(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 113535739 A
(43)申请公布日 2021.10.22

(21)申请号 202111088634.3
(22)申请日 2021.09.16
(71)申请人 国网浙江省电力有限公司信息通信分公司
地址 310007 浙江省杭州市黄龙路8号641室
申请人 国网浙江省电力有限公司
国网浙江综合能源服务有限公司
(72)发明人 黄宇腾 王红凯 夏洪涛 倪阳旦 孔晓昀 应张驰 潘司晨
(74)专利代理机构 杭州华鼎知识产权代理事务所(普通合伙) 33217
代理人 魏亮
(51)Int.Cl.
G06F 16/22(2019.01)
G06F 16/28(2019.01)
G06F 16/215(2019.01)
G06Q 10/10(2012.01)
G06Q 50/06(2012.01)

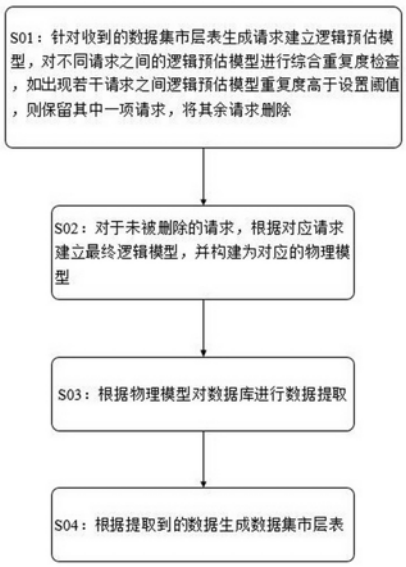
权利要求书1页 说明书6页 附图2页

(54)发明名称

一种基于电网能源数据的数据集市层表建立方法

(57)摘要

本发明公开了一种基于电网能源数据的数据集市层表建立方法,包括以下步骤:S01:针对收到的数据集市层表生成请求建立逻辑预估模型,对不同请求之间的逻辑预估模型进行综合重复度检查,如出现若干请求之间的逻辑预估模型重复度高于设置阈值,则保留其中一项请求,将其余请求删除;S02:对于未被删除的请求,根据该请求建立最终逻辑模型,并构建为对应的物理模型;S03:根据物理模型对数据库进行数据提取;S04:根据提取到的数据生成数据集市层表。本发明在任务前期通过判断实体重复度以及关系重复度,并汇总为综合重复度,以识别出内容相似的请求并删除,避免创建重复度过高的数据集市层表,防止运算资源浪费,提高数据处理效率。



CN 113535739 A

1.一种基于电网能源数据的数据集市层表建立方法,其特征在于,包括以下步骤:

S01:针对收到的数据集市层表生成请求建立逻辑预估模型,对不同请求之间的逻辑预估模型进行综合重复度检查,如出现若干请求之间的逻辑预估模型重复度高于设置阈值,则保留其中一项请求,将其余请求删除;

S02:对于未被删除的请求,根据该请求建立最终逻辑模型,并构建为对应的物理模型;

S03:根据物理模型对数据库进行数据提取;

S04:根据提取到的数据生成数据集市层表;

其中逻辑预估模型包含逻辑实体及对应的关联关系,所述综合重复度检查的过程包括:每次对比两个不同逻辑预估模型中的逻辑实体,计算重复的逻辑实体数与逻辑实体总数的比值,得到实体重复度;将这两个不同逻辑预估模型中的关联关系分别进行图像转化,然后对图像转化得到的关系图像进行比对,得到的相似度值为关系重复度;将关系重复度与实体重复度带入公式中计算,得到综合重复度。

2.根据权利要求1所述的一种基于电网能源数据的数据集市层表建立方法,其特征在于,所述逻辑预估模型的建立过程包括:将数据集市层表生成请求中涉及的逻辑实体随机分为两组,遍历其中一组中各逻辑实体与另一组中各逻辑实体的关联关系并记录,得到逻辑预估模型。

3.根据权利要求2所述的一种基于电网能源数据的数据集市层表建立方法,其特征在于,所述逻辑实体包括用户台账、计量点、计量点日用电量统计、计量点日用水量统计、计量点日用气量统计、计量点日用油量统计和/或计量点日用煤量统计。

4.根据权利要求3所述的一种基于电网能源数据的数据集市层表建立方法,其特征在于,所述关系重复度的获得过程包括:根据每种逻辑实体存在的关联关系的数量,对逻辑预估模型中的逻辑实体进行升序或降序排序并编号;在极坐标系中,从0度开始正向每隔一定度数确定一个极角并对应一个编号,并将该编号对应的逻辑实体拥有的关联关系数量作为极径,在极坐标中描点;待全部描点完成后,将相邻的点连接得到曲线,作为关系图像;随后,利用关系图像进行图形相似度对比,得到的相似度值为关系重复度。

5.根据权利要求3所述的一种基于电网能源数据的数据集市层表建立方法,其特征在于,所述综合重复度由以下公式计算得到:

$T=ix+(1-i)y$;

其中,x为实体重复度,y为关系重复度,i为比例系数,T为综合重复度。

6.根据权利要求1所述的一种基于电网能源数据的数据集市层表建立方法,其特征在于,所述物理模型的构建过程包括:确定所述最终逻辑模型中逻辑实体对应的数据结构,并整合至逻辑实体中,作为物理模型的数据域。

7.根据权利要求6所述的一种基于电网能源数据的数据集市层表建立方法,其特征在于,所述根据物理模型对数据库进行数据提取,包括:根据逻辑实体对应的数据结构,从数据库中进行数据提取。

8.根据权利要求6所述的一种基于电网能源数据的数据集市层表建立方法,其特征在于,所述根据提取到的数据生成数据集市层表,包括:将任务名称作为表头,数据实体的名称作为属性,逻辑实体对应的数据结构作为描述,数据实体的填充数据作为内容,生成数据集市层表。

一种基于电网能源数据的数据集市层表建立方法

技术领域

[0001] 本发明涉及数据处理领域，特别涉及一种基于电网能源数据的数据集市层表建立方法。

背景技术

[0002] 在能源大数据中心建设过程中，累计了大量公共能源数据模型，按照能源大数据中心的标准架构，数据集市层表需要基于公共能源数据模型建设。目前公共能源数据模型在规范使用方面仍存在一些问题，导致上层的数据集市层表无法规范建设，亟需开展公共能源数据模型可视化设计方法探索与开发，进而提升数据集市层表的建设规范性。

[0003] 现有技术中，如公开号CN108959356A的发明公开了一种智能配用电大数据应用系统数据集市建立方法，将用户用电数据、电网运行数据等一系列基础数据通过不同的数据分析和挖掘手段，建成数据集市，为各个应用模块提供应用数据。数据集市基于数据处理为核心，进行数据采集、数据存储、数据清洗、数据分析，最终实现数据分析结果可视化展示的目的，实现大数据应用系统对数据的高速查询和检索。

[0004] 电网的各业务部门在根据公共能源数据模型来建立数据集市层表的过程中，由于业务领域相近，经常会出现数据集市层表重复度过高的问题，导致计算资源的浪费，降低了数据处理效率，因此为了提高系统的数据处理效率，避免相似信息的重复处理显得尤为重要，而相比于一般的数据信息，数据集市层表通常涉及逻辑实体和实体之间的关联关系，因此简单地对比名称或字词并不能准确获知重复情况。

发明内容

[0005] 针对现有技术中相似信息的重复处理导致数据处理效率低下的问题，本发明提供了一种基于电网能源数据的数据集市层表建立方法，在任务前期通过判断实体重复度以及关系重复度，并汇总为综合重复度，以识别出内容相似的请求并删除，避免创建重复度过高的数据集市层表，防止运算资源浪费，提高数据处理效率。

[0006] 以下是本发明的技术方案。

[0007] 一种基于电网能源数据的数据集市层表建立方法，包括以下步骤：

S01：针对收到的数据集市层表生成请求建立逻辑预估模型，对不同请求之间的逻辑预估模型进行综合重复度检查，如出现若干请求之间的逻辑预估模型重复度高于设置阈值，则保留其中一项请求，将其余请求删除；

S02：对于未被删除的请求，根据该请求建立最终逻辑模型，并构建为对应的物理模型；

S03：根据物理模型对数据库进行数据提取；

S04：根据提取到的数据生成数据集市层表；

其中逻辑预估模型包含逻辑实体及对应的关联关系，所述综合重复度检查的过程包括：每次对比两个不同逻辑预估模型中的逻辑实体，计算重复的逻辑实体数与逻辑实体

总数的比值，得到实体重复度；将这两个不同逻辑预估模型中的关联关系分别进行图像转化，以去除逻辑实体的影响，然后对图像转化得到的关系图像进行比对，得到的相似度值为关系重复度；将关系重复度与实体重复度带入公式中计算，得到综合重复度。

[0008] 其中逻辑实体在本发明中一般为公共能源数据模型中的业务实体，而逻辑实体之间如果存在相同属性则认为这些逻辑实体之间具有关联关系。实体重复度的对比采用的方式较为普遍，可以用词向量的形式进行对比来实现；而关系重复度需要将关联关系的表示方式转化为图像后进行对比得到，以去除逻辑实体不同带来的影响，最后进行逻辑实体和关联关系的汇总，得到综合重复度。

[0009] 本发明主要用于模型开发前期的数据集市层表生成过程，通过综合重复度检查判断数据集市层表生成请求的重复情况，引入了对实体重复度以及关系重复度的判别，以适应逻辑模型或数据集市所特有的逻辑实体以及关联关系的特点。通过将重复的请求删除可以在源头上阻止重复度过高的任务，避免出现重复计算，提高数据处理效率。

[0010] 作为优选，所述逻辑预估模型的建立过程包括：将数据集市层表生成请求中涉及的逻辑实体随机分为两组，遍历其中一组中各逻辑实体与另一组中各逻辑实体的关联关系并记录，得到逻辑预估模型。一般的逻辑模型至少包括了逻辑实体以及逻辑实体之间的完整关联关系，但由于建立逻辑预估模型的目的并不是确定最终的逻辑模型，而是为了方便后续进行的综合重复度检查，因此这里采用的逻辑预估模型是一种简化版的逻辑模型，具有完整的逻辑实体但关联关系仅考虑一部分，即这里关联关系仅是示意性质的，用于为数据集市层表生成请求建立一个初步的框架，以便于与其他数据对比。

[0011] 作为优选，所述逻辑实体包括用户台账、计量点、计量点日用电量统计、计量点日用水量统计、计量点日用气量统计、计量点日用油量统计和/或计量点日用煤量统计。

[0012] 作为优选，所述关系重复度的获得过程包括：根据每种逻辑实体存在的关联关系的数量，对逻辑预估模型中的逻辑实体进行升序或降序排序并编号；在极坐标系中，从0度开始正向每隔一定度数确定一个极角并对应一个编号，并将该编号对应的逻辑实体拥有的关联关系数量作为极径，在极坐标中描点；待全部描点完成后，将相邻的点连接得到曲线，作为关系图像；随后，利用关系图像进行图形相似度对比，得到的相似度值为关系重复度。在不同的模型中，逻辑实体之间的关联关系可能会有不同，前述的实体重复度已经表示了逻辑实体的重复程度，因此关联关系的重复程度需要尽可能剔除逻辑实体的影响，仅保留关联关系本身；在图像转化的过程中，逻辑实体之间的区别被消除，生成的关系图像中，曲线上的点的极角对应逻辑实体的编号，点的极径对应了该逻辑实体对应的关联关系数量，通过对比曲线的相似度，可以得到关联关系的相似程度，从而获得关系重复度。另外这种关系图像相比于柱形图，作图过程灵活性更高，调整余地大且更容易进行相似性的区分。

[0013] 作为优选，所述综合重复度由以下公式计算得到：

$$T=ix+(1-i)y;$$

其中，x为实体重复度，y为关系重复度，i为比例系数，T为综合重复度。

[0014] 作为优选，所述物理模型的构建过程包括：确定所述最终逻辑模型中逻辑实体对应的数据结构，并整合至逻辑实体中，作为物理模型的数据域。

[0015] 作为优选，所述根据物理模型对数据库进行数据提取，包括：根据逻辑实体对应的数据结构，从数据库中进行数据提取。

[0016] 作为优选，所述根据提取到的数据生成数据集市层表，包括：将任务名称作为表头，数据实体的名称作为属性，逻辑实体对应的数据结构作为描述，数据实体的填充数据作为内容，生成数据集市层表。

[0017] 本发明的实质效果包括：利用逻辑实体和关联关系的特殊性，分别通过实体重复度以及关系重复度，对数据集市层表生成请求的重复度进行审查，删除重复的请求，防止后续有重复内容过多的任务出现，提高数据处理效率。

附图说明

[0018] 图1是本发明实施例的流程图；

图2是本发明实施例的关系图像示意图。

具体实施方式

[0019] 下面将结合实施例，对本申请的技术方案进行描述。另外，为了更好的说明本发明，在下文中给出了众多的具体细节。本领域技术人员应当理解，没有某些具体细节，本发明同样可以实施。在一些实例中，对于本领域技术人员熟知的方法、手段、元件和电路未做详细描述，以便于凸显本发明的主旨。另外对于数据集市层表的开发环境或开发方式，可以采用直接编程的方式进行，也可以采用无代码开发平台，通过拖拉预先封装好的模块的方式来进行。应当理解，在本发明中，“多个”是指两个或两个以上。“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，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。字符“/”一般表示前后关联对象是一种“或”的关系。“包含A、B和C”、“包含A、B、C”是指A、B、C三者都包含，“包含A、B或C”是指包含A、B、C三者之一，“包含A、B和/或C”是指包含A、B、C三者中任1个或任2个或3个。

[0020] 实施例：

一种基于电网能源数据的数据集市层表建立方法，如图1所示，包括以下步骤：

S01：针对收到的数据集市层表生成请求建立逻辑预估模型，对不同请求之间的逻辑预估模型进行综合重复度检查，如出现若干请求之间逻辑预估模型重复度高于设置阈值，则保留其中一项请求，将其余请求删除。其中数据集市层表生成请求由各部门的用户终端或用户模块发出，数据中心的服务器接收对应请求并进行综合重复度检查及后续工作。

[0021] 逻辑预估模型的建立过程包括：将数据集市层表生成请求中涉及的逻辑实体随机分为两组，遍历其中一组中各逻辑实体与另一组中逻辑实体的关联关系并记录，得到逻辑预估模型。一般的逻辑模型至少包括了逻辑实体以及逻辑实体之间的完整关联关系，但由于建立逻辑预估模型的目的并不是确定最终的逻辑模型，而是为了方便后续进行的综合重复度检查，因此这里采用的逻辑预估模型是一种简化版的逻辑模型，具有完整的逻辑实体但关联关系仅考虑一部分，即这里关联关系仅是示意性质的，用于为数据集市层表生成请求建立一个初步的框架，以便于与其他数据对比。

[0022] 以某能源大数据中心的能源公共数据模型为例，需要开发“某企业日用能情况”相关的能源数据逻辑模型或数据集市层表，则逻辑实体包括了能源公共数据模型中的用户台账、计量点、计量点日用电量统计、计量点日用水量统计、计量点日用气量统计、计量点日用油量统计、计量点日用煤量等数据业务实体。关联关系需结合实体的属性，即存在相同属

性的逻辑实体之间具有关联关系，例如，因用户台账实体与计量点台账实体中，均存在“用户编号”属性，其中用户台账中的“用户编号”为主键，计量点中的“用户编号”为外键；计量点实体与计量点用能量系列实体中，均存在“计量点编号”属性，其中计量点中的“计量点编号”为主键，计量点用能量系列实体中的“计量点编号”为外键。以主键和外键为依据，生成关联关系。在本实施例中，对每个逻辑实体对应的关联关系数量进行记录。

[0023] 综合重复度检查的过程包括：每次对比两个不同逻辑预估模型中的逻辑实体，计算重复的逻辑实体数与逻辑实体总数的比值，得到实体重复度；将这两个不同逻辑预估模型中的关联关系分别进行图像转化，然后对图像转化得到关系图像进行比对，得到的相似度值为关系重复度；将关系重复度与实体重复度带入公式中计算，得到综合重复度。其中实体重复度的对比采用的方式较为普遍，可以用词向量的形式进行对比来实现；而关系重复度需要将关联关系的表示方式转化为图像后进行对比得到，以去除逻辑实体不同带来的影响，最后进行逻辑实体和关联关系的汇总，得到综合重复度。

[0024] 关系重复度的获得过程包括：根据每种逻辑实体存在的关联关系的数量，对逻辑预估模型中的逻辑实体进行升序或降序排序并编号；在极坐标系中，从0度开始正向每隔一定度数确定一个极角并对应一个编号，并将该编号对应的逻辑实体拥有的关联关系数量作为极径，在极坐标中描点，如图2所示；待全部描点完成后，将相邻的点连接得到曲线，作为关系图像；随后，利用关系图像进行图形相似度对比，得到的相似度值为关系重复度。

[0025] 在不同的模型中，逻辑实体之间的关联关系可能会有不同，前述的实体重复度已经表示了逻辑实体的重复程度，因此关联关系的重复程度需要尽可能剔除逻辑实体的影响，仅保留关联关系本身；在图像转化的过程中，逻辑实体之间的区别被消除，生成的关系图像中，曲线上的点的极角对应逻辑实体的编号，点的极径对应了该逻辑实体对应的关联关系数量，通过对比曲线的相似度，可以得到关联关系的相似程度，从而获得关系重复度。另外这种关系图像相比于柱形图，作图过程灵活性更高，调整余地大且更容易进行相似性的区分。

[0026] 综合重复度由以下公式即综合重复度指标函数计算得到：

$$T=ix+(1-i)y;$$

其中，x为实体重复度，y为关系重复度，i为比例系数，T为综合重复度。

[0027] S02：对于未被删除的请求，根据对应请求建立最终逻辑模型，并构建为对应的物理模型。其中物理模型的构建过程包括：确定最终逻辑模型中逻辑实体对应的数据结构，并整合至逻辑实体中，作为物理模型的数据域。物理模型的构建主要是根据逻辑实体的属性，将关联关系转换到映射关系中，仍然以某能源大数据中心的能源公共数据模型为例，第一步，将类映射为表：将类分为两种，一般类及继承类，一般类直接转换成表，继承类中父类不转换为数据库中的表，父类属性下落到子类，子类转换成数据库物理表，父类属性存储在子类映射成数据库物理表中对应的字段。继承类情况以断路器和空气开关为例，断路器是开关的父类，即断路器是开关的一个子集，那么断路器作为父类不独立建表，其属性作为一个字段标签存储在开关的信息表中，用于描述该开关是否为断路器。

[0028] 第二步，将属性映射为字段：首先判断字段数据类型，后续按照字段类型进行转换，字段数据类型分为普通数据类型、复合数据类型及枚举类型，普通属性数据类型直接转换成通用数据类型；复合数据类型，默认为字符串类型；针对枚举类型，直接作为String类

型转换成通用数据类型，生成一张公共码表，存放枚举值。所述码表字段包括枚举字段、初始值、名称和别名，所述码表内容为枚举类属性。

[0029] 第三步，进行能源数据对应实体关系的映射：例如用户台账实体与计量点台账实体是一对多关系，一个用户可能有一个或者多个计量点，那就将用户台账中的主键放在计量点台账中做外键，例如计量点台账实体与计量点用电量实体是一对一关系，则以计量点台账中的主键放到计量点用电量实体中做外键，例如用户台账实体与企业统一信用代码实体是多对多关系，则开发中间表，将两表的主键均放到中间表中，共用两个类表的属性作为主键。

[0030] S03：根据物理模型对数据库进行数据提取。根据逻辑实体对应的数据结构，从数据库中进行数据提取。

[0031] S04：根据提取到的数据生成数据集市层表。将任务名称作为表头，数据实体的名称作为属性，逻辑实体对应的数据结构作为描述，数据实体的填充数据作为内容，生成数据集市层表。在物理模型的字段信息、字段类型均确定之后，数据集市层表的结构已经确定，即可调用工具，生成数据集市层表建表语句，导入数据库后即可自动生成表结构，得到数据集市层表。另外由于数据集市层表是由能源公共信息模型裁剪而成，因此本实施例也能够按照映射关系生成能源公共信息模型向数据集市层表的数据集成语句。

[0032] 本实施例主要用于模型开发前期的数据集市层表生成过程，通过综合重复度检查判断数据集市层表生成请求与现有模型的重复情况，引入了对实体重复度以及关系重复度的判别，以适应逻辑模型或数据集市所特有的逻辑实体以及关联关系的特点。随后在通过重复度检查的前提下进行后续的处理过程。可以在源头上阻止重复度过高的任务，避免出现与已有模型撞车。

[0033] 本实施例的实质效果包括：利用逻辑实体和关联关系的特殊性，分别通过实体重复度以及关系重复度，对数据集市层表生成请求的重复度进行审查，删除重复的请求，防止重复内容过多的任务出现，提高数据处理效率。

[0034] 通过以上实施方式的描述，所属领域的技术人员可以了解到，为描述的方便和简洁，仅以上述各功能模块的划分进行举例说明，实际应用中可以根据需要而将上述功能分配由不同的功能模块完成，即将具体装置的内部结构划分成不同的功能模块，以完成以上描述的全部或者部分功能。

[0035] 在本申请所提供的实施例中，应该理解到，所揭露的结构和方法，可以通过其它的方式实现。例如，以上所描述的关于结构的实施例仅仅是示意性的，例如，模块或单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，例如多个单元或组件可以结合或者可以集成到另一个结构，或一些特征可以忽略，或不执行。另一点，所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口，结构或单元的间接耦合或通信连接，可以是电性，机械或其它的形式。

[0036] 作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是一个物理单元或多个物理单元，即可以位于一个地方，或者也可以分布到多个不同地方。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

[0037] 另外，在本申请实施例中的各功能单元可以集成在一个处理单元中，也可以是各

个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现，也可以采用软件功能单元的形式实现。

[0038] 集成的单元如果以软件功能单元的形式实现并作为独立的产品销售或使用时，可以存储在一个可读取存储介质中。基于这样的理解，本申请实施例的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的全部或部分可以以软件产品的形式体现出来，该软件产品存储在一个存储介质中，包括若干指令用以使得一个设备(可以是单片机，芯片等)或处理器(processor)执行本申请各个实施例方法的全部或部分步骤。而前述的存储介质包括：U盘、移动硬盘、只读存储器(read only memory，ROM)、随机存取存储器(random access memory，RAM)、磁碟或者光盘等各种可以存储程序代码的介质。

[0039] 以上内容，仅为本申请的具体实施方式，但本申请的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本申请揭露的技术范围内，可轻易想到变化或替换，都应涵盖在本申请的保护范围之内。因此，本申请的保护范围应以权利要求的保护范围为准。

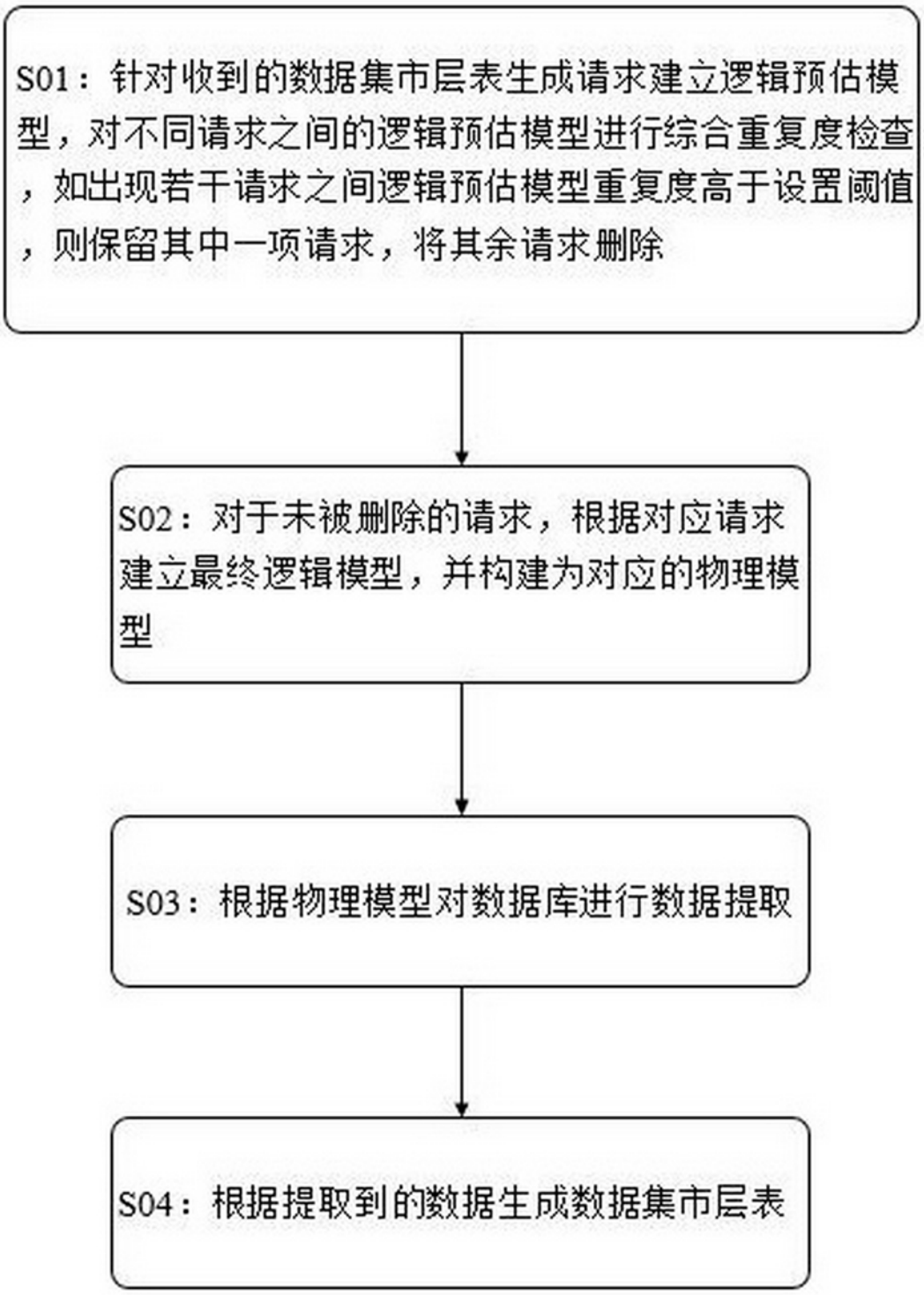

S01：针对收到的数据集市层表生成请求建立逻辑预估模型，对不同请求之间的逻辑预估模型进行综合重复度检查，如出现若干请求之间逻辑预估模型重复度高于设置阈值，则保留其中一项请求，将其余请求删除
S02：对于未被删除的请求，根据对应请求建立最终逻辑模型，并构建为对应的物理模型
S03：根据物理模型对数据库进行数据提取
S04：根据提取到的数据生成数据集市层表

图1

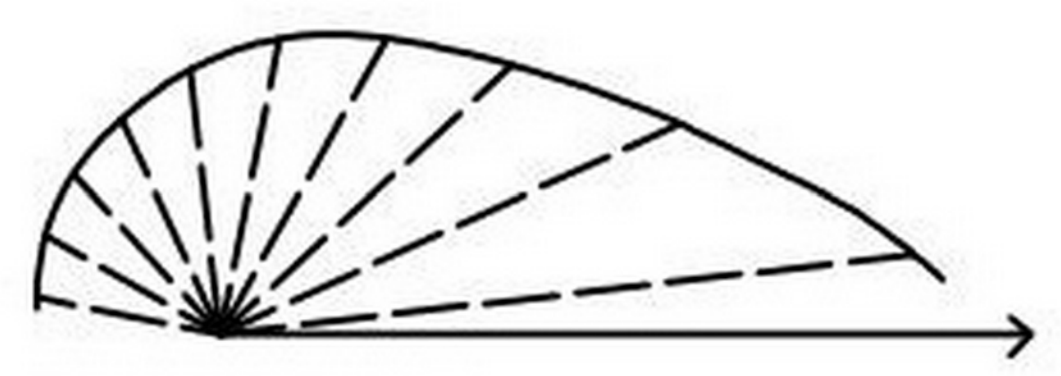


图2