(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 111460798 B
(45)授权公告日 2024.10.18

(21)申请号 202010136905.7

(22)申请日 2020.03.02

(65)同一申请的已公布的文献号
申请公布号 CN 111460798 A

(43)申请公布日 2020.07.28

(73)专利权人 平安科技(深圳)有限公司
地址 518000 广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼

(72)发明人 陈林 金戈 徐亮

(74)专利代理机构 深圳市赛恩倍吉知识产权代理有限公司 44334
专利代理师 钟良

(51)Int.Cl.
G06F 16/332(2019.01)
G06F 40/247(2020.01)
G06F 40/284(2020.01)
G06F 40/30(2020.01)
G06N 5/022(2023.01)
G06Q 10/1053(2023.01)

(56)对比文件
CN 109597988 A,2019.04.09
CN 109947922 A,2019.06.28

审查员 漆丽娟

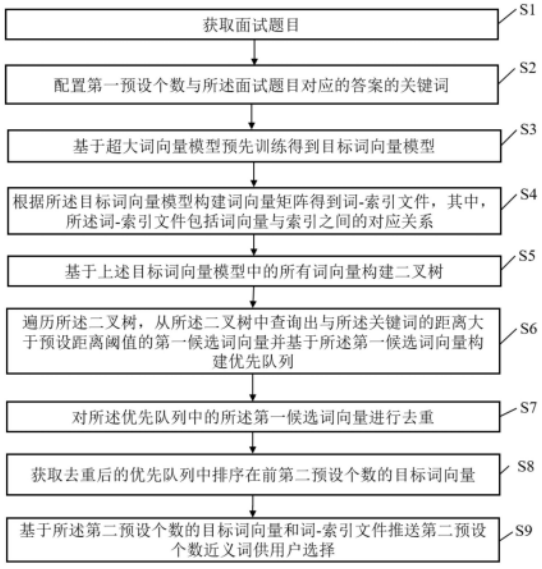
权利要求书3页 说明书13页 附图3页

(54)发明名称

近义词推送方法、装置、电子设备及介质

(57)摘要

本发明公开一种近义词推送方法,包括:获取面试题目;配置第一预设个数与面试题目对应的答案的关键词;基于超大词向量模型预先训练得到目标词向量模型;根据目标词向量模型构建词向量矩阵得到词-索引文件;基于所有词向量构建二叉树;遍历所述二叉树,从所述二叉树中查询出与所述关键词的距离大于预设距离阈值的第一候选词向量并构建优先队列;对优先队列中的所述第一候选词向量进行去重;获取去重后的优先队列中排序在前第二预设个数的目标词向量;基于目标词向量和词-索引文件推送第二预设个数近义词供用户选择。本发明还提供一种近义词推送装置、电子设备及存储介质。通过本发明可以为用户快速推送近义词。



CN 111460798 B

1.一种近义词推送方法,其特征在于,所述方法包括:

获取面试题目;

配置第一预设个数与所述面试题目对应的答案的关键词;

基于超大词向量模型预先训练得到目标词向量模型,包括:扩充所述超大词向量模型中的机器人面试场景语料,其中,包括对所述机器人面试场景语料进行分词、去停用词及基于CBOW模式增量训练词向量操作;根据扩充语料后的超大词向量模型训练得到目标词向量模型;

根据所述目标词向量模型构建词向量矩阵得到词-索引文件,包括:以每个词的维度为行数,以所述目标词向量模型中所有词的总数为列数构建词向量矩阵;所述词向量矩阵中的每一行对应一个索引;根据所述词向量矩阵构建词-索引文件,并输出所述词-索引文件,其中,所述词-索引文件包括词向量与索引之间的对应关系;

基于所述目标词向量模型中的所有词向量构建二叉树,包括:(1)随机选择两个点为初始节点,连接两个初始节点形成一个等距超平面;(2)根据所述两个初始节点的连线的中点垂直线构建一个等距垂直超平面,将所述目标词向量模型中的所有词向量对应的数据空间分成两部分,并得到两个子空间;(3)分别将每个子空间中的每个点与等距超平面的法向量相乘,求出每个点与法向量夹角的正负,以正负来分出其属于二叉树的左子树还是右子树;依此类推,分别在所述两个子空间内重复上述(1)至(3),可以将所述数据空间切分为多个子空间,并根据所述多个子空间构建所述二叉树;

遍历所述二叉树,从所述二叉树中查询出与所述关键词的距离大于预设距离阈值的第一候选词向量并基于所述第一候选词向量构建优先队列;

对所述优先队列中的所述第一候选词向量进行去重;

获取去重后的优先队列中排序在前第二预设个数的目标词向量;

基于所述第二预设个数的目标词向量和词-索引文件推送第二预设个数近义词供用户选择。

2.如权利要求1所述的近义词推送方法,其特征在于,所述配置第一预设个数与所述面试题目对应的答案的关键词的步骤包括:

根据预先构建的题目解析模型分析所述面试题目得到对应的题目意图;

根据所述题目意图和预先建立的知识库,确定所述面试题目对应的答案;及

根据所述对应的答案提取第一预设个数关键词。

3.如权利要求1所述的近义词推送方法,其特征在于,所述基于所述第二预设个数的目标词向量和词-索引文件推送第二预设个数近义词供用户选择包括:

获取所述第二预设个数的目标词向量对应的目标索引;

根据所述词-索引文件查询与所述目标索引对应的词向量;

推送所述词向量对应的近义词供用户选择。

4.如权利要求1所述的近义词推送方法,其特征在于,所述从所述二叉树中查询出与所述关键词的距离大于预设距离阈值的第一候选词向量并基于所述第一候选词向量构建优先队列包括:

以所述关键词作为所述二叉树的根节点;

遍历所述根节点下的所有中间节点;

计算所述根节点与每一个中间节点之间的距离；

确定大于预设距离阈值的目标距离对应的中间节点为第一层目标节点；

遍历所述第一层目标节点下的所有中间节点直至最后一层叶子节点；

将所有叶子节点中的词向量作为第一候选词向量；

计算所述第一候选词向量与所述关键词之间的相似度；

根据相似度的大小顺序将所述第一候选词向量插入优先队列中。

5.如权利要求1所述的近义词推送方法，其特征在于：根据预设规则筛选查找到的第二预设个数近义词，其中，所述预设规则包括一下规则中的至少一种：

根据词语字数调整查询到的第二预设个数近义词的顺序；

按词汇的类型来筛选查询到的第二预设个数近义词；

去除所述第二预设个数近义词中字数多于所述关键词的字数预设个数的词汇。

6.一种近义词推送装置，其特征在于，所述装置包括：

获取模块，用于获取面试题目；

配置模块，用于配置第一预设个数与所述面试题目对应的答案的关键词；

训练模块，用于基于超大词向量模型预先训练得到目标词向量模型，包括：扩充所述超大词向量模型中的机器人面试场景语料，其中，包括对所述机器人面试场景语料进行分词、去停用词及基于CBOW模式增量训练词向量操作；根据扩充语料后的超大词向量模型训练得到目标词向量模型；

构建模块，用于根据所述目标词向量模型构建词向量矩阵得到词-索引文件，包括：以每个词的维度为行数，以所述目标词向量模型中所有词的总数为列数构建词向量矩阵；所述词向量矩阵中的每一行对应一个索引；根据所述词向量矩阵构建词-索引文件，并输出所述词-索引文件，其中，所述词-索引文件包括词向量与索引之间的对应关系；

所述构建模块，还用于基于所述目标词向量模型中的所有词向量构建二叉树，包括：(1)随机选择两个点为初始节点，连接两个初始节点形成一个等距超平面；(2)根据所述两个初始节点的连线的中点垂直线构建一个等距垂直超平面，将所述目标词向量模型中的所有词向量对应的数据空间分成两部分，并得到两个子空间；(3)分别将每个子空间中的每个点与等距超平面的法向量相乘，求出每个点与法向量夹角的正负，以正负来分出其属于二叉树的左子树还是右子树；依此类推，分别在所述两个子空间内重复上述(1)至(3)，可以将所述数据空间切分为多个子空间，并根据所述多个子空间构建所述二叉树；

遍历模块，用于遍历所述二叉树，从所述二叉树中查询出与所述关键词的距离大于预设距离阈值的第一候选词向量并基于所述第一候选词向量构建优先队列；

去重模块，用于对所述优先队列中的所述第一候选词向量进行去重；

所述获取模块，还用于获取去重后的优先队列中排序在前第二预设个数的目标词向量；及

推送模块，用于基于所述第二预设个数的目标词向量和词-索引文件推送第二预设个数近义词供用户选择。

7.一种电子设备，其特征在于，所述电子设备包括处理器和存储器，所述处理器用于执行所述存储器中存储的计算机程序时实现如权利要求1至5中任意一项所述的近义词推送方法。

8.一种计算机可读存储介质,所述计算机可读存储介质上存储有计算机程序,其特征在于,所述计算机程序被处理器执行时实现如权利要求1至5中任意一项所述的近义词推送方法。

近义词推送方法、装置、电子设备及介质

技术领域

[0001] 本发明涉及计算机技术领域,具体涉及一种近义词推送方法、装置、电子设备及存储介质。

背景技术

[0002] 项目需求为人工智能(AI)面试规则配置系统中,部分公司的用户可实时更新专家规则中的回答关键词。但用户在填写回答关键词时,需要手动、纯人力的输入大量信息,系统无法对用户输入关键词时提供帮助,如近义词的推荐等。这种操作降低了用户的编写效率,也极度依赖用户对回答关键词的个人理解,无法保证用户输入的关键词是否较为全量且客观。

发明内容

[0003] 鉴于以上内容,有必要提出一种近义词推送方法、装置、电子设备及存储介质,可以为用户在进行AI面试时提供快速地近义词推送。

[0004] 本发明的第一方面提供一种近义词推送方法,所述方法包括:

[0005] 获取面试题目;

[0006] 配置第一预设个数与所述面试题目对应的答案的关键词;

[0007] 基于超大词向量模型预先训练得到目标词向量模型;

[0008] 根据所述目标词向量模型构建词向量矩阵得到词-索引文件,其中,所述词-索引文件包括词向量与索引之间的对应关系;

[0009] 基于所述目标词向量模型中的所有词向量构建二叉树;

[0010] 遍历所述二叉树,从所述二叉树中查询出与所述关键词的距离大于预设距离阈值的第一候选词向量并基于所述第一候选词向量构建优先队列;

[0011] 对所述优先队列中的所述第一候选词向量进行去重;

[0012] 获取去重后的优先队列中排序在前第二预设个数的目标词向量;

[0013] 基于所述第二预设个数的目标词向量和词-索引文件推送第二预设个数近义词供用户选择。

[0014] 优选地,所述配置第一预设个数与所述面试题目对应的答案的关键词的步骤包括:

[0015] 根据预先构建的题目解析模型分析所述面试题目得到对应的题目意图;

[0016] 根据所述题目意图和预先建立的知识库,确定所述面试题目对应的答案;及

[0017] 根据所述对应的答案提取第一预设个数关键词。

[0018] 优选地,所述基于超大词向量模型预先训练得到目标词向量模型的步骤包括:

[0019] 扩充所述超大词向量模型中的机器人面试场景语料,其中,包括对所述机器人面试场景语料进行分词、去停用词及基于CBOW模式增量训练词向量操作;

[0020] 根据扩充语料后的超大词向量模型训练得到目标词向量模型。

[0021] 优选地,根据所述目标词向量模型构建词向量矩阵得到词-索引文件的步骤包括:

[0022] 以每个词的维度为行数,以所述目标词向量模型中所有词的总数为列数构建词向量矩阵;

[0023] 所述词向量矩阵中的每一行对应一个索引;

[0024] 根据所述词向量矩阵构建词-索引文件,并输出所述词-索引文件。

[0025] 优选地,所述基于所述第二预设个数的目标词向量和词-索引文件推送第二预设个数近义词供用户选择包括:

[0026] 获取所述第二预设个数的目标词向量对应的目标索引;

[0027] 根据所述词-索引文件查询与所述目标索引对应的词向量;

[0028] 推送所述词向量对应的近义词供用户选择。

[0029] 优选地,所述从所述二叉树中查询出与所述关键词的距离大于预设距离阈值的第一候选词向量并基于所述第一候选词向量构建优先队列包括:

[0030] 以所述关键词作为所述二叉树的根节点;

[0031] 遍历所述根节点下的所有中间节点;

[0032] 计算所述根节点与每一个中间节点之间的距离;

[0033] 确定大于预设距离阈值的目标距离对应的中间节点为第一层目标节点;

[0034] 遍历所述第一层目标节点下的所有中间节点直至最后一层叶子节点;

[0035] 将所有叶子节点中的词向量作为第一候选词向量;

[0036] 计算所述第一候选词向量与所述关键词之间的相似度;

[0037] 根据相似度的大小顺序将所述第一候选词向量插入优先队列中。

[0038] 优选地,根据预设规则筛选查找到的第二预设个数近义词,其中,所述预设规则包括一下规则中的至少一种:

[0039] 根据词语字数调整查询到的第二预设个数近义词的顺序;

[0040] 按词汇的类型来筛选查询到的第二预设个数近义词;

[0041] 去除所述第二预设个数近义词中字数多于所述关键字的字数预设个数的词汇。

[0042] 本发明的第二方面提供一种近义词推送装置,所述装置包括:

[0043] 获取模块,用于获取面试题目;

[0044] 配置模块,用于配置第一预设个数与所述面试题目对应的答案的关键词;

[0045] 训练模块,用于基于超大词向量模型预先训练得到目标词向量模型;

[0046] 构建模块,用于根据所述目标词向量模型构建词向量矩阵得到词-索引文件,其中,所述词-索引文件包括词向量与索引之间的对应关系;

[0047] 所述构建模块,还用于基于所述目标词向量模型中的所有词向量构建二叉树;

[0048] 遍历模块,用于遍历所述二叉树,从所述二叉树中查询出与所述关键词的距离大于预设距离阈值的第一候选词向量并基于所述第一候选词向量构建优先队列;

[0049] 去重模块,用于对所述优先队列中的所述第一候选词向量进行去重;

[0050] 所述获取模块,还用于获取去重后的优先队列中排序在前第二预设个数的目标词向量;及

[0051] 推送模块,用于基于所述第二预设个数的目标词向量和词-索引文件推送第二预设个数近义词供用户选择。

[0052] 本发明的第三方面提供一种电子设备,所述电子设备包括处理器和存储器,所述处理器用于执行所述存储器中存储的计算机程序时实现所述近义词推送方法。

[0053] 本发明的第四方面提供一种计算机可读存储介质,所述计算机可读存储介质上存储有计算机程序,所述计算机程序被处理器执行时实现所述近义词推送方法。

[0054] 本发明所述的近义词推送方法、装置、电子设备及存储介质。通过配置第一预设个数与所述面试题目对应的答案的关键词,在预先训练的词向量模型中查找与每个关键词对应的第二预设个数近义词,推送所述第二预设个数近义词供用户选择。可以为机器人面试过程中配置更多面试题目对应的答案的关键字的近义词。方便人力资源HR在对求职者进行面试时,为面试题目配置更加全面的答案。从而在接收到求职者针对面试题目的答案时,可以更准确地分析求职者的答案,方便人力资源给出对求职者的更全面的分析。

附图说明

[0055] 为了更清楚地说明本发明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,下面描述中的附图仅仅是本发明的实施例,对于本领域普通技术人员来讲,在不付出创造性劳动的前提下,还可以根据提供的附图获得其他的附图。

[0056] 图1是本发明实施例一提供的近义词推送方法的流程图。

[0057] 图2是本发明实施例二提供的推送装置的功能模块图。

[0058] 图3是本发明实施例三提供的电子设备的示意图。

[0059] 如下具体实施方式将结合上述附图进一步说明本发明。

具体实施方式

[0060] 为了能够更清楚地理解本发明的上述目的、特征和优点,下面结合附图和具体实施例对本发明进行详细描述。需要说明的是,在不冲突的情况下,本发明的实施例及实施例中的特征可以相互组合。

[0061] 在下面的描述中阐述了很多具体细节以便于充分理解本发明,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做出创造性劳动前提下所获得的所有其他实施例,都属于本发明保护的范围。

[0062] 除非另有定义,本文所使用的所有的技术和科学术语与属于本发明的技术领域的技术人员通常理解的含义相同。本文中在本发明的说明书中所使用的术语只是为了描述具体的实施例的目的,不是旨在于限制本发明。

[0063] 本发明的说明书和权利要求书及上述附图中的术语“第一”、“第二”和“第三”等是用于区别不同对象,而非用于描述特定顺序。此外,术语“包括”以及它们任何变形,意图在于覆盖不排他的包含。例如包含了一系列步骤或单元的过程、方法、系统、产品或设备没有限定于已列出的步骤或单元,而是可选地还包括没有列出的步骤或单元,或可选地还包括对于这些过程、方法、产品或设备固有的其它步骤或单元。

[0064] 本发明实施例的近义词推送方法应用在电子设备中。所述对于需要进近义词推送的电子设备,可以直接在电子设备上集成本发明的方法所提供的近义词推送功能,或者安装用于实现本发明的方法的客户端。再如,本发明所提供的方法还可以以软件开发工具包

(Software Development Kit,SDK)的形式运行在服务器等设备上,以SDK的形式提供近义词推送功能的接口,电子设备或其他设备通过提供的接口即可实现近义词推送功能。

[0065] 实施例一

[0066] 图1是本发明实施例一提供的近义词推送方法的流程图。根据不同的需求,所述流程图中的执行顺序可以改变,某些步骤可以省略。

[0067] 为了在机器人面试过程中,通过机器人更好的判定求职者在回答面试过程中的面试题目是否正确,并根据回答结果给求职者评分时。需要根据所述面试题目对应的答案配置关键词,并在接收到求职者输入的答案后,根据输入的答案提取关键词。将提取的关键词与配置的关键词进行匹配,得到匹配结果,根据匹配结果对所述求职者进行评分。而在根据所述面试题目对应的答案配置关键词时,为了避免关键词不够全面的情况出现,本申请提供了一种在配置关键词时,对面试官输入的关键词进行拓展,推送同近/义词的方法。所述方法包括:

[0068] 步骤S1,获取面试题目。

[0069] 在机器人面试过程中,会根据不同的岗位配置不同的面试题目。例如,根据研发岗位配置的面试题目包括“你熟悉哪些编程语言”、“在Java中,如何跳出当前的多重嵌套循环”和“Java中会存在内存泄漏吗,请简单描述”等等。

[0070] 在本实施方式中,机器人面试需要预先配置好面试题目和答案。然而,不同的求职者面对同样的面试题目时给出的答案也不相同。为了全面的评判求职者的能力,在配置面试题目和答案时需要根据所述面试题目配置详尽完整且全面的答案。

[0071] 步骤S2,配置第一预设个数与所述面试题目对应的答案的关键词。

[0072] 在一实施方式中,所述配置第一预设个数与所述面试题目对应的答案的关键词的步骤包括:

[0073] 在预先建立的面试题目与答案对应表中查询所述面试题目配置对应的答案,得到查询结果;

[0074] 提取所述查询结果中的关键词,其中,所述关键词为第一预设个数。

[0075] 可以理解的是,所述关键词还可以是根据所述查询结果进行语义分析得到的与所述查询结果相关的关键词。

[0076] 在另一实施方式中,所述配置第一预设个数与所述面试题目对应的答案的关键词的步骤包括:

[0077] (1)根据预先构建的题目解析模型分析所述面试题目得到对应的题目意图。

[0078] 在本实施方式中,所述题目解析模型可以对所述面试题目的题目特征进行分析。所述题目特征可以包括题干意图和关键信息。例如,当面试题目为“你所擅长的编程语言有哪些”,那么题干意图是擅长的编程语言,关键信息可以是编程语言。

[0079] (2)根据所述题目意图和预先建立的知识库,确定所述面试题目对应的答案。

[0080] 例如,当面试题目为“你所擅长的编程语言有哪些”,那么所述预先建立的知识库中可能包括C/C++、Java、C#和SQL等。

[0081] (3)根据所述对应的答案提取第一预设个数关键词。

[0082] 步骤S3,基于超大词向量模型预先训练得到目标词向量模型。

[0083] 在本实施方式中,基于超大词向量模型进行预先训练得到合适的目标词向量模

型。具体包括:扩充所述超大词向量模型中的机器人面试场景语料,其中,包括对所述机器人面试场景语料进行分词、去停用词及基于CBOW模式增量训练词向量操作;根据扩充语料后的超大词向量模型训练得到目标词向量模型。

[0084] 具体地,所述超大词向量模型的训练语料涵盖了大量新闻、网页、小说、百度百科、维基百科等不同维度的语料。而针对机器人面试场景,超大词向量模型中的特定化场景的语料不足。因此,在超大词向量模型基础上融合机器人面试场景的语料,扩充机器人面试中的问答文本、相似问题文本等语料。所述目标词向量模型为包含了机器人面试预料的词向量模型。

[0085] 再先对机器人面试场景语料进行分词、去停用词、基于CBOW模式增量训练词向量等操作,以扩充它在机器人面试场景下的性能表现。最终训练好的目标词向量模型涵盖了超过800万个词,每个词的维度约有200维。从而使所述目标词向量模型语料广泛,并且其中的每个词向量都能很好地反应出每个词的语义。同时800万个词的数量级能够完全顶替传统的构建近义词词典的方式,很好地解决找不到词的问题。

[0086] 需要说明的是,基于超大词向量模型预先训练得到目标词向量模型的方法为现有技术,在此不再赘述。

[0087] 步骤S4,根据所述目标词向量模型构建词向量矩阵得到词-索引文件,其中,所述词-索引文件包括词向量与索引之间的对应关系。

[0088] 在本实施方式中,所述根据所述目标词向量模型构建词向量矩阵得到词-索引文件可以包括:

[0089] (a1)以每个词的维度为行数,以所述目标词向量模型中所有词的总数为列数构建一个词向量矩阵;

[0090] (a2)所述词向量矩阵中的每一行对应一个索引;

[0091] (a3)根据所述词向量矩阵构建词-索引文件,并输出所述词-索引文件。

[0092] 具体地,所述词向量矩阵以每个词的维度为行数,以所有词的总数为列数组成的一个矩阵。在本实施方式中,每个词的维度为200,所述目标词向量模型包括800万个词。那么,可以得到一个200列,800万行的词向量矩阵。

[0093] 而所述词向量矩阵中的每一行都有一个索引,那么,可以得到每个词对应的索引。从而根据所述词向量矩阵输出词-索引文件。同时,也可以得到每个索引与每个词向量之间的对应关系。

[0094] 步骤S5,基于所述目标词向量模型中的所有词向量构建二叉树。

[0095] 在本实施方式中,根据所述目词向量模型中的所有词向量构建二叉树结构。

[0096] 所述词向量是一个200维的向量,即是200维的高维数据空间,每个词向量在高维数据空间表示一个点,所述目标词向量模型中的所有词向量对应的数据空间可以表示为800万个点。通过以下方法根据所述目标词向量模型构建二叉树:

[0097] (1)随机选择两个点为初始节点,连接两个初始节点形成一个等距超平面;

[0098] (2)根据所述两个初始节点的连线的中点垂直线构建一个等距垂直超平面,将所述目标词向量模型中的所有词向量对应的数据空间分成两部分,并得到两个子空间;

[0099] (3)分别将每个子空间中的每个点与等距超平面的法向量相乘(向量点积),求出每个点与法向量夹角的正负,以正负来分出其属于二叉树的左子树还是右子树;

[0100] (4)依此类推,分别在所述两个子空间内重复上述步骤(1)至(3),可以将所述数据空间切分为多个子空间,并根据所述多个子空间构建二叉树结构。

[0101] 优选地,当每个子空间最多只剩下k个点时,不再对所述子空间进行切分。优选地,所述k大于等于8且小于等于10。在本实施方式中,所述k的取值为10。

[0102] 上述的二叉树结构中的每个节点的分割条件就是那些等距垂直超平面,最终,词向量即为二叉树上的叶子节点。即,所述二叉树包括根节点及多层中间节点和最后一层叶子节点,其中,每一个叶子节点代表一个词向量。在本申请中,无需在所述叶子节点上保存词向量,只需要保存词向量对应的索引即可。如此,相似的词向量在二叉树上的位置更近,为后续查询近义词提供了更快的速度。

[0103] 步骤S6,遍历所述二叉树,从所述二叉树中查询出与所述关键词的距离大于预设距离阈值的第一候选词向量并基于所述第一候选词向量构建优先队列。

[0104] 具体的构建优先队列的方法为:以所述关键词作为所述二叉树的根节点;遍历所述根节点下的所有中间节点;计算所述根节点与每一个中间节点之间的距离;确定大于预设距离阈值的目标距离对应的中间节点为第一层目标节点;遍历所述第一层目标节点下的所有中间节点直至最后一层叶子节点;将所有叶子节点中的词向量作为第一候选词向量;计算所述第一候选词向量与所述关键词之间的相似度;根据相似度的大小顺序将所述第一候选词向量插入优先队列中。

[0105] 步骤S7,对所述优先队列中的所述第一候选词向量进行去重。

[0106] 步骤S8,获取去重后的优先队列中排序在前第二预设个数的目标词向量。

[0107] 步骤S9,基于所述第二预设个数的目标词向量和词-索引文件推送第二预设个数近义词供用户选择。

[0108] 在本实施方式中,基于所述第二预设个数的目标词向量和词-索引文件推送第二预设个数近义词供用户选择的方法包括:获取所述第二预设个数的目标词向量对应的目标索引;根据所述词-索引文件查询与所述目标索引对应的词向量;推送所述词向量对应的近义词供用户选择。

[0109] 在本实施方式中,将二叉树结构文件和词-索引文件一起保存,需要进行查询某个关键词的相近邻Top N的词汇的时候,只需要利用这两个文件进行索引即可。

[0110] 在本实施方式中,通过推送第二预设个数近义词供用户筛选,可以方便用户更全面的配置面试题目对应的答案的关键词。从而在求职者答题时,不会片面的根据求职者的答案给求职者评分。本申请支持的近义词查找功能更加创新和便利,一次性可生成5个关键词的近义词,并且每次推送8个近义词,支持用户点击“换一批”更换另外一轮的8个近义词,便于用户查看和使用。例如,在推送界面显示一个“换一批”按钮,在用户点击所述按钮后,可以更新原来的近义词,推送更多的近义词。

[0111] 优选地,由于很多词语并不是所述面试题目的答案,所以增加了预设规则筛选查询到的词汇,其中,所述预设规则包括一下规则中的至少一种:

[0112] (1)根据词语字数调整查询到的词汇的顺序。例如,优先返回与所述关键词字数一致的词汇。而对于与所述关键词字数不一致的词汇,每增加/减少1个字,则在将查询到的词汇进行排序时增加预设距离(如0.1)。

[0113] (2)按词汇的类型来筛选查询到的词汇,所述类型包括中文,英文和数字。例如,优

先返回与所述关键词类型一致的词汇。另外对于输入中文,返回英文;或输入英文,返回中文的情况,正常返回。但对于输入中文,返回数字;或输入英文,返回数字的情况,则直接删除该近义词。需要说明的是,单个英文字母或单个中文字代表1个字符。

[0114] (3)去除字数多于所述关键字的字数预设个数的词汇。例如,多于所述关键词的字数5个字以上的词汇。

[0115] 可以理解的是,上述方法同样可以用于推送同义词。

[0116] 综上所述,本发明提供的近义词推送方法包括,获取面试题目;配置第一预设个数与所述面试题目对应的答案的关键词;在预先训练的词向量模型中查找与每个关键词对应的第二预设个数近义词;及推送所述第二预设个数近义词供用户选择。本申请采用的词向量涵盖面广泛,表征词的向量维度200维,每个词的向量都能很好地反应出每个词的实际语义;本申请的词向量模型包括800万个词,很好地解决传统的匹配不到词汇(out-of-word)的问题。本申请采用的词向量模型内存占用大大减少,通过采样词-索引文件极大地降低的内存占用率,并且极大增加系统稳定性。另外,本申请查询返回的速度大大增加,原来一个词需要十几秒左右的查询时间,现在降低到0.01s以内返回。最后本申请可以为机器人面试过程中配置更多面试题目对应的答案的关键字的近义词。从而在接收到求职者针对面试题目的答案时,可以更准确地分析求职者的答案,方便人力资源给出对求职者的更全面的分析。

[0117] 以上所述,仅是本发明的具体实施方式,但本发明的保护范围并不局限于此,对于本领域的普通技术人员来说,在不脱离本发明创造构思的前提下,还可以做出改进,但这些均属于本发明的保护范围。

[0118] 下面结合图2和图3,分别对实现上述近义词推送方法的电子设备的功能模块及硬件结构进行介绍。

[0119] 实施例二

[0120] 图2为本发明近义词推送装置较佳实施例中的功能模块图。

[0121] 在一些实施例中,所述近义词推送装置20(为便于描述,简称为“推送装置”)运行于电子设备中。所述推送装置20可以包括多个由程序代码段所组成的功能模块。所述推送装置20中的各个程序段的程序代码可以存储于存储器中,并由至少一个处理器所执行,以执行近义词推送功能。

[0122] 为了在机器人面试过程中,通过机器人更好的判定求职者在回答面试过程中的面试题目是否正确,并根据回答结果给求职者评分时。需要根据所述面试题目对应的答案配置关键词,并在接收到求职者输入的答案后,根据输入的答案提取关键词。将提取的关键词与配置的关键词进行匹配,得到匹配结果,根据匹配结果对所述求职者进行评分。而在根据所述面试题目对应的答案配置关键词时,为了避免关键词不够全面的情况出现,本申请提供了一种在配置关键词时,对面试官输入的关键词进行拓展,推送同近/义词的所述推送装置20。所述推送装置20的功能模块可以包括:获取模块201、配置模块202、训练模块203、构建模块204、遍历模块205、去重模块206及推送模块207。关于各模块的功能将在后续的实施例中详述。本发明所称的模块是指一种能够被至少一个处理器所执行并且能够完成固定功能的一系列计算机程序段,其存储在存储器中。

[0123] 所述获取模块201用于获取面试题目。

[0124] 在机器人面试过程中,会根据不同的岗位配置不同的面试题目。例如,根据研发岗位配置的面试题目包括“你熟悉哪些编程语言”、“在Java中,如何跳出当前的多重嵌套循环”和“Java中会存在内存泄漏吗,请简单描述”等等。

[0125] 在本实施方式中,机器人面试需要预先配置好面试题目和答案。然而,不同的求职者面对同样的面试题目时给出的答案也不相同。为了全面的评判求职者的能力,在配置面试题目和答案时需要根据所述面试题目配置详尽完整且全面的答案。

[0126] 所述配置模块202用于配置第一预设个数与所述面试题目对应的答案的关键词。

[0127] 在一实施方式中,所述配置第一预设个数与所述面试题目对应的答案的关键词包括:

[0128] 在预先建立的面试题目与答案对应表中查询所述面试题目配置对应的答案,得到查询结果;

[0129] 提取所述查询结果中的关键词,其中,所述关键词为第一预设个数。

[0130] 可以理解的是,所述关键词还可以是根据所述查询结果进行语义分析得到的与所述查询结果相关的关键词。

[0131] 在另一实施方式中,所述配置第一预设个数与所述面试题目对应的答案的关键词包括:

[0132] (1)根据预先构建的题目解析模型分析所述面试题目得到对应的题目意图。

[0133] 在本实施方式中,所述题目解析模型可以对所述面试题目的题目特征进行分析。所述题目特征可以包括题干意图和关键信息。例如,当面试题目为“你所擅长的编程语言有哪些”,那么题干意图是擅长的编程语言,关键信息可以是编程语言。

[0134] (2)根据所述题目意图和预先建立的知识库,确定所述面试题目对应的答案。

[0135] 例如,当面试题目为“你所擅长的编程语言有哪些”,那么所述预先建立的知识库中可能包括C/C++、Java、C#和SQL等。

[0136] (3)根据所述对应的答案提取第一预设个数关键词。

[0137] 所述训练模块203用于基于超大词向量模型预先训练得到目标词向量模型。

[0138] 在本实施方式中,基于超大词向量模型进行预先训练得到合适的目标词向量模型。具体包括:扩充所述超大词向量模型中的机器人面试场景语料,其中,包括对所述机器人面试场景语料进行分词、去停用词及基于CBOW模式增量训练词向量操作;根据扩充语料后的超大词向量模型训练得到目标词向量模型。

[0139] 具体地,所述超大词向量模型的训练语料涵盖了大量新闻、网页、小说、百度百科、维基百科等不同维度的语料。而针对机器人面试场景,超大词向量模型中的特定化场景的语料不足。因此,在超大词向量模型基础上融合机器人面试场景的语料,扩充机器人面试中的问答文本、相似问题文本等语料。所述目标词向量模型为包含了机器人面试预料的词向量模型。

[0140] 再先对机器人面试场景语料进行分词、去停用词、基于CBOW模式增量训练词向量等操作,以扩充它在机器人面试场景下的性能表现。最终训练好的目标词向量模型涵盖了超过800万个词,每个词的维度约有200维。从而使所述目标词向量模型语料广泛,并且其中的每个词向量都能很好地反应出每个词的语义。同时800万个词的数量级能够完全顶替传统的构建近义词词典的方式,很好地解决找不到词的问题。

[0141] 需要说明的是,基于超大词向量模型预先训练得到目标词向量模型的方法为现有技术,在此不再赘述。

[0142] 所述构建模块204用于根据所述目标词向量模型构建词向量矩阵得到词-索引文件,其中,所述词-索引文件包括词向量与索引之间的对应关系。

[0143] 在本实施方式中,所述根据所述目标词向量模型构建词向量矩阵得到词-索引文件可以包括:

[0144] (a1)以每个词的维度为行数,以所述目标词向量模型中所有词的总数为列数构建一个词向量矩阵;

[0145] (a2)所述词向量矩阵中的每一行对应一个索引;

[0146] (a3)根据所述词向量矩阵构建词-索引文件,并输出所述词-索引文件。

[0147] 具体地,所述词向量矩阵以每个词的维度为行数,以所有词的总数为列数组成的一个矩阵。在本实施方式中,每个词的维度为200,所述目标词向量模型包括800万个词。那么,可以得到一个200列,800万行的词向量矩阵。

[0148] 而所述词向量矩阵中的每一行都有一个索引,那么,可以得到每个词对应的索引。从而根据所述词向量矩阵输出词-索引文件。同时,也可以得到每个索引与每个词向量之间的对应关系。

[0149] 所述构建模块204还用于基于所述目标词向量模型中的所有词向量构建二叉树。

[0150] 在本实施方式中,将所述目标词向量模型中的所有词向量构建二叉树结构。

[0151] 所述词向量是一个200维的向量,即是200维的高维数据空间,每个词向量在高维数据空间表示一个点,所述目标词向量模型中的所有词向量对应的数据空间可以表示为800万个点。通过以下方法根据所述目标词向量模型构建二叉树:

[0152] (1)随机选择两个点为初始节点,连接两个初始节点形成一个等距超平面。

[0153] (2)根据所述两个初始节点的连线的中点垂直线构建一个等距垂直超平面,将所述目标词向量模型中的所有词向量对应的数据空间分成两部分,并得到两个子空间。

[0154] (3)分别将每个子空间中的每个点与等距超平面的法向量相乘(向量点积),求出每个点与法向量夹角的正负,以正负来分出其属于二叉树的左子树还是右子树。

[0155] (4)依此类推,分别在所述两个子空间内重复上述步骤(1)至(3),可以将所述数据空间切分为多个子空间,并根据所述多个子空间构建二叉树结构。

[0156] 优选地,当每个子空间最多只剩下k个点时,不再对所述子空间进行切分。优选地,所述k大于等于8且小于等于10。在本实施方式中,所述k的取值为10。

[0157] 上述的二叉树结构中的每个节点的分割条件就是那些等距垂直超平面,最终,词向量即为二叉树上的叶子节点。在本申请中,无需在所述叶子节点上保存词向量,只需要保存词向量对应的索引即可。如此,相似的词向量在二叉树上的位置更近,为后续查询近义词提供了更快的速度。

[0158] 所述遍历模块205用于遍历所述二叉树,从所述二叉树中查询出与所述关键词的距离大于预设距离阈值的第一候选词向量并基于所述第一候选词向量构建优先队列。

[0159] 具体的构建优先队列的方法为:以所述关键词作为所述二叉树的根节点;遍历所述根节点下的所有中间节点;计算所述根节点与每一个中间节点之间的距离;确定大于预设距离阈值的目标距离对应的中间节点为第一层目标节点;遍历所述第一层目标节点下的

所有中间节点直至最后一层叶子节点;将所有叶子节点中的词向量作为第一候选词向量;计算所述第一候选词向量与所述关键词之间的相似度;根据相似度的大小顺序将所述第一候选词向量插入优先队列中。

[0160] 所述去重模块206用于对所述优先队列中的所述第一候选词向量进行去重。

[0161] 所述获取模块201还用于获取去重后的优先队列中排序在前第二预设个数的目标词向量。

[0162] 所述推送模块207用于基于所述第二预设个数的目标词向量和词-索引文件推送第二预设个数近义词供用户选择。

[0163] 在本实施方式中,基于所述第二预设个数的目标词向量和词-索引文件推送第二预设个数近义词供用户选择的方法包括:获取所述第二预设个数的目标词向量对应的目标索引;根据所述词-索引文件查询与所述目标索引对应的词向量;推送所述词向量对应的近义词供用户选择。

[0164] 在本实施方式中,将二叉树结构文件和词-索引文件一起保存,需要进行查询某个关键词的相近邻Top N的词汇的时候,只需要利用这两个文件进行索引即可。

[0165] 在本实施方式中,通过推送第二预设个数近义词供用户筛选,可以方便用户更全面的配置面试题目对应的答案的关键词。从而在求职者答题时,不会片面的根据求职者的答案给求职者评分。本申请支持的近义词查找功能更加创新和便利,一次性可生成5个关键词的近义词,并且每次推送8个近义词,支持用户点击“换一批”更换另外一轮的8个近义词,便于用户查看和使用。例如,在推送界面显示一个“换一批”按钮,在用户点击所述按钮后,可以更新原来的近义词,推送更多的近义词。

[0166] 优选地,由于很多词语并不是所述面试题目的答案,所以增加了预设规则筛选查询到的词汇,其中,所述预设规则包括一下规则中的至少一种:

[0167] (1)根据词语字数调整查询到的词汇的顺序。例如,优先返回与所述关键词字数一致的词汇。而对于与所述关键词字数不一致的词汇,每增加/减少1个字,则在将查询到的词汇进行排序时增加预设距离(如0.1)。

[0168] (2)按词汇的类型来筛选查询到的词汇,所述类型包括中文,英文和数字。例如,优先返回与所述关键词类型一致的词汇。另外对于输入中文,返回英文;或输入英文,返回中文的情况,正常返回。但对于输入中文,返回数字;或输入英文,返回数字的情况,则直接删除该近义词。需要说明的是,单个英文字母或单个中文字代表1个字符。

[0169] (3)去除字数多于所述关键字的字数预设个数的词汇。例如,多于所述关键词的字数5个字以上的词汇。

[0170] 可以理解的是,上述推送装置20同样可以用于推送同义词。

[0171] 综上所述,本发明所述的推送装置20,包括获取模块201、配置模块202、训练模块203、构建模块204、遍历模块205、去重模块206及推送模块207。所述获取模块201用于获取面试题目;所述配置模块202用于配置第一预设个数与所述面试题目对应的答案的关键词;所述训练模块203用于基于超大词向量模型预先训练得到目标词向量模型;所述构建模块204用于根据所述目标词向量模型构建词向量矩阵得到词-索引文件,其中,所述词-索引文件包括词向量与索引之间的对应关系;所述构建模块204还用于基于所述目标词向量模型中的所有词向量构建二叉树;所述遍历模块205用于遍历所述二叉树,从所述二叉树中查询

出与所述关键词的距离大于预设距离阈值的第一候选词向量并基于所述第一候选词向量构建优先队列;所述去重模块206用于对所述优先队列中的所述第一候选词向量进行去重;所述获取模块201还用于获取去重后的优先队列中排序在前第二预设个数的目标词向量;及所述推送模块207用于基于所述第二预设个数的目标词向量和词-索引文件推送第二预设个数近义词供用户选择。

[0172] 本申请采用的词向量涵盖面广泛,表征词的向量维度200维,每个词的向量都能很好地反应出每个词的实际语义;本申请的词向量模型包括800万个词,很好地解决传统的匹配不到词汇(out-of-word)的问题。本申请采用的词向量模型内存占用大大减少,通过采样词-索引文件极大地降低的内存占用率,并且极大增加系统稳定性。另外,本申请查询返回的速度大大增加,原来一个词需要十几秒左右的查询时间,现在降低到0.01s以内返回。最后本申请可以为机器人面试过程中配置更多面试题目对应的答案的关键字的近义词。方便人力资源HR在对求职者进行面试时,为面试题目配置更加全面的答案。从而在接收到求职者针对面试题目的答案时,可以更准确地分析求职者的答案,方便人力资源给出对求职者的更全面的分析。

[0173] 上述以软件功能模块的形式实现的集成的单元,可以存储在一个计算机可读取存储介质中。上述软件功能模块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,双屏设备,或者网络设备等)或处理器(processor)执行本发明各个实施例所述方法的部分。

[0174] 图3为本发明实施例三提供的电子设备的示意图。

[0175] 所述电子设备3包括:存储器31、至少一个处理器32、存储在所述存储器31中并可在所述至少一个处理器32上运行的计算机程序33、至少一条通讯总线34及数据库35。

[0176] 所述至少一个处理器32执行所述计算机程序33时实现上述近义词推送方法实施例中的步骤。

[0177] 示例性的,所述计算机程序33可以被分割成一个或多个模块/单元,所述一个或者多个模块/单元被存储在所述存储器31中,并由所述至少一个处理器32执行,以完成本发明。所述一个或多个模块/单元可以是能够完成特定功能的一系列计算机程序指令段,所述指令段用于描述所述计算机程序33在所述电子设备3中的执行过程。

[0178] 所述电子设备3可以是手机、平板电脑、个人数字助理(Personal Digital Assistant,PDA)等安装有应用程序的设备。本领域技术人员可以理解,所述示意图3仅仅是电子设备3的示例,并不构成对电子设备3的限定,可以包括比图示更多或更少的部件,或者组合某些部件,或者不同的部件,例如所述电子设备3还可以包括输入输出设备、网络接入设备、总线等。

[0179] 所述至少一个处理器32可以是中央处理单元(Central Processing Unit,CPU),还可以是其他通用处理器、数字信号处理器(Digital Signal Processor,DSP)、专用集成电路(Application Specific Integrated Circuit,ASIC)、现成可编程门阵列(Field-Programmable Gate Array,FPGA)或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。所述处理器32可以是微处理器或者所述处理器32也可以是任何常规的处理器等,所述处理器32是所述电子设备3的控制中心,利用各种接口和线路连接整个电子设备3的各个部分。

[0180] 所述存储器31可用于存储所述计算机程序33和/或模块/单元,所述处理器32通过运行或执行存储在所述存储器31内的计算机程序和/或模块/单元,以及调用存储在存储器31内的数据,实现所述电子设备3的各种功能。所述存储器31可主要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、至少一个功能所需的应用程序(比如声音播放功能、图像播放功能等)等;存储数据区可存储根据电子设备3的使用所创建的数据(比如音频数据等)等。此外,存储器31可以包括高速随机存取存储器,还可以包括非易失性存储器,例如硬盘、内存、插接式硬盘,智能存储卡(Smart Media Card,SMC),安全数字(Secure Digital,SD)卡,闪存卡(Flash Card)、至少一个磁盘存储器件、闪存器件、或其他非易失性固态存储器件。

[0181] 所述存储器31中存储有程序代码,且所述至少一个处理器32可调用所述存储器31中存储的程序代码以执行相关的功能。例如,图2中所述的各个模块(获取模块201、配置模块202、训练模块203、构建模块204、遍历模块205、去重模块206及推送模块207)是存储在所述存储器31中的程序代码,并由所述至少一个处理器32所执行,从而实现所述各个模块的功能以达到近义词推送的目的。

[0182] 所述获取模块201用于获取面试题目;

[0183] 所述配置模块202用于配置第一预设个数与所述面试题目对应的答案的关键词;

[0184] 所述训练模块203用于基于超大词向量模型预先训练得到目标词向量模型;

[0185] 所述构建模块204用于根据所述目标词向量模型构建词向量矩阵得到词-索引文件,其中,所述词-索引文件包括词向量与索引之间的对应关系;

[0186] 所述构建模块204还用于基于所述目标词向量模型中的所有词向量构建二叉树;

[0187] 所述遍历模块205用于遍历所述二叉树,从所述二叉树中查询出与所述关键词的距离大于预设距离阈值的第一候选词向量并基于所述第一候选词向量构建优先队列;

[0188] 所述去重模块206用于对所述优先队列中的所述第一候选词向量进行去重;

[0189] 所述获取模块201还用于获取去重后的优先队列中排序在前第二预设个数的目标词向量;及

[0190] 所述推送模块207用于所述第二预设个数的目标词向量和词-索引文件推送第二预设个数近义词供用户选择。

[0191] 所述数据库(Database)35是按照数据结构来组织、存储和管理数据的建立在所述电子设备3上的仓库。数据库通常分为层次式数据库、网络式数据库和关系式数据库三种。在本实施方式中,所述数据库35用于存储面试题目等信息。

[0192] 所述电子设备3集成的模块/单元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时,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,本发明实现上述实施例方法中的全部或部分流程,也可以通过计算机程序来指令相关的硬件来完成,所述的计算机程序可存储于一计算机可读存储介质中,所述计算机程序在被处理器执行时,可实现上述各个方法实施例的步骤。其中,所述计算机程序包括计算机程序代码,所述计算机程序代码可以为源代码形式、对象代码形式、可执行文件或某些中间形式等。所述计算机可读介质可以包括:能够携带所述计算机程序代码的任何实体或装置、记录介质、U盘、移动硬盘、磁碟、光盘、计算机存储器、只读存储器(ROM,Read-Only Memory)。

[0193] 在本发明所提供的几个实施例中,应所述理解到,所揭露的电子设备和方法,可以

通过其它的方式实现。例如,以上所描述的电子设备实施例仅仅是示意性的,例如,所述单元的划分,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。

[0194] 另外,在本发明各个实施例中的各功能单元可以集成在相同处理单元中,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相同单元中。上述集成的单元既可以采用硬件的形式实现,也可以采用硬件加软件功能模块的形式实现。

[0195] 对于本领域技术人员而言,显然本发明不限于上述示范性实施例的细节,而且在不背离本发明的精神或基本特征的情况下,能够以其他的具体形式实现本发明。因此,无论从哪一点来看,均应将实施例看作是示范性的,而且是非限制性的,本发明的范围由所附权利要求而不是上述说明限定,因此旨在将落在权利要求的等同要件的含义和范围内的所有变化涵括在本发明内。不应将权利要求中的任何附图标记视为限制所涉及的权利要求。此外,显然“包括”一词不排除其他单元或,单数不排除复数。系统权利要求中陈述的多个单元或装置也可以由一个单元或装置通过软件或者硬件来实现。第一,第二等词语用来表示名称,而并不表示任何特定的顺序。

[0196] 最后应说明的是,以上实施例仅用以说明本发明的技术方案而非限制,尽管参照较佳实施例对本发明进行了详细说明,本领域的普通技术人员应当理解,可以对本发明的技术方案进行修改或等同替换,而不脱离本发明技术方案的精神范围。

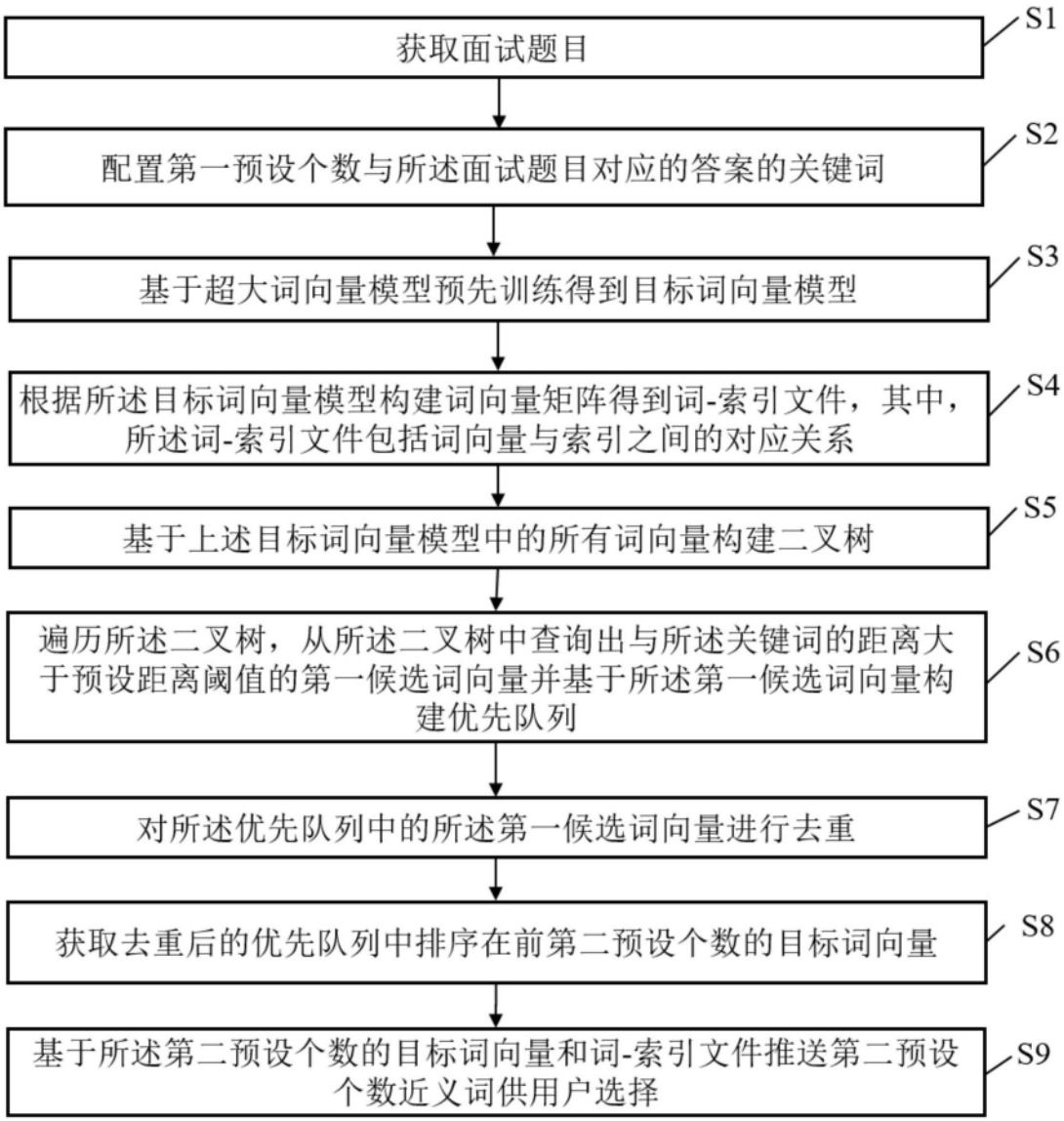

获取面试题目
S1
配置第一预设个数与所述面试题目对应的答案的关键词
S2
基于超大词向量模型预先训练得到目标词向量模型
S3
根据所述目标词向量模型构建词向量矩阵得到词-索引文件，其中，所述词-索引文件包括词向量与索引之间的对应关系
S4
基于上述目标词向量模型中的所有词向量构建二叉树
S5
遍历所述二叉树，从所述二叉树中查询出与所述关键词的距离大于预设距离阈值的第一候选词向量并基于所述第一候选词向量构建优先队列
S6
对所述优先队列中的所述第一候选词向量进行去重
S7
获取去重后的优先队列中排序在前第二预设个数的目标词向量
S8
基于所述第二预设个数的目标词向量和词-索引文件推送第二预设个数近义词供用户选择
S9

图1


推送装置 20
获取模块 201
配置模块 202
训练模块 203
构建模块 204
遍历模块 205
去重模块 206
推送模块 207

图2

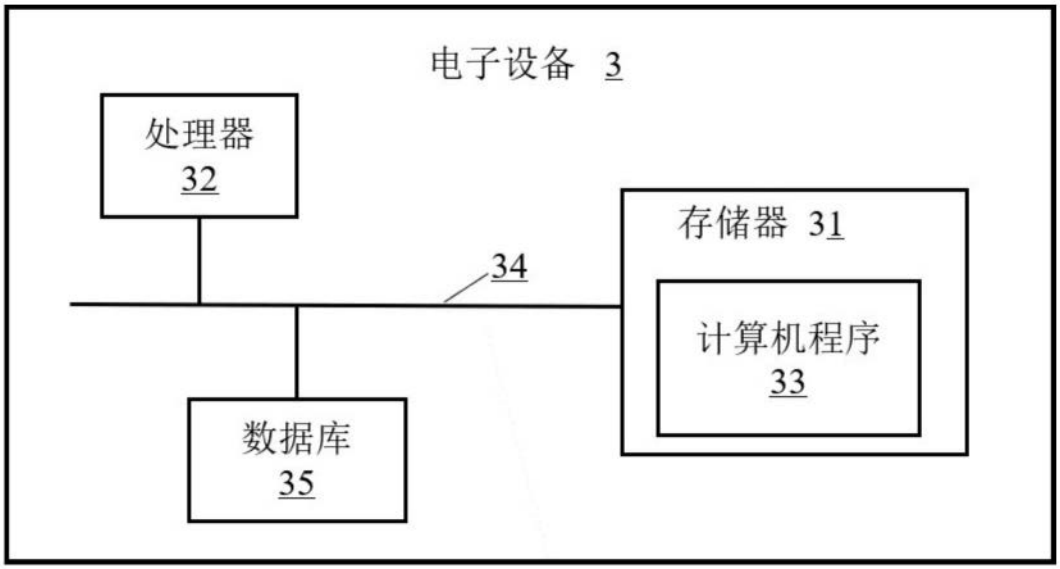


图3