



(12) 发明专利申请

(10) 申请公布号 CN 114706970 A

(43) 申请公布日 2022. 07. 05

(21) 申请号 202210333795.2

(22) 申请日 2022.03.30

(71) 申请人 北京明略昭辉科技有限公司

地址 100098 北京市海淀区北三环西路25
号27号楼二层2020室

(72) 发明人 林阔

(74) 专利代理机构 北京华夏泰和知识产权代理
有限公司 11662

专利代理师 李曼

(51) Int.Cl.

G06F 16/335 (2019.01)

G06F 40/289 (2020.01)

G06F 40/30 (2020.01)

G06K 9/62 (2022.01)

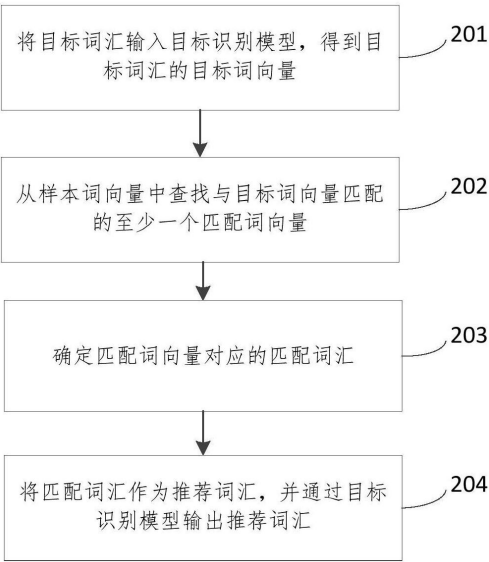
权利要求书2页 说明书9页 附图2页

(54) 发明名称

一种确定推荐词汇的方法、装置、电子设备和存储介质

(57) 摘要

本申请提供了一种确定推荐词汇的方法、装置、电子设备和存储介质。所述方法包括：将目标词汇输入目标识别模型，得到所述目标词汇的目标词向量，其中，所述目标识别模型能够缩小所述目标词汇和所述目标词汇的同义词之间的词向量距离，所述目标识别模型中具有至少一个样本词向量；从所述样本词向量中查找与所述目标词向量匹配的至少一个匹配词向量；确定所述匹配词向量对应的匹配词汇；将所述匹配词汇作为推荐词汇，并通过所述目标识别模型输出所述推荐词汇。本申请使推荐词汇更加准确。



1. 一种确定推荐词汇的方法,其特征在于,所述方法包括:

将目标词汇输入目标识别模型,得到所述目标词汇的目标词向量,其中,所述目标识别模型能够缩小所述目标词汇和所述目标词汇的同义词之间的词向量距离,所述目标识别模型中具有至少一个样本词向量;

从所述样本词向量中查找与所述目标词向量匹配的至少一个匹配词向量;

确定所述匹配词向量对应的匹配词汇;

将所述匹配词汇作为推荐词汇,并通过所述目标识别模型输出所述推荐词汇。

2. 根据权利要求1所述的方法,其特征在于,将目标词汇输入所述目标识别模型之前,所述方法还包括:

获取样本文档,其中,所述样本文档中包括至少一个样本词汇;

获取样本词汇和所述样本词汇的至少一个同义词汇,其中,所述同义词汇与所述样本词汇的含义相同;

将所述样本词汇和所述至少一个同义词汇输入词向量模型,其中,所述词向量模型用于生成词汇的词向量;

通过所述至少一个同义词汇,缩短所述样本词汇的词向量和所述同义词汇的词向量之间的向量距离,得到训练后的词向量模型;

将每个样本词汇输入所述训练后的词向量模型,得到目标识别模型。

3. 根据权利要求2所述的方法,其特征在于,所述将每个样本词汇输入所述训练后的词向量模型,得到目标识别模型包括:

将每个样本词汇输入所述训练后的词向量模型,得到每个样本词汇的初始词向量;

根据预设筛选方案,从所述初始词向量中筛选出与业务场景相关联的样本词向量;

将包含所述样本词向量的模型作为所述目标识别模型。

4. 根据权利要求2所述的方法,其特征在于,获取样本文档之前,所述方法还包括:

识别预设文档中的待选实体和所述待选实体的实体类别;

根据实体类别从所述待选实体中筛选样本实体,得到包含至少一个样本实体的词表;

通过分词工具将所述词表进行分词,得到包含样本词汇的样本文档,其中,所述样本词汇为样本实体。

5. 根据权利要求1所述的方法,其特征在于,将目标词汇输入所述目标识别模型之前,所述方法还包括:

获取查询词汇;

在所述查询词汇中包含目标实体的情况下,将所述查询词汇作为所述目标词汇;

在所述查询词汇中不包含目标实体、或在所述样本词向量中未匹配到所述目标实体的目标词向量的情况下,将所述查询词汇进行分词,得到分词后的目标词汇。

6. 根据权利要求2所述的方法,其特征在于,所述通过所述至少一个同义词汇,缩短所述样本词汇的词向量和所述同义词汇的词向量之间的向量距离包括:

在所述同义词汇为一个的情况下,将所述同义词汇作为监督数据,缩短所述样本词汇的词向量和所述同义词汇的词向量之间的向量距离;

在所述同义词汇为至少两个的情况下,将所述至少两个同义词汇作为监督数据,缩短所述样本词汇的词向量和每个所述同义词汇的词向量之间的向量距离。

7. 一种确定推荐词汇的装置,其特征在于,所述装置包括:

输入模块,用于将目标词汇输入目标识别模型,得到所述目标词汇的目标词向量,其中,所述目标识别模型能够缩小所述目标词汇和所述目标词汇的同义词之间的词向量距离,所述目标识别模型中具有至少一个样本词向量;

查找模块,用于从所述样本词向量中查找与所述目标词向量匹配的至少一个匹配词向量;

确定模块,用于确定所述匹配词向量对应的匹配词汇;

输出模块,用于将所述匹配词汇作为推荐词汇,并通过所述目标识别模型输出所述推荐词汇。

8. 根据权利要求7所述的装置,其特征在于,该装置还用于:

获取样本文档,其中,所述样本文档中包括至少一个样本词汇;

获取样本词汇和所述样本词汇的至少一个同义词汇,其中,所述同义词汇与所述样本词汇的含义相同;

将所述样本词汇和所述至少一个同义词汇输入词向量模型,其中,所述词向量模型用于生成词汇的词向量;

通过所述至少一个同义词汇,缩短所述样本词汇的词向量和所述同义词汇的词向量之间的向量距离,得到训练后的词向量模型;

将每个样本词汇输入所述训练后的词向量模型,得到目标识别模型。

9. 一种电子设备,其特征在于,包括处理器、通信接口、存储器和通信总线,其中,处理器,通信接口,存储器通过通信总线完成相互间的通信;

存储器,用于存放计算机程序;

处理器,用于执行存储器上所存放的程序时,实现权利要求1-6任一所述的方法步骤。

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内存储有计算机程序,所述计算机程序被处理器执行时实现权利要求1-6任一所述的方法步骤。

一种确定推荐词汇的方法、装置、电子设备和存储介质

技术领域

[0001] 本申请涉及推荐算法技术领域,尤其涉及一种确定推荐词汇的方法、装置、电子设备和存储介质。

背景技术

[0002] 推荐算法是根据用户的行为,背靠大量的用户日志和知识图谱,推测出用户可能感兴趣的词汇的一种算法。推荐算法在上世纪九十年代被提出,现在已经被广泛应用在搜索引擎中。

[0003] 目前的推荐方法主要为:收集用户的搜索记录,通过统计的方式从搜索记录中挖掘出推荐词汇,其中,推荐词汇可以相同含义但不同表达方式的词汇,也可以字面意思相近的词汇,也可以为含义相关联的词汇。

[0004] 但利用搜索记录确定推荐词汇,只能满足统计学上的相似,即字面层面上的相似,并不能保证实际语义上的相似。例如知识图谱和知识分子只有字面层面上的相似,并没有实际语义上的相似。因此,采用搜索记录推荐词汇的方式并不准确。

发明内容

[0005] 本申请实施例的目的在于提供一种确定推荐词汇的方法、装置、电子设备和存储介质,以解决推荐词汇不准确的问题。具体技术方案如下:

[0006] 第一方面,提供了一种确定推荐词汇的方法,所述方法包括:

[0007] 将目标词汇输入目标识别模型,得到所述目标词汇的目标词向量,其中,所述目标识别模型能够缩小所述目标词汇和所述目标词汇的同义词之间的词向量距离,所述目标识别模型中具有至少一个样本词向量;

[0008] 从所述样本词向量中查找与所述目标词向量匹配的至少一个匹配词向量;

[0009] 确定所述匹配词向量对应的匹配词汇;

[0010] 将所述匹配词汇作为推荐词汇,并通过所述目标识别模型输出所述推荐词汇。

[0011] 可选地,将目标词汇输入所述目标识别模型之前,所述方法还包括:

[0012] 获取样本文档,其中,所述样本文档中包括至少一个样本词汇;

[0013] 获取样本词汇和所述样本词汇的至少一个同义词汇,其中,所述同义词汇与所述样本词汇的含义相同;

[0014] 将所述样本词汇和所述至少一个同义词汇输入词向量模型,其中,所述词向量模型用于生成词汇的词向量;

[0015] 通过所述至少一个同义词汇,缩短所述样本词汇的词向量和所述同义词汇的词向量之间的向量距离,得到训练后的词向量模型;

[0016] 将每个样本词汇输入所述训练后的词向量模型,得到目标识别模型。

[0017] 可选地,所述将每个样本词汇输入所述训练后的词向量模型,得到目标识别模型包括:

- [0018] 将每个样本词汇输入所述训练后的词向量模型,得到每个样本词汇的初始词向量;
- [0019] 根据预设筛选方案,从所述初始词向量中筛选出与业务场景相关联的样本词向量;
- [0020] 将包含所述样本词向量的模型作为所述目标识别模型。
- [0021] 可选地,获取样本文档之前,所述方法还包括:
- [0022] 识别预设文档中的待选实体和所述待选实体的实体类别;
- [0023] 根据实体类别从所述待选实体中筛选样本实体,得到包含至少一个样本实体的词表;
- [0024] 通过分词工具将所述词表进行分词,得到包含样本词汇的样本文档,其中,所述样本词汇为样本实体。
- [0025] 可选地,将目标词汇输入所述目标识别模型之前,所述方法还包括:
- [0026] 获取查询词汇;
- [0027] 在所述查询词汇中包含目标实体的情况下,将所述查询词汇作为所述目标词汇;
- [0028] 在所述查询词汇中不包含目标实体、或在所述样本词向量中未匹配到所述目标实体的目标词向量的情况下,将所述查询词汇进行分词,得到分词后的目标词汇。
- [0029] 可选地,所述通过所述至少一个同义词汇,缩短所述样本词汇的词向量和所述同义词汇的词向量之间的向量距离包括:
- [0030] 在所述同义词汇为一个的情况下,将所述同义词汇作为监督数据,缩短所述样本词汇的词向量和所述同义词汇的词向量之间的向量距离;
- [0031] 在所述同义词汇为至少两个的情况下,将所述至少两个同义词汇作为监督数据,缩短所述样本词汇的词向量和每个所述同义词汇的词向量之间的向量距离。
- [0032] 第二方面,提供了一种确定推荐词汇的装置,所述装置包括:
- [0033] 输入模块,用于将目标词汇输入目标识别模型,得到所述目标词汇的目标词向量,其中,所述目标识别模型能够缩小所述目标词汇和所述目标词汇的同义词之间的词向量距离,所述目标识别模型中具有至少一个样本词向量;
- [0034] 查找模块,用于从所述样本词向量中查找与所述目标词向量匹配的至少一个匹配词向量;
- [0035] 确定模块,用于确定所述匹配词向量对应的匹配词汇;
- [0036] 输出模块,用于将所述匹配词汇作为推荐词汇,并通过所述目标识别模型输出所述推荐词汇。
- [0037] 可选地,该装置还用于:
- [0038] 获取样本文档,其中,所述样本文档中包括至少一个样本词汇;
- [0039] 获取样本词汇和所述样本词汇的至少一个同义词汇,其中,所述同义词汇与所述样本词汇的含义相同;
- [0040] 将所述样本词汇和所述至少一个同义词汇输入词向量模型,其中,所述词向量模型用于生成词汇的词向量;
- [0041] 通过所述至少一个同义词汇,缩短所述样本词汇的词向量和所述同义词汇的词向量之间的向量距离,得到训练后的词向量模型;

[0042] 将每个样本词汇输入所述训练后的词向量模型,得到目标识别模型。

[0043] 第三方面,提供了一种电子设备,包括处理器、通信接口、存储器和通信总线,其中,处理器,通信接口,存储器通过通信总线完成相互间的通信;

[0044] 存储器,用于存放计算机程序;

[0045] 处理器,用于执行存储器上所存放的程序时,实现任一所述的确定推荐词汇的方法步骤。

[0046] 第四方面,提供了一种计算机可读存储介质,所述计算机可读存储介质内存储有计算机程序,所述计算机程序被处理器执行时实现任一所述的确定推荐词汇的方法步骤。

[0047] 本申请实施例有益效果:

[0048] 本申请应用于推荐技术领域的推荐算法,在本申请中,目标识别模型得到目标词汇的目标词向量,然后根据目标词向量得到匹配词向量,再反推出匹配词向量的匹配词汇。目标识别模型能够缩小目标词汇和目标词汇的同义词之间的词向量距离,一方面使输出的目标词向量更加准确,另一方面目标识别模型得到的目标词向量在向量距离上也更加靠近匹配词向量,这样目标词向量对应的匹配词向量也更加准确,从而使匹配词向量对应的匹配词汇(推荐词汇)也更加准确。

[0049] 当然,实施本申请的任一产品或方法并不一定需要同时达到以上的所有优点。

附图说明

[0050] 为了更清楚地说明本申请实施例或现有技术中的技术方案,下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍,显而易见地,对于本领域普通技术人员而言,在不付出创造性劳动性的前提下,还可以根据这些附图获得其他的附图。

[0051] 图1为本申请实施例提供的一种确定推荐词汇的方法硬件环境示意图;

[0052] 图2为本申请实施例提供的一种确定推荐词汇的方法流程图;

[0053] 图3为本申请实施例提供的一种确定推荐词汇的装置的结构示意图;

[0054] 图4为本申请实施例提供的一种电子设备的结构示意图。

具体实施方式

[0055] 为使本申请实施例的目的、技术方案和优点更加清楚,下面将结合本申请实施例中的附图,对本申请实施例中的技术方案进行清楚、完整地描述,显然,所描述的实施例是本申请的一部分实施例,而不是全部的实施例。基于本申请中的实施例,本领域普通技术人员在没有做出创造性劳动的前提下所获得的所有其他实施例,都属于本申请保护的范围。

[0056] 在后续的描述中,使用用于表示元件的诸如“模块”、“部件”或“单元”的后缀仅为了有利于本申请的说明,其本身并没有特定的意义。因此,“模块”与“部件”可以混合地使用。

[0057] 为了解决背景技术中提及的问题,根据本申请实施例的一方面,提供了一种确定推荐词汇的方法的实施例。

[0058] 可选地,在本申请实施例中,上述确定推荐词汇的方法可以应用于如图1所示的由终端101和服务器103所构成的硬件环境中。如图1所示,服务器103通过网络与终端101进行连接,可用于为终端或终端上安装的客户端提供服务,可在服务器上或独立于服务器设置

数据库105,用于为服务器103提供数据存储服务,上述网络包括但不限于:广域网、城域网或局域网,终端101包括但不限于PC、手机、平板电脑等。

[0059] 本申请实施例中的一种确定推荐词汇的方法可以由服务器103来执行,还可以是由终端101来执行,用于确定提高生成推荐词汇的准确性。

[0060] 下面将结合具体实施方式,以应用于服务器为例,对本申请实施例提供的一种确定推荐词汇的方法进行详细的说明,如图2所示,具体步骤如下:

[0061] 步骤201:将目标词汇输入目标识别模型,得到目标词汇的目标词向量。

[0062] 其中,目标识别模型能够缩小目标词汇和目标词汇的同义词之间的词向量距离,目标识别模型中具有至少一个样本词向量。

[0063] 在本申请实施例中,目标词汇为用户在搜索框中输入的查询词汇,服务器将目标词汇输入目标识别模型,目标识别模型用于将目标词汇转换为目标词向量,目标识别模型还可以在向量转换过程中,缩小目标词汇和目标词汇的同义词之间的词向量距离,这样可以使目标词向量更加准确。其中,目标词向量为目标词汇的真实含义的向量。

[0064] 样本文档中包括至少一个样本词汇,目标识别模型包含样本词汇的样本词向量。

[0065] 步骤202:从样本词向量中查找与目标词向量匹配的至少一个匹配词向量。

[0066] 在本申请实施例中,服务器确定目标词汇的目标词向量后,从样本词向量中查找与目标词向量匹配的至少一个匹配词向量,其中,匹配指样本词向量和匹配词向量之间的向量距离小于预设距离阈值。

[0067] 步骤203:确定匹配词向量对应的匹配词汇。

[0068] 在本申请实施例中,服务器通过目标识别网络反推出匹配词向量对应的匹配词汇。

[0069] 步骤204:将匹配词汇作为推荐词汇,并通过目标识别模型输出推荐词汇。

[0070] 在本申请实施例中,由于样本词向量和匹配词向量之间的向量距离小于预设距离阈值,表示样本词向量和匹配词向量很接近,那么样本词汇和匹配词语的语义也接近,则服务器将匹配词汇作为推荐词汇,并通过目标识别模型输出推荐词汇。

[0071] 其中,本申请实施例中的推荐词汇,不局限于含义相同的词汇,还可以为具有其他关联关系的词汇,示例性地,两个动物同属于家养动物,两个球员同属于一个球队,两个人均任职过一个职位等。

[0072] 本申请中缩小目标词汇和目标词汇的同义词之间的词向量距离,目的是为了调整目标识别模型的模型参数,使目标识别模型能够查找和目标词向量最接近的匹配词向量,但该匹配词向量对应的匹配词汇(推荐词汇)并不局限于同义词,还可以为相关联的词汇。

[0073] 在本申请中,目标识别模型得到目标词汇的目标词向量,然后根据目标词向量得到匹配词向量,再反推出匹配词向量的匹配词汇。目标识别模型能够缩小目标词汇和目标词汇的同义词之间的词向量距离,一方面使输出的目标词向量更加准确,另一方面目标识别模型得到的目标词向量在向量距离上也更加靠近匹配词向量,这样目标词向量对应的匹配词向量也更加准确,从而使匹配词向量对应的匹配词汇(推荐词汇)也更加准确。

[0074] 另外,相对于现有技术中构建领域知识库从而控制推荐词汇和目标词汇的语义相关度,本申请无需构建领域知识库,降低了成本,也省去了将领域知识转化为机器学习算法的难题。

[0075] 作为一种可选的实施方式,将目标词汇输入目标识别模型之前,方法还包括:获取样本文档,其中,所述样本文档中包括至少一个样本词汇;获取样本词汇和所述样本词汇的至少一个同义词汇,其中,所述同义词汇与所述样本词汇的含义相同;将所述样本词汇和所述至少一个同义词汇输入词向量模型,其中,所述词向量模型用于生成词汇的词向量;通过所述至少一个同义词汇,缩短所述样本词汇的词向量和所述同义词汇的词向量之间的向量距离,得到训练后的词向量模型;将每个样本词汇输入所述训练后的词向量模型,得到目标识别模型。

[0076] 服务器获取包含至少一个样本词汇的样本文档,然后获取样本文档中的样本词汇和该样本词汇的至少一个同义词汇,其中,同义词汇与样本词汇的含义相同而表述方式不同,服务器将样本词汇和至少一个同义词汇输入词向量模型,词向量模型用于生成词汇的词向量,服务器将至少一个同义词汇作为监督数据,这样缩短样本词汇的词向量和同义词汇的词向量之间的向量距离,从而使样本词向量和同义词汇的词向量之间的向量距离更接近。其中,词向量模型可以为word2vec,也可以为glove,本申请对词向量模型不做具体限制。

[0077] 其中,若同义词汇的数量为一个,则服务器将该同义词汇作为监督数据,缩短样本词汇的词向量和该同义词汇的词向量之间的向量距离;若同义词汇的数量为至少两个,则服务器将该至少两个同义词汇作为监督数据,缩短样本词汇的词向量和每个同义词汇的词向量之间的向量距离。这样样本词汇的词向量是接近每个同义词汇的词向量的,提高了样本词汇的词向量的准确性。

[0078] 服务器采用上述方式对词向量模型进行训练,得到训练后的词向量模型。服务器每个样本词汇输入训练后的词向量模型,得到每个样本词汇的样本向量,从而得到包含至少一个样本向量的目标识别模型。

[0079] 在本申请中,服务器通过采用同义词汇作为监督数据,缩小样本词向量和同义词汇的词向量之间的向量距离,可以使训练后的词向量模型一方面能够修正样本词汇的词向量,提高样本词汇的词向量的准确性,另一方面通过词向量模型的训练,可以使得到的目标识别模型,能够查找到和目标词向量距离接近的匹配词向量。

[0080] 作为一种可选的实施方式,将每个样本词汇输入所述训练后的词向量模型,得到目标识别模型包括:将每个样本词汇输入所述训练后的词向量模型,得到每个样本词汇的初始词向量;根据预设筛选方案,从所述初始词向量中筛选出与业务场景相关联的样本词向量;将包含所述样本词向量的模型作为所述目标识别模型。

[0081] 服务器将每个样本词汇输入训练后的词向量模型,训练后的词向量模型能够准确的得到样本词汇的初始词向量。服务器对训练后的词向量模型进行解析,通过提前设定的与业务场景相关联的筛选方案,从初始词向量中筛选出与业务场景相关联的样本词向量,然后将包含与业务场景相关联的样本词向量的模型作为目标识别模型。

[0082] 在本申请中,服务器对对训练后的词向量模型进行解析,去除掉和业务场景无关的词向量,能够实现对词向量模型的数据压缩,提高查询的响应速度。

[0083] 作为一种可选的实施方式,获取样本文档之前,方法还包括:识别预设文档中的待选实体和待选实体的实体类别;根据实体类别从待选实体中筛选样本实体,得到包含至少一个样本实体的词表;通过分词工具将词表进行分词,得到包含样本词汇的样本文档,其

中,样本词汇为样本实体。

[0084] 服务器获取预设文档,预设文档中包含查询词汇,服务器对预设文档进行实体识别,得到文档中的待选实体和待选实体的实体类别,示例性地,实体类别可以为人名、地名、行业名称、项目名称、电话号码等,本申请对实体类别不做具体限制。服务器对待选实体进行去重,避免实体重复。服务器根据实体类别,从待选实体中筛选样本实体,得到包含至少一个样本实体的词表,然后通过分词工具将词表进行分词,由于已经事先确认了样本实体,因此分词工具不会再对实体进行分词,可以保证实体的完整性,分词结束后,服务器得到至少一个样本词汇,每个样本词汇都是一个样本实体,然后根据该至少一个样本词汇构建样本文档。

[0085] 在本申请中,服务器进行分词,可以使得到的分词满足词向量模型的输入格式。另外,本申请先识别实体再分词,可以避免对实体进行分词,保证了实体的完整性。

[0086] 可选地,实体识别是为了选取预设文档中有效的样本词汇,本申请不仅限于采用实体识别的方式选取样本词汇,还可以采用预设选取条件从预设文档中的词汇中选取样本词汇。

[0087] 作为一种可选的实施方式,将目标词汇输入目标识别模型之前,方法还包括:获取查询词汇;在查询词汇中包含目标实体的情况下,将查询词汇作为目标词汇。在查询词汇中不包含目标实体、或在样本词向量中未匹配到目标实体的目标词向量的情况下,将查询词汇进行分词,得到分词后的目标词汇。

[0088] 服务器获取查询词汇,服务器对查询词汇进行实体识别,若能够识别到实体,确定该实体的实体类别,并根据实体类别筛选需要的目标实体。服务器将目标实体作为目标词汇,然后通过目标识别模型得到目标词汇的目标词向量,再在样本词向量中查找匹配词向量,从而得到推荐词汇。

[0089] 若服务器未识别到实体,或在样本词向量中未查找到该目标词向量的匹配词向量,则无法得到推荐词汇,则服务器将查询词汇进行分词处理,得到查询词汇更为细化的分词,则有可能在目标识别模型中查找到与分词的词向量对应的匹配词向量,从而得到推荐词汇。提高了得到推荐词汇的概率。

[0090] 可选的,本申请实施例还提供了一种确定推荐词汇方法的处理流程,具体步骤如下:

[0091] 步骤1:对预设文档进行实体识别、去重、筛选和分词,得到样本文档,其中,样本文档中包括至少一个样本词汇。

[0092] 步骤2:采用同义词汇作为样本词汇的监督数据,对词向量模型进行训练,得到训练后的词向量模型。

[0093] 步骤3:采用训练后的词向量模型,得到每个样本词汇的样本词向量,形成目标识别模型。

[0094] 步骤4:将查询词汇进行实体识别。

[0095] 步骤5:判断是否识别到实体,若否,则执行步骤6,若是,则执行步骤7。

[0096] 步骤6:则将查询词汇进行分词,再输入目标识别模型,得到推荐词汇。

[0097] 步骤7:将目标实体输入目标识别模型,判断是否得到输出的推荐词汇,若是,则执行步骤8。若否,则返回步骤6。

[0098] 步骤8:推荐结束。

[0099] 基于相同的技术构思,本申请实施例还提供了一种确定推荐词汇的装置,如图3所示,该装置包括:

[0100] 第一输入模块301,用于将目标词汇输入目标识别模型,得到目标词汇的目标词向量,其中,目标识别模型能够缩小目标词汇和目标词汇的同义词之间的词向量距离,目标识别模型中具有至少一个样本词向量;

[0101] 查找模块302,用于从样本词向量中查找与目标词向量匹配的至少一个匹配词向量;

[0102] 确定模块303,用于确定匹配词向量对应的匹配词汇;

[0103] 输出模块304,用于将匹配词汇作为推荐词汇,并通过目标识别模型输出推荐词汇。

[0104] 可选地,该装置还包括:

[0105] 第一获取模块,用于获取样本文档,其中,所述样本文档中包括至少一个样本词汇;

[0106] 第二获取模块,用于获取样本词汇和样本词汇的至少一个同义词汇,其中,同义词汇与样本词汇的含义相同;

[0107] 第二输入模块,用于将样本词汇和至少一个同义词汇输入词向量模型,其中,词向量模型用于生成词汇的词向量;

[0108] 第一得到模块,用于通过至少一个同义词汇,缩短样本词汇的词向量和同义词汇的词向量之间的向量距离,得到训练后的词向量模型;

[0109] 第二得到模块,用于将每个样本词汇输入训练后的词向量模型,得到目标识别模型,其中,样本文档中包括至少一个样本词汇。

[0110] 可选地,第二得到模块用于:

[0111] 将每个样本词汇输入所述训练后的词向量模型,得到每个样本词汇的初始词向量;

[0112] 根据预设筛选方案,从所述初始词向量中筛选出与业务场景相关联的样本词向量;

[0113] 将包含所述样本词向量的模型作为所述目标识别模型。

[0114] 可选地,该装置还用于:

[0115] 识别预设文档中的待选实体和待选实体的实体类别;

[0116] 根据实体类别从待选实体中筛选样本实体,得到包含至少一个样本实体的词表;

[0117] 通过分词工具将词表进行分词,得到包含样本词汇的样本文档,其中,样本词汇为样本实体。

[0118] 可选地,该装置还用于:

[0119] 获取查询词汇;

[0120] 在查询词汇中包含目标实体的情况下,将查询词汇作为目标词汇;

[0121] 在查询词汇中不包含目标实体、或在样本词向量中未匹配到目标实体的目标词向量的情况下,将查询词汇进行分词,得到分词后的目标词汇。

[0122] 可选地,第一得到模块用于:

[0123] 在同义词汇为一个的情况下,将同义词汇作为监督数据,缩短样本词汇的词向量和同义词汇的词向量之间的向量距离;

[0124] 在同义词汇为至少两个的情况下,将至少两个同义词汇作为监督数据,缩短样本词汇的词向量和每个同义词汇的词向量之间的向量距离。

[0125] 根据本申请实施例的另一方面,本申请提供了一种电子设备,如图4所示,包括存储器403、处理器401、通信接口402及通信总线404,存储器403中存储有可在处理器401上运行的计算机程序,存储器403、处理器401通过通信接口402和通信总线404进行通信,处理器401执行计算机程序时实现上述方法的步骤。

[0126] 上述电子设备中的存储器、处理器通过通信总线和通信接口进行通信。所述通信总线可以是外设部件互连标准(Peripheral Component Interconnect,简称PCI)总线或扩展工业标准结构(Extended Industry Standard Architecture,简称EISA)总线等。该通信总线可以分为地址总线、数据总线、控制总线等。

[0127] 存储器可以包括随机存取存储器(Random Access Memory,简称RAM),也可以包括非易失性存储器(non-volatile memory),例如至少一个磁盘存储器。可选的,存储器还可以是至少一个位于远离前述处理器的存储装置。

[0128] 上述的处理器可以是通用处理器,包括中央处理器(Central Processing Unit,简称CPU)、网络处理器(Network Processor,简称NP)等;还可以是数字信号处理器(Digital Signal Processing,简称DSP)、专用集成电路(Application Specific Integrated Circuit,简称ASIC)、现场可编程门阵列(Field-Programmable Gate Array,简称FPGA)或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件。

[0129] 根据本申请实施例的又一方面还提供了一种具有处理器可执行的非易失的程序代码的计算机可读介质。

[0130] 可选地,在本申请实施例中,计算机可读介质被设置为存储用于所述处理器执行上述方法的程序代码:

[0131] 可选地,本实施例中的具体示例可以参考上述实施例中所描述的示例,本实施例在此不再赘述。

[0132] 本申请实施例在具体实现时,可以参阅上述各个实施例,具有相应的技术效果。

[0133] 可以理解的是,本文描述的这些实施例可以用硬件、软件、固件、中间件、微码或其组合来实现。对于硬件实现,处理单元可以实现在一个或多个专用集成电路(Application Specific Integrated Circuits,ASIC)、数字信号处理器(Digital Signal Processing, DSP)、数字信号处理设备(DSP Device,DSPD)、可编程逻辑设备(Programmable Logic Device,PLD)、现场可编程门阵列(Field-Programmable Gate Array,FPGA)、通用处理器、控制器、微控制器、微处理器、用于执行本申请所述功能的其它电子单元或其组合中。

[0134] 对于软件实现,可通过执行本文所述功能的单元来实现本文所述的技术。软件代码可存储在存储器中并通过处理器执行。存储器可以在处理器中或在处理器外部实现。

[0135] 本领域普通技术人员可以意识到,结合本文中所公开的实施例描述的各示例的单元及算法步骤,能够以电子硬件、或者计算机软件和电子硬件的结合来实现。这些功能究竟以硬件还是软件方式来执行,取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能,但是这种实现不应认为超出

本申请的范围。

[0136] 所属领域的技术人员可以清楚地了解到,为描述的方便和简洁,上述描述的系统、装置和单元的具体工作过程,可以参考前述方法实施例中的对应过程,在此不再赘述。

[0137] 在本申请所提供的实施例中,应该理解到,所揭露的装置和方法,可以通过其它的方式实现。例如,以上所描述的装置实施例仅仅是示意性的,例如,所述模块的划分,仅仅为一种逻辑功能划分,实际实现时可以有另外的划分方式,例如多个模块或组件可以结合或者可以集成到另一个系统,或一些特征可以忽略,或不执行。另一点,所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口,装置或单元的间接耦合或通信连接,可以是电性,机械或其它的形式。

[0138] 所述作为分离部件说明的单元可以是或者也可以不是物理上分开的,作为单元显示的部件可以是或者也可以不是物理单元,即可以位于一个地方,或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

[0139] 另外,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单元中,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单元中。

[0140] 所述功能如果以软件功能单元的形式实现并作为独立的产品销售或使用,可以存储在一个计算机可读取存储介质中。基于这样的理解,本申请实施例的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的部分可以以软件产品的形式体现出来,该计算机软件产品存储在一个存储介质中,包括若干指令用以使得一台计算机设备(可以是个人计算机,服务器,或者网络设备等)执行本申请各个实施例所述方法的全部或部分步骤。而前述的存储介质包括:U盘、移动硬盘、ROM、RAM、磁碟或者光盘等各种可以存储程序代码的介质。需要说明的是,在本文中,诸如“第一”和“第二”等之类的关系术语仅仅用来将一个实体或者操作与另一个实体或操作区分开来,而不一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。而且,术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含,从而使得包括一系列要素的过程、方法、物品或者设备不仅包括那些要素,而且还包括没有明确列出的其他要素,或者是还包括为这种过程、方法、物品或者设备所固有的要素。在没有更多限制的情况下,由语句“包括一个……”限定的要素,并不排除在包括所述要素的过程、方法、物品或者设备中还存在另外的相同要素。

[0141] 以上所述仅是本申请的具体实施方式,使本领域技术人员能够理解或实现本申请。对这些实施例的多种修改对本领域的技术人员来说将是显而易见的,本文中所定义的一般原理可以在不脱离本申请的精神或范围的情况下,在其它实施例中实现。因此,本申请将不会被限制于本文所示的这些实施例,而是要符合与本文所申请的原理和新颖特点相一致的最宽的范围。

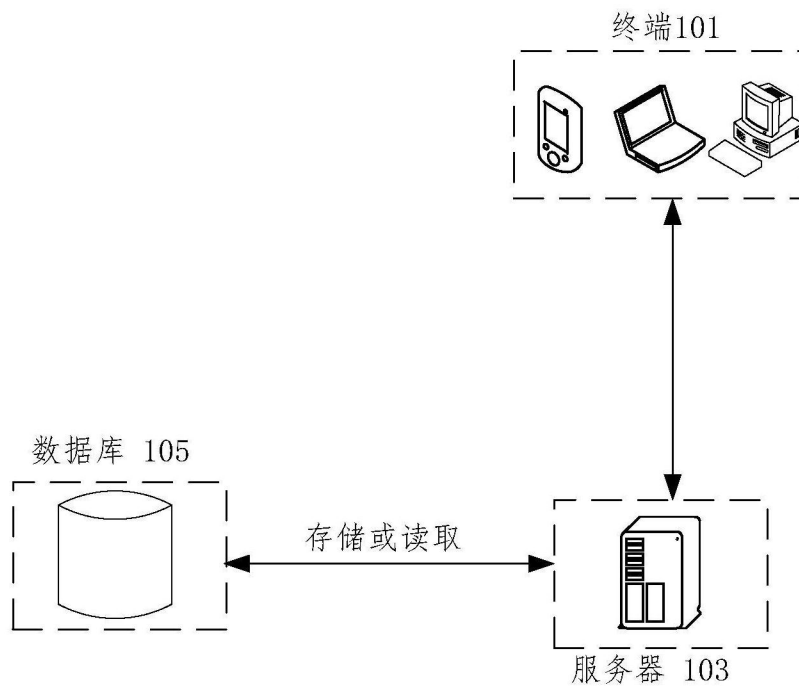


图1

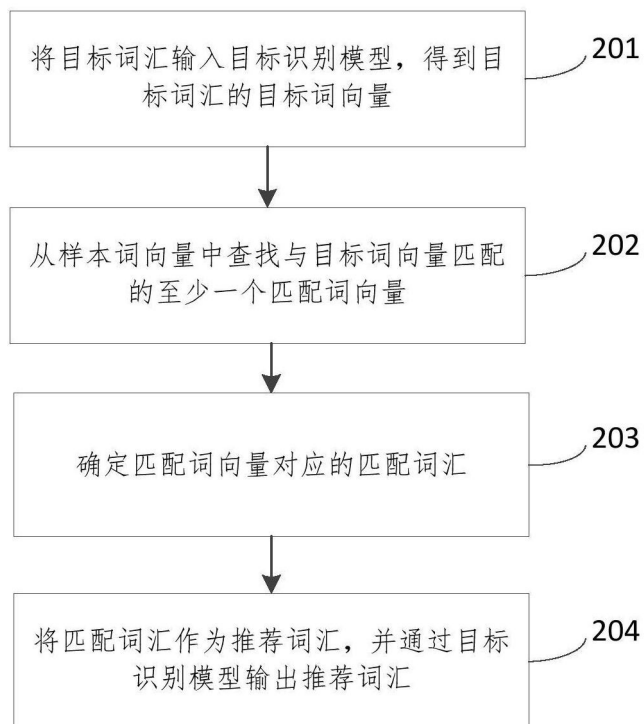


图2

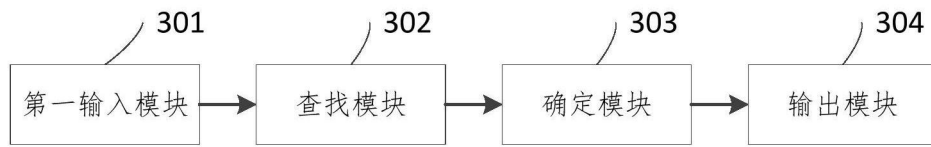


图3

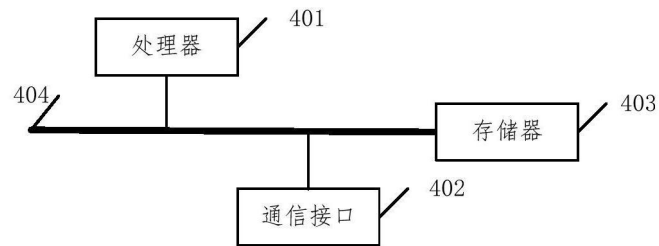


图4