(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 114490926 A
(43)申请公布日 2022.05.13

(21)申请号 202111668984.7

(22)申请日 2021.12.30

(71)申请人 特斯联科技集团有限公司
地址 101100 北京市通州区滨惠北一街3号院1号楼1-6室

(72)发明人 冯琰一 邹游 张睿 刘跃

(74)专利代理机构 北京辰权知识产权代理有限公司 11619
专利代理师 李小朋

(51)Int.Cl.
G06F 16/33(2019.01)
G06F 16/332(2019.01)
G06F 16/35(2019.01)
G06F 40/211(2020.01)
G06F 40/289(2020.01)
G06F 40/30(2020.01)
G06K 9/62(2022.01)

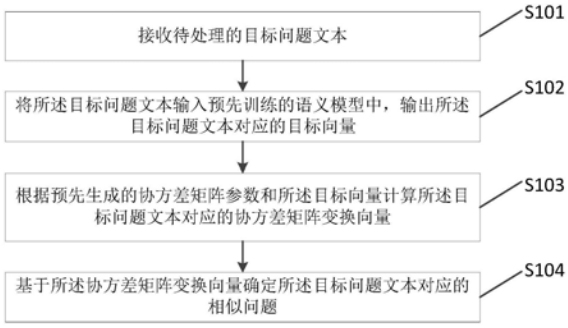
权利要求书2页 说明书12页 附图3页

(54)发明名称

一种相似问题的确定方法、装置、存储介质及终端

(57)摘要

本发明公开了一种相似问题的确定方法、装置、存储介质及终端，方法包括：接收待处理的目标问题文本；将目标问题文本输入预先训练的语义模型中，输出目标问题文本对应的目标向量；根据预先生成的协方差矩阵参数和目标向量计算目标问题文本对应的协方差矩阵变换向量；基于协方差矩阵变换向量确定目标问题文本对应的相似问题。由于本申请将问题文本转化为句向量，并采用预先生成的协方差矩阵参数对句向量进行协方差矩阵变换，从而保障了句向量的各向同性，即句向量不会因其它影响因子而发生变化，进而提升了相似问题推荐的精确度。



CN 114490926 A

1.一种相似问题的确定方法,其特征在于,所述方法包括:

接收待处理的目标问题文本;

将所述目标问题文本输入预先训练的语义模型中,输出所述目标问题文本对应的目标向量;

根据预先生成的协方差矩阵参数和所述目标向量计算所述目标问题文本对应的协方差矩阵变换向量;

基于所述协方差矩阵变换向量确定所述目标问题文本对应的相似问题。

2.根据权利要求1所述的方法,其特征在于,按照以下步骤生成预先训练的语义模型,包括:

获取bert网络,并初始化bert网络的权重后得到语义模型;

获取无标签数据集和问题文本库,根据所述无标签数据集和问题文本库对所述语义模型进行预训练,得到预训练后的语义模型;

对所述问题文本库中的每个问题文本构造正样本和负样本,生成多个训练样本;

将每个训练样本输入预训练后的语义模型中,输出多个样本参数向量;

根据所述多个样本参数向量计算损失值;

当所述损失值到达预设阈值时,生成预先训练的语义模型。

3.根据权利要求2所述的方法,其特征在于,所述根据所述无标签数据集和问题文本库对所述语义模型进行预训练,得到预训练后的语义模型,包括:

将所述无标签数据集中每个无标签数据进行分词处理,得到每个无标签数据的子词序列;

将所述无标签数据集输入预设word2vec网络中进行负采样方式训练,输出每个词的词向量;

计算所述子词序列中每个子词与所述每个词的词向量之间的余弦相似度,并根据余弦相似度确定出每个子词的相似度集合;

根据所述每个子词的相似度集合将与其对应的子词序列中的词进行替换,得到最终的无标签数据;

将最终的无标签数据和所述问题文本库中所有问题句输入所述语义模型中进行训练,训练结束后得到初始语义模型;

将所述无标签数据集中每个无标签数据与所述问题文本库的所有问题文本随机组合后输入所述初始语义模型中进行训练,训练结束后得到预训练后的语义模型。

4.根据权利要求2所述的方法,其特征在于,所述预训练后的语义模型包括bert网络、GRU网络以及池化层;

所述将每个训练样本输入预训练后的语义模型中,输出多个样本参数向量,包括:

计算所述每个训练样本中各参数的最终向量;

将所述各参数的最终向量依次输入bert网络、GRU网络以及池化层,输出每个样本参数向量;

生成多个样本参数向量。

5.根据权利要求2所述的方法,其特征在于,按照以下步骤得到预先生成的协方差矩阵参数,包括:

将所述问题文本库中的所有问题句分别输入所述预先训练的语义模型中，输出句向量集合；

根据预设协方差矩阵变换公式将所述句向量集合中每个句向量进行变换，得到变换后的数据协方差矩阵；

求解变换后的数据协方差矩阵，得到第一求解参数μ和第二求解参数W；

将所述第一求解参数μ和第二求解参数W确定为预先生成的协方差矩阵参数。

6.根据权利要求5所述的方法，其特征在于，所述方法还包括：

根据所述预先生成的协方差矩阵参数计算所述句向量集合中每个句向量对应的协方差矩阵变换结果；

将每个句向量对应的协方差矩阵变换结果保存至数据库，得到问题库的协方差矩阵变换结果集。

7.根据权利要求6所述的方法，其特征在于，所述基于所述协方差矩阵变换向量确定所述目标问题文本对应的相似问题，包括：

将所述问题库的协方差矩阵变换结果集平均分配至预先设定的多个服务节点；

计算所述协方差矩阵变换向量与每个服务节点上多个协方差矩阵变换结果之间的余弦相似度，生成每个服务节点对应的多个余弦相似度；

将每个服务节点对应的多个余弦相似度进行排序，并提取预设数量的余弦相似度，得到初始相似度集合；

将所述初始相似度集合中相似度进行排序，并提取预设数量的余弦相似度，得到多个目标相似度；

将所述多个目标相似度对应的问题文本确定为所述目标问题文本对应的相似问题。

8.一种相似问题的确定装置，其特征在于，所述装置包括：

问题文本接收模块，用于接收待处理的目标问题文本；

问题文本输入模块，用于将所述目标问题文本输入预先训练的语义模型中，输出所述目标问题文本对应的目标向量；

协方差矩阵变换向量计算模块，用于根据预先生成的协方差矩阵参数和所述目标向量计算所述目标问题文本对应的协方差矩阵变换向量；

相似问题确定模块，用于基于所述协方差矩阵变换向量确定所述目标问题文本对应的相似问题。

9.一种计算机存储介质，其特征在于，所述计算机存储介质存储有多条指令，所述指令适于由处理器加载并执行如权利要求1-7任意一项的方法步骤。

10.一种终端，其特征在于，包括：处理器和存储器；其中，所述存储器存储有计算机程序，所述计算机程序适于由所述处理器加载并执行如权利要求1-7任意一项的方法步骤。

一种相似问题的确定方法、装置、存储介质及终端

技术领域

[0001] 本发明涉及机器学习技术领域，特别涉及一种相似问题的确定方法、装置、存储介质及终端。

背景技术

[0002] 在问答领域的建设过程之中，语料是非常重要的核心资产。有了语料才能训练一个好的模型，让属于这个领域的语料都能被模型识别出来。对于问答型的任务，问答对语料的数量就更加的重要，更多的语料就能让产品更加的智能化，能回答用户各种千奇百怪的问题。所以不难发现，问答语料的数量和质量对于问答领域的端到端影响和用户的体验是起了决定影响力的，对问答型的任务是显得尤其重要，语料的数量指的是语料要够多，数量要够大，语料的质量指的是语料的质量要好，要能包含用户的各种方式的问法。

[0003] 现有技术向用户推荐相关问题的时候，通常是采用检索式的召回推荐，一般是通过搜索引擎进行检索、召回，然后推荐。比如，用户输入了一个问答对，通常会到数据库中进行检索，看看数据库中有哪些相似的问题，可以推荐给企业用户。由于目前相似问题推荐系统，很多没有将问题转化为句向量处理，仅仅从关键词等角度处理，即使有将问题转化为句向量的做法，往往没有后续的处理保证句向量的各项同性，即会随着其他影响因子会导致句向量发生变化，从而严重影响相似问题的推荐精确度。

发明内容

[0004] 本申请实施例提供了一种相似问题的确定方法、装置、存储介质及终端。为了对披露的实施例的一些方面有一个基本的理解，下面给出了简单的概括。该概括部分不是泛泛评述，也不是要确定关键/重要组成元素或描绘这些实施例的保护范围。其唯一目的是用简单的形式呈现一些概念，以此作为后面的详细说明的序言。

[0005] 第一方面，本申请实施例提供了一种相似问题的确定方法，方法包括：

[0006] 接收待处理的目标问题文本；

[0007] 将目标问题文本输入预先训练的语义模型中，输出目标问题文本对应的目标向量；

[0008] 根据预先生成的协方差矩阵参数和目标向量计算目标问题文本对应的协方差矩阵变换向量；

[0009] 基于协方差矩阵变换向量确定目标问题文本对应的相似问题。

[0010] 可选的，按照以下步骤生成预先训练的语义模型，包括：

[0011] 获取bert网络，并初始化bert网络的权重后得到语义模型；

[0012] 获取无标签数据集和问题文本库，根据无标签数据集和问题文本库对语义模型进行预训练，得到预训练后的语义模型；

[0013] 对问题文本库中的每个问题文本构造正样本和负样本，生成多个训练样本；

[0014] 将每个训练样本输入预训练后的语义模型中，输出多个样本参数向量；

[0015] 根据多个样本参数向量计算损失值；

[0016] 当损失值到达预设阈值时，生成预先训练的语义模型。

[0017] 可选的，根据无标签数据集和问题文本库对语义模型进行预训练，得到预训练后的语义模型，包括：

[0018] 将无标签数据集中每个无标签数据进行分词处理，得到每个无标签数据的子词序列；

[0019] 将无标签数据集输入预设word2vec网络中进行负采样方式训练，输出每个词的词向量；

[0020] 计算子词序列中每个子词与每个词的词向量之间的余弦相似度，并根据余弦相似度确定出每个子词的相似度集合；

[0021] 根据每个子词的相似度集合将与其对应的子词序列中的词进行替换，得到最终的无标签数据；

[0022] 将最终的无标签数据和问题文本库中所有问题句输入语义模型中进行训练，训练结束后得到初始语义模型；

[0023] 将无标签数据集中每个无标签数据与问题文本库的所有问题文本随机组合后输入初始语义模型中进行训练，训练结束后得到预训练后的语义模型。

[0024] 可选的，预训练后的语义模型包括bert网络、GRU网络以及池化层；

[0025] 将每个训练样本输入预训练后的语义模型中，输出多个样本参数向量，包括：

[0026] 计算每个训练样本中各参数的最终向量；

[0027] 将各参数的最终向量依次输入bert网络、GRU网络以及池化层，输出每个样本参数向量；生成多个样本参数向量。

[0028] 可选的，按照以下步骤得到预先生成的协方差矩阵参数，包括：

[0029] 将问题文本库中的所有问题句分别输入预先训练的语义模型中，输出句向量集合；

[0030] 根据预设协方差矩阵变换公式将句向量集合中每个句向量进行变换，得到变换后的数据协方差矩阵；

[0031] 求解变换后的数据协方差矩阵，得到第一求解参数μ和第二求解参数W；

[0032] 将第一求解参数μ和第二求解参数W确定为预先生成的协方差矩阵参数。

[0033] 可选的，方法还包括：

[0034] 根据预先生成的协方差矩阵参数计算句向量集合中每个句向量对应的协方差矩阵变换结果；

[0035] 将每个句向量对应的协方差矩阵变换结果保存至数据库，得到问题库的协方差矩阵变换结果集。

[0036] 可选的，基于协方差矩阵变换向量确定目标问题文本对应的相似问题，包括：

[0037] 将问题库的协方差矩阵变换结果集平均分配至预先设定的多个服务节点；

[0038] 计算协方差矩阵变换向量与每个服务节点上多个协方差矩阵变换结果之间的余弦相似度，生成每个服务节点对应的多个余弦相似度；

[0039] 将每个服务节点对应的多个余弦相似度进行排序，并提取预设数量的余弦相似度，得到初始相似度集合；

[0040] 将初始相似度集合中相似度进行排序，并提取预设数量的余弦相似度，得到多个目标相似度；

[0041] 将多个目标相似度对应的问题文本确定为目标问题文本对应的相似问题。

[0042] 第二方面，本申请实施例提供了一种相似问题的确定装置，装置包括：

[0043] 问题文本接收模块，用于接收待处理的目标问题文本；

[0044] 问题文本输入模块，用于将目标问题文本输入预先训练的语义模型中，输出目标问题文本对应的目标向量；

[0045] 协方差矩阵变换向量计算模块，用于根据预先生成的协方差矩阵参数和目标向量计算目标问题文本对应的协方差矩阵变换向量；

[0046] 相似问题确定模块，用于基于协方差矩阵变换向量确定目标问题文本对应的相似问题。

[0047] 第三方面，本申请实施例提供一种计算机存储介质，计算机存储介质存储有多条指令，指令适于由处理器加载并执行上述的方法步骤。

[0048] 第四方面，本申请实施例提供一种终端，可包括：处理器和存储器；其中，存储器存储有计算机程序，计算机程序适于由处理器加载并执行上述的方法步骤。

[0049] 本申请实施例提供的技术方案可以包括以下有益效果：

[0050] 在本申请实施例中，相似问题的确定装置首先接收待处理的目标问题文本；然后将目标问题文本输入预先训练的语义模型中，输出目标问题文本对应的目标向量，其次根据预先生成的协方差矩阵参数和目标向量计算目标问题文本对应的协方差矩阵变换向量，最后基于协方差矩阵变换向量确定目标问题文本对应的相似问题。由于本申请将问题文本转化为句向量，并采用预先生成的协方差矩阵参数对句向量进行协方差矩阵变换，从而保障了句向量的各向同性，即句向量不会因其它影响因子而发生变化，进而提升了相似问题推荐的精确度。

[0051] 应当理解的是，以上的一般描述和后文的细节描述仅是示例性和解释性的，并不能限制本发明。

附图说明

[0052] 此处的附图被并入说明书中并构成本说明书的一部分，示出了符合本发明的实施例，并与说明书一起用于解释本发明的原理。

[0053] 图1是本申请实施例提供的一种相似问题的确定方法的流程示意图；

[0054] 图2是本申请实施例提供的一种语义模型训练方法的流程示意图；

[0055] 图3是本申请实施例提供的一种语义模型的模型结构图；

[0056] 图4是本申请实施例提供的一种相似问题的确定装置的结构示意图；

[0057] 图5是本申请实施例提供的一种终端的结构示意图。

具体实施方式

[0058] 以下描述和附图充分地示出本发明的具体实施方案，以使本领域的技术人员能够实践它们。

[0059] 应当明确，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作出创造性劳动前提下所获得的所有其它实施例，都属于本发明保护的范围。

[0060] 下面的描述涉及附图时，除非另有表示，不同附图中的相同数字表示相同或相似的要素。以下示例性实施例中所描述的实施方式并不代表与本发明相一致的所有实施方式。相反，它们仅是如所附权利要求书中所详述的、本发明的一些方面相一致的装置和方法的例子。

[0061] 在本发明的描述中，需要理解的是，术语“第一”、“第二”等仅用于描述目的，而不能理解为指示或暗示相对重要性。对于本领域的普通技术人员而言，可以具体情况理解上述术语在本发明中的具体含义。此外，在本发明的描述中，除非另有说明，“多个”是指两个或两个以上。“和/或”，描述关联对象的关联关系，表示可以存在三种关系，例如，A和/或B，可以表示：单独存在A，同时存在A和B，单独存在B这三种情况。字符“/”一般表示前后关联对象是一种“或”的关系。

[0062] 本申请提供了一种相似问题的确定方法、装置、存储介质及终端，以解决上述相关技术问题中存在的问题。本申请提供的技术方案中，由于本申请将问题文本转化为句向量，并采用预先生成的协方差矩阵参数对句向量进行协方差矩阵变换，从而保障了句向量的各向同性，即句向量不会因其它影响因子而发生变化，进而提升了相似问题推荐的精确度，下面采用示例性的实施例进行详细说明。

[0063] 下面将结合附图1-附图3，对本申请实施例提供的相似问题的确定方法进行详细介绍。该方法可依赖于计算机程序实现，可运行于基于冯诺依曼体系的相似问题的确定装置上。该计算机程序可集成在应用中，也可作为独立的工具类应用运行。

[0064] 请参见图1，为本申请实施例提供了一种相似问题的确定方法的流程示意图。如图1所示，本申请实施例的方法可以包括以下步骤：

[0065] S101，接收待处理的目标问题文本；

[0066] 其中，文本，是指书面语言的表现形式，通常是具有完整、系统含义的一个句子或多个句子的组合，一个文本可以是一个句子、一个段落或者一个篇章。

[0067] 通常，文本是由几个字符组成的词语或者由几个词语构成的一句话，还可以是由几句话构成的一个段落，用户可以将自己的思想通过语言文本进行描述，利用文本进行描述可以使复杂的思想变成让其他人容易理解的指令。针对文本，可以使用不同的表达方式让复杂的思想变得通俗易懂，使得沟通更加容易理解。文本包含的一条或多条自然语言可以简称为语句，也可以通俗的称为句子，也可以根据文本中的标点将文本拆分成句子，即，将以句号、问号、感叹号、逗号等结尾的内容作为一句。

[0068] 目标问题文本指用户输入到用户终端的语言文本，可以是用户通过用户终端文字编辑软件编辑生成的语言文本，也可以是用户通过用户终端语音录取软件录取的语音信息生成的语言文本，目标语言文本的生成具有多种方式，此处不做限定。

[0069] 在一种可能的实现方式中，用户通过点击用户终端上安装的具备问答系统的软件进入到聊天界面，然后通过点击文本输入框后弹出文本编辑器，当弹出文本编辑器之后用户可以将自己表达的思想通过文字描述的方式输入到文本编辑框中，用户终端针对用户的操作生成目标问题文本。需要说明的是，获取待理解的目标语言文本方式有多种，此处不做限定。

[0070] S102,将目标问题文本输入预先训练的语义模型中,输出目标问题文本对应的目标向量;

[0071] 其中,预先训练的语义模型是将用户输入的问题文本转化为向量的数学模型。该模型包括bert网络、GRU网络以及池化层。

[0072] 在本申请实施例中,在生成预先训练的语义模型时,首先获取bert网络,并初始化bert网络的权重后得到语义模型,然后获取无标签数据集和问题文本库,根据无标签数据集和问题文本库对语义模型进行预训练,得到预训练后的语义模型,再对问题文本库中的每个问题文本构造正样本和负样本,生成多个训练样本,其次将每个训练样本输入预训练后的语义模型中,输出多个样本参数向量,再根据多个样本参数向量计算损失值,最后当损失值到达预设阈值时,生成预先训练的语义模型。

[0073] 在一种可能的实现方式中,在接收到待处理的目标问题文本后,可将该待处理的目标问题文本输入预先训练的语义模型中,在经过模型的bert网络、GRU网络以及池化层处理之后,可输出目标问题文本对应的目标向量,记为out_{q_1}。

[0074] S103,根据预先生成的协方差矩阵参数和目标向量计算目标问题文本对应的协方差矩阵变换向量;

[0075] 在一种可能的实现方式中,预先生成的协方差矩阵参数分别为第一求解参数μ和第二求解参数W,在得到out_{q_1}后,可根据预设协方差矩阵变换公式$\widetilde{q}_1=\left(out_{q_1}-\mu\right)W$计算出目标问题文本对应的协方差矩阵变换向量$\widetilde{q}_1$。

[0076] S104,基于协方差矩阵变换向量确定目标问题文本对应的相似问题。

[0077] 在本申请实施例中,在得到$\widetilde{q}_1$后,使用$\widetilde{q}_1$与问题库的协方差矩阵变换结果集$\{\widetilde{e}\}_{i=1}^{N}$按照余弦相似度算分,即可得到最相似的问题推荐给用户。

[0078] 在一种可能的实现方式中,首先将问题库的协方差矩阵变换结果集平均分配至预先设定的多个服务节点,然后计算协方差矩阵变换向量与每个服务节点上多个协方差矩阵变换结果之间的余弦相似度,生成每个服务节点对应的多个余弦相似度,再将每个服务节点对应的多个余弦相似度进行排序,并提取预设数量的余弦相似度,得到初始相似度集合,其次将初始相似度集合中相似度进行排序,并提取预设数量的余弦相似度,得到多个目标相似度,最后将多个目标相似度对应的问题文本确定为目标问题文本对应的相似问题。

[0079] 例如,将$\{\widetilde{e}\}_{i=1}^{N}$分别放在N个服务节点master,node1,node2,node3,…,$node_{n-1}$上。假设需要给用户推荐的相似问题数为K,则对于每个节点,通过余弦相似度取出Top-K个结果。然后master服务节点合并所有N个节点的Top-K个结果,将相似度排序取出最终前Top-K个结果推荐给用户。

[0080] 在另一种可能的实现方式中,将$\{\widetilde{e}\}_{i=1}^{N}$构建成球树,具体包括以下步骤:1)先构建一个超球体,这个超球体是可以包含所有样本的最小球体。2)从球中选择一个离球的中心最远的点,然后选择第二个点离第一个点最远,将球中所有的点分配到离这两个聚类中心最近的一个上,然后计算每个聚类的中心,以及聚类能够包含它所有数据点所需的最小半

径。这样得到了两个子超球体，和KD树里面的左右子树对应。3)对于这两个子超球体，递归执行步骤2)最终得到了一个球树。每个节点在计算topK个结果的时候，通过在上述步骤构建的球体中去搜寻，可大大提高计算速度。

[0081] 需要说明的是，通过以上两种方式可以大大缩短相似问题推荐的响应时间，使得在海量数据的情况下，系统也能有好的响应时间，大大提升实际工程环境可用性，提升用户体验。

[0082] 在本申请实施例中，相似问题的确定装置首先接收待处理的目标问题文本，然后将目标问题文本输入预先训练的语义模型中，输出目标问题文本对应的目标向量，其次根据预先生成的协方差矩阵参数和目标向量计算目标问题文本对应的协方差矩阵变换向量，最后基于协方差矩阵变换向量确定目标问题文本对应的相似问题。由于本申请将问题文本转化为句向量，并采用预先生成的协方差矩阵参数对句向量进行协方差矩阵变换，从而保障了句向量的各向同性，即句向量不会因其它影响因子而发生变化，进而提升了相似问题推荐的精确度。

[0083] 请参见图2，为本申请实施例提供了一种语义模型训练方法的流程示意图。

[0084] 如图2所示，本申请实施例的方法可以包括以下步骤：

[0085] S201，获取bert网络，并初始化bert网络的权重后得到语义模型；

[0086] 在一种可能的实现方式中，首先获取bert网络的模型框架，然后将模型的权重进行初始化，初始化使用谷歌的bert-base-chinese模型权重进行初始化。

[0087] S202，获取无标签数据集和问题文本库，根据无标签数据集和问题文本库对语义模型进行预训练，得到预训练后的语义模型；

[0088] 在本申请实施例中，首先获取无标签数据集和问题文本库，然后将无标签数据集中每个无标签数据进行分词处理，得到每个无标签数据的子词序列，再将无标签数据集输入预设word2vec网络中进行负采样方式训练，输出每个词的词向量，然后计算子词序列中每个子词与每个词的词向量之间的余弦相似度，并根据余弦相似度确定出每个子词的相似度集合，再根据每个子词的相似度集合将与其对应的子词序列中的词进行替换，得到最终的无标签数据，其次将最终的无标签数据和问题文本库中所有问题句输入语义模型中进行训练，训练结束后得到初始语义模型，最后将无标签数据集中每个无标签数据与问题文本库的所有问题文本随机组合后输入初始语义模型中进行训练，训练结束后得到预训练后的语义模型。

[0089] 在一种可能的实现方式中，首先根据问题推荐系统的当前所属领域，搜集大量的领域内无标签数据，得到无标签数据集$\{x_{\text{domain}}\}_{i=1}^{N}$，利用CRF算法将每一个无标签数据$\{x_{\text{doma}}i_{n}\}_{i}$进行分词处理，分词处理结束后可得到每个无标签数据的子词序列$\{piece_1, piece_2, piece_3, ..., piece_M\}$，将子词序列按照N-Gram的掩盖策略进行候选tokens的掩盖，从一个piece到连续四piece掩盖的百分比设置为45%，35%，15%，5%。其中4元模型可表示为：

[0090] $$P(piece_1, piece_2, ...piece_m) = \prod_{i=1}^{m} P(piece_i \mid piece_{i-3} piece_{i-2} piece_{i-1})$$

[0091] 在对子词序列进行掩盖时，采用相似的子词进行掩盖，首先对无标签数据使用

word2vec算法，并采用负采样方式训练得到每个词的词向量，计算每个piece与每个词的词向量之间的余弦相似度，得到相似度集合$\{p\}_i$，取集合$\{p\}_i$升序排序后第一四分位数为闸值M。

[0092] 具体的，子词掩盖时，通过余弦相似度计算得到的最相似的词相似度为P_{sim}，当P_{sim}>M则用那个最相似词替换进行掩盖，当P_{sim}<＝M则采用随机替换掩盖的方式。输入句子中的piece有15％概率进行掩盖，其中90％按上述方式替换为相似词，5％替换为随机词，剩下的保持不变，最终替换后得到最终的无标签数据，将最终的无标签数据输入语义模型进行预训练，epoch为80。预训练完毕之后，将问题库的所有句子$\{x_{query}\}_{i=1}^{N2}$，按照领域内预训练同样的方式输入预训练模型继续进行预训练，epoch为80。待预训练结束之后，将收集的无标签数据集$\{x_{domain}\}_{i=1}^{N}$与问题库的所有句子$\{x_{query}\}_{i=1}^{N2}$重组打乱之后，再次输入模型按照相同的方式进行预训练，epoch为100。按照如上方式预训练结束之后，得到预训练后的语义模型。

[0093] S203，对问题文本库中的每个问题文本构造正样本和负样本，生成多个训练样本；

[0094] 在一种可能的实现方式中，对于问题库中的所有问题$\{q\}_{i=1}^{N}$，首先对每个问题q_i选择一个正样本$positive_i$和两个负样本negative 1_i与negative 2_i，正负样本分别从问题库中筛选，若无通过人工构造，最终可得到N个训练样本，每个训练样本可表示为(q_i，$positive_i$，negative 1_i，negative 2_i)。

[0095] S204，将每个训练样本输入预训练后的语义模型中，输出多个样本参数向量；

[0096] 其中，预训练后的语义模型包括bert网络、GRU网络以及池化层。

[0097] 在本申请实施例中，首先计算每个训练样本中各参数的最终向量，然后将各参数的最终向量依次输入bert网络、GRU网络以及池化层，输出每个样本参数向量，最后生成多个样本参数向量。

[0098] 具体的，在计算每个训练样本中各参数的最终向量时，首先将每个训练样本中各参数经过token的embdding(embdding表示向量化的表示)和位置embedding还有segment(段)的embedding，得到每个参数的多个embedding结果，最后将每个参数的多个embedding结果做和后得到最终embedding，即每个训练样本中各参数的最终向量。

[0099] 例如，计算训练样本(q_i，$positive_i$，negative 1_i，negative 2_i)中的q_i的最终embedding时，q_i经过token、位置、segment(段)的embedding，公式为：Embedding(q_i)＝TokenEmbedding(q_i)+PosEmbedding(q_i)+SegEmbedding(q_i)具体的，例如图3所示，每个训练样本中各参数的最终向量为embedding-q，embedding-positive，embedding-negative1，embedding-negative2，分别依次输入bert网络、GRU网络以及池化层，输出每个样本参数向量，最后生成多个样本参数向量，多个样本参数向量为out-q，out-positive，out-negative1，out-negative2。

[0100] 例如，在生成out-q时，将embedding-q输入bert得到$bert-encoder_{q_i}=\mathrm{bert}(Embedding(q_i))$，再将$bert-encoder_{q_i}$通过GRU进一步的编码$GRU-encoder_{q_i}=GRU(bert-encoder_{q_i})$，将$GRU-encoder_{q_i}$经过平均池化得到最后的

输出$out_{q_i} = avgpooling(GRU - encoder_{q_i})$。

[0101] S205,根据多个样本参数向量计算损失值;

[0102] 在一种可能的实现方式中,例如图3所示在得到out-q,out-positive,out-negative1,out-negative2后,使用提出的Quaternary Contrast Loss损失函数计算损失。

[0103] 具体的,公式如下:

[0104] $QCL = \max\left(\left\|out_{q_i} - out_{positive_i}\right\| - \alpha\left(\left\|out_{q_i} - out_{negative1_i}\right\| + \left\|out_{q_i} - out_{negative2_i}\right\|\right) + \varepsilon, 0\right)$;

[0105] 其中out_{q_i}、$out_{positive_i}$、$out_{negative1_i}$、$out_{negative2_i}$分别为(q_i,$positive_i$,$negative\ 1_i$,$negative\ 2_i$)经过模型后的输出向量表示,即out-q,out-positive,out-negative1,out-negative2,ε表示边距,边距大小为1,||·||表示距离度量,这里使用欧氏距离,α表示缩放系数,这里取值范围为0～1。

[0106] 需要说明的是,采用了新的预训练模式首先对模型进行预训练。并使用图3所示的基于预训练模型的多塔模型结构,并创造性提出了Quaternary Contrast Loss损失函数对模型进行训练,大大提高了模型精准的捕捉用户问题深层次的语义信息,挖掘用户问题和其他问题的潜在联系的能力。

[0107] S206,当损失值到达预设阈值时,生成预先训练的语义模型。

[0108] 在一种可能的实现方式中,当损失值到达预设阈值时,生成预先训练的语义模型。

[0109] 在另一种可能的实现方式中,当损失值未到达预设阈值时,将损失值反向传播已更新模型参数,并继续执行将每个训练样本输入预训练后的语义模型中的步骤,继续训练图3的模型。

[0110] 进一步地,在得到预先训练的语义模型后,可按照以下步骤得到预先生成的协方差矩阵参数,首先将问题文本库中的所有问题句分别输入预先训练的语义模型中,输出句向量集合,然后根据预设协方差矩阵变换公式将句向量集合中每个句向量进行变换,得到变换后的数据协方差矩阵,其次求解变换后的数据协方差矩阵,得到第一求解参数μ和第二求解参数W,最后将第一求解参数μ和第二求解参数W确定为预先生成的协方差矩阵参数。

[0111] 进一步地,还根据预先生成的协方差矩阵参数计算句向量集合中每个句向量对应的协方差矩阵变换结果,最后将每个句向量对应的协方差矩阵变换结果保存至数据库,得到问题库的协方差矩阵变换结果集。

[0112] 具体的,将问题库中的所有问题句$\{q\}_{i=1}^{N}$分别经过预先训练的语义模型的bert,GRU和pooling层之后得到输出的句向量表示为$\{e\}_{i=1}^{N}$,将$\{e\}_{i=1}^{N}$执行变换公式如下:$\tilde{e}_i = (e_i - \mu)W$,使得$\{\tilde{e}\}_{i=1}^{N}$的均值为0,协方差矩阵为单位矩阵。其中$\mu = \frac{1}{N}\sum_{i=1}^{N} e_i$,这里将原始数据的协方差矩阵记为:

[0113] $$\Sigma = \frac{1}{N}\sum_{i=1}^{N}(e_i - \mu)^T(e_i - \mu) = \left(\frac{1}{N}\sum_{i=1}^{N} e_i^T e_i\right) - \mu^T\mu$$

[0114] 则变换后的数据协方差矩阵为$\widetilde{\Sigma}=W^{T}\Sigma W$,因此实际需要解方程

[0115] $W^{T}\Sigma W=I$可推出$\Sigma=(W^{T})^{-1}W^{-1}=(W^{-1})^{T}W^{-1}$,由于协方差矩阵Σ是一个半正定的对称矩阵,具有如下的SVD分解形式:

[0116] $\Sigma=U\Lambda U^{T}$;其中U是一个正交矩阵,而Λ是一个对角阵,并且对角线元素都是正的,因此直接让$W^{-1}=\sqrt{\Lambda}U^{T}$就可以完成求解:

[0117] $$W=U\sqrt{\Lambda^{-1}}$$

[0118] 按照上述方式求得μ和W之后,分别将$\{e\}_{i=1}^{N}$通过$\widetilde{e}_{i}=(e_{i}-\mu)W$求得$\{\widetilde{e}\}_{i=1}^{N}$即为问题库中的所有问题句$\{q\}_{i=1}^{N}$经过cov-transform层进行协方差矩阵变换的结果,将$\{\widetilde{e}\}_{i=1}^{N}$的句向量保存在数据库中召回阶段使用。

[0119] 在本申请实施例中,相似问题的确定装置首先接收待处理的目标问题文本,然后将目标问题文本输入预先训练的语义模型中,输出目标问题文本对应的目标向量,其次根据预先生成的协方差矩阵参数和目标向量计算目标问题文本对应的协方差矩阵变换向量,最后基于协方差矩阵变换向量确定目标问题文本对应的相似问题。由于本申请将问题文本转化为句向量,并采用预先生成的协方差矩阵参数对句向量进行协方差矩阵变换,从而保障了句向量的各向同性,即句向量不会因其它影响因子而发生变化,进而提升了相似问题推荐的精确度。

[0120] 下述为本发明装置实施例,可以用于执行本发明方法实施例。对于本发明装置实施例中未披露的细节,请参照本发明方法实施例。

[0121] 请参见图4,其示出了本发明一个示例性实施例提供的相似问题的确定装置的结构示意图。该相似问题的确定装置可以通过软件、硬件或者两者的结合实现成为终端的全部或一部分。该装置1包括问题文本接收模块10、问题文本输入模块20、协方差矩阵变换向量计算模块30、相似问题确定模块40。

[0122] 问题文本接收模块10,用于接收待处理的目标问题文本;

[0123] 问题文本输入模块20,用于将目标问题文本输入预先训练的语义模型中,输出目标问题文本对应的目标向量;

[0124] 协方差矩阵变换向量计算模块30,用于根据预先生成的协方差矩阵参数和目标向量计算目标问题文本对应的协方差矩阵变换向量;

[0125] 相似问题确定模块40,用于基于协方差矩阵变换向量确定目标问题文本对应的相似问题。

[0126] 需要说明的是,上述实施例提供的相似问题的确定装置在执行相似问题的确定方法时,仅以上述各功能模块的划分进行举例说明,实际应用中,可以根据需要而将上述功能分配由不同的功能模块完成,即将设备的内部结构划分成不同的功能模块,以完成以上描述的全部或者部分功能。另外,上述实施例提供的相似问题的确定装置与相似问题的确定方法实施例属于同一构思,其体现实现过程详见方法实施例,这里不再赘述。

[0127] 上述本申请实施例序号仅仅为了描述,不代表实施例的优劣。

[0128] 在本申请实施例中,相似问题的确定装置首先接收待处理的目标问题文本,然后

将目标问题文本输入预先训练的语义模型中，输出目标问题文本对应的目标向量，其次根据预先生成的协方差矩阵参数和目标向量计算目标问题文本对应的协方差矩阵变换向量，最后基于协方差矩阵变换向量确定目标问题文本对应的相似问题。由于本申请将问题文本转化为句向量，并采用预先生成的协方差矩阵参数对句向量进行协方差矩阵变换，从而保障了句向量的各向同性，即句向量不会因其它影响因子而发生变化，进而提升了相似问题推荐的精确度。

[0129] 本发明还提供一种计算机可读介质，其上存储有程序指令，该程序指令被处理器执行时实现上述各个方法实施例提供的相似问题的确定方法。

[0130] 本发明还提供了一种包含指令的计算机程序产品，当其在计算机上运行时，使得计算机执行上述各个方法实施例的相似问题的确定方法。

[0131] 请参见图5，为本申请实施例提供了一种终端的结构示意图。如图5所示，终端1000可以包括：至少一个处理器1001，至少一个网络接口1004，用户接口1003，存储器1005，至少一个通信总线1002。

[0132] 其中，通信总线1002用于实现这些组件之间的连接通信。

[0133] 其中，用户接口1003可以包括显示屏(Display)、摄像头(Camera)，可选用户接口1003还可以包括标准的有线接口、无线接口。

[0134] 其中，网络接口1004可选的可以包括标准的有线接口、无线接口(如WI-FI接口)。

[0135] 其中，处理器1001可以包括一个或者多个处理核心。处理器1001利用各种接口和线路连接整个电子设备1000内的各个部分，通过运行或执行存储在存储器1005内的指令、程序、代码集或指令集，以及调用存储在存储器1005内的数据，执行电子设备1000的各种功能和处理数据。可选的，处理器1001可以采用数字信号处理(Digital Signal Processing，DSP)、现场可编程门阵列(Field-Programmable Gate Array，FPGA)、可编程逻辑阵列(Programmable Logic Array，PLA)中的至少一种硬件形式来实现。处理器1001可集成中央处理器(Central Processing Unit，CPU)、图像处理器(Graphics Processing Unit，GPU)和调制解调器等中的一种或几种的组合。其中，CPU主要处理操作系统、用户界面和应用程序等；GPU用于负责显示屏所需要显示的内容的渲染和绘制；调制解调器用于处理无线通信。可以理解的是，上述调制解调器也可以不集成到处理器1001中，单独通过一块芯片进行实现。

[0136] 其中，存储器1005可以包括随机存储器(Random Access Memory，RAM)，也可以包括只读存储器(Read-Only Memory)。可选的，该存储器1005包括非瞬时性计算机可读介质(non-transitory computer-readable storage medium)。存储器1005可用于存储指令、程序、代码、代码集或指令集。存储器1005可包括存储程序区和存储数据区，其中，存储程序区可存储用于实现操作系统的指令、用于至少一个功能的指令(比如触控功能、声音播放功能、图像播放功能等)、用于实现上述各个方法实施例的指令等；存储数据区可存储上面各个方法实施例中涉及到的数据等。存储器1005可选的还可以是至少一个位于远离前述处理器1001的存储装置。如图5所示，作为一种计算机存储介质的存储器1005中可以包括操作系统、网络通信模块、用户接口模块以及相似问题的确定应用程序。

[0137] 在图5所示的终端1000中，用户接口1003主要用于为用户提供输入的接口，获取用户输入的数据；而处理器1001可以用于调用存储器1005中存储的相似问题的确定应用程

序,并具体执行以下操作:

[0138] 接收待处理的目标问题文本;

[0139] 将目标问题文本输入预先训练的语义模型中,输出目标问题文本对应的目标向量;

[0140] 根据预先生成的协方差矩阵参数和目标向量计算目标问题文本对应的协方差矩阵变换向量;

[0141] 基于协方差矩阵变换向量确定目标问题文本对应的相似问题。

[0142] 在一个实施例中,处理器1001在生成预先训练的语义模型时,具体执行以下操作:

[0143] 获取bert网络,并初始化bert网络的权重后得到语义模型;

[0144] 获取无标签数据集和问题文本库,根据无标签数据集和问题文本库对语义模型进行预训练,得到预训练后的语义模型;

[0145] 对问题文本库中的每个问题文本构造正样本和负样本,生成多个训练样本;

[0146] 将每个训练样本输入预训练后的语义模型中,输出多个样本参数向量;

[0147] 根据多个样本参数向量计算损失值;

[0148] 当损失值到达预设阈值时,生成预先训练的语义模型。

[0149] 在一个实施例中,处理器1001在根据无标签数据集和问题文本库对语义模型进行预训练,得到预训练后的语义模型时,具体执行以下操作:

[0150] 将无标签数据集中每个无标签数据进行分词处理,得到每个无标签数据的子词序列;

[0151] 将无标签数据集输入预设word2vec网络中进行负采样方式训练,输出每个词的词向量;

[0152] 计算子词序列中每个子词与每个词的词向量之间的余弦相似度,并根据余弦相似度确定出每个子词的相似度集合;

[0153] 根据每个子词的相似度集合将与其对应的子词序列中的词进行替换,得到最终的无标签数据;

[0154] 将最终的无标签数据和问题文本库中所有问题句输入语义模型中进行训练,训练结束后得到初始语义模型;

[0155] 将无标签数据集中每个无标签数据与问题文本库的所有问题文本随机组合后输入初始语义模型中进行训练,训练结束后得到预训练后的语义模型。

[0156] 在一个实施例中,处理器1001在执行将每个训练样本输入预训练后的语义模型中,输出多个样本参数向量时,具体执行以下操作:

[0157] 计算每个训练样本中各参数的最终向量;

[0158] 将各参数的最终向量依次输入bert网络、GRU网络以及池化层,输出每个样本参数向量;

[0159] 生成多个样本参数向量。

[0160] 在一个实施例中,处理器1001在生成预先生成的协方差矩阵参数时,具体执行以下操作:

[0161] 将问题文本库中的所有问题句分别输入预先训练的语义模型中,输出句向量集合;

[0162] 根据预设协方差矩阵变换公式将句向量集合中每个句向量进行变换，得到变换后的数据协方差矩阵；

[0163] 求解变换后的数据协方差矩阵，得到第一求解参数μ和第二求解参数W；

[0164] 将第一求解参数μ和第二求解参数W确定为预先生成的协方差矩阵参数。

[0165] 在一个实施例中，处理器1001还执行以下操作：

[0166] 根据预先生成的协方差矩阵参数计算句向量集合中每个句向量对应的协方差矩阵变换结果；

[0167] 将每个句向量对应的协方差矩阵变换结果保存至数据库，得到问题库的协方差矩阵变换结果集。

[0168] 在一个实施例中，处理器1001在执行基于协方差矩阵变换向量确定目标问题文本对应的相似问题时，具体执行以下操作：

[0169] 将问题库的协方差矩阵变换结果集平均分配至预先设定的多个服务节点；

[0170] 计算协方差矩阵变换向量与每个服务节点上多个协方差矩阵变换结果之间的余弦相似度，生成每个服务节点对应的多个余弦相似度；

[0171] 将每个服务节点对应的多个余弦相似度进行排序，并提取预设数量的余弦相似度，得到初始相似度集合；

[0172] 将初始相似度集合中相似度进行排序，并提取预设数量的余弦相似度，得到多个目标相似度；

[0173] 将多个目标相似度对应的问题文本确定为目标问题文本对应的相似问题。

[0174] 在本申请实施例中，相似问题的确定装置首先接收待处理的目标问题文本，然后将目标问题文本输入预先训练的语义模型中，输出目标问题文本对应的目标向量，其次根据预先生成的协方差矩阵参数和目标向量计算目标问题文本对应的协方差矩阵变换向量，最后基于协方差矩阵变换向量确定目标问题文本对应的相似问题。由于本申请将问题文本转化为句向量，并采用预先生成的协方差矩阵参数对句向量进行协方差矩阵变换，从而保障了句向量的各向同性，即句向量不会因其它影响因子而发生变化，进而提升了相似问题推荐的精确度。

[0175] 本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程，是可以通过计算机程序来指令相关的硬件来完成，相似问题的确定的程序可存储于计算机可读取存储介质中，该程序在执行时，可包括如上述各方法的实施例的流程。其中，的存储介质可为磁碟、光盘、只读存储记忆体或随机存储记忆体等。

[0176] 以上所揭露的仅为本申请较佳实施例而已，当然不能以此来限定本申请之权利范围，因此依本申请权利要求所作的等同变化，仍属本申请所涵盖的范围。

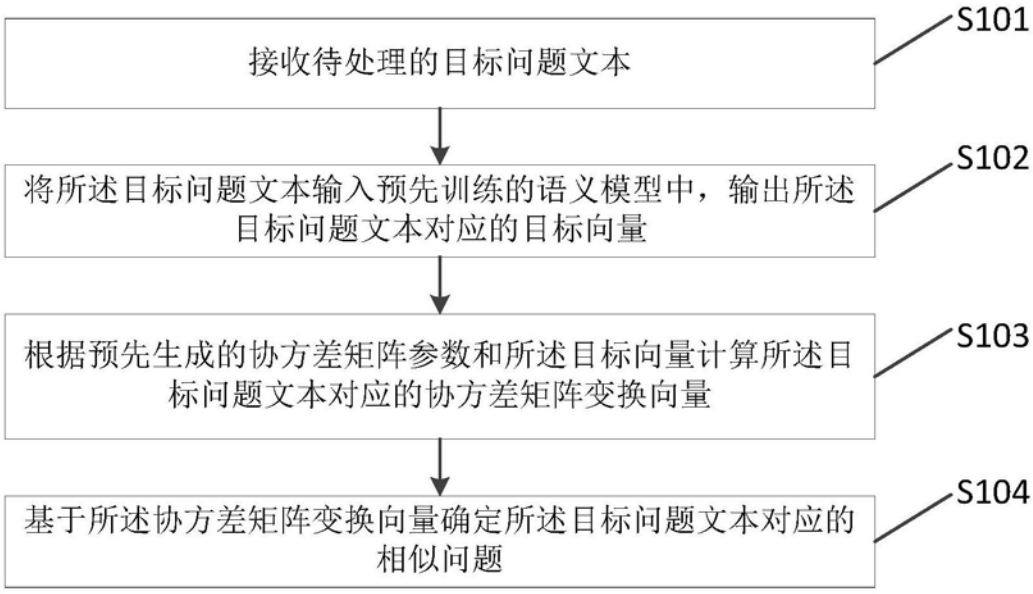

S101
接收待处理的目标问题文本
S102
将所述目标问题文本输入预先训练的语义模型中，输出所述目标问题文本对应的目标向量
S103
根据预先生成的协方差矩阵参数和所述目标向量计算所述目标问题文本对应的协方差矩阵变换向量
S104
基于所述协方差矩阵变换向量确定所述目标问题文本对应的相似问题

图1

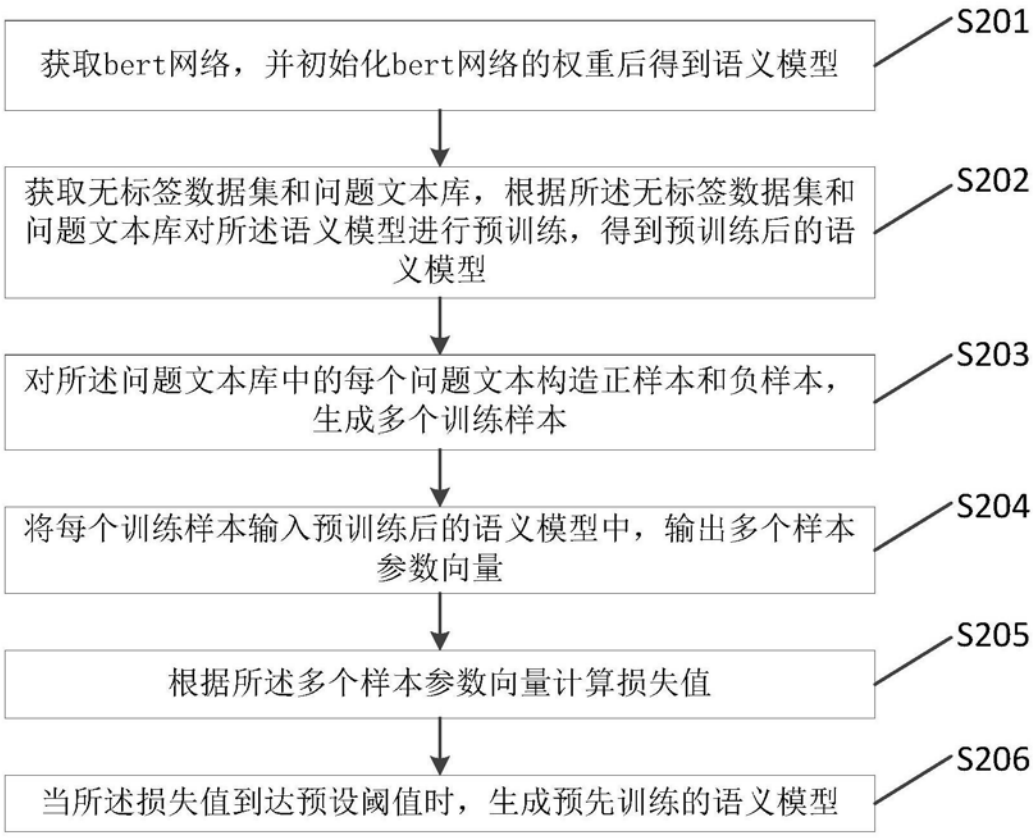

S201
获取bert网络，并初始化bert网络的权重后得到语义模型
S202
获取无标签数据集和问题文本库，根据所述无标签数据集和问题文本库对所述语义模型进行预训练，得到预训练后的语义模型
S203
对所述问题文本库中的每个问题文本构造正样本和负样本，生成多个训练样本
S204
将每个训练样本输入预训练后的语义模型中，输出多个样本参数向量
S205
根据所述多个样本参数向量计算损失值
S206
当所述损失值到达预设阈值时，生成预先训练的语义模型

图2

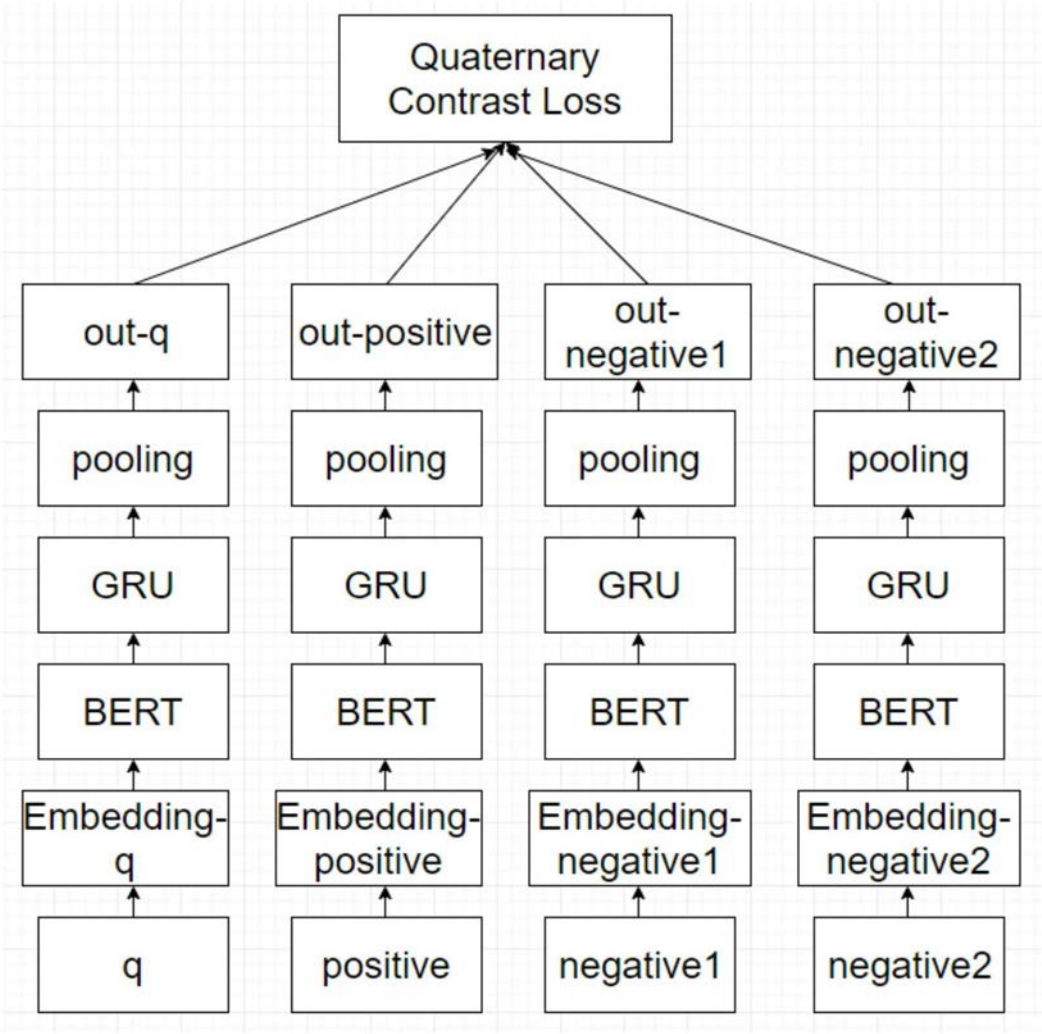

Quaternary Contrast Loss
out-q
out-positive
out-negative1
out-negative2
pooling
pooling
pooling
pooling
GRU
GRU
GRU
GRU
BERT
BERT
BERT
BERT
Embedding-q
Embedding-positive
Embedding-negative1
Embedding-negative2
q
positive
negative1
negative2

图3

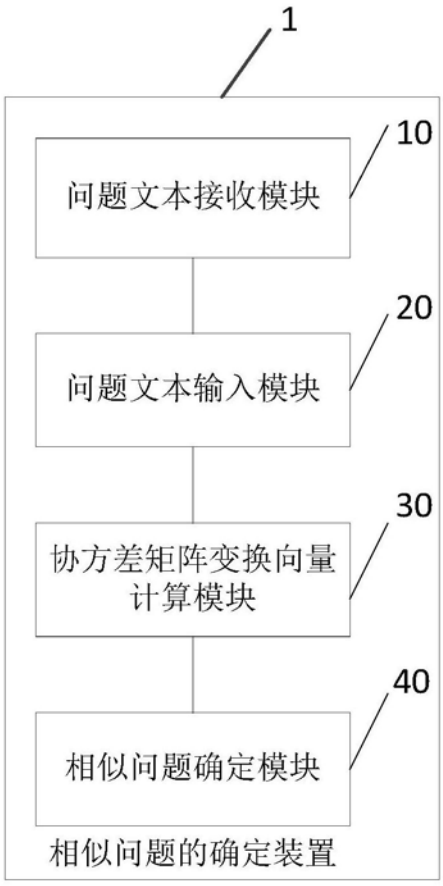

1
10
问题文本接收模块
20
问题文本输入模块
30
协方差矩阵变换向量计算模块
40
相似问题确定模块
相似问题的确定装置

图4

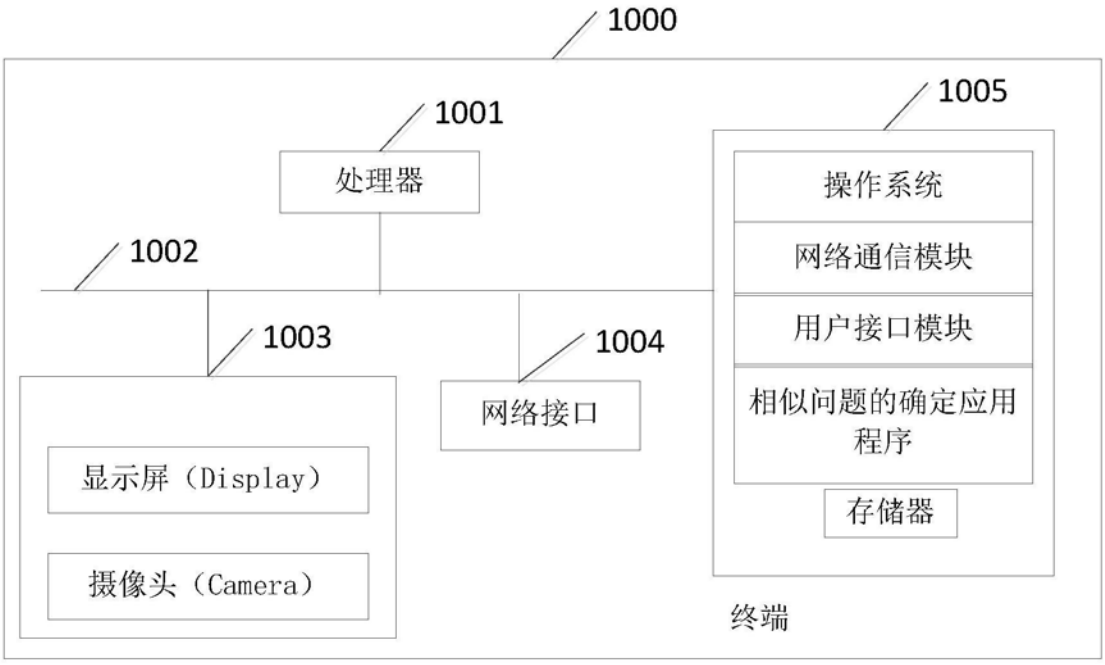

1000
1001
处理器
1002
1003
显示屏（Display）
摄像头（Camera）
1004
网络接口
1005
操作系统
网络通信模块
用户接口模块
相似问题的确定应用程序
存储器
终端

图5