(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 114238564 B
(45)授权公告日 2025.05.13

(21)申请号 202111495205.8
(22)申请日 2021.12.09
(65)同一申请的已公布的文献号
申请公布号 CN 114238564 A
(43)申请公布日 2022.03.25
(73)专利权人 阳光保险集团股份有限公司
地址 518000 广东省深圳市福田区红荔西路7002号第一世界广场A座17层
(72)发明人 韩佳 杜新凯 吕超 谷姗姗 张晗 李文灏
(74)专利代理机构 北京超凡宏宇知识产权代理有限公司 11463
专利代理师 刘凤
(51)Int.Cl.
G06F 16/3331(2025.01)
G06F 40/216(2020.01)
G06F 40/284(2020.01)
G06F 18/214(2023.01)
(56)对比文件
CN 108520033 A,2018.09.11
CN 109522392 A,2019.03.26
审查员 倪赛华

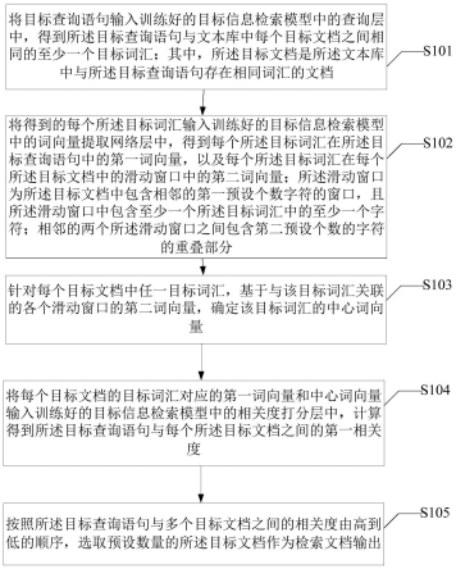
权利要求书3页 说明书11页 附图4页

(54)发明名称

一种信息检索方法、装置、电子设备及存储介质

(57)摘要

本申请提供了一种信息检索方法、装置、电子设备及存储介质,所述信息检索方法基于训练好的目标信息检索模型,确定目标词汇在目标查询语句的第一词向量,以及确定目标词汇在每个目标文档的中的中心词向量,根据第一词向量与中心词向量确定查询语句与目标文档的第一相关度,由于本申请中目标词汇的中心词向量是基于每个目标文档中的多个滑动窗口确定的,使得目标词汇对应的目标查询语句与得到的检索文档的语义和词义更加匹配,提高了信息检索的准确度。



CN 114238564 B

1.一种信息检索方法,其特征在于,所述信息检索方法包括:

将目标查询语句输入训练好的目标信息检索模型中的查询层中,得到所述目标查询语句与文本库中每个目标文档之间相同的至少一个目标词汇;其中,所述目标文档是所述文本库中与所述目标查询语句存在相同词汇的文档;

将得到的每个所述目标词汇输入训练好的目标信息检索模型中的词向量提取网络层中,得到每个所述目标词汇在所述目标查询语句中的第一词向量,以及每个所述目标词汇在每个所述目标文档中的滑动窗口中的第二词向量;所述滑动窗口为所述目标文档中包含相邻的第一预设个数字符的窗口,且所述滑动窗口中包含至少一个所述目标词汇中的至少一个字符;相邻的两个所述滑动窗口之间包含第二预设个数的字符的重叠部分;

针对每个目标文档中任一目标词汇,基于与该目标词汇关联的各个滑动窗口的第二词向量,确定该目标词汇的中心词向量;

将每个目标文档的目标词汇对应的第一词向量和中心词向量输入训练好的目标信息检索模型中的相关度打分层中,计算得到所述目标查询语句与每个所述目标文档之间的第一相关度;

按照所述目标查询语句与多个目标文档之间的相关度由高到低的顺序,选取预设数量的所述目标文档作为检索文档输出;

其中,通过以下方式确定训练好的目标信息检索模型:

获取样本查询语句和与所述样本查询语句对应的样本文档;

将所述样本查询语句与所述样本文档按照预设相关度进行相关度划分,确定所述相关度大于等于所述预设相关度的所述样本查询语句与所述样本文档为样本相关文本,以及确定所述相关度小于所述预设相关度的所述样本查询语句与所述样本文档为样本不相关文本;

根据所述样本相关文本以及所述样本不相关文本,对初始信息检索模型进行训练,确定训练好的目标信息检索模型。

2.根据权利要求1所述的信息检索方法,其特征在于,所述针对每个目标文档中任一目标词汇,基于与该目标词汇关联的各个滑动窗口的第二词向量,确定该目标词汇的中心词向量,包括:

针对每个目标文档中任一目标词汇,获取该目标词汇在各个滑动窗口的第二词向量;

针对各个所述第二词向量进行求和计算,并在求和计算后进行平均值计算,将所述平均值确定为该目标词汇的中心词向量。

3.根据权利要求1所述的信息检索方法,其特征在于,通过以下方式计算得到所述目标查询语句与每个所述目标文档之间的第一相关度:

针对每个目标词汇对应的第一词向量与每个所述目标词汇对应的每个中心词向量进行点乘计算,并将点乘结果确定为每个所述目标词汇与每个所述目标文档之间的第二相关度;

针对每个目标文档中的各个所述目标词汇对应的各个所述第二相关度进行求和计算,将求和结果确定为目标查询语句与每个所述目标文档之间的第一相关度。

4.根据权利要求3所述的信息检索方法,其特征在于,所述针对每个目标词汇对应的第一词向量与每个所述目标词汇对应的每个中心词向量进行点乘计算,并将点乘结果确定为

每个所述目标词汇与每个所述目标文档之间的第二相关度,包括:

针对每个目标词汇对应的第一词向量与每个所述目标词汇对应的每个中心词向量进行点乘计算,选取最大点乘值为结果相关度;

将所述结果相关度确定为每个所述目标词汇与每个所述目标文档之间的第二相关度。

5.一种信息检索装置,其特征在于,所述信息检索装置包括:

第一确定模块,用于将目标查询语句输入训练好的目标信息检索模型中的查询层中,得到所述目标查询语句与文本库中每个目标文档之间相同的至少一个目标词汇;其中,所述目标文档是所述文本库中与所述目标查询语句存在相同词汇的文档;其中,通过以下方式确定训练好的目标信息检索模型:

获取样本查询语句和与所述样本查询语句对应的样本文档;

将所述样本查询语句与所述样本文档按照预设相关度进行相关度划分,确定所述相关度大于等于所述预设相关度的所述样本查询语句与所述样本文档为样本相关文本,以及确定所述相关度小于所述预设相关度的所述样本查询语句与所述样本文档为样本不相关文本;

根据所述样本相关文本以及所述样本不相关文本,对初始信息检索模型进行训练,确定训练好的目标信息检索模型;

第二确定模块,用于将得到的每个所述目标词汇输入训练好的目标信息检索模型中的词向量提取网络层中,得到每个所述目标词汇在所述目标查询语句中的第一词向量,以及每个所述目标词汇在每个所述目标文档中的滑动窗口中的第二词向量;所述滑动窗口为所述目标文档中包含相邻的第一预设个数字符的窗口,且所述滑动窗口中包含至少一个所述目标词汇中的至少一个字符;相邻的两个所述滑动窗口之间包含第二预设个数的字符的重叠部分;

第三确定模块,用于针对每个目标文档中任一目标词汇,基于与该目标词汇关联的各个滑动窗口的第二词向量,确定该目标词汇的中心词向量;

计算模块,用于将每个目标文档的目标词汇对应的第一词向量和中心词向量输入训练好的目标信息检索模型中的相关度打分层中,计算得到所述目标查询语句与每个所述目标文档之间的第一相关度;

第四确定模块,用于按照所述目标查询语句与多个目标文档之间的相关度由高到低的顺序,选取预设数量的所述目标文档作为检索文档输出。

6.根据权利要求5所述的信息检索装置,其特征在于,所述第三确定模块中针对每个目标文档中任一目标词汇,基于与该目标词汇关联的各个滑动窗口的第二词向量,确定该目标词汇的中心词向量,包括:

针对每个目标文档中任一目标词汇,获取该目标词汇在各个滑动窗口的第二词向量;

针对各个所述第二词向量进行求和计算,并在求和计算后进行平均值计算,将所述平均值确定为该目标词汇的中心词向量。

7.根据权利要求5所述的信息检索装置,其特征在于,所述计算模块中通过以下方式计算得到所述目标查询语句与每个所述目标文档之间的第一相关度:

针对每个目标词汇对应的第一词向量与每个所述目标词汇对应的每个中心词向量进行点乘计算,并将点乘结果确定为每个所述目标词汇与每个所述目标文档之间的第二相关

度;

针对每个目标文档中的各个所述目标词汇对应的各个所述第二相关度进行求和计算,将求和结果确定为目标查询语句与每个所述目标文档之间的第一相关度。

8.一种电子设备,其特征在于,包括:处理器、存储器和总线,所述存储器存储有所述处理器可执行的机器可读指令,当电子设备运行时,所述处理器与所述存储器之间通过总线通信,所述机器可读指令被所述处理器运行时执行如上述权利要求1至4中任一所述的信息检索方法的步骤。

9.一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上存储有计算机程序,所述计算机程序被处理器运行时执行如上述权利要求1至4中任一所述的信息检索方法的步骤。

一种信息检索方法、装置、电子设备及存储介质

技术领域

[0001] 本申请涉及电网优化技术领域,尤其是涉及一种信息检索方法、装置、电子设备及存储介质。

背景技术

[0002] 现有技术中的信息检索方式主要是基于词汇频率的匹配算法对搜索文本在数据库中进行检索,这种基于词频的信息检索方式依旧是当今检索系统的主流方法,但是,基于词频的信息检索只考虑到了词汇在数据库中出现的次数,并未结合语义环境以及词义环境,导致检索出来的文档匹配程度和准确度较低。

发明内容

[0003] 有鉴于此,本申请的目的在于提供一种信息检索方法、装置、电子设备及存储介质,本申请基于训练好的目标信息检索模型,确定目标词汇在目标查询语句的第一词向量,以及确定目标词汇在每个目标文档的中的中心词向量,根据第一词向量与中心词向量确定查询语句与目标文档的第一相关度,由于本申请中目标词汇的中心词向量是基于每个目标文档中的多个滑动窗口确定的,使得目标词汇对应的目标查询语句与得到的检索文档的语义和词义更加匹配,提高了信息检索的准确度。

[0004] 本申请实施例提供了一种信息检索方法,所述信息检索方法包括:

[0005] 将目标查询语句输入训练好的目标信息检索模型中的查询层中,得到所述目标查询语句与文本库中每个目标文档之间相同的至少一个目标词汇;其中,所述目标文档是所述文本库中与所述目标查询语句存在相同词汇的文档;

[0006] 将得到的每个所述目标词汇输入训练好的目标信息检索模型中的词向量提取网络层中,得到每个所述目标词汇在所述目标查询语句中的第一词向量,以及每个所述目标词汇在每个所述目标文档中的滑动窗口中的第二词向量;所述滑动窗口为所述目标文档中包含相邻的第一预设个数字符的窗口,且所述滑动窗口中包含至少一个所述目标词汇中的至少一个字符;相邻的两个所述滑动窗口之间包含第二预设个数的字符的重叠部分;

[0007] 针对每个目标文档中任一目标词汇,基于与该目标词汇关联的各个滑动窗口的第二词向量,确定该目标词汇的中心词向量;

[0008] 将每个目标文档的目标词汇对应的第一词向量和中心词向量输入训练好的目标信息检索模型中的相关度打分层中,计算得到所述目标查询语句与每个所述目标文档之间的第一相关度;

[0009] 按照所述目标查询语句与多个目标文档之间的相关度由高到低的顺序,选取预设数量的所述目标文档作为检索文档输出。

[0010] 进一步的,针对每个目标文档中任一目标词汇,基于与该目标词汇关联的各个滑动窗口的第二词向量,确定该目标词汇的中心词向量,包括:

[0011] 针对每个目标文档中任一目标词汇,获取该目标词汇在各个滑动窗口的第二词向

量；

[0012] 针对各个所述第二词向量进行求和计算，并在求和计算后进行平均值计算，将所述平均值确定为该目标词汇的中心词向量。

[0013] 进一步的，通过以下方式计算得到所述目标查询语句与每个所述目标文档之间的第一相关度：

[0014] 针对每个目标词汇对应的第一词向量与每个所述目标词汇对应的每个中心词向量进行点乘计算，并将点乘结果确定为每个所述目标词汇与每个所述目标文档之间的第二相关度；

[0015] 针对每个目标文档中的各个所述目标词汇对应的各个所述第二相关度进行求和计算，将求和结果确定为目标查询语句与每个所述目标文档之间的第一相关度。

[0016] 进一步的，所述针对每个目标词汇对应的第一词向量与每个所述目标词汇对应的每个中心词向量进行点乘计算，并将点乘结果确定为每个所述目标词汇与每个所述目标文档之间的第二相关度，包括：

[0017] 针对每个目标词汇对应的第一词向量与每个所述目标词汇对应的每个中心词向量进行点乘计算，选取最大点乘值为结果相关度；

[0018] 将所述结果相关度确定为每个所述目标词汇与每个所述目标文档之间的第二相关度。

[0019] 进一步的，通过以下方式确定训练好的目标信息检索模型；

[0020] 获取样本查询语句和与所述样本查询语句对应的样本文档；

[0021] 将所述样本查询语句与所述样本文档按照预设相关度进行相关度划分，确定所述相关度大于等于所述预设相关度的所述样本查询语句与所述样本文档为样本相关文本，以及确定所述相关度小于所述预设相关度的所述样本查询语句与所述样本文档为样本不相关文本；

[0022] 根据所述样本相关文本以及所述样本不相关文本，对初始信息检索模型进行训练，确定训练好的目标信息检索模型。

[0023] 本申请实施例还提供了一种信息检索装置，所述信息检索装置包括：

[0024] 第一确定模块，用于将目标查询语句输入训练好的目标信息检索模型中的查询层中，得到所述目标查询语句与文本库中每个目标文档之间相同的至少一个目标词汇；其中，所述目标文档是所述文本库中与所述目标查询语句存在相同词汇的文档；

[0025] 第二确定模块，用于将得到的每个所述目标词汇输入训练好的目标信息检索模型中的词向量提取网络层中，得到每个所述目标词汇在所述目标查询语句中的第一词向量，以及每个所述目标词汇在每个所述目标文档中的滑动窗口中的第二词向量；所述滑动窗口为所述目标文档中包含相邻的第一预设个数字符的窗口，且所述滑动窗口中包含至少一个所述目标词汇中的至少一个字符；相邻的两个所述滑动窗口之间包含第二预设个数的字符的重叠部分；

[0026] 第三确定模块，用于针对每个目标文档中任一目标词汇，基于与该目标词汇关联的各个滑动窗口的第二词向量，确定该目标词汇的中心词向量；

[0027] 计算模块，用于将每个目标文档的目标词汇对应的第一词向量和中心词向量输入训练好的目标信息检索模型中的相关度打分层中，计算得到所述目标查询语句与每个所述

目标文档之间的第一相关度；

[0028] 第四确定模块，用于按照所述目标查询语句与多个目标文档之间的相关度由高到低的顺序，选取预设数量的所述目标文档作为检索文档输出。

[0029] 进一步的，所述第三确定模块中针对每个目标文档中任一目标词汇，基于与该目标词汇关联的各个滑动窗口的第二词向量，确定该目标词汇的中心词向量，包括：

[0030] 针对每个目标文档中任一目标词汇，获取该目标词汇在各个滑动窗口的第二词向量；

[0031] 针对各个所述第二词向量进行求和计算，并在求和计算后进行平均值计算，将所述平均值确定为该目标词汇的中心词向量。

[0032] 进一步的，所述计算模块中通过以下方式计算得到所述目标查询语句与每个所述目标文档之间的第一相关度：

[0033] 针对每个目标词汇对应的第一词向量与每个所述目标词汇对应的每个中心词向量进行点乘计算，并将点乘结果确定为每个所述目标词汇与每个所述目标文档之间的第二相关度；

[0034] 针对每个目标文档中的各个所述目标词汇对应的各个所述第二相关度进行求和计算，将求和结果确定为目标查询语句与每个所述目标文档之间的第一相关度。

[0035] 本申请实施例还提供一种电子设备，包括：处理器、存储器和总线，所述存储器存储有所述处理器可执行的机器可读指令，当电子设备运行时，所述处理器与所述存储器之间通过总线通信，所述机器可读指令被所述处理器执行时执行如上述的信息检索方法的步骤。

[0036] 本申请实施例还提供一种计算机可读存储介质，该计算机可读存储介质上存储有计算机程序，该计算机程序被处理器运行时执行如上述的信息检索方法的步骤。

[0037] 本申请实施例提供的信息检索方法、装置、电子设备及存储介质，与现有技术相比，本申请基于训练好的目标信息检索模型，确定目标词汇在目标查询语句的第一词向量，以及确定目标词汇在每个目标文档的中的中心词向量，根据第一词向量与中心词向量确定查询语句与目标文档的第一相关度，由于本申请中目标词汇的中心词向量是基于每个目标文档中的多个滑动窗口确定的，使得目标词汇对应的目标查询语句与得到的检索文档的语义和词义更加匹配，提高了信息检索的准确度。

[0038] 为使本申请的上述目的、特征和优点能更明显易懂，下文特举较佳实施例，并配合所附附图，作详细说明如下。

附图说明

[0039] 为了更清楚地说明本申请实施例的技术方案，下面将对实施例中所需要使用的附图作简单地介绍，应当理解，以下附图仅示出了本申请的某些实施例，因此不应被看作是对范围的限定，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，还可以根据这些附图获得其他相关的附图。

[0040] 图1示出了本申请实施例所提供的一种信息检索方法的流程图；

[0041] 图2示出了本申请实施例所提供的一种信息检索方法中的滑动窗口与目标文档之间关系的结构图；

[0042] 图3示出了本申请实施例所提供的另一种信息检索方法的流程图；

[0043] 图4示出了本申请实施例所提供的一种信息检索装置的结构示意图；

[0044] 图5示出了本申请实施例所提供的一种电子设备的结构示意图。

[0045] 图中：

[0046] 400-信息检索装置；410-第一确定模块；420-第二确定模块；430-第三确定模块；440-计算模块；450-第四确定模块；500-电子设备；510-处理器；520-存储器；530-总线。

具体实施方式

[0047] 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，所描述的实施例仅仅是本申请一部分实施例，而不是全部的实施例。通常在此处附图中描述和示出的本申请实施例的组件可以以各种不同的配置来布置和设计。因此，以下对在附图中提供的本申请的实施例的详细描述并非旨在限制要求保护的本申请的范围，而是仅仅表示本申请的选定实施例。基于本申请的实施例，本领域技术人员在没有做出创造性劳动的前提下所获得的每个其他实施例，都属于本申请保护的范围。

[0048] 首先，对本申请可适用的应用场景进行介绍，经研究发现，现有技术中的信息检索方式主要是基于词汇频率的匹配算法对搜索文本在数据库中进行检索，这种基于词频的信息检索方式依旧是当今检索系统的主流方法，但是，基于词频的信息检索只考虑到了词汇在数据库中出现的次数，并未结合语义环境以及词义环境，导致检索出来的文档匹配程度和准确度较低。

[0049] 基于此，本申请实施例提供了一种信息检索方法、装置、电子设备及存储介质，基于训练好的目标信息检索模型，确定目标词汇在目标查询语句的第一词向量，以及确定目标词汇在每个目标文档的中的中心词向量，根据第一词向量与中心词向量确定查询语句与目标文档的第一相关度，由于本申请中目标词汇的中心词向量是基于每个目标文档中的多个滑动窗口确定的，使得目标词汇对应的目标查询语句与得到的检索文档的语义和词义更加匹配，提高了信息检索的准确度。

[0050] 请参阅图1，图1为本申请实施例所提供的一种信息检索方法的流程图。所如图1中所示，本申请实施例提供的信息检索方法，包括：

[0051] S101、将目标查询语句输入训练好的目标信息检索模型中的查询层中，得到所述目标查询语句与文本库中每个目标文档之间相同的至少一个目标词汇；其中，所述目标文档是所述文本库中与所述目标查询语句存在相同词汇的文档。

[0052] 该步骤中，在用户产生想要获取查询的目标查询语句对应的一系列检索文档的需求后，将目标查询语句输入训练好的目标信息检索模型中的查询网络层中，得到目标查询语句与文本库中每个目标文档之间一个目标词汇，这里，会对文本库中的所述文档进行筛选，剔除掉那些与目标查询语句没有任何关联的文档，并将文本库中与目标查询语句存在相同词汇的文档筛选出来，确定为目标文档。

[0053] 这里，目标词汇定义为同时存在于目标查询语句与目标文档中的相同词，且目标词汇的字符个数不唯一，目标词汇的划分为可根据中文的表达特点进行自定义划分，在本申请提供的实施例中，使用q来表示目标词汇，其中，qi表示为第i个标记的目标词汇，且查

询网络层用于表征目标信息检索模型中用于确定目标词汇的网络结构层。

[0054] 其中,目标信息检索模型为对初始信息检索模型训练得到的,通过以下方式确定训练好的目标信息检索模型;

[0055] 获取样本查询语句和与所述样本查询语句对应的样本文档。

[0056] 这里,首先,获取初始样本查询语句和所述初始样本查询语句对应的初始样本文档,初始样本查询语句为从日志或其他外部样本数据库中获取的,然后通过人工标注初始样本查询语句对应的初始样本文档,将初始查询语句以及初始样本文档进行去噪处理,删除掉无意义的特殊字符,如空格以及乱码等字符,这里,使用正则表达式进行清理。

[0057] 其中,正则表达式是对字符串操作的一种逻辑公式,就是用事先定义好的一些特定字符、及这些特定字符的组合,组成一个“规则字符串”,这个“规则字符串”用来表达对字符串的一种过滤逻辑。

[0058] 将所述样本查询语句与所述样本文档按照预设相关度进行相关度划分,确定相关度大于等于所述预设相关度的所述样本查询语句与所述样本文档为样本相关文本,以及确定所述相关度小于所述预设相关度的所述样本查询语句与所述样本文档为样本不相关文本。

[0059] 上述中,样本相关文本以及样本不相关文本均包括训练初始信息检模型用的样本训练集以及样本验证集,使用python编程语言将样本查询语句与样本文档按照预设相关度进行相关度划分,并将相关度大于等于预设相关度样本查询语句与样本文档作为样本相关文本,其中,样本相关文本通过进行样本查询语句与样本文档的相关度标注而实现与预设相关度比较,将相关度小于预设相关度样本查询语句与样本文档作为样本不相关文本,并且将根据概率检索算法得到的预测相关度较高的样本查询语句和样本文档,但是实际上并不相关的样本确定为困难不相关样本。

[0060] 这样,概率检索算法是基于概率检索模型提出的算法,概率检索算法包括但不限制于使用BM25信息检索算法。

[0061] 根据所述样本训练集以及所述样本验证集,对所述初始信息检索模型进行训练,确定训练好的目标信息检索模型。

[0062] 这里,根据样本训练集,对初始信息检索模型进行训练,并在训练的过程中,使用样本验证集对初始信息检索模型中的网络结构参数进行实时的更新和替换,进而实现对初始信息检索模型的训练效果的验证,并且在训练结束后,将训练好的目标信息检索模型中的模型文档进行对应存储,便于后续的信息检索任务或对目标文档的重排序任务。

[0063] 上述,可以将训练好的目标信息检索模型应用到线上的不同的检索系统中,具体需要首先使用文档编码器对文本库中的文档进行离线处理,实现基于目标查询语句,对文档的预处理,删除掉文档中的一些噪声;然后在检索系统中加载训练好的目标信息检索模型,并对训练好的目标信息检索模型进行初始化,接着,输入去噪后的目标查询语句,得到与目标查询语句对应的预设数量的目标文档作为检索文档输出。

[0064] 再次,可以将训练好的目标信息检索模型用于召回文本库外的其他文档,并对其他文档进行重新排序,具体过程为:首先,加载训练好的目标信息检索模型,并对训练好的目标信息检索模型进行初始化,其次使用文档编码器对文本库中的文档进行离线处理,然后利用搜索引擎中BM25信息检索算法召回指定数量的文本库外的其他文本库的目标文档,

并基于初始化后的目标信息检索文档,对召回后的文本库外的目标文档进行相关度匹配,并将预设数量的目标文档作为召回的检索文档输出。

[0065] 其中,训练好的目标信息检索模型的结构图,这里,使用初始信息检索模型首先对样本查询语句代表的字符串和样本查询语句对应的样本文档代表的字符串分别进行编码,其中,本申请提供的实施例中的初始信息检索模型可以为基于BERT编码的为语言模型,但不仅仅限制于使用BERT编码,这里,使用Adam优化器对BERT编码器进行微调,实现调整训练过程中的初始信息检索模型的参数的功能。

[0066] 其中,可以使用负对数似然函数来对初始信息检索模型进行训练,确定训练好的目标信息检索模型,似然函数是一种关于模型中参数的函数,“似然性(likelihood)”和"概率(probability)"词意相似,但在统计学中它们有着完全不同的含义:概率用于在已知参数的情况下,预测接下来的观测结果;似然性用于根据一些观测结果,估计给定模型的参数可能值。

[0067] 这里,使用Token来表示样本查询语句以及样本文档生成的一串字符串,作为客户端进行请求的一个令牌,当第一次登录后,服务器生成一个Token便将此Token返回给客户端,以后客户端只需带上这个Token前来请求数据即可。Token其实说的更通俗点可以叫暗号,即在一些数据传输之前,要先进行暗号的核对,不同的暗号被授权不同的数据操作,在本申请提供的实施例中,Token表示为样本查询语句以及样本文档中的任意一个字符或文字,即每个样本查询语句以及每个样本文档中包含了多个Token,Token的个数取决于字符或文字的个数。

[0068] 这里,字符串Token,具体表示为:

[0069] $W_{tok}^{n_t \times n_{lm}}$;

[0070] 这里,$W_{tok}^{n_t \times n_{lm}}$是一个矩阵,它将初始信息检索模型的nlm维输出映射为低维nt的向量,训练好的目标信息检索模型为添加了预训练语言模型(Bidirectional Encoder Representations from Transformers,BERT)中[CLS]匹配的信息检索模型,其中,预训练语言模型是通过预训练来学习无标注数据中的深度双向表示,且在预训练语言模型中添加[CLS]是将该符号对应的输出字符向量作为整篇文本或语句的语义表示。

[0071] 进一步的,通过以下方式获取标准样本查询文本和所述标准样本查询文本对应的标准样本文档:

[0072] 获取初始样本查询文本和所述初始样本查询文本对应的初始标准样本文档。

[0073] 针对所述初始样本查询文本以及所述初始标准样本文档进行去噪处理,确定标准样本查询文本以及标准样本文档。

[0074] S102、将得到的每个所述目标词汇输入训练好的目标信息检索模型中的词向量提取网络层中,得到每个所述目标词汇在所述目标查询语句中的第一词向量,以及每个所述目标词汇在每个所述目标文档中的滑动窗口中的第二词向量;所述滑动窗口为所述目标文档中包含相邻的第一预设个数字符的窗口,且所述滑动窗口中包含至少一个所述目标词汇中的至少一个字符;相邻的两个所述滑动窗口之间包含第二预设个数的字符的重叠部分。

[0075] 该步骤中,将目标查询语句与文本库中每个目标文档之间的每个目标词汇输入训练好的目标信息检索模型中的词向量提取网络层中,得到该目标词汇在目标查询语句中的

第一词向量,以及上述目标词汇在每个目标文档中的滑动窗口中的第二词向量,其中,第一词向量基于Token的编码具体表示为:

[0076] $$v_i^q = W_{tok}LM(q,i) + b_{tok};$$

[0077] 这里,目标查询语句的英文名称为query,式中,使用缩写q来表示,v_i^q表示第i个标记的目标查询语句的字符串,b_{tok}为系数,W_{tok}为矩阵形式的系数。

[0078] 且其中,需要考虑到目标查询语句中不同目标词汇的相关度,我们利用BERT中的[CLS]匹配来获取第一词向量,此时第一词向量的具体表达式为:

[0079] $$v_{cls}^q = W_{cls}LM(q,CLS) + b_{cls};$$

[0080] 这里,第二词向量基于Token的编码具体表示为:

[0081] $$v_j^d = W_{tok}LM(d,j) + b_{tok};$$

[0082] 这里,目标文档的英文名称为document,式中,使用缩写d来表示,v_j^d第i个标记的目标文档字符串,b_{tok}为系数。

[0083] 且其中,需要考虑到目标查询语句中不同目标词汇的相关度,我们利用BERT中的[CLS]匹配来获取第二词向量,此时第二词向量的具体表达式为:

[0084] $$v_{cls}^d = W_{cls}LM(d,CLS) + b_{cls};$$

[0085] 上述中,v_{cls}^q和v_{cls}^d之间的相关度可以提供高层次的语义匹配信息,缓解词汇不匹配的问题。

[0086] 这样,滑动窗口为在每个目标文档中按照预设方向进行滑动的窗口,且滑动窗口中包括第一预设个数的字符,这里,第一预设个数的设定并不固定,可根据中文汉子的表达特点进行自定义个数的字符设置,且每个滑动窗口中包括至少一个目标词汇中的至少一个字符,且相邻的两个滑动窗口之间包含第二预设个数的字符的重叠部分,这里,本申请提供的实施例部分,设定第二预设个数的字符的重叠部分为一个字符,也就是说,当滑动窗口在任一目标文档中按照预设方向进行滑动时,当有滑动窗口中出现目标词汇时,其相邻的下一个滑动窗口为与上述滑动窗口相距一个重叠字符的滑动窗口。

[0087] 如图2所示,图2为本申请实施例所提供的一种信息检索方法中的滑动窗口与目标文档之间关系的结构图。

[0088] 其中,设定滑动窗口包括四个字符,即第一预设个数的字符为四个,例如:本实施例提供的目标查询语句为“骄傲使人落后”,其对应的文本库中的一个目标文档为“身为中国人,女排取得奥运会冠军,我感到无比骄傲和自豪”。

[0089] 这里,目标查询语句与目标文档中的目标词汇只有一个,目标词汇具体为“骄傲”,此时,获取目标词汇中出现包括目标词汇“骄傲”中任一字符“骄”或“傲”的所有滑动窗口,使用编码器获取目标词汇在每个滑动窗口中的基于语义的第二词向量,下面具体表示图2中包括目标词汇“骄傲”的滑动窗口为:“到无比骄”、“无比骄傲”、“比骄傲和”以及“骄傲和自”。

[0090] 若目标文档中存在多个目标词汇,则对每个目标词汇设置滑动窗口的过程同上述

过程,在此处不在赘述。

[0091] S103、针对每个目标文档中任一目标词汇,基于与该目标词汇关联的各个滑动窗口的第二词向量,确定该目标词汇的中心词向量。

[0092] 该步骤中,将任一目标词汇关联的各个滑动窗口的第二词向量进行求和平均计算,确定该目标词汇在目标文档中的中心词向量。

[0093] 这样,目标词汇在目标文档中的中心词向量用CK表示。

[0094] S104、将每个目标文档的目标词汇对应的第一词向量和中心词向量输入训练好的目标信息检索模型中的相关度打分层中,计算得到所述目标查询语句与每个所述目标文档之间的第一相关度。

[0095] 该步骤中,在将每个目标文档的目标词汇对应的第一词向量和中心词向量输入训练好的目标信息检索模型中的相关度计算网络层中后,针对每个目标词汇对应的第一词向量与每个所述目标词汇对应的每个中心词向量进行点乘计算,并选取最大点乘值为结果相关度,点乘结果确定为每个所述目标词汇与每个所述目标文档之间的第二相关度。

[0096] 这里,针对每个目标文档中的各个所述目标词汇对应的各个所述第二相关度进行求和计算,将求和结果确定为目标查询语句与每个所述目标文档之间的第一相关度。

[0097] 其中,计算目标查询语句与每个所述目标文档之间的第一相关度的表达式具体为:

[0098] $$s_{tok}(q,d)=\sum_{q_i\in q\cap d}\max_{d_j=q_i}(v_i^{q^T}v_{\mathrm{k}}^{c})$$;

[0099] 式中,$q_i\in q\cap d$表示为第i个标记的目标词汇为在目标查询语句与目标文档中共有的相同词,这里,选取最大点乘值,即进行max运算是为了捕捉目标文档中上述目标词汇的重要语义信息。

[0100] 上述为未引入BERT中的[CLS]匹配方式的计算目标查询语句与每个所述目标文档之间的第一相关度的方式,下列公式为添加了[CLS]匹配方式的计算第一相关度的表达式:

[0101] $$s_{full}(q,d)=s_{tok}(q,d)+v_{cls}^{q\ T}v_{cls}^{d}$$;

[0102] 式中,full表示为添加[CLS]匹配的训练好的目标信息检索模型中相关度计算网络层的输出。

[0103] S105、按照所述目标查询语句与多个目标文档之间的相关度由高到低的顺序,选取预设数量的所述目标文档作为检索文档输出。

[0104] 该步骤中,当用户对目标查询语句进行查询时,通过将目标查询语句输入到训练好的目标信息检索模型中,会输出按照相关度由高到低进行排序的预设数量的目标文档作为检索文档输出,这里的检索文档可以作为该目标查询语句的答案文档。

[0105] 本申请实施例提供的信息检索方法,与现有技术中的信息检索方法相比,本申请基于训练好的目标信息检索模型,确定目标词汇在目标查询语句的第一词向量,以及确定目标词汇在每个目标文档的中的中心词向量,根据第一词向量与中心词向量确定查询语句与目标文档的第一相关度,由于本申请中目标词汇的中心词向量是基于每个目标文档中的多个滑动窗口确定的,使得目标词汇对应的目标查询语句与得到的检索文档的语义和词义

更加匹配,提高了信息检索的准确度。

[0106] 请参阅图3,图3为本申请另一实施例提供的信息检索方法的流程图。如图3中所示,本申请实施例提供的信息检索方法,包括:

[0107] S201、将目标查询语句输入训练好的目标信息检索模型中的查询层中,得到所述目标查询语句与文本库中每个目标文档之间相同的至少一个目标词汇;其中,所述目标文档是所述文本库中与所述目标查询语句存在相同词汇的文档。

[0108] S202、将得到的每个所述目标词汇输入训练好的目标信息检索模型中的词向量提取网络层中,得到每个所述目标词汇在所述目标查询语句中的第一词向量,以及每个所述目标词汇在每个所述目标文档中的滑动窗口中的第二词向量;所述滑动窗口为所述目标文档中包含相邻的第一预设个数字符的窗口,且所述滑动窗口中包含至少一个所述目标词汇中的至少一个字符;相邻的两个所述滑动窗口之间包含第二预设个数的字符的重叠部分。

[0109] S203、针对每个目标文档中任一目标词汇,获取该目标词汇在各个滑动窗口的第二词向量。

[0110] 该步骤中,针对每个目标文档中的多个目标词汇,均获取任一目标词汇在各个滑动窗口的第二词向量。

[0111] 其中,在每个目标文档中按照预设方向进行滑动的窗口,且滑动窗口中包括第一预设个数的字符,这里,第一预设个数的设定并不固定,可根据中文汉子的表达特点进行自定义个数的字符设置,且每个滑动窗口中包括至少一个目标词汇中的至少一个字符,且相邻的两个滑动窗口之间包含第二预设个数的字符的重叠部分,这里,本申请提供的实施例部分,设定第二预设个数的字符的重叠部分为一个字符,也就是说,当滑动窗口在任一目标文档中按照预设方向进行滑动时,当有滑动窗口中出现目标词汇时,其相邻的下一个滑动窗口为与上述滑动窗口相距一个重叠字符的滑动窗口。

[0112] S204、针对各个所述第二词向量进行求和计算,并在求和计算后进行平均值计算,将所述平均值确定为该目标词汇的中心词向量。

[0113] 该步骤中,每个目标词汇的中心词向量均通过对滑动窗口的第二词向量进行求和后却平均值后得到的。

[0114] S205、将每个目标文档的目标词汇对应的第一词向量和中心词向量输入训练好的目标信息检索模型中的相关度打分层中,计算得到所述目标查询语句与每个所述目标文档之间的第一相关度。

[0115] S206、按照所述目标查询语句与多个目标文档之间的相关度由高到低的顺序,选取预设数量的所述目标文档作为检索文档输出。

[0116] 其中,S201至S202以及S205至S206的描述可以参照S101至S102以及S104至S105的描述,并且能达到相同的技术效果,对此不做赘述。

[0117] 本申请实施例提供的信息检索方法,与现有技术中的信息检索方法相比,本申请基于训练好的目标信息检索模型,确定目标词汇在目标查询语句的第一词向量,以及确定目标词汇在每个目标文档的中的中心词向量,根据第一词向量与中心词向量确定查询语句与目标文档的第一相关度,由于本申请中目标词汇的中心词向量是基于每个目标文档中的多个滑动窗口确定的,使得目标词汇对应的目标查询语句与得到的检索文档的语义和词义更加匹配,提高了信息检索的准确度。

[0118] 请参阅图4,图4为本申请实施例所提供的一种信息检索装置的结构示意图。如图4中所示,所述信息检索装置400包括:

[0119] 第一确定模块410,用于将目标查询语句输入训练好的目标信息检索模型中的查询层中,得到所述目标查询语句与文本库中每个目标文档之间相同的至少一个目标词汇;其中,所述目标文档是所述文本库中与所述目标查询语句存在相同词汇的文档。

[0120] 第二确定模块420,用于将得到的每个所述目标词汇输入训练好的目标信息检索模型中的词向量提取网络层中,得到每个所述目标词汇在所述目标查询语句中的第一词向量,以及每个所述目标词汇在每个所述目标文档中的滑动窗口中的第二词向量;所述滑动窗口为所述目标文档中包含相邻的第一预设个数字符的窗口,且所述滑动窗口中包含至少一个所述目标词汇中的至少一个字符;相邻的两个所述滑动窗口之间包含第二预设个数的字符的重叠部分。

[0121] 第三确定模块430,用于针对每个目标文档中任一目标词汇,基于与该目标词汇关联的各个滑动窗口的第二词向量,确定该目标词汇的中心词向量。

[0122] 进一步的,第三确定模块中针对每个目标文档中任一目标词汇,基于与该目标词汇关联的各个滑动窗口的第二词向量,确定该目标词汇的中心词向量,包括:

[0123] 针对每个目标文档中任一目标词汇,获取该目标词汇在各个滑动窗口的第二词向量。

[0124] 针对各个所述第二词向量进行求和计算,并在求和计算后进行平均值计算,将所述平均值确定为该目标词汇的中心词向量。

[0125] 计算模块440,用于将每个目标文档的目标词汇对应的第一词向量和中心词向量输入训练好的目标信息检索模型中的相关度打分层中,计算得到所述目标查询语句与每个所述目标文档之间的第一相关度。

[0126] 进一步的,计算模块440中通过以下方式计算得到所述目标查询语句与每个所述目标文档之间的第一相关度:

[0127] 针对每个目标词汇对应的第一词向量与每个所述目标词汇对应的每个中心词向量进行点乘计算,并将点乘结果确定为每个所述目标词汇与每个所述目标文档之间的第二相关度。

[0128] 针对每个目标文档中的各个所述目标词汇对应的各个所述第二相关度进行求和计算,将求和结果确定为目标查询语句与每个所述目标文档之间的第一相关度。

[0129] 第四确定模块440,用于按照所述目标查询语句与多个目标文档之间的相关度由高到低的顺序,选取预设数量的所述目标文档作为检索文档输出。

[0130] 本申请实施例提供的信息检索装置400,与现有技术中的相比,本申请实施例提供的信息检索方法,与现有技术中的信息检索方法相比,本申请基于训练好的目标信息检索模型,确定目标词汇在目标查询语句的第一词向量,以及确定目标词汇在每个目标文档的中的中心词向量,根据第一词向量与中心词向量确定查询语句与目标文档的第一相关度,由于本申请中目标词汇的中心词向量是基于每个目标文档中的多个滑动窗口确定的,使得目标词汇对应的目标查询语句与得到的检索文档的语义和词义更加匹配,提高了信息检索的准确度。

[0131] 请参阅图5,图5为本申请实施例所提供的一种电子设备的结构示意图。如图5中所

示,所述电子设备500包括处理器510、存储器520和总线530。

[0132] 所述存储器520存储有所述处理器510可执行的机器可读指令,当电子设备500运行时,所述处理器510与所述存储器520之间通过总线530通信,所述机器可读指令被所述处理器510执行时,可以执行如上述图1以及图4所示方法实施例中的信息检索方法的步骤,具体实现方式可参见方法实施例,在此不再赘述。

[0133] 本申请实施例还提供一种计算机可读存储介质,该计算机可读存储介质上存储有计算机程序,该计算机程序被处理器运行时可以执行如上述图1以及图4所示方法实施例中的信息检索方法的步骤,具体实现方式可参见方法实施例,在此不再赘述。

[0134] 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。

[0135] 在本申请所提供的几个实施例中,应该理解到,所揭露的系统、装置和方法,可以通过其它的方式实现。以上所描述的装置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,又例如,多个单元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通信接口,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。

[0136] 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。

[0137] 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。

[0138] 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时,可以存储在一个处理器可执行的非易失的计算机可读取存储介质中。基于这样的理解,本申请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、只读存储器(Read-Only Memory,ROM)、随机存取存储器(Random Access Memory,RAM)、磁碟或者光盘等各种可以存储程序代码的介质。

[0139] 最后应说明的是:以上所述实施例,仅为本申请的具体实施方式,用以说明本申请的技术方案,而非对其限制,本申请的保护范围并不局限于此,尽管参照前述实施例对本申请进行了详细的说明,本领域的普通技术人员应当理解:任何熟悉本技术领域的技术人员在本申请揭露的技术范围内,其依然可以对前述实施例所记载的技术方案进行修改或可轻易想到变化,或者对其中部分技术特征进行等同替换;而这些修改、变化或者替换,并不使相应技术方案的本质脱离本申请实施例技术方案的精神和范围,都应涵盖在本申请的保护范围之内。因此,本申请的保护范围应以权利要求的保护范围为准。

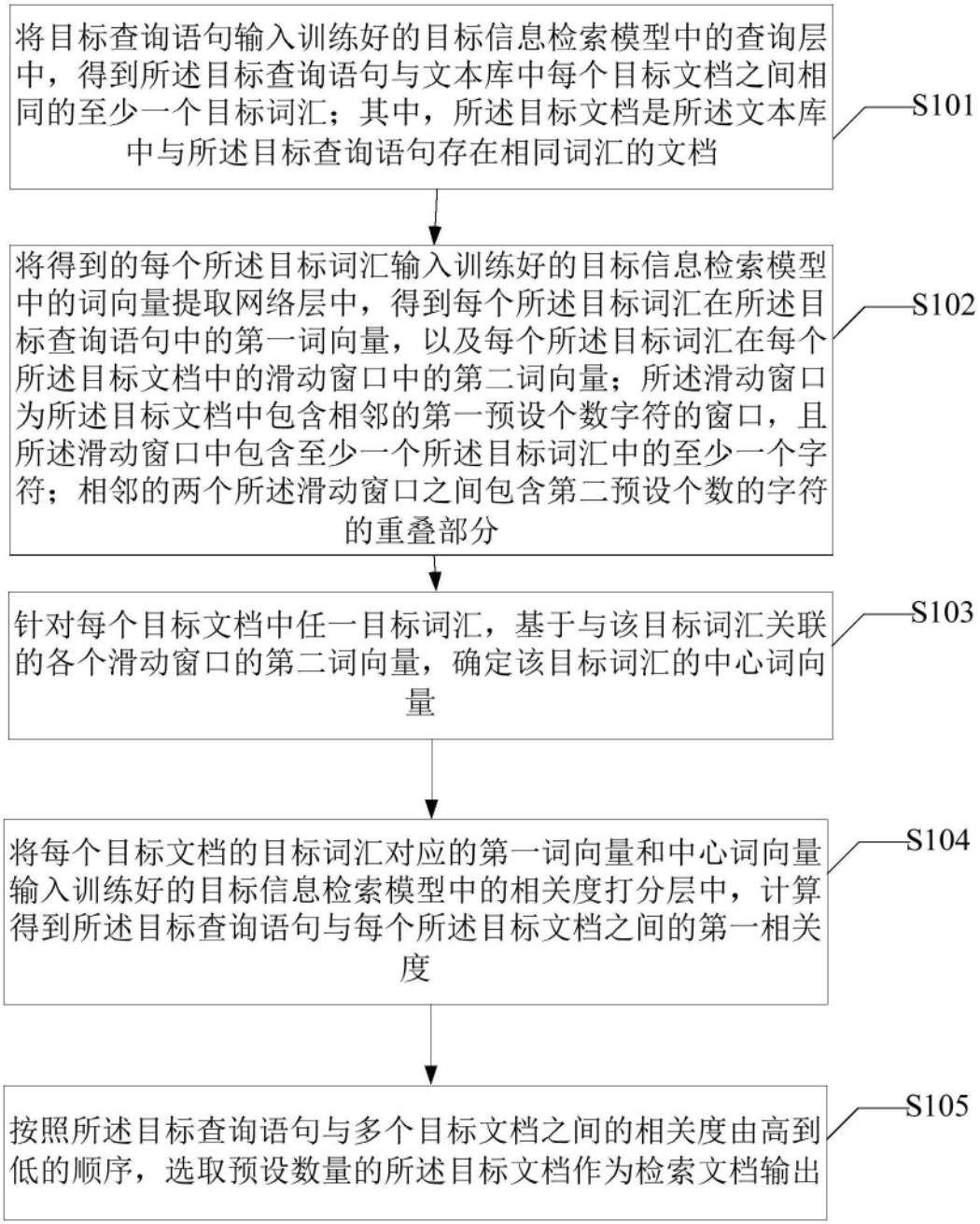

将目标查询语句输入训练好的目标信息检索模型中的查询层中，得到所述目标查询语句与文本库中每个目标文档之间相同的至少一个目标词汇；其中，所述目标文档是所述文本库中与所述目标查询语句存在相同词汇的文档
S101
将得到的每个所述目标词汇输入训练好的目标信息检索模型中的词向量提取网络层中，得到每个所述目标词汇在所述目标查询语句中的第一词向量，以及每个所述目标词汇在每个所述目标文档中的滑动窗口中的第二词向量；所述滑动窗口为所述目标文档中包含相邻的第一预设个数字符的窗口，且所述滑动窗口中包含至少一个所述目标词汇中的至少一个字符；相邻的两个所述滑动窗口之间包含第二预设个数的字符的重叠部分
S102
针对每个目标文档中任一目标词汇，基于与该目标词汇关联的各个滑动窗口的第二词向量，确定该目标词汇的中心词向量
S103
将每个目标文档的目标词汇对应的第一词向量和中心词向量输入训练好的目标信息检索模型中的相关度打分层中，计算得到所述目标查询语句与每个所述目标文档之间的第一相关度
S104
按照所述目标查询语句与多个目标文档之间的相关度由高到低的顺序，选取预设数量的所述目标文档作为检索文档输出
S105

图1

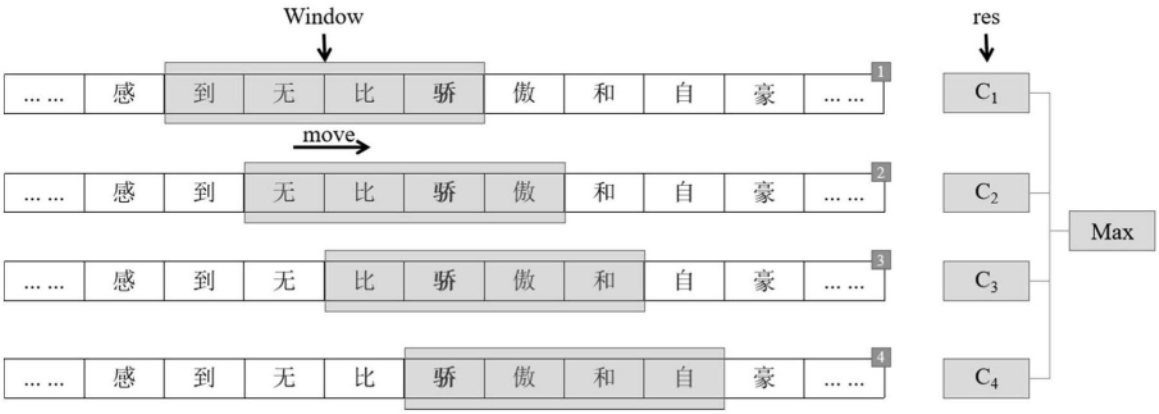


图2

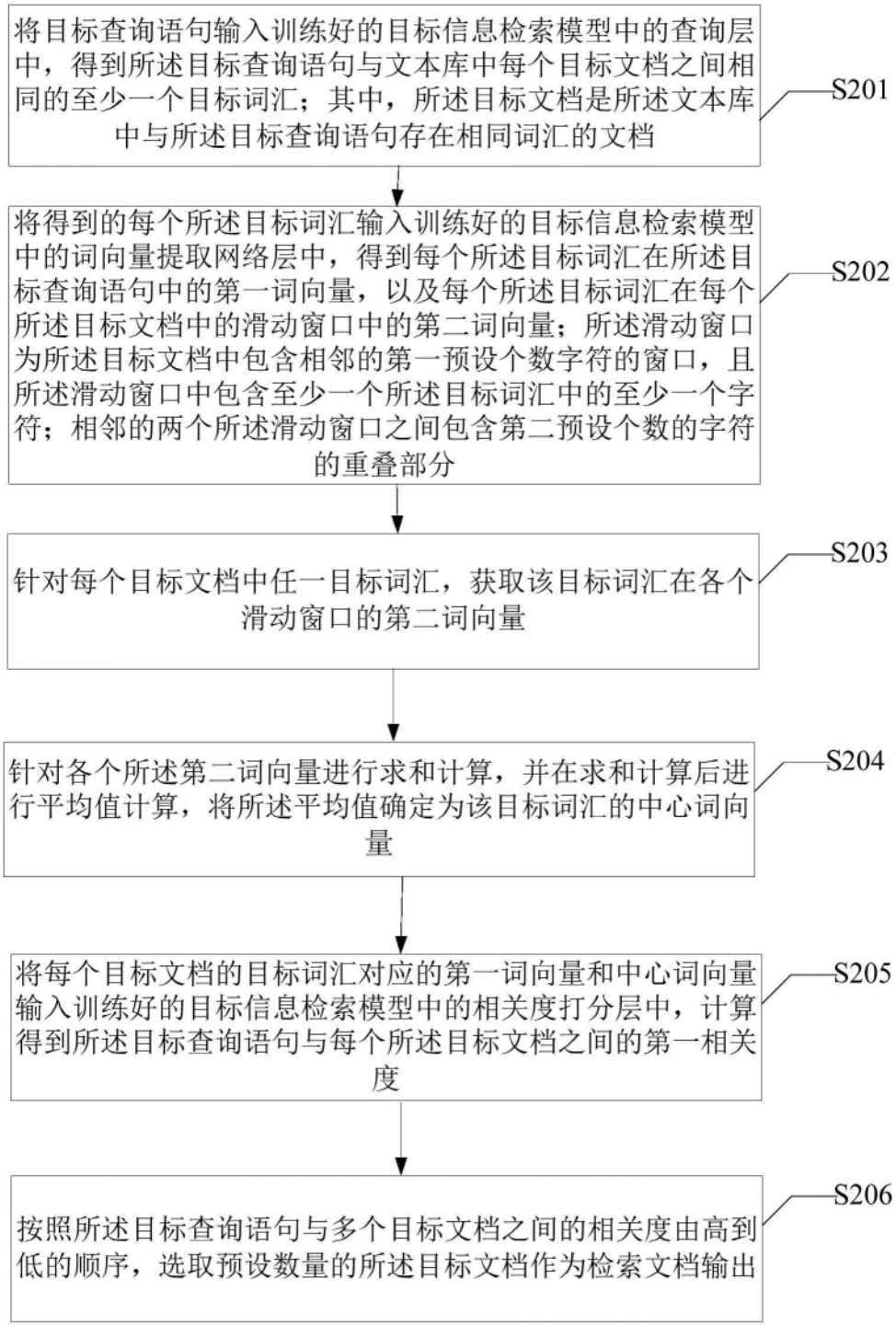

将目标查询语句输入训练好的目标信息检索模型中的查询层中，得到所述目标查询语句与文本库中每个目标文档之间相同的至少一个目标词汇；其中，所述目标文档是所述文本库中与所述目标查询语句存在相同词汇的文档
S201
将得到的每个所述目标词汇输入训练好的目标信息检索模型中的词向量提取网络层中，得到每个所述目标词汇在所述目标查询语句中的第一词向量，以及每个所述目标词汇在每个所述目标文档中的滑动窗口中的第二词向量；所述滑动窗口为所述目标文档中包含相邻的第一预设个数字符的窗口，且所述滑动窗口中包含至少一个所述目标词汇中的至少一个字符；相邻的两个所述滑动窗口之间包含第二预设个数的字符的重叠部分
S202
针对每个目标文档中任一目标词汇，获取该目标词汇在各个滑动窗口的第二词向量
S203
针对各个所述第二词向量进行求和计算，并在求和计算后进行平均值计算，将所述平均值确定为该目标词汇的中心词向量
S204
将每个目标文档的目标词汇对应的第一词向量和中心词向量输入训练好的目标信息检索模型中的相关度打分层中，计算得到所述目标查询语句与每个所述目标文档之间的第一相关度
S205
按照所述目标查询语句与多个目标文档之间的相关度由高到低的顺序，选取预设数量的所述目标文档作为检索文档输出
S206

图3

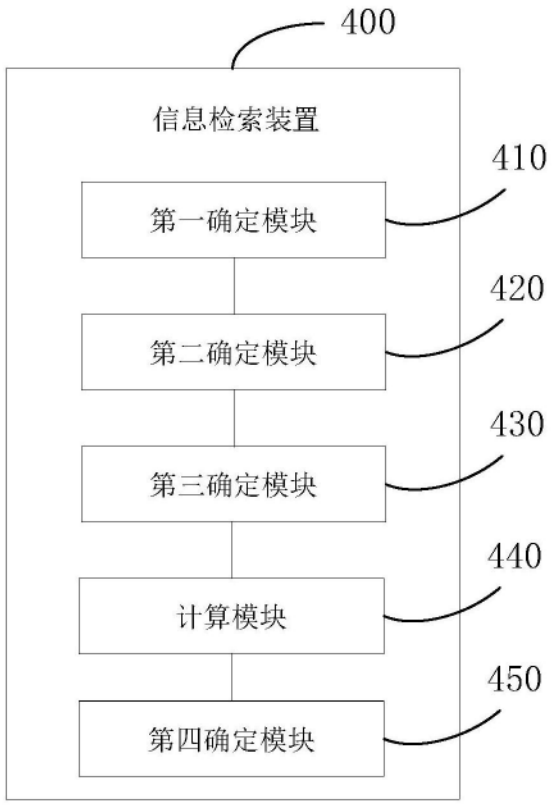

400
信息检索装置
410
第一确定模块
420
第二确定模块
430
第三确定模块
440
计算模块
450
第四确定模块

图4

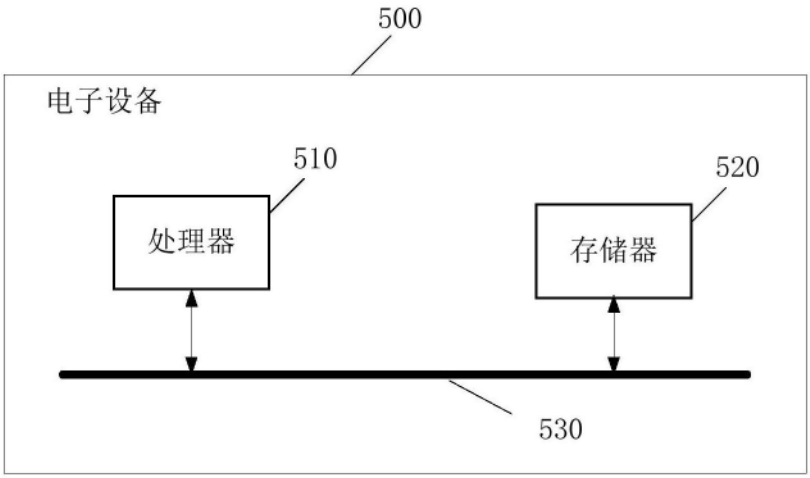

500
电子设备
510
处理器
520
存储器
530

图5