(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 117591639 A
(43)申请公布日 2024.02.23

(21)申请号 202311368282.6

(22)申请日 2023.10.20

(71)申请人 北京猎户星空科技有限公司
地址 100025 北京市朝阳区姚家园南路一号惠通时代广场8号

(72)发明人 张大成 李永强 陈都 韩堃 孙晋喜 樊扬

(74)专利代理机构 北京同达信恒知识产权代理有限公司 11291
专利代理师 杜秀科

(51)Int.Cl.
G06F 16/332(2019.01)
G06F 40/35(2020.01)
G06F 40/284(2020.01)
G06F 40/205(2020.01)
G06F 40/289(2020.01)
G06F 18/22(2023.01)

权利要求书3页 说明书18页 附图6页

(54)发明名称

一种问题答复方法、装置、设备及介质

(57)摘要

本申请实施例涉及一种问题答复方法、装置、设备及介质，用以提高问题答复的准确性，同时无需事先定义规则和关键词，节省人力，且能够处理复杂问题和新的问题。所述方法包括：获取用户提出的问题；依据预设的多种检索策略，分别对所述问题进行处理，获得每种检索策略对应的检索用词；在预先建立的向量数据库中，基于获得的每种检索策略对应的检索用词，采用相应的检索策略进行检索，获得检索内容，所述向量数据库中至少包括多个文档及对应的文档向量，所述文档向量是对文档进行词向量转换得到的；基于所述检索内容答复所述问题。

CN 117591639 A

1.一种问题答复方法,其特征在于,包括:

获取用户提出的问题;

依据预设的多种检索策略,分别对所述问题进行处理,获得每种检索策略对应的检索用词;

在预先建立的向量数据库中,基于获得的每种检索策略对应的检索用词,采用相应的检索策略进行检索,获得检索内容,所述向量数据库中至少包括多个文档及对应的文档向量,所述文档向量是对文档进行词向量转换得到的;

基于所述检索内容答复所述问题。

2.根据权利要求1所述的方法,其特征在于,所述依据预设的多种检索策略,分别对所述问题进行处理,获得每种检索策略对应的检索用词,包括以下操作的至少一种:

若检索策略为语义检索,则对所述问题进行词向量转换,获得对应的语义向量,将获得的语义向量作为检索用词;

若检索策略为关键词检索,则对所述问题进行分词处理,获得所述问题包含的词语,在获得的词语中选择第一预设数量的关键词作为检索用词;

若检索策略为候选问题检索,则利用预先训练的大语言模型对所述问题进行处理,生成第二预设数量的候选问题,将生成的候选问题作为检索用词。

3.根据权利要求2所述的方法,其特征在于,所述在预先建立的向量数据库中,基于获得的每种检索策略对应的检索用词,采用相应的检索策略进行检索,获得检索内容,包括以下操作中的至少一种:

若检索策略为语义检索,在获得所述问题对应的语义向量作为检索用词之后,在所述向量数据库中,检索与所述问题对应的语义向量的相似度大于第一预设阈值的文档向量,获得检索内容;

若检索策略为关键词检索,在获得第一预设数量的关键词作为检索用词之后,在所述向量数据库中,检索与所述关键词相关联的文档,获得检索内容;

若检索策略为候选问题检索,在获得候选问题作为检索用词之后,将所述候选问题作为待答复的问题,利用语义检索策略和关键词检索策略,在向量数据库中检索与所述候选问题相关联的文档,获得检索内容。

4.根据权利要求3所述的方法,其特征在于,所述向量数据库中,还包括:扩展内容向量,所述扩展内容向量是对扩展内容进行词向量转换生成的,所述扩展内容是利用预先训练的大语言模型,基于文档生成的,所述扩展内容包括以下至少一种:所述文档的摘要、对所述文档进行提问生成的多个文档问题、以及对所述摘要进行提问生成的多个摘要问题;

所述在所述向量数据库中,检索与所述问题对应的语义向量的相似度大于第一预设阈值的文档向量,获得检索内容,包括:

在所述向量数据库中,检索与所述问题对应的语义向量的相似度大于第一预设阈值的文档向量和/或扩展内容向量,获得检索内容。

5.根据权利要求4所述的方法,其特征在于,所述扩展内容向量,采用如下方式生成:

在确定所述文档包含的字符数量大于预设数量阈值时,将所述文档切分为多个文档片段;

若所述扩展内容包括所述文档的摘要,则利用所述大语言模型,生成所述多个文档片

段对应的多级摘要;

若所述扩展内容包括对所述文档进行提问生成的多个文档问题,则利用所述大语言模型,对每个文档片段进行提问,生成每个文档片段对应的文档问题;

若所述扩展内容包括对所述文档的摘要进行提问生成的多个摘要问题,则利用所述大语言模型,生成所述多个文档片段对应的多级摘要,并对每个摘要提问,生成每个摘要对应的摘要问题;

其中,所述多级摘要中每一级中包含至少一个摘要,每个摘要均是基于上一级中至少一个文档片段或者至少一个摘要生成的。

6.根据权利要求5所述的方法,其特征在于,所述在确定所述文档包含的字符数量大于预设数量阈值时,将所述文档切分为多个文档片段,包括:

基于预设切分方式,对所述文档进行初次切分,获得多个文本块;

针对获得的每个文本块,执行以下操作进行二次切分,得到多个文档片段:提取所述文本块的语义特征,并基于所述语义特征,将所述文本块切分为文档片段。

7.根据权利要求4所述的方法,其特征在于,所述向量数据库中还包括:与所述文档向量对应存储所述文档的元信息,与所述扩展内容向量对应存储所述扩展内容的元信息;

所述基于所述检索内容答复所述问题,包括:

基于所述检索内容对应存储的元信息,确定所述检索内容归属的源文档或者源摘要;

基于所述源文档或者所述源摘要生成提示词,将所述提示词发送至预先训练的大语言模型,以使所述大语言模型基于所述提示词答复所述问题。

8.根据权利要求7所述的方法,其特征在于,所述基于所述检索内容对应存储的元信息,确定所述检索内容归属的源文档或者源摘要,包括:

若所述检索内容为文档向量,则基于所述文档向量对应的元信息,确定所述文档向量归属的源文档;

若所述检索内容为摘要向量,则基于所述摘要向量对应的元信息,确定所述摘要向量归属的源摘要;

若所述检索内容为文档问题向量,则基于所述文档问题向量对应的元信息,确定所述文档问题向量归属的源文档;

若所述检索内容为摘要问题向量,则基于所述摘要问题向量对应的元信息,确定所述摘要问题向量归属的源摘要。

9.根据权利要求2-8中任一项所述的方法,其特征在于,所述在获得的词语中选择第一预设数量的关键词作为检索用词,包括:

在获得的词语中,筛选出存在于预先确定的关键词汇表中的词语,作为关键词,其中,所述关键词汇表是对文档切分之后,基于切分得到的文档片段中所包含词语的逆文档频率IDF值确定的。

10.根据权利要求1-8中任一项所述的方法,其特征在于,所述向量数据库中还包括:多个预先生成的干预问题的语义向量以及每个干预问题对应的答复文本;

所述获取用户提出的问题之后,所述方法还包括:

对所述问题进行词向量转换,获得对应的语义向量;

计算所述问题对应的语义向量,与每个干预问题的语义向量之间的相似度;

在确定目标干预问题的语义向量与所述问题对应的语义向量之间的相似度大于第二预设阈值时,基于所述目标干预问题对应的答复文本,答复所述问题。

11.一种问题答复装置,其特征在于,包括:

获取单元,用于获取用户提出的问题;

处理单元,用于依据预设的多种检索策略,分别对所述问题进行处理,获得每种检索策略对应的检索用词;

检索单元,用于在预先建立的向量数据库中,基于获得的每种检索策略对应的检索用词,采用相应的检索策略进行检索,获得检索内容,所述向量数据库中至少包括多个文档及对应的文档向量,所述文档向量是对文档进行词向量转换得到的;

答复单元,用于基于所述检索内容答复所述问题。

12.一种电子设备,其特征在于,包括:至少一个处理器、至少一个存储器以及存储在所述存储器中的计算机程序指令,当所述计算机程序指令被所述处理器执行时实现如权利要求1-10中任一项所述的方法。

13.一种计算机可读存储介质,其上存储有计算机程序指令,其特征在于,当所述计算机程序指令被处理器执行时实现如权利要求1-10中任一项所述的方法。

一种问题答复方法、装置、设备及介质

技术领域

[0001] 本申请涉及人工智能领域,尤其是涉及一种问题答复方法、装置、设备及介质。

背景技术

[0002] 随着人工智能技术的发展,各种各样的智能产品,如智能客服、智能机器人等得到广泛的应用,此类智能产品中配置的对话系统(或者问答系统),可以与用户进行对话,例如,通过语音或者文字的方式,接收用户提出的问题,并给出相应的答复。

[0003] 目前,传统的问答系统,主要是基于规则或关键词匹配的方案,具体地,通过事先定义一系列规则或关键词,获取到用户提出的问题之后,利用单一任务的小模型,对用户的意图进行理解,然后基于用户的意图,通过关键词或者规则匹配相关的文档内容,从而基于匹配到的文档内容回答用户提出的问题。

[0004] 上述方案在实际应用中,存在以下缺点:1)无法理解文档的语义和上下文信息,导致回答的准确性不高;2)只能基于事先定义的规则或关键词进行匹配,无法处理复杂问题和新的问题;3)需要大量的人工整理并定义规则和关键字。

发明内容

[0005] 本申请实施例提供一种问题答复方法、装置、设备及介质,用以提高问题答复的准确性,同时无需事先定义规则和关键词,节省人力,且能够处理复杂问题和新的问题。

[0006] 第一方面,本申请实施例提供一种问题答复方法,包括:

[0007] 获取用户提出的问题;

[0008] 依据预设的多种检索策略,分别对所述问题进行处理,获得每种检索策略对应的检索用词;

[0009] 在预先建立的向量数据库中,基于获得的每种检索策略对应的检索用词,采用相应的检索策略进行检索,获得检索内容,所述向量数据库中至少包括多个文档及对应的文档向量,所述文档向量是对文档进行词向量转换得到的;

[0010] 基于所述检索内容答复所述问题。

[0011] 在一种可能的实施方式中,本申请实施例提供的上述方法中,所述依据预设的多种检索策略,分别对所述问题进行处理,获得每种检索策略对应的检索用词,包括以下操作的至少一种:

[0012] 若检索策略为语义检索,则对所述问题进行词向量转换,获得对应的语义向量,将获得的语义向量作为检索用词;

[0013] 若检索策略为关键词检索,则对所述问题进行分词处理,获得所述问题包含的词语,在获得的词语中选择第一预设数量的关键词作为检索用词;

[0014] 若检索策略为候选问题检索,则利用预先训练的大语言模型对所述问题进行处理,生成第二预设数量的候选问题,将生成的候选问题作为检索用词。

[0015] 在一种可能的实施方式中,本申请实施例提供的上述方法中,所述在预先建立的

向量数据库中,基于获得的每种检索策略对应的检索用词,采用相应的检索策略进行检索,获得检索内容,包括以下操作中的至少一种:

[0016] 若检索策略为语义检索,在获得所述问题对应的语义向量作为检索用词之后,在所述向量数据库中,检索与所述问题对应的语义向量的相似度大于第一预设阈值的文档向量,获得检索内容;

[0017] 若检索策略为关键词检索,在获得第一预设数量的关键词作为检索用词之后,在所述向量数据库中,检索与所述关键词相关联的文档,获得检索内容;

[0018] 若检索策略为候选问题检索,在获得候选问题作为检索用词之后,将所述候选问题作为待答复的问题,利用语义检索策略和关键词检索策略,在向量数据库中检索与所述候选问题相关联的文档,获得检索内容。

[0019] 在一种可能的实施方式中,本申请实施例提供的上述方法中,所述向量数据库中,还包括:扩展内容向量,所述扩展内容向量是对扩展内容进行词向量转换生成的,所述扩展内容是利用预先训练的大语言模型,基于文档生成的,所述扩展内容包括以下至少一种:所述文档的摘要、对所述文档进行提问生成的多个文档问题、以及对所述摘要进行提问生成的多个摘要问题;

[0020] 所述在所述向量数据库中,检索与所述问题对应的语义向量的相似度大于第一预设阈值的文档向量,获得检索内容,包括:

[0021] 在所述向量数据库中,检索与所述问题对应的语义向量的相似度大于第一预设阈值的文档向量和/或扩展内容向量,获得检索内容。

[0022] 在一种可能的实施方式中,本申请实施例提供的上述方法中,所述扩展内容向量,采用如下方式生成:

[0023] 在确定所述文档包含的字符数量大于预设数量阈值时,将所述文档切分为多个文档片段;

[0024] 若所述扩展内容包括所述文档的摘要,则利用所述大语言模型,生成所述多个文档片段对应的多级摘要;

[0025] 若所述扩展内容包括对所述文档进行提问生成的多个文档问题,则利用所述大语言模型,对每个文档片段进行提问,生成每个文档片段对应的文档问题;

[0026] 若所述扩展内容包括对所述文档的摘要进行提问生成的多个摘要问题,则利用所述大语言模型,生成所述多个文档片段对应的多级摘要,并对每个摘要提问,生成每个摘要对应的摘要问题;

[0027] 其中,所述多级摘要中每一级中包含至少一个摘要,每个摘要均是基于上一级中至少一个文档片段或者至少一个摘要生成的。

[0028] 在一种可能的实施方式中,本申请实施例提供的上述方法中,所述在确定所述文档包含的字符数量大于预设数量阈值时,将所述文档切分为多个文档片段,包括:

[0029] 基于预设切分方式,对所述文档进行初次切分,获得多个文本块;

[0030] 针对获得的每个文本块,执行以下操作进行二次切分,得到多个文档片段:提取所述文本块的语义特征,并基于所述语义特征,将所述文本块切分为文档片段。

[0031] 在一种可能的实施方式中,本申请实施例提供的上述方法中,所述向量数据库中还包括:与所述文档向量对应存储所述文档的元信息,与所述扩展内容向量对应存储所述

扩展内容的元信息；

[0032] 所述基于所述检索内容答复所述问题，包括：

[0033] 基于所述检索内容对应存储的元信息，确定所述检索内容归属的源文档或者源摘要；

[0034] 基于所述源文档或者所述源摘要生成提示词，将所述提示词发送至预先训练的大语言模型，以使所述大语言模型基于所述提示词答复所述问题。

[0035] 在一种可能的实施方式中，本申请实施例提供的上述方法中，所述基于所述检索内容对应存储的元信息，确定所述检索内容归属的源文档或者源摘要，包括：

[0036] 若所述检索内容为文档向量，则基于所述文档向量对应的元信息，确定所述文档向量归属的源文档；

[0037] 若所述检索内容为摘要向量，则基于所述摘要向量对应的元信息，确定所述摘要向量归属的源摘要；

[0038] 若所述检索内容为文档问题向量，则基于所述文档问题向量对应的元信息，确定所述文档问题向量归属的源文档；

[0039] 若所述检索内容为摘要问题向量，则基于所述摘要问题向量对应的元信息，确定所述摘要问题向量归属的源摘要。

[0040] 在一种可能的实施方式中，本申请实施例提供的上述方法中，所述在获得的词语中选择第一预设数量的关键词作为检索用词，包括：

[0041] 在获得的词语中，筛选出存在于预先确定的关键词汇表中的词语，作为关键词，其中，所述关键词汇表是对文档切分之后，基于切分得到的文档片段中所包含词语的逆文档频率(Inverse Document Frequency，IDF)值确定的。

[0042] 在一种可能的实施方式中，本申请实施例提供的上述方法中，所述向量数据库中还包括：多个预先生成的干预问题的语义向量以及每个干预问题对应的答复文本；

[0043] 所述获取用户提出的问题之后，所述方法还包括：

[0044] 对所述问题进行词向量转换，获得对应的语义向量；

[0045] 计算所述问题对应的语义向量，与每个干预问题的语义向量之间的相似度；

[0046] 在确定目标干预问题的语义向量与所述问题对应的语义向量之间的相似度大于第二预设阈值时，基于所述目标干预问题对应的答复文本，答复所述问题。

[0047] 第二方面，本申请实施例提供一种问题答复装置，包括：

[0048] 获取单元，用于获取用户提出的问题；

[0049] 处理单元，用于依据预设的多种检索策略，分别对所述问题进行处理，获得每种检索策略对应的检索用词；

[0050] 检索单元，用于在预先建立的向量数据库中，基于获得的每种检索策略对应的检索用词，采用相应的检索策略进行检索，获得检索内容，所述向量数据库中至少包括多个文档及对应的文档向量，所述文档向量是对文档进行词向量转换得到的；

[0051] 答复单元，用于基于所述检索内容答复所述问题。

[0052] 第三方面，本申请实施例提供一种电子设备，包括：至少一个处理器、至少一个存储器以及存储在存储器中的计算机程序指令，当计算机程序指令被处理器执行时实现如本申请实施例第一方面所提供的方法。

[0053] 第四方面,本申请实施例提供一种计算机可读存储介质,其上存储有计算机程序指令,当计算机程序指令被处理器执行时实现如本申请实施例第一方面所提供的方法。

[0054] 本申请实施例提供的问题答复方法、装置、设备及介质,获取用户提出的问题之后,依据预设的多种检索策略,如语义检索策略、关键词检索策略等,分别对问题进行处理,获得每种检索策略对应的检索用词,然后在预先建立的包括文档和文档向量的向量数据库中,基于获得的每种检索策略对应的检索用词,采用相应的检索策略进行检索,获得检索内容,最后基于检索内容答复用户提出的问题。

[0055] 本申请实施例,与现有技术中基于规则或关键词匹配的方案相比,在预先建立的向量数据库中存储文档和文档向量,在针对用户提出的问题进行检索时,一方面,基于关键词检索策略,可以从文档中检索出与问题中包含的关键词相关联的文档,另一方面,文档向量是对文档进行词向量转换得到的,其包含了文档的语义和上下文信息,利用语义检索策略检索时,基于用户所提问题对应的语义向量,在文档向量中进行检索,能够基于用户所提问题和文档的语义检索得到检索内容,提高检索结果的准确性,同时基于语义的检索能够处理复杂问题和新的问题,且无需人工事先定义规则或者关键词,节省人力。

[0056] 本申请的其它特征和优点将在随后的说明书中阐述,并且,部分地从说明书中变得显而易见,或者通过实施本申请而了解。本申请的目的和其他优点可通过在所写的说明书、权利要求书、以及附图中所特别指出的结构来实现和获得。

附图说明

[0057] 图1为本申请实施例中的一种应用场景的一个可选的示意图;

[0058] 图2为本申请实施例提供的问题答复方法的示意流程图;

[0059] 图3为本申请实施例提供的一种文档切分结果的示意图;

[0060] 图4为本申请实施例提供的另一文档切分结果的示意图;

[0061] 图5为本申请实施例提供的问题答复方法的具体流程的示意流程图;

[0062] 图6为本申请实施例提供的问题答复装置的结构示意图;

[0063] 图7为本申请实施例提供的一种电子设备的结构示意图;

[0064] 图8为本申请实施例提供的另一电子设备的结构示意图。

具体实施方式

[0065] 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技术方案的一部分实施例,而不是全部的实施例。基于本申请文件中记载的实施例,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他实施例,都属于本申请技术方案保护的范围。

[0066] 下面对本申请实施例的设计思想进行简要介绍:

[0067] 随着人工智能技术的发展,各种各样的智能产品,如智能客服、智能机器人等得到广泛的应用,此类智能产品中配置的对话系统(或者问答系统),可以与用户进行对话,例如,通过语音或者文字的方式,接收用户提出的问题,并给出相应的答复。

[0068] 目前,传统的问答系统,主要是基于规则或关键词匹配的方案,具体地,通过事先

定义一系列规则或关键词,获取到用户提出的问题之后,利用单一任务的小模型,对用户的意图进行理解,然后基于用户的意图,通过关键词或者规则匹配相关的文档内容,从而基于匹配到的文档内容回答用户提出的问题。

[0069] 上述方案在实际应用中,存在以下缺点:1)无法理解文档的语义和上下文信息,导致回答的准确性不高;2)只能基于事先定义的规则或关键词进行匹配,无法处理复杂问题和新的问题;3)需要大量的人工整理并定义规则和关键字。

[0070] 有鉴于此,本申请实施例提供一种问题答复方法、装置、设备及介质,获取用户提出的问题之后,依据预设的多种检索策略,如语义检索策略、关键词检索策略等,分别对问题进行处理,获得每种检索策略对应的检索用词,然后在预先建立的包括文档和文档向量的向量数据库中,基于获得的每种检索策略对应的检索用词,采用相应的检索策略进行检索,获得检索内容,最后基于检索内容答复用户提出的问题。

[0071] 本申请实施例,在预先建立的向量数据库中存储文档和文档向量,在针对用户提出的问题进行检索时,一方面,基于关键词检索策略,可以从文档中检索出与问题中包含的关键词相关联的文档,另一方面,文档向量是对文档进行词向量转换得到的,其包含了文档的语义和上下文信息,利用语义检索策略检索时,基于用户所提问题对应的语义向量,在文档向量中进行检索,能够基于用户所提问题和文档的语义检索得到检索内容,提高检索结果的准确性,同时基于语义的检索能够处理复杂问题和新的问题,且无需人工事先定义规则或者关键词,节省人力。

[0072] 以下结合说明书附图对本申请的优选实施例进行说明,应当理解,此处所描述的优选实施例仅用于说明和解释本申请,并不用于限定本申请,并且在不冲突的情况下,本申请中的实施例及实施例中的特征可以相互组合。

[0073] 如图1所示,其为本申请实施例的应用场景示意图。该应用场景图中包括多个智能产品中的任一智能产品110和多个服务器中的任一个服务器120。

[0074] 在本申请实施例中,智能产品110包括但不限于手机、电脑、智能机器人等产品;服务器120则是智能产品的后台服务端。服务器120可以是独立的物理服务器,也可以是多个物理服务器构成的服务器集群或者分布式系统,还可以是提供云服务、云数据库、云计算、云函数、云存储、网络服务、云通信、中间件服务、域名服务、安全服务、内容分发网络(Content Delivery Network,CDN)、以及大数据和人工智能平台等基础云计算服务的云服务器。

[0075] 需要说明的是,本申请实施例中的问题答复方法,可以由服务器120执行,也可以由智能产品110执行,本申请实施例对此不做限定。

[0076] 以在智能产品110中执行为例,智能产品110获取用户提出的问题之后,依据预设的多种检索策略,如语义检索策略、关键词检索策略等,分别对问题进行处理,获得每种检索策略对应的检索用词,然后在预先建立的包括文档和文档向量的向量数据库中,基于获得的每种检索策略对应的检索用词,采用相应的检索策略进行检索,获得检索内容,最后基于检索内容答复用户提出的问题。

[0077] 在一种可选的实施方式中,智能产品110与服务器120之间可以通过通信网络进行通信,通信网络是有线网络或无线网络。

[0078] 需要说明的是,图1所示只是举例说明,实际上智能产品和服务器的数量和通信方

式均不受限制,当服务器的数量为多个时,多个服务器可组成为一区块链,而服务器为区块链上的节点,在本申请实施例中不做具体限定。

[0079] 下面结合上述描述的应用场景,参考附图来描述本申请示例性实施方式提供的问题答复方法,需要注意的是,上述应用场景仅是为了便于理解本申请的精神和原理而示出,本申请的实施方式在此方面不受任何限制。

[0080] 如图2所示,其为本申请实施例中的一种问题答复方法的实施流程图,该方法的具体实施流程如下S201-S204:

[0081] S201,获取用户提出的问题。

[0082] 具体实施时,获取用户提出的问题,可以是接收用户以语音形式或者文字形式提出的问题,本申请实施例对此不做限定。

[0083] S202,依据预设的多种检索策略,分别对问题进行处理,获得每种检索策略对应的检索用词。

[0084] 其中,预设的多种检测策略,可以包括但不限于:语义检索策略,关键词检索策略以及候选问题检测策略等。

[0085] 具体实施时,若检索策略为语义检索,则对问题进行词向量转换,获得对应的语义向量,将获得的语义向量作为检索用词;若检索策略为关键词检索,则对问题进行分词处理,获得问题包含的词语,在获得的词语中选择第一预设数量的关键词作为检索用词;若检索策略为候选问题检索,则利用预先训练的大语言模型对问题进行处理,生成第二预设数量的候选问题,将生成的候选问题作为检索用词。其中,对问题进行词向量转换,可以使用现有的词向量转换模型,例如:word2vec等,本申请实施例对此不做限定。第一预设数量和第二预设数量均可以根据经验值设定,例如,第一预设数量为3,第二预设数量为5。

[0086] 需要说明的是,预先训练的大语言模型,可以利用现有的大语言模型,也可以单独训练大语言模型,本申请实施例对此不做限定。

[0087] S203,在预先建立的向量数据库中,基于获得的每种检索策略对应的检索用词,采用相应的检索策略进行检索,获得检索内容,向量数据库中至少包括多个文档及对应的文档向量,文档向量是对文档进行词向量转换得到的。

[0088] 具体实施时,若检索策略为语义检索,在获得问题对应的语义向量作为检索用词之后,在向量数据库中,检索与问题对应的语义向量的相似度大于第一预设阈值的文档向量,获得检索内容,其中,第一预设阈值可以根据经验设定,例如,第一预设阈值设置为0.9或者0.95等。

[0089] 若检索策略为关键词检索,在获得第一预设数量的关键词作为检索用词之后,在向量数据库中,检索与关键词相关联的文档,例如,检索包含关键词的文档,获得检索内容。

[0090] 若检索策略为候选问题检索,在获得候选问题作为检索用词之后,将候选问题作为待答复的问题,利用语义检索策略和关键词检索策略,在向量数据库中检索与候选问题相关联的文档,获得检索内容。

[0091] 需要说明的是,检索策略为候选问题检索时,在获得候选问题作为检索用词之后,一方面,需要对候选问题进行词向量转换,得到候选问题对应的语义向量,然后以候选问题对应的语义向量为检索用词,利用语义检索策略进行检索;另一方面,需要对候选问题进行分词处理,得到候选问题包含的关键词,然后以候选问题包含的关键词为检索用词,利用关

键词检索策略进行检索。

[0092] S204,基于检索内容答复问题。

[0093] 具体实施时,由于不同的检索策略检索得到的检索内容可能不同,也可能相同,具体在基于检索内容答复问题时,可以在每种检索策略的检索结果中选取一定数量的检索内容,基于选取的检索内容答复问题;也可以选取多种检索策略的检索结果中公共的检索内容,基于确定出的公共检索内容答复问题,本申请实施例对此不做限定。

[0094] 下面结合具体示例,分别对本申请实施例中提到的语义检索策略和关键词检索策略分别进行详细说明。

[0095] 一、语义检索策略。

[0096] 此种检索策略主要是以问题对应的语义向量为检索用词,在向量数据库中包括的向量中进行检索,向量数据库中除包括文档向量之外,还可以包括其他内容,下面将展开说明。

[0097] 为了便于在向量数据库中检索内容,向量数据库中,还可以包括:扩展内容向量,扩展内容向量是对扩展内容进行词向量转换生成的,扩展内容是利用预先训练的大语言模型,基于文档生成的,扩展内容包括以下至少一种:文档的摘要、对文档进行提问生成的多个文档问题、以及对摘要进行提问生成的多个摘要问题。其中,对扩展内容进行词向量转换,同样可以使用现有的词向量转换模型,例如:word2vec等,本申请实施例对此不做限定。

[0098] 实际应用中,由于大语言模型通常有输入字符的限制,因此,若文档包含的字符数量大于预设数量阈值,则可以将文档切分为多个文档片段,其中,预设数量阈值可以根据大语言模型支持的最大输入字符数量设定,例如,大语言模型支持的最大输入字符数量为100,预设数量阈值可以设置为100或者99等。

[0099] 具体将文档切分为多个文档片段时,可以基于预设切分方式,对文档进行初次切分,获得多个文本块,然后针对获得的每个文本块,执行以下操作进行二次切分,得到多个文档片段:提取文本块的语义特征,并基于语义特征,将文本块切分为文档片段,将得到的多个文档片段作为切分结果。

[0100] 基于预设切分方式,对文档进行初次切分,其目的是将文档切分为语言模型支持的最大输入长度,因此,文本块中包含的字符数量可以根据语言模型支持的最大输入长度设定,其可以小于或者等于语言模型支持的最大输入长度。例如,语言模型的最大输入长度为100个字符,则文本块中包含的字符数量可以设置为100,也可以设置为95。

[0101] 其中,预设切分方式,可以是按照段落进行切分,当然,预设切分方式,也可以是其他方式,例如按照章节进行切分等。

[0102] 具体实施时,在基于预设切分方式,对文档进行初次切分,获得多个文本块之后,还可以记录多个文本块在文档中的顺序,作为预先记录的顺序,以避免文本块顺序混乱,且可以便于后续对文档片段进行合并时的排序。

[0103] 具体实施时,在对文档进行初次切分得到多个文本块之后,可以对每个文本块进行二次切分,将每个文本块切分为至少一个文档片段,具体在将文本块切分为文档片段时,可以将文本块输入预先配置的语义分割模型中,提取文本块的语义特征,然后基于文本块的语义特征,将文本块切分为语义完整的文档片段。

[0104] 需要说明的是,预先配置的语义分割模型,可以采用如下方式训练生成:首先获取

训练文档数据,并对训练文档数据进行预处理和清洗,保证训练文档数据的质量,然后可以以训练文档数据为输入和输出,训练基础模型,最后以训练文档数据为输入,以训练文档数据切分后的结果为输出,对基础模型中的参数进行微调,直至模型的损失函数满足设定的收敛条件,此时,将参数调整后的基础模型作为语义分割模型。

[0105] 在本申请其他实施例中,预先配置的语义分割模型,也可以采用现有的开放接口的模型,本申请实施例对此不做限定。

[0106] 具体实施时,在对获得的多个文本块进行二次切分时,采用预先记录的顺序进行二次切分。

[0107] 具体实施时,在将多个文本块切分为文档片段之后,本申请实施例可以将切分得到的文档片段作为切分结果,也可以对部分相邻的文档片段进行合并处理,得到合并后的目标文档片段,然后将合并得到的目标文档片段和未合并的文档片段,作为对文档的切分结果。

[0108] 需要说明的是,由于合并之前各个文档片段均是语义完整的文档片段,因此,合并处理得到的目标文档片段仍然是语义完整的文档片段,则对部分相邻的文档片段进行合并处理,不仅能够使得单个目标文档片段尽可能达到语言模型的最大输入长度,而且合并后的目标文档片段仍是语义完整的文档片段,保证切分的质量。

[0109] 实际应用中,在将多个文本块切分为文档片段之后,可以将多个文档片段,依照预先记录的顺序排列,得到排列结果,然后按照预先设定的文档片段长度要求,将排列结果中部分相邻的文档片段合并为一个目标文档片段,得到至少一个目标文档片段,将合并得到的目标文档片段和未合并的文档片段,作为对文档的切分结果。其中,预先设定的文档片段长度要求,可以根据语言模型支持的最大输入长度设定,预先设定的文档片段长度可以小于或等于语言模型支持的最大输入长度。

[0110] 在一个示例中,如图3所示,假设待切分的文档X经过初次切分,得到文本块A、文本块B、文本块C和文本块D,文本块A经过二次切分,得到文档片段A1、A2、A3;文本块B经过二次切分,得到文档片段B1、B2;文本块C经过二次切分,得到文档片段C1、C2、C3;文本块D经过二次切分,得到文档片段D1、D2;此时可以将文档片段A1、A2、A3、B1、B2、C1、C2、C3、D1和D2作为切分结果。

[0111] 在另一示例中,如图4所示,仍沿用上述示例的切分方案,还可以对相邻的部分文档片段进行合并处理,例如,将文档片段A3和文档片段B1合并为目标文档片段E1,将文档片段B2和文档片段C1合并为目标文档片段E2,则此时可以将文档片段A1、A2、E1、E2、C2、C3、D1和D2作为切分结果。

[0112] 在将文档切分为多个文档片段之后,在利用对话系统中的大语言模型,生成与文档相关联的扩展内容时,具体包括以下情况:

[0113] 情况1、扩展内容包括文档的摘要,则利用大语言模型,生成多个文档片段对应的多级摘要,其中,多级摘要中每一级中包含至少一个摘要,每个摘要均是基于上一级中至少一个文档片段或者至少一个摘要生成的。

[0114] 具体实施时,基于上一级中至少一个文档片段或者至少一个摘要生成下一级摘要时,所选择的文档片段数量或者摘要数量,可以基于每个文档片段或者摘要的长度,以及大语言模型支持的最大输入字符数量灵活设置,本申请实施例对此不做限定。

[0115] 在一个示例中,假设文档C,被切分为多个文档片段C1、C2、C3、C4和C5,则在生成文档片段对应的摘要时,第一级摘要可以分别生成(C1、C2、C3)、(C2、C3、C4)和(C3、C4、C5)对应的摘要,得到S1-1、S1-2和S1-3三个第一级摘要,第二级摘要可以分别生成(S1-1、S1-2)和(S1-2、S1-3)对应的摘要,得到S2-1和S2-2两个二级摘要,第三级摘要可以生成(S2-1、S2-2)对应的摘要,得到一个第三级摘要S3-1。

[0116] 在另一示例中,仍假设文档C,被切分为多个文档片段C1、C2、C3、C4和C5,则在生成文档片段对应的摘要时,第一级摘要可以分别生成(C1、C2)、(C2、C3)、(C3、C4)和(C4、C5)对应的摘要,得到X1-1、X1-2、X1-3和X1-4四个第一级摘要,第二级摘要可以分别生成(X1-1、X1-2)、(X1-2、X1-3)和(X1-3、X1-4)对应的摘要,得到X2-1、X2-2和X2-3三个二级摘要,第三级摘要可以生成(X2-1、X2-2)、(X2-2、X2-3)对应的摘要,得到两个第三级摘要X3-1和X3-2,第四级摘要可以生成(X3-1、X3-2)对应的摘要,得到一个第四级摘要X4-1。

[0117] 需要说明的是,文档切分为多个文档片段的情况下,针对多个文档片段生成多级摘要,可以在存储多级摘要对应的摘要向量之后,获取到用户提出的问题,基于用户提出的问题对应的语义向量,在摘要向量中进行检索时,减少检索的次数,提高检索效率。

[0118] 情况2、若扩展内容包括对文档进行提问生成的多个文档问题,则利用大语言模型,对每个文档片段进行提问,生成每个文档片段对应的文档问题,其中,每个文档片段对应的文档问题,可以是一个,也可以是多个,本申请实施例对此不做限定。

[0119] 情况3、若扩展内容包括对文档的摘要进行提问生成的多个摘要问题,则利用大语言模型,生成多个文档片段对应的多级摘要,并对每个摘要提问,生成每个摘要对应的摘要问题,其中,每个摘要对应的摘要问题,可以是一个,也可以是多个,本申请实施例对此不做限定。

[0120] 具体利用大语言模型,生成多个文档片段对应的多级摘要的方式,可以采用上述情况1中介绍的方式,此处不再赘述。

[0121] 实际应用中,为了方便对文档进行归类,确定文档所属的文档集合,本申请实施例,还可以获取文档的元信息,并基于文档的元信息,生成扩展内容的元信息,然后在向量数据库中,与文档向量对应存储文档的元信息,与扩展内容向量对应存储扩展内容的元信息,并建立索引。

[0122] 其中,元信息可以包括但不限于:文本内容、文本的信息摘要算法md5值、文档作者、文档标题、文档标识、文档来源、来源类型、摘要相关信息等信息。

[0123] 另外,为了支持多语种,本申请实施例针对向量数据库中存储的文档,还可以利用大语言模型,将文档和扩展内容分别转换为同一其他语种,并分别对转换后的文档和转换后的扩展内容进行词向量转换,获得转换后文档对应的文档向量和转换后扩展内容对应的扩展内容向量,然后在向量数据库中存储转换后文档对应的文档向量和转换后扩展内容对应的扩展内容向量。

[0124] 需要说明的是,可以将文档和扩展内容转换为多个语种的版本,在每次转换时,可以将文档和扩展内容分别转换为同一语种。

[0125] 向量数据库中包括扩展内容向量的情况下,具体实施时,获得用户提出的问题对应的语义向量之后,可以在向量数据库中,检索与问题对应的语义向量的相似度大于第一预设阈值的文档向量和/或扩展内容向量作为备选向量,并将检索到的备选向量作为获得

的检索内容。

[0126] 实际应用中,如果扩展内容向量包括:文档的摘要对应的摘要向量,对问题进行提问生成的文档问题对应的文档问题向量,以及对文档的摘要进行提问生成的摘要问题对应的摘要问题向量,则确定备选向量时,可以在文档向量中确定第一批备选向量,在摘要向量中确定第二批备选向量,在文档问题中确定第三批备选向量,在摘要问题向量中确定第四批备选向量,然后将确定出的第一批备选向量、第二批备选向量、第三批备选向量、以及第四批备选向量作为确定出的备选向量。

[0127] 具体实施时,若确定出一个备选向量,则将确定出的备选向量作为检索内容,若确定出的备选向量为多个,则基于预设策略在多个备选向量中选择出第三预设数量个备选向量作为检索内容。

[0128] 具体基于预设策略在多个备选向量中,选择第三预设数量个备选向量时,预设策略,可以包括以下至少一种:

[0129] 策略1、基于每个备选向量与问题对应的语义向量之间的相似度,对确定出的备选向量进行降序排列,得到排列结果,在排列结果中选取前第三预设数量个备选向量。

[0130] 策略2、在类型为文档向量的备选向量中选择第一数量的备选向量,类型为扩展内容向量的备选向量中选择第二数量的备选向量,第一数量与第二数量之和为第三预设数量。

[0131] 在一个示例中,假设扩展内容向量包括:文档的摘要对应的摘要向量,对问题进行提问生成的文档问题对应的文档问题向量,以及对文档的摘要进行提问生成的摘要问题对应的摘要问题向量,第三预设数量为6,则可以从类型为文档向量的备选向量中选择2个备选向量,从类型为摘要向量的备选向量中选择2个备选向量,从类型为文档问题向量的备选向量中选择1个备选向量,从类型为摘要问题向量的备选向量中选择1个备选向量,一共选取6个备选向量。

[0132] 策略3、若备选向量对应存储有元信息,则确定任一备选向量对应的元信息,基于元信息确定备选向量归属的文档所属的文档集合,在由文档集合中文档生成的备选向量中,选择第三预设数量个备选向量。

[0133] 具体实施时,根据备选向量对应的元信息,确定备选向量归属的文档所属的文档集合,然后在文档集合中文档生成的备选向量中,选择第三预设数量个备选向量,可以缩小备选向量的选择范围,提高备选向量的选择准确性。

[0134] 例如,某一备选向量对应的元信息中文档作者为张三,则可以基于元信息确定文档作者为张三的所有文档,组成文档集合,进而在所有基于文档作者为张三的文档,相关联的备选向量中,选择第三预设数量个备选向量。

[0135] 具体实施时,在向量数据库中还包括:与文档向量对应存储文档的元信息,与扩展内容向量对应存储扩展内容的元信息时,基于检索内容答复问题,包括:基于检索内容对应存储的元信息,确定检索内容归属的源文档或者源摘要;基于源文档或者源摘要生成提示词,将提示词发送至预先训练的大语言模型,以使大语言模型基于提示词答复问题。

[0136] 具体来说,元信息可以包括但不限于:文本内容、文本的信息摘要算法md5值、文档作者、文档标题、文档标识、文档来源、来源类型、摘要相关信息等信息。

[0137] 若检索内容为文档向量,则基于文档向量对应的元信息,确定文档向量归属的源

文档时,可以直接基于元信息中的文本内容,确定检索内容归属的源文档。

[0138] 若检索内容为摘要向量,则基于摘要向量对应的元信息,确定摘要向量归属的源摘要时,同样可以直接基于元信息中的文本内容,确定检索内容归属的源摘要。

[0139] 若检索内容为文档问题向量,则基于文档问题向量对应的元信息,确定文档问题向量归属的源文档,可以基于元信息中的文档作者、文档标题或者文档标识等,确定检索内容归属的源文档。

[0140] 若检索内容为摘要问题向量,则基于摘要问题向量对应的元信息,确定摘要问题向量归属的源摘要时,可以基于元信息中与摘要相关的信息,确定检索内容归属的源摘要。

[0141] 二、关键词检索策略。

[0142] 此种检索策略中,主要是以从问题包含的词语中选出的关键词为检索用词,在向量数据库中包括的文档中进行检索。

[0143] 此种检索策略,在对问题进行分词处理之后,在获得的词语中选择第一预设数量的关键词作为检索用词,可以在获得的词语中,筛选出存在于预先确定的关键词汇表中的词语,作为关键词,其中,关键词汇表是对文档切分之后,基于切分得到的文档片段中所包含词语的IDF值确定的。

[0144] 其中,预先确定的关键词汇表,具体可以采用如下方式确定:对向量数据库中包含的文档进行切分,得到多个文档片段,计算每个文档片段中所包含的词语的IDF值,选取IDF值满足预设条件的词语,作为向量数据库中所有文档对应的关键词汇表。

[0145] 预设条件可以是IDF值大于设定值,或者IDF值在所有词语的IDF降序排列结果中位于前n%,n的取值为大于等于1且小于等于100的自然数,当然,本申请其他实施例中还可以是其他条件,此处不用于具体限定。

[0146] 具体实施时,对问题进行分词处理,可以使用现有的方式,本申请实施例对此不做限定。在获得问题包含的词语之后,在获得的词语中,使用预先确定的关键词汇表,对问题中包含的词语进行过滤,筛选出存在于预先确定的关键词汇表中的关键词,进而基于关键词在向量数据库包含的文档中搜索,获得检索内容。

[0147] 需要说明的是,在本申请其他实施例中,还可以通过维护停用词库的方式,对问题包含的词语进行过滤,过滤一些无关紧要的停用词,留下重要的词语作为关键词。

[0148] 在基于关键词检索到文档之后,可以对文档进行排序,为了保证重要的词语评分更高,利用预设算法,基于文档中包含的关键词,确定每个文档对应的评分值,可以先对每个关键词的IDF值进行预处理,例如,计算IDF值的平方值,然后将每个文档中包含的所有关键词的IDF值的平方值之和,作为每个文档对应的评分值。当然,在本申请其他实施例中,预设算法也可以是其他算法,本申请实施例对此不做限定。

[0149] 另外,针对问题中可能包含同义词的情况,本申请实施例,可以先确定每个词语的同义词,然后将同义词也作为关键词语进行检索,具体来说:在获得问题包含的词语之后,可以对获得的词语进行词向量转换,获得对应的词向量,进而基于获得的词向量,分别确定每个词语的同义词,其中,每个词语的同义词的词向量与每个词语对应的词向量相似度大于预设相似度阈值;然后在获得的问题包含的词语以及每个词语的同义词中,筛选出存在于预先确定的关键词汇表中的关键词。其中,预设相似度阈值可以根据经验值设定,例如:0.9或0.95等。

[0150] 另外,具体实施时,若确定向量数据库中的文档更新,例如,新增文档、删除文档或者文档版本更新,本申请实施例可以更新对应的关键词汇表。

[0151] 实际应用中,为了避免用户所提出的问题不准确,本申请实施例还可以预先生成一些干预问题以及对应的答复文本,然后存储在向量数据库中,也即向量数据库中还包括:多个预先生成的干预问题的语义向量以及每个干预问题对应的答复文本。

[0152] 如此,获取用户提出的问题之后,本申请实施例还可以对问题进行词向量转换,获得对应的语义向量,计算问题对应的语义向量,与每个干预问题的语义向量之间的相似度,在确定目标干预问题的语义向量与问题对应的语义向量之间的相似度大于第二预设阈值时,基于目标干预问题对应的答复文本,答复问题。其中,第二预设阈值可以根据经验值设定,例如,第二预设阈值为0.9或者0.97等。

[0153] 具体实施时,在确定多个干预问题中每个干预问题的语义向量,与问题对应的语义向量之间的相似度均小于或等于第二预设阈值时,利用预先训练的大语言模型答复问题。

[0154] 需要说明的是,若有多个干预问题的语义向量与问题对应的语义向量之间的相似度大于第二预设阈值,则可以将最大相似度对应的干预问题确定为目标干预问题,当然,本申请其他实施例中,也可以随机选择一个干预问题作为目标干预问题,本申请实施例对此不做限定。

[0155] 实际应用中,基于目标干预问题对应的答复文本,答复问题之后,本申请实施例还可以呈现用于请求用户针对问题答复进行评价的交互界面,例如,交互界面中可以展示“满意”和“不满意”两个按键供用户选择,再例如,交互界面中可以展示1-10分的打分选项供用户选择评分值。

[0156] 为了扩展干预问题,在根据用户在交互界面的选择操作,获取用户针对问题答复的评价之后,若确定用户针对问题答复的评价满足第一预设条件,将问题作为干预问题进行存储,并将针对问题的答复存储为问题对应的答复文本。其中,第一预设条件可以是用户针对问题答复的评价为满意,或者评分值大于预设分值(根据经验值设定,例如:10分制下,预设分值为8分)。

[0157] 当然,若确定用户针对问题答复的评价满足第二预设条件,记录问题和目标干预问题,以请求其他用户对目标干预问题进行更新和优化。其中,其他用户可以是维护人员等,第二预设条件可以是用户针对问题答复的评价为不满意,或者评分值小于预设分值。

[0158] 下面结合图5对本申请实施例提供的问题答复方法的具体实施过程进行详细说明。

[0159] 如图5所示,本申请实施例提供的问题答复方法,整体可以分为两部分:离线部分和在线部分。

[0160] 其中,离线部分对文档进行解析,并在文档包含的字符数量较多时,对文档进行切分,将文档切分为文档片段。需要说明的是,文档可以是用户上传的文档,也可以是预置的一些文档,本申请实施例对此不做限定。

[0161] 在对文档进行切分获得文档片段之后,可以通过大语言模型,生成文档的摘要,并分别对文档片段和文档摘要进行提问,获得关联问题,其中,关联问题包括文档问题和摘要问题。

[0162] 本申请实施例还可以利用词向量模型,对文档片段、文档摘要和关联问题分别进行词向量转换,得到文档片段对应的文档向量,文档摘要对应的摘要向量,以及关联问题对应的关联问题向量,并将文档片段、文档向量、摘要向量以及关联问题向量均存储到预先建立的向量数据库中。

[0163] 当然,在本申请其他实施例中,还可以将文档片段、文档摘要和关联问题转换为其他语种,并计算其他语种下文档片段、文档摘要和关联问题对应的向量,存储到向量数据库中;还可以确定文档片段、文档摘要和关联问题的元信息,并存储到向量数据库中。

[0164] 而在线部分,是在获取到用户提出的问题之后,利用多种检索策略,分别对问题进行处理,获得每种检索策略对应的检索用词。

[0165] 具体实施时,若检索策略为语义检索,则对问题进行词向量转换,获得对应的语义向量,将获得的语义向量作为检索用词;若检索策略为关键词检索,则对问题进行分词处理,获得问题包含的词语,在获得的词语中选择第一预设数量的关键词作为检索用词;若检索策略为候选问题检索,则利用预先训练的大语言模型对问题进行处理,生成第二预设数量的候选问题,将生成的候选问题作为检索用词。

[0166] 针对以上三种检索策略得到的检索用词,采用对应的检索策略,分别在向量数据库中进行检索,可以获得检索内容,检索内容中可以包含文档片段、文档摘要、以及关联问题中的一种或多种。

[0167] 在得到检索内容之后,可以基于检索内容生成提示词,并将提示词发送至预先训练的大语言模型,由大语言模型针对用户提出的问题进行答复。

[0168] 基于同样的发明构思,如图6所示,本申请实施例提供一种问题答复装置,包括:

[0169] 获取单元601,用于获取用户提出的问题;

[0170] 处理单元602,用于依据预设的多种检索策略,分别对问题进行处理,获得每种检索策略对应的检索用词;

[0171] 检索单元603,用于在预先建立的向量数据库中,基于获得的每种检索策略对应的检索用词,采用相应的检索策略进行检索,获得检索内容,向量数据库中至少包括多个文档及对应的文档向量,文档向量是对文档进行词向量转换得到的;

[0172] 答复单元604,用于基于检索内容答复问题。

[0173] 在一种可能的实施方式中,处理单元602具体用于:

[0174] 若检索策略为语义检索,则对问题进行词向量转换,获得对应的语义向量,将获得的语义向量作为检索用词;

[0175] 若检索策略为关键词检索,则对问题进行分词处理,获得问题包含的词语,在获得的词语中选择第一预设数量的关键词作为检索用词;

[0176] 若检索策略为候选问题检索,则利用预先训练的大语言模型对问题进行处理,生成第二预设数量的候选问题,将生成的候选问题作为检索用词。

[0177] 在一种可能的实施方式中,检索单元603具体用于:

[0178] 若检索策略为语义检索,在获得问题对应的语义向量作为检索用词之后,在向量数据库中,检索与问题对应的语义向量的相似度大于第一预设阈值的文档向量,获得检索内容;

[0179] 若检索策略为关键词检索,在获得第一预设数量的关键词作为检索用词之后,在

向量数据库中,检索与关键词相关联的文档,获得检索内容;

[0180] 若检索策略为候选问题检索,在获得候选问题作为检索用词之后,将候选问题作为待答复的问题,利用语义检索策略和关键词检索策略,在向量数据库中检索与候选问题相关联的文档,获得检索内容。

[0181] 在一种可能的实施方式中,向量数据库中,还包括:扩展内容向量,扩展内容向量是对扩展内容进行词向量转换生成的,扩展内容是利用预先训练的大语言模型,基于文档生成的,扩展内容包括以下至少一种:文档的摘要、对文档进行提问生成的多个文档问题、以及对摘要进行提问生成的多个摘要问题;

[0182] 检索单元603具体用于:

[0183] 在向量数据库中,检索与问题对应的语义向量的相似度大于第一预设阈值的文档向量和/或扩展内容向量,获得检索内容。

[0184] 在一种可能的实施方式中,处理单元602预先采用如下方式生成扩展内容向量:

[0185] 在确定文档包含的字符数量大于预设数量阈值时,将文档切分为多个文档片段;

[0186] 若扩展内容包括文档的摘要,则利用大语言模型,生成多个文档片段对应的多级摘要;

[0187] 若扩展内容包括对文档进行提问生成的多个文档问题,则利用大语言模型,对每个文档片段进行提问,生成每个文档片段对应的文档问题;

[0188] 若扩展内容包括对文档的摘要进行提问生成的多个摘要问题,则利用大语言模型,生成多个文档片段对应的多级摘要,并对每个摘要提问,生成每个摘要对应的摘要问题;

[0189] 其中,多级摘要中每一级中包含至少一个摘要,每个摘要均是基于上一级中至少一个文档片段或者至少一个摘要生成的。

[0190] 在一种可能的实施方式中,处理单元602具体用于:

[0191] 基于预设切分方式,对文档进行初次切分,获得多个文本块;

[0192] 针对获得的每个文本块,执行以下操作进行二次切分,得到多个文档片段:提取文本块的语义特征,并基于语义特征,将文本块切分为文档片段。

[0193] 在一种可能的实施方式中,向量数据库中还包括:与文档向量对应存储文档的元信息,与扩展内容向量对应存储扩展内容的元信息;

[0194] 答复单元604具体用于:

[0195] 基于检索内容对应存储的元信息,确定检索内容归属的源文档或者源摘要;

[0196] 基于源文档或者源摘要生成提示词,将提示词发送至预先训练的大语言模型,以使大语言模型基于提示词答复问题。

[0197] 在一种可能的实施方式中,答复单元604具体用于:

[0198] 若检索内容为文档向量,则基于文档向量对应的元信息,确定文档向量归属的源文档;

[0199] 若检索内容为摘要向量,则基于摘要向量对应的元信息,确定摘要向量归属的源摘要;

[0200] 若检索内容为文档问题向量,则基于文档问题向量对应的元信息,确定文档问题向量归属的源文档;

[0201] 若检索内容为摘要问题向量,则基于摘要问题向量对应的元信息,确定摘要问题向量归属的源摘要。

[0202] 在一种可能的实施方式中,处理单元602具体用于:

[0203] 在获得的词语中,筛选出存在于预先确定的关键词汇表中的词语,作为关键词,其中,关键词汇表是对文档切分之后,基于切分得到的文档片段中所包含词语的逆文档频率IDF值确定的。

[0204] 在一种可能的实施方式中,向量数据库中还包括:多个预先生成的干预问题的语义向量以及每个干预问题对应的答复文本;

[0205] 处理单元602还用于:对问题进行词向量转换,获得对应的语义向量,计算问题对应的语义向量,与每个干预问题的语义向量之间的相似度;

[0206] 答复单元604还用于:在确定目标干预问题的语义向量与问题对应的语义向量之间的相似度大于第二预设阈值时,基于目标干预问题对应的答复文本,答复问题。

[0207] 与上述方法实施例基于同一发明构思,本申请实施例中还提供了一种电子设备。该电子设备可以对用户提出的问题进行答复。在一种实施例中,该电子设备可以是服务器,如图1所示的服务器120。在该实施例中,电子设备的结构可以如图7所示,包括存储器701,通讯模块703以及一个或多个处理器702。

[0208] 存储器701,用于存储处理器702执行的计算机程序。存储器701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,以及运行即时通讯功能所需的程序等;存储数据区可存储各种即时通讯信息和操作指令集等。

[0209] 存储器701可以是易失性存储器(volatile memory),例如随机存取存储器(random-access memory,RAM);存储器701也可以是非易失性存储器(non-volatile memory),例如只读存储器,快闪存储器(flash memory),硬盘(hard disk drive,HDD)或固态硬盘(solid-state drive,SSD);或者存储器701是能够用于携带或存储具有指令或数据结构形式的期望的程序代码并能够由计算机存取的任何其他介质,但不限于此。存储器701可以是上述存储器的组合。

[0210] 处理器702,可以包括一个或多个中央处理单元(central processing unit,CPU)或者为数字处理单元等等。处理器702,用于调用存储器701中存储的计算机程序时实现上述问题答复方法。

[0211] 通讯模块703用于与终端设备和其他服务器进行通信。

[0212] 本申请实施例中不限定上述存储器701、通讯模块703和处理器702之间的具体连接介质。本公开实施例在图7中以存储器701和处理器702之间通过总线704连接,总线704在图7中以粗线表示,其它部件之间的连接方式,仅是进行示意性说明,并不引以为限。总线704可以分为地址总线、数据总线、控制总线等。为便于表示,图7中仅用一条粗线表示,但并不表示仅有一根总线或一种类型的总线。

[0213] 存储器701中存储有计算机存储介质,计算机存储介质中存储有计算机可执行指令,计算机可执行指令用于实现本申请实施例的问题答复方法。处理器702用于执行上述的问题答复方法,如图2所示。

[0214] 在另一种实施例中,电子设备也可以是其他电子设备,如图1所示的智能产品110。在该实施例中,电子设备的结构可以如图8所示,包括:通信组件810、存储器820、显示单元

830、摄像头840、传感器850、音频电路860、蓝牙模块870、处理器880等部件。

[0215] 通信组件810用于与服务器进行通信。在一些实施例中,可以包括电路无线保真(Wireless Fidelity,WiFi)模块,WiFi模块属于短距离无线传输技术,电子设备通过WiFi模块可以帮助用户收发信息。

[0216] 存储器820可用于存储软件程序及数据。处理器880通过运行存储在存储器820的软件程序或数据,从而执行智能产品110的各种功能以及数据处理。存储器820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,例如至少一个磁盘存储器件、闪存器件、或其他易失性固态存储器件。存储器820存储有使得智能产品110能运行的操作系统。本申请中存储器820可以存储操作系统及各种应用程序,还可以存储执行本申请实施例问题答复方法的代码。

[0217] 显示单元830还可用于显示由用户输入的信息或提供给用户的信息以及智能产品110的各种菜单的图形用户界面(graphical user interface,GUI)。具体地,显示单元830可以包括设置在智能产品110正面的显示屏832。其中,显示屏832可以采用液晶显示器、发光二极管等形式来配置。显示单元830可以用于呈现本申请实施例中的图像或文字。

[0218] 显示单元830还可用于接收输入的数字或字符信息,产生与智能产品110的用户设置以及功能控制有关的信号输入,具体地,显示单元830可以包括设置在智能产品110正面的触控屏831,可收集用户在其上或附近的触摸操作,例如点击按钮,拖动滚动框等。

[0219] 其中,触控屏831可以覆盖在显示屏832之上,也可以将触控屏831与显示屏832集成而实现智能产品110的输入和输出功能,集成后可以简称触摸显示屏。本申请中显示单元830可以显示用户提出的问题及针对问题给出的答复。

[0220] 摄像头840可用于捕获静态图像,用户可以将摄像头840拍摄的图像通过客户端发送给聊天对方的用户。摄像头840可以是一个,也可以是多个。物体通过镜头生成光学图像投射到感光元件。感光元件可以是电荷耦合器件(charge coupled device,CCD)或互补金属氧化物半导体(complementary metal-oxide-semiconductor,CMOS)光电晶体管。感光元件把光信号转换成电信号,之后将电信号传递给处理器870转换成数字图像信号。

[0221] 智能产品还可以包括至少一种传感器850,比如加速度传感器851、距离传感器852、指纹传感器853、温度传感器854。智能产品110还可配置有陀螺仪、气压计、湿度计、温度计、红外线传感器、光传感器、运动传感器等其他传感器。

[0222] 音频电路860、扬声器861、传声器862可提供用户与智能产品110之间的音频接口。音频电路860可将接收到的音频数据转换后的电信号,传输到扬声器861,由扬声器861转换为声音信号输出。智能产品110还可配置音量按钮,用于调节声音信号的音量。另一方面,传声器862将收集的声音信号转换为电信号,由音频电路860接收后转换为音频数据,再将音频数据输出至通信组件810以发送给比如另一智能产品110,或者将音频数据输出至存储器820以便进一步处理。

[0223] 蓝牙模块870用于通过蓝牙协议来与其他具有蓝牙模块的蓝牙设备进行信息交互。例如,智能产品可以通过蓝牙模块870与同样具备蓝牙模块的可穿戴电子设备(例如智能手表)建立蓝牙连接,从而进行数据交互。

[0224] 处理器880是智能产品的控制中心,利用各种接口和线路连接整个终端的各个部分,通过运行或执行存储在存储器820内的软件程序,以及调用存储在存储器820内的数据,

执行智能产品的各种功能和处理数据。在一些实施例中,处理器880可包括一个或多个处理单元;处理器880还可以集成应用处理器和基带处理器,其中,应用处理器主要处理操作系统、用户界面和应用程序等,基带处理器主要处理无线通信。可以理解的是,上述基带处理器也可以不集成到处理器880中。本申请中处理器880可以运行操作系统、应用程序、用户界面显示及触控响应,以及本申请实施例的问题答复方法。另外,处理器880与显示单元830耦接。

[0225] 在一些可能的实施方式中,本申请提供的问题答复方法的各个方面还可以实现为一种程序产品的形式,其包括程序代码,当程序产品在计算机设备上运行时,程序代码用于使计算机设备执行本说明书上述描述的根据本申请各种示例性实施方式的问题答复方法中的步骤,例如,计算机设备可以执行如图2中所示的步骤。

[0226] 程序产品可以采用一个或多个可读介质的任意组合。可读介质可以是可读信号介质或者可读存储介质。可读存储介质例如可以是但不限于电、磁、光、电磁、红外线、或半导体的系统、装置或器件,或者任意以上的组合。可读存储介质的更具体的例子(非穷举的列表)包括:具有一个或多个导线的电连接、便携式盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM或闪存)、光纤、便携式紧凑盘只读存储器(CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。

[0227] 本申请的实施方式的程序产品可以采用便携式紧凑盘只读存储器(CD-ROM)并包括程序代码,并可以在计算装置上运行。然而,本申请的程序产品不限于此,在本申请实施例中,可读存储介质可以是任何包含或存储程序的有形介质,该程序可以被命令执行系统、装置或者器件使用或者与其结合使用。

[0228] 可读信号介质可以包括在基带中或者作为载波一部分传播的数据信号,其中承载了可读程序代码。这种传播的数据信号可以采用多种形式,包括但不限于电磁信号、光信号或上述的任意合适的组合。可读信号介质还可以是可读存储介质以外的任何可读介质,该可读介质可以发送、传播或者传输用于由命令执行系统、装置或者器件使用或者与其结合使用的程序。

[0229] 可读介质上包含的程序代码可以用任何适当的介质传输,包括但不限于无线、有线、光缆、RF等等,或者上述的任意合适的组合。

[0230] 可以以一种或多种程序设计语言的任意组合来编写用于执行本申请操作的程序代码,程序设计语言包括面向对象的程序设计语言—诸如Java、C++等,还包括常规的过程式程序设计语言—诸如“C”语言或类似的程序设计语言。程序代码可以完全地在用户计算装置上执行、部分地在用户设备上执行、作为一个独立的软件包执行、部分在用户计算装置上部分在远程计算装置上执行、或者完全在远程计算装置或服务器上执行。在涉及远程计算装置的情形中,远程计算装置可以通过任意种类的网络包括局域网(LAN)或广域网(WAN)连接到用户计算装置,或者,可以连接到外部计算装置(例如利用因特网服务提供商来通过因特网连接)。

[0231] 应当注意,尽管在上文详细描述中提及了装置的若干单元或子单元,但是这种划分仅仅是示例性的并非强制性的。实际上,根据本申请的实施方式,上文描述的两个或更多单元的特征和功能可以在一个单元中具体化。反之,上文描述的一个单元的特征和功能可以进一步划分为由多个单元来具体化。

[0232] 此外,尽管在附图中以特定顺序描述了本申请方法的操作,但是,这并非要求或者暗示必须按照该特定顺序来执行这些操作,或是必须执行全部所示的操作才能实现期望的结果。附加地或备选地,可以省略某些步骤,将多个步骤合并为一个步骤执行,和/或将一个步骤分解为多个步骤执行。

[0233] 本领域内的技术人员应明白,本申请的实施例可提供为方法、系统、或计算机程序产品。因此,本申请可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且,本申请可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质(包括但不限于磁盘存储器、CD-ROM、光学存储器等)上实施的计算机程序产品的形式。

[0234] 本申请是参照根据本申请实施例的方法、设备(系统)、和计算机程序产品的流程图和/或方框图来描述的。应理解可由计算机程序命令实现流程图和/或方框图中的每一流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序命令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产生一个机器,使得通过计算机或其他可编程数据处理设备的处理器执行的命令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

[0235] 这些计算机程序命令也可存储在能引导计算机或其他可编程数据处理设备以特定方式工作的计算机可读存储器中,使得存储在该计算机可读存储器中的命令产生包括命令装置的制造品,该命令装置实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能。

[0236] 这些计算机程序命令也可装载到计算机或其他可编程数据处理设备上,使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理,从而在计算机或其他可编程设备上执行的命令提供用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的步骤。

[0237] 尽管已描述了本申请的优选实施例,但本领域内的技术人员一旦得知了基本创造性概念,则可对这些实施例做出另外的变更和修改。所以,所附权利要求意欲解释为包括优选实施例以及落入本申请范围的所有变更和修改。

[0238] 显然,本领域的技术人员可以对本申请进行各种改动和变型而不脱离本申请的精神和范围。这样,倘若本申请的这些修改和变型属于本申请权利要求及其等同技术的范围之内,则本申请也意图包含这些改动和变型在内。

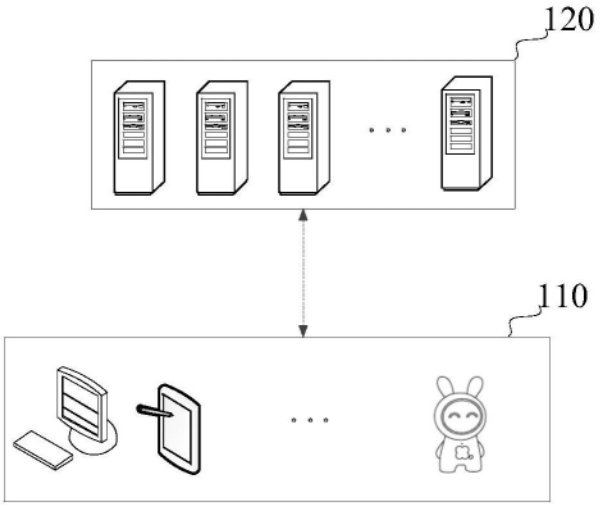
120
110

图1

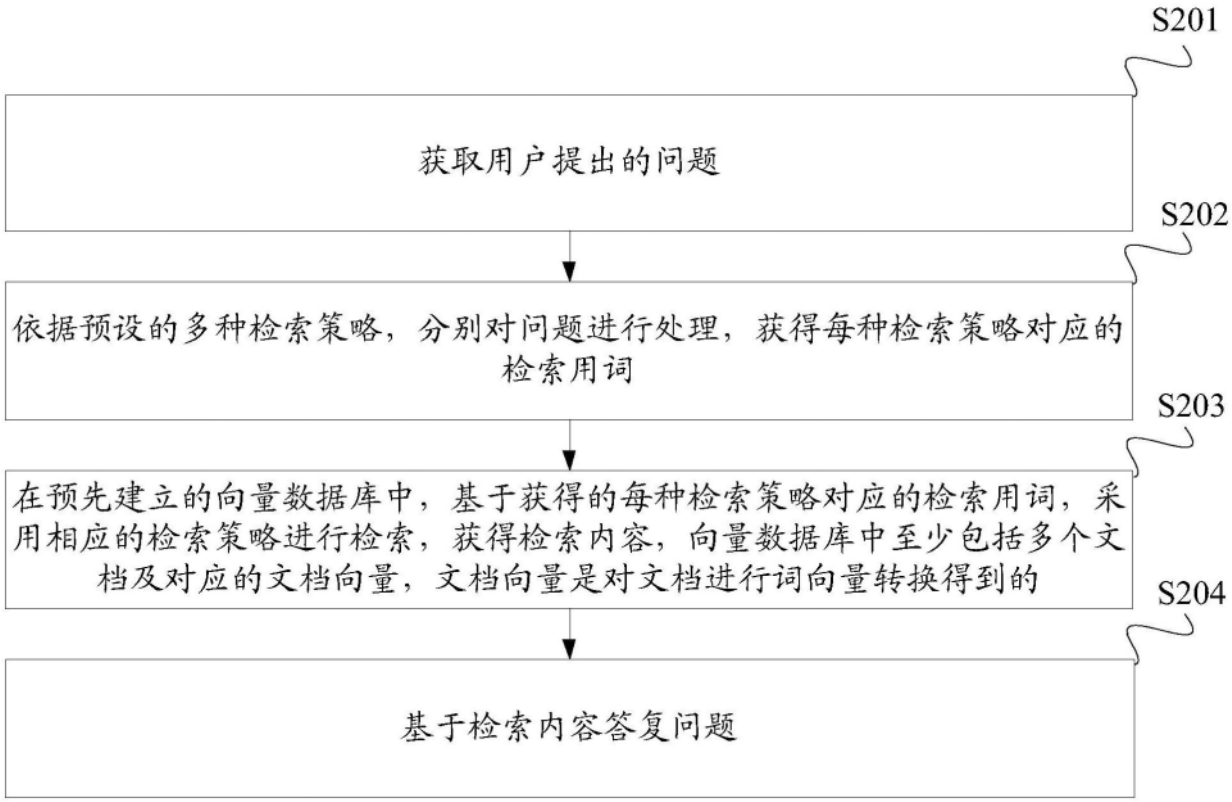
S201
获取用户提出的问题
S202
依据预设的多种检索策略，分别对问题进行处理，获得每种检索策略对应的检索用词
S203
在预先建立的向量数据库中，基于获得的每种检索策略对应的检索用词，采用相应的检索策略进行检索，获得检索内容，向量数据库中至少包括多个文档及对应的文档向量，文档向量是对文档进行词向量转换得到的
S204
基于检索内容答复问题

图2

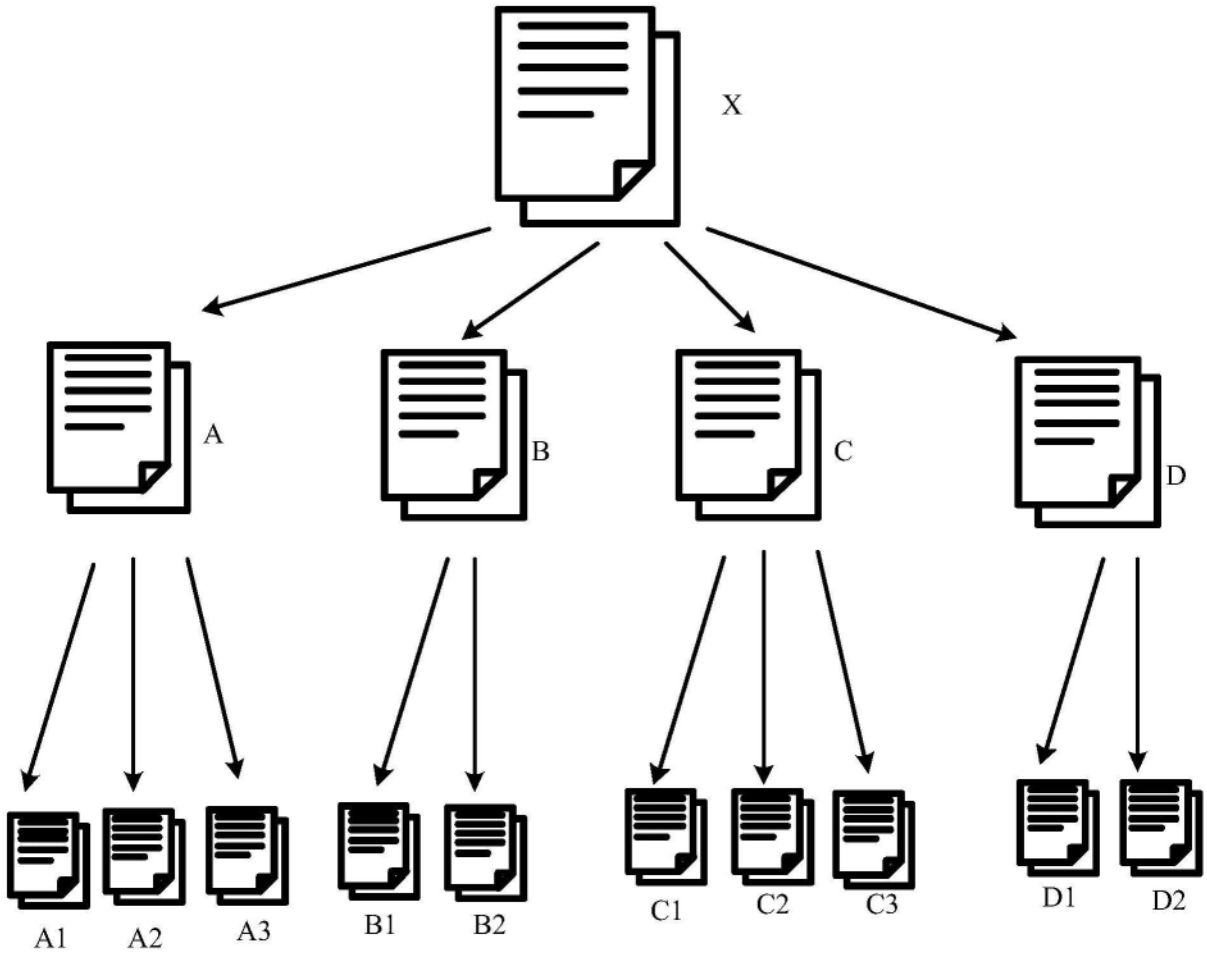

图3

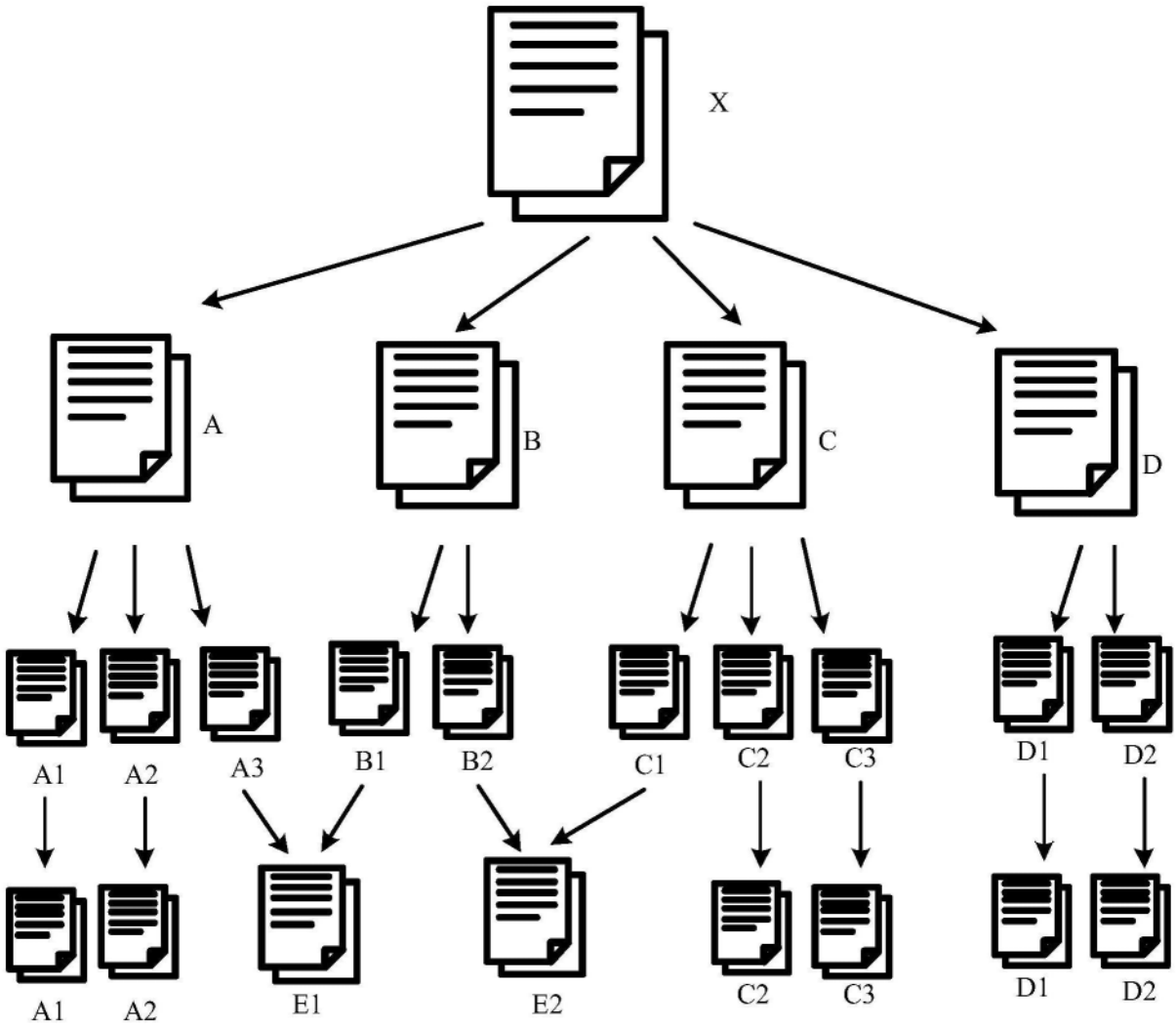

图4

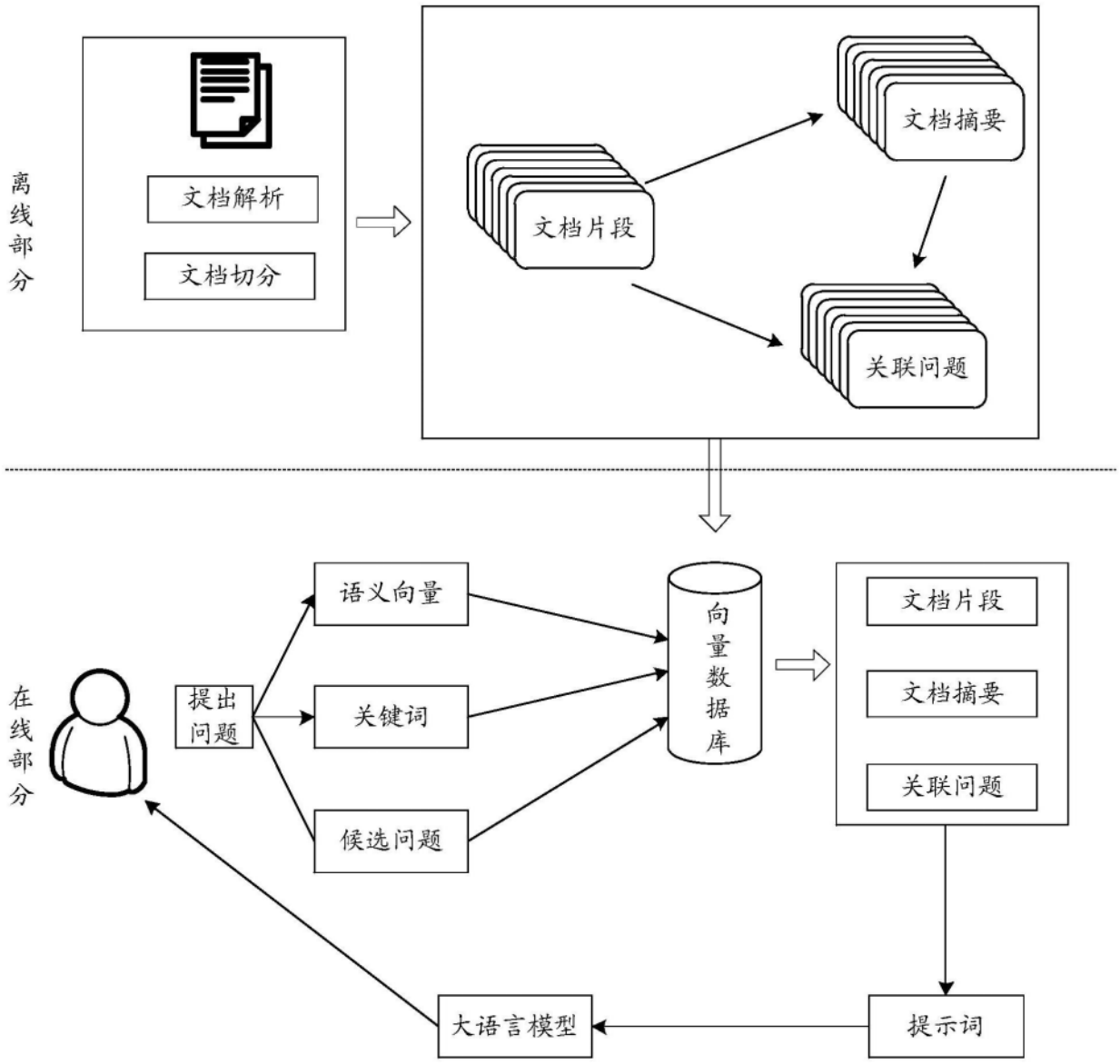
离线部分
文档解析
文档切分
文档片段
文档摘要
关联问题
在线部分
提出问题
语义向量
关键词
候选问题
向量数据库
文档片段
文档摘要
关联问题
大语言模型
提示词

图5

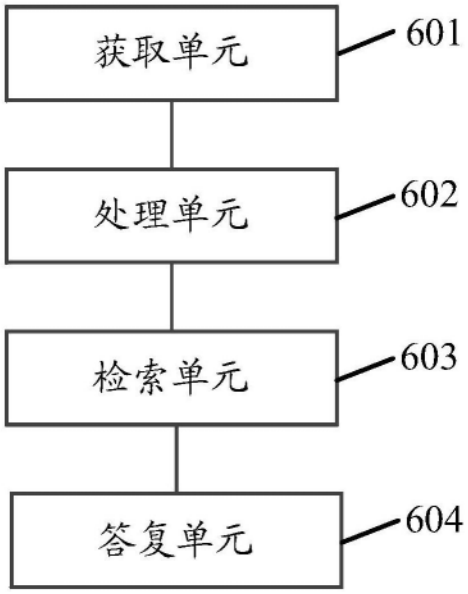
获取单元
601
处理单元
602
检索单元
603
答复单元
604

图6

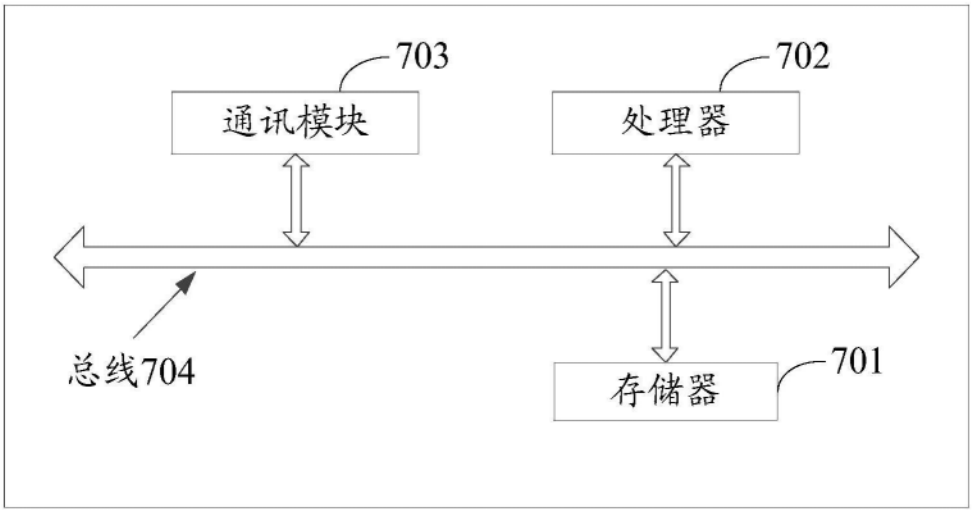
703
通讯模块
702
处理器
总线704
存储器
701

图7

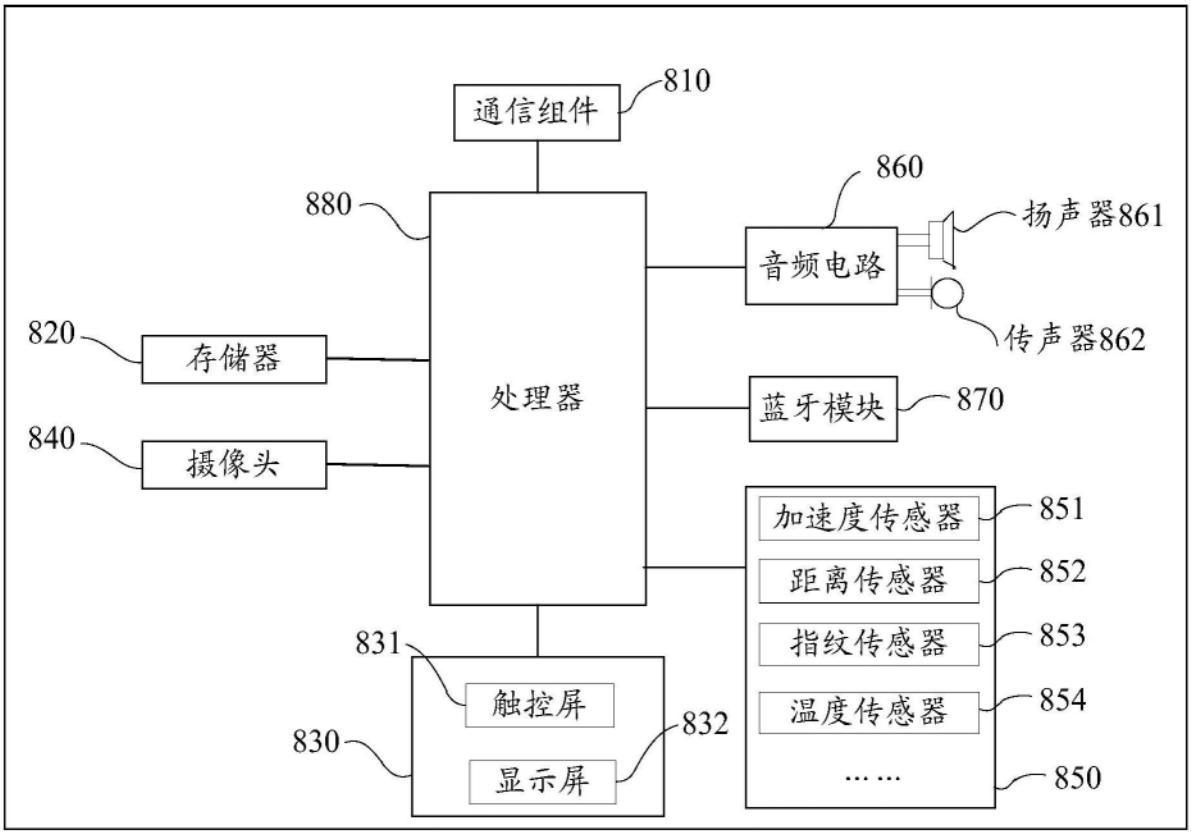
通信组件
810
880
处理器
860
音频电路
扬声器861
传声器862
820
存储器
840
摄像头
蓝牙模块
870
加速度传感器
851
距离传感器
852
指纹传感器
853
温度传感器
854
……
850
831
触控屏
830
显示屏
832

图8