理表集的json存储文件形式如下:

[0078]
```
{
    "name": [
        "DWD_AST_APP_TYPE",
        "应用软件分类"
    ],
    "area": "资产域",
    "secondary area": "040401:实物资产",
    "type": "哈希分布表",
```

[0079]

"description": "1）定义：应用软件分类\n2）业务用途：用于应用软件分类\n3）责任部门：国网互联网部\n4）所属域：资产域\n5）来源系统：I6000\n6）来源表：App_Type",
"department": "国网互联网部",
"source system": "信息通信一体化调度运行支撑平台",
"source table": [
"App_Type",
"应用软件分类"
],
"more": "",
"fields": [
{
"name": [
"ID",
"编号"
],
"datatype": "VARCHAR2(64)",
"description": "1）字段描述：编号\n2）来源系统：信息通信一体化调度运行支撑平台\n3）来源表：APP_TYPE\n4）来源字段：id ",
"standard code": "",
"storage format": "",
"hash column": "Y",
"department": "国网互联网部",
"source system": [
"I6000",
"信息通信一体化调度运行支撑平台"
],
"source table": [
"App_Type",
"应用软件分类"

```
        ],
        "source field": [
            "ID",
            "编号"
[0080]
        ],
        "source datatype": "VARCHAR2(64)",
        "more": ""
    }
```

[0081] 下面结合图2对步骤S103进行详细地描述。

[0082] 如图2所示，作为第一数据源的excel文档包括“数据表信息”、“数据表对照表”、“字段对照表”三个部分，其中“数据表信息”指示所有表对象的表名称和字段的相关信息，“数据表对照表”指示表对象的来源系统的表来源相关信息，“字段对照表”指示字段的来源系统的字段来源相关信息。来源系统可以例如包括各种可能的电力知识方面的数据平台。由于这些信息中存在表名称大小写不一致的问题，因此在建立模型的过程中采取大小写不敏感原则，将表的英文名称一律做大写处理。具体过程如下：首先将“表英文名称+表中文名称”作为表对象的判别标准，按“表英文名称+表中文名称”聚合得到每个表对象的表名称信息；其次，对于每个表对象，对其表名称进行规范化处理，例如包括去掉其中的空格和换行符等；再次，对该表对象的所有字段的字段名称、字段数据类型等进行规范化处理，例如包括去掉其中的空格和换行符等，按照字段定义整理得到一个规范化的字段，再基于“字段对照表”对该字段的信息进行补充，例如从“字段对照表”获取字段的来源系统的名称、该来源系统的表名称、该来源系统的字段名称、该来源系统的字段类型等并对这些信息进行规范化处理，再将规范化处理的信息整合到该字段信息中；当该表对象的所有字段处理完毕后，按照表定义生成该表对象的表及其字段信息；然后，基于“数据表对照表”对该表及其字段信息进行补充，例如从“数据表对照表”获取表对象的来源系统的名称、该数据来源系统的表名称等并对这些信息进行规范化处理，再将规范化处理的信息整合到该表对象的信息中；最后，将整合后的该表对象的所有信息整理成相应的表信息，该表信息可以由一个json串来表示。重复以上步骤，直到处理完所有的表对象，由此得到包含所有表对象的物理表集。该包含所有表对象的物理表集可以json存储文件导出。

[0083] 在步骤S104中，确定用于定义源表对象和目标表对象之间的关系的关系模式。本文中关系模式也可以被称为关系定义，其用于限定表对象对之间的关系的组成成员，例如可以表示源表对象与目标表对象之间的相互关联。源表对象的表名称和目标表对象的表名称可以包括相应的表英文名称和表中文名称中的至少一个。例如，可以json格式确定关系模式，如下：

[0084] {

[0085] 'entity1':[源表对象表名(英文),源表对象表名(中文)],

[0086] 'entity2':[目标表对象表名(英文),目标表对象表名(中文)],

[0087] 'relation':关联关系,

[0088] 'field':关联字段,

[0089] 'area':主题域,

[0090] 'secondary area':二级主题域

[0091] }

[0092] 在步骤S105中,接收包括源表对象和目标表对象之间的关系的关系相关信息的第二数据源,所述第二数据源包括多对源表对象和目标表对象,对于每对源表对象和目标表对象,基于所述第二数据源按照所述关系模式生成该对源表对象和目标表对象的表关系信息,从而得到所述第二数据源包括的全部关系的表关系信息集合,以生成包括所述表关系信息集合的所述物理模型的关系集。这里,第二数据源可以被广义地理解为囊括各种可能形式的数据源,包括结构化、半结构化和非结构化形式的数据源,例如关系型数据库、数仓、非关系型数据库、文档库、各类报表等。优选地,本发明的第二数据源包括excel文件形式的数据源。本发明的第一数据源和第二数据源可以包括基于电网下各业务系统的数据以及智能电网上采集到的时序数据,主要包括公司营销数据、量化采集数据、运检数据及一些图形化的图像网页数据,可以针对这些结构化、半结构化和非结构化的三种不同形式的数据进行处理、知识抽取和融合规范。根据需要,物理模型的关系集可以各种合适的文件形式存储,例如json存储文件形式。在一个实施例中,物理模型的关系集的json存储文件形式如下:

[0093]
```
{
    "entity1": [
        "DWD_FIN_SGACCOUNTINGDOCUMENT",
        "会计凭证"
    ],
    "entity2": [
        "DWD_FIN_SGCOMPANYCODEGLAC",
        "公司代码总账科目"
    ],
    "relation": "1..*-1",
    "field":
"DWD_FIN_SGACCOUNTINGDOCUMENT.SGACCDO_SGGLAC_ID=DWD
_FIN_SGCOMPANYCODEGLAC.SGCOMPANYCODEGLAC_ID",
    "area": "财务域",
    "secondary area": "040200:财力资源公共"
}
```

[0094] 在一个实施例中,步骤S105可以包括:对于所述第二数据源中的一对源表对象和目标表对象以及另一对源表对象和目标表对象,如果其各自的源表对象的表名称、目标表对象的表名称、源表对象与目标表对象之间的关联关系和源表对象与目标表对象之间的关联字段都相同,则判定该一对源表对象和目标表对象之间的关系与该另一对源表对象和目

标表对象的关系相同，对于相同的关系，只对其中的一个关系进行规范化处理并按照所述关系模式生成相应的源表对象和目标表对象的表关系信息。即，将源表对象的表名称、目标表对象的表名称、源表对象与目标表对象之间的关联关系和源表对象与目标表对象之间的关联字段作为识别一条关系的标识符。

[0095] 下面结合图3对步骤S105进行详细地描述。

[0096] 如图3所示，读取作为第二数据源的excel文档并获取其中的“关联关系”的数据。由于该“关联关系”的数据中存在表名大小写不一致的问题，因此在建立模型的过程中采取大小写不敏感原则，将表英文名称一律做大写处理。具体过程如下：以如下六项，即源表对象的表名(英文)、源表对象的表名(中文)、目标表对象的表名(英文)、目标表对象的关联表名(中文)、相应的关联关系、相应的关联字段作为每条关系的判别标准，针对每条关系将上述六项聚合以得到多个表关系的聚合标识符，将具有相同的聚合标识符的表关系判定为同一个聚合组；对于同一个聚合组的表关系，只针对其中的第一条数据(因此可以对表关系进行去重复以避免冗余)对其表名称、关联关系、关联字段等进行规范化处理，例如去掉其中的空格、换行符、多余的横杠和等号等；基于规范化处理的信息按照关系定义整理成相应的表关系信息，该表关系信息可以由一个json串来表示。重复以上步骤，直到处理完所有的表关系，由此得到包含所有表关系的关系集。该包含所有表关系的关系集可以json存储文件导出。

[0097] 在步骤S106中，基于所述物理模型的物理表集和所述物理模型的关系集，建立包括表对象、字段和关系的物理模型。

[0098] 在一个实施例中，本发明的方法还包括：基于物理模型的物理表集计算该物理模型中的表对象对之间的相似度，对相似度超过预定阈值的表对象对进行去重复处理以生成冗余度较低的物理模型的物理表集。这样的物理模型的物理表集可以与相应的逻辑模型进行匹配以实现模型的一致性检测，从而提高现有模型(例如，国家电网公司企业公共数据模型SG-CIM4.0)静态语义的合理性和完备性并有效地减少冗余，最终将难以观察的非空间知识数据转化为空间图谱，便于相关领域人员的认知和理解，为跨域实体的关联贯通提供了有效的解决方案。同时，可以使得知识图谱技术描述实体、属性与关系的这种强大的语义处理能力得到很好的体现。

[0099] 作为本发明第二方面，提供一种用于建立用于电网知识图谱的物理模型的系统。图4示意性示出根据本发明一个实施例的用于建立用于电网知识图谱的物理模型的系统200。系统200可以包括物理表集生成单元201、关系集生成单元202和处理单元203。处理单元203与物理表集生成单元201和关系集生成单元202通信地耦合。

[0100] 物理表集生成单元201可以被配置为：

[0101] 确定用于定义表对象及其字段的表模式；

[0102] 接收包括多个表对象的第一数据源，所述第一数据源包括表对象相关信息、字段相关信息、表对象来源相关信息和/或字段来源相关信息；

[0103] 对于每个表对象，基于所述第一数据源按照所述表模式生成相应的表信息，从而得到所述第一数据源包括的全部表对象的表信息集合，以生成包括所述表信息集合的所述物理模型的物理表集，其中，所述表信息至少指示表对象的表名称、字段、表对象来源和字段来源。

[0104] 关系集生成单元202可以被配置为：

[0105] 确定用于定义源表对象和目标表对象之间的关系的关系模式；

[0106] 接收包括源表对象和目标表对象之间的关系的关系相关信息的第二数据源，所述第二数据源包括多对源表对象和目标表对象，对于每对源表对象和目标表对象，基于所述第二数据源按照所述关系模式生成该对源表对象和目标表对象的表关系信息，从而得到所述第二数据源包括的全部关系的表关系信息集合，以生成包括所述表关系信息集合的所述物理模型的关系集。

[0107] 在一个实施例中，对于所述第二数据源中的一对源表对象和目标表对象以及另一对源表对象和目标表对象，如果其各自的源表对象的表名称、目标表对象的表名称、源表对象与目标表对象之间的关联关系、源表对象与目标表对象之间的关联字段都相同，则判定该一对源表对象和目标表对象之间的关系与该另一对源表对象和目标表对象的关系相同，对于相同的关系，只对其中的一个关系进行规范化处理并按照所述关系模式生成相应的源表对象和目标表对象的表关系信息。

[0108] 处理单元203可以被配置为：基于所述物理模型的物理表集和所述物理模型的关系集，建立包括表对象、字段和关系的物理模型。

[0109] 在一个实施例中，所述字段基于所述第一数据源中的字段相关信息和字段来源相关信息按照预定义的字段模式确定，所述字段模式包括字段的字段名称、字段数据类型、字段描述、标准代码、数据存储格式、哈希列、责任部门、数据来源系统的名称、数据来源系统的表名称、数据来源系统的字段名称和数据来源系统的字段类型。

[0110] 在一个实施例中，所述表模式包括表对象的表名称、主题域、二级主题域、表类型、表描述、责任部门、数据来源系统的名称、数据来源系统的表名称和字段列表。

[0111] 在一个实施例中，所述关系模式包括源表对象的表名称、目标表对象的表名称、源表对象与目标表对象之间的关联关系、源表对象与目标表对象之间的关联字段、主题域和二级主题域。

[0112] 应理解，本文中前述第一方面的关于用于建立用于电网知识图谱的物理模型的方法所描述的具体特征也可类似地应用于第二方面的用于建立用于电网知识图谱的物理模型的系统以进行类似扩展。为简化起见，未对其进行详细描述。

[0113] 应理解，本发明的用于建立用于电网知识图谱的物理模型的系统200的各个单元可全部或部分地通过软件、硬件、固件或其组合来实现。所述各单元各自可以硬件或固件形式内嵌于计算机设备的处理器中或独立于所述处理器，也可以软件形式存储于计算机设备的存储器中以供处理器调用来执行所述各单元的操作。所述各单元各自可以实现为独立的部件或模块，或者两个或更多个单元可实现为单个部件或模块。

[0114] 本领域普通技术人员应理解，图4中示出的系统200的示意图仅仅是与本发明的方案相关的部分结构的示例性说明框图，并不构成对体现本发明的方案的计算机设备、处理器或计算机程序的限定。具体的计算机设备、处理器或计算机程序可以包括比图中所示更多或更少的部件或模块，或者组合或拆分某些部件或模块，或者可具有不同的部件或模块布置。

[0115] 在本发明中，设有表对象及其字段的表模式的库，从所述表模式的库确定用于定义表对象及其字段的表模式。

[0116] 在本发明中，设有表示表对象之间的关系的关系模式的库，从所述关系模式的库确定用于定义源表对象和目标表对象之间的关系的关系模式。

[0117] 在本发明中，设有表对象、其字段和表对象之间的关系的别名集库，所述别名集库包括既往记录的别名及其出现频次，将所述第一数据源和所述第二数据源中出现的表对象、其字段、表对象之间的关系记录到所述别名库中，并将出现的频次累加；显示的表对象、其字段、表对象之间的关系为出现频次最大的表对象、其字段、表对象之间的关系。

[0118] 在一个优选的方案中，对于所述表对象别名集库、其字段的别名集库和表对象之间的关系的别名集库，对于每次记录设有标签，用于区分不同次采集。这样的话，可以对不同来源，例如不同部门的别名库进行合并，如果两个记录具有相同的标签，则认为它们来自同一次采集，不累计计算。所述标签例如包括日期、时间、随机序列。日期采用8位数模式例如20201030，时间精确到分或秒，例如1830或183025，随机序列例如是6-10位的随机数字，用于校验。通过记录采集日期可以跟踪表对象、其字段和表对象之间的关系的名称变迁，一般显示的是最流行、最大规模使用的名称，对统一名称有规范作用。

[0119] 作为本发明第三方面，提供一种计算机可读存储介质，其上存储有计算机程序，所述计算机程序在被处理器执行时实现本发明第一方面的方法的步骤。在一个实施例中，所述计算机程序被分布在网络耦合的多个计算机设备或处理器上，以使得所述计算机程序由一个或多个计算机设备或处理器以分布式方式存储、访问和执行。单个方法步骤/操作，或者两个或更多个方法步骤/操作，可以由单个计算机设备或处理器或由两个或更多个计算机设备或处理器执行。一个或多个方法步骤/操作可以由一个或多个计算机设备或处理器执行，并且一个或多个其他方法步骤/操作可以由一个或多个其他计算机设备或处理器执行。一个或多个计算机设备或处理器可以执行单个方法步骤/操作，或执行两个或更多个方法步骤/操作。

[0120] 本领域普通技术人员可以理解，本发明的用于建立用于电网知识图谱的物理模型的方法的全部或部分步骤可以通过计算机程序来指示相关的硬件如计算机设备或处理器完成，所述的计算机程序可存储于非暂时性计算机可读存储介质中，该计算机程序被执行时实现本发明的辅助方法的步骤。根据情况，本文中对存储器、存储、数据库或其它介质的任何引用可包括非易失性和/或易失性存储器。非易失性存储器的示例包括只读存储器(ROM)、可编程ROM(PROM)、电可编程ROM(EPROM)、电可擦除可编程ROM(EEPROM)、闪存、磁带、软盘、磁光数据存储装置、光学数据存储装置、硬盘、固态盘等。易失性存储器的示例包括随机存取存储器(RAM)、外部高速缓冲存储器等。

[0121] 以上描述的各技术特征可以任意地组合。尽管未对这些技术特征的所有可能组合进行描述，但这些技术特征的任何组合都应当被认为由本说明书涵盖，只要这样的组合不存在矛盾。

[0122] 尽管结合实施例对本发明进行了描述，但本领域技术人员应理解，上文的描述和附图仅是示例性而非限制性的，本发明不限于所公开的实施例。在不偏离本发明的精神的情况下，各种改型和变体是可能的。

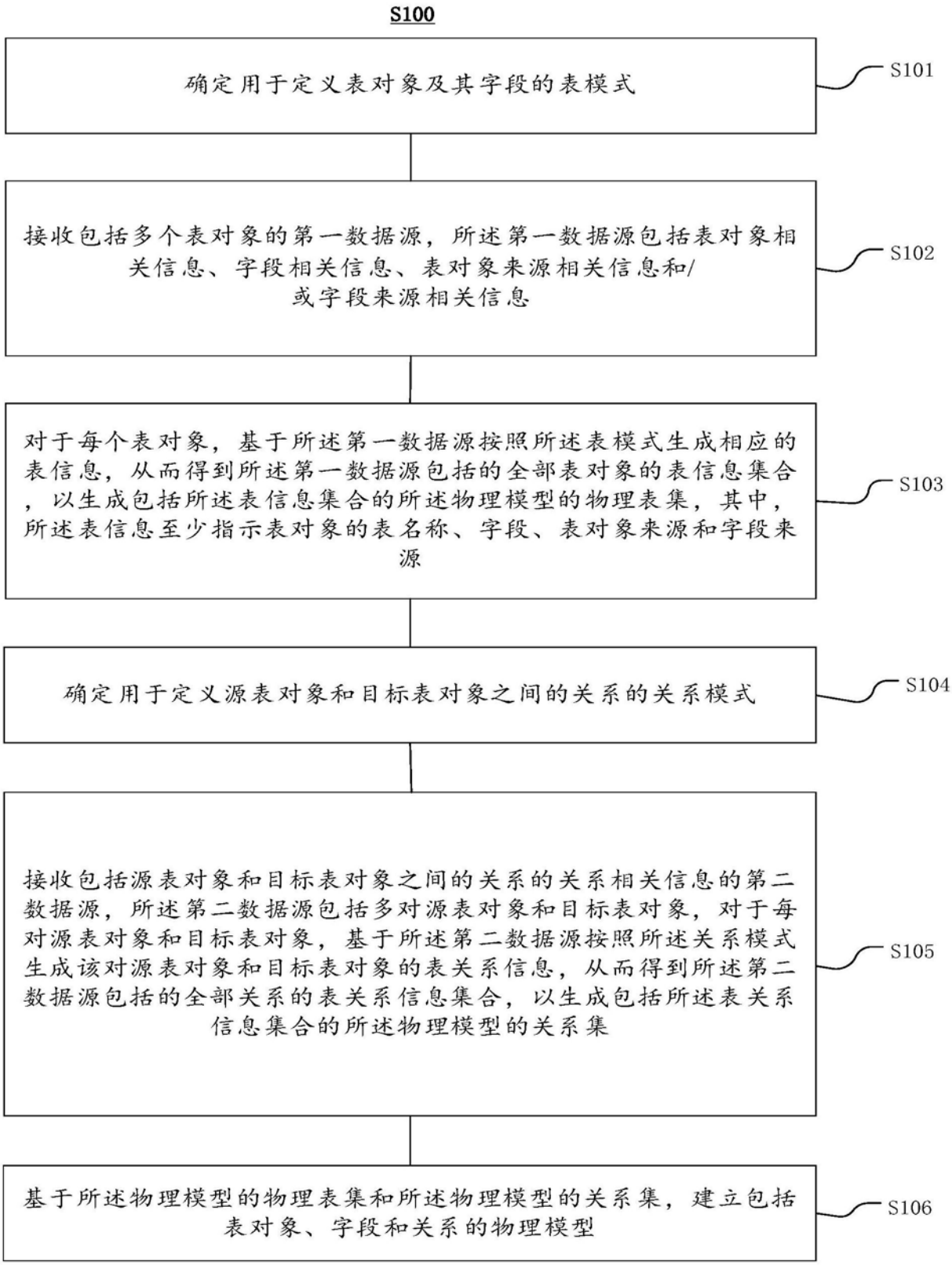

S100
确定用于定义表对象及其字段的表模式
S101
接收包括多个表对象的第一数据源，所述第一数据源包括表对象相关信息、字段相关信息、表对象来源相关信息和/或字段来源相关信息
S102
对于每个表对象，基于所述第一数据源按照所述表模式生成相应的表信息，从而得到所述第一数据源包括的全部表对象的表信息集合，以生成包括所述表信息集合的所述物理模型的物理表集，其中，所述表信息至少指示表对象的表名称、字段、表对象来源和字段来源
S103
确定用于定义源表对象和目标表对象之间的关系的关系模式
S104
接收包括源表对象和目标表对象之间的关系的关系相关信息的第二数据源，所述第二数据源包括多对源表对象和目标表对象，对于每对源表对象和目标表对象，基于所述第二数据源按照所述关系模式生成该对源表对象和目标表对象的表关系信息，从而得到所述第二数据源包括的全部关系的表关系信息集合，以生成包括所述表关系信息集合的所述物理模型的关系集
S105
基于所述物理模型的物理表集和所述物理模型的关系集，建立包括表对象、字段和关系的物理模型
S106

图1

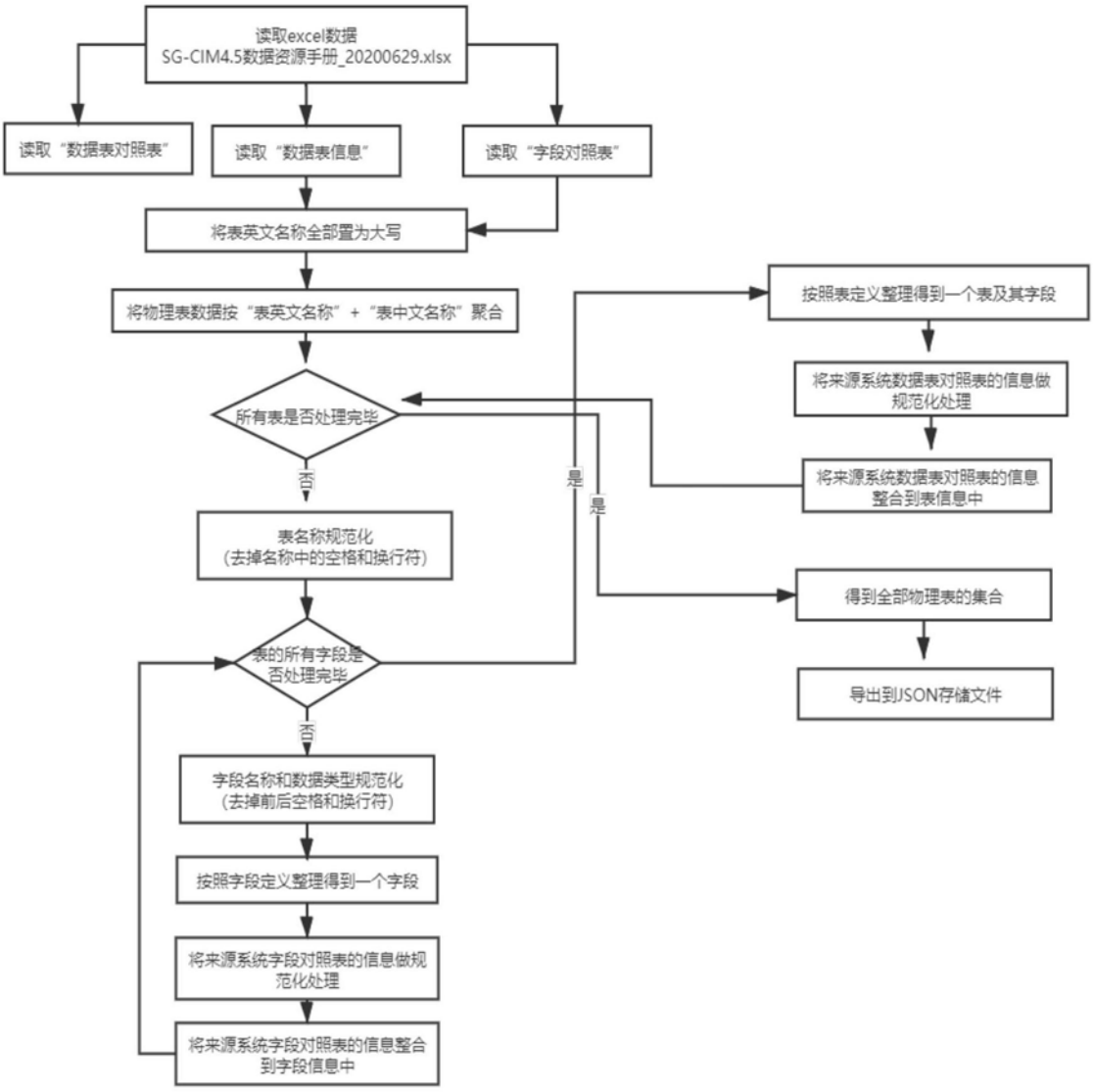

读取excel数据
SG-CIM4.5数据资源手册_20200629.xlsx
读取“数据表对照表”
读取“数据表信息”
读取“字段对照表”
将表英文名称全部置为大写
将物理表数据按“表英文名称”+“表中文名称”聚合
所有表是否处理完毕
否
表名称规范化
（去掉名称中的空格和换行符）
表的所有字段是否处理完毕
否
字段名称和数据类型规范化
（去掉前后空格和换行符）
按照字段定义整理得到一个字段
将来源系统字段对照表的信息做规范化处理
将来源系统字段对照表的信息整合到字段信息中
是
是
按照表定义整理得到一个表及其字段
将来源系统数据表对照表的信息做规范化处理
将来源系统数据表对照表的信息整合到表信息中
得到全部物理表的集合
导出到JSON存储文件

图2

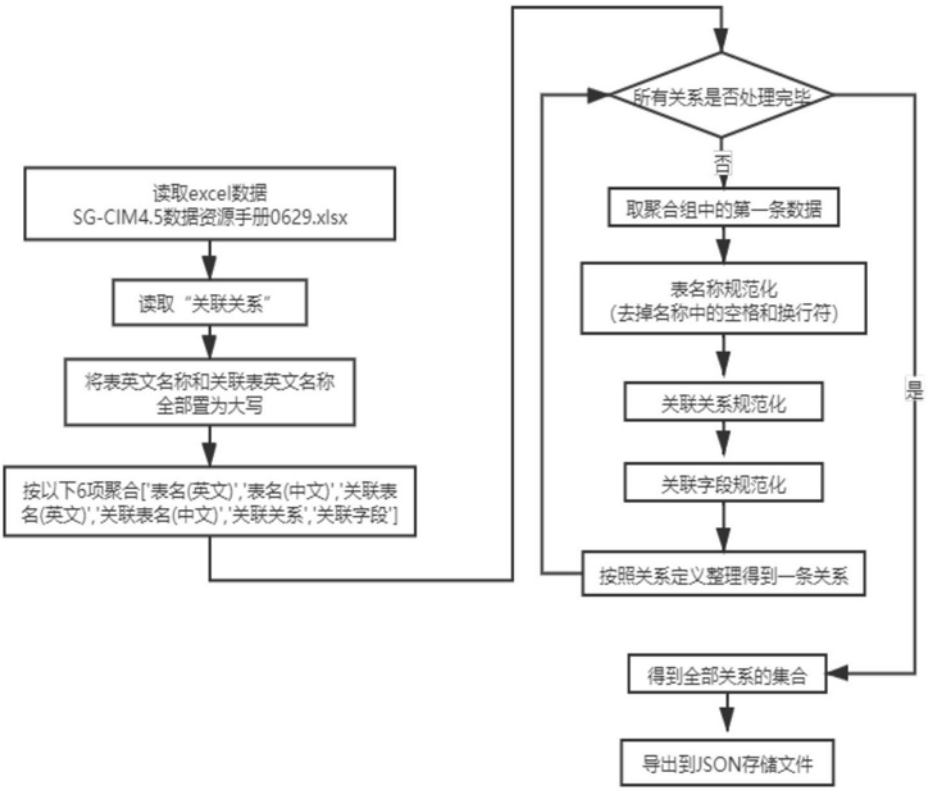

读取excel数据
SG-CIM4.5数据资源手册0629.xlsx
读取“关联关系”
将表英文名称和关联表英文名称
全部置为大写
按以下6项聚合['表名(英文)','表名(中文)','关联表名(英文)','关联表名(中文)','关联关系','关联字段']
所有关系是否处理完毕
否
是
取聚合组中的第一条数据
表名称规范化
(去掉名称中的空格和换行符)
关联关系规范化
关联字段规范化
按照关系定义整理得到一条关系
得到全部关系的集合
导出到JSON存储文件

图3

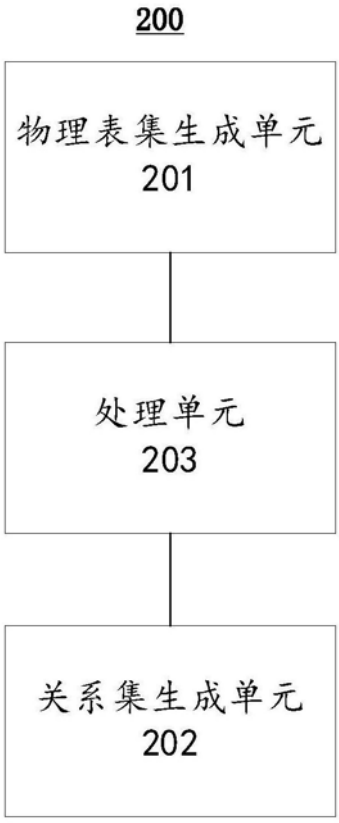


图4