(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 113505889 B
(45)授权公告日 2024.08.02

(21)申请号 202110838039.0

(22)申请日 2021.07.23

(65)同一申请的已公布的文献号
申请公布号 CN 113505889 A

(43)申请公布日 2021.10.15

(73)专利权人 中国平安人寿保险股份有限公司
地址 518000 广东省深圳市福田区益田路5033号平安金融中心14、15、16、37、41、44、45、46层

(72)发明人 周元笙 蒋佳惟 马龙

(74)专利代理机构 北京辰权知识产权代理有限公司 11619
专利代理师 付婧

(51)Int.Cl.
G06N 5/025(2023.01)
G06N 5/022(2023.01)
G06N 5/02(2023.01)

(56)对比文件
CN 111291135 A,2020.06.16

审查员 王澜

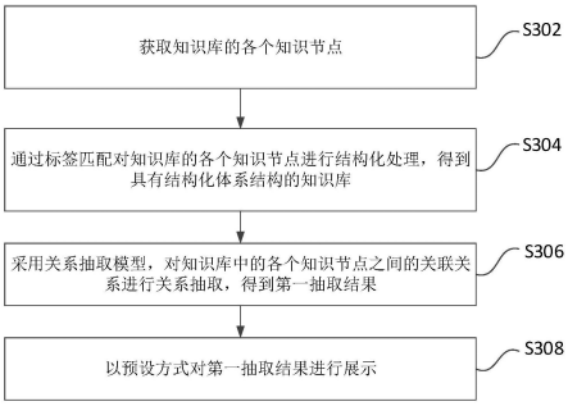
权利要求书2页 说明书10页 附图2页

(54)发明名称

图谱化知识库的处理方法、装置、计算机设备和存储介质

(57)摘要

本发明公开了一种图谱化知识库的处理方法、装置、计算机设备和存储介质。所述方法包括:获取知识库的各个知识节点;通过标签匹配对知识库的各个知识节点进行结构化处理,得到具有结构化体系结构的知识库;采用关系抽取模型,对知识库中的各个知识节点之间的关联关系进行关系抽取,得到第一抽取结果;以及以预设方式对第一抽取结果进行展示。由于引入了关系抽取模型,这样,能够对知识库中的各个知识节点之间的关联关系进行关系抽取,得到第一抽取结果,并对该第一抽取结果进行展示,这样,展示出来的各个知识节点之间是有一定关联度的,且以用户可以直观地看到各个知识节点之间的关联关系的预设方式展示,从而大大地提高了用户的体验度。



CN 113505889 B

1.一种图谱化知识库的处理方法,其特征在于,所述方法包括:

获取知识库的各个知识节点;

通过标签匹配对所述知识库的各个知识节点进行结构化处理,得到具有结构化体系结构的知识库;

所述通过标签匹配对所述知识库的各个知识节点进行结构化处理包括:

对所述知识库的各个知识节点进行抽取,得到第二抽取结果,所述第二抽取结果用于标识关键实体列表;

基于所述第二抽取结果中的各个数据,构建具有分类标签的字典;

基于具有所述分类标签的所述字典进行结构化处理,得到符合预设条件的知识集合;

所述对所述知识库的各个知识节点进行抽取包括:

通过预设数量的人工标注对序列模型进行训练,得到训练后的序列模型;

基于所述训练后的序列模型,对所述知识库的各个知识节点的关键内容进行识别,得到识别结果,所述识别结果至少包括用于识别所述知识库的各个知识节点的标签;

基于预设标签分类规则和所述知识库的各个知识节点的标签,判断出所述知识库的各个知识节点的标签所属的标签类别;

基于所述知识库的各个知识节点的标签所属的所述标签类别,对所述知识库的各个知识节点进行标签归类;

在所述对所述知识库的各个知识节点进行抽取之前,所述方法还包括:

读取所述识别结果,

所述识别结果还包括以下至少一项:

所述知识库的各个知识节点的关键内容,所述知识库的各个知识节点与对应的标签、对应的关键内容之间的映射关系;

采用关系抽取模型,对所述知识库中的各个知识节点之间的关联关系进行关系抽取,得到第一抽取结果;

以预设方式对所述第一抽取结果进行展示。

2.根据权利要求1所述的方法,其特征在于,所述基于所述第二抽取结果中的各个数据,构建具有分类标签的字典包括:

配置进行筛选的筛选条件,所述筛选条件至少包括预设高频条件;

根据所述筛选条件,对所述第二抽取结果中的各个数据进行比对和数据清洗,得到清洗后的数据;

获取与所述知识库的各个知识节点关联的各种关联数据;

对各种关联数据进行数据融合,得到数据融合结果;

基于所述数据融合结果,构建具有分类标签的字典。

3.根据权利要求1所述的方法,其特征在于,所述基于具有所述分类标签的所述字典进行结构化处理,得到符合预设条件的知识集合包括:

选取待检索的目标知识;

基于具有所述分类标签的所述字典,对所述待检索的目标知识进行结构化处理,得到结构化抽取结果;

获取符合预设条件的标签组合;

基于所述标签组合,对所述结构化抽取结果进行筛选,得到符合所述预设条件的知识集合。

4.一种图谱化知识库的处理装置,其特征在于,所述装置包括:

获取模块,用于获取知识库的各个知识节点;

处理模块,用于通过标签匹配对所述获取模块获取的所述知识库的各个知识节点进行结构化处理,得到具有结构化体系结构的知识库;

所述通过标签匹配对所述知识库的各个知识节点进行结构化处理包括:

对所述知识库的各个知识节点进行抽取,得到第二抽取结果,所述第二抽取结果用于标识关键实体列表;

基于所述第二抽取结果中的各个数据,构建具有分类标签的字典;

基于具有所述分类标签的所述字典进行结构化处理,得到符合预设条件的知识集合;

所述对所述知识库的各个知识节点进行抽取包括:

通过预设数量的人工标注对序列模型进行训练,得到训练后的序列模型;

基于所述训练后的序列模型,对所述知识库的各个知识节点的关键内容进行识别,得到识别结果,所述识别结果至少包括用于识别所述知识库的各个知识节点的标签;

基于预设标签分类规则和所述知识库的各个知识节点的标签,判断出所述知识库的各个知识节点的标签所属的标签类别;

基于所述知识库的各个知识节点的标签所属的所述标签类别,对所述知识库的各个知识节点进行标签归类;

在所述对所述知识库的各个知识节点进行抽取之前,还包括:

读取所述识别结果,

所述识别结果还包括以下至少一项:

所述知识库的各个知识节点的关键内容,所述知识库的各个知识节点与对应的标签、对应的关键内容之间的映射关系;

抽取模块,用于采用关系抽取模型,对所述处理模块得到的所述知识库中的各个知识节点之间的关联关系进行关系抽取,得到第一抽取结果;

展示模块,用于以预设方式对所述抽取模块抽取的所述第一抽取结果进行展示。

5.一种计算机设备,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3中任一项权利要求所述处理方法的步骤。

6.一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存储有计算机程序,所述计算机程序被一个或多个处理器执行时,实现如权利要求1至3中任一项权利要求所述处理方法的步骤。

图谱化知识库的处理方法、装置、计算机设备和存储介质

技术领域

[0001] 本发明涉及人工智能技术领域,特别涉及图谱化知识库的处理方法、装置、计算机设备和存储介质。

背景技术

[0002] 知识图谱是一种用图模型来描述知识和建模世界万物之间的关联关系的方法。知识图谱由节点和边组成。节点可以是实体,如一个人、一本书等,或是抽象的概念。边可以是实体的属性,如姓名、书名,或是实体之间的关系,如朋友。

[0003] 常用的知识图谱由知识图谱Schema,该知识图谱Schema定义了知识图谱的基本类、术语、属性和关系等本体层概念。cnSchema.ORG是OpenKG发起和完成的开放的知识图谱Schema标准。cnSchema的词汇集包括了上千种概念分类、数据类型、属性和关系等常用概念定义,以支持知识图谱数据的通用性、复用性和流动性。结合中文的特点,复用、连接并扩展了Schema.org、Wikidata、Wikipedia等已有的知识图谱Schema标准,为中文领域的开放知识图谱、聊天机器人、搜索引擎优化提供可供参考的数据描述和接口定义标准。通过cnSchema,开发者可以快速对接大量基于Schema.org定义的网站,以及Bot的知识图谱数据API。cnSchema主要解决如下三个问题:第一,Bots是新兴的人机接口,对话中的信息粒度缩小到短文本、实体和关系,不仅需要文本与结构化数据的结合,还需要更丰富的上下文处理机制;第二,知识图谱Schema.缺乏对中文的支持;第三,知识图谱的构建成本高,需要分摊成本。

[0004] 现有的知识图谱方法涉及知识表示、知识获取、知识处理和知识利用多个方面。一般处理流程为:首先,确定知识表示模型,然后根据数据来源选择不同的知识获取手段导入知识,接着综合利用知识推理、知识融合、知识挖掘等技术对构建的知识图谱进行不断优化,最后根据不同应用场景需求设计不同的知识访问与呈现方法,例如,语义搜索、问答交互、图谱可视化分析等。

[0005] 上述现有的基于知识图谱的方法构建的知识库,每条知识以行的形式进行存储和管理,以至于每条知识之间缺乏关联性,难以根据语义进行相关问题的关联和访问,从而导致造成基于构建的知识库进行索引的索引效率低,而且索引得到的索引结果也不精准,索引结果往往也不是用户想要的,用户体验度低。

发明内容

[0006] 基于此,有必要针对现有构建的知识库中的各个知识之间低关联度的问题,提供一种图谱化知识库的处理方法、装置、计算机设备和存储介质。

[0007] 第一方面,本申请实施例提供了一种图谱化知识库的处理方法,所述方法包括:

[0008] 获取知识库的各个知识节点;

[0009] 通过标签匹配对所述知识库的各个知识节点进行结构化处理,得到具有结构化体系结构的知识库;

[0010] 采用关系抽取模型,对所述知识库中的各个知识节点之间的关联关系进行关系抽取,得到第一抽取结果;

[0011] 以预设方式对所述第一抽取结果进行展示。

[0012] 在一种实施方式中,所述对知识库的各个知识节点通过标签匹配进行结构化处理包括:

[0013] 对所述知识库的各个知识节点进行抽取,得到第二抽取结果,所述第二抽取结果用于标识关键实体列表;

[0014] 基于所述第二抽取结果中的各个数据,构建具有分类标签的字典;

[0015] 基于具有所述分类标签的所述字典进行结构化处理,得到符合预设条件的知识集合。

[0016] 在一种实施方式中,所述对所述知识库的各个知识节点进行抽取包括:

[0017] 通过预设数量的人工标注对序列模型进行训练,得到训练后的序列模型;

[0018] 基于所述训练后的序列模型,对所述知识库的各个知识节点的关键内容进行识别,得到识别结果,所述识别结果至少包括用于识别所述知识库的各个知识节点的标签;

[0019] 基于预设标签分类规则和所述知识库的各个知识节点的标签,判断出所述知识库的各个知识节点的标签所属的标签类别;

[0020] 基于所述知识库的各个知识节点的标签所属的所述标签类别,对所述知识库的各个知识节点进行标签归类。

[0021] 在一种实施方式中,在所述对所述知识库的各个知识节点进行抽取之前,所述方法还包括:

[0022] 读取所述识别结果,

[0023] 所述识别结果还包括以下至少一项:

[0024] 所述知识库的各个知识节点的关键内容,所述知识库的各个知识节点与对应的标签、对应的关键内容之间的映射关系。

[0025] 在一种实施方式中,所述基于所述第二抽取结果中的各个数据,构建具有分类标签的字典包括:

[0026] 配置进行筛选的筛选条件,所述筛选条件至少包括预设高频条件;

[0027] 根据所述筛选条件,对所述第二抽取结果中的各个数据进行比对和数据清洗,得到清洗后的数据;

[0028] 获取与所述知识库的各个知识节点关联的各种关联数据;

[0029] 对各种关联数据进行数据融合,得到数据融合结果;

[0030] 基于所述数据融合结果,构建具有分类标签的字典。

[0031] 在一种实施方式中,所述基于具有所述分类标签的所述字典进行结构化处理,得到符合预设条件的知识集合包括:

[0032] 选取待检索的目标知识;

[0033] 基于具有所述分类标签的所述字典,对所述待检索的目标知识进行结构化处理,得到结构化抽取结果;

[0034] 获取符合预设条件的标签组合;

[0035] 基于所述标签组合,对所述结构化抽取结果进行筛选,得到符合所述预设条件的

知识集合。

[0036] 第二方面,本申请实施例提供了一种图谱化知识库的处理装置,所述装置包括:

[0037] 获取模块,用于获取知识库的各个知识节点;

[0038] 处理模块,用于通过标签匹配对所述获取模块获取的所述知识库的各个知识节点进行结构化处理,得到具有结构化体系结构的知识库;

[0039] 抽取模块,用于采用关系抽取模型,对所述处理模块得到的所述知识库中的各个知识节点之间的关联关系进行关系抽取,得到第一抽取结果;

[0040] 展示模块,用于以预设方式对所述抽取模块抽取的所述第一抽取结果进行展示。

[0041] 在一种实施方式中,所述处理模块用于:

[0042] 对所述知识库的各个知识节点进行抽取,得到第二抽取结果,所述第二抽取结果用于标识关键实体列表;

[0043] 基于所述第二抽取结果中的各个数据,构建具有分类标签的字典;

[0044] 基于具有所述分类标签的所述字典进行结构化处理,得到符合预设条件的知识集合。

[0045] 第三方面,本申请实施例提供一种计算机设备,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的方法步骤。

[0046] 第四方面,本申请实施例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的方法步骤。

[0047] 本申请实施例提供的技术方案可以包括以下有益效果:

[0048] 在本申请实施例中,获取知识库的各个知识节点;通过标签匹配对知识库的各个知识节点进行结构化处理,得到具有结构化体系结构的知识库;采用关系抽取模型,对知识库中的各个知识节点之间的关联关系进行关系抽取,得到第一抽取结果;以及以预设方式对第一抽取结果进行展示。因此,采用本申请实施例,由于引入了关系抽取模型,这样,能够对知识库中的各个知识节点之间的关联关系进行关系抽取,得到第一抽取结果,并对该第一抽取结果进行展示,这样,展示出来的各个知识节点之间是有一定关联度的,且以用户可以直观地看到各个知识节点之间的关联关系的预设方式展示,从而大大地提高了用户的体验度。应当理解的是,以上的一般描述和后文的细节描述仅是示例性和解释性的,并不能限制本发明。

附图说明

[0049] 此处的附图被并入说明书中并构成本说明书的一部分,示出了符合本发明的实施例,并与说明书一起用于解释本发明的原理。

[0050] 图1为一个实施例中提供的一种图谱化知识库的处理方法的实施环境图;

[0051] 图2为一个实施例中计算机设备的内部结构框图;

[0052] 图3是本公开实施例提供的一种图谱化知识库的处理方法的流程示意图;

[0053] 图4是本公开实施例提供的一种图谱化知识库的处理装置的结构示意图。

具体实施方式

[0054] 以下描述和附图充分地示出本发明的具体实施方案,以使本领域的技术人员能够实践它们。

[0055] 应当明确,所描述的实施例仅仅是本发明一部分实施例,而不是全部的实施例。基于本发明中的实施例,本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其它实施例,都属于本发明保护的范围。

[0056] 下面结合附图详细说明本公开的可选实施例。

[0057] 图1为一个实施例中提供的一种图谱化知识库的处理方法的实施环境图,如图1所示,在该实施环境中,包括计算机设备110以及终端120。

[0058] 需要说明的是,终端120以及计算机设备110可为智能手机、平板电脑、笔记本电脑、台式计算机等,但并不局限于此。计算机设备110以及终端110可以通过蓝牙、USB(Universal Serial Bus,通用串行总线)或者其他通讯连接方式进行连接,本发明在此不做限制。

[0059] 图2为一个实施例中计算机设备的内部结构示意图。如图2所示,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图谱化知识库的处理方法。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图谱化知识库的处理方法。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2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。

[0060] 如图3所示,本公开实施例提供一种图谱化知识库的处理方法,该图谱化知识库的处理方法具体包括以下方法步骤:

[0061] S302:获取知识库的各个知识节点。

[0062] S304:通过标签匹配对知识库的各个知识节点进行结构化处理,得到具有结构化体系结构的知识库。

[0063] 在一种可能的实现方式中,对知识库的各个知识节点通过标签匹配进行结构化处理包括以下步骤:

[0064] 对知识库的各个知识节点进行抽取,得到第二抽取结果,第二抽取结果用于标识关键实体列表;

[0065] 基于第二抽取结果中的各个数据,构建具有分类标签的字典;

[0066] 基于具有分类标签的字典进行结构化处理,得到符合预设条件的知识集合。

[0067] 在一种可能的实现方式中,对知识库的各个知识节点进行抽取包括以下步骤:

[0068] 通过预设数量的人工标注对序列模型进行训练,得到训练后的序列模型;

[0069] 在本申请实施例中,通过少量的人工标注对序列模型进行训练,得到训练后的序列模型,例如,人工标注的数量为500个,在此对人工标注的数量并不做限制,可以根据对训练模型精度的要求,增加人工标注的数量,在此不再赘述。

[0070] 基于训练后的序列模型,对知识库的各个知识节点的关键内容进行识别,得到识别结果,识别结果至少包括用于识别知识库的各个知识节点的标签;

[0071] 基于预设标签分类规则和知识库的各个知识节点的标签,判断出知识库的各个知识节点的标签所属的标签类别;

[0072] 基于知识库的各个知识节点的标签所属的标签类别,对知识库的各个知识节点进行标签归类。

[0073] 通过上述抽取过程,能够做到:以较少的人工标注,尽可能多的覆盖知识库中的各类知识;此外,抽取的结果用来标识关键实体列表。

[0074] 在一种可能的实现方式中,在对知识库的各个知识节点进行抽取之前,所述方法还包括以下步骤:

[0075] 读取识别结果,

[0076] 识别结果还包括以下至少一项:

[0077] 知识库的各个知识节点的关键内容,知识库的各个知识节点与对应的标签、对应的关键内容之间的映射关系。

[0078] 在一种可能的实现方式中,基于第二抽取结果中的各个数据,构建具有分类标签的字典包括步骤:

[0079] 配置进行筛选的筛选条件,筛选条件至少包括预设高频条件;,预设高频条件包括特定词的出现次数,例如,在某一具体应用场景下,可以将预设高频条件配置为:包括某一特定词的出现次数大于k次。上述仅仅是示例,可以根据不同应用场景的需求,对预设高频条件进行调整,在此不再赘述。

[0080] 根据筛选条件,对第二抽取结果中的各个数据进行比对和数据清洗,得到清洗后的数据;

[0081] 例如,在某一具体应用场景下,若配置的筛选条件为预设高频条件,且配置的预设高频条件包括特定词的出现次数为至少30次,则基于该筛选条件,对上述得到的抽取结果进行比对和数据清洗,得到清洗后的数据。

[0082] 获取与知识库的各个知识节点关联的各种关联数据;

[0083] 在本申请实施例中,与各个知识节点关联的关联数据包括:用于标识各个知识节点的关键属性的标签数据、与各个知识节点对应的抽取结果中的关键实体列表数据、以及与各个知识节点对应的外部抓取的关键词列表数据。

[0084] 在本申请实施例中,标签数据包括用于标识各个知识节点所属产品类别的产品标签数据,用于标识各个知识节点的关联疾病的疾病标签数据,用于标识各个知识节点的关联职业的职业标签数据,以及用于标识各个知识节点的关联城市名称的城市名称标签数据。

[0085] 对各种关联数据进行数据融合,得到数据融合结果;

[0086] 在本申请实施例中,对上述获取的各种与各个知识节点关联的关联数据进行数据融合的融合方法为常规方法,在此不再赘述。

[0087] 基于数据融合结果,构建具有分类标签的字典。

[0088] 其中,该字典具有与各个知识节点对应的分类标签;这样,便于根据字典中各个分类标签,快速且精准地对各个知识节点进行索引。

[0089] 在一种可能的实现方式中,基于具有分类标签的字典进行结构化处理,得到符合预设条件的知识集合包括以下步骤:

[0090] 选取待检索的目标知识;

[0091] 基于具有分类标签的字典,对待检索的目标知识进行结构化处理,得到结构化抽取结果;

[0092] 获取符合预设条件的标签组合;例如,在某一具体应用场景下,符合条件的标签组合为:“产品”+“属性”。

[0093] 基于标签组合,对结构化抽取结果进行筛选,得到符合预设条件的知识集合。

[0094] 在某一具体应用场景中,对上述得到的检索结果的正确性进行验证,得到其在通用库中的覆盖率达到90%。

[0095] S306:采用关系抽取模型,对知识库中的各个知识节点之间的关联关系进行关系抽取,得到第一抽取结果。

[0096] 在本申请实施例所采用的关系抽取模型为开放域实体关系抽取模型,该关系抽取模型为改进的TextRunner开放域实体关系抽取模型。

[0097] 如下简单介绍一下改进的TextRunner开放域实体关系抽取模型所采用的TextRunner系统的工作原理如下所述:

[0098] TextRunner能够直接从网页纯文本中抽取实体关系。TextRunner通过一些简单的启发式规则自动从宾州树库里面获取实体关系三元组的正负样本,根据它们的一些浅层句法特征训练一个分类器来判断两个实体间是否存在语义关系;然后将网络文本进行一定的处理后作为候选句子,提取其浅层句法特征,利用分类器判断所抽取的关系三元组是否可信,最后利用网络数据的冗余信息,对初步认定可信的关系进行评估。对于关系名称的抽取,TextRunner把动词作为关系名称,通过动词链接两个论元,从而挖掘论元之间的关系,其抽取过程类似于语义角色标注。

[0099] 本申请实施例中的关系抽取模型所采用的系统为改进的TextRunne,该系统使用启发式规则在宾州树库中自动标注语料,不需要人工预先定义关系类别体系。

[0100] 本申请实施例所采用的抽取步骤具体包括以下步骤:

[0101] 步骤1、语料的自动生成和分类器训练

[0102] 1.1语料的自动生成:主要是通过依存句法分析结合启发式规则自动生成语料。

[0103] 常用的启发式规则示例如下:

[0104] 两个实体的依存路径长度不能大于指定值。

[0105] 实体不能是代词。

[0106] 关系指示词是两个实体之间依存路径上的动词或动词短语。

[0107] 两个实体必须在同一个句子中。

[0108] 1.2分类器的训练:TextRunner利用朴素贝叶斯分类器进行训练,得到初始关系抽取模板,上述训练过程所使用的特征示例如下:

[0109] 关系指示词的词性关系指示词的长度;

[0110] 实体的类型;

[0111] 实体是否是专有名词;

[0112] 左实体左边词语的词性;

[0113] 右实体右边词语的词性。

[0114] 步骤2、对步骤1得到的初始关系抽取模板不断地进行迭代,得到最终关系抽取模型所采用的关系抽取器,以及所采用的最终抽取模板。具体迭代过程如下所述:

[0115] 获取语料库中的数据;

[0116] 统计数据中出现的多个高频词,并将多个高频词作为触发词;

[0117] 根据触发词匹配出候选语料;

[0118] 根据候选语料得出元模板,对元模板进行多次迭代后,得到最终关系抽取模型所采用的关系抽取器,以及最终抽取模板。

[0119] 步骤3、通过步骤2得到的关系抽取器,以及最终抽取模板,对语料库中的数据进行关系三元组的抽取,得到大量的三元组,并将得到的三元组进行存储。

[0120] 在本申请实施例中,为了进行大规模关系三元组的抽取,需要对语料库中的数据进行预处理,将语料库中的数据转换成能够进行批量处理的文本数据。

[0121] 具体转换的方法为常规方法,在此不再赘述。

[0122] 步骤4:对步骤3得到的关系三元组的可信度进行计算,得到对应的可信度数值。

[0123] 读取步骤3中存储的各个三元组,并对相似的三元组进行合并,得到合并后的关系三元组;

[0124] 根据预置的筛选条件,该筛选条件用于排除掉合并后的关系三元组中的重复且冗余的数据,得到精简的优化后的合并关系三元组;

[0125] 对优化后的合并关系三元组在文本中出现的频次,得到对应的关系三元组的可信度数值。

[0126] 步骤5:根据预置的可信度阈值和各个关系三元组的可信度数值,依次确定各个关系三元组是否可以作为抽取出的关系三元组。

[0127] 从各个关系三元组中随机选取任意一个关系三元组作为当前关系三元组;

[0128] 读取该关系三元组的可信度数值;

[0129] 将该关系三元组的可信度数值和预置的可信度阈值进行比较,若该关系三元组的可信度数值大于或者等于预置的可信度阈值,则确定该关系三元组可以作为抽取出的关系三元组。

[0130] 例如,在某一具体应用场景下,抽取出的关系三元组可以为:

[0131] (保险名称,例如,小福星20,相关问题A,与相关问题A对应的保费1);或者,

[0132] (保险名称,例如,小福星20,相关问题B,与相关问题B对应的保费2);或者,

[0133] (保险名称,例如,小福星20,相关问题C,与相关问题C对应的保费3)。

[0134] 通过上述抽取出的三元组关系,当接收到用户的携带有“保险名称,例如,小福星20”的检索指令时,会自动展示出上述三元组关系中有小福星20的检索结果,大大地提高了基于关键词的检索效率。

[0135] 此外,为了进一步地提高检索结果的精准度,还可以对上述检索结果进行进一步地提炼,例如,引入新的检索词,例如,保费的金额范围,形成新的检索指令:“保险名称,例如,小福星20”+“保费的金额范围”;这样,可以大大地提高检索结果的精准度。

[0136] 本申请实施例提供的处理方法,通过对初始关系抽取模板进行多次迭代,得到最终关系抽取模型所采用的关系抽取器,以及所采用的最终抽取模板;以及,基于关系抽取

器,以及最终抽取模板,对语料库中的数据进行关系三元组的抽取,得到大量的关系三元组,并计算各个关系三元组的可信度数值;根据预置的可信度阈值计算出的各个关系三元组的可信度数值,精准地判断哪一个关系三元组可以作为抽取出的关系三元组;这样,能够大大地提高以关系三元组中的任意一个元素作为主关键词进行检索的效率和精准度;此外,也为后续基于精准地检索结果进行推荐提供了可能性。

[0137] S308:以预设方式对第一抽取结果进行展示。

[0138] 在本申请实施例中,对上述抽取结果可以以主关键词“保险名称,例如,小福星20”的形式展示。

[0139] 例如,在某一具体应用场景下,抽取出的关系三元组为:

[0140] (保险名称,例如,小福星20,相关问题A,与相关问题A对应的保费1);或者,

[0141] (保险名称,例如,小福星20,相关问题B,与相关问题B对应的保费2);或者,

[0142] (保险名称,例如,小福星20,相关问题C,与相关问题C对应的保费3)时,则可以以以主关键词“保险名称,例如,小福星20”的形式展示。

[0143] 在此对展现形式不做具体限制,优先选择关系图的展现方式,在关系图中,每一个节点对应上述关系三元组中的一个元素,例如,保险名称“小福星20,相关问题A,与相关问题A对应的保费1,各个元素之间的有向边代表节点间具有一定的关系。

[0144] 在本公开实施例中,获取知识库的各个知识节点;通过标签匹配对知识库的各个知识节点进行结构化处理,得到具有结构化体系结构的知识库;采用关系抽取模型,对知识库中的各个知识节点之间的关联关系进行关系抽取,得到第一抽取结果;以及以预设方式对第一抽取结果进行展示。因此,采用本申请实施例,由于引入了关系抽取模型,这样,能够对知识库中的各个知识节点之间的关联关系进行关系抽取,得到第一抽取结果,并对该第一抽取结果进行展示,这样,展示出来的各个知识节点之间是有一定关联度的,且以用户可以直观地看到各个知识节点之间的关联关系的预设方式展示,从而大大地提高了用户的体验度。

[0145] 下述为本发明图谱化知识库的处理装置实施例,可以用于执行本发明图谱化知识库的处理方法实施例。对于本发明图谱化知识库的处理装置实施例中未披露的细节,请参照本发明图谱化知识库的处理方法实施例。

[0146] 请参见图4,其示出了本发明一个示例性实施例提供的图谱化知识库的处理装置的结构示意图。该图谱化知识库的处理装置可以通过软件、硬件或者两者的结合实现成为终端的全部或一部分。该图谱化知识库的处理装置包括获取模块401、处理模块402、抽取模块403和展示模块404。

[0147] 具体而言,获取模块401,用于获取知识库的各个知识节点;

[0148] 处理模块402,用于通过标签匹配对获取模块401获取的知识库的各个知识节点进行结构化处理,得到具有结构化体系结构的知识库;

[0149] 抽取模块403,用于采用关系抽取模型,对处理模块402得到的知识库中的各个知识节点之间的关联关系进行关系抽取,得到第一抽取结果;

[0150] 展示模块404,用于以预设方式对抽取模块403抽取的第一抽取结果进行展示。

[0151] 可选的,处理模块402用于:

[0152] 对知识库的各个知识节点进行抽取,得到第二抽取结果,第二抽取结果用于标识

关键实体列表；

[0153] 基于第二抽取结果中的各个数据，构建具有分类标签的字典；

[0154] 基于具有分类标签的字典进行结构化处理，得到符合预设条件的知识集合。

[0155] 可选的，处理模块402具体用于：

[0156] 通过预设数量的人工标注对序列模型进行训练，得到训练后的序列模型；

[0157] 基于训练后的序列模型，对知识库的各个知识节点的关键内容进行识别，得到识别结果，识别结果至少包括用于识别知识库的各个知识节点的标签；

[0158] 基于预设标签分类规则和知识库的各个知识节点的标签，判断出知识库的各个知识节点的标签所属的标签类别；

[0159] 基于知识库的各个知识节点的标签所属的标签类别，对知识库的各个知识节点进行标签归类。

[0160] 可选的，所述装置还包括：

[0161] 读取模块(在图4中未示出)，用于在抽取模块403对知识库的各个知识节点进行抽取之前，读取识别结果，读取模块读取的识别结果还包括以下至少一项：知识库的各个知识节点的关键内容，知识库的各个知识节点与对应的标签、对应的关键内容之间的映射关系。

[0162] 可选的，处理模块402具体用于：

[0163] 配置进行筛选的筛选条件，筛选条件至少包括预设高频条件；

[0164] 根据筛选条件，对第二抽取结果中的各个数据进行比对和数据清洗，得到清洗后的数据；

[0165] 获取与知识库的各个知识节点关联的各种关联数据；

[0166] 对各种关联数据进行数据融合，得到数据融合结果；

[0167] 基于数据融合结果，构建具有分类标签的字典。

[0168] 可选的，处理模块402具体用于：

[0169] 选取待检索的目标知识；

[0170] 基于具有分类标签的字典，对待检索的目标知识进行结构化处理，得到结构化抽取结果；

[0171] 获取符合预设条件的标签组合；

[0172] 基于标签组合，对结构化抽取结果进行筛选，得到符合预设条件的知识集合。

[0173] 需要说明的是，上述实施例提供的图谱化知识库的处理装置在执行图谱化知识库的处理方法时，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，即将设备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。另外，上述实施例提供的图谱化知识库的处理装置与图谱化知识库的处理方法实施例属于同一构思，其体现实现过程详见图谱化知识库的处理方法实施例，这里不再赘述。

[0174] 在本公开实施例中，获取模块用于获取知识库的各个知识节点；处理模块用于通过标签匹配对获取模块获取的知识库的各个知识节点进行结构化处理，得到具有结构化体系结构的知识库；抽取模块用于采用关系抽取模型，对处理模块得到的知识库中的各个知识节点之间的关联关系进行关系抽取，得到第一抽取结果；以及展示模块用于以预设方式对抽取模块抽取的第一抽取结果进行展示。因此，采用本申请实施例，由于引入了关系抽取

模型,这样,能够对知识库中的各个知识节点之间的关联关系进行关系抽取,得到第一抽取结果,并对该第一抽取结果进行展示,这样,展示出来的各个知识节点之间是有一定关联度的,且以用户可以直观地看到各个知识节点之间的关联关系的预设方式展示,从而大大地提高了用户的体验度。

[0175] 在一个实施例中,提出了一种计算机设备,计算机设备包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序,处理器执行计算机程序时实现以下步骤:获取知识库的各个知识节点;通过标签匹配对知识库的各个知识节点进行结构化处理,得到具有结构化体系结构的知识库;采用关系抽取模型,对知识库中的各个知识节点之间的关联关系进行关系抽取,得到第一抽取结果;以及以预设方式对第一抽取结果进行展示。

[0176] 在一个实施例中,提出了一种存储有计算机可读指令的存储介质,该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以下步骤:获取知识库的各个知识节点;通过标签匹配对知识库的各个知识节点进行结构化处理,得到具有结构化体系结构的知识库;采用关系抽取模型,对知识库中的各个知识节点之间的关联关系进行关系抽取,得到第一抽取结果;以及以预设方式对第一抽取结果进行展示。

[0177] 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。

[0178] 以上实施例的各技术特征可以进行任意的组合,为使描述简洁,未对上述实施例中的各个技术特征所有可能的组合都进行描述,然而,只要这些技术特征的组合不存在矛盾,都应当认为是本说明书记载的范围。

[0179] 以上实施例仅表达了本发明的几种实施方式,其描述较为具体和详细,但并不能因此而理解为对本发明专利范围的限制。应当指出的是,对于本领域的普通技术人员来说,在不脱离本发明构思的前提下,还可以做出若干变形和改进,这些都属于本发明的保护范围。因此,本发明专利的保护范围应以所附权利要求为准。

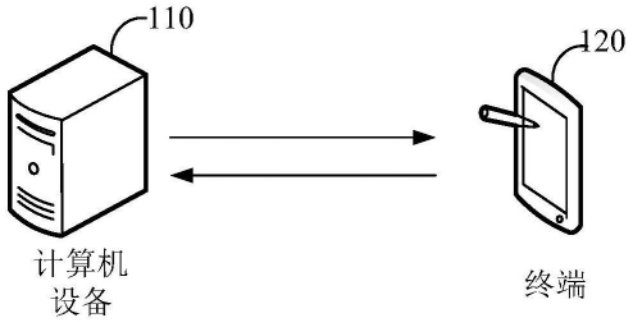


图1

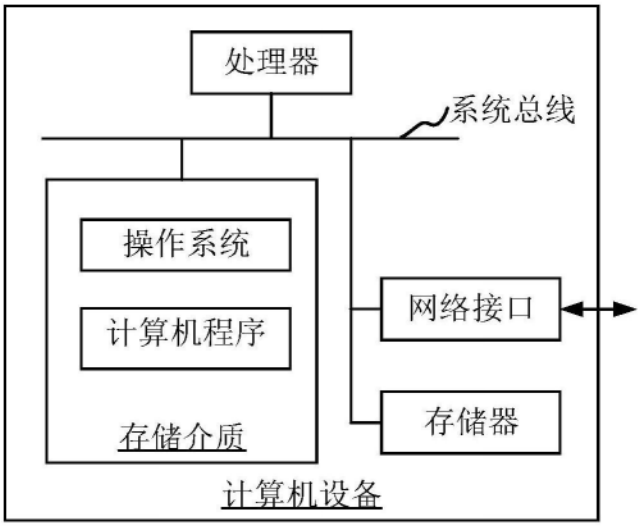


图2

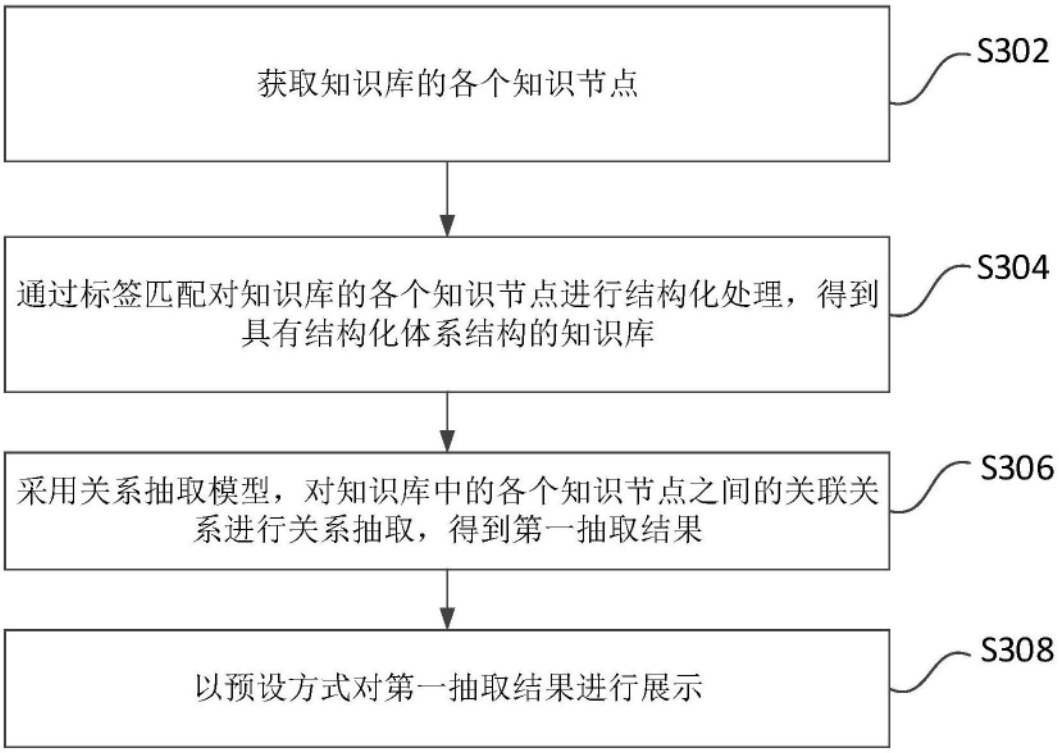

获取知识库的各个知识节点
S302
通过标签匹配对知识库的各个知识节点进行结构化处理，得到具有结构化体系结构的知识库
S304
采用关系抽取模型，对知识库中的各个知识节点之间的关联关系进行关系抽取，得到第一抽取结果
S306
以预设方式对第一抽取结果进行展示
S308

图3

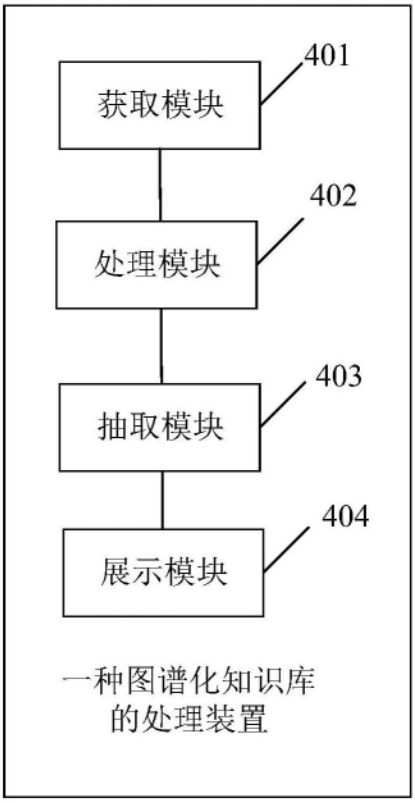

401
获取模块
402
处理模块
403
抽取模块
404
展示模块
一种图谱化知识库的处理装置

图4