(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 112307134 B
(45)授权公告日 2024.02.06

(21)申请号 202011196563.4

(22)申请日 2020.10.30

(65)同一申请的已公布的文献号
申请公布号 CN 112307134 A

(43)申请公布日 2021.02.02

(73)专利权人 北京百度网讯科技有限公司
地址 100085 北京市海淀区上地十街10号百度大厦2层

(72)发明人 骆金昌 万凡 王海威 王杰 陈坤斌 刘准 和为

(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 单冠飞

(51)Int.Cl.
G06F 16/28(2019.01)
G06F 16/36(2019.01)
G06F 16/35(2019.01)
G06F 40/30(2020.01)
G06F 40/295(2020.01)
G06N 3/0455(2023.01)

(56)对比文件
CN 110263318 A,2019.09.20
CN 111723575 A,2020.09.29
CN 106776711 A,2017.05.31
CN 110334211 A,2019.10.15
CN 107861939 A,2018.03.30
CN 106909655 A,2017.06.30
CN 110277149 A,2019.09.24
KR 100877477 B1,2009.01.07
US 2018137404 A1,2018.05.17
US 2013311467 A1,2013.11.21
US 9785696 B1,2017.10.10
熊玲;徐增壮;王潇斌;洪宇;朱巧明.基于共指消解的实体搜索模型研究.中文信息学报.2018,(05),94-101.
陆伟;武川.实体链接研究综述.情报学报.2015,(01),107-114.
杨一帆;陈文亮.旅游场景下的实体别名抽取联合模型.中文信息学报.2020,(06),59-67.

审查员 单娟

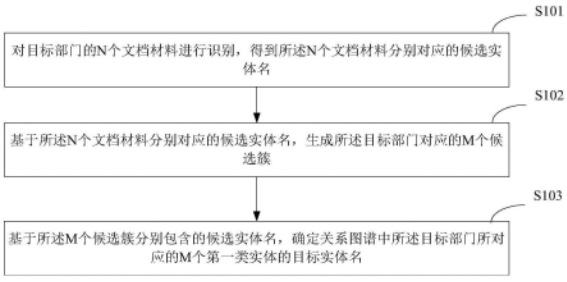
权利要求书3页 说明书10页 附图3页

(54)发明名称

实体信息处理方法、装置、电子设备及存储介质

(57)摘要

本公开提供了实体信息处理方法、装置、电子设备及存储介质,涉及深度学习等领域。具体实现方案为:对目标部门的N个文档材料进行识别,得到所述N个文档材料分别对应的候选实体名;N为大于等于1的整数;基于所述N个文档材料分别对应的候选实体名,生成所述目标部门对应的M个候选簇;基于所述M个候选簇分别包含的候选实体名,确定关系图谱中所述目标部门所对应的M个第一类实体的目标实体名。



CN 112307134 B

1.一种实体信息处理方法,包括:

对目标部门的N个文档材料进行识别,得到所述N个文档材料分别对应的候选实体名;N为大于等于1的整数;

基于所述N个文档材料分别对应的候选实体名,生成所述目标部门对应的M个候选簇;M为大于等于1的整数;

基于所述M个候选簇分别包含的候选实体名,确定关系图谱中所述目标部门所对应的M个第一类实体的目标实体名,所述第一类实体包括所述关系图谱中的事实体;

其中,所述基于所述M个候选簇分别包含的候选实体名,确定关系图谱中所述目标部门所对应的M个第一类实体的目标实体名,包括:

获取所述M个候选簇中的第i个候选簇中包含的候选实体名的频次信息;其中,i为大于等于1且小于等于M的整数;

将所述第i个候选簇中包含的候选实体名中频次信息最高的候选实体名作为所述第i个候选簇所对应的第i个第一类实体的目标实体标准名,将所述第i个候选簇中除所述目标实体标准名外的其他实体名作为所述第i个第一类实体的目标实体别名。

2.根据权利要求1所述的方法,其中,所述对目标部门的N个文档材料进行识别,得到所述N个文档材料分别对应的候选实体名,包括:

将所述目标部门的所述N个文档材料中的第j个文档材料及其对应的目标部门输入至预设模型,得到所述预设模型输出的所述第j个文档材料所对应的候选实体名;其中,j为大于等于1且小于等于N的整数。

3.根据权利要求2所述的方法,其中,所述方法还包括:

获取多个部门分别对应的历史候选实体名;

将所述多个部门中各个部门的历史文档材料与其对应的部门的历史候选实体名进行匹配,得到所述各个部门的历史文档材料所对应的历史实体名;

基于所述各个部门的历史文档材料及其对应的历史实体名生成训练集。

4.根据权利要求3所述的方法,其中,所述方法还包括:

基于所述训练集中包含的所述各个部门的历史文档材料及其对应的历史实体名对所述预设模型进行训练,得到训练后的所述预设模型。

5.根据权利要求1所述的方法,其中,所述基于所述N个文档材料分别对应的候选实体名,确定所述目标部门对应的M个候选簇,包括:

从所述N个文档材料分别对应的N个候选实体名中筛选得到L个候选实体名;L为大于等于1且小于等于N的整数;

对所述L个候选实体名进行聚类,得到所述目标部门对应的M个候选簇;其中,所述M个候选簇中不同的候选簇包含不同的候选实体名。

6.根据权利要求1-5任一项所述的方法,其中,所述方法还包括:

从所述M个第一类实体中的第k个第一类实体的目标实体名所分别对应的文档材料中,获取所述第k个第一类实体关联的第二类实体,所述第二类实体包括所述关系图谱中的人实体;基于所述第k个第一类实体所关联的所述第二类实体,在所述关系图谱中建立所述第k个第一类实体与所述第二类实体之间的关联关系;其中,k为大于等于1且小于等于M的整数。

7.一种实体信息处理装置,包括:

识别模块,用于对目标部门的N个文档材料进行识别,得到所述N个文档材料分别对应的候选实体名;N为大于等于1的整数;

聚簇模块,用于基于所述N个文档材料分别对应的候选实体名,生成所述目标部门对应的M个候选簇;M为大于等于1的整数;

实体名确定模块,用于基于所述M个候选簇分别包含的候选实体名,确定关系图谱中所述目标部门所对应的M个第一类实体的目标实体名,所述第一类实体包括所述关系图谱中的事实体;

其中,所述实体名确定模块,用于获取所述M个候选簇中的第i个候选簇中包含的候选实体名的频次信息;其中,i为大于等于1且小于等于M的整数;将所述第i个候选簇中包含的候选实体名中频次信息最高的候选实体名作为所述第i个候选簇所对应的第i个第一类实体的目标实体标准名,将所述第i个候选簇中除所述目标实体标准名外的其他实体名作为所述第i个第一类实体的目标实体别名。

8.根据权利要求7所述的装置,其中,所述识别模块,用于将所述目标部门的所述N个文档材料中的第j个文档材料及其对应的目标部门输入至预设模型,得到所述预设模型输出的所述第j个文档材料所对应的候选实体名;其中,j为大于等于1且小于等于N的整数。

9.根据权利要求8所述的装置,其中,所述装置还包括:

训练集构建模块,用于获取多个部门分别对应的历史候选实体名;将所述多个部门中各个部门的历史文档材料与其对应的部门的历史候选实体名进行匹配,得到所述各个部门的历史文档材料所对应的历史实体名;基于所述各个部门的历史文档材料及其对应的历史实体名生成训练集。

10.根据权利要求9所述的装置,其中,所述装置还包括:

模型训练模块,用于基于所述训练集中包含的所述各个部门的历史文档材料及其对应的历史实体名对所述预设模型进行训练,得到训练后的所述预设模型。

11.根据权利要求7所述的装置,其中,所述聚簇模块,用于从所述N个文档材料分别对应的N个候选实体名中筛选得到L个候选实体名;L为大于等于1且小于等于N的整数;对所述L个候选实体名进行聚类,得到所述目标部门对应的M个候选簇;其中,所述M个候选簇中不同的候选簇包含不同的候选实体名。

12.根据权利要求7-11任一项所述的装置,其中,所述装置还包括:

关系构建模块,用于从所述M个第一类实体中的第k个第一类实体的目标实体名所分别对应的文档材料中,获取所述第k个第一类实体关联的第二类实体,所述第二类实体包括所述关系图谱中的人实体;基于所述第k个第一类实体所关联的所述第二类实体,在所述关系图谱中建立所述第k个第一类实体与所述第二类实体之间的关联关系;其中,k为大于等于1且小于等于M的整数。

13. 一种电子设备,包括:

至少一个处理器;以及

与所述至少一个处理器通信连接的存储器;其中,

所述存储器存储有可被所述至少一个处理器执行的指令,所述指令被所述至少一个处理器执行,以使所述至少一个处理器能够执行权利要求1-6中任一项所述的方法。

14.一种存储有计算机指令的非瞬时计算机可读存储介质,所述计算机指令用于使所述计算机执行权利要求1-6中任一项所述的方法。

实体信息处理方法、装置、电子设备及存储介质

技术领域

[0001] 本公开涉及计算机技术领域。本公开尤其涉及深度学习领域。

背景技术

[0002] 关系图谱在企业中越来越广泛的被使用,在所述关系图谱中可以包含第一类实体(即“事”)以及第二类实体(即“人”),以及第一类实体以及第二类实体之间的关系等内容。所述关系图谱可以提供更多功能,比如可以搜索事的负责人,查看人的相关信息等等。然而,如何高效并准确的构建关系图谱中的第一类实体就成为需要解决的问题。

发明内容

[0003] 本公开提供了一种实体信息处理方法、装置、电子设备及存储介质。

[0004] 根据本公开的第一方面,提供了一种实体信息处理方法,包括:

[0005] 对目标部门的N个文档材料进行识别,得到所述N个文档材料分别对应的候选实体名;N为大于等于1的整数;

[0006] 基于所述N个文档材料分别对应的候选实体名,生成所述目标部门对应的M个候选簇;M为大于等于1的整数;

[0007] 基于所述M个候选簇分别包含的候选实体名,确定关系图谱中所述目标部门所对应的M个第一类实体的目标实体名。

[0008] 根据本公开的第二方面,提供了一种实体信息处理装置,包括:

[0009] 识别模块,用于对目标部门的N个文档材料进行识别,得到所述N个文档材料分别对应的候选实体名;N为大于等于1的整数;

[0010] 聚簇模块,用于基于所述N个文档材料分别对应的候选实体名,生成所述目标部门对应的M个候选簇;M为大于等于1的整数;

[0011] 实体名确定模块,用于基于所述M个候选簇分别包含的候选实体名,确定关系图谱中所述目标部门所对应的M个第一类实体的目标实体名。

[0012] 根据本公开的第三方面,提供了一种电子设备,包括:

[0013] 至少一个处理器;以及

[0014] 与所述至少一个处理器通信连接的存储器;其中,

[0015] 所述存储器存储有可被所述至少一个处理器执行的指令,所述指令被所述至少一个处理器执行,以使所述至少一个处理器能够执行前述方法。

[0016] 根据本公开的第四方面,提供了一种存储有计算机指令的非瞬时计算机可读存储介质,所述计算机指令用于使所述计算机执行前述方法。

[0017] 根据本公开的第五方面,提供了一种计算机程序产品,包括计算机程序,所述计算机程序在被处理器执行时实现如上所述的方法。

[0018] 采用本申请,可以基于目标部门的文档材料,确定文档材料对应的候选实体名,进而基于候选实体名确定在关系图谱中的目标部门对应的一个或多个第一类实体的目标实

体名,如此,可以避免人工分析实体名所带来的效率较低、时效性较差以及结果不准确等问题,保证了获取目标实体名的处理效率以及准确性,进而也保证了构建或更新关系图谱的效率以及准确性。

[0019] 应当理解,本部分所描述的内容并非旨在标识本公开的实施例的关键或重要特征,也不用于限制本公开的范围。本公开的其它特征将通过以下的说明书而变得容易理解。

附图说明

[0020] 附图用于更好地理解本方案,不构成对本公开的限定。其中:

[0021] 图1是根据本公开实施例的实体信息处理方法流程示意图;

[0022] 图2是根据本公开实施例的信息处理方法中构建候选簇的处理流程示意图;

[0023] 图3是根据本公开实施例的信息处理装置组成结构示意图一;

[0024] 图4是根据本公开实施例的信息处理装置组成结构示意图二;

[0025] 图5是用来实现本公开实施例的信息处理方法的电子设备的框图。

具体实施方式

[0026] 以下结合附图对本公开的示范性实施例做出说明,其中包括本公开实施例的各种细节以助于理解,应当将它们认为仅仅是示范性的。因此,本领域普通技术人员应当认识到,可以对这里描述的实施例做出各种改变和修改,而不会背离本公开的范围和精神。同样,为了清楚和简明,以下的描述中省略了对公知功能和结构的描述。

[0027] 本公开实施例提供了一种实体信息处理方法,如图1所示,包括:

[0028] S101:对目标部门的N个文档材料进行识别,得到所述N个文档材料分别对应的候选实体名;N为大于等于1的整数;

[0029] S102:基于所述N个文档材料分别对应的候选实体名,生成所述目标部门对应的M个候选簇;M为大于等于1的整数;

[0030] S103:基于所述M个候选簇分别包含的候选实体名,确定关系图谱中所述目标部门所对应的M个第一类实体的目标实体名。

[0031] 本发明实施例可以应用于电子设备,比如可以是服务器,或者可以是终端设备等。

[0032] 上述目标部门可以为一个单位或企业中多个部门中的任意一个,针对每一个部门都可以采用本实施例提供的方案进行处理,这里,将任意一个部门称作目标部门,其余部门的处理与所述目标部门相同,因此这里不做一一赘述。

[0033] 所述目标部门的N个文档材料,具体可以是,所述目标部门的周报、晋升材料等文档材料中至少之一。

[0034] 获取目标部门的N个文档材料的方法,可以是在部门内部进行材料收集,可以是将目标部门各个员工全部上传的文档材料收集起来作为所述目标部门的所述N个文档材料;又或者,从所述目标部门各个员工上传的文档材料中随机抽取得到所述目标部门的所述N个文档材料。

[0035] 对目标部门的N个文档材料进行识别,得到所述N个文档材料分别对应的第一类实体的候选实体名,可以包括:将所述目标部门的N个文档材料分别输入至预设模型,得到所述预设模型分别输出的候选实体名。

[0036] 基于所述N个文档材料分别对应的候选实体名,生成所述目标部门对应的M个候选簇,可以包括:将所述目标部门的所述N个文档材料分别对应的候选实体名进行聚类,得到所述目标部门对应的所述M个候选簇。

[0037] 进一步地,可以从每一个候选簇中包含的一个或多个候选实体名中选取一个候选实体名作为每一个候选簇所对应的目标实体名;将该目标实体名作为所述目标部门的一个第一类实体的目标实体名。

[0038] 需要理解的是,上述M个候选簇的具体数量可以根据实际情况不同而不同。假设目标部门最终可以得到一个第一类实体的目标实体名,则M等于1;假设目标部门对应的2个或更多第一类实体分别对应的目标实体名,则M为2个或更多;这里不对全部可能的情况进行穷举。

[0039] 上述第一类实体可以指的是关系图谱中的事实体,“事”实体可以包含多种内容,比如,可以包括:项目、平台、工具等等;需要理解的是,第一类实体可以包含一个或多个,也就是说在所述关系图谱中可以包括有一个或多个事实体。

[0040] 相应的第一类实体的目标实体名或候选实体名,可以指的是关系图谱中所要使用到的“事”一种属性或信息,比如,“事”的实体名可以为:项目的名称、平台的名称、工具的名称等等。

[0041] 通过上述方案,可以基于部门为单位输入部门的文档材料,确定文档材料对应的候选实体名,进而基于候选实体名确定在关系图谱中的每一个部门对应的一个或多个目标实体名,如此,仅通过收集部门的文档材料就可以最终确定关系图谱中包含的部门的目标实体名,可以避免人工分析所带来的效率较低、时效性较差以及结果不准确等问题,保证了获取目标实体名的处理效率以及准确性,进而也保证了构建或更新关系图谱的效率以及准确性。

[0042] 具体来说,上述S101中,所述对目标部门的N个文档材料进行识别,得到所述N个文档材料分别对应的候选实体名,包括:

[0043] 将所述目标部门的所述N个文档材料中的第j个文档材料及其对应的目标部门输入至预设模型,得到所述预设模型输出的所述第j个文档材料所对应的候选实体名;其中,j为大于等于1且小于等于N的整数。

[0044] 所述N个文档材料可以是从企业内部文档中抽取的,包含周报、晋升材料、述职报告、项目立案材料等等文档材料。因为这些材料在企业内部大量存在,可以很低的成本获取;而且这些材料往往时效性较新,例如周报需要每周写一次,因此通过收集这部分文档材料可满足时效性的要求。

[0045] 所述第j个文档材料为上述N个文档材料中任意之一。针对所述N个文档材料均采用相同的方式进行处理以得到对应的候选实体名,因此不对全部N个文档材料的处理一一赘述。

[0046] 需要理解的是,所述预设模型的输入信息具体来说可以为所述目标部门的名称以及所述第j个文档材料;再进一步地,可以预先对所述第j个文档材料进行切分,得到至少一个切分后的句子,将所述至少一个切分后的句子以及所述目标部门的名称作为所述预设模型的输入信息;相应的,所述预设模型的输出信息可以是候选实体名。

[0047] 可见,本实施例提供采用预设模型对文档材料进行分析得到文档材料所对应的候

选实体名,如此,可以避免人工分析或简单的文字匹配所带来的效率较低以及准确率较差的问题,提升了后续确定目标实体名的准确性以及提升了处理效率。

[0048] 进一步地,针对上述预设模型,可以为采用训练集包含的样本数据进行训练得到预设模型。关于构建训练集的方式,可以包括:

[0049] 获取多个部门分别对应的历史候选实体名;

[0050] 将所述多个部门中各个部门的历史文档材料与其对应的部门的历史候选实体名进行匹配,得到所述各个部门的历史文档材料所对应的历史实体名;

[0051] 基于所述各个部门的历史文档材料及其对应的历史实体名生成训练集。

[0052] 具体来说,历史文档材料的获取方式可以是从企业内部文档中抽取历史文档材料;比如,可以包含周报、晋升材料等等包含部门项目名的历史文档材料。因为这些历史文档材料在企业内部大量存在,可以很低的成本获取。

[0053] 基于所述各个部门的历史文档材料及其对应的历史实体名生成训练集,可以是将各个历史文档材料及其对应的历史实体名以及对应的部门作为各个样本数据,将所述各个样本数据添加至所述训练集。最终所述训练集中可以包括上述全部的样本数据。

[0054] 需要指出的是,在训练集的构建中,需要将同一个部门的历史实体名与同一个部门的历史文档材料进行匹配,以对历史文档材料进行标注,这样可以降低噪声,提高训练集的质量。上述确定每一个历史文档材料所对应的历史实体名可以为对所述历史文档材料进行标注,即将历史文档材料所匹配的历史实体名作为历史文档材料的标注。在相关技术中,训练集中的样本数据一般是人工标注的,这样成本较大;本实施例中,只需要将相同部门内的历史实体名以及历史文档材料进行匹配即可自动的完成对历史文档材料标注对应的历史实体名的处理,避免人工标注的成本过大的问题,并且相比于人工标注效率更高、准确率更高。

[0055] 可见,上述关于训练集的数据的标注工作是由设备自动完成的,并且由于样本数据的标注的时候采用同一个部门的历史实体名对同一个部门的历史文档材料进行的标注,这样,将部门作为信息的粒度或作为全局信息,可提升实体抽取的效果,如此可以提升训练集的样本数据的质量的同时能够降低噪声。

[0056] 然后,基于所述训练集中包含的所述各个部门的历史文档材料及其对应的历史实体名对所述预设模型进行训练,得到训练后的所述预设模型。

[0057] 也就是基于上述构建的包含有多个部门中各个部门的历史文档材料及其对应的部门内的历史实体名(比如项目名)的样本数据的训练集,对预设模型进行训练;在训练中,将所述训练集中的各个样本数据中包含的历史文档材料可以划分为一个或多个句子,将划分得到的一个或多个句子以及部门的名称作为预设模型的输入、将样本数据中历史文档材料所对应的历史实体名作为输出,以对预设模型进行训练。比如,所述预设模型训练的时候,其输入层和特征包含:历史文档材料的句子、部门,其表示方式可以是:句子+<SEP>+部门。

[0058] 预设模型的训练中的收敛条件可以是迭代次数达到预设门限值和/或损失函数小于预设门限值。具体的收敛条件可能还包括更多,本实施例中不做穷举。

[0059] 上述预设模型可以是采用BERT(Bidirectional Encoder Representation from Transformers,双向变形的编码器)以及条件随机场(Conditional Random Field,CRF)模

型来构建。其中,通过采用预设训练语言模型BERT来做语义向量抽取,可对句子实现精确的语义提取,且能提升语义迁移性,在训练集较小的情况下,也能取得较好的结果。

[0060] 可见,上述训练预设模型的处理中,由于训练集的数据的标注工作是由设备自动完成的,并且样本数据的标注的时候采用同一个部门的历史实体名对同一个部门的历史文档材料进行的标注,可以提升训练集的样本数据的质量的同时能够降低噪声,进而基于上述训练集进行预设模型的训练也可以保证最终得到的预设模型的识别准确度。

[0061] 采用上述处理,可以基于预设模型对当前输入的文档材料进行分析,得到当前输入的每一个文档材料所对应的实体名,将其作为每一个文档材料的候选实体名。然后执行前述S102的处理,基于所述N个文档材料分别对应的候选实体名,确定所述目标部门对应的M个候选簇,如图2所示,可以包括:

[0062] S201:从所述N个文档材料分别对应的N个候选实体名中筛选得到L个候选实体名;

[0063] S202:对所述L个候选实体名进行聚类,得到所述目标部门对应的M个候选簇;其中,所述M个候选簇中不同的候选簇包含不同的候选实体名。

[0064] 关于S201,具体可以包括以下几种处理方式:

[0065] 方式1、获取所述N个候选实体名的频次信息,从所述N个候选实体名中选取频次信息大于预设频次阈值的L个候选实体名;

[0066] 或者,

[0067] 方式2、基于预设规则对所述N个文档材料的N个候选实体名进行过滤,保留不满足所述预设规则的L个候选实体名;

[0068] 又或者,

[0069] 方式3、将上述方式1与方式2结合进行处理,可以为:

[0070] 基于预设规则对所述N个文档材料的N个候选实体名进行过滤,保留不满足所述预设规则的至少一个候选实体名;获取所述至少一个候选实体名的频次信息,从所述至少一个候选实体名中选取频次信息大于预设频次阈值的L个候选实体名。

[0071] 分别来说,方式1中,首先针对每个部门的候选实体名,进行频次统计得到每一个候选实体名对应的频次信息,然后结合频次信息将低频次的候选实体名过滤掉。如此可以提升后续聚类的准确率。

[0072] 关于预设频次阈值可以根据实际情况设置,比如,可以是出现3次作为所述预设频次阈值,又或者,可以是出现4次作为所述预设频次阈值。

[0073] 方式2中,所述预设规则可以包括:与预设关键词相同。其中,所述预设关键词可以根据实际情况进行设置,比如,可以将“商业化”作为一个预设关键词,相应的,在候选实体名中包含“商业化”这个预设关键词的候选实体名删除。

[0074] 在方式3中,可以将上述两种方式结合使用,首先删除满足预设规则的一部分候选实体名,然后再过滤掉一部分频次较低的候选实体名。当然,还可以先过滤掉频次低于预设频次门限值的一部分候选实体名之后,再将剩余的候选实体名中满足预设规则的候选实体名删除,最终得到所述目标部门对应的L个候选实体名。

[0075] S202中,对所述L个候选实体名进行聚类,得到所述目标部门对应的M个候选簇,具体可以包括:对所述L个候选实体名进行相似度计算,将相似度小于预设相似度阈值的候选实体名添加至同一个簇内,最终得到目标部门对应的M个候选簇。

[0076] 进一步地,上述相似度计算可以为:编辑距离相似度的计算,和/或语义相似度的计算。相应的,所述预设相似度阈值可以包括:预设编辑距离相似度阈值、预设语义相似度阈值中至少一种。

[0077] 比如,一种示例中,可以运用DBSCAN近邻聚类算法对候选实体名进行聚类,这一步是为了解决实体融合问题。候选实体名的相似度可以为字面的编辑距离,也就是当两个候选实体名之间的编辑距离小于预设编辑距离相似度阈值时,即聚到同一簇下。

[0078] 又一示例中,可以采用深度语义模型(Deep Structured Sematic models,DSSM)或其他模型进行语义相似度的计算,将语义距离小于预设语义相似度阈值的候选实体名作为同一类,聚到同一个簇下。

[0079] 再一示例中,可以是任意两个候选实体名的编辑距离小于预设编辑距离相似度阈值、并且语义距离小于预设语义相似度阈值的时候,将这两个候选实体名聚到同一个簇下。

[0080] 当然,还可以采用其他相似度计算来确定候选实体名之间的相似度,均可以在本实施例的保护范围内,这里不做穷举。

[0081] 通过预先对文档材料得到的候选实体名进行过滤,得到过滤后的候选实体名进一步进行聚簇,得到目标部门对应的M个候选簇,如此,可以减少部分候选实体名对最终识别目标实体名的影响,使得最终基于候选簇进行目标实体名的确定更加准确。

[0082] S103中,基于所述M个候选簇分别包含的候选实体名,确定关系图谱中所述目标部门所对应的M个第一类实体的目标实体名,包括:

[0083] 获取所述M个候选簇中的第i个候选簇中包含的候选实体名的频次信息;

[0084] 将所述第i个候选簇中包含的候选实体名中频次信息最高的候选实体名作为所述第i个候选簇所对应的第i个第一类实体的目标实体标准名,将所述第i个候选簇中除所述目标标准实体名外的其他实体名作为所述第i个第一类实体的目标实体别名。

[0085] 第i个候选簇可以是M个候选簇中任意之一,由于针对每一个候选簇都采用相同的方式确定对应的目标实体名,因此这里仅针对其中之一进行说明,其余的候选簇的处理方式相同,不做一一赘述。

[0086] 采用上述处理,可以基于第i个候选簇中的每一个候选实体名的频次信息,选取出现频次最高的一个候选实体名作为第i个候选簇所对应的第i个第一类实体的目标实体标准名,将第i个候选簇中的其余候选实体名均作为第i个第一类实体的实体别名。这样,每一个候选簇都可以得到针对相应的第一类实体的一个或多个实体别名,但是仅可以得到一个目标实体标准名。

[0087] 由于一个目标部门可以构建得到多个候选簇,可以认为每一个候选簇对应了一个第一类实体,基于一个候选簇可以确定一个第一类实体的目标实体标准名以及目标实体别名;最终可以得到目标部门的多个第一类实体分别对应的目标实体标准名以及目标实体别名。

[0088] 可见,通过上述方案,可以基于构建的候选簇最终确定针对一个事的目标实体标准名以及一个或多个目标实体别名,从而可以为构建关系图谱中的事的实体的时候提供更加准确的表述方式,并且由于增加了目标实体别名的信息,从而在后续泛化进行搜索的时候提供更多的参考信息进行搜索,从而使得关系图谱更加准确并且使用更加方便。

[0089] 基于以上处理可以得到关系图谱中的第一类实体的目标实体名,进一步地,还可

以获取事与相关第二类实体之间的关系,从而构建所述关系图谱中第一类实体的目标实体名与所述相关第二类实体之间的关系。具体的可以包括:

[0090] 从所述M个第一类实体中的第k个第一类实体的目标实体名所分别对应的文档材料中,获取所述第k个第一类实体关联的第二类实体,基于所述第k个第一类实体所关联的所述第二类实体,在所述关系图谱中建立所述第k个第一类实体与所述第二类实体之间的关联关系;其中,k为大于等于1且小于等于M的整数。

[0091] 具体来说,M个第一类实体中每一个第一类实体可以包含一个目标实体标准名以及一个或多个目标实体别名;可以查找目标实体标准名以及一个或多个目标实体别名所对应的一个或多个文档材料,从一个或多个文档材料中提取一个或多个第二类实体。这样就可以获取与每一个第一类实体具备关系的相关第二类实体。

[0092] 其中,所述第二类实体具体可以指的是关系图谱中的“人”实体。

[0093] 进一步地,可以在关系图谱中建立第一类实体与其具备关联关系的相关第二类实体之间的关系。也就是说,先得到与每一个第一类实体具备关系的一个或多个第二类实体,然后将每一个第一类实体与其相关的一个或多个第二类实体之间的关联关系添加到所述关系图谱中。

[0094] 其中,所述第二类实体可以为人,人在关系图谱中可以表现为人的名称;另外,第二类实体,比如人,也可以包括相关的属性信息或称为实体信息,比如,可以包括人的职位、职称等等,这里不做穷举。

[0095] 如此,可以通过事的名字确定关系图谱中的相关第二类实体,从而完善关系图谱的构建,由于事的实体名称的构建与获取相关第二类实体的材料是相同的,因此仅需要预先分析出来事的实体就可以构建关系图谱中的事跟相关第二类实体之间的关系,从而提升了构建关系图谱的效率。

[0096] 本发明实施例还提供一种实体信息处理装置,如图3所示,包括:

[0097] 识别模块31,用于对目标部门的N个文档材料进行识别,得到所述N个文档材料分别对应的候选实体名;N为大于等于1的整数;

[0098] 聚簇模块32,用于基于所述N个文档材料分别对应的候选实体名,生成所述目标部门对应的M个候选簇;M为大于等于1的整数;

[0099] 实体名确定模块33,用于基于所述M个候选簇分别包含的候选实体名,确定关系图谱中所述目标部门所对应的M个第一类实体的目标实体名。

[0100] 所述识别模块31,用于将所述目标部门的所述N个文档材料中的第j个文档材料及其对应的目标部门输入至预设模型,得到所述预设模型输出的所述第j个文档材料所对应的候选实体名;其中,j为大于等于1且小于等于N的整数。

[0101] 在图3的基础上,本实施例提供的信息处理装置,如图4所示,所述装置还包括:

[0102] 训练集构建模块34,用于获取多个部门分别对应的历史候选实体名;将所述多个部门中各个部门的历史文档材料与其对应的部门的历史候选实体名进行匹配,得到所述各个部门的历史文档材料所对应的历史实体名;基于所述各个部门的历史文档材料及其对应的历史实体名生成训练集。

[0103] 如图4所示,所述装置还包括:

[0104] 模型训练模块35,用于基于所述训练集中包含的所述各个部门的历史文档材料及

其对应的历史实体名对所述预设模型进行训练,得到训练后的所述预设模型。

[0105] 所述聚簇模块32,用于从所述N个文档材料分别对应的N个候选实体名中筛选得到L个候选实体名;L为大于等于1且小于等于N的整数;对所述L个候选实体名进行聚类,得到所述目标部门对应的M个候选簇;其中,所述M个候选簇中不同的候选簇包含不同的候选实体名。

[0106] 所述实体名确定模块33,用于获取所述M个候选簇中的第i个候选簇中包含的候选实体名的频次信息;其中,i为大于等于1且小于等于M的整数;将所述第i个候选簇中包含的候选实体名中频次信息最高的候选实体名作为所述第i个候选簇所对应的第i个第一类实体的目标实体标准名,将所述第i个候选簇中除所述目标标准实体名外的其他实体名作为所述第i个第一类实体的目标实体别名。

[0107] 如图4所示,所述装置还包括:

[0108] 关系构建模块36,用于从所述M个第一类实体中的第k个第一类实体的目标实体名所分别对应的文档材料中,获取所述第k个第一类实体关联的第二类实体;基于所述第k个第一类实体所关联的所述第二类实体,在所述关系图谱中建立所述第k个第一类实体与所述第二类实体之间的关联关系;其中,k为大于等于1且小于等于M的整数。

[0109] 根据本申请的实施例,本申请还提供了一种电子设备、一种可读存储介质和一种计算机程序产品。

[0110] 如图5所示,是根据本申请实施例的信息处理方法的电子设备的框图。电子设备旨在表示各种形式的数字计算机,诸如,膝上型计算机、台式计算机、工作台、个人数字助理、服务器、刀片式服务器、大型计算机、和其它适合的计算机。电子设备还可以表示各种形式的移动装置,诸如,个人数字处理、蜂窝电话、智能电话、可穿戴设备和其它类似的计算装置。本文所示的部件、它们的连接和关系、以及它们的功能仅仅作为示例,并且不意在限制本文中描述的和/或者要求的本申请的实现。

[0111] 如图5所示,该电子设备包括:一个或多个处理器701、存储器702,以及用于连接各部件的接口,包括高速接口和低速接口。各个部件利用不同的总线互相连接,并且可以被安装在公共主板上或者根据需要以其它方式安装。处理器可以对在电子设备内执行的指令进行处理,包括存储在存储器中或者存储器上以在外部输入/输出装置(诸如,耦合至接口的显示设备)上显示GUI的图形信息的指令。在其它实施方式中,若需要,可以将多个处理器和/或多条总线与多个存储器和多个存储器一起使用。同样,可以连接多个电子设备,各个设备提供部分必要的操作(例如,作为服务器阵列、一组刀片式服务器、或者多处理器系统)。图5中以一个处理器701为例。

[0112] 存储器702即为本申请所提供的非瞬时计算机可读存储介质。其中,所述存储器存储有可由至少一个处理器执行的指令,以使所述至少一个处理器执行本申请所提供的信息处理方法方法。本申请的非瞬时计算机可读存储介质存储计算机指令,该计算机指令用于使计算机执行本申请所提供的实体信息处理方法。

[0113] 存储器702作为一种非瞬时计算机可读存储介质,可用于存储非瞬时软件程序、非瞬时计算机可执行程序以及模块,如本申请实施例中的信息处理方法对应的程序指令/模块(例如,图4所示的识别模块、聚簇模块、实体名确定模块、训练集构建模块、模型训练模块)。处理器701通过运行存储在存储器702中的非瞬时软件程序、指令以及模块,从而执行

服务器的各种功能应用以及数据处理,即实现上述方法实施例中的信息处理方法。

[0114] 存储器702可以包括存储程序区和存储数据区,其中,存储程序区可存储操作系统、至少一个功能所需要的应用程序;存储数据区可存储根据信息处理方法的电子设备的使用所创建的数据等。此外,存储器702可以包括高速随机存取存储器,还可以包括非瞬时存储器,例如至少一个磁盘存储器件、闪存器件、或其他非瞬时固态存储器件。在一些实施例中,存储器702可选包括相对于处理器701远程设置的存储器,这些远程存储器可以通过网络连接至信息处理方法的电子设备。上述网络的实例包括但不限于互联网、企业内部网、局域网、移动通信网及其组合。

[0115] 信息处理方法的电子设备还可以包括:输入装置703和输出装置704。处理器701、存储器702、输入装置703和输出装置704可以通过总线或者其他方式连接,图5中以通过总线连接为例。

[0116] 输入装置703可接收输入的数字或字符信息,以及产生与电子设备的用户设置以及功能控制有关的键信号输入,例如触摸屏、小键盘、鼠标、轨迹板、触摸板、指示杆、一个或者多个鼠标按钮、轨迹球、操纵杆等输入装置。输出装置704可以包括显示设备、辅助照明装置(例如,LED)和触觉反馈装置(例如,振动电机)等。该显示设备可以包括但不限于,液晶显示器(LCD)、发光二极管(LED)显示器和等离子体显示器。在一些实施方式中,显示设备可以是触摸屏。

[0117] 此处描述的系统和技术的各种实施方式可以在数字电子电路系统、集成电路系统、专用ASIC(专用集成电路)、计算机硬件、固件、软件、和/或它们的组合中实现。这些各种实施方式可以包括:实施在一个或者多个计算机程序中,该一个或者多个计算机程序可在包括至少一个可编程处理器的可编程系统上执行和/或解释,该可编程处理器可以是专用或者通用可编程处理器,可以从存储系统、至少一个输入装置、和至少一个输出装置接收数据和指令,并且将数据和指令传输至该存储系统、该至少一个输入装置、和该至少一个输出装置。

[0118] 这些计算程序(也称作程序、软件、软件应用、或者代码)包括可编程处理器的机器指令,并且可以利用高级过程和/或面向对象的编程语言、和/或汇编/机器语言来实施这些计算程序。如本文使用的,术语“机器可读介质”和“计算机可读介质”指的是用于将机器指令和/或数据提供给可编程处理器的任何计算机程序产品、设备、和/或装置(例如,磁盘、光盘、存储器、可编程逻辑装置(PLD)),包括,接收作为机器可读信号的机器指令的机器可读介质。术语“机器可读信号”指的是用于将机器指令和/或数据提供给可编程处理器的任何信号。

[0119] 为了提供与用户的交互,可以在计算机上实施此处描述的系统和技术,该计算机具有:用于向用户显示信息的显示装置(例如,CRT(阴极射线管)或者LCD(液晶显示器)监视器);以及键盘和指向装置(例如,鼠标或者轨迹球),用户可以通过该键盘和该指向装置来将输入提供给计算机。其它种类的装置还可以用于提供与用户的交互;例如,提供给用户的反馈可以是任何形式的传感反馈(例如,视觉反馈、听觉反馈、或者触觉反馈);并且可以用任何形式(包括声输入、语音输入、或者触觉输入)来接收来自用户的输入。

[0120] 可以将此处描述的系统和技术实施在包括后台部件的计算系统(例如,作为数据服务器)、或者包括中间件部件的计算系统(例如,应用服务器)、或者包括前端部件的计算

系统(例如,具有图形用户界面或者网络浏览器的用户计算机,用户可以通过该图形用户界面或者该网络浏览器来与此处描述的系统和技术的实施方式交互)、或者包括这种后台部件、中间件部件、或者前端部件的任何组合的计算系统中。可以通过任何形式或者介质的数字数据通信(例如,通信网络)来将系统的部件相互连接。通信网络的示例包括:局域网(LAN)、广域网(WAN)和互联网。

[0121] 计算机系统可以包括客户端和服务器。客户端和服务器一般远离彼此并且通常通过通信网络进行交互。通过在相应的计算机上运行并且彼此具有客户端-服务器关系的计算机程序来产生客户端和服务器的关系。服务器可以是云服务器,又称为云计算服务器或云主机,是云计算服务体系中的一项主机产品,以解决传统物理主机与虚拟专用服务器(VPS)服务中,存在的管理难度大,业务扩展性弱的缺陷。服务器也可以为分布式系统的服务器,或者是结合了区块链的服务器。

[0122] 根据本申请实施例的技术方案,基于部门为单位输入部门的文档材料,确定文档材料对应的候选实体名,进而基于候选实体名确定在关系图谱中的每一个部门对应的一个或多个目标实体名,如此,仅通过收集部门的文档材料就可以最终确定关系图谱中包含的部门的目标实体名,可以避免人工分析所带来的效率较低、时效性较差以及结果不准确等问题,保证了获取目标实体名的处理效率以及准确性,进而也保证了构建或更新关系图谱的效率以及准确性。

[0123] 应该理解,可以使用上面所示的各种形式的流程,重新排序、增加或删除步骤。例如,本申请中记载的各步骤可以并行地执行也可以顺序地执行也可以不同的次序执行,只要能够实现本申请公开的技术方案所期望的结果,本文在此不进行限制。

[0124] 上述具体实施方式,并不构成对本申请保护范围的限制。本领域技术人员应该明白的是,根据设计要求和其他因素,可以进行各种修改、组合、子组合和替代。任何在本申请的精神和原则之内所作的修改、等同替换和改进等,均应包含在本申请保护范围之内。

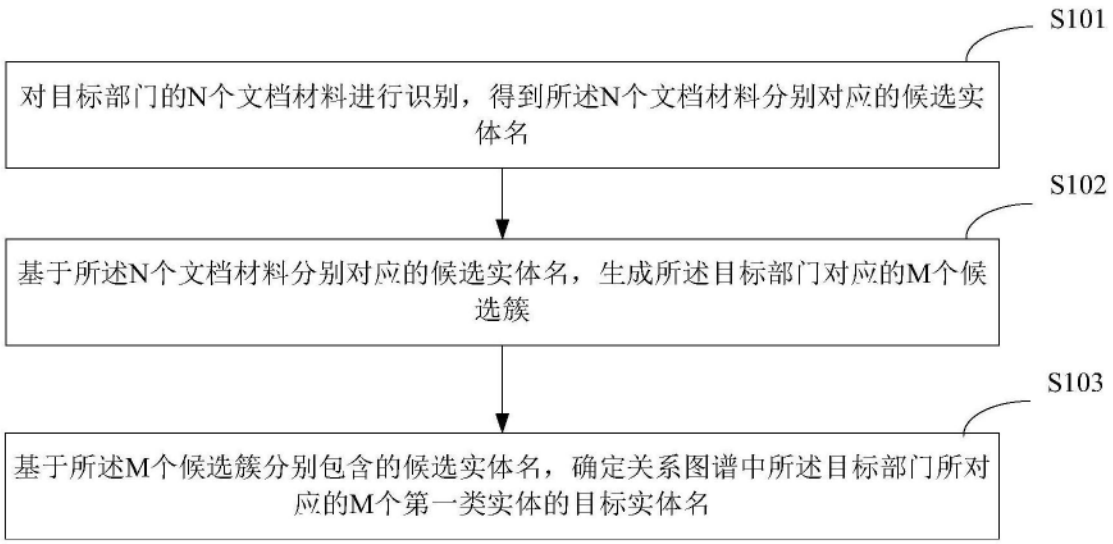

S101
对目标部门的N个文档材料进行识别，得到所述N个文档材料分别对应的候选实体名
S102
基于所述N个文档材料分别对应的候选实体名，生成所述目标部门对应的M个候选簇
S103
基于所述M个候选簇分别包含的候选实体名，确定关系图谱中所述目标部门所对应的M个第一类实体的目标实体名

图1

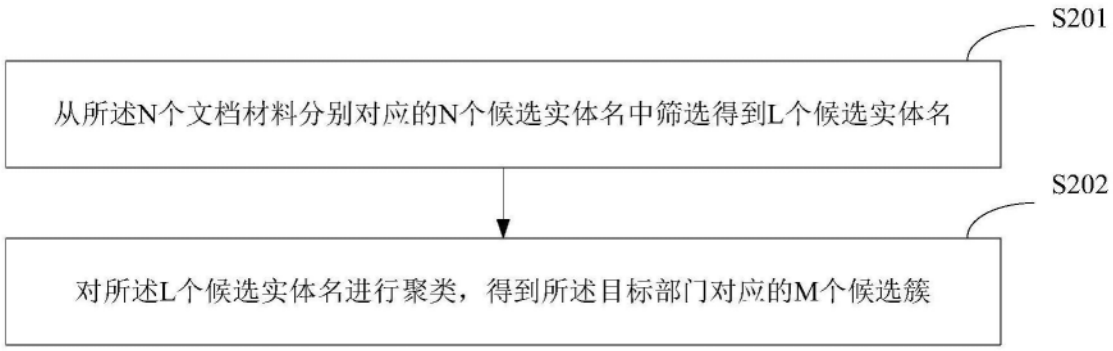

S201
从所述N个文档材料分别对应的N个候选实体名中筛选得到L个候选实体名
S202
对所述L个候选实体名进行聚类，得到所述目标部门对应的M个候选簇

图2

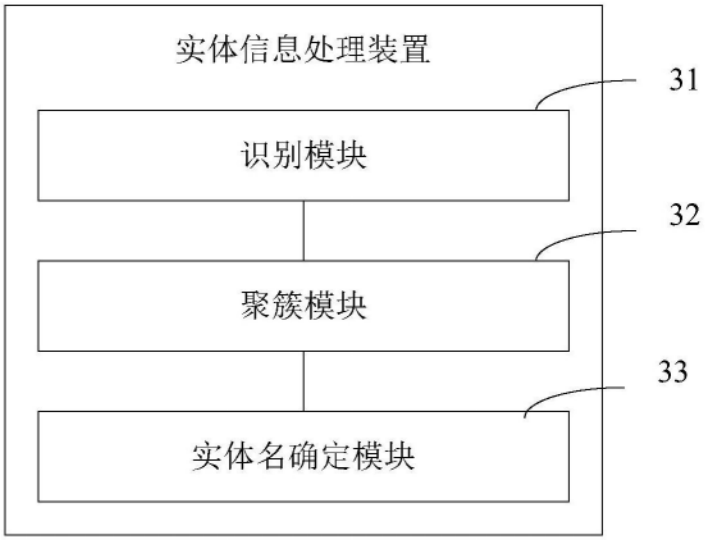

实体信息处理装置
31
识别模块
32
聚簇模块
33
实体名确定模块

图3

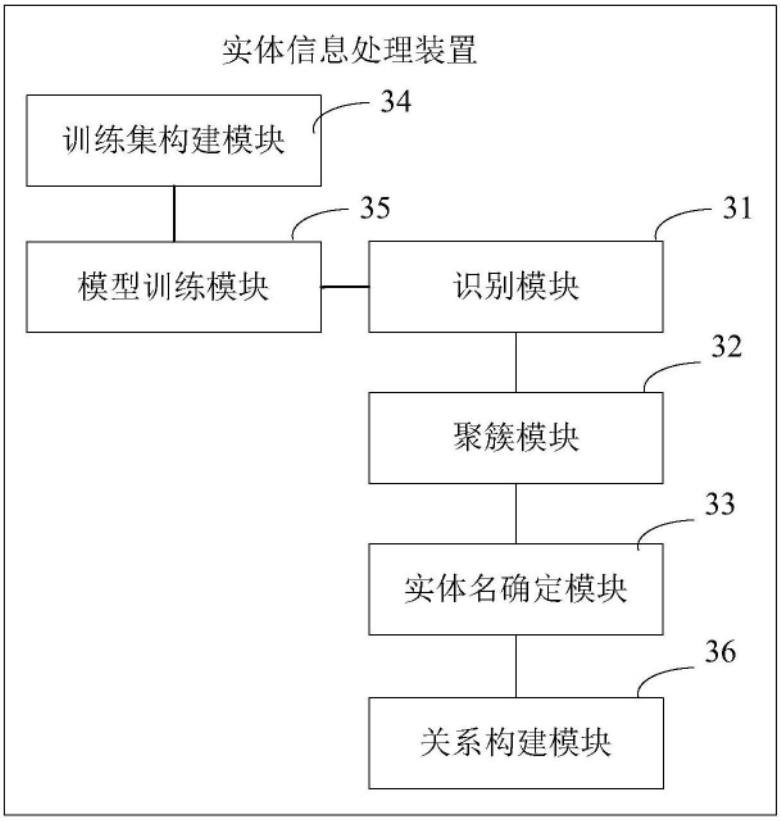

实体信息处理装置
34
训练集构建模块
35
模型训练模块
31
识别模块
32
聚簇模块
33
实体名确定模块
36
关系构建模块

图4

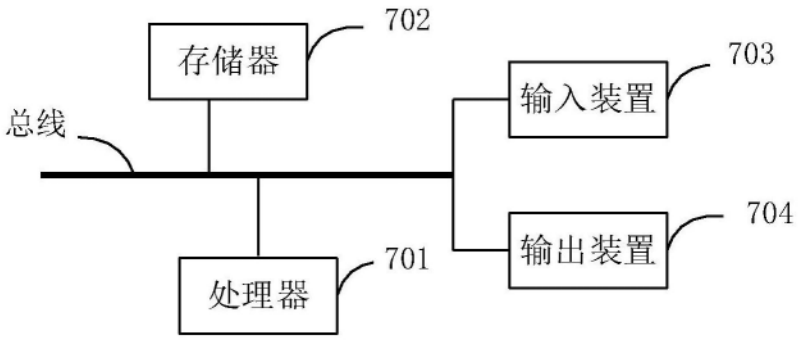


图5