(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 110781213 B
(45)授权公告日 2022.04.22

(21)申请号 201910911014.1
(22)申请日 2019.09.25
(65)同一申请的已公布的文献号
申请公布号 CN 110781213 A
(43)申请公布日 2020.02.11
(73)专利权人 中国电子进出口有限公司
地址 100036 北京市海淀区复兴路17号A座6-23层
(72)发明人 马万里
(74)专利代理机构 北京君尚知识产权代理有限公司 11200
代理人 邱晓锋
(51)Int.Cl.
G06F 16/2455(2019.01)
G06F 16/25(2019.01)
G06F 16/22(2019.01)

(56)对比文件
Fuyu Chen.Joint Power Optimization for Multi-Source Multi-Destination Relay Networks.《IEEE Transactions on Signal Processing》.2011,

审查员 蔡秀梅

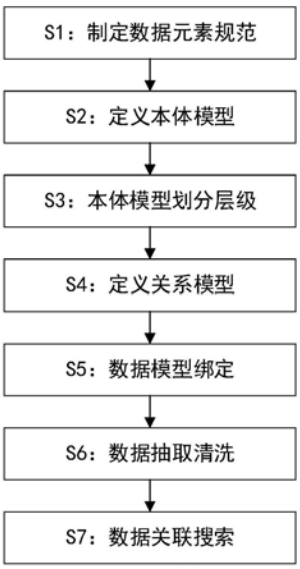
权利要求书2页 说明书6页 附图1页

(54)发明名称

一种以人员为中心的多源海量数据关联搜索方法和系统

(57)摘要

本发明涉及一种以人员为中心的多源海量数据关联搜索方法和系统。该方法包括以下步骤:定义数据元素规范、本体模型和关系模型,构建以人员为中心的知识森林体系;将数据源与本体模型、关系模型进行绑定,并将多源海量数据融入知识森林体系;在构建的知识森林体系中进行跨数据源的关联搜索。进一步可以对搜索结果进行分类统计、字段排序、条件筛选和二次搜索,帮助用户快速、精准地定位到目标结果。本发明的以人员为中心的知识森林体系和多源海量数据关联搜索方法,能够接入更多种类的数据源,实现更高效的聚合搜索,支撑更丰富的数据应用,极大提升大数据应用系统的数据兼容性和业务扩展性。



CN 110781213 B

1.一种以人员为中心的多源海量数据关联搜索方法,其特征在于,包括以下步骤:

定义数据元素规范、本体模型和关系模型,构建以人员为中心的知识森林体系;

将数据源与本体模型、关系模型进行绑定,并将多源海量数据融入知识森林体系;

在构建的知识森林体系中进行跨数据源的关联搜索;

所述将多源海量数据融入知识森林体系,包括以下步骤:

为数据源添加流水编号自增序列,作为数据增量式导入的依据;

从最后完成导入的流水编号开始,计算剩余待导入的数据量;

对待导入数据进行分包封装,将待导入任务拆分为若干个数据包的导入任务;

将数据包导入任务分发至大数据集群节点,实现多个数据包的并行导入;

对于本体模型数据导入任务,首先验证待导入本体在知识森林体系中是否已经存在,若尚未存在则创建一个新的本体节点,否则跳过创建操作,然后将本体的数据元素信息追加至知识森林体系中,本体节点与数据元素之间通过唯一标识创建关联索引;

对于关系模型数据导入任务,首先验证待导入关系在知识森林体系中是否已经存在,若尚未存在则创建一条新的关系边,否则跳过创建操作,然后将关系的数据元素信息追加至知识森林体系中,关系边与数据元素之间通过唯一标识创建关联索引;

对数据包的导入状态进行记录,支持重新执行失败的数据包导入任务。

2.根据权利要求1所述的方法,其特征在于,对搜索结果进行分类统计、字段排序、条件筛选和二次搜索,帮助用户快速、精准地定位到目标结果。

3.根据权利要求1所述的方法,其特征在于,所述数据元素规范的内容包括:数据元素中文名称、数据元素英文名称、数据元素存储类型、数据元素正则表达式;然后依据所述数据元素规范定义所述本体模型,所述本体模型包括本体模型代码、本体中文名称、本体英文名称、本体数据元素、本体标签库,并指定一个或多个数据元素作为本体模型实例的唯一标识;所述本体模型分为本体大类模型和本体小类模型,本体大类模型是指对一类客观事物的抽象概念描述,本体小类模型是在本体大类模型的基础上具象出来的对特定客观事物的描述,自动继承本体大类模型的所有数据元素,并允许按需追加数据元素。

4.根据权利要求3所述的方法,其特征在于,所述本体大类模型划分为三个层级,根据与人员本体大类的关联程度进行划分;其中,人员本体大类作为第一层级,与人员直接关联的本体大类作为第二层级,与人员无直接关联的本体大类作为第三层级。

5.根据权利要求1所述的方法,其特征在于,在所述本体模型之间定义所述关系模型,包括关系模型代码、关系中文名称、关系英文名称、起点本体模型、终点本体模型、关系数据元素;关系模型实例的唯一标识由以下三元组表达:(关系模型代码,起点本体模型实例唯一标识,终点本体模型实例唯一标识)。

6.根据权利要求1所述的方法,其特征在于,所述将多源海量数据融入知识森林体系,包括:

根据构建的数据绑定关系,对数据源中的数据进行抽取,并按照对应数据元素的规范要求进行格式转换,将多源海量数据导入统一的知识森林体系,数据导入工作支持增量式更新,更新频率能够按需设定。

7.根据权利要求1所述的方法,其特征在于,所述在构建的知识森林体系中进行跨数据源的关联搜索,包括:

1)如果用户未指定搜索范围或指定的搜索范围中包含第一层级本体大类,则使用搜索条件对第一层级本体大类进行搜索;

2)如果用户未指定搜索范围或指定的搜索范围中包含第二层级本体大类,则使用搜索条件对第二层级本体大类进行搜索;

3)在知识森林体系中对步骤2)的搜索结果进行关联查找,找出与第二层级本体大类存在关联的第一层级本体大类结果;

4)合并步骤1)与步骤3)的第一层级本体大类搜索结果,并对结果进行去重;

5)在知识森林体系中对步骤4)的搜索结果进行关联查找,找出与第一层级本体大类存在关联的第二层级本体大类结果;

6)合并步骤2)与步骤5)的第二层级本体大类搜索结果,并对结果进行去重;

7)将步骤4)与步骤6)得到的搜索结果按域分类组织,并按照与搜索条件的匹配度排序。

8.一种采用权利要求1～7中任一权利要求所述方法的以人员为中心的多源海量数据关联搜索系统,其特征在于,包括模型管理模块、数据治理模块和数据搜索模块;

所述模型管理模块负责定义数据元素、本体模型、关系模型,构建以人员为中心的知识森林体系;

所述数据治理模块负责将数据源与本体模型、关系模型进行绑定,并将多源海量数据融入知识森林体系;

所述数据搜索模块负责在构建的知识森林体系中进行跨数据源的关联搜索,支持跨域搜索和多域联搜。

9.根据权利要求8所述的系统,其特征在于,还包括结果筛选模块,所述结果筛选模块负责对搜索结果进行分类统计、字段排序、条件筛选和二次搜索,帮助用户快速、精准地定位到目标结果。

一种以人员为中心的多源海量数据关联搜索方法和系统

技术领域

[0001] 本发明涉及一种数据搜索方法，尤其涉及一种以人员为中心的多源海量数据关联搜索方法和系统。

背景技术

[0002] 随着社会信息化水平的不断提升，特别是公共安全领域的快速发展，各类传感器和感知源已经遍布了城市的每个角落。日常生活中的出行、住宿、交易、社交等活动已经全面进入了数字化时代，在为人们提供便捷服务的同时，也为公安机关积累了宝贵的数据资源。

[0003] 这些数据资源具有来源广、种类多、规模大、更新快的特点，如何有效组织这些数据、充分发挥数据价值，成为了公安机关近年来最为关注的课题。现有技术中，尚未有将多源海量数据融合成为统一的以人员为中心的知识体系的技术方案，从而无法实现跨数据源的关联搜索。

发明内容

[0004] 本发明公开一种以人员为中心的数据关联搜索方法和系统，包括数据存储方法和数据搜索方法，能够实现多源海量数据的动态关联和高效搜索。

[0005] 基于以上所述，本发明提供如下技术方案：

[0006] 一种以人员为中心的多源海量数据关联搜索方法，其步骤包括：

[0007] 定义数据元素规范、本体模型和关系模型，构建以人员为中心的知识森林体系；

[0008] 将数据源与本体模型、关系模型进行绑定，并将多源海量数据融入知识森林体系；

[0009] 在构建的知识森林体系中进行跨数据源的关联搜索。

[0010] 进一步地，上述方法具体包括以下步骤：

[0011] S1：制定数据元素规范，对于需要建模的数据字段进行统一命名、格式转换，如姓名、性别、出生日期、身份证号等字段。数据元素规范的内容包括：数据元素中文名称、数据元素英文名称、数据元素存储类型、数据元素正则表达式。

[0012] S2：依据步骤S1制定的数据元素规范定义本体模型，包括本体模型代码、本体中文名称、本体英文名称、本体数据元素、本体标签库，可以指定一个或多个数据元素作为本体模型实例的唯一标识(如身份证号、护照号、驾驶证号等均可作为人员本体模型实例“张三”的唯一标识)。其中，本体模型可分为本体大类模型和本体小类模型，本体大类模型是指对一类客观事物的抽象概念描述，如人员、车辆、手机等；本体小类模型是在本体大类模型的基础上具象出来的对特定客观事物的描述，自动继承本体大类模型的所有数据元素，并允许按需追加数据元素，如小轿车、客车、货车等本体小类模型都继承了车辆本体大类模型的所有数据元素，并允许追加核定载客、核定载重等特有数据元素。

[0013] S3：将步骤S2定义的本体大类模型划分为三个层级，根据与人员本体大类的关联程度进行划分，构建以人员为中心的知识森林体系。其中，人员本体大类作为第一层级，与

人员直接关联的本体大类作为第二层级(如证件、手机、车辆等),与人员无直接关联的本体大类作为第三层级(如车站、机场、基站等)。

[0014] S4:在步骤S2定义的本体模型之间定义关系模型,包括关系模型代码、关系中文名称、关系英文名称、起点本体模型、终点本体模型、关系数据元素。关系模型实例的唯一标识由以下三元组表达:(关系模型代码,起点本体模型实例唯一标识,终点本体模型实例唯一标识)。

[0015] S5:将步骤S2定义的本体模型或步骤S4定义的关系模型与数据源建立绑定关系,将模型中的数据元素与数据源的数据库表中的具体字段进行逐一对应。

[0016] S6:根据步骤S5构建的数据绑定关系,对数据源中的数据进行抽取,并按照对应数据元素的规范要求进行格式转换,将多源海量数据融入统一的知识森林体系(由本体模型和关系模型构成),数据导入工作支持增量式更新,更新频率可以按需设定。

[0017] S7:在步骤S6构建的知识森林体系中,用户可以通过关键字、筛选条件、本体标签等多种方式对数据进行搜索,支持跨域搜索(如通过车牌号关联搜索归属人,通过手机号关联搜索归属人等)和多域联搜(如通过姓名同时搜出相关人员、车辆、手机等),真正实现跨数据源的关联搜索。

[0018] 优选的,所述步骤S6中,数据导入方法为:

[0019] S61:为数据源添加流水编号自增序列,作为数据增量式导入的依据;

[0020] S62:从最后完成导入的流水编号开始,计算剩余待导入的数据量;

[0021] S63:对待导入数据进行分包封装(如2万条数据封装为1包),将待导入任务拆分为若干个数据包的导入任务;

[0022] S64:将数据包导入任务分发至大数据集群节点,实现多个数据包的并行导入;

[0023] S65:对于本体模型数据导入任务,首先验证待导入本体在知识森林体系中是否已经存在,若尚未存在则创建一个新的本体节点,否则跳过创建操作(如某个手机号码在通话清单中出现过100次,也仅会在第一次出现时创建本体节点),然后将本体的数据元素信息追加至知识森林体系中,本体节点与数据元素之间通过唯一标识创建关联索引;

[0024] S66:对于关系模型数据导入任务,首先验证待导入关系在知识森林体系中是否已经存在,若尚未存在则创建一条新的关系边,否则跳过创建操作(如两个手机号码之间通话过100次,也仅会在第一次出现时创建关系边),然后将关系的数据元素信息追加至知识森林体系中,关系边与数据元素之间通过唯一标识创建关联索引;

[0025] S67:对数据包的导入状态(成功/失败)进行记录,支持重新执行失败的数据包导入任务。

[0026] 优选的,所述步骤S7中,数据搜索方法为:

[0027] S71:如果用户未指定搜索范围或指定的搜索范围中包含第一层级本体大类,则使用搜索条件对第一层级本体大类进行搜索;

[0028] S72:如果用户未指定搜索范围或指定的搜索范围中包含第二层级本体大类,则使用搜索条件对第二层级本体大类进行搜索;

[0029] S73:在知识森林体系中对步骤S72的搜索结果进行关联查找,找出与第二层级本体大类存在关联的第一层级本体大类结果;

[0030] S74:合并步骤S71与步骤S73的第一层级本体大类搜索结果,并对结果进行去重;

[0031] S75:在知识森林体系中对步骤S74的搜索结果进行关联查找,找出与第一层级本体大类存在关联的第二层级本体大类结果;

[0032] S76:合并步骤S72与步骤S75的第二层级本体大类搜索结果,并对结果进行去重;

[0033] S77:将步骤S74与步骤S76得到的搜索结果按域分类组织(人员、车辆、手机等),并按照与搜索条件的匹配度排序。

[0034] 一种采用上面所述方法的以人员为中心的多源海量数据聚合搜索系统,包括模型管理模块、数据治理模块和数据搜索模块;

[0035] 所述模型管理模块定义数据元素、本体模型、关系模型,构建知识森林体系;

[0036] 所述数据治理模块负责将数据源与本体模型、关系模型进行绑定,对数据进行抽取、清洗和转换,将多源海量数据融入知识森林体系;

[0037] 所述数据搜索模块通过关键字、筛选条件、本体标签等多种方式对知识森林体系进行搜索,支持跨域搜索和多域联搜。

[0038] 进一步地,该系统还包括结果筛选模块,所述结果筛选模块负责对搜索结果进行分类统计、字段排序、条件筛选和二次搜索,帮助用户快速、精准地定位到目标结果。

[0039] 与现有技术相比,本发明的以人员为中心的知识森林体系和多源海量数据关联搜索方法,能够接入更多种类的数据源,实现更高效的聚合搜索,支撑更丰富的数据应用,极大提升大数据应用系统的数据兼容性和业务扩展性。

附图说明

[0040] 图1为本发明公开的以人员为中心的多源海量数据关联搜索方法流程示意图;

[0041] 图2为本发明原理结构示意图。

具体实施方式

[0042] 下面通过具体实施例和附图,对本发明做进一步详细说明。

[0043] 参照图1,本实施例提供的技术方案,具体步骤如下:

[0044] S1:制定数据元素规范,对于需要建模的数据字段进行统一命名、格式转换。数据元素规范的内容包括:数据元素中文名称、数据元素英文名称、数据元素存储类型、数据元素正则表达式。

[0045] 其中,数据元素存储类型是指数字、日期、文本等类型;数据元素正则表达式是指用于校验数据元素内容合法性的正则表达式,例如校验手机号码合法性的正则表达式为“^(\+86)?1[3-9]\d{9}$”。

[0046] S2:依据步骤S1制定的数据元素规范定义本体模型,包括本体模型代码、本体中文名称、本体英文名称、本体数据元素、本体标签库,可以指定一个或多个数据元素作为本体模型实例的唯一标识。其中,本体模型可分为本体大类模型和本体小类模型,本体大类模型是指对一类客观事物的抽象概念描述;本体小类模型是在本体大类模型的基础上具象出来的对特定客观事物的描述,自动继承本体大类模型的所有数据元素,并允许按需追加数据元素。

[0047] 其中,本体模型代码是指本体模型的全局唯一编码,如11000001;本体数据元素是指本体模型中的数据元素,如身份证号、姓名、性别、出生日期等;本体标签库是指本体模型

可用的分类标签，如服刑人员、涉毒人员、涉黑人员等。

[0048] S3：将步骤S2定义的本体大类模型划分为三个层级，根据与人员本体大类的关联程度进行划分，构建以人员为中心的知识森林体系。其中，人员本体大类作为第一层级，与人员直接关联的本体大类作为第二层级，与人员无直接关联的本体大类作为第三层级。

[0049] S4：在步骤S2定义的本体模型之间定义关系模型，包括关系模型代码、关系中文名称、关系英文名称、起点本体模型、终点本体模型、关系数据元素，关系模型实例的唯一标识由以下三元组表达：(关系模型代码，起点本体模型实例唯一标识，终点本体模型实例唯一标识)。

[0050] 其中，关系是指购买、拥有、驾驶等关系；关系模型代码是指关系模型的全局唯一编码，如21010003；起点本体模型是指关系起点所属的本体模型代码，如11000001(人员)；终点本体模型是指关系终点所属的本体模型代码，如13000001(车辆)；关系数据元素是指关系模型中的数据元素，如购买时间、购买金额等。

[0051] S5：将步骤S2定义的本体模型或步骤S4定义的关系模型与数据源建立绑定关系，将模型中的数据元素与数据源的数据库表中的具体字段进行逐一对应。

[0052] S6：根据步骤S5构建的数据绑定关系，对数据源中的数据进行抽取，并按照对应数据元素的规范要求进行格式转换，将多源海量数据融入统一的知识森林体系，数据导入工作支持增量式更新，更新频率可以按需设定，具体方法如下：

[0053] S61：为数据源添加流水编号自增序列，作为数据增量式导入的依据；其中，流水编号自增序列是指每新增一条数据都会自动增长的序列，如{10000,10001,10002,10003,…}；

[0054] S62：从最后完成导入的流水编号开始，计算剩余待导入的数据量；

[0055] S63：对待导入数据进行分包封装，将待导入任务拆分为若干个数据包的导入任务；

[0056] S64：将数据包导入任务分发至大数据集群节点，实现多个数据包的并行导入；其中，大数据集群节点是指部署了知识森林体系数据导入服务的大数据平台计算节点；

[0057] S65：对于本体模型数据导入任务，首先验证待导入本体在知识森林体系中是否已经存在，若尚未存在则创建一个新的本体节点，否则跳过创建操作，然后将本体的数据元素信息追加至知识森林体系中，本体节点与数据元素之间通过唯一标识创建关联索引；

[0058] S66：对于关系模型数据导入任务，首先验证待导入关系在知识森林体系中是否已经存在，若尚未存在则创建一条新的关系边，否则跳过创建操作，然后将关系的数据元素信息追加至知识森林体系中，关系边与数据元素之间通过唯一标识创建关联索引；

[0059] S67：对数据包的导入状态进行记录，支持重新执行失败的数据包导入任务；

[0060] S7：在步骤S6构建的知识森林体系中，用户可以通过关键字、筛选条件、本体标签等多种方式对数据进行搜索，支持跨域搜索和多域联搜，真正实现跨数据源的关联搜索，具体方法如下：

[0061] S71：如果用户未指定搜索范围或指定的搜索范围中包含第一层级本体大类，则使用搜索条件对第一层级本体大类进行搜索；

[0062] S72：如果用户未指定搜索范围或指定的搜索范围中包含第二层级本体大类，则使用搜索条件对第二层级本体大类进行搜索；

[0063] S73:在知识森林体系中对步骤S72的搜索结果进行关联查找,找出与第二层级本体大类存在关联的第一层级本体大类结果;

[0064] S74:合并步骤S71与步骤S73的第一层级本体大类搜索结果,并对结果进行去重;

[0065] S75:在知识森林体系中对步骤S74的搜索结果进行关联查找,找出与第一层级本体大类存在关联的第二层级本体大类结果;

[0066] S76:合并步骤S72与步骤S75的第二层级本体大类搜索结果,并对结果进行去重;

[0067] S77:将步骤S74与步骤S76得到的搜索结果按域分类组织,并按照与搜索条件的匹配度排序。

[0068] 如图2所示,本发明另一实施例提供一种以人员为中心的多源海量数据聚合搜索系统,该系统包括模型管理模块、数据治理模块、数据搜索模块和结果筛选模块。所述模型管理模块定义数据元素、本体模型、关系模型,构建知识森林体系。所述数据治理模块负责将数据源与本体模型、关系模型进行绑定,对数据进行抽取、清洗和转换,将多源海量数据融入知识森林体系。所述数据搜索模块通过关键字、筛选条件、本体标签等多种方式对知识森林体系进行搜索,支持跨域搜索和多域联搜。所述结果筛选模块负责对搜索结果进行分类统计、字段排序、条件筛选和二次搜索,帮助用户快速、精准地定位到目标结果。

[0069] 以表5中的实验数据为例,本发明的具体实施步骤如下:

[0070] S1:制定数据元素规范,具体内容见表1。

[0071] 表1

[0072]

中文名称	英文名称	存储类型	正则表达式
姓名	Name	文本	[\s\S]*
身份证号	IDNumber	文本	^(\d{6})(\d{4})(\d{2})(\d{2})(\d{3})([0-9]|X)$
手机号	PhoneNumber	文本	^(\+86)?1[3-9]\d{9}$
车牌号	PlateNumber	文本	^[\u4e00-\u9fa5]{1}[A-Z]{1}[A-Z0-9]{5}$
…	…	…	…

[0073] S2:定义本体模型,具体内容见表2。

[0074] 表2

[0075]

模型代码	**中文名称**	**英文名称**	**数据元素**	**标签库**
11000001	人员	Person	**身份证号**、姓名、性别、出生日期	服刑人员、在逃人员、涉毒人员、…
12000001	手机	Phone	**手机号**、IMSI、归属地	推销号码、诈骗号码、外卖号码、…
13000001	车辆	Vehicle	**车牌号**、车辆类型、归属地	涉案车辆、被盗车辆、…

[0076] S3:划分本体模型层级,其中人员本体模型作为第一层级,手机本体模型和车辆本体模型作为第二层级,暂无第三层级。

[0077] S4:定义关系模型,具体内容见表3。

[0078] 表3

[0079]

模型代码	中文名称	英文名称	起点本体模型	终点本体模型	数据元素
21000001	人员拥有手机关系	PersonOwnPhone	11000001	12000001	注册时间、注册单位
21000002	人员拥有车辆关系	PersonOwnVehicle	11000001	13000001	登记时间、登记单位

[0080] S5:将本体模型、关系模型与数据源(表5)建立绑定关系,具体内容见表4;其中,StartNode为起点本体模型实例唯一标识,EndNode为终点本体模型实例唯一标识。

[0081] 表4

[0082]

模型代码	数据元素	绑定字段
11000001	IDNumber	表5.身份证号
11000001	Name	表5.姓名
12000001	PhoneNumber	表5.手机号
13000001	PlateNumber	表5.车牌号
21000001	StartNode	表5.身份证号
21000001	EndNode	表5.手机号
21000002	StartNode	表5.身份证号
21000002	EndNode	表5.车牌号

[0083] S6:根据表4中的数据绑定关系对数据源中的数据进行抽取,并按照对应数据元素的规范要求进行格式转换,将数据融入统一的知识森林体系。

[0084] S7:用户可以通过关键字、筛选条件、本体标签等多种方式对数据进行搜索,支持跨域搜索(通过车牌号搜人、通过手机号搜人、通过身份证号搜车、通过身份证号搜手机等)和多域联搜(通过同一组关键字同时搜索人员、车辆、手机三个数据域),真正实现跨数据源的关联搜索。

[0085] 表5.实验数据

[0086]

姓名	身份证号	手机号	车牌号
张三	210103198603254817	13922438657	辽A35636
李四	110105199212123328	15801026678	京B62008
王五	450302199607221936	13662868530	桂A99096
…	…	…	…

[0087] 本发明中知识森林体系的层级划分方式可以根据不同业务场景需要进行灵活调整,例如可以将“以人员为中心”调整为“以案件为中心”,即案件本体大类作为第一层级,与案件直接关联的本体大类作为第二层级,与案件无直接关联的本体大类作为第三层级。

[0088] 以上所述,仅为本发明的最佳优选实施方式,对于本技术领域的普通技术人员而言,在不脱离本发明的原理的前提下,可以对上述实施细则进行多种变化、修饰、变型等,这些润色和改进也应视为本发明的保护范围。

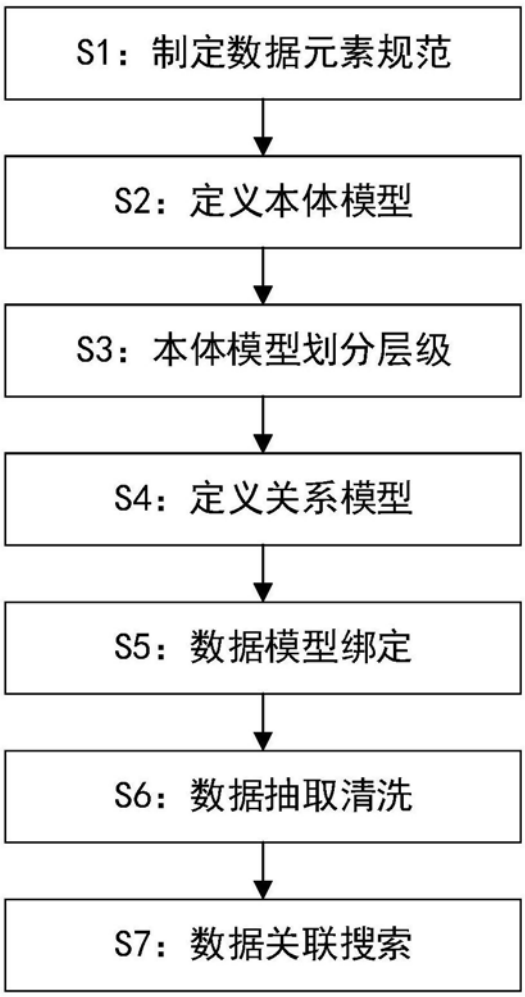

S1：制定数据元素规范
S2：定义本体模型
S3：本体模型划分层级
S4：定义关系模型
S5：数据模型绑定
S6：数据抽取清洗
S7：数据关联搜索

图1

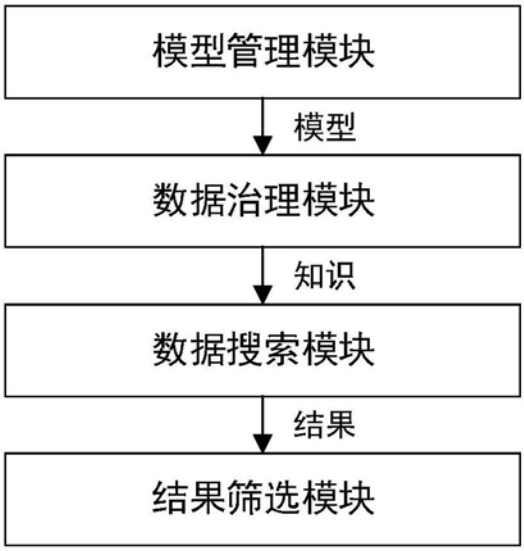

模型管理模块
模型
数据治理模块
知识
数据搜索模块
结果
结果筛选模块

图2