(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 119294488 A
(43)申请公布日 2025.01.10

(21)申请号 202411310963.1

(22)申请日 2024.09.19

(71)申请人 中煤科工开采研究院有限公司
地址 101399 北京市顺义区中关村科技园区顺义园临空二路1号

(72)发明人 吕依濛

(74)专利代理机构 北京清亦华知识产权代理事务所(普通合伙) 11201
专利代理师 汤宝平

(51)Int.Cl.
G06N 5/022(2023.01)

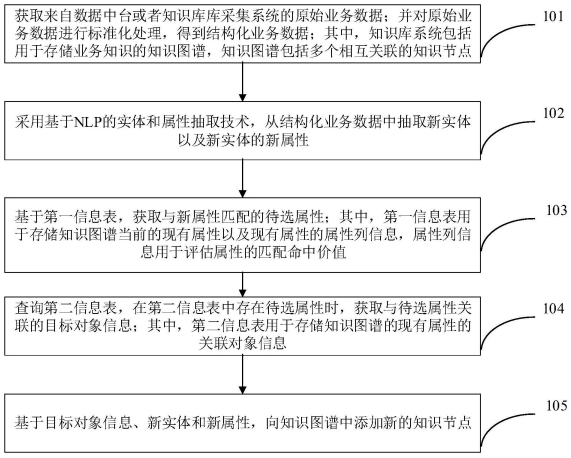
权利要求书3页 说明书10页 附图2页

(54)发明名称

用于知识库系统的图谱实体自动关联构建的方法以及装置

(57)摘要

本申请提出一种用于知识库系统的图谱实体自动关联构建的方法以及装置,其中,方法包括:获取来自数据中台或者知识库库采集系统的原始业务数据;对原始业务数据进行标准化处理,得到结构化业务数据;知识库系统包括用于存储业务知识的知识图谱;从结构化业务数据中抽取新实体以及新实体的新属性;基于第一信息表获取与新属性匹配的待选属性;第一信息表用于存储知识图谱当前的现有属性以及现有属性的属性列信息;查询第二信息表获取与待选属性关联的目标对象信息;基于目标对象信息、新实体和新属性,向知识图谱中添加新的知识节点。实现了图谱实体的自动关联构建,提升了图谱的构建效率。



CN 119294488 A

1.一种用于知识库系统的图谱实体自动关联构建的方法,其特征在于,包括以下步骤:

获取来自数据中台或者知识库库采集系统的原始业务数据;并对所述原始业务数据进行标准化处理,得到结构化业务数据;其中,所述知识库系统包括用于存储业务知识的知识图谱,所述知识图谱包括多个相互关联的知识节点,每个所述知识节点对应一个实体,每个实体包括至少一个属性,实体之间通过关系关联;

采用基于NLP的实体和属性抽取技术,从所述结构化业务数据中抽取新实体以及所述新实体的新属性;

基于第一信息表,获取与所述新属性匹配的待选属性;其中,所述第一信息表用于存储所述知识图谱当前的现有属性以及所述现有属性的属性列信息,所述属性列信息用于评估属性的匹配命中价值;

查询第二信息表,在所述第二信息表中存在所述待选属性时,获取与所述待选属性关联的目标对象信息;其中,所述第二信息表用于存储所述知识图谱的现有属性的关联对象信息;

基于所述目标对象信息、所述新实体和所述新属性,向所述知识图谱中添加新的知识节点。

2.根据权利要求1所述的方法,其特征在于,所述基于第一信息表,获取与所述新属性匹配的待选属性;包括:

从所述第一信息表中的现有属性中,获取与所述新属性相似度满足相似度阈值的多个相似属性;

对于所述多个相似属性中的各相似属性,基于该相似属性的属性列信息,计算多个构建参数值;

基于所述多个构建参数值与各构建参数值对应的预设阈值,得到该相似属性的匹配值;

基于所述多个相似属性各自的匹配值,从所述多个相似属性中确定匹配值满足预设条件的待选属性。

3.根据权利要求2所述的方法,其特征在于,所述从所述第一信息表中的现有属性中,获取与所述新属性相似度满足相似度阈值的多个相似属性;包括:

通过距离相似度算法计算所述新属性与所述第一信息表中现有属性中各属性的相似度,获取与所述新属性相似度满足相似度阈值的多个相似属性;

获取所述多个相似属性中各相似属性与所述新属性的距离值L。

4.根据权利要求3所述的方法,其特征在于,所述属性列信息包括实体值、属性值、抽取时间、构建量、构建时间、人工纠错量、人工纠错时间和是否高价值段,其中,所述实体值用于记录图谱中实体的值,所述属性值用于记录该实体中某个属性的值;所述抽取时间用于记录该属性抽取的时间,所述构建量用于记录该属性构建图谱的次数,所述构建时间用于记录该属性最近一次构建时的时间;所述人工纠错量用于记录该属性构建图谱错误时的字段进行人工干预的次数;所述人工纠错时间用于记录最近一次人工纠错量变更时的指标修改时间;所述是否高价值段用于识别该属性是否为高价值数据。

5.根据权利要求4所述的方法,其特征在于,所述基于该相似属性的属性列信息,计算多个构建参数值;包括:

基于所述构建时间,通过第一公式得到构建量参数y1;所述第一公式表示如下:

$y1=2\mathrm{arccot}(build\_number(x))/\pi$

其中,build_number表示构建时间,x表示相似属性;

基于所述人工纠错量,通过第二公式得到人工纠错量参数y2;所述第二公式表示如下:

$y2=2\arctan(correct\_number(x))/\pi$

其中,correct_number表示人工纠错量,x表示相似属性;

基于所述构建时间和所述抽取时间,通过第三公式得到构建时间参数t1;所述第三公式表示如下:

$t1=2\mathrm{arccot}(build\_time(x)-time(x))/\pi$

其中,build_time表示构建时间,time表示抽取时间,x表示相似属性;

基于所述人工纠错时间和所述抽取时间,通过第四公式得到人工纠错时间参数t2;所述第四公式表示如下:

$t2=2\arctan(correct\_time(x)-time(x))/\pi$

其中,correct_time表示人工纠错时间,time表示抽取时间,x表示相似属性;

基于所述是否高价值段,通过第五公式得到高价值参数m;所述第五公式表示如下:

$m=is\_high$

其中,is_high表示是否高价值段;

基于所述距离值L,通过第六公式得到距离参数s;所述第六公式表示如下:

$s=(threshold-L)/threshold$

其中,threshold表示相似度阈值。

6.根据权利要求1所述的方法,其特征在于,所述第二信息表包括第一实体、第一属性、创建时间、关系、对象实体和对象属性,所述创建时间用于记录所述关系的创建时间,所述关系用于记录所述第一实体与所述对象实体之间的关系类型;所述查询第二信息表,在所述第二信息表中存在所述待选属性时,获取与所述待选属性关联的目标对象信息;包括:

将所述待选属性的值作为所述第一属性的值查询所述第二信息表,确定所述第二信息表中是否存在所述待选属性;

在确定所述第二信息表中存在所述待选属性时,获取目标对象实体、目标对象属性和目标关系。

7.根据权利要求6所述的方法,其特征在于,所述基于所述目标对象信息、所述新实体和所述新属性,向所述知识图谱中添加新的知识节点;包括:

将所述新实体及所述新属性通过所述关系建立与所述对象实体及所述对象属性相关联的图谱关系。

8.一种用于知识库系统的图谱实体自动关联构建的装置,其特征在于,包括:

数据获取模块,用于获取来自数据中台或者知识库库采集系统的原始业务数据;并对所述原始业务数据进行标准化处理,得到结构化业务数据;其中,所述知识库系统包括用于存储业务知识的知识图谱,所述知识图谱包括多个相互关联的知识节点,每个所述知识节点对应一个实体,每个实体包括至少一个属性,实体之间通过关系关联;

所述数据获取模块,还用于采用基于NLP的实体和属性抽取技术,从所述结构化业务数据中抽取新实体以及所述新实体的新属性;

属性匹配模块,用于基于第一信息表,获取与所述新属性匹配的待选属性;其中,所述第一信息表用于存储所述知识图谱当前的现有属性以及所述现有属性的属性列信息,所述属性列信息用于评估属性的匹配命中价值;

对象获取模块,用于基于第二信息表,在所述第二信息表中存在所述待选属性时,获取与所述待选属性关联的目标对象信息;其中,所述第二信息表用于存储所述知识图谱的现有属性的关联对象信息;

图谱构建模块,用于基于所述目标对象信息、所述新实体和所述新属性,向所述知识图谱中添加新的知识节点。

9.一种电子设备,其特征在于,包括:处理器,以及与所述处理器通信连接的存储器;

所述存储器存储计算机执行指令;

所述处理器执行所述存储器存储的计算机执行指令,以实现如权利要求1-6中任一项所述的方法。

10.一种计算机可读存储介质,其特征在于,所述计算机可读存储介质中存储有计算机执行指令,所述计算机执行指令被处理器执行时用于实现如权利要求1-6中任一项所述的方法。

用于知识库系统的图谱实体自动关联构建的方法以及装置

技术领域

[0001] 本申请涉及图谱构建技术领域,尤其涉及一种用于知识库系统的图谱实体自动关联构建的方法、装置、电子设备及存储介质。

背景技术

[0002] 随着信息化技术的迅速发展,知识库应用和管理系统成为了学术研究和技术开发的重要工具。这些系统存储了大量的研究报告、技术文档和数据资料,为科研人员提供了便捷的资源共享和知识管理平台。

[0003] 煤知识图谱作为一种新兴的数据表示和组织方式,已经在多个领域取得成功应用,如互联网搜索、推荐系统、问答系统等。图数据库的发展,使得存储和查询大规模图谱数据变得更加高效和便捷。矿开采涉及大量的实体和关系,包括矿井、设备、工人、安全规程、环境监测等。这些实体和关系之间的信息复杂且多样,传统的手工管理和记录方式难以有效处理和利用这些信息。

[0004] 而煤矿企业内部通常有大量的数据分散在不同系统和部门之间,通过知识图谱可以将这些数据进行整合,形成一个统一的知识库。知识库可以支持智能搜索、自动推荐等应用,提高信息利用效率。

[0005] 由此,需要提供一种可以实现图谱自动化构建的技术方案,提升图谱构建的效率和准确性,以提高知识库管理和应用系统图谱的性能。

发明内容

[0006] 本申请旨在至少在一定程度上解决相关技术中的技术问题之一。

[0007] 为此,本申请的第一个目的在于提出一种用于知识库系统的图谱实体自动关联构建的方法,以实现图谱自动化构建,解决相关技术中图谱的构建效率低的问题。

[0008] 本申请的第二个目的在于提出一种用于知识库系统的图谱实体自动关联构建的装置。

[0009] 本申请的第三个目的在于提出一种电子设备。

[0010] 本申请的第四个目的在于提出一种计算机可读存储介质。

[0011] 本申请的第五个目的在于提出一种计算机程序产品。

[0012] 为达上述目的,本申请第一方面实施例提出了一种用于知识库系统的图谱实体自动关联构建的方法,包括:

[0013] 获取来自数据中台或者知识库库采集系统的原始业务数据;并对所述原始业务数据进行标准化处理,得到结构化业务数据;其中,所述知识库系统包括用于存储业务知识的知识图谱,所述知识图谱包括多个相互关联的知识节点,每个所述知识节点对应一个实体,每个实体包括至少一个属性,实体之间通过关系关联;

[0014] 采用基于NLP的实体和属性抽取技术,从所述结构化业务数据中抽取新实体以及所述新实体的新属性;

[0015] 基于第一信息表,获取与所述新属性匹配的待选属性;其中,所述第一信息表用于存储所述知识图谱当前的现有属性以及所述现有属性的属性列信息,所述属性列信息用于评估属性的匹配命中价值;

[0016] 查询第二信息表,在所述第二信息表中存在所述待选属性时,获取与所述待选属性关联的目标对象信息;其中,所述第二信息表用于存储所述知识图谱的现有属性的关联对象信息;

[0017] 基于所述目标对象信息、所述新实体和所述新属性,向所述知识图谱中添加新的知识节点。

[0018] 为达上述目的,本申请第二方面实施例提出了一种用于知识库系统的图谱实体自动关联构建的装置,包括:

[0019] 数据获取模块,用于获取来自数据中台或者知识库库采集系统的原始业务数据;并对所述原始业务数据进行标准化处理,得到结构化业务数据;其中,所述知识库系统包括用于存储业务知识的知识图谱,所述知识图谱包括多个相互关联的知识节点,每个所述知识节点对应一个实体,每个实体包括至少一个属性,实体之间通过关系关联;

[0020] 所述数据获取模块,还用于采用基于NLP的实体和属性抽取技术,从所述结构化业务数据中抽取新实体以及所述新实体的新属性;

[0021] 属性匹配模块,用于基于第一信息表,获取与所述新属性匹配的待选属性;其中,所述第一信息表用于存储所述知识图谱当前的现有属性以及所述现有属性的属性列信息,所述属性列信息用于评估属性的匹配命中价值;

[0022] 对象获取模块,用于基于第二信息表,在所述第二信息表中存在所述待选属性时,获取与所述待选属性关联的目标对象信息;其中,所述第二信息表用于存储所述知识图谱的现有属性的关联对象信息;

[0023] 图谱构建模块,用于基于所述目标对象信息、所述新实体和所述新属性,向所述知识图谱中添加新的知识节点。

[0024] 为达上述目的,本申请第三方面实施例提出了一种电子设备,包括:处理器,以及与所述处理器通信连接的存储器;所述存储器存储计算机执行指令;所述处理器执行所述存储器存储的计算机执行指令,以实现第一方面所述的方法。

[0025] 为达上述目的,本申请第四方面实施例提出了一种计算机可读存储介质,所述计算机可读存储介质中存储有计算机执行指令,所述计算机执行指令被处理器执行时用于实现第一方面所述的方法。

[0026] 为达上述目的,本申请第五方面实施例提出了一种计算机程序产品,包括计算机程序,该计算机程序被处理器执行时实现第一方面所述的方法。

[0027] 本申请提供的用于知识库系统的图谱实体自动关联构建的方法、装置、电子设备及存储介质,知识库系统收集数据并进行预处理,再通过实体属性抽取技术得到新实体和新属性,从知识图谱的现有属性中获取与新属性匹配的待选属性;再从获取与所述待选属性匹配的对象信息,添加新的图谱关系;从而实现图谱实体的自动关联构建,提升了图谱的构建效率。

[0028] 本申请附加的方面和优点将在下面的描述中部分给出,部分将从下面的描述中变得明显,或通过本申请的实践了解到。

附图说明

[0029] 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:

[0030] 图1为本申请实施例所提供的一种用于知识库系统的图谱实体自动关联构建的方法的流程示意图;

[0031] 图2为本申请实施例所提供的另一种用于知识库系统的图谱实体自动关联构建的方法的流程示意图;

[0032] 图3为本申请实施例所提供的一种用于知识库系统的图谱实体自动关联构建的装置的框图;

[0033] 图4为本申请实施例所提供的一种电子设备的框图。

具体实施方式

[0034] 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,旨在用于解释本申请,而不能理解为对本申请的限制。

[0035] 术语解释:

[0036] 图谱中的实体:指的是现实世界中的事物,比如:人、地名、公司、机构、概念、物品等。

[0037] 图谱中的关系:指的是实体之间的某种联系,比如:人-“出生于”-上海、张三和李四是“配偶”等等。

[0038] 图谱中的属性:实体和关系会拥有各自的属性,比如:人可以有“民族”和“年龄”,配偶关系可以有“结婚日期”等等。

[0039] 下面参考附图描述本申请实施例的用于知识库系统的图谱实体自动关联构建的方法、装置及设备。

[0040] 图1为本申请实施例所提供的一种用于知识库系统的图谱实体自动关联构建的方法的流程示意图。

[0041] 需要说明的是,本申请实施例的用于知识库系统的图谱实体自动关联构建的方法的执行主体为本申请实施例的用于知识库系统的图谱实体自动关联构建的装置,该用于知识库系统的图谱实体自动关联构建的装置可被配置于电子设备中,以使该电子设备可以执行用于知识库系统的图谱实体自动关联构建的功能。

[0042] 如图1所示,该用于知识库系统的图谱实体自动关联构建的方法包括以下步骤:

[0043] 步骤101,获取来自数据中台或者知识库库采集系统的原始业务数据;并对原始业务数据进行标准化处理,得到结构化业务数据;其中,知识库系统包括用于存储业务知识的知识图谱,知识图谱包括多个相互关联的知识节点,每个知识节点对应一个实体,每个实体包括至少一个属性,实体之间通过关系关联。

[0044] 在一些实施例中,知识库系统从数据中台的数据湖以及知识库库采集系统收集原始非结构化数据,处理该数据中的噪音和错误,如去除重复项、处理缺失值等,将不同格式的数据按照一定的数据标准转换为统一的格式的结构化数据。

[0045] 这里需要说明的是,知识库系统包括用于存储业务知识的知识图谱,即知识库系

统通过知识图谱存储收集的知识节点。由此,在知识库系统收集数据之后,需要更新知识图谱,也就是随着数据的增加,不断构建知识图谱的结构。在获取新的数据之后,需要抽取能够构建知识图谱的实体和属性。

[0046] 步骤102,采用基于NLP的实体和属性抽取技术,从结构化业务数据中抽取新实体以及新实体的新属性。

[0047] 作为一种实现方式,知识库系统在完成数据收集之后,进行实体识别和抽取、实体消歧与融合等步骤得到新实体和对应的新属性。

[0048] 在一些实施例中,实体识别与抽取的实现方式,包括:利用自然语言处理(NLP,Natuarl Language Processing)技术从文本中识别出实体(如人名、地名、机构名等),从文本中提取相关的属性和关系。例如从一句话中提取出人名和他们所属的公司,此步骤采用NLP实体、属性识别及抽取算法,属于现在技术,不在详述。

[0049] 在一些实施例中,实体消歧与融合的实现方式,包括:处理实体的同名异义现象,确保不同上下文中的同名实体被正确区分,将相同的实体进行合并。例如,将不同文档中的同一个人物合并为一个统一的实体。

[0050] 示例1,知识库系统经过数据收集、实体识别与抽取、实体消歧与融合形成的新实体为E1,其中某个新属性为A1。

[0051] 步骤103,基于第一信息表,获取与新属性匹配的待选属性;其中,第一信息表用于存储知识图谱当前的现有属性以及现有属性的属性列信息,属性列信息用于评估属性的匹配命中价值。

[0052] 可以理解为,根据知识图谱当前的现有属性以及所述现有属性的属性列信息,从现有属性中确定与新属性匹配度最高的待选属性。

[0053] 可以基于所有现有属性的属性列信息判断各现有属性与新属性的匹配值,从而根据匹配值确定待选属性。为了提高节约计算资源、提高计算效率,可以先通过距离相似度算法从现有属性中初步筛选出多个相似属性,跟根据多个相似属性的属性列信息判断各相似属性与新属性的匹配值,从而根据匹配值确定待选属性。

[0054] 步骤104,查询第二信息表,在第二信息表中存在待选属性时,获取与待选属性关联的目标对象信息;其中,第二信息表用于存储知识图谱的现有属性的关联对象信息。

[0055] 作为一种实现方式,将所述待选属性的值作为所述第一属性的值查询所述第二信息表,确定所述第二信息表中是否存在所述待选属性;

[0056] 在确定所述第二信息表中存在所述待选属性时,获取目标对象实体、目标对象属性和目标关系。

[0057] 可以理解为,第一信息表中已经存在的属性不一定已经建立图谱并已经存储于第二信息表第二信息表。在确定待选属性之后,需要查询第二信息表第二信息表,判断第二信息表第二信息表中是否存在该待选属性,如果第二信息表第二信息表中存在该待选属性,则从第二信息表第二信息表中获取与待选属性匹配的对象信息。

[0058] 在一些实施例中,第二信息表包括:

[0059] 第一实体(entity1),用于记录当前实体的值;

[0060] 第一属性(attribute1),用于记录该实体中当前属性的值;

[0061] 创建时间(time),用于记录关系的创建时间;

[0062] 关系(relationship),用于记录第一实体与对象实体之间的关系类型,该关系的类型,例如包含、属于等;

[0063] 对象实体(entity2),;用于记录与entity1建立关系的实体的值;

[0064] 对象属性(attribute2),用于记录与attribute1建立关系的属性的值。

[0065] 从第二信息表中获取与待选属性匹配的对象信息;包括:将待选属性的值作为第一属性的值查询第二信息表,获取目标对象实体、目标对象属性和目标关系。

[0066] 通过待选属性值查询第二信息表,如命中该待选属性,查询对应的对象实体entity2、对象属性attribute2,以关系relationship。

[0067] 步骤105,基于目标对象信息、新实体和新属性,向知识图谱中添加新的知识节点。

[0068] 作为一种实现方式,将新实体及新属性通过关系建立与对象实体及对象属性相关联的图谱关系,即将新实体、关系和对象实体组成的图谱三元组;并更新第二信息表。

[0069] 示例性的,查询得到待选属性的对象实体entity2、对象属性attribute2,以关系relationship后,便可以添加新的图谱关系;并将新实体E1,新属性A1的相应信息存入第一信息表和第二信息表。

[0070] 需要说明的是,当通过人工检查到新建立的图谱关系错误时,修正图谱关系,并修改第一信息表和第二信息表的相应值,尤其是第一信息表中的人工纠错量和人工纠错时间两个字段信息。

[0071] 本申请实施例的用于知识库系统的图谱实体自动关联构建的方法,通过知识库系统收集数据并进行预处理,再通过实体属性抽取技术得到新实体和新属性,从知识图谱的现有属性中获取与新属性匹配的待选属性;再从获取与所述待选属性匹配的对象信息,添加新的图谱关系;从而实现图谱实体的自动关联构建,提升了图谱的构建效率。

[0072] 在上述实施例的基础上,下面对步骤103的基于第一信息表获取与所述新属性匹配的待选属性的具体实现方式进行详细描述。

[0073] 图2为本申请实施例所提供的另一种用于知识库系统的图谱实体自动关联构建的方法的流程示意图。如图2所示,该用于知识库系统的图谱实体自动关联构建的方法包括以下步骤:

[0074] 步骤201,从第一信息表中的现有属性中,获取与新属性相似度满足相似度阈值的多个相似属性。

[0075] 作为一种实现方式,通过距离相似度算法计算新属性与第一信息表中的现有属性中各属性的相似度,获取与新属性相似度满足相似度阈值的多个相似属性;获取多个相似属性中各相似属性与新属性的距离值L。

[0076] 可以理解为,设定相似度阈值threshold,如果距离小于或等于该相似度阈值,则认为当前属性与新属性相似,并记录距离值L。

[0077] 继续以示例1为例,利用Levenshtein距离相似度算法,计算新属性A1与知识库系统的现有属性中各属性的相似度。

[0078] 步骤202,对于多个相似属性中的各相似属性,基于该相似属性的属性列信息,计算多个构建参数值。

[0079] 可以理解为,基于相似属性的属性列信息,得到各相似属性的多个构建参数值,从而根据多个构建参数值和各构建参数值对应的预设阈值,得到评价该相似属性与新属性的

匹配度的匹配值。

[0080] 在一些实施例中,第一信息表用于存储所述知识图谱当前的现有属性以及所述现有属性的属性列信息,该属性列信息可以包括但不限于:实体值、属性值、抽取时间、构建量、构建时间、人工纠错量、人工纠错时间和是否高价值段,属性列信息中的每一项对应第一信息表中的一列,当任一项数据存在更新是,第一信息表对应更新。

[0081] 其中,实体值(entity),用于记录图谱中实体的值;

[0082] 属性值(attribute),用于记录该实体中某个属性的值;

[0083] 抽取时间(time),用于记录该属性抽取的时间,

[0084] 构建量(build_number),用于记录该属性构建图谱的次数,默认值为0,每构建一次,构建量+1,如其中某个构建关系消失构建量-1,用于得到多个构建参数值中的构建量参数y1。

[0085] 构建时间(build_time),用于记录该属性最近一次构建时的时间;用于结合抽取时间计算多个构建参数值中的构建时间参数t1。

[0086] 人工纠错量(correct_number),用于记录该属性构建图谱错误时的字段进行人工干预的次数;默认值为0,每修改一次,该值+1,该数值只增不减;用于计算多个构建参数值中的人工纠错量参数y2。

[0087] 人工纠错时间(correct_time),用于记录最近一次人工纠错量变更时的指标修改时间;用于结合抽取时间计算多个构建参数值中的人工纠错时间参数t2。

[0088] 是否高价值段(is_high),用于识别该属性是否为高价值数据,示例性的,0表示低价值数据,1表示高价值数据,默认为0;用于获取多个构建参数值中的高价值参数m。

[0089] 需要说明的是,在获取各相似属性的属性列信息之后,可以计算各相似属性的多个构建参数值,以便通过多个构建参数值评价各相似属性与新属性的匹配度。

[0090] 在一些实施例中,多个构建参数值包括构建量参数y1、人工纠错量参数y2、构建时间参数t1、人工纠错时间参数t2、高价值参数m和距离参数s,各构建参数值的计算公式如下:

[0091] $y1=2\mathrm{arccot}(build\_number(x))/\pi$;

[0092] $y2=2\arctan(correct\_number(x))/\pi$;

[0093] $t1=2\mathrm{arccot}(build\_time(x)-time(x))/\pi$;

[0094] $t2=2\arctan(correct\_time(x)-time(x))/\pi$;

[0095] $m=is\_high$;

[0096] $s=(threshold-L)/threshold$;

[0097] 其中,x表示相似属性。

[0098] 通过上述公式,可以得到各相似属性的多个构建参数值。

[0099] 得到多个构建参数值之后,以便根据多个构建参数值后续步骤得到相似属性的匹配值。

[0100] 步骤203,基于多个构建参数值与各构建参数值对应的预设阈值,得到该相似属性的匹配值。

[0101] 作为一种实现方式:获取多个构建参数值与对应各构建参数值的预设阈值的加权平均值;将加权平均值作为该相似属性的匹配值。

[0102] 示例性的,y1、y2、t1、t2、m和s的预设阈值分别为p1、p2、p3、p4、p5和p6,则相似属性的匹配值V的计算公式如下:

[0103] V＝(p1*y1+p2*y2+p3*t1+p4*t2+p5*m+p6*s)/(p1+p2+p3+p4+p5+p6)

[0104] 需要说明的是,V为取值在(0,1]的值,几个相似属性的V值越大,匹配命中价值越高,默认取V值最大的相似属性,作为后续步骤的待选属性值。

[0105] 步骤204,基于多个相似属性各自的匹配值,从多个相似属性中确定匹配值满足预设条件的待选属性。

[0106] 在一些实施例中,获取待选属性的方法,包括:将多个相似属性中匹配值最大的相似属性确定为待选属性。即预设条件为匹配值最大。

[0107] 本申请实施例的用于知识库系统的图谱实体自动关联构建的方法,从现有属性中获取与新属性相似的多个相似属性;再通过各相似属性的属性列信息计算多个构建参数值;从而根据多个构建参数值得到各相似属性与新属性的匹配值;基于匹配值确定待选属性,从而提高属性的匹配精度和匹配效率,提高图谱构建的成功率,以提升图谱构建效率。

[0108] 为了实现上述实施例,本申请还提出一种用于知识库系统的图谱实体自动关联构建的装置。图3为本申请实施例提供的一种用于知识库系统的图谱实体自动关联构建的装置的框图。如图3所示,该用于知识库系统的图谱实体自动关联构建的装置可以包括:数据获取模块301、属性匹配模块302、对象获取模块303和图谱构建模块304。

[0109] 其中,数据获取模块301,用于获取来自数据中台或者知识库库采集系统的原始业务数据;并对所述原始业务数据进行标准化处理,得到结构化业务数据;其中,所述知识库系统包括用于存储业务知识的知识图谱,所述知识图谱包括多个相互关联的知识节点,每个所述知识节点对应一个实体,每个实体包括至少一个属性,实体之间通过关系关联;

[0110] 数据获取模块301,还用于采用基于NLP的实体和属性抽取技术,从所述结构化业务数据中抽取新实体以及所述新实体的新属性;

[0111] 属性匹配模块302,用于基于第一信息表,获取与所述新属性匹配的待选属性;其中,所述第一信息表用于存储所述知识图谱当前的现有属性以及所述现有属性的属性列信息,所述属性列信息用于评估属性的匹配命中价值;

[0112] 对象获取模块303,用于基于第二信息表,在所述第二信息表中存在所述待选属性时,获取与所述待选属性关联的目标对象信息;其中,所述第二信息表用于存储所述知识图谱的现有属性的关联对象信息;

[0113] 图谱构建模块304,用于基于所述目标对象信息、所述新实体和所述新属性,向所述知识图谱中添加新的知识节点。

[0114] 进一步地,在本申请实施例的一种可能的实现方式中,属性匹配模块302,具体用于:

[0115] 从所述第一信息表中的现有属性中,获取与所述新属性相似度满足相似度阈值的多个相似属性;

[0116] 对于所述多个相似属性中的各相似属性,基于该相似属性的属性列信息,计算多个构建参数值;

[0117] 基于所述多个构建参数值与各构建参数值对应的预设阈值,得到该相似属性的匹配值;

[0118] 基于所述多个相似属性各自的匹配值,从所述多个相似属性中确定匹配值满足预设条件的待选属性。

[0119] 进一步地,在本申请实施例的一种可能的实现方式中,属性匹配模块302在从所述第一信息表中的现有属性中,获取与所述新属性相似度满足相似度阈值的多个相似属性时,具体用于:

[0120] 通过距离相似度算法计算所述新属性与所述第一信息表中现有属性中各属性的相似度,获取与所述新属性相似度满足相似度阈值的多个相似属性;

[0121] 获取所述多个相似属性中各相似属性与所述新属性的距离值L。

[0122] 进一步地,在本申请实施例的一种可能的实现方式中,属性列信息包括实体值、属性值、抽取时间、构建量、构建时间、人工纠错量、人工纠错时间和是否高价值段,其中,实体值用于记录图谱中实体的值,属性值用于记录该实体中某个属性的值;抽取时间用于记录该属性抽取的时间,构建量用于记录该属性构建图谱的次数,构建时间用于记录该属性最近一次构建时的时间;人工纠错量用于记录该属性构建图谱错误时的字段进行人工干预的次数;人工纠错时间用于记录最近一次人工纠错量变更时的指标修改时间;是否高价值段用于识别该属性是否为高价值数据。

[0123] 进一步地,在本申请实施例的一种可能的实现方式中,属性匹配模块302在基于该相似属性的属性列信息,计算多个构建参数值时;具体用于:

[0124] 基于构建时间,通过第一公式得到构建量参数y1;第一公式表示如下:

[0125] $y1=2\operatorname{arccot}(build\_number(x))/\pi$

[0126] 其中,build_number表示构建时间,x表示相似属性;

[0127] 基于人工纠错量,通过第二公式得到人工纠错量参数y2;第二公式表示如下:

[0128] $y2=2\arctan(correct\_number(x))/\pi$

[0129] 其中,correct_number表示人工纠错量,x表示相似属性;

[0130] 基于构建时间和抽取时间,通过第三公式得到构建时间参数t1;第三公式表示如下:

[0131] $t1=2\operatorname{arccot}(build\_time(x)-time(x))/\pi$

[0132] 其中,build_time表示构建时间,time表示抽取时间,x表示相似属性;

[0133] 基于人工纠错时间和抽取时间,通过第四公式得到人工纠错时间参数t2;第四公式表示如下:

[0134] $t2=2\arctan(correct\_time(x)-time(x))/\pi$

[0135] 其中,correct_time表示人工纠错时间,time表示抽取时间,x表示相似属性;

[0136] 基于是否高价值段,通过第五公式得到高价值参数m;第五公式表示如下:

[0137] $m=is\_high$

[0138] 其中,is_high表示是否高价值段;

[0139] 基于距离值L,通过第六公式得到距离参数s;第六公式表示如下:

[0140] $s=(threshold-L)/threshold$

[0141] 其中,threshold表示相似度阈值。

[0142] 进一步地,在本申请实施例的一种可能的实现方式中,第二信息表包括第一实体、第一属性、创建时间、关系、对象实体和对象属性,创建时间用于记录关系的创建时间,关系

用于记录第一实体与对象实体之间的关系类型;对象获取模块303,具体用于:

[0143] 将所述待选属性的值作为所述第一属性的值查询所述第二信息表,确定所述第二信息表中是否存在所述待选属性;

[0144] 在确定所述第二信息表中存在所述待选属性时,获取目标对象实体、目标对象属性和目标关系。

[0145] 进一步地,在本申请实施例的一种可能的实现方式中,图谱构建模块304,具体用于:

[0146] 将所述新实体及所述新属性通过所述关系建立与所述对象实体及所述对象属性相关联的图谱关系。

[0147] 需要说明的是,前述对用于知识库系统的图谱实体自动关联构建的方法实施例的解释说明也适用于该实施例的用于知识库系统的图谱实体自动关联构建的装置,此处不再赘述。

[0148] 为了实现上述实施例,本申请还提出一种电子设备。请参见图4,图4是本申请实施例提供的电子设备的框图。如图4所示,电子设备400包括:处理器401,以及与所述处理器401通信连接的存储器402;所述存储器402存储计算机执行指令;所述处理器401执行所述存储器存储的计算机执行指令,以实现执行前述实施例所提供的方法。

[0149] 为了实现上述实施例,本申请还提出一种计算机可读存储介质,计算机可读存储介质中存储有计算机执行指令,所述计算机执行指令被处理器执行时用于实现前述实施例所提供的方法。

[0150] 为了实现上述实施例,本申请还提出一种计算机程序产品,包括计算机程序,该计算机程序被处理器执行时实现前述实施例所提供的方法。

[0151] 本申请中所涉及的用户个人信息的收集、存储、使用、加工、传输、提供和公开等处理,均符合相关法律法规的规定,且不违背公序良俗。

[0152] 需要说明的是,来自用户的个人信息应当被收集用于合法且合理的用途,并且不在这些合法使用之外共享或出售。此外,应在收到用户知情同意后进行此类采集/共享,包括但不限于在用户使用该功能前,通知用户阅读用户协议/用户通知,并签署包括授权相关用户信息的协议/授权。此外,还需采取任何必要步骤,保卫和保障对此类个人信息数据的访问,并确保有权访问个人信息数据的其他人遵守其隐私政策和流程。

[0153] 本申请预期可提供用户选择性阻止使用或访问个人信息数据的实施方案。即本公开预期可提供硬件和/或软件,以防止或阻止对此类个人信息数据的访问。一旦不再需要个人信息数据,通过限制数据收集和删除数据可最小化风险。此外,在适用时,对此类个人信息去除个人标识,以保护用户的隐私。

[0154] 在前述各实施例描述中,参考术语“一个实施例”、“一些实施例”、“示例”、“具体示例”、或“一些示例”等的描述意指结合该实施例或示例描述的具体特征、结构、材料或者特点包含于本申请的至少一个实施例或示例中。在本说明书中,对上述术语的示意性表述不必须针对的是相同的实施例或示例。而且,描述的具体特征、结构、材料或者特点可以在任一个或多个实施例或示例中以合适的方式结合。此外,在不相互矛盾的情况下,本领域的技术人员可以将本说明书中描述的不同实施例或示例以及不同实施例或示例的特征进行结合和组合。

[0155] 此外,术语“第一”、“第二”仅用于描述目的,而不能理解为指示或暗示相对重要性或者隐含指明所指示的技术特征的数量。由此,限定有“第一”、“第二”的特征可以明示或者隐含地包括至少一个该特征。在本申请的描述中,“多个”的含义是至少两个,例如两个,三个等,除非另有明确具体的限定。

[0156] 流程图中或在此以其他方式描述的任何过程或方法描述可以被理解为,表示包括一个或更多个用于实现定制逻辑功能或过程的步骤的可执行指令的代码的模块、片段或部分,并且本申请的优选实施方式的范围包括另外的实现,其中可以不按所示出或讨论的顺序,包括根据所涉及的功能按基本同时的方式或按相反的顺序,来执行功能,这应被本申请的实施例所属技术领域的技术人员所理解。

[0157] 在流程图中表示或在此以其他方式描述的逻辑和/或步骤,例如,可以被认为是用于实现逻辑功能的可执行指令的定序列表,可以具体实现在任何计算机可读介质中,以供指令执行系统、装置或设备(如基于计算机的系统、包括处理器的系统或其他可以从指令执行系统、装置或设备取指令并执行指令的系统)使用,或结合这些指令执行系统、装置或设备而使用。就本说明书而言,"计算机可读介质"可以是任何可以包含、存储、通信、传播或传输程序以供指令执行系统、装置或设备或结合这些指令执行系统、装置或设备而使用的装置。计算机可读介质的更具体的示例(非穷尽性列表)包括以下:具有一个或多个布线的电连接部(电子装置),便携式计算机盘盒(磁装置),随机存取存储器(RAM),只读存储器(ROM),可擦除可编辑只读存储器(EPROM或闪速存储器),光纤装置,以及便携式光盘只读存储器(CDROM)。另外,计算机可读介质甚至可以是可在其上打印所述程序的纸或其他合适的介质,因为可以例如通过对纸或其他介质进行光学扫描,接着进行编辑、解译或必要时以其他合适方式进行处理来以电子方式获得所述程序,然后将其存储在计算机存储器中。

[0158] 应当理解,本申请的各部分可以用硬件、软件、固件或它们的组合来实现。在上述实施方式中,多个步骤或方法可以用存储在存储器中且由合适的指令执行系统执行的软件或固件来实现。如,如果用硬件来实现和在另一实施方式中一样,可用本领域公知的下列技术中的任一项或他们的组合来实现:具有用于对数据信号实现逻辑功能的逻辑门电路的离散逻辑电路,具有合适的组合逻辑门电路的专用集成电路,可编程门阵列(PGA),现场可编程门阵列(FPGA)等。

[0159] 本技术领域的普通技术人员可以理解实现上述实施例方法携带的全部或部分步骤是可以通过程序来指令相关的硬件完成,所述的程序可以存储于一种计算机可读存储介质中,该程序在执行时,包括方法实施例的步骤之一或其组合。

[0160] 此外,在本申请各个实施例中的各功能单元可以集成在一个处理模块中,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模块中。上述集成的模块既可以采用硬件的形式实现,也可以采用软件功能模块的形式实现。所述集成的模块如果以软件功能模块的形式实现并作为独立的产品销售或使用时,也可以存储在一个计算机可读取存储介质中。

[0161] 上述提到的存储介质可以是只读存储器,磁盘或光盘等。尽管上面已经示出和描述了本申请的实施例,可以理解的是,上述实施例是示例性的,不能理解为对本申请的限制,本领域的普通技术人员在本申请的范围内可以对上述实施例进行变化、修改、替换和变型。

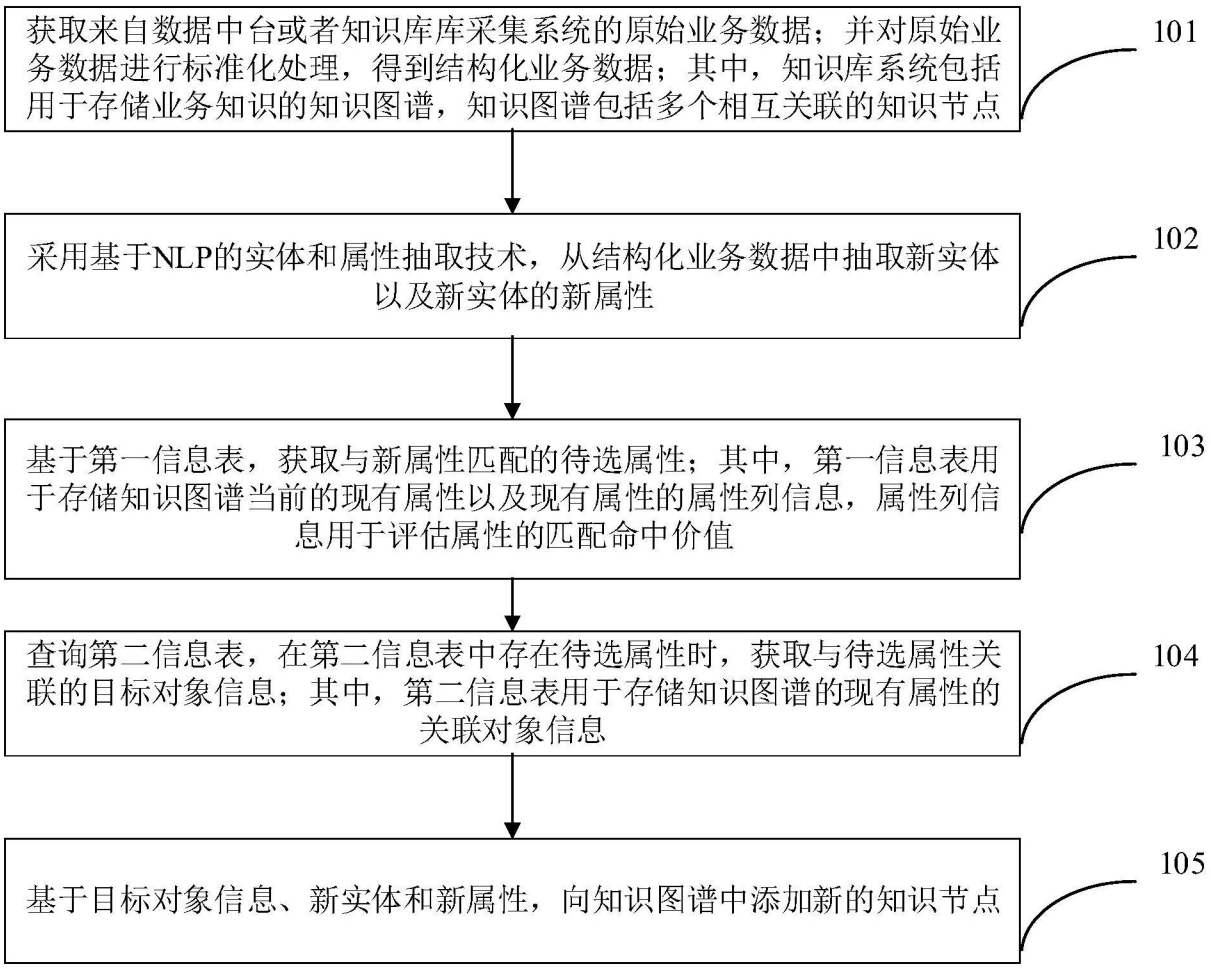

获取来自数据中台或者知识库库采集系统的原始业务数据；并对原始业务数据进行标准化处理，得到结构化业务数据；其中，知识库系统包括用于存储业务知识的知识图谱，知识图谱包括多个相互关联的知识节点
101
采用基于NLP的实体和属性抽取技术，从结构化业务数据中抽取新实体以及新实体的新属性
102
基于第一信息表，获取与新属性匹配的待选属性；其中，第一信息表用于存储知识图谱当前的现有属性以及现有属性的属性列信息，属性列信息用于评估属性的匹配命中价值
103
查询第二信息表，在第二信息表中存在待选属性时，获取与待选属性关联的目标对象信息；其中，第二信息表用于存储知识图谱的现有属性的关联对象信息
104
基于目标对象信息、新实体和新属性，向知识图谱中添加新的知识节点
105

图1

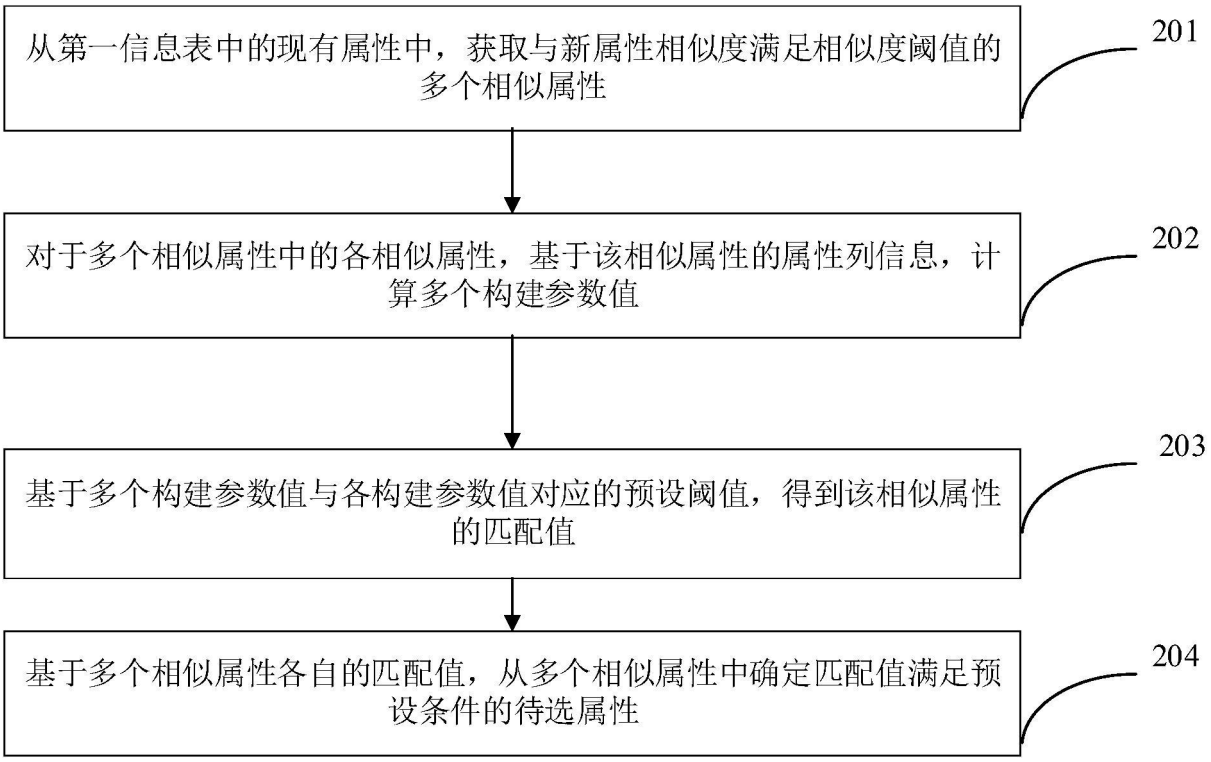

从第一信息表中的现有属性中，获取与新属性相似度满足相似度阈值的多个相似属性
201
对于多个相似属性中的各相似属性，基于该相似属性的属性列信息，计算多个构建参数值
202
基于多个构建参数值与各构建参数值对应的预设阈值，得到该相似属性的匹配值
203
基于多个相似属性各自的匹配值，从多个相似属性中确定匹配值满足预设条件的待选属性
204

图2

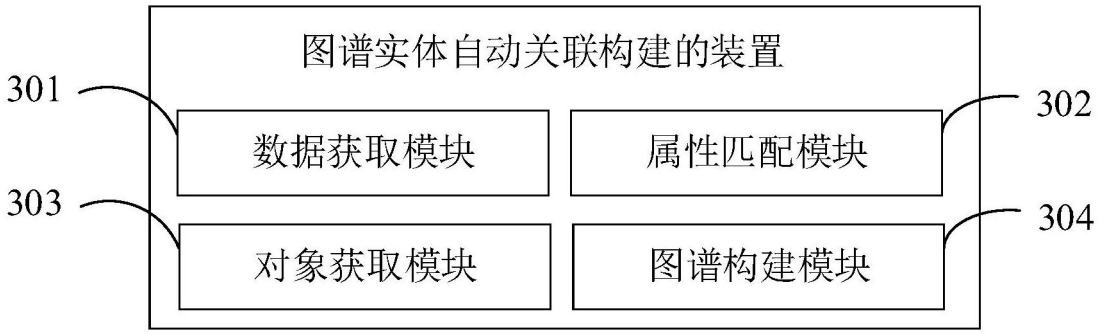

图谱实体自动关联构建的装置
301
数据获取模块
302
属性匹配模块
303
对象获取模块
304
图谱构建模块

图3

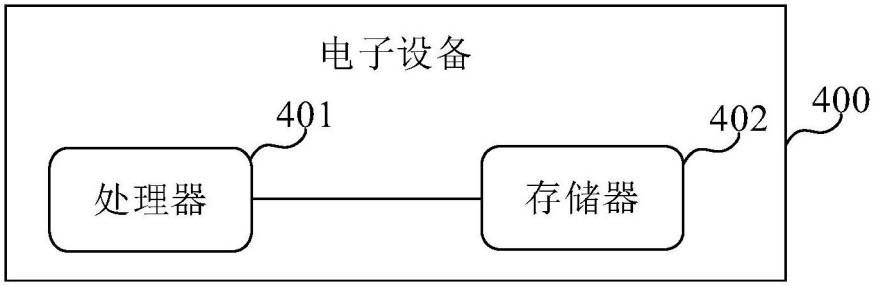

电子设备
401
处理器
402
存储器
400

图4