(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 110795524 A
(43)申请公布日 2020.02.14

(21)申请号 201911052600.1

(22)申请日 2019.10.31

(71)申请人 北京东软望海科技有限公司
地址 100176 北京市大兴区北京经济技术开发区宏达北路12号B楼二区夹07室

(72)发明人 龙乐乐

(74)专利代理机构 北京市立方律师事务所 11330
代理人 张筱宁

(51)Int.Cl.
G06F 16/30(2019.01)
G06F 16/24(2019.01)

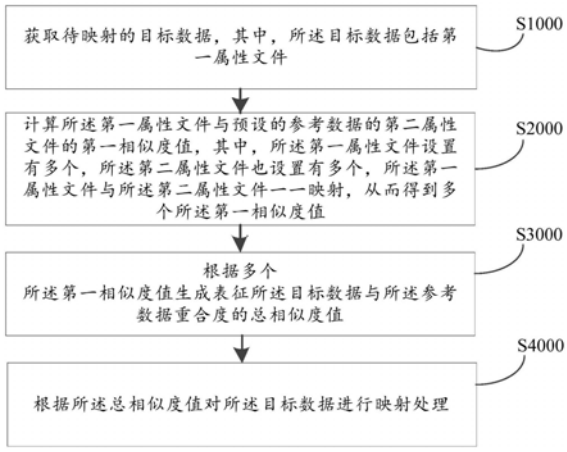
权利要求书2页 说明书11页 附图6页

(54)发明名称

主数据映射处理方法、装置、计算机设备及存储介质

(57)摘要

本申请公开了一种主数据映射处理方法、装置、计算机设备及存储介质，包括，获取待映射的目标数据，其中，目标数据包括第一属性文件；计算第一属性文件与预设的参考数据的第二属性文件的第一相似度值，其中，第一属性文件设置有多个，第二属性文件也设置有多个，第一属性文件与第二属性文件一一映射，从而得到多个第一相似度值；根据多个第一相似度值生成表征目标数据与参考数据重合度的总相似度值；根据总相似度值对目标数据进行映射处理。本申请通过对不同的属性文件分别进行相似度计算，再计算总的相似度，使获得的总相似度值，使相似度计算更为客观，消除人为干扰，且映射方式更方便、快捷。



CN 110795524 A

1.一种主数据映射处理方法,其特征在于,包括:

获取待映射的目标数据,其中,所述目标数据包括第一属性文件;

计算所述第一属性文件与预设的参考数据的第二属性文件的第一相似度值,其中,所述第一属性文件设置有多个,所述第二属性文件也设置有多个,所述第一属性文件与所述第二属性文件一一映射,从而得到多个所述第一相似度值;

根据多个所述第一相似度值生成表征所述目标数据与所述参考数据重合度的总相似度值;

根据所述总相似度值对所述目标数据进行映射处理。

2.根据权利要求1所述的数据映射方法,其特征在于,所述第一属性文件包括第一标识信息和第一文本信息,所述第一标识信息与所述第一文本信息一一映射,所述第二属性文件包括第二标识信息和第二文本信息,所述第二标识信息与所述第二文本信息一一映射,其中,所述第一标识信息为表征所述目标数据的类型参数,所述第二标识信息为表征所述参考数据的类型参数;所述计算所述第一属性文件与预设的参考数据的第二属性文件的第一相似度值的方法包括:

提取类型参数相同的第一标识信息和第二标识信息所分别映射的第一文本信息和第二文本信息;

对所提取的所述第一文本信息和所述第二文本信息进行比对,以得到所述第一相似度值。

3.根据权利要求2所述的主数据映射处理方法,其特征在于,所述对所提取的所述第一文本信息和所述第二文本信息进行比对,以得到所述第一相似度值的方法包括:

调取规则数据库,在所述规则数据库中查找确定与所述第一标识信息匹配的比较规则;

根据所述比较规则计算所述第一文本信息与所述第二文本信息的第一相似度值。

4.根据权利要求3所述的主数据映射处理方法,其特征在于,所述根据多个所述第一相似度值生成表征所述目标数据与所述参考数据重合度的总相似度值的方法包括:

获取所述第一标识信息所映射的权重值;

将所述第一标识信息所映射的权重值与其对应的第一相似度值相乘以得到第二相似度值;

将所述目标数据中所包含的所有第一标识信息对应的第二相似度值相加得到所述总相似度值。

5.根据权利要求3所述的主数据映射处理方法,其特征在于,所述比较规则包括全等算法,其中,所述全等算法为判断所述第一文本信息与所述第二文本信息是否完全相同。

6.根据权利要求3-5任意一项所述的主数据映射处理方法,其特征在于,所述比较规则包括相似度算法,其中,所述相似度算法为判断所述第一文本信息与所述第二文本信息的相似概率。

7.根据权利要求1所述的主数据映射处理方法,其特征在于,当目标数据有多个时,所述根据所述总相似度值对所述目标数据进行映射处理的方法包括:

对多个所述目标数据的总相似度值的大小进行排序;

根据排序结果,提取所述总相似度值大于或等于预设阈值的目标数据生成相似数据列

表；

从所述相似数据列表中选取一条或多条所述目标数据与所述参考数据关联映射。

8.一种数据映射处理装置，其特征在于，包括：

获取模块：被配置为执行获取待映射的目标数据，其中，所述目标数据包括多个第一属性文件；

第一计算模块：被配置为执行计算各第一属性文件与预设的参考数据的第二属性文件的第一相似度值；

第二计算模块：被配置为执行根据所述第一相似度值生成表征所述目标数据与所述参考数据重合度的总相似度值；

执行模块：被配置为执行根据所述总相似度值对所述目标数据进行映射处理。

9.一种计算机设备，包括存储器和处理器，所述存储器中存储有计算机可读指令，所述计算机可读指令被所述处理器执行时，使得所述处理器执行如权利要求1至7中任一项权利要求所述主数据映射处理方法的步骤。

10.一种存储有计算机可读指令的存储介质，所述计算机可读指令被一个或多个处理器执行时，使得一个或多个处理器执行如权利要求1至7中任一项权利要求所述主数据映射处理方法的步骤。

主数据映射处理方法、装置、计算机设备及存储介质

技术领域

[0001] 本申请涉及企业信息化的数据处理技术领域，具体而言，本申请涉及一种主数据映射处理方法、装置、计算机设备及存储介质。

背景技术

[0002] 主数据是企业内跨业务、能共享的高价值核心业务实体，是企业的关键数据，例如：人员、产品、供应商、物料等。主数据管理帮助企业建立主数据单一视图并进行数据共享。

[0003] 主数据管理整合企业各业务系统的主数据然后进行数据治理。数据治理的一个重要技术手段是主数据映射，主数据映射的目的是把重复的、疑似重复的两条或多条数据找出来，进行筛选与修改，然后与标准主数据建立对照关系，这样能提高主数据共享的数据质量。

[0004] 现有的主数据映射技术主要包括：一、利用数据库能力，写SQL使用where语句的“=”、“LIKE”或特定函数来去重数据，人工比较后写SQL直接进行映射关系的更新；二、利用EXCEL等工具进行人工去重比对并与标准主数据建立映射关系，然后直接导入到系统。

[0005] 以上两种传统方案的缺陷有：1)忽略了数据相似判断是业务与技术相结合的过程，既要技术手段还需要业务手段，一般情况使用技术手段去重后，需要业务人员进行稽核，确定是否要去重或修改；2)仅使用传统的数据库能力很难找到疑似重复、不同词但同义的主数据记录，例如供应商地址：“辽宁省沈阳市”与“沈阳市”就是一个地址。3)两条主数据记录有时要进行多字段属性内容的综合比较，来确定二者的相似度，而非单个属性。

发明内容

[0006] 基于以上问题，本申请公开一种主数据映射处理方法、装置、计算机设备及存储介质，采用计算机对多个数据多个属性文件进行客观、准确、快速地相似度识别和数据映射。

[0007] 本申请的实施例根据第一个方面，提供了一种主数据映射处理方法，包括：

[0008] 获取待映射的目标数据，其中，所述目标数据包括第一属性文件；

[0009] 计算所述第一属性文件与预设的参考数据的第二属性文件的第一相似度值，其中，所述第一属性文件设置有多个，所述第二属性文件也设置有多个，所述第一属性文件与所述第二属性文件一一映射，从而得到多个所述第一相似度值；

[0010] 根据多个所述第一相似度值生成表征所述目标数据与所述参考数据重合度的总相似度值；

[0011] 根据所述总相似度值对所述目标数据进行映射处理。

[0012] 可选的，所述第一属性文件包括第一标识信息和第一文本信息，所述第一标识信息与所述第一文本信息一一映射，所述第二属性文件包括第二标识信息和第二文本信息，所述第二标识信息与所述第二文本信息一一映射，其中，所述第一标识信息为表征所述目标数据的类型参数，所述第二标识信息为表征所述参考数据的类型参数；所述计算所述第

一属性文件与预设的参考数据的第二属性文件的第一相似度值的方法包括：

[0013] 提取类型参数相同的第一标识信息和第二标识信息所分别映射的第一文本信息和第二文本信息；

[0014] 对所提取的所述第一文本信息和所述第二文本信息进行比对，以得到所述第一相似度值。

[0015] 可选的，所述对所提取的所述第一文本信息和所述第二文本信息进行比对，以获得到所述第一相似度值的方法包括：

[0016] 调取规则数据库，在所述规则数据库中查找确定与所述第一标识信息匹配的比较规则；

[0017] 根据所述比较规则计算所述第一文本信息与所述第二文本信息的第一相似度值。

[0018] 可选的，所述根据多个所述第一相似度值生成表征所述目标数据与所述参考数据重合度的总相似度值的方法包括：

[0019] 获取所述第一标识信息所映射的权重值；

[0020] 将所述第一标识信息所映射的权重值与其对应的第一相似度值相乘以得到第二相似度值；

[0021] 将所述目标数据中所包含的所有第一标识信息对应的第二相似度值相加得到所述总相似度值。

[0022] 可选的，所述比较规则包括全等算法，其中，所述全等算法为判断所述第一文本信息与所述第二文本信息是否完全相同。

[0023] 可选的，所述比较规则包括相似度算法，其中，所述相似度算法为判断所述第一文本信息与所述第二文本信息的相似概率。

[0024] 可选的，当目标数据有多个时，所述根据所述总相似度值对所述目标数据进行映射处理的方法包括：

[0025] 对所述目标数据的总相似度值的大小进行排序；

[0026] 根据排序结果，提取所述总相似度值大于或等于预设阈值的目标具生成相似数据列表；

[0027] 从所述相似数据列表中选取一条或多条所述目标数据与所述参考数据关联映射。

[0028] 另一方面，本申请公开一种主数据映射处理装置，包括：

[0029] 获取模块：被配置为执行获取待映射的目标数据，其中，所述目标数据包括第一属性文件；

[0030] 第一计算模块：被配置为执行计算所述第一属性文件与预设的参考数据的第二属性文件的第一相似度值，其中，所述第一属性文件设置有多个，所述第二属性文件也设置有多个，所述第一属性文件与所述第二属性文件一一映射，从而得到多个所述第一相似度值；

[0031] 第二计算模块：被配置为执行根据多个所述第一相似度值生成表征所述目标数据与所述参考数据重合度的总相似度值；

[0032] 执行模块：被配置为执行根据所述总相似度值对所述目标数据进行映射处理。

[0033] 可选的，所述第一属性文件包括第一标识信息和第一文本信息，所述第一标识信息与所述第一文本信息一一映射，所述第二属性文件包括第二标识信息和第二文本信息，所述第二标识信息与所述第二文本信息一一映射，其中，所述第一标识信息为表征所述目

标数据的类型参数，所述第二标识信息为表征所述参考数据的类型参数；所述第一计算模块包括：

[0034] 提取模块：被配置为执行提取类型参数相同的第一标识信息和第二标识信息所分别映射的第一文本信息和第二文本信息；

[0035] 第一比对模块：被配置为执行对所提取的所述第一文本信息和所述第二文本信息进行比对，以获得到所述第一相似度值。

[0036] 可选的，所述第一比对模块包括：

[0037] 规则匹配模块，被配置为执行调取规则数据库，在所述规则数据库中查找确定与所述第一标识信息匹配的比较规则；；

[0038] 第一计算子模块：被配置为执行根据所述比较规则计算所述第一文本信息与所述第二文本信息的第一相似度值。

[0039] 可选的，所述第二计算模块包括：

[0040] 权重获取模块：被配置为执行获取所述第一标识信息所映射的权重值；

[0041] 乘积模块：被配置为执行将所述第一标识信息所映射的权重值与其对应的第一相似度值相乘以得到第二相似度值；

[0042] 第二计算子模块：被配置为执行将所述目标数据中所包含的所有第一标识信息对应的第二相似度值相加得到所述总相似度值。

[0043] 可选的，所述比较规则包括全等算法，其中，所述全等算法为判断所述第一文本信息与所述第二文本信息是否完全相同。

[0044] 可选的，所述比较规则包括相似度算法，其中，所述相似度算法为判断所述第一文本信息与所述第二文本信息的相似概率。

[0045] 可选的，当目标数据有多个时，所述执行模块包括：

[0046] 排序模块：被配置为执行对所述目标数据的总相似度值的大小进行排序；

[0047] 列表生成模块：被配置为执行根据排序结果，提取所述总相似度值大于或等于预设阈值的目标具生成相似数据列表；

[0048] 映射模块：被配置为执行从所述相似数据列表中选取一条或多条所述目标数据与所述参考数据关联映射。

[0049] 本申请的实施例根据第三个方面，还提供了一种计算机设备，包括存储器和处理器，所述存储器中存储有计算机可读指令，所述计算机可读指令被所述处理器执行时，使得所述处理器执行上述所述主数据映射处理方法的步骤。

[0050] 本申请的实施例根据第四个方面，还提供了一种存储有计算机可读指令的存储介质，所述计算机可读指令被一个或多个处理器执行时，使得一个或多个处理器执行上述所述主数据映射处理方法的步骤。

[0051] 本申请实施例的有益效果是：本申请公开一种主数据映射处理方法、装置、计算机设备及存储介质，通过获取待映射的目标数据，识别该数据中的属性文件信息与参考数据中的属性文件信息进行比对，计算其相似度值，通过相似度值来协助用户进行数据映射，采用该映射方式使主数据映射更方便、快捷，且通过不同的属性文件具有不同的特性，通过对不同的属性文件分别进行相似度计算，再计算总的相似度，使获得的总相似度值更为客观，消除人为干扰。

附图说明

[0052] 本申请上述的和/或附加的方面和优点从下面结合附图对实施例的描述中将变得明显和容易理解，其中：

[0053] 图1为本申请一个实施例的主数据映射处理方法流程示意图；

[0054] 图2为本申请一个具体实施例的计算第一相似度值的方法示意图；

[0055] 图3为本申请各属性文件的具体实施例；

[0056] 图4为本申请一个具体实施例根据文本信息获取第一相似度值的方法示意图；

[0057] 图5为本申请一个具体实施例的计算总相似度值的方法示意图；

[0058] 图6为本申请一具体实施例总相似度计算流程示意图；

[0059] 图7为本申请一个实施例的目标主数据映射处理方法示意图；

[0060] 图8为本申请一个实施例的目标主数据映射整体流程示意图；

[0061] 图9为本申请一具体实施例的数据稽核显示界面；

[0062] 图10为本申请一具体实施例的数据相似度及映射报告显示界面；

[0063] 图11为本申请一个实施例的主数据映射处理装置模块示意图；

[0064] 图12为本申请一个实施例的计算机设备基本结构框图。

具体实施方式

[0065] 下面详细描述本申请的实施例，所述实施例的示例在附图中示出，其中自始至终相同或类似的标号表示相同或类似的元件或具有相同或类似功能的元件。下面通过参考附图描述的实施例是示例性的，仅用于解释本申请，而不能解释为对本申请的限制。

[0066] 本技术领域技术人员可以理解，除非特意声明，这里使用的单数形式“一”、“一个”、“所述”和“该”也可包括复数形式。应该进一步理解的是，本申请的说明书中使用的措辞“包括”是指存在所述特征、整数、步骤、操作、元件和/或组件，但是并不排除存在或添加一个或多个其他特征、整数、步骤、操作、元件、组件和/或它们的组。

[0067] 本技术领域技术人员可以理解，除非另外定义，这里使用的所有术语(包括技术术语和科学术语)，具有与本申请所属领域中的普通技术人员的一般理解相同的意义。还应该理解的是，诸如通用字典中定义的那些术语，应该被理解为具有与现有技术的上下文中的意义一致的意义，并且除非像这里一样被特定定义，否则不会用理想化或过于正式的含义来解释。

[0068] 随着计算机的发展，很多原本由人工完成的工作都由计算机进行，计算机根据指定的规定进行操作和信息处理，错误率低，速度快。本申请基于计算机的这种特性，公开一种主数据映射处理方法，请参阅图1，具体包括：

[0069] S1000、获取待映射的目标数据，其中，所述目标数据包括第一属性文件；

[0070] S2000、计算所述第一属性文件与预设的参考数据的第二属性文件的第一相似度值，其中，所述第一属性文件设置有多个，所述第二属性文件也设置有多个，所述第一属性文件与所述第二属性文件一一映射，从而得到多个所述第一相似度值；

[0071] 在本申请中，待映射的目标数据为任意需要与参考数据进行相似度比对和映射的数据，参考数据即主数据，为预先录入的针对每一款产品的标准信息输入格式，其包括多个属性文件，每个属性文件包括标识信息和文本信息，标识信息与文本信息一一映射。目标数

据的录入规则通常与参考数据相同，因此在目标数据中也包括多个属性文件，每个属性文件下也都包括标识信息和对应的文本信息。为了区分，在本申请中，将目标数据的属性文件称之为第一属性文件，将第一属性文件下的标识信息称之为第一标识信息，第一标识信息映射的文本信息称之为第一文本信息；将参考数据的属性文件称之为第二属性文件，将第二属性文件下的标识信息称之为第二标识信息，将第二标识信息映射的文本信息称之为第二文本信息。

[0072] 在一实施例中，请参阅图2，所述第一属性文件包括第一标识信息和第一文本信息，所述第一标识信息与所述第一文本信息一一映射，所述第二属性文件包括第二标识信息和第二文本信息，所述第二标识信息与所述第二文本信息一一映射，其中，所述第一标识信息为表征所述目标数据的类型参数，所述第二标识信息为表征所述参考数据的类型参数；所述计算所述第一属性文件与预设的参考数据的第二属性文件的第一相似度值的方法包括：

[0073] S2100、提取类型参数相同的第一标识信息和第二标识信息所分别映射的第一文本信息和第二文本信息；

[0074] S2200、对所提取的所述第一文本信息和所述第二文本信息进行比对，以得到所述第一相似度值。

[0075] 在本申请中，主数据映射处理的应用场景可以为：在多条待映射的目标数据中，根据每条待映射的目标数据中的第一标识信息与参考数据中的第二标识信息依次进行对比，看是否相同，若相同则将该目标数据的第一文本信息与参考数据的第二文本信息进行比对，获得第一相似度，例如，以对药品数据进行管理过程中的主数据映射处理为例，药品管理类数据中通常需要记载药品的名称、编码、型号、厂商、生产日期等类型的内容，这些类型信息称之为标识信息，这些标识信息下具体的内容称之为文本信息，标识信息与文本信息结合起来称之为属性文件，在一条待映射的目标数据中包含多个类型，每个类型下都映射有具体信息，通过获取目标数据与参考数据中每个相同类型下的第一文本信息与第二文本信息的相似度，来判断二者是否相同或相似，并以此作为判断该目标数据与参考数据是否可建立映射关系的基础。

[0076] 在一实施例中，不同的标识信息在判断目标数据与参考数据之间是否相似的过程中所起的重要性和识别规则不同，例如，请参阅图3，为以具体应用图示，当属性文件的标识信息为“其他信息”时，该标识信息映射的文本信息中只记录一些投诉内容、投诉的数目等一些跟区分产品的实质内容无关的信息时，即使该文本信息内容不同，也不能认为该条属性文件所属的目标数据跟参考数据不一样，这类型属性文件可称之为无效属性文件。与无效属性文件对应的为有效属性文件，即能够对数据映射判断构成影响的科目数据，比如编码、产品名称、型号、厂商名称、厂商地址、所属业务系统等属性文件，当目标数据中的型号信息与参考数据中的型号信息不同时，即使产品名称一样，产商等信息都一样，也可认为二者不是相同的数据，而当目标数据中其他属性文件下的文本信息都与参考数据中相同属性文件下的中文本信息一样，只是产品名称不一样时，目标数据与参考数据之间可能相同也可能不同，以药品管理为例，对于药品而言，有学名、中文名、英文名之分，例如“青霉素”，又叫“盘尼西林”，英文名称为“benzylpenicillin”，在产品名称中记载“青霉素”、“盘尼西林”或“benzylpenicillin”实际上都是一个东西，除了药品名称和表征该条数据代号的编码不

同,其他的标识信息下的文本信息都一样时,二者属于相关联的数据,应当进行映射处理,例如图3中的目标数据的编码为AS123,其与参考数据中的标准编码MD12345的数据在产品型号、厂商地址等信息都一样,可以认为二者是同一种数据,具有映射关系。而对于一些不相关的名称,例如“青霉素”与“阿莫西林”,名称不一样,则为不同的产品,不可建立映射关系。另外,对于编码,为了便于商品管理,厂家在生产产品过程中,会对不同的产品编制不同的编码,例如对“青霉素”类产品,编码是A123,对该“阿莫西林”类产品,编码是B123,因此,当识别出编码不同时,对应的其他属性文件即使不进行识别也可认为二者是不相同的数据。

[0077] 基于属性文件本身存在以上区别,在本申请中,需要对目标数据与参考数据中每个对应的标识信息下的文本信息分别进行相似度判断,才可从整体上识别二者是否相同或相似,单个属性文件对应文本信息的相似度称之为第一相似度。

[0078] 在一实施例中,请参阅图4,所述对所提取的所述第一文本信息和所述第二文本信息进行比对,以得到所述第一相似度值的方法包括:

[0079] S2210、调取规则数据库,在所述规则数据库中查找确定与所述第一标识信息匹配对应的比较规则;

[0080] S2220、根据所述比较规则计算所述第一文本文信息与所述第二文本信息的第一相似度值。

[0081] 不同的第一标识信息具有不同的特性,有的第一标识信息中的第一文本信息必须与第二文本信息完全一样才能被定义为相同,有的第一标识信息的第一文本信息即使与第二文本信息不一样,也可通过近义词或关联词匹配定义二者相同,例如,当识别“型号”、“编码”等属性文件下的数据时,需采用判断文本信息是否完全相同的规则进行判断,完全相同则认为一样,不完全相同则认为不一样;而当识别“产品名称”的第一标识信息中的第一文本信息时,可采用是否相似,以将“产品名称”下的第一文本信息与预设名称数据库中的相似名称进行比对,判断二者是否相同。而当识别“产商地址”的第一标识信息的第一文本信息时,由于地址分为国家、省、市、区(县)、乡、镇、村、组等不同的级别,级别越大,相同的概率越大,级别越小,相同的概率越小,当越小级别的地址信息相同时,该地址相同的概率越大。而对于地址,可能输入的信息不完全,例如只输入市区,或者只输入了乡镇,而没有输对应的市区,但是乡镇与市区和省级之间会有相关的关联性,因此可采用关键字识别的方式来判断该科目下的数据是否相同,例如,识别输入的地址中是否有省、市、区(县)、乡、镇、村、组、路、街等相关的关键字,当有这几个关键字时,判断关键字之前的字是否与参考数据中的相同,在相同的情况下,判断关键字是否为路或者街,当为不为路或者街,则表示输入的地址区域范围过大,不能识别,当关键字为路或者街,关键字前面的文本信息又相同,则可判断二者相同。

[0082] 因此,不同的第一标识信息在计算第一相似度值时具有不同的计算规则,要更客观地计算目标数据的相似度,需要根据不同的标识信息采用不同的比较计算策略,即采用不同的计算规则。在本申请中,建立有规则数据库,将各个标识信息映射不同的相似度比较规则,当需要计算第一相似度值时,首先调取该规则数据库,然后再在该规则数据库中匹配属性文件对应的比较规则。

[0083] S3000、根据多个所述第一相似度值生成表征所述目标数据与所述参考数据重合

度的总相似度值；

[0084] 由于目标数据中具有多个属性文件，在计算总相似度值时，直接将通过步骤S2000计算得到的每个标识信息的第一相似度值相加即可得到总相似度值。

[0085] 在判断整个目标数据的总相似度值的过程中每个标识信息所起的作用不一样，为了更客观地体现总相似度值，根据标识信息的特性，以及该标识信息在整个目标数据的重要程度，请参阅图5，所述根据多个所述第一相似度值生成表征所述目标数据与所述参考数据重合度的总相似度值的方法包括：

[0086] S3100、获取所述第一标识信息所映射的权重值；

[0087] S3200、将所述第一标识信息所映射的权重值与其对应的第一相似度值相乘以得到第二相似度值；

[0088] S3300、将所述目标数据中所包含的所有第一标识信息对应的第二相似度值相加得到所述总相似度值。

[0089] 对各个第一标识信息设置一个权重值，该目标数据中的所有第一标识信息的权重值之和为1，通过第一标识信息所匹配的比较规则得到各自的第一相似度值后，将第一标识信息的第一相似度与对应的权重值相乘，得到第二相似度值，将所有的第二相似度值相加，则可得到该待映射的目标数据的总相似度值。例如，比较规则包括相似度算法和全等算法，请参阅图6，在某个目标数据中，分析得到对应的第一标识信息为：属性1、属性2……属性n，再对各个第一标识信息分配比较规则及对应的权重值，得到属性1采用相似度算法较规则，权重值为B1，属性2采用全等算法规则，权重值为B2，属性n也采用相似度算法规则，权重值为Bn，根据相似度算法规则得到算出属性1的第一相似度值为A1，根据全等算法规则算出属性2的第一相似度值为A2，根据相似度算法规则，计算得到属性n的第一相似度值为An，其中，B1+B2+……+Bn＝1，计算得到总相似度＝A1*B1+A2*B2+……+An*Bn。

[0090] 上述公开的比较规则包括全等算法，其中，所述全等算法为判断所述属性文件中的文本信息与参考数据对应的属性文件的文本信息是否完全相同，完全相同输出1，不完全相同输出0。全等算法可以直接使用数据库SQL的where语句的“＝”进行判断或者使用普通程序实现。

[0091] 所述比较规则包括相似度算法，其中，所述相似度算法为判断所述属性文件中的文本信息与参考数据对应的属性文件的文本信息是否相似，相似度的输出结果为1至0之间的数值；相似度算法包括但不局限于采用余弦相似度算法，如果数据量大的话，还可以采用大数据相关技术。在一实施例中，在采用余弦相似度算法过程中，先提取文本信息，列出所有的词，进行分词编码，之后进行分词向量化，再采用余弦函数计量两个文本的相似度。

[0092] 文本分词编码可以采用JAVA或PYTHON的Ikanalyzer、Jcseg、Jieba等开源技术。分词向量化为将文本分词后的数据转换为计算机认识的形式，其包括几种方案，例如以字或单词为单位进行向量化，或者以句子为单位进行向量化，以字或词为单位的文本向量化方法包括词集模型、词袋模型、n-gram、TF-IDF和word2vec等算法；以句子为单位的的向量化方法包括LSA、NMF、pLSA、LDA等算法。

[0093] 例如，以TF-IDF(Term Frequency-Inverse Document Frequency，词频-逆文本频率)为例，它由两部分组成，TF和IDF；前面的TF也就是词频，词频向量化也就是做了文本中各个词的出现频率统计，并作为文本特征，后面的IDF，即“逆文本频率”，由于几乎所有文本

都会出现的"的"其词频虽然高,但是重要性却应该比一些词频低的词要低,因此通过IDF来反应这个词的重要性的,进而修正仅仅用词频表示的词特征值。所以一个词的定量化表示比较合理的是(词频X词的权重),其计算规则为:TF-IDF(x)＝TF(x)*IDF(x),其中,x表示需要统计的某个词或字,TF(x)为改词的频率,IDF(x)为该词的权重。

[0094] 分词向量化后得到第一文本信息的向量化值X和参考数据的第二文本信息的向量化值Y两个词频向量化数组,计算得到的cos值为:

[0095] $$\cos(\theta) = \frac{\sum_{i=1}^{n}(X_i \times Y_i)}{\sqrt{\sum_{i=1}^{n}(X_i)^2} \times \sqrt{\sum_{i=1}^{n}(Y_i)^2}}$$

[0096] 其中,n为第一文本信息中字的数量,Xi表示第一文本信息中第i个字的向量化值,Yi表示参考数据中第二文本信息中第I个字的向量化值。

[0097] 计算得到的COS取值范围是【-1,1】,当值COS<0,我们取值＝0,因此相似度为【0,1】,COS值数值越大,表示越相似,COS值数值越小,表示相似度越低,当计算出的数据为负数时,cos值取0。

[0098] S4000、根据所述总相似度值对所述目标数据进行映射处理。

[0099] 通过步骤S3000得出待映射的目标数据的总相似度值后,则可对相关数据进行映射处理。总相似度值是判断两个或多个数据之间是否内容相同或相似的数值,总相似度值越高,则表示二者之间的数据越相似,总相似度值越低,则表示两组数据之间相差越大。

[0100] 映射的过程实际上也可理解成数据稽核的过程,数据稽核的方式包括人工和计算机批量处理两种形式,根据目标数据的数量进行选择,例如,当目标数据有多个对应的总相似度大于或等于预设阈值时,若采用手工稽核映射,效率较低,因此可采用计算机批量稽核的方式进行映射,在一实施例中,计算机批量稽核映射为将总相似度值大于或等于预设阈值的目标数据全部标记,与参考数据一一进行自动映射。这个预设阈值为判断目标数据与参考数据是否需要映射的最小的总相似度值,例如将预设阈值设置成98%,当总相似度值大于或等于98%时,表示相似度较高,需要进行下一步的映射处理,当总相似度值小于98%时,则表示相似度不太高,两个数据的相关参数信息相差较多,不属于同样或相似的数据,可不进行映射处理。当用户选择的是不进行批量稽核映射时,则又可分为是否人工映射,当为人工映射时,由用户自己根据稽核条件,对照参考数据进行修改,并手动进行映射关联,当用户选择非人工映射时,则计算机根据获取的总相似度值对所述目标数据进行映射处理。

[0101] 请参阅图7,所述根据所述总相似度值对所述目标数据进行映射处理的方法包括:

[0102] S4100、对多个所述目标数据的总相似度值的大小进行排序;

[0103] S4200、根据排序结果,提取所述总相似度值大于或等于预设阈值的目标数据生成相似数据列表;

[0104] S4300、从所述相似数据列表中选取一条或多条所述目标数据与所述参考数据关联映射。

[0105] 对获取的各个目标数据的总相似度值进行排名,并设置一个预设阈值,采用预设

阈值来划分需要映射处理的数据和不许映射处理的数据，在一实施例中，将所述总相似度值大于或等于预设阈值的目标数据罗列以生成相似数据列表，且在映射处理过程中，只显示相似数据列表中的目标数据，以减少后续映射处理的工作量，并使后续映射处理数据的数量更少，映射处理界面更简洁，采用计算机对符合预设阈值的目标数据进行排序，从而方便用户根据目标数据的总相似度值的大小来选择需要映射的数据，并建立映射关系。

[0106] 在一实施例中，目标数据总相似度计算是通过WEB页面进行操作的，在一实施例中，请参阅图8，目标数据的映射方法整体流程包括：

[0107] S4310、开始：对所述目标数据开始执行映射作业；

[0108] S4320、设置条件，判断所述总相似度值大于或等于某个阈值，当小于该阈值，则直接结束流程，进入S4370，当大于或等于该阈值，则进入步骤S4330；

[0109] S4330、判断是否需要批量进行稽核映射处理，当需要则进入步骤S4340，当不需要则进入步骤S4350；

[0110] S4340、对所有大于或等于某个阈值的目标数据批量建立映射关系；

[0111] S4350、选择判断方式，判断是否需要人工判断，当需要人工判断进入步骤S4351，当无需人工判断进入步骤S4352；

[0112] S4352、对总相似度值进行排序；

[0113] S4353、依据总相似度值的排列顺序对目标数据进行修改和映射；

[0114] S4360、稽核数据是否准确，正确则进入步骤S4370，不正确则重新进入S4310步骤进行映射筛选；

[0115] S4370、稽核通过，生成相似度及映射报告；

[0116] S4380、结束流程。

[0117] 当进行人工映射后，具体的步骤为：

[0118] S4351、用户自己查找需要修改和映射的目标数据以进行修改和映射，完成该步骤后进入S4360步骤。

[0119] 在一具体的实施例中，请参阅图9，为用户对数据进行稽核的显示界面，分别对“标准编码”、“产品名称”、“型号”、“厂商地址”、“其他信息”的第一标识信息进行匹配，该第一标识信息下对应的数据“cc”“a”、“b”、“c”“ac”等为第一文本信息，在未稽核数据中，通过计算机自动匹配了总相似度值大于80％的所有目标数据，只需要选择“稽核通过”按钮，则可以将所显示的总相似度值大于80％的所有目标数据一键映射。进一步的，在显示界面中，还公开了当不选用人工进行判断的示例界面，由计算机根据对应的算法匹配结果，并按照总相似度数据排序，罗列出包括对应编码，显示该编码对应的匹配度、产品名称、型号和厂商地址等信息，以便于用户查看，进行手动映射。或者是通过“自己来匹配”，完全人工查找相关的数据进行手动映射匹配。

[0120] 当稽核完成，生成的生成相似度及映射报告如图10所示，该示例性报告中，包括文件名、上传日期、上传条数、上传失败条数、匹配规则及其权重值信息、各个匹配度的数量及其分布图或柱形、树形图，进一步的，还可通过“下载报告EXCEL”将上述数据生成EXCEL表格。

[0121] 需要说明的是生成的相似度及映射报告包括文字报告、图形报告和表格报告等多种形式，以可视化地展示相关的相似度处理结果和映射结果。

[0122] 另一方面，请参阅图11，本申请公开一种主数据映射处理装置，包括：

[0123] 获取模块：被配置为执行获取待映射的目标数据，其中，所述目标数据包括第一属性文件；

[0124] 第一计算模块：被配置为执行计算所述第一属性文件与预设的参考数据的第二属性文件的第一相似度值，其中，所述第一属性文件设置有多个，所述第二属性文件也设置有多个，所述第一属性文件与所述第二属性文件一一映射，从而得到多个所述第一相似度值；

[0125] 第二计算模块：被配置为执行根据多个所述第一相似度值生成表征所述目标数据与所述参考数据重合度的总相似度值；

[0126] 执行模块：被配置为执行根据所述总相似度值对所述目标数据进行映射处理。

[0127] 可选的，所述第一属性文件包括第一标识信息和第一文本信息，所述第一标识信息与所述第一文本信息一一映射，所述第二属性文件包括第二标识信息和第二文本信息，所述第二标识信息与所述第二文本信息一一映射，其中，所述第一标识信息为表征所述目标数据的类型参数，所述第二标识信息为表征所述参考数据的类型参数；所述第一计算模块包括：

[0128] 提取模块：被配置为执行提取类型参数相同的第一标识信息和第二标识信息所分别映射的第一文本信息和第二文本信息；

[0129] 第一比对模块：被配置为执行对所提取的所述第一文本信息和所述第二文本信息进行比对，以获得到所述第一相似度值。

[0130] 可选的，所述第一比对模块包括：

[0131] 规则匹配模块，被配置为执行调取规则数据库，在所述规则数据库中查找确定与所述第一标识信息匹配的比较规则；；

[0132] 第一计算子模块：被配置为执行根据所述比较规则计算所述第一文本信息与所述第二文本信息的第一相似度值。

[0133] 可选的，所述第二计算模块包括：

[0134] 权重获取模块：被配置为执行获取所述第一标识信息所映射的权重值；

[0135] 乘积模块：被配置为执行将所述第一标识信息所映射的权重值与其对应的第一相似度值相乘以得到第二相似度值；

[0136] 第二计算子模块：被配置为执行将所述目标数据中所包含的所有第一标识信息对应的第二相似度值相加得到所述总相似度值。

[0137] 可选的，所述比较规则包括全等算法，其中，所述全等算法为判断所述第一文本信息与所述第二文本信息是否完全相同。

[0138] 可选的，所述比较规则包括相似度算法，其中，所述相似度算法为判断所述第一文本信息与所述第二文本信息的相似概率。

[0139] 可选的，当目标数据有多个时，所述执行模块包括：

[0140] 排序模块：被配置为执行对所述目标数据的总相似度值的大小进行排序；

[0141] 列表生成模块：被配置为执行根据排序结果，提取所述总相似度值大于或等于预设阈值的目标具生成相似数据列表；

[0142] 映射模块：被配置为执行从所述相似数据列表中选取一条或多条所述目标数据与所述参考数据关联映射。

[0143] 由于上述数据映射处理装置是主数据映射处理方法一一对应的装置,其实现原理与主数据映射处理方法一样,此处不再赘述。

[0144] 本发明实施例提供计算机设备基本结构框图请参阅图12。

[0145] 该计算机设备包括通过系统总线连接的处理器、非易失性存储介质、存储器和网络接口。其中,该计算机设备的非易失性存储介质存储有操作系统、数据库和计算机可读指令,数据库中可存储有控件信息序列,该计算机可读指令被处理器执行时,可使得处理器实现一种主数据映射处理方法。该计算机设备的处理器用于提供计算和控制能力,支撑整个计算机设备的运行。该计算机设备的存储器中可存储有计算机可读指令,该计算机可读指令被处理器执行时,可使得处理器执行一种主数据映射处理方法。该计算机设备的网络接口用于与终端连接通信。本领域技术人员可以理解,图12中示出的结构,仅仅是与本申请方案相关的部分结构的框图,并不构成对本申请方案所应用于其上的计算机设备的限定,具体的计算机设备可以包括比图中所示更多或更少的部件,或者组合某些部件,或者具有不同的部件布置。

[0146] 计算机设备通过接收关联的客户端发送的提示行为的状态信息,即关联终端是否开启提示以及贷款人是否关闭该提示任务。通过验证上述任务条件是否达成,进而向关联终端发送对应的预设指令,以使关联终端能够根据该预设指令执行相应的操作,从而实现了对关联终端的有效监管。同时,在提示信息状态与预设的状态指令不相同时,服务器端控制关联终端持续进行响铃,以防止关联终端的提示任务在执行一段时间后自动终止的问题。

[0147] 本发明还提供一种存储有计算机可读指令的存储介质,所述计算机可读指令被一个或多个处理器执行时,使得一个或多个处理器执行上述任一实施例所述主数据映射处理方法。

[0148] 本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程,是可以通过计算机程序来指令相关的硬件来完成,该计算机程序可存储于一计算机可读取存储介质中,该程序在执行时,可包括如上述各方法的实施例的流程。其中,前述的存储介质可为磁碟、光盘、只读存储记忆体(Read-Only Memory,ROM)等非易失性存储介质,或随机存储记忆体(Random Access Memory,RAM)等。

[0149] 应该理解的是,虽然附图的流程图中的各个步骤按照箭头的指示依次显示,但是这些步骤并不是必然按照箭头指示的顺序依次执行。除非本文中有明确的说明,这些步骤的执行并没有严格的顺序限制,其可以以其他的顺序执行。而且,附图的流程图中的至少一部分步骤可以包括多个子步骤或者多个阶段,这些子步骤或者阶段并不必然是在同一时刻执行完成,而是可以在不同的时刻执行,其执行顺序也不必然是依次进行,而是可以与其他步骤或者其他步骤的子步骤或者阶段的至少一部分轮流或者交替地执行。

[0150] 以上所述仅是本发明的部分实施方式,应当指出,对于本技术领域的普通技术人员来说,在不脱离本发明原理的前提下,还可以做出若干改进和润饰,这些改进和润饰也应视为本发明的保护范围。

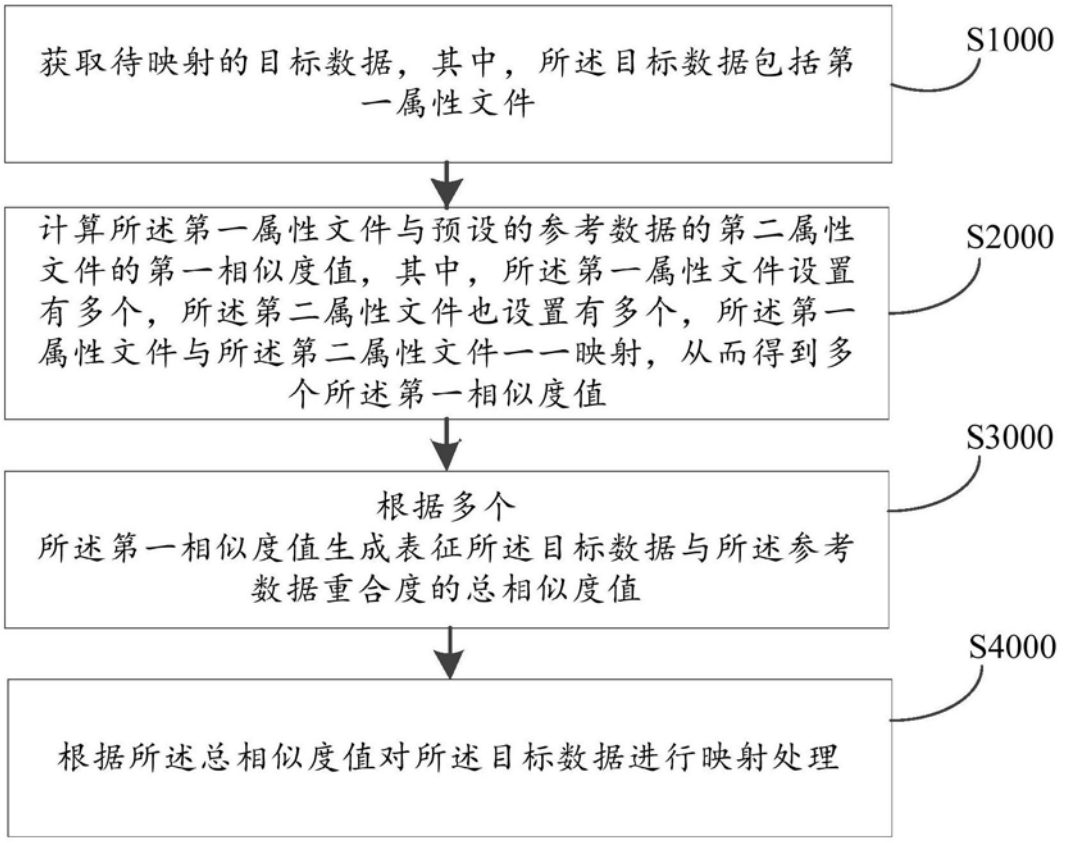

获取待映射的目标数据，其中，所述目标数据包括第一属性文件
S1000
计算所述第一属性文件与预设的参考数据的第二属性文件的第一相似度值，其中，所述第一属性文件设置有多个，所述第二属性文件也设置有多个，所述第一属性文件与所述第二属性文件一一映射，从而得到多个所述第一相似度值
S2000
根据多个
所述第一相似度值生成表征所述目标数据与所述参考数据重合度的总相似度值
S3000
根据所述总相似度值对所述目标数据进行映射处理
S4000

图1

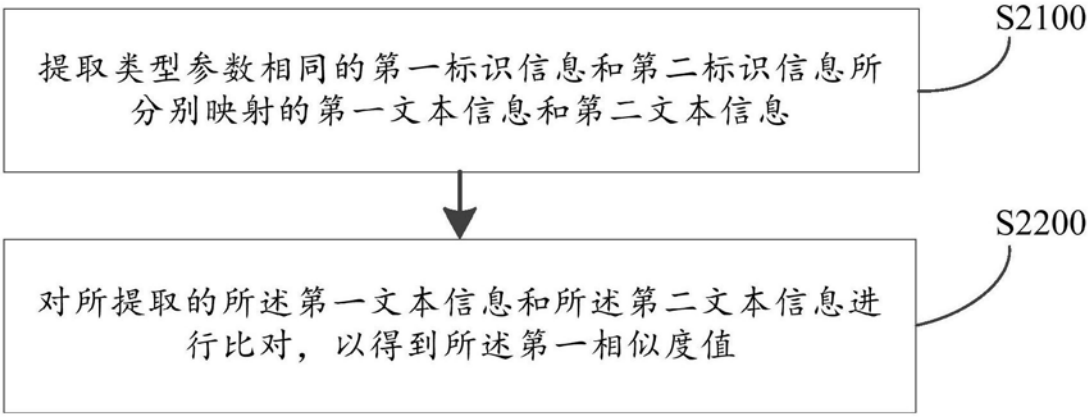

提取类型参数相同的第一标识信息和第二标识信息所分别映射的第一文本信息和第二文本信息
S2100
对所提取的所述第一文本信息和所述第二文本信息进行比对，以得到所述第一相似度值
S2200

图2


1
2
标准编码 产品名称 型号 厂商地址 其他信息
MD12345 盘尼西林 100克/袋 辽宁省沈阳市 ...
编码 产品名称 型号 厂商地址 其他信息 业务系统
AS123 青霉素 100克/袋 沈阳市 ... A
编码 产品名称 型号 厂商地址 其他信息 业务系统
BS456 盘尼西林 200克/袋 中国湖南 ... A

图3

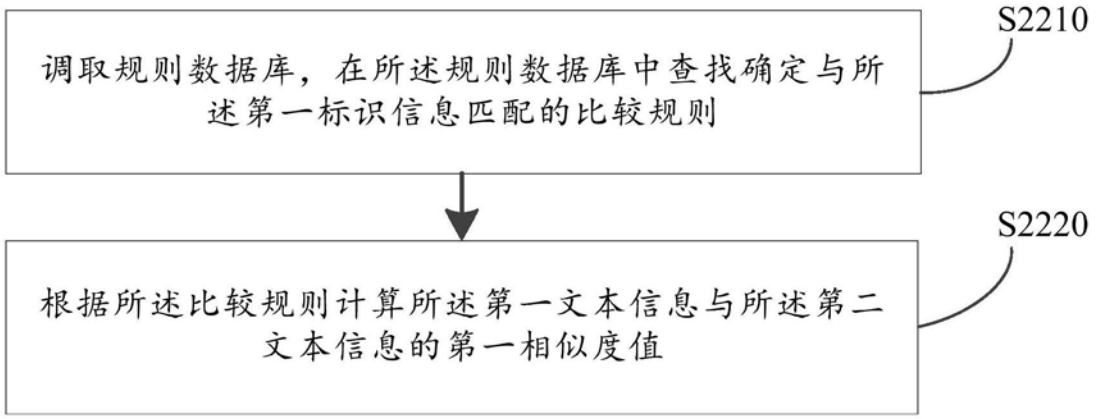

调取规则数据库，在所述规则数据库中查找确定与所述第一标识信息匹配的比较规则
S2210
根据所述比较规则计算所述第一文本信息与所述第二文本信息的第一相似度值
S2220

图4

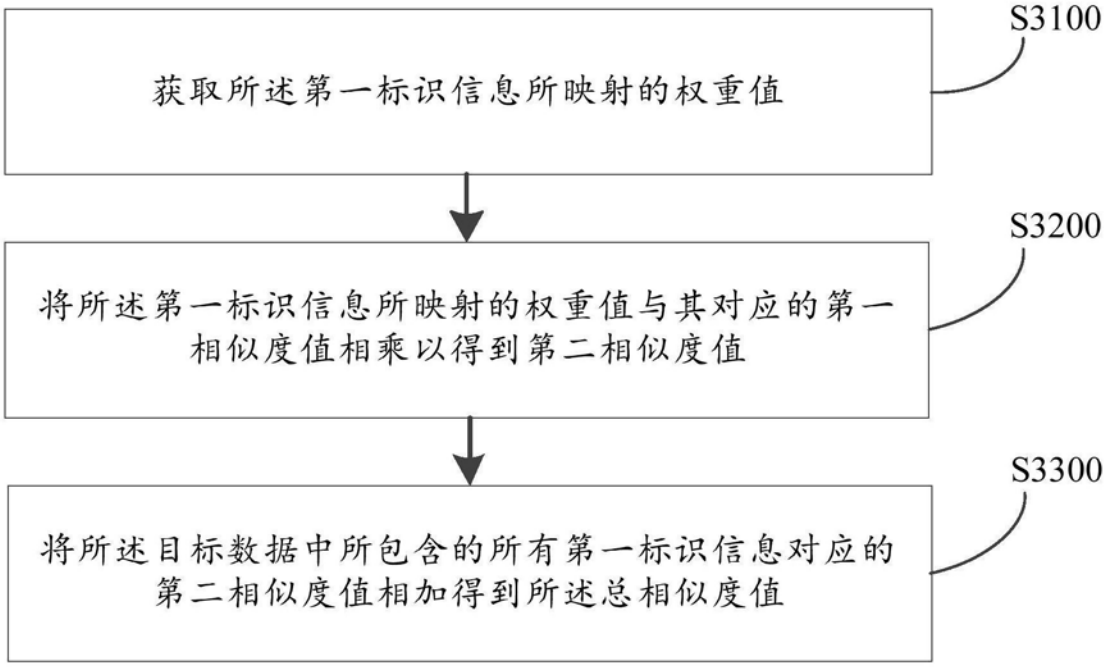

获取所述第一标识信息所映射的权重值
S3100
将所述第一标识信息所映射的权重值与其对应的第一相似度值相乘以得到第二相似度值
S3200
将所述目标数据中所包含的所有第一标识信息对应的第二相似度值相加得到所述总相似度值
S3300

图5

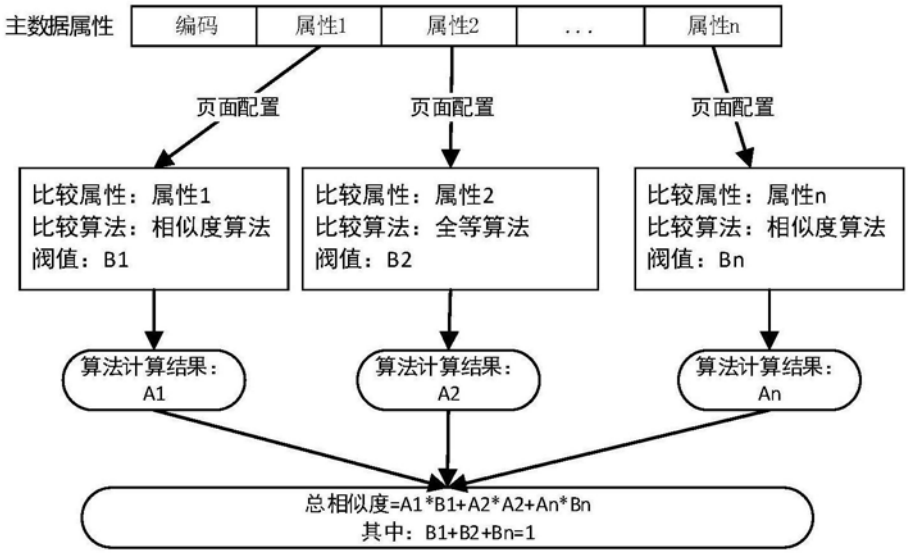

主数据属性
编码
属性1
属性2
...
属性n
页面配置
页面配置
页面配置
比较属性：属性1
比较算法：相似度算法
阀值：B1
比较属性：属性2
比较算法：全等算法
阀值：B2
比较属性：属性n
比较算法：相似度算法
阀值：Bn
算法计算结果：
A1
算法计算结果：
A2
算法计算结果：
An
总相似度=A1*B1+A2*A2+An*Bn
其中：B1+B2+Bn=1

图6

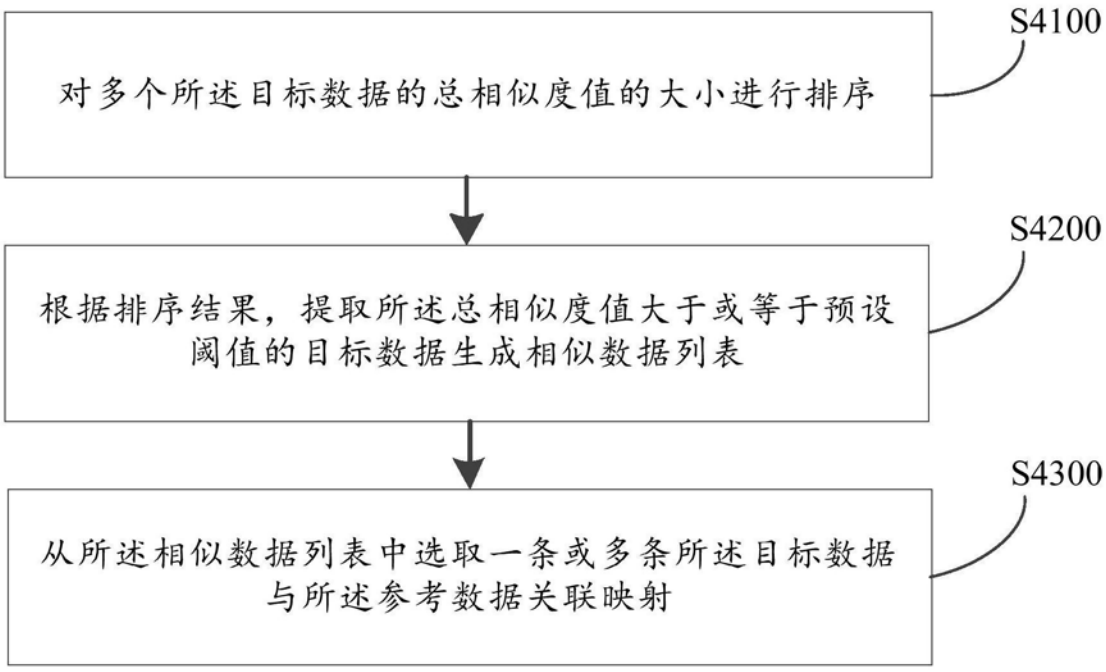

S4100
对多个所述目标数据的总相似度值的大小进行排序
S4200
根据排序结果，提取所述总相似度值大于或等于预设阈值的目标数据生成相似数据列表
S4300
从所述相似数据列表中选取一条或多条所述目标数据与所述参考数据关联映射

图7

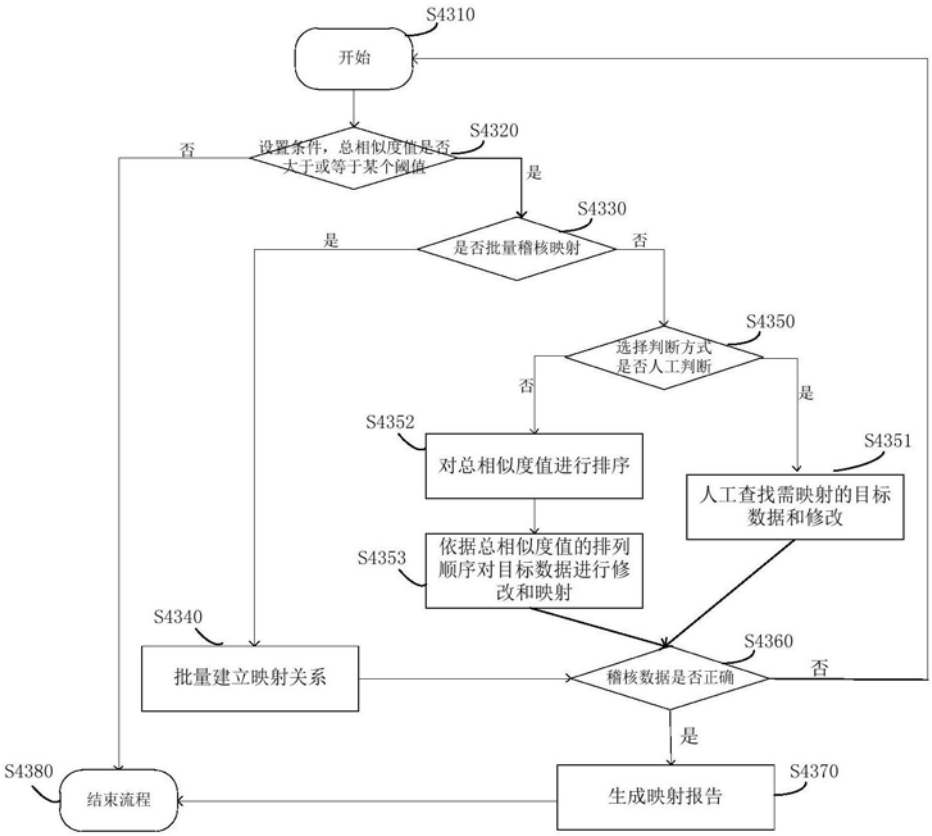

S4310
开始
S4320
设置条件，总相似度值是否大于或等于某个阈值
否
是
S4330
是否批量稽核映射
是
否
S4350
选择判断方式是否人工判断
否
是
S4352
对总相似度值进行排序
S4351
人工查找需映射的目标数据和修改
S4353
依据总相似度值的排列顺序对目标数据进行修改和映射
S4340
批量建立映射关系
S4360
稽核数据是否正确
否
是
S4380
结束流程
生成映射报告
S4370

图8


未稽核（小计数）
已稽核（小计数）
匹配过滤
>80%
查询
批量稽核
稽核通过
标准编码 产品名称 型号 厂商地址 其他信息 匹配个数
√ cc c c c c 2
a a c ac ac 1
b b b b b 0
分页控件
算法匹配结果
自己来匹配
匹配度 编码 产品名称 型号 厂商地址 其他信息 操作
98% ABC c c ac c 对照标准修改 稽核
82% ABD c c abc cde 对照标准修改 稽核
分页控件

图9

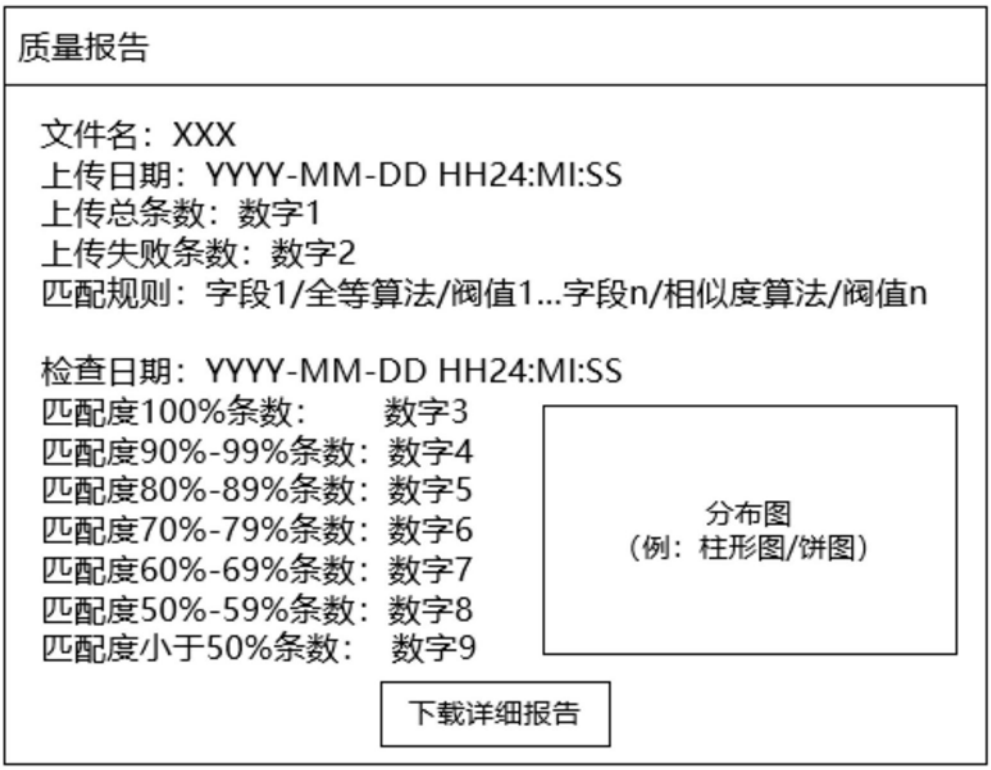

质量报告
文件名：XXX
上传日期：YYYY-MM-DD HH24:MI:SS
上传总条数：数字1
上传失败条数：数字2
匹配规则：字段1/全等算法/阀值1...字段n/相似度算法/阀值n
检查日期：YYYY-MM-DD HH24:MI:SS
匹配度100%条数： 数字3
匹配度90%-99%条数：数字4
匹配度80%-89%条数：数字5
匹配度70%-79%条数：数字6
匹配度60%-69%条数：数字7
匹配度50%-59%条数：数字8
匹配度小于50%条数： 数字9
分布图
（例：柱形图/饼图）
下载详细报告

图10

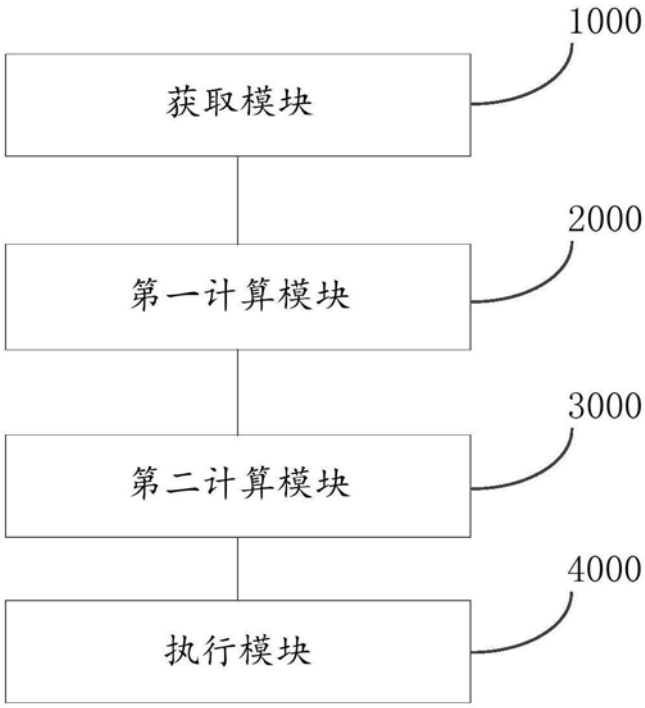

1000
获取模块
2000
第一计算模块
3000
第二计算模块
4000
执行模块

图11

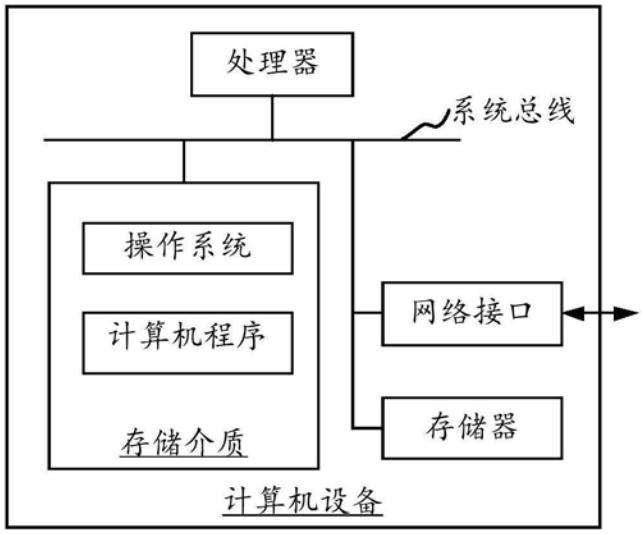


图12