(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 108549731 A
(43)申请公布日 2018.09.18

(21)申请号 201810754428.3

(22)申请日 2018.07.11

(71)申请人 中国电子科技集团公司第二十八研究所
地址 210003 江苏省南京市白下区苜蓿园东街1号

(72)发明人 朱峰 李磊 鲁兴河 李青山

(74)专利代理机构 南京苏高专利商标事务所(普通合伙) 32204
代理人 向文

(51)Int.Cl.
G06F 17/30(2006.01)

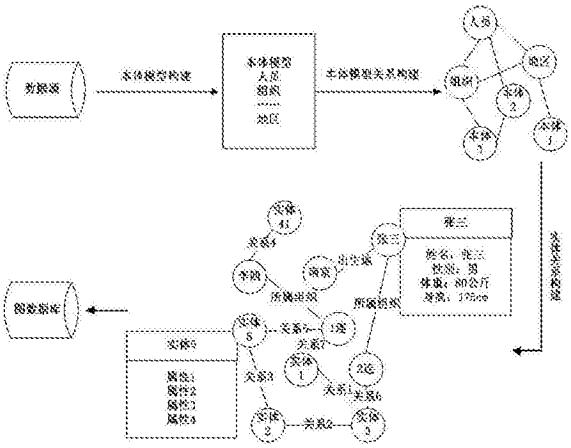
权利要求书1页 说明书6页 附图3页

(54)发明名称

一种基于本体模型的知识图谱构建方法

(57)摘要

本发明公开了一种基于本体模型的知识图谱构建方法，从充分利用现有指挥信息系统内数据价值的角度出发，通过统一的数据访问方法，将存储在关系型数据库内数据构建成多个本体模型，然后利用原有数据库表内字段之间的关系构建本体模型间关系，完成现有数据架构下的数据关系图谱构建，接着获取本体模型下所有的实体数据，利用本体模型关联参数构建实体数据关系，形成实体数据关系网，最终将本体模型、实体数据和关系按照邻接表的方式存入到图数据库中，并实现基于图结构的索引技术。本发明能够实现对任一关系型数据库内数据本体和实体构建关系，快速构建知识图谱，辅助业务人员掌握数据关系，充分实现对现有数据的利用。



CN 108549731 A

1.一种基于本体模型的知识图谱构建方法,其特征在于:包括以下步骤:

1)通过统一数据访问方法获取关系型数据库的表结构信息,并结合这些信息依据数据实际意义构建本体数据模型;

2)依据本体模型字段间的关系构建本体模型关系;

3)依据本体模型关系采用统一数据访问方法获取本体模型对应的实体数据并依据字段关系构建实体数据关系;

4)将上述步骤产生的图顶点和边信息按照邻接表的形式存入图数据库;

5)对图数据中的顶点数据和边数据分别按照其数据特点构建索引,提高对图数据的访问速度。

2.根据权利要求1所述的一种基于本体模型的知识图谱构建方法,其特征在于:所述步骤1中会显示数据库内所有原始信息,包括所有的表名称、每张表的字段信息、表的注释和所有字段的注释,利用这些信息按照表的实际意义构建本体模型,形成包括多个要素的本体模型,并设定每个模型的输出字段。

3.根据权利要求1或2所述的一种基于本体模型的知识图谱构建方法,其特征在于:所述步骤2中本体模型关系的构建方法步骤如下:

2.1)选择需要建立关系的多个数据模型。

2.2)选择每个模型的关联字段,建立字段之间的关系。

2.3)将本体模型关系存储入图数据库中,存入的信息包括本体模型的字段信息、参与关联的模型名称、关联的参数。

4.根据权利要求1所述的一种基于本体模型的知识图谱构建方法,其特征在于:所述步骤3中根据本体模型对应的表结构信息通过统一数据访问方法获取存储在数据库中的所有数据,然后依据设定的字段关系构建实体关系。

5.根据权利要求1或4所述的一种基于本体模型的知识图谱构建方法,其特征在于:所述步骤3中实体关系的构建方法步骤如下:

3.1)对参与构建关系的每个本体模型通过统一的数据访问接口获取所有的数据;

3.2)由数据库表中对于表的注释和对于表中字段的注释,将实体数据由英文属性名转为中文属性名;

3.3)将所有本体模型的实体数据存入图数据库中;

3.4)利用本体模型的关联参数构建实体关系;

3.5)重复步骤3.1到步骤3.4,直至所有的本体模型关系都完成了对应实体关系的构建。

6.根据权利要求1所述的一种基于本体模型的知识图谱构建方法,其特征在于:所述步骤4以顶点为单位按照邻接表的方式进行图数据的存储,顶点按照其id值按序进行存储,顶点关联的边信息会存储在顶点属性之后,并按照固定的格式存储。

7.根据权利要求1所述的一种基于本体模型的知识图谱构建方法,其特征在于:所述步骤5对图数据中的边和顶点分别建立索引,按照属性值是数值类型还是文本类型采用不同的索引方法,同时针对确定性查询和范围性查询采用不同的索引技术以根据查询条件快速定位到符合条件的边或者顶点。

一种基于本体模型的知识图谱构建方法

技术领域

[0001] 本发明涉及信息系统数据管理技术领域，具体涉及一种基于本体模型的知识图谱构建方法。

背景技术

[0002] 通过多年作战指挥信息系统的建设，在信息服务中心内存储、接入了各类基础数据、业务信息、战略支援信息等、已经初步显现了数据/信息的汇集作用。但是，现有的数据汇聚手段，支撑决策指挥的知识作用仍然严重不足，静态数据与动态数据之间、各类数据之间并未建立完整的关联，面向任务的数据检索能力较弱，数据价值未充分挖掘，相关的理论、方法和技术研究还比较薄弱，急需开展基于现有数据的知识图谱构建研究工作。

[0003] 知识图谱是在传统知识工程的基础上以及语义Web的发展中孕育并发展而来的知识表示技术，其旨在描述客观世界的概念、实体、事件及其之间的关系。本质上，知识图谱是一种揭示实体之间关系的语义网络，可以对现实世界的事物及其相互关系进行形式化地描述。知识图谱亦可被看做是一张巨大的图，图中的节点表示实体或概念，而图中的边则由属性或关系构成。现在的知识图谱已被用来泛指各种大规模的知识库。知识图谱技术逐步渗透到各个领域。同时，随着作战保障和业务处理系统稳步发展，各类数据资源逐渐丰富。各领域军事应用需求的不断增长，作战指挥、作战保障和日常业务处理信息系统建设投入不断加大，各类作战保障和业务处理信息系统规模逐步扩展，积累形成了一批可用、实用的信息资源，成为构建知识图谱的重要支撑。因此，建立军事知识图谱，实现对数据的高效组织管理和智能检索服务势在必行

发明内容

[0004] 发明目的：为了克服现有技术中存在的不足，在现有的数据基础上通过本体构建工具建立起包括组织、人员、设施等本体概念，同时提供基于本体的统一数据访问能力，然后通过对本体间建立关系，利用本体概念下的参数关系，构建实体知识图谱，为信息系统的数据资源利用提供技术保障。

[0005] 技术方案：为实现上述目的，本发明提供一种基于本体模型的知识图谱构建方法，其具体包括如下技术要点：

[0006] 1、提供基于关系型数据库的本体模型构建功能

[0007] 能够通过配置数据库连接获取库内所有的表结构信息，通过选择对应的数据库表，并通过表字段关联构建本体模型。

[0008] 2、本体模型关系构建功能

[0009] 利用建立好的本体模型，通过本体模型之间的关联字段构建关系。

[0010] 3、本体模型下的实体数据采集功能

[0011] 通过统一的访问接口，可以以本体为单位获取存储在数据库内的实体数据。

[0012] 4、本体模型下的实体数据关系构建功能

[0013] 本体模型关系构建好之后，通过数据访问服务获取存储在数据库内的实体数据信息，将这些实体数据存储在图数据中，然后根据本体模型对应的关联字段构建实体数据间的关系。

[0014] 5、图数据存储和索引方法

[0015] 针对图数据的特点，对图中边和顶点分别设计了存储方式，减少了对存储空间的要求，同时，为方便对图中数据的访问，对边和顶点分别设计了合理有效的索引方式，提高了图数据库的访问速度。

[0016] 一种基于本体模型的知识图谱构建方法，其包括以下步骤：

[0017] 1)通过统一数据访问方法获取关系型数据库的表结构信息，并结合这些信息依据数据实际意义构建本体数据模型；

[0018] 2)依据本体模型字段间的关系构建本体模型关系；

[0019] 3)依据本体模型关系采用统一数据访问方法获取本体模型对应的实体数据并依据字段关系构建实体数据关系；

[0020] 4)将上述步骤产生的图顶点和边信息按照邻接表的形式存入图数据库；

[0021] 5)对图数据中的顶点数据和边数据分别按照其数据特点构建索引，提高对图数据的访问速度，提高对图数据的访问速度。

[0022] 进一步地，所述步骤1中显示数据库内所有原始信息，包括所有的表名称、每张表的字段信息、表的注释和所有字段的注释，业务人员利用这些信息按照表的实际意义构建本体模型，形成包括组织、地名、人员、设施等本体模型，并设定每个模型的输出字段。

[0023] 进一步地，所述步骤2依据步骤1构建好的本体模型，利用本体模型的输出字段构建关系，如组织模型中的组织内码字段等于组织人员关系模型中的组织内码字段、组织人员关系模型中的人员内码字段等于人员模型中的人员内码字段。

[0024] 进一步地，所述步骤3中根据本体模型对应的表结构信息通过统一数据访问方法获取存储在数据库中的所有数据，如获取所有的组织数据、地名数据、人员数据等，然后依据设定的字段关系构建实体关系，如组织A和人员B之间有组织下属人员关系，人员B和地名C有出生地的关系等。

[0025] 进一步地，所述步骤4将以顶点为单位按照邻接表的方式进行图数据的存储，顶点按照其id值按序进行存储，顶点关联的边信息会存储在顶点属性之后，并按照固定的格式存储。

[0026] 进一步地，所述步骤5会对图数据中的边和顶点分别建立索引，按照属性值是数值类型还是文本类型采用不同的索引方法，同时针对确定性查询和范围性查询采用不同的索引技术以根据查询条件快速定位到符合条件的边或者顶点。

[0027] 有益效果：本发明与现有技术相比，具备如下优点：

[0028] 1、可以创建各个领域内的知识图谱，对数据种类没有要求。

[0029] 2、能够利用数据表内的注释将本体和实体转为更容易理解的中文表示方式。

[0030] 3、能够已知识图谱的方式将掌握在少数人手中的知识传播给他人。

[0031] 4、能够有效的节省图谱存储空间。

[0032] 5、能够快速查询到图谱中的数据。

附图说明

[0033] 图1为基于本体模型构建知识图谱方法流程示意图；

[0034] 图2为人员本体模型示意图；

[0035] 图3为人员与组织本体模型关系示意图；

[0036] 图4为图存储采用的邻接表示意图；

[0037] 图5为图边和属性存储格式示意图；

[0038] 图6为图索引方法示意图。

具体实施方式

[0039] 下面结合附图和具体实施例，进一步阐明本发明。

[0040] 首先给出本发明中使用的本体和实例概念的定义，以帮助理解本发明的实施方式。

[0041] 实体：是数据空间中基本的数据表示单位，描述了现实世界中的一个对象，由一个或多个<属性，值>对集合组成。实体属性值的三元组的表示形式如下：(实体，属性，值)。如某人员实体张三，其包含姓名、年龄、性别等属性，每个属性有对应的值。

[0042] 本体：描述了现实世界中同种类型实体的一个抽象概念，指出一个实体应该属于的类型或类，即任何一个实体类都属于某个本体，一种本体类包含一个或多个实体。如人员这个本体包含了张三、李四等多个实体。

[0043] 关联关系：描述两个实体之间或两个本体之间的各种语义关系(联系)，分别称作实体关联关系和本体关联关系。

[0044] 如图1所示，本实施例中首先利用数据访问工具获取数据库内所有的表结构信息，利用这些表构建出一个个本体模型，然后对这些本体模型构建关系，指明参与关联的本体和关联参数，接着获取本体对应的所有实体数据并利用关联参数信息构建关系，最后将所有的本体数据、实体数据和关系数据存入图数据库中。

[0045] 根据上述内容，本实施例的方法中可依次分为本体模型构建方法、本体模型关系构建方法、实体关系构建方法、基于邻接表的图数据存储方法和基于图结构的索引技术五个技术要点，下面对这五个技术要点做详细说明。

[0046] 1、本实施例中本体模型构建方法具体的解释和操作如下：

[0047] 参照图1，存在于数据库中的各基础和业务数据通常包含各种本体模型，如人员、设施、地名等，这些本体多以表为单位进行存储，本体之间的关系通过主外键进行关联。用户在访问数据时需要对数据库中的表结构有一定的了解，在此基础上获取数据，支撑自己业务系统的运转，这样就增加了数据的使用和维护成本。为解决此问题，本实施例提供了一种配置化的本体模型构建工具，此工具首先获取数据库用户空间下的所有表结构，用户根据表的存储信息构建本体模型，再通过字段关联将关联信息加入到本体模型中，这样就形成了多个独立的本体模型，具体如图2所示，使得数据库使用人员能够迅速获取到数据库内的数据结构信息，然后根据需求进行数据的访问。

[0048] 2、本实施例中本体模型关系构建方法具体的解释和操作如下：

[0049] 参照图1，本体模型构建完成后会形成多个独立的本体，例如人员、设施、组织、地

区等，这些本体之间存在诸多的关系，如人员的出生地、人员所属组织、组织所属地区等，这些关系在数据库中一般通过关联表实现，如建立一张人员与组织的关系表，表结构为人员内码、组织内码，一行数据就表示了某个人员所属组织。利用这些关系表构建出的关系网隐藏在数据库表结构中，使用不便，也无法直观的表现本体之间的关系，为解决此问题，本实施例提出了一种基于知识图谱的本体模型关系构建方法，其实施步骤如下所示：

[0050] 1)选择需要建立关系的多个数据模型，模型的数量不定，如M_1、M_2、M_3……M_N。

[0051] 2)选择每个模型的关联字段，建立字段之间的关系，此关系可以是相等关系，如内码相等，也可以是其它复杂关系，如子字符串、取模计算等，支持的关系类型如表1所示。

[0052] 3)将本体模型关系存储入图数据库中，存入的信息包括本体模型的字段信息、参与关联的模型名称、关联的参数，图3为建立好的人员本体与组织本体关系示意图。

[0053]

数学计算	+，-，*，/，%	加减乘除
	abs()	取绝对值
	sqrt()	开根号
	sin()，cos()，tan()，cot()，asin()，acos()，atan	三角函数
	log10()，log()，exp()，e()	对数函数
布尔计算	and，or，xor，not	条件运算
字符串计算	substring(original,begin,sublength)	获取子字符串
	left(original,sublength)	左子字符串
	right(original,sublength)	右子字符串
	upper()，lower()	字符串大小写转换
	length(string)	长度计算
	reverse(original)	字符串反转

[0054] 表1

[0055] 3、本实施例中实体关系构建方法具体的解释和操作如下：

[0056] 在本体模型关系构建之后，就可以根据关系参数构建实体关系，构建方法如下：

[0057] A)对参与构建关系的每个本体模型通过统一的数据访问接口获取所有的数据；

[0058] B)由数据库表中对于表的注释和对于表中字段的注释，将实体数据由英文属性名转为中文属性名，如组织实体中英文字段“zzmc”转为中文字段名称“组织名称”，使所有数据表现更为直观；

[0059] C)将所有本体模型的实体数据存入图数据库中；

[0060] D)利用本体模型的关联参数构建实体关系，例如对于组织、人员组织关系、人员这

三个本体，如果某个组织实体的组织内码等于人员组织关系实体的组织内码且此人员组织关系实体的人员内码等于某个人员实体的人员内码，则在此组织实体和人员实体之间构建组织下属人员关系；

[0061] E)重复步骤A到步骤D，直至所有的本体模型关系都完成了对应实体关系的构建。

[0062] 4、本实施例中基于邻接表的图数据存储方法具体的解释和操作如下：

[0063] 图数据常用的存储结构有邻接矩阵、邻接表。邻接矩阵以一个二维数组来表示图中顶点的连通性，优点是直观简洁，能够快速查找到两个顶点的连通性，但是存储代价高昂，即时顶点间没有连接也需要空间去存储，在大数据量下，空间浪费尤其严重。邻接表将顶点的连接顶点以链表方式存储，存储开销小，逻辑简单，便于分割处理，在海量图数据库中有较大优势，所以本实施例提出了一种采用邻接表作为图的存储方法。

[0064] 本实施例采用的邻接表如图4所示，每一行代表一个顶点，以顶点id作为存储的key，顶点的属性和邻接边作为value的独立单元，以方便属性和边的删除、修改操作，value的长度并没有限制。顶点按照顶点id进行按序存储，以实现顶点的快速查询。

[0065] 图中边和属性的存储方式如图5所示，每条边和属性都独立存储在连接表中。每个参数都被序列化以减少存储开支。边的标签id、连接顶点id、边id和属性的key id、属性id都以二进制数字的格式被编码，其实际值以索引方式存储在独立的空间中，这样能有效的减少它们占用的存储空间，而排序字段和属性值以字符串的方式存储。标签id后一位比特用于表示边的方向，连接顶点id并部存储顶点的实际id，而是相对与邻接顶点的差值，这样也能减少其占用空间。因为边的属性个数是不定的，所以存储边的长度也是可变的

[0066] 5、基于图结构的索引技术具体的解释和操作如下：

[0067] 为实现对图数据的快速查询，必须对图中的顶点和边建立索引。常见的索引技术如B+树索引、Hash索引和位图索引等被广泛应用于数据存储和查询中，但这些索引技术都没有结合图数据结构的连接性特点。如图6所示，本实施例结合自身数据存储格式的特点，通过建立全图索引和顶点内索引，有效的支撑了图的遍历查询，实现了对关系和数据的快速查询。

[0068] 针对大部分对图的查询都是基于某个属性查询符合条件的顶点和边的特点，首先对图中的顶点和边属性进行全图索引。一般查询条件分为两种，一种为确定性查询，如判断字符串和数值的相等、大于、小于，另一种为范围性查询，如判断字符串是否包含某个特定子串，某个数值在一定的范围内。为提高建立索引的速度，针对这两种情况分别提供复合索引和混合索引两种不同的索引建立方法。同时提供建立联合索引的方法，即在多个属性上建立关联索引，通过对属性值的联合查询能够快速定位到符合条件的边和顶点。

[0069] 由于顶点和边的属性大部分都为文本格式，所以在对图进行查询时很多都是基于文本匹配的查询，因此，针对文本索引进行了特殊化的处理。在对文本建立索引时，必须指定建立的是全文索引还是字符串索引。全文索引在建立时会将字符串值进行标签化，标签化的方法用户可以自己指定，默认情况下会将字符串以非字符文本进行切割，然后去除长度小于2的标签。全文索引支持三种查询方式：标签化后的字符串某个标签包含某个指定的子串、以某个子串开头和结尾、复合某个指定的正则表达式。字符串索引不会对文本进行标签化，以整个文本的值建立索引，其支持的查询方式有四种：文本相等、文本不等、文本以某个给定字符开头和结尾、文本匹配给定的正则表达式。通过对文本建立不同的索引方式，可

以大大减少建立索引的开销，同时也会加快对文本的查询。

[0070] 顶点内索引是针对每个顶点的数据进行索引。在海量图数据中，一个顶点可能有成千上万条边，对这个顶点进行边遍历时非常耗时，因为需要查看每条边是否符合查询条件，通过建立顶点内索引可以解决这个问题。在顶点的存储模型中，每个顶点都存储了它所有的相邻边，相同标签的边会存储在一起，通过指定相同标签表的排序属性，可以按照某个属性值对边进行降序或者增序排序，这样，在根据属性值查找复合条件的边时可以通过二分查找、递归查找等算法进行加速。参与排序的属性值也可以多个，此时会在第一个属性值相等条件下按第二个属性值进行排序，以此类推。

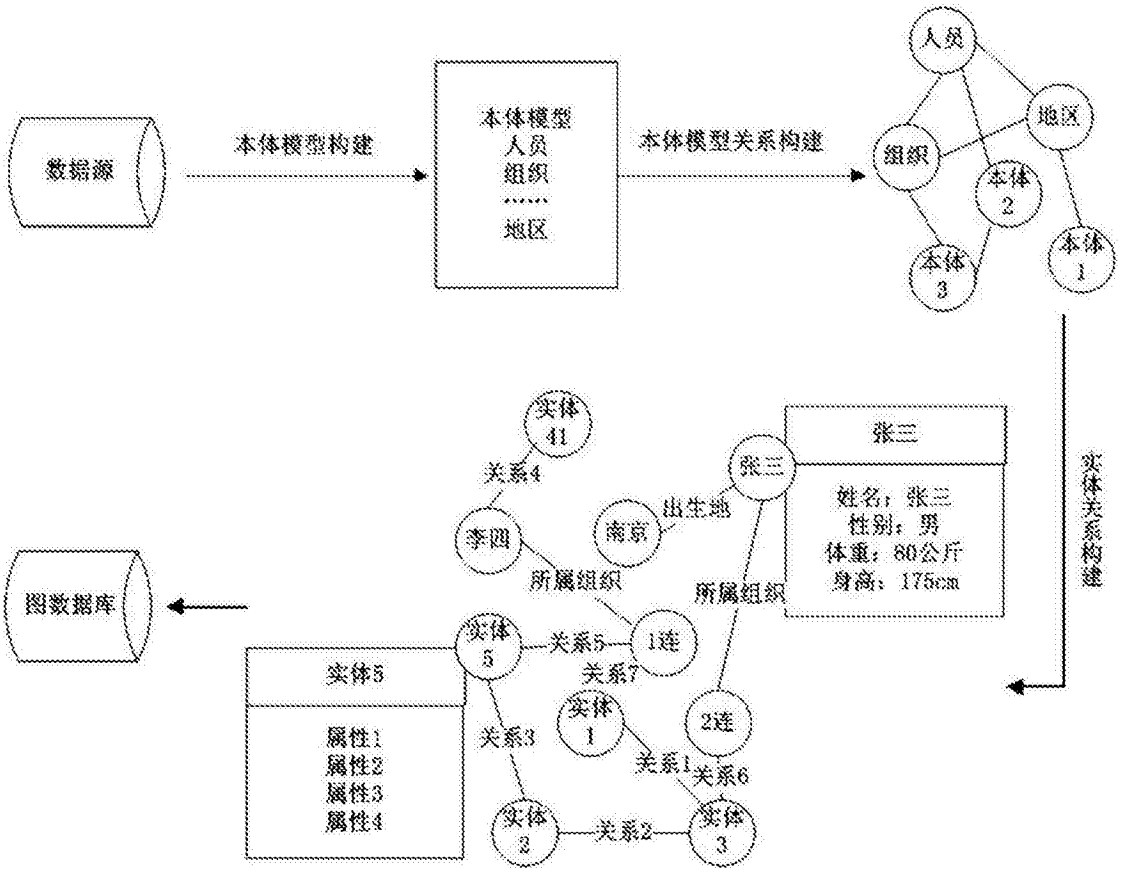

数据源
本体模型构建
本体模型
人员
组织
.....
地区
本体模型关系构建
人员
地区
组织
本体2
本体3
本体1
实体关系构建
图数据库
实体41
关系4
李四
南京
出生地
张三
姓名：张三
性别：男
体重：80公斤
身高：175cm
所属组织
所属组织
实体5
关系5
1连
关系7
实体5
属性1
属性2
属性3
属性4
关系3
实体1
2连
关系1
关系6
实体2
关系2
实体3

图1

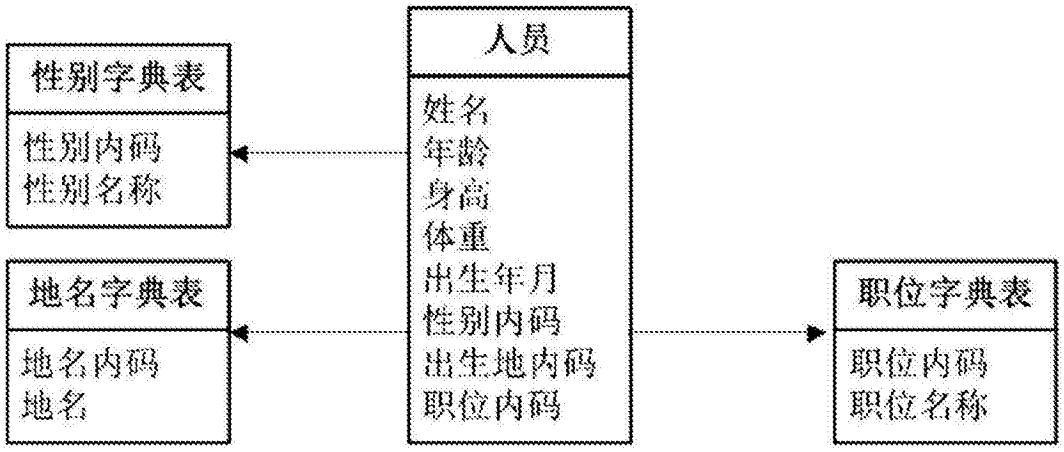

性别字典表
性别内码
性别名称
人员
姓名
年龄
身高
体重
出生年月
性别内码
出生地内码
职位内码
地名字典表
地名内码
地名
职位字典表
职位内码
职位名称

图2

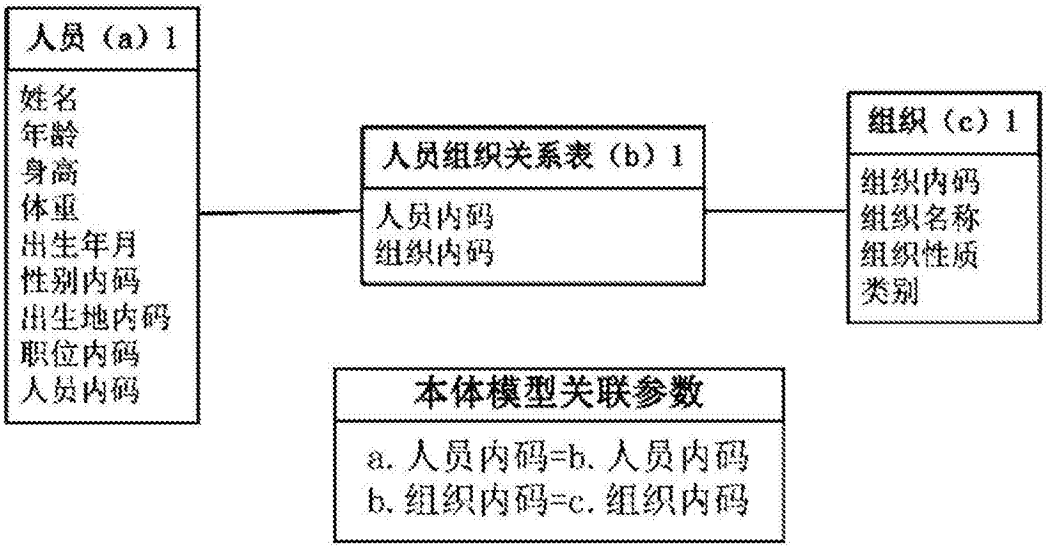

人员（a）1
姓名
年龄
身高
体重
出生年月
性别内码
出生地内码
职位内码
人员内码
人员组织关系表（b）1
人员内码
组织内码
组织（c）1
组织内码
组织名称
组织性质
类别
本体模型关联参数
a. 人员内码=b. 人员内码
b. 组织内码=c. 组织内码

图3

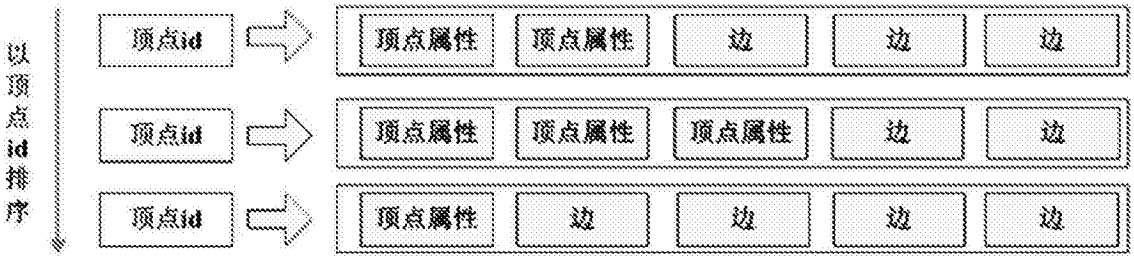

以顶点id排序
顶点id
顶点属性
顶点属性
边
边
边
顶点id
顶点属性
顶点属性
顶点属性
边
边
顶点id
顶点属性
边
边
边
边

图4

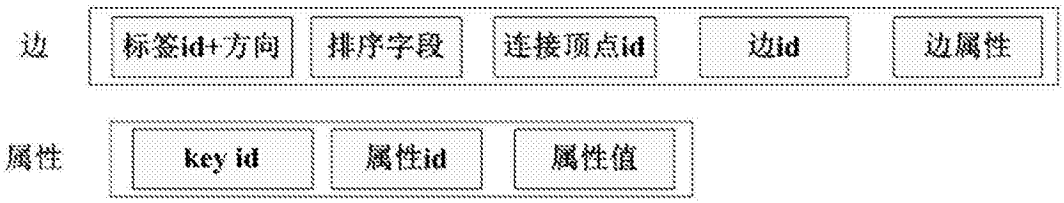

边
标签id+方向
排序字段
连接顶点id
边id
边属性
属性
key id
属性id
属性值

图5

<table>
<tr><th colspan="3">图索引</th></tr>
<tr><th colspan="2">全图索引</th><th>顶点内索引</th></tr>
<tr><td rowspan="2">数值型属性</td><td>复合索引（确定性查询）
混合查询（范围性查询）</td><td rowspan="2">按边的属性值进行排序</td></tr>
<tr><td>文本型属性</td><td>全文索引
字符串索引</td></tr>
</table>

图6