(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 114282528 A
(43)申请公布日 2022.04.05

(21)申请号 202110961476.1

(22)申请日 2021.08.20

(71)申请人 腾讯科技(深圳)有限公司
地址 518044 广东省深圳市南山区高新区科技中一路腾讯大厦35层

(72)发明人 黄剑辉

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

(51)Int.Cl.
G06F 40/253(2020.01)
G06F 40/258(2020.01)
G06F 40/279(2020.01)
G06F 40/30(2020.01)
G06F 16/33(2019.01)
G06F 16/9532(2019.01)
G06N 3/04(2006.01)
G06N 3/08(2006.01)

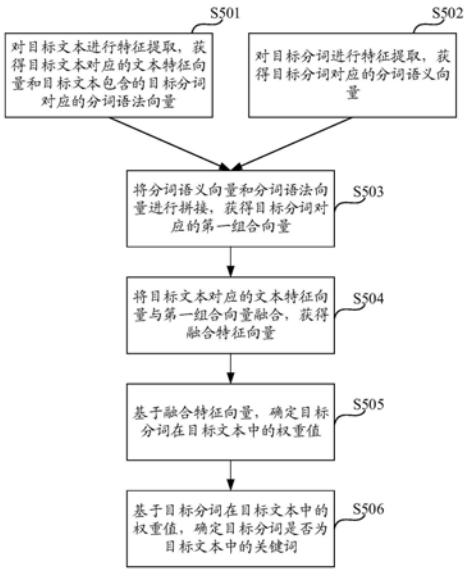
权利要求书3页 说明书16页 附图11页

(54)发明名称

一种关键词提取方法、装置、设备及存储介质

(57)摘要

本申请实施例提供了一种关键词提取方法、装置、设备及存储介质，涉及人工智能技术领域，该方法包括：对目标文本进行特征提取，获得目标文本对应的文本特征向量和目标文本包含的目标分词的分词语法向量。对目标分词进行特征提取，获得目标分词的分词语义向量，然后将分词语义向量和分词语法向量进行拼接获得第一组合向量。由于第一组合向量中包含了目标分词的语义信息和语法信息，故将目标文本的文本特征向量与第一组合向量融合后获得的融合特征向量，可以更好地表征目标文本中的核心成分。基于融合特征向量确定目标分词在目标文本中的权重值，并基于权重值确定目标分词是否为目标文本中的关键词时，可以有效提高提取目标文本中关键词的准确性。



CN 114282528 A

1.一种关键词提取方法,其特征在于,包括:

对目标文本进行特征提取,获得所述目标文本对应的文本特征向量和所述目标文本包含的目标分词对应的分词语法向量,以及对所述目标分词进行特征提取,获得所述目标分词对应的分词语义向量;

将所述分词语义向量和所述分词语法向量进行拼接,获得所述目标分词对应的第一组合向量;将所述目标文本对应的文本特征向量与所述第一组合向量融合,获得融合特征向量;

基于所述融合特征向量,确定所述目标分词在所述目标文本中的权重值,所述权重值用于表征所述目标分词对于所述目标文本的语义理解的影响程度;

基于所述目标分词在所述目标文本中的权重值,确定所述目标分词是否为所述目标文本中的关键词。

2.如权利要求1所述的方法,其特征在于,所述对目标文本进行特征提取,获得所述目标文本对应的文本特征向量和所述目标文本包含的目标分词对应的分词语法向量,包括:

分别提取所述目标文本中各个分词各自对应的分词语法向量、位置向量以及分割向量;其中,每个分词对应一个分词语法向量、一个位置向量和一个分割向量,每个分词语法向量用于表征相应的一个分词在所述目标文本中的语法信息,每个位置向量用于表征相应的一个分词与所述目标文本中其他分词之间的相对位置关系,每个分割向量用于表征相应的一个分词所属语句的语句类型;

分别基于所述各个分词各自对应的分词语法向量、位置向量以及分割向量,获得相应分词对应的第二组合向量;

对获得的各个第二组合向量进行特征提取,获得所述目标文本对应的文本特征向量;

从所述各个分词各自对应的分词语法向量中,获取所述目标分词对应的分词语法向量。

3.如权利要求2所述的方法,其特征在于,所述分别基于所述各个分词各自对应的分词语法向量、位置向量以及分割向量,获得相应分词对应的第二组合向量,包括:

分别针对所述各个分词,执行以下操作:将一个分词对应的分词语法向量、位置向量以及分割向量进行叠加,获得所述一个分词对应的第二组合向量。

4.如权利要求2所述的方法,其特征在于,所述对获得的各个第二组合向量进行特征提取,获得所述目标文本对应的文本特征向量,包括:

根据所述各个第二组合向量与相应的注意力权重矩阵,获得所述各个分词各自对应的注意力权重向量,其中,一个分词对应的注意力权重向量包含的各个值,分别表征所述各个分词各自相对于所述一个分词的注意力权重;

根据所述各个分词各自对应的注意力权重向量,以及所述各个第二组合向量,获得所述目标文本对应的文本特征向量,其中,所述文本特征向量包括所述各个分词各自对应的分词特征向量,每个分词特征向量是根据相应的一个注意力权重向量中各个注意力权重,与相应第二组合向量进行加权求和获得的。

5.如权利要求4所述的方法,其特征在于,所述根据所述各个第二组合向量与相应的注意力权重矩阵,获得所述各个分词各自对应的注意力权重向量,包括:

根据所述各个第二组合向量与相应的注意力权重矩阵,获得所述各个分词各自对应的

至少一个注意力向量，其中，所述至少一个注意力向量包括请求向量和键向量；

基于所述各个分词各自对应的至少一个注意力向量，获取所述各个分词各自对应的注意力权重向量，所述各个分词各自相对于所述一个分词的注意力权重为所述各个分词各自对应的键向量分别与所述一个分词的请求向量的相似度。

6.如权利要求1至5任一所述的方法，其特征在于，所述将所述目标文本对应的文本特征向量与所述目标分词对应的第一组合向量融合，获得融合特征向量，包括：

对所述目标文本对应的文本特征向量与所述目标分词对应的第一组合向量进行点乘处理，获得相应的融合特征向量。

7.如权利要求1至5任一所述的方法，其特征在于，所述基于所述目标分词在所述目标文本中的权重值，确定所述目标分词是否为所述目标文本中的关键词，包括：

若所述目标分词对应的权重值大于等于预设阈值，则确定所述目标分词为所述目标文本中的关键词；

若所述目标分词对应的权重值小于所述预设阈值，则确定所述目标分词不是所述目标文本中的关键词。

8.如权利要求7所述的方法，其特征在于，所述基于所述目标分词在所述目标文本中的权重值，确定所述目标分词是否为所述目标文本中的关键词之后，所述方法还包括：

获取待匹配的目标关键词；

将所述目标关键词与视频标题库中各个视频标题进行关键词匹配，获得至少一个候选视频标题；

根据所述至少一个候选视频标题中各个关键词的权重值，对所述至少一个候选视频标题进行排序；

根据排序结果确定所述目标关键词的匹配视频标题。

9.一种关键词提取装置，其特征在于，包括：

特征提取模块，用于对目标文本进行特征提取，获得所述目标文本对应的文本特征向量和所述目标文本包含的目标分词对应的分词语法向量，以及对所述目标分词进行特征提取，获得所述目标分词对应的分词语义向量；

拼接模块，用于将所述分词语义向量和所述分词语法向量进行拼接，获得所述目标分词对应的第一组合向量；

融合模块，用于将所述目标文本对应的文本特征向量与所述第一组合向量融合，获得融合特征向量；

预测模块，用于基于所述融合特征向量，确定所述目标分词在所述目标文本中的权重值，所述权重值用于表征所述目标分词对于所述目标文本的语义理解的影响程度；

判决模块，用于基于所述目标分词在所述目标文本中的权重值，确定所述目标分词是否为所述目标文本中的关键词。

10.一种计算机设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，其特征在于，所述处理器执行所述程序时实现权利要求1～8任一权利要求所述方法的步骤。

11.一种计算机可读存储介质，其特征在于，其存储有可由计算机设备执行的计算机程序，当所述程序在计算机设备上运行时，使得所述计算机设备执行权利要求1～8任一所述

方法的步骤。

一种关键词提取方法、装置、设备及存储介质

技术领域

[0001] 本发明实施例涉及人工智能技术领域，尤其涉及一种关键词提取方法、装置、设备及存储介质。

背景技术

[0002] 随着信息技术的发展，互联网上无时无刻不在产生大量的数据，面对大量的数据，用户难以快速地从中找到比较重要且关键的内容，因此标题词权重任务(term-weights)应运而生，标题词权重任务指提取文本中的核心成分，消除冗余成分的影响的主要方式。

[0003] 目前，词权重任务大多基于全局语料进行词频统计，来提取文本中的关键词，比如，词频-逆文本频率指数(term frequency-inverse document frequency)。然而，基于统计方式提取文本中的关键词时，其准确性较低。

发明内容

[0004] 本申请实施例提供了一种关键词提取方法、装置、设备及存储介质，用于提高提取文本中关键词的准确性。

[0005] 一方面，本申请实施例提供了一种关键词提取方法，该方法包括：

[0006] 对目标文本进行特征提取，获得所述目标文本对应的文本特征向量和所述目标文本包含的目标分词对应的分词语法向量，以及对所述目标分词进行特征提取，获得所述目标分词对应的分词语义向量；

[0007] 将所述分词语义向量和所述分词语法向量进行拼接，获得所述目标分词对应的第一组合向量；将所述目标文本对应的文本特征向量与所述第一组合向量融合，获得融合特征向量；

[0008] 基于所述融合特征向量，确定所述目标分词在所述目标文本中的权重值，所述权重值用于表征所述目标分词对于所述目标文本的语义理解的影响程度；

[0009] 基于所述目标分词在所述目标文本中的权重值，确定所述目标分词是否为所述目标文本中的关键词。

[0010] 一方面，本申请实施例提供了一种关键词提取装置，该装置包括：

[0011] 特征提取模块，用于对目标文本进行特征提取，获得所述目标文本对应的文本特征向量和所述目标文本包含的目标分词对应的分词语法向量，以及对所述目标分词进行特征提取，获得所述目标分词对应的分词语义向量；

[0012] 拼接模块，用于将所述分词语义向量和所述分词语法向量进行拼接，获得所述目标分词对应的第一组合向量；

[0013] 融合模块，用于将所述目标文本对应的文本特征向量与所述第一组合向量融合，获得融合特征向量；

[0014] 预测模块，用于基于所述融合特征向量，确定所述目标分词在所述目标文本中的权重值，所述权重值用于表征所述目标分词对于所述目标文本的语义理解的影响程度；

[0015] 判决模块，用于基于所述目标分词在所述目标文本中的权重值，确定所述目标分词是否为所述目标文本中的关键词。

[0016] 可选地，所述特征提取模块具体用于：

[0017] 分别提取所述目标文本中各个分词各自对应的分词语法向量、位置向量以及分割向量；其中，每个分词对应一个分词语法向量、一个位置向量和一个分割向量，每个分词语法向量用于表征相应的一个分词在所述目标文本中的语法信息，每个位置向量用于表征相应的一个分词与所述目标文本中其他分词之间的相对位置关系，每个分割向量用于表征相应的一个分词所属语句的语句类型；

[0018] 分别基于所述各个分词各自对应的分词语法向量、位置向量以及分割向量，获得相应分词对应的第二组合向量；

[0019] 对获得的各个第二组合向量进行特征提取，获得所述目标文本对应的文本特征向量；

[0020] 从所述各个分词各自对应的分词语法向量中，获取所述目标分词对应的分词语法向量。

[0021] 可选地，所述特征提取模块具体用于：

[0022] 分别针对所述各个分词，执行以下操作：将一个分词对应的分词语法向量、位置向量以及分割向量进行叠加，获得所述一个分词对应的第二组合向量。

[0023] 可选地，所述特征提取模块具体用于：

[0024] 根据所述各个第二组合向量与相应的注意力权重矩阵，获得所述各个分词各自对应的注意力权重向量，其中，一个分词对应的注意力权重向量包含的各个值，分别表征所述各个分词各自相对于所述一个分词的注意力权重；

[0025] 根据所述各个分词各自对应的注意力权重向量，以及所述各个第二组合向量，获得所述目标文本对应的文本特征向量，其中，所述文本特征向量包括所述各个分词各自对应的分词特征向量，每个分词特征向量是根据相应的一个注意力权重向量中各个注意力权重，与相应第二组合向量进行加权求和获得的。

[0026] 可选地，所述特征提取模块具体用于：

[0027] 根据所述各个第二组合向量与相应的注意力权重矩阵，获得所述各个分词各自对应的至少一个注意力向量，其中，所述至少一个注意力向量包括请求向量和键向量；

[0028] 基于所述各个分词各自对应的至少一个注意力向量，获取所述各个分词各自对应的注意力权重向量，所述各个分词各自相对于所述一个分词的注意力权重为所述各个分词各自对应的键向量分别与所述一个分词的请求向量的相似度。

[0029] 可选地，所述融合模块具体用于：

[0030] 对所述目标文本对应的文本特征向量与所述目标分词对应的第一组合向量进行点乘处理，获得相应的融合特征向量。

[0031] 可选地，所述判别模块具体用于：

[0032] 若所述目标分词对应的权重值大于等于预设阈值，则确定所述目标分词为所述目标文本中的关键词；

[0033] 若所述目标分词对应的权重值小于所述预设阈值，则确定所述目标分词不是所述目标文本中的关键词。

[0034] 可选地，还包括关键词匹配模块；

[0035] 所述关键词匹配模块具体用于：

[0036] 基于所述目标分词在所述目标文本中的权重值，确定所述目标分词是否为所述目标文本中的关键词之后，获取待匹配的目标关键词；

[0037] 将所述目标关键词与视频标题库中各个视频标题进行关键词匹配，获得至少一个候选视频标题；

[0038] 根据所述至少一个候选视频标题中各个关键词的权重值，对所述至少一个候选视频标题进行排序；

[0039] 根据排序结果确定所述目标关键词的匹配视频标题。

[0040] 一方面，本申请实施例提供了一种计算机设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，所述处理器执行所述程序时实现上述关键词提取方法的步骤。

[0041] 一方面，本申请实施例提供了一种计算机可读存储介质，其存储有可由计算机设备执行的计算机程序，当所述程序在计算机设备上运行时，使得所述计算机设备执行上述关键词提取方法的步骤。

[0042] 本申请实施例中，将目标文本包含的目标分词对应的分词语义向量和分词语法向量进行拼接，获得目标分词对应的第一组合向量，由于第一组合向量中包含了目标分词的语义信息和语法信息，故将目标文本对应的文本特征向量与第一组合向量融合后获得的融合特征向量，可以更好地表征目标文本中的核心成分。基于融合特征向量，确定目标分词在目标文本中的权重值，并基于目标文本中的权重值确定目标分词是否为目标文本中的关键词时，可以有效提高提取目标文本中关键词的准确性。

附图说明

[0043] 为了更清楚地说明本发明实施例中的技术方案，下面将对实施例描述中所需要使用的附图作简要介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域的普通技术人员来讲，在不付出创造性劳动性的前提下，还可以根据这些附图获得其他的附图。

[0044] 图1为本申请实施例适用的一种系统架构示意图；

[0045] 图2为本申请实施例适用的一种搜索应用的搜索结果界面示意图；

[0046] 图3为本申请实施例适用的一种词权重模型的网络结构示意图；

[0047] 图4为本申请实施例适用的一种词权重模型训练方法的流程示意图；

[0048] 图5为本申请实施例适用的一种关键词提取方法的流程示意图；

[0049] 图6为本申请实施例适用的一种预测分词权重值的方法的流程示意图；

[0050] 图7为本申请实施例适用的一种预测分词权重值的方法的流程示意图；

[0051] 图8为本申请实施例适用的一种获取第二组合向量的方法的示意图；

[0052] 图9为本申请实施例适用的一种获取第二组合向量的方法的示意图；

[0053] 图10为本申请实施例适用的一种获取第二组合向量的方法的示意图；

[0054] 图11为本申请实施例适用的一种获取第二组合向量的方法的示意图；

[0055] 图12为本申请实施例适用的一种视频应用的搜索结果界面示意图；

[0056] 图13为本申请实施例适用的一种关键词提取装置的结构示意图；

[0057] 图14为本申请实施例适用的一种计算机设备的结构示意图。

具体实施方式

[0058] 为了使本发明的目的、技术方案及有益效果更加清楚明白，以下结合附图及实施例，对本发明进行进一步详细说明。应当理解，此处所描述的具体实施例仅仅用以解释本发明，并不用于限定本发明。

[0059] 为了方便理解，下面对本发明实施例中涉及的名词进行解释。

[0060] 人工智能(Artificial Intelligence,AI)是利用数字计算机或者数字计算机控制的机器模拟、延伸和扩展人的智能，感知环境、获取知识并使用知识获得最佳结果的理论、方法、技术及应用系统。换句话说，人工智能是计算机科学的一个综合技术，它企图了解智能的实质，并生产出一种新的能以人类智能相似的方式做出反应的智能机器。人工智能也就是研究各种智能机器的设计原理与实现方法，使机器具有感知、推理与决策的功能。

[0061] 人工智能技术是一门综合学科，涉及领域广泛，既有硬件层面的技术也有软件层面的技术。人工智能基础技术一般包括如传感器、专用人工智能芯片、云计算、分布式存储、大数据处理技术、操作/交互系统、机电一体化等技术。人工智能软件技术主要包括计算机视觉技术、语音处理技术、自然语言处理技术以及机器学习/深度学习等几大方向。

[0062] 自然语言处理(Nature Language processing,NLP)是计算机科学领域与人工智能领域中的一个重要方向。它研究能实现人与计算机之间用自然语言进行有效通信的各种理论和方法。自然语言处理是一门融语言学、计算机科学、数学于一体的科学。因此，这一领域的研究将涉及自然语言，即人们日常使用的语言，所以它与语言学的研究有着密切的联系。自然语言处理技术通常包括文本处理、语义理解、机器翻译、机器人问答、知识图谱等技术。

[0063] 机器学习(Machine Learning,ML)是一门多领域交叉学科，涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。专门研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能。机器学习是人工智能的核心，是使计算机具有智能的根本途径，其应用遍及人工智能的各个领域。机器学习和深度学习通常包括人工神经网络、置信网络、强化学习、迁移学习、归纳学习、式教学习等技术。本申请实施例提供的方案涉及人工智能的NLP以及ML等技术。

[0064] BERT:(Bidirectional Encoder Representations from Transformers)，即双向Transformer的Encoder，增加词向量模型泛化能力，充分描述字符级、词级、句子级甚至句间关系特征。Transformer模型是2018年5月提出的，可以替代传统循环神经网络(Recurrent Neural Network,RNN)和卷积神经网络(Convolutional Neural Networks,CNN)的一种新的架构，用来实现机器翻译。Transformer模型包括Encoder(编码器)和Decoder(解码器)。

[0065] 下面对本申请实施例的设计思想进行介绍。

[0066] 在自然语言处理技术中，标题词权重任务是提取句子核心语义成分，消除冗余成分的影响的主要方式。目前，词权重任务大多采用统计方式完成，而基于统计的方式仅仅是基于全局语料进行词频统计，无法将词语和句子具体的语义建立联系，从而导致提取句子

中关键词的准确性较低。

[0067] 通过分析发现，在对整个句子进行特征提取时，不仅会提取句子中各个分词的语义信息，同时可以获得各个分词在句子中的语法信息，若结合各个分词的语义信息和语法信息确定各个分词在句子中的权重值，并基于获得的权重值提取句子中的关键词，将有效提高提取句子中关键词的准确性。

[0068] 鉴于此，本申请实施例提供了一种关键词提取方法，在该方法中，对目标文本进行特征提取，获得目标文本对应的文本特征向量和目标文本包含的目标分词对应的分词语法向量，以及对目标分词进行特征提取，获得目标分词对应的分词语义向量。然后将分词语义向量和分词语法向量进行拼接，获得目标分词对应的第一组合向量，将目标文本对应的文本特征向量与第一组合向量融合，获得融合特征向量。再基于融合特征向量，确定目标分词在目标文本中的权重值，其中，权重值用于表征目标分词对于目标文本的语义理解的影响程度。之后再基于目标分词在目标文本中的权重值，确定目标分词是否为目标文本中的关键词。

[0069] 本申请实施例中，将目标文本包含的目标分词对应的分词语义向量和分词语法向量进行拼接，获得目标分词对应的第一组合向量，由于第一组合向量中包含了目标分词的语义信息和语法信息，故将目标文本对应的文本特征向量与第一组合向量融合后获得的融合特征向量，可以更好地表征目标文本中的核心成分。基于融合特征向量，确定目标分词在目标文本中的权重值，并基于目标文本中的权重值确定目标分词是否为目标文本中的关键词时，可以有效提高提取目标文本中关键词的准确性。

[0070] 参考图1，其为本申请实施例提供的关键词提取方法所适用的系统架构图，该架构至少包括终端设备101以及关键词提取设备102。

[0071] 终端设备101中可以安装具备关键词提取功能的目标应用，其中，目标应用可以是客户端应用、网页版应用、小程序应用等。终端设备101可以是智能手机、平板电脑、笔记本电脑、台式计算机、智能音箱、智能手表等，但并不局限于此。

[0072] 关键词提取设备102可以是目标应用的后台服务器，为目标应用提供相应的服务，关键词提取设备102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。终端设备101与关键词提取设备102可以通过有线或无线通信方式进行直接或间接地连接，本申请在此不做限制。

[0073] 本申请实施例中的关键词提取方法可以由终端设备101执行，也可以由关键词提取设备102执行。

[0074] 以上述关键词提取方法由关键词提取设备102执行展开来说：

[0075] 终端设备101获取目标文本，并将目标文本发送至关键词提取设备102，关键词提取设备102对目标文本进行特征提取，获得目标文本对应的文本特征向量和目标文本包含的目标分词对应的分词语法向量。以及，对目标分词进行特征提取，获得目标分词对应的分词语义向量。然后将分词语义向量和分词语法向量进行拼接，获得目标分词对应的第一组合向量，将目标文本对应的文本特征向量与第一组合向量融合，获得融合特征向量。再基于融合特征向量，确定目标分词在目标文本中的权重值，其中，权重值用于表征目标分词对于

目标文本的语义理解的影响程度。之后再基于目标分词在目标文本中的权重值，确定目标分词是否为目标文本中的关键词。关键词提取设备102采用同样的方式确定目标文本中其他分词是否为目标文本中的关键词，最终提取出目标文本中包含的所有关键词。

[0076] 在实际应用中，本申请实施例提供的方案可以适用于所有需要理解文本核心词、词权重场景，比如标题理解、篇章句式理解、文本搜索、视频搜索、内容推荐等场景。

[0077] 举例来说，在文本检索场景中，关键词提取设备采用本申请实施例中的方法提取文章标题库中每个文章标题中的关键词以及每个关键词对应的权重值，然后将文章标题中的关键词以及每个关键词对应的权重值与文章标题对应保存在文章标题库中。

[0078] 终端设备响应于用户在搜索应用中触发的文章检索操作，发送用户输入的待匹配关键词“XX公园”至关键词提取设备。关键词提取设备将待匹配关键词与文章标题库中各个文章标题进行关键词匹配，确定待匹配关键词的匹配文章标题为文章标题A“XX公园官网”和文章标题B“XX公园中的樱花图片”。关键词提取设备将文章标题A和文章标题A对应的封面图像，以及文章标题B和文章标题B对应的封面图像发送至终端设备。终端设备在搜索应用的搜索结果界面展示文章标题A和文章标题B以及相应的封面图像。

[0079] 参见图2，在搜索结果界面的第一区域201展示文章标题A和文章标题A对应的封面图像，在搜索结果界面的第二区域202展示文章标题B和文章标题B对应的封面图像。

[0080] 除此之外，用户也可以在搜索应用中输入一个句子，终端设备响应于用户在搜索应用中触发的文章检索操作，发送用户输入的句子至关键词提取设备。关键词提取设备可以先对用户输入的句子进行关键词提取，然后基于提取的关键字与标题库中各个文章标题进行关键词匹配，确定用户输入的句子的匹配文章标题。

[0081] 本申请实施例提供的关键词提取方法可以通过图1中的终端设备101或关键词提取设备102来执行，该方法的具体流程介绍如下：

[0082] 本申请实施例中，关键词提取的过程可以包括标题词权重任务和关键词提取任务两个过程，其中，标题词权重任务是指获取目标文本中各个分词的权重值，关键词提取任务指根据获得的权重值进行关键词的提取，其中，权重值可用于表征各个分词对于目标文本的语义理解的影响程度。

[0083] 本申请实施例中，为了提升标题词权重任务的准确性，可以采用基于深度学习的神经网络模型来获取目标文本中各个分词的权重值。下面，以一种可能的神经网络模型为例对本申请实施例的技术方案进行介绍。

[0084] 参见图3所示，为本申请实施例提供的词权重模型的网络结构示意图，其中，该词权重模型可以包括第一编码器(encoder)、第二编码器、拼接层、融合层以及分类层，第一编码器用于对目标文本进行特征编码，获得目标文本的文本特征向量和目标文本包含的n个分词对应的分词语法向量，其中，n大于等于1。第二编码器用于对目标文本包含的分词i进行特征编码，获得分词i的分词语义向量x，$1 \leqslant x \leqslant n$，$1 \leqslant i \leqslant n$。

[0085] 拼接层用于拼接第一编码器输出分词i的分词语法向量y和第二编码器输出的分词i的分词语义向量，获得第一组合向量。融合层用于融合第一编码器输出的目标文本的文本特征向量和拼接层输出的第一组合向量，获得融合特征向量。分类层基于融合特征向量进行分类，从而获得目标文本中的分词i在目标文本中的权重值。由于各个层所执行的过程将在后续详细进行介绍，因而在此先不过多进行介绍。

[0086] 在词权重模型投入使用之前，首先需要对该模型进行训练，因此，下面先对词权重模型的训练过程进行介绍。请参见图4，为词权重模型的训练流程示意图。

[0087] 步骤401：获取多个训练样本。

[0088] 本申请实施例中，每个训练样本可以包括一个标题和该标题中的一个分词，并且每个训练样本标注了该训练样本中分词是否为关键词的标签(label)，其中，标签可以通过两种不同的值来表征是否为关键词，例如可以用0和1进行表示，当分词为关键词时，则该训练样本的标签为1，当分词不是关键词时，则该训练样本的标签为0，或者，也可以当分词为关键词时，则该训练样本的标签为0，否则，当分词不是关键词时，则该训练样本的标签为1，当然，也可以采用其他可能的值来进行表示，本申请实施例对此并不进行限制。

[0089] 如表1所示，为训练样本的数据格式的示意，其中，以标题为“跳一跳，教你上600分的攻略”为例，其可以构成多个训练样本。通过对该标题的语义理解可知，该标题的核心内容在于跳一跳的攻略，因此，将“跳一跳”作为训练样本的分词时，由于“跳一跳”对于该标题的语义理解的作用是很大的，即“跳一跳”为核心词，因此其对应的训练样本的标签为1，同理，“攻略”对应的训练样本的标签也为1，那么{“跳一跳，教你上600分的攻略”，“跳一跳”}和{“跳一跳，教你上600分的攻略”，“攻略”}作为标注的正样本。

[0090] 而对于标题中的分词“教你”或者“的”相对而言，其对于该标题的语义理解的作用较小，即“教你”或者“的”为非核心词，因而其对应的训练样本的标签为0，那么{“跳一跳，教你上600分的攻略”，“教你”}和{“跳一跳，教你上600分的攻略”，“的”作为标注的负样本。

[0091] 表1.

[0092]

标题	当前词	标签
跳一跳，教你上600分的攻略	跳一跳	1
跳一跳，教你上600分的攻略	攻略	1
跳一跳，教你上600分的攻略	教你	0
跳一跳，教你上600分的攻略	的	0

[0093] 通过上述的过程，可以获得多个正样本和负样本，从而通过构造的训练样本对词权重模型进行训练。训练样本的格式可以如下：

[0094] 正样本：{“标题”：“跳一跳，教你上600分的攻略”，“当前词”：“跳一跳”，“label”：1}

[0095] 负样本：{“标题”：“跳一跳，教你上600分的攻略”，“当前词”：“教你”，“label”：0}

[0096] 当然，也可以采用其他可能的样本格式，本申请实施例对此不做限制。

[0097] 步骤402：利用词权重模型确定各个训练样本中分词的预测权重值。

[0098] 本申请实施例中，每个训练样本均包括一个标题和该标题中的一个分词，那么通过词权重模型可以确定每个训练样本中分词的权重值，其中，词权重模型确定权重值的过程将在后续具体进行介绍，因此这里先不过多赘述。

[0099] 步骤403：根据获得的各个预测权重值以及训练样本中的标签指示的权重值，确定词权重模型的损失函数。

[0100] 具体地，可以采用负对数交叉熵作为词权重模型的损失函数，如以下公式(1)所示：

[0101] $$loss = -\sum_{i=1}^{n}(y_i log a_i) \quad (1)$$

[0102] 其中，loss表示词权重模型的损失值，y_i表示第i个训练样本的标签指示的权重值，a_i表示第i个训练样本的预测权重值，n表示训练样本的数量。

[0103] 一般而言，当获得的权重值与标签指示的权重值之间的差异程度越小，例如标签为1时，而词权重模型获得的预测权重值为0.95，或者，标签为0时，而词权重模型获得的预测权重值为0.02，那么采用上述公式(1)获得的词权重模型的损失值越小，说明词权重模型预测的权重值更为接近真实值，因此其准确程度是更高的。

[0104] 步骤404：根据损失函数确定词权重模型是否收敛。

[0105] 步骤405：当确定词权重模型未收敛时，根据损失值对词权重模型进行模型参数的调整。

[0106] 步骤406：当确定词权重模型收敛时，结束训练。

[0107] 本申请实施例中，当损失值小于设定的损失阈值时，则表明词权重模型的准确度能够达到要求，因而可以确定词权重模型收敛，相反的，当损失值不小于设定的损失阈值时，则表明词权重模型的准确度未能达到要求，那么还需要进一步的对词权重模型进行参数调整，并通过参数调整后的词权重模型进行后续的训练流程，即重复进行步骤402～404的过程。

[0108] 本申请实施例中，在训练获得词权重模型之后，则可以利用已训练的词权重模型进行关键词提取，请参见图5，包括以下步骤：

[0109] 步骤S501，对目标文本进行特征提取，获得目标文本对应的文本特征向量和目标文本包含的目标分词对应的分词语法向量。

[0110] 具体地，目标文本可以是文章或者视频等多媒体内容的标题，也可以是文章或者视频等多媒体内容中的句子。文本特征向量用于表征目标文本的语义信息，目标分词对应的分词语法向量用于表征目标分词在目标文本中的语法信息。

[0111] 在获得目标文本之后，对目标文本进行分词操作，获得目标文本中的各个分词。分词(Word Segmentation)操作的过程是指将一个句子切分成一个个单独的词，可以通过任意可能的分词方法进行分词操作，例如字符匹配方法、理解法或者统计法，也可以采用相应的分词工具进行分词操作，例如结巴(jieba)分词等。

[0112] 本申请实施例中，可以通过第一编码器对目标文本进行特征提取，获得目标文本对应的文本特征向量和目标文本包含的各个分词对应的分词语法向量，其中，第一编码器可以采用任意可能的语义编码方式进行特征编码，以获得目标文本的文本特征向量以及各个分词对应的分词语法向量，例如可以采用BERT(Bidirectional Encoder Representation from Transformers)、卷积神经网络(Convolutional Neural Networks，CNN)、长短期记忆人工神经网络(Long-Short Term Memory，LSTM)或者LSTM结合注意力(Attention)机制等多种方式完成。

[0113] 步骤S502，对目标分词进行特征提取，获得目标分词对应的分词语义向量。

[0114] 具体的，目标分词对应的分词语义向量用于表征目标分词的语义信息。可以通过第二编码器对目标文本包含的各个分词进行特征编码，获得目标分词对应的分词语义向量，其中，第二编码器采用任意可能的词编码方式进行特征编码，例如可以采用深度神经网络(Deep Neural Networks，DNN)等方式完成特征空间的映射变换。

[0115] 上述第一编码器和第二编码器可以通过共同训练获得，也可以单独进行训练。同时，步骤501的过程与步骤502的过程可以是同时进行的，也可以是先后顺序进行的，对此，本申请不做具体限定。

[0116] 步骤S503，将分词语义向量和分词语法向量进行拼接，获得目标分词对应的第一组合向量。

[0117] 本申请实施例中，采用词权重模型中的拼接层将分词语义向量和分词语法向量进行拼接，输出目标分词对应的第一组合向量。

[0118] 具体地，可以将分词语义向量拼接在分词语法向量的末尾，也可以将分词语法向量拼接在分词语义向量的末尾。

[0119] 举例来说，第一组合向量的格式为：

[0120] Lword＝[tokenj:wordemb]

[0121] 其中，Lword表示第一组合向量，tokeni表示分词i对应的分词语法向量，wordemb表示分词i对应的分词语义向量。

[0122] 步骤S504，将目标文本对应的文本特征向量与第一组合向量融合，获得融合特征向量。

[0123] 本申请实施例中，采用词权重模型中的融合层，将目标文本对应的文本特征向量与第一组合向量融合，获得融合特征向量。

[0124] 具体地，可以对目标文本对应的文本特征向量与目标分词对应的第一组合向量进行点乘处理，获得相应的融合特征向量；也可以采用注意力机制将目标文本对应的文本特征向量与目标分词对应的第一组合向量融合，获得相应的融合特征向量；还可以采用双线性池化方式，将目标文本对应的文本特征向量与目标分词对应的第一组合向量融合，获得相应的融合特征向量。

[0125] 步骤S505，基于融合特征向量，确定目标分词在目标文本中的权重值。

[0126] 本申请实施例中，采用词权重模型中的分类层，基于融合特征向量，确定目标分词在目标文本中的权重值。分类层可以通过任何可能的分类算法来实现，例如可以通过softmax算法、逻辑回归(Logistic)或者全连接层等来进行分类，以获得各个分词所对应的分类结果，分类结果即为各个分词的权重值。

[0127] 具体地，权重值用于表征目标分词对于目标文本的语义理解的影响程度，权重值越大，说明目标分词对于目标文本的语义理解的影响程度越高，权重值越小，说明目标分词对于目标文本的语义理解的影响程度越低。

[0128] 采用词权重模型也可以确定目标文本中其他分词在目标文本中的权重值，此处不再赘述。

[0129] 步骤S506，基于目标分词在目标文本中的权重值，确定目标分词是否为目标文本中的关键词。

[0130] 在一种可能的实施方式中，若目标分词对应的权重值大于等于预设阈值，则确定目标分词为所述目标文本中的关键词。若目标分词对应的权重值小于预设阈值，则确定目标分词不是目标文本中的关键词。采用相同的方式可以确定目标文本中其他分词是否为目标文本中的关键词，进而可以得到目标文本中的所有关键词。

[0131] 在另一种可能的实施方式中，采用词权重模型获得目标分词中各个分词对应的权

重值，然后按照权重值从大到小的顺序排序，将排在前N位的权重值对应的分词作为目标文本中的关键词，其中，N为预设正整数。

[0132] 举例来说，如图6所示，在文章搜索场景中，将文章标题“第一公园的樱花又盛开了，等你来观赏”作为目标标题。将该目标标题输入至词权重模型之后，可以得到各个分词的权重值，其中，“第一公园”的权重值为0.91，“樱花”的权重值为0.81，“盛开”的权重值为0.7，“等你”的权重值为0.3，“观赏”的权重值为0.2，“来”的权重值为0.2，“又”的权重值为0.1。

[0133] 基于权重值对各个分词进行排序，将排在前3位的分词“第一公园”、“樱花”“盛开”作为目标标题的关键词，提取的关键词可以应用于文章搜索过程。举例来说，如图7所示，将视频标题“暖心宝宝安慰刚领养回家的狗狗获高赞”作为目标标题输入至词权重模型之后，可以得到各个分词的权重值，其中，“宝宝”的权重值为0.85，“狗狗”的权重值为0.83，“安慰”的权重值为0.82，“获高赞”的权重值为0.51，“暖心”的权重值为0.3，“领养”的权重值为0.1，“回家”的权重值为0.01。基于权重值对各个分词进行排序，将排在前3位的分词“宝宝”、“狗狗”“安慰”作为目标标题的关键词，提取的关键词可以应用于视频搜索过程。

[0134] 本申请实施例中，将目标文本包含的目标分词对应的分词语义向量和分词语法向量进行拼接，获得目标分词对应的第一组合向量，由于第一组合向量中包含了目标分词的语义信息和语法信息，故将目标文本对应的文本特征向量与第一组合向量融合后获得的融合特征向量，可以更好地表征目标文本中的核心成分。基于融合特征向量，确定目标分词在目标文本中的权重值，并基于目标文本中的权重值确定目标分词是否为目标文本中的关键词时，可以有效提高提取目标文本中关键词的准确性。

[0135] 可选地，在上述步骤S501中，以其中一种方式为例介绍对目标文本进行特征提取的过程，包括以下步骤：

[0136] 步骤S5011，分别提取目标文本中各个分词各自对应的分词语法向量、位置向量以及分割向量。

[0137] 具体地，每个分词对应一个分词语法向量、一个位置向量和一个分割向量，每个分词语法向量用于表征相应的一个分词在目标文本中的语法信息。每个位置向量用于表征相应的一个分词与目标文本中其他分词之间的相对位置关系，其可以通过分词在目标文本中的序号进行表示，也可以通过其前后存在的词向量进行表示。每个分割向量用于表征相应的一个分词所属语句的语句类型。

[0138] 步骤S5012，分别基于各个分词各自对应的分词语法向量、位置向量以及分割向量，获得相应分词对应的第二组合向量。

[0139] 参见图8所示，为各个分词的第二组合向量获取示意图，其中，目标文本中包括n个分词，分别为分词1、分词2、…、分词n，其中，分词1对应的位置向量、分割向量以及分词语法向量分别表示为Ec1、Eb1和Ea1。分词2对应的位置向量、分割向量以及分词语法向量分别表示为Ec2、Eb2和Ea2，依次类推。

[0140] 在获取各个分词的位置向量、分割向量以及分词语法向量之后，则可以基于位置向量、分割向量以及分词语法向量，获取相应分词的第二组合向量，其中，分词1对应的第二组合向量表示为E1，分词2对应的第二组合向量表示为E2，分词n对应的第二组合向量表示为En。第二组合向量为能够同时体现出位置向量、分割向量以及分词语法向量所包含的信

息的向量。

[0141] 具体的，针对每个分词，可以对位置向量、分割向量以及分词语法向量进行叠加，从而获得相应分词的第二组合向量。

[0142] 以分词1为例，如图9所示，将分词1的位置向量Ec1、分割向量Eb1和分词语法向量Ea1在相同位置上的值进行叠加后，获得分词1的第二组合向量E1。

[0143] 具体的，还可以对分词语法向量、位置向量以及分割向量进行拼接，从而获得相应分词的第二组合向量。

[0144] 以分词1为例，如图10所示，可以将分词1的分割向量Eb1拼接至位置向量Ec1的后面，再将分词语法向量Ea1拼接至分割向量Eb1的后面，从而获得分词1的第二组合向量E1。

[0145] 具体的，还可以对分词语法向量、位置向量以及分割向量进行池化处理，从而获得相应分词的第二组合向量。

[0146] 如图11所示，同样以分词1为例，在进行最大池化处理时，将分词1的位置向量Ec1、分割向量Eb1和分词语法向量Ea1在相同位置上的值取最大值，从而获得分词1的第二组合向量E1。

[0147] 步骤S5013，对获得的各个第二组合向量进行特征提取，获得目标文本对应的文本特征向量。

[0148] 本申请实施例中，可以将目标文本的各个分词的第二组合向量进行组合，从而获得目标文本对应的文本特征向量。

[0149] 或者，仅进行组合可能无法体现出各个分词之间的语义关系，因此，还可以对各个第二组合向量进行特征提取，从而将提取后的向量进行组合，从而得到目标文本对应的文本特征向量。

[0150] 在一种可能的实施方式中，可以采用自注意力机制对各个第二组合向量进行特征提取，获得目标文本对应的文本特征向量。

[0151] 具体地，根据各个第二组合向量与相应的注意力权重矩阵，获得各个分词各自对应的注意力权重向量，其中，一个分词对应的注意力权重向量包含的各个值，分别表征各个分词各自相对于一个分词的注意力权重。

[0152] 具体实施中，注意力权重矩阵可以包括请求(query)向量矩阵、键(key)向量矩阵，根据各个第二组合向量与相应的注意力权重矩阵，获得各个分词各自对应的至少一个注意力向量，至少一个注意力向量包括query向量、key向量。

[0153] 进而，可以基于各个分词各自对应的至少一个注意力向量，获取各个分词各自对应的注意力权重向量，其中，一个分词对应的注意力权重向量包含的各个值，分别表征各个分词各自相对于一个分词的注意力权重。例如，目标文本包含4个分词，那么对于其中的分词1而言，分词1的注意力权重向量包含4个值，每个值表示目标文本包含的一个分词对于分词1的注意力权重。

[0154] 可选地，各个分词各自相对于一个分词的注意力权重，为各个分词各自对应的键向量分别与一个分词的请求向量的相似度。

[0155] 具体的，分词2对分词1的注意力权重可以通过分词2的key向量与分词1的query向量之间的相似度来获得，同理，其他分词也是如此，而分词1对分词1的注意力权重可以通过分词1的key向量与分词1的query向量之间的相似度来获得。

[0156] 接着，可以根据各个分词各自对应的注意力权重向量，以及各个第二组合向量，获得目标文本对应的文本特征向量，其中，文本特征向量包括各个分词各自对应的分词特征向量，每个分词特征向量是根据相应的一个注意力权重向量中各个注意力权重，与相应第二组合向量进行加权求和获得的。

[0157] 以分词1举例来说，将分词1对应的注意力权重向量中各个注意力权重，与分词1对应的第二组合向量中各个值加权求和，获得分词1的分词特征向量。

[0158] 可选地，注意力权重矩阵还包括值(value)向量矩阵，相应地，至少一个注意力向量还包括value向量。

[0159] 可以根据各个分词各自对应的注意力权重向量，以及各个分词各自对应的至少一个注意力向量，获得目标文本对应的文本特征向量，其中，文本特征向量包括各个分词各自对应的分词特征向量，每个分词特征向量是根据相应的一个注意力权重向量中各个注意力权重，与相应注意力向量进行加权求和获得的。

[0160] 以分词1举例来说，将分词1对应的注意力权重向量中各个注意力权重，与分词1的value向量中各个值加权求和，获得分词1的分词特征向量。

[0161] 步骤S5014，从各个分词各自对应的分词语法向量中，获取目标分词对应的分词语法向量。

[0162] 具体地，词权重模型预测目标文本中的目标分词在目标文本中的权重值时，第一编码器分别提取目标文本中各个分词各自对应的分词语法向量、位置向量以及分割向量，第二编码器对目标分词进行特征编码，获得目标分词对应的分词语义向量。第一编码器中提取的各个分词各自对应的分词语法向量中包含目标分词对应的分词语法向量，因此，可以从各个分词各自对应的分词语法向量中获取目标分词对应的分词语法向量，然后结合目标分词对应的分词语法向量、分词语义向量以及目标文本对应的文本特征向量，预测目标分词在目标文本中的权重值。

[0163] 本申请实施例中，提取分词对应的分词语法向量、位置向量以及分割向量，并将多个维度的特征向量进行拼接，获得分词的第二组合向量。然后采用自注意力机制对各个第二组合向量进行特征提取，获得目标文本对应的文本特征向量，使得文本特征向量可以更好地表征各个分词之间的语义关系和语法关系，进而提高确定的分词权重值的准确性。

[0164] 本申请实施例中，以上描述的关键词提取方法可以适用于所有需要理解文本核心词、词权重场景，比如标题理解、篇章句式理解、文本检索、视频搜索、内容推荐等场景。

[0165] 以视频搜索举例来说，先获取待匹配的目标关键词，然后将目标关键词与视频标题库中各个视频标题进行关键词匹配，获得至少一个候选视频标题。根据至少一个候选视频标题中各个关键词的权重值，对至少一个候选视频标题进行排序。根据排序结果确定目标关键词的匹配视频标题。

[0166] 具体地，视频标题库中每个视频标题对应的关键词以及关键词对应的权重值均可以采用本申请实施例中的关键词提取方法获得。将目标关键词与视频标题库中各个视频标题进行关键词匹配，若视频标题中包含目标关键词，则将该视频标题作为候选视频标题，同时获得目标关键词在候选视频标题中的权重值。然后按照权重值从大到小的顺序，对各个候选视频标题进行排序，将排序结果中排在前M位的候选视频标题作为目标关键词的匹配视频标题，其中，M为预设正整数。之后再将获得的匹配视频标题以及相应的视频封面图像

发送至终端设备，终端设备可以在视频应用中展示获得的匹配视频标题以及相应的视频封面图像。

[0167] 举例来说，用户在视频应用的搜索界面输入目标关键词“足球比赛”后提交，终端设备响应于用户在视频应用中触发的视频搜索操作，发送用户输入的目标关键词“足球比赛”至关键词提取设备。关键词提取设备将目标关键词“足球比赛”与视频标题库中各个视频标题进行关键词匹配，获得4个包含该目标关键词的候选视频标题，分别为候选视频标题1、候选视频标题2、候选视频标题3和候选视频标题4，其中，目标关键词在候选视频标题1中的权重值为0.7，目标关键词在候选视频标题2中的权重值为0.8，目标关键词在候选视频标题3中的权重值为0.9，目标关键词在候选视频标题4中的权重值为0.6。

[0168] 按照权重值从大到小的顺序，对各个候选视频标题进行排序，获得的排序结果为：候选视频标题3、候选视频标题2、候选视频标题1和候选视频标题4。将排在前两位的候选视频标题3和候选视频标题2，作为目标关键词的匹配视频标题。

[0169] 关键词提取设备将候选视频标题3和候选视频标题2以及相应的视频封面图像发送至终端设备，终端设备在视频应用中的搜索结果界面中展示候选视频标题3和候选视频标题2以及相应的视频封面图像。

[0170] 参见如图12，在搜索结果界面的第一区域1201展示候选视频标题3和候选视频标题3对应的视频封面图像，在搜索结果界面的第二区域1202展示候选视频标题2和候选视频标题2对应的视频封面图像，其中，候选视频标题3具体内容为“足球比赛：A队对战B队”，候选视频标题2具体内容为“足球比赛集锦”。

[0171] 本申请实施例中，同时获取句子的语义向量和句中词的语法信息，融合语法和语义信息的模型在实验中能更好的表征句子中的核心信息。

[0172] 采用词权重模型同时获取目标文本的语义向量，以及目标文本中各个分词的分词语义向量和分词语法向量，将目标文本的语义向量以及分词的分词语义向量和分词语法向量融合后的融合特征向量可以更好地表征目标文本中的核心信息，故基于融合特征向量预测分词在目标文本中的权重值时，提高了获得的分词权重值的准确性，进而提高了基于分词权重值提取目标文本中关键词的准确性，也使得在各类场景下对文本进行语义理解的准确性。

[0173] 基于相同的技术构思，本申请实施例提供了一种关键词提取装置的结构示意图，如图13所示，该装置1300包括：

[0174] 特征提取模块1301，用于对目标文本进行特征提取，获得所述目标文本对应的文本特征向量和所述目标文本包含的目标分词对应的分词语法向量，以及对所述目标分词进行特征提取，获得所述目标分词对应的分词语义向量；

[0175] 拼接模块1302，用于将所述分词语义向量和所述分词语法向量进行拼接，获得所述目标分词对应的第一组合向量；

[0176] 融合模块1303，用于将所述目标文本对应的文本特征向量与所述第一组合向量融合，获得融合特征向量；

[0177] 预测模块1304，用于基于所述融合特征向量，确定所述目标分词在所述目标文本中的权重值，所述权重值用于表征所述目标分词对于所述目标文本的语义理解的影响程度；

[0178] 判决模块1305，用于基于所述目标分词在所述目标文本中的权重值，确定所述目标分词是否为所述目标文本中的关键词。

[0179] 可选地，所述特征提取模块1301具体用于：

[0180] 分别提取所述目标文本中各个分词各自对应的分词语法向量、位置向量以及分割向量；其中，每个分词对应一个分词语法向量、一个位置向量和一个分割向量，每个分词语法向量用于表征相应的一个分词在所述目标文本中的语法信息，每个位置向量用于表征相应的一个分词与所述目标文本中其他分词之间的相对位置关系，每个分割向量用于表征相应的一个分词所属语句的语句类型；

[0181] 分别基于所述各个分词各自对应的分词语法向量、位置向量以及分割向量，获得相应分词对应的第二组合向量；

[0182] 对获得的各个第二组合向量进行特征提取，获得所述目标文本对应的文本特征向量；

[0183] 从所述各个分词各自对应的分词语法向量中，获取所述目标分词对应的分词语法向量。

[0184] 可选地，所述特征提取模块1301具体用于：

[0185] 分别针对所述各个分词，执行以下操作：将一个分词对应的分词语法向量、位置向量以及分割向量进行叠加，获得所述一个分词对应的第二组合向量。

[0186] 可选地，所述特征提取模块1301具体用于：

[0187] 根据所述各个第二组合向量与相应的注意力权重矩阵，获得所述各个分词各自对应的注意力权重向量，其中，一个分词对应的注意力权重向量包含的各个值，分别表征所述各个分词各自相对于所述一个分词的注意力权重；

[0188] 根据所述各个分词各自对应的注意力权重向量，以及所述各个第二组合向量，获得所述目标文本对应的文本特征向量，其中，所述文本特征向量包括所述各个分词各自对应的分词特征向量，每个分词特征向量是根据相应的一个注意力权重向量中各个注意力权重，与相应第二组合向量进行加权求和获得的。

[0189] 可选地，所述特征提取模块1301具体用于：

[0190] 根据所述各个第二组合向量与相应的注意力权重矩阵，获得所述各个分词各自对应的至少一个注意力向量，其中，所述至少一个注意力向量包括请求向量和键向量；

[0191] 基于所述各个分词各自对应的至少一个注意力向量，获取所述各个分词各自对应的注意力权重向量，所述各个分词各自相对于所述一个分词的注意力权重为所述各个分词各自对应的键向量分别与所述一个分词的请求向量的相似度。

[0192] 可选地，所述融合模块1303具体用于：

[0193] 对所述目标文本对应的文本特征向量与所述目标分词对应的第一组合向量进行点乘处理，获得相应的融合特征向量。

[0194] 可选地，所述判别模块1305具体用于：

[0195] 若所述目标分词对应的权重值大于等于预设阈值，则确定所述目标分词为所述目标文本中的关键词；

[0196] 若所述目标分词对应的权重值小于所述预设阈值，则确定所述目标分词不是所述目标文本中的关键词。

[0197] 可选地，还包括关键词匹配模块1306；

[0198] 所述关键词匹配模块1306具体用于：

[0199] 基于所述目标分词在所述目标文本中的权重值，确定所述目标分词是否为所述目标文本中的关键词之后，获取待匹配的目标关键词；

[0200] 将所述目标关键词与视频标题库中各个视频标题进行关键词匹配，获得至少一个候选视频标题；

[0201] 根据所述至少一个候选视频标题中各个关键词的权重值，对所述至少一个候选视频标题进行排序；

[0202] 根据排序结果确定所述目标关键词的匹配视频标题。

[0203] 本申请实施例中，将目标文本包含的目标分词对应的分词语义向量和分词语法向量进行拼接，获得目标分词对应的第一组合向量，由于第一组合向量中包含了目标分词的语义信息和语法信息，故将目标文本对应的文本特征向量与第一组合向量融合后获得的融合特征向量，可以更好地表征目标文本中的核心成分。基于融合特征向量，确定目标分词在目标文本中的权重值，并基于目标文本中的权重值确定目标分词是否为目标文本中的关键词时，可以有效提高提取目标文本中关键词的准确性。

[0204] 基于相同的技术构思，本申请实施例提供了一种计算机设备，如图14所示，包括至少一个处理器1401，以及与至少一个处理器连接的存储器1402，本申请实施例中不限定处理器1401与存储器1402之间的具体连接介质，图14中处理器1401和存储器1402之间通过总线连接为例。总线可以分为地址总线、数据总线、控制总线等。

[0205] 在本申请实施例中，存储器1402存储有可被至少一个处理器1401执行的指令，至少一个处理器1401通过执行存储器1402存储的指令，可以执行上述关键词提取方法的步骤。

[0206] 其中，处理器1401是计算机设备的控制中心，可以利用各种接口和线路连接计算机设备的各个部分，通过运行或执行存储在存储器1402内的指令以及调用存储在存储器1402内的数据，从而提取目标文本中的关键词。可选的，处理器1401可包括一个或多个处理单元，处理器1401可集成应用处理器和调制解调处理器，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，调制解调处理器主要处理无线通信。可以理解的是，上述调制解调处理器也可以不集成到处理器1401中。在一些实施例中，处理器1401和存储器1402可以在同一芯片上实现，在一些实施例中，它们也可以在独立的芯片上分别实现。

[0207] 处理器1401可以是通用处理器，例如中央处理器(CPU)、数字信号处理器、专用集成电路(Application Specific Integrated Circuit,ASIC)、现场可编程门阵列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，可以实现或者执行本申请实施例中公开的各方法、步骤及逻辑框图。通用处理器可以是微处理器或者任何常规的处理器等。结合本申请实施例所公开的方法的步骤可以直接体现为硬件处理器执行完成，或者用处理器中的硬件及软件模块组合执行完成。

[0208] 存储器1402作为一种非易失性计算机可读存储介质，可用于存储非易失性软件程序、非易失性计算机可执行程序以及模块。存储器1402可以包括至少一种类型的存储介质，例如可以包括闪存、硬盘、多媒体卡、卡型存储器、随机访问存储器(Random Access Memory,RAM)、静态随机访问存储器(Static Random Access Memory,SRAM)、可编程只读存

储器(Programmable Read Only Memory,PROM)、只读存储器(Read Only Memory,ROM)、带电可擦除可编程只读存储器(Electrically Erasable Programmable Read-Only Memory,EEPROM)、磁性存储器、磁盘、光盘等等。存储器1402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。本申请实施例中的存储器1402还可以是电路或者其它任意能够实现存储功能的装置,用于存储程序指令和/或数据。

[0209] 基于同一发明构思,本申请实施例提供了一种计算机可读存储介质,其存储有可由计算机设备执行的计算机程序,当程序在计算机设备上运行时,使得计算机设备执行上述关键词提取方法的步骤。

[0210] 本领域内的技术人员应明白,本发明的实施例可提供为方法、或计算机程序产品。因此,本发明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,本发明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。

[0211] 本发明是参照根据本发明实施例的方法、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指令实现流程图和/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指令到通用计算机、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指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。

[0212] 这些计算机程序指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指令产生包括指令装置的制造品,该指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。

[0213] 这些计算机程序指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指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。

[0214] 尽管已描述了本发明的优选实施例,但本领域内的技术人员一旦得知了基本创造性概念,则可对这些实施例作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发明范围的所有变更和修改。

[0215] 显然,本领域的技术人员可以对本发明进行各种改动和变型而不脱离本发明的精神和范围。这样,倘若本发明的这些修改和变型属于本发明权利要求及其等同技术的范围之内,则本发明也意图包含这些改动和变型在内。

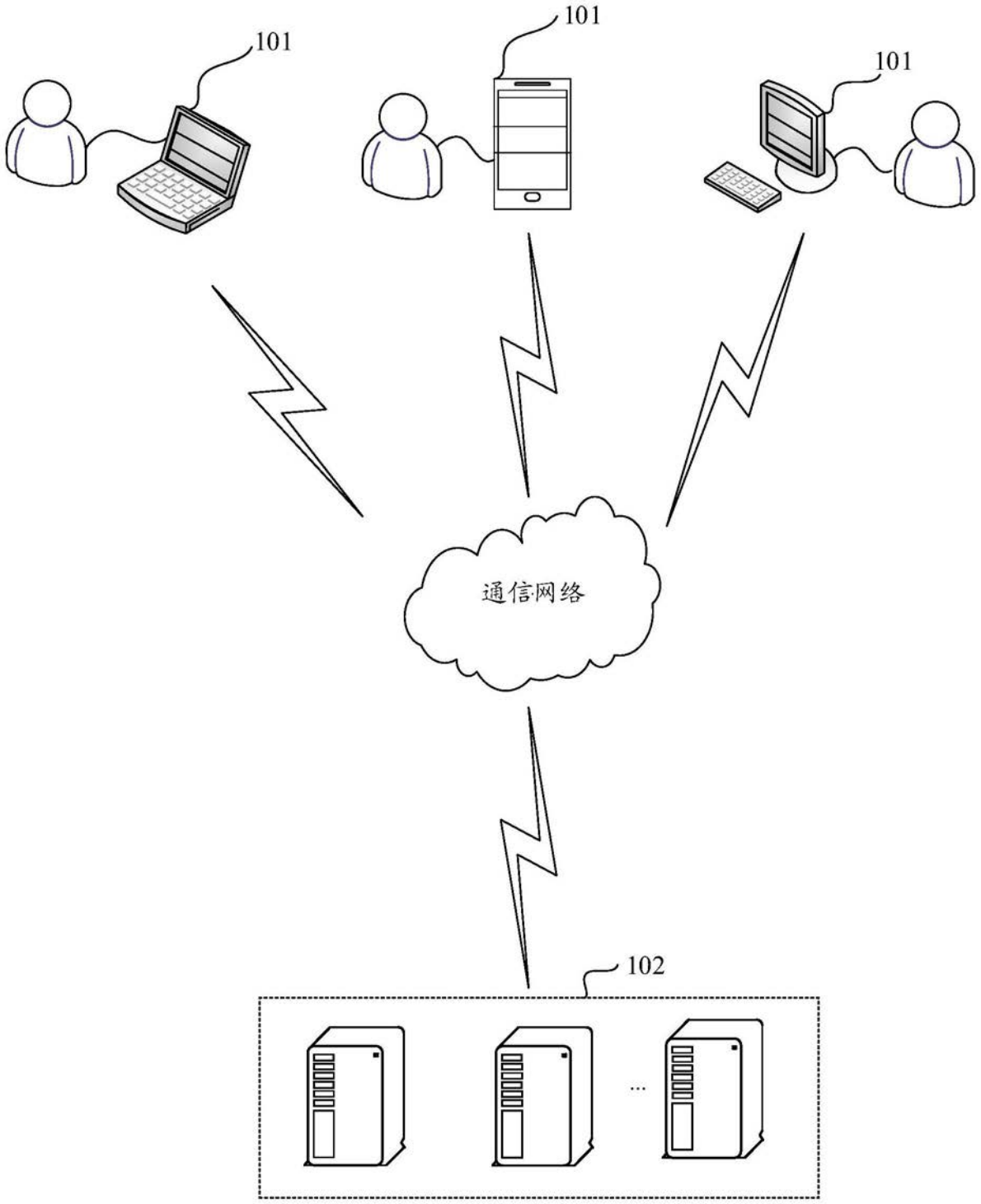


图1

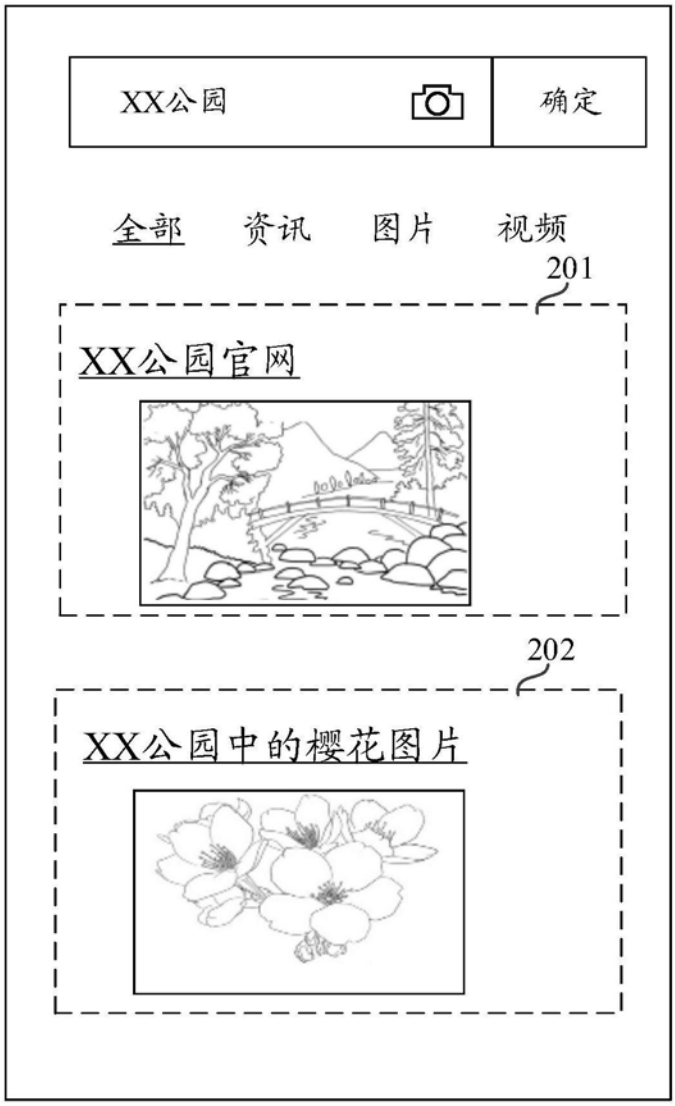


图2

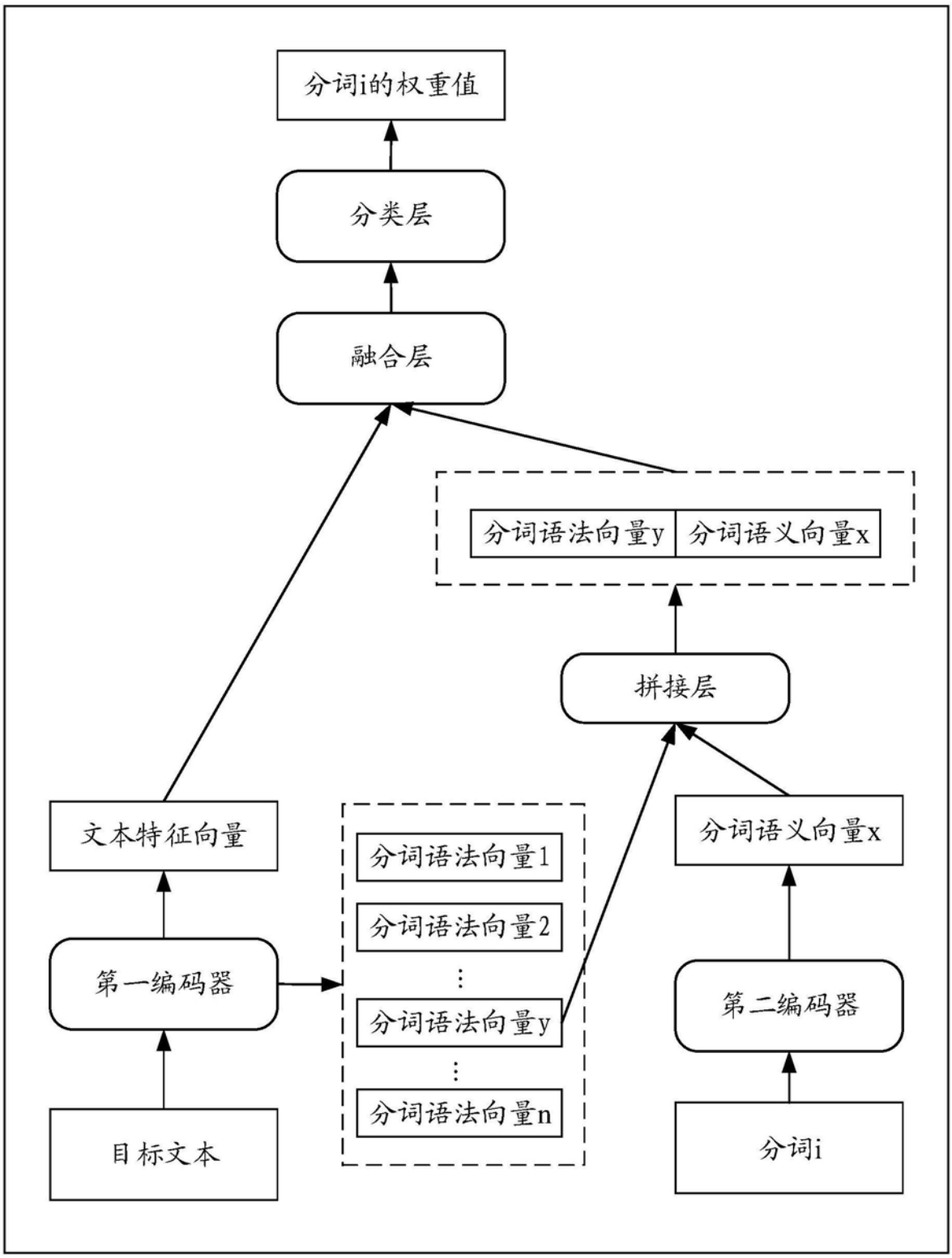

分词i的权重值
分类层
融合层
分词语法向量y
分词语义向量x
拼接层
文本特征向量
分词语法向量1
分词语法向量2
分词语法向量y
分词语法向量n
分词语义向量x
第一编码器
第二编码器
目标文本
分词i

图3

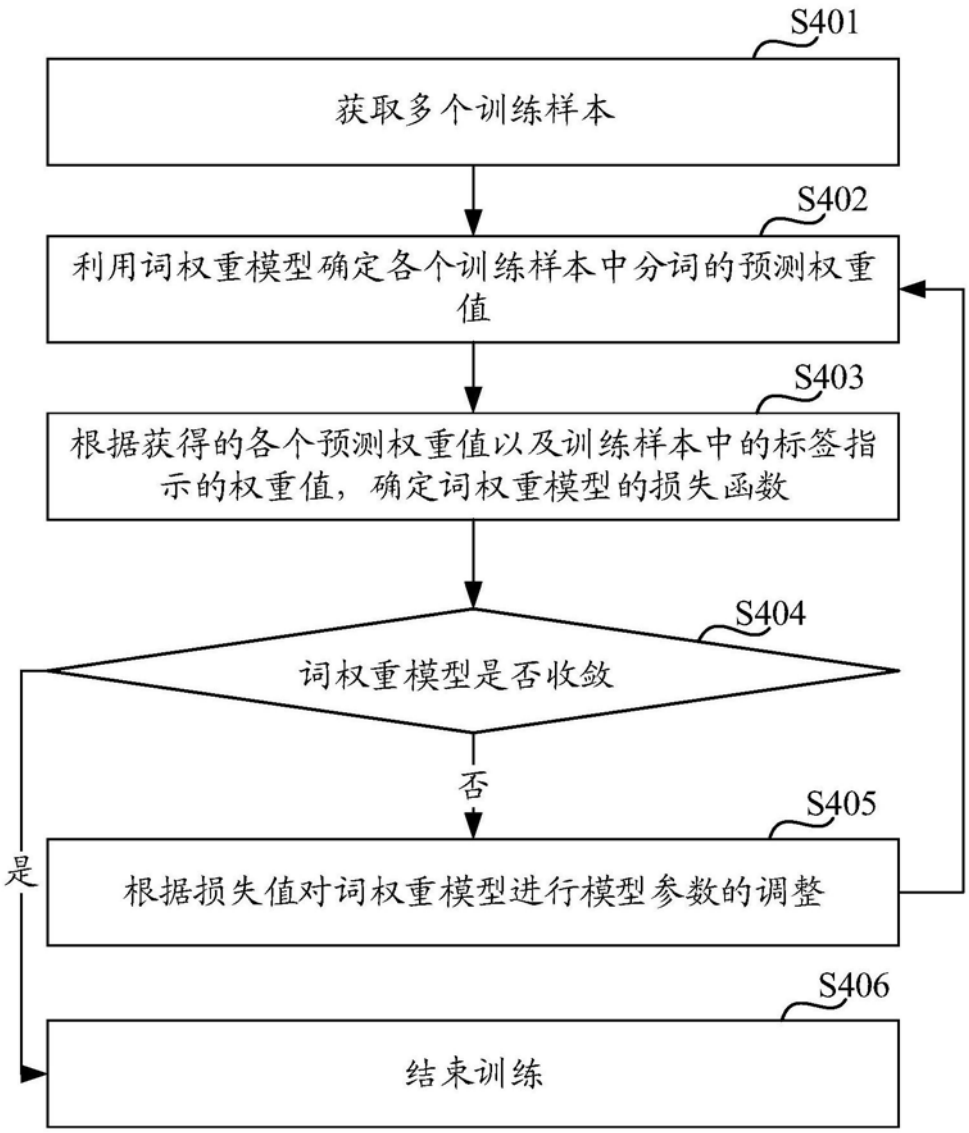

S401
获取多个训练样本
S402
利用词权重模型确定各个训练样本中分词的预测权重值
S403
根据获得的各个预测权重值以及训练样本中的标签指示的权重值，确定词权重模型的损失函数
S404
词权重模型是否收敛
否
是
S405
根据损失值对词权重模型进行模型参数的调整
S406
结束训练

图4

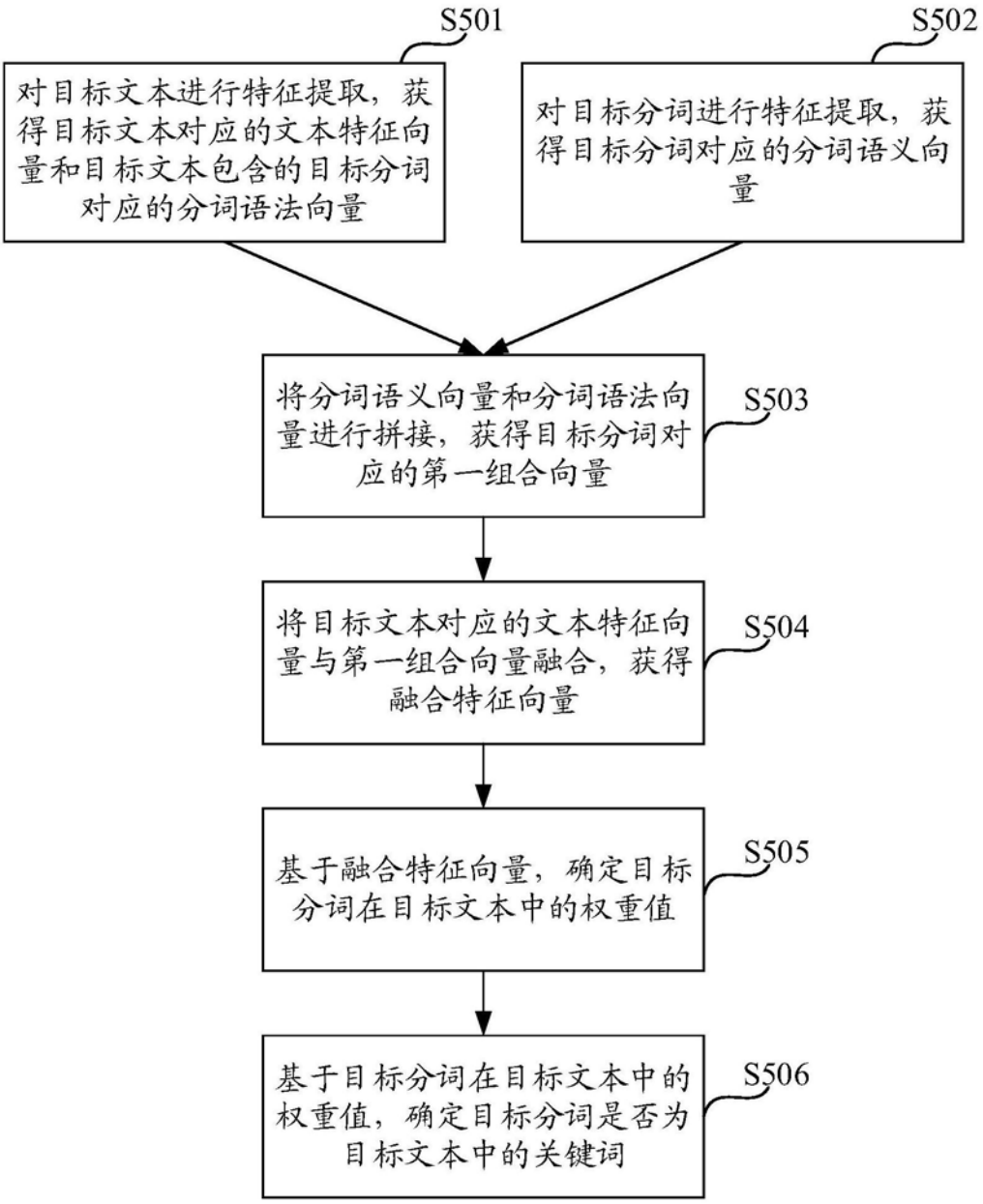

S501
对目标文本进行特征提取，获得目标文本对应的文本特征向量和目标文本包含的目标分词对应的分词语法向量
S502
对目标分词进行特征提取，获得目标分词对应的分词语义向量
将分词语义向量和分词语法向量进行拼接，获得目标分词对应的第一组合向量
S503
将目标文本对应的文本特征向量与第一组合向量融合，获得融合特征向量
S504
基于融合特征向量，确定目标分词在目标文本中的权重值
S505
基于目标分词在目标文本中的权重值，确定目标分词是否为目标文本中的关键词
S506

图5

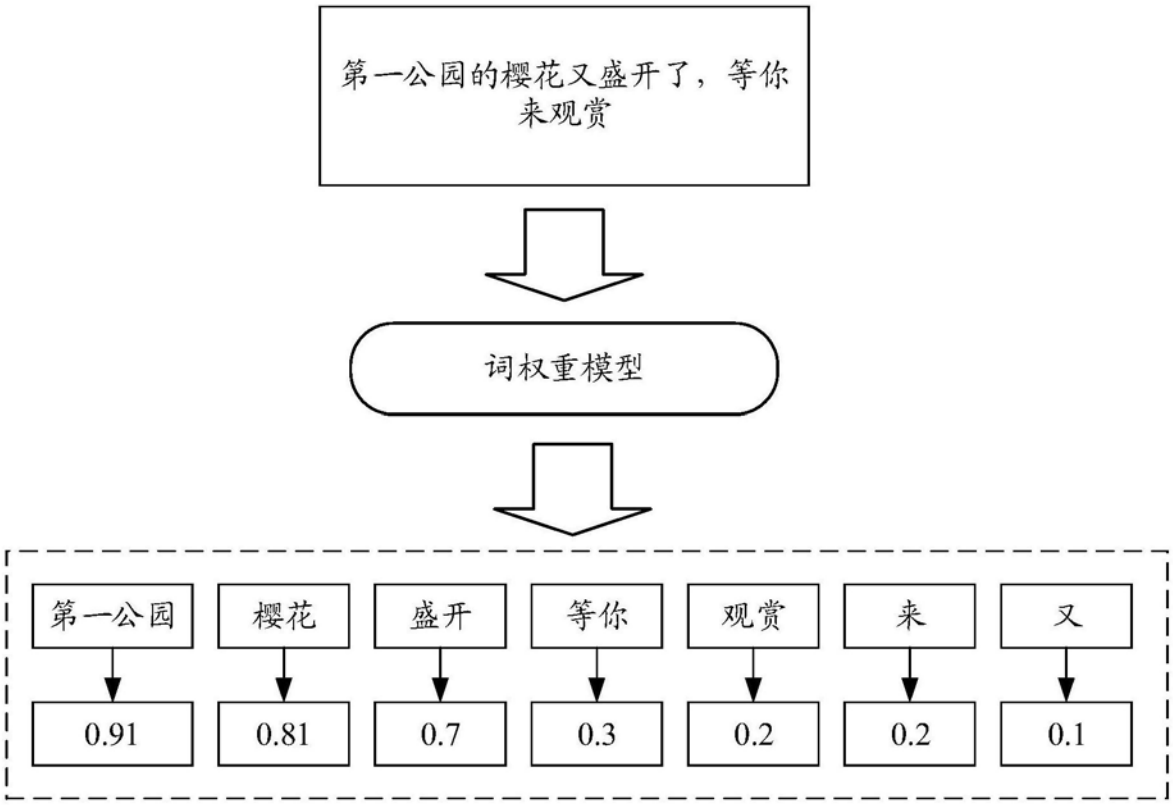

第一公园的樱花又盛开了，等你来观赏
词权重模型
第一公园
樱花
盛开
等你
观赏
来
又
0.91
0.81
0.7
0.3
0.2
0.2
0.1

图6

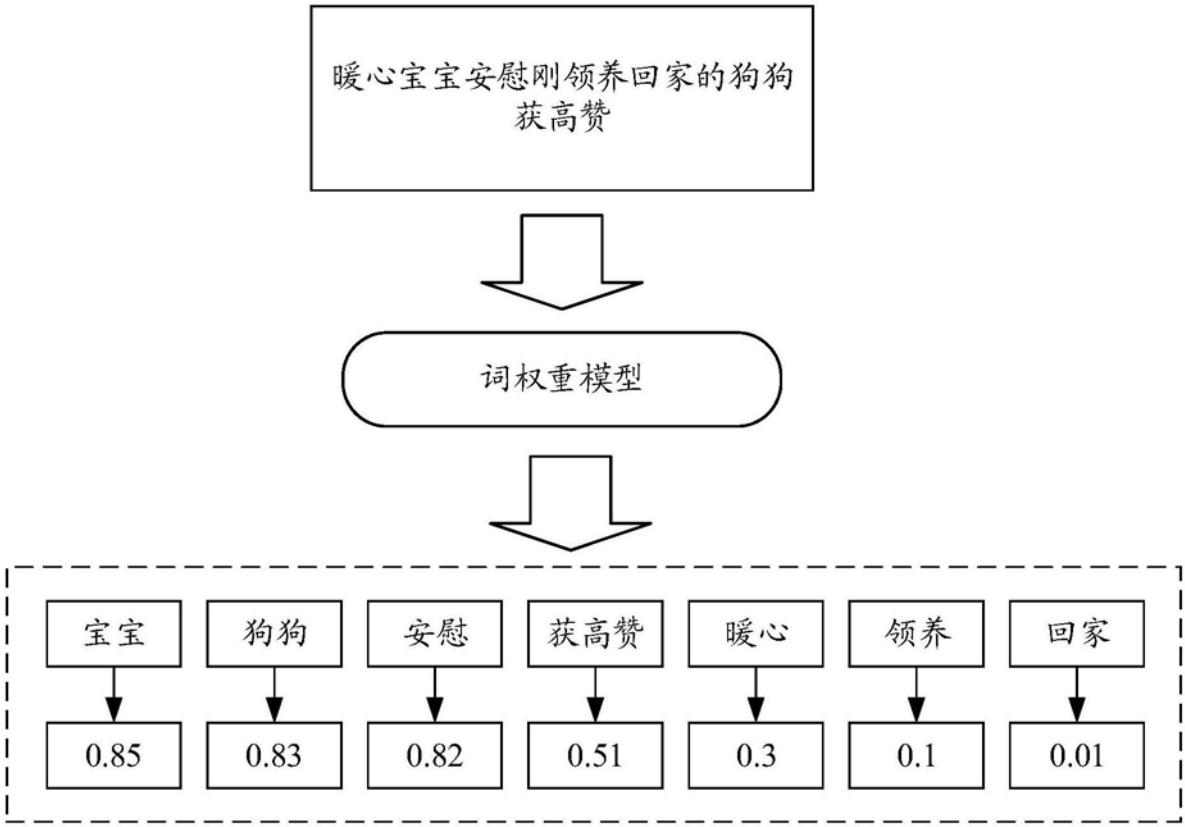

暖心宝宝安慰刚领养回家的狗狗获高赞
词权重模型
宝宝
狗狗
安慰
获高赞
暖心
领养
回家
0.85
0.83
0.82
0.51
0.3
0.1
0.01

图7

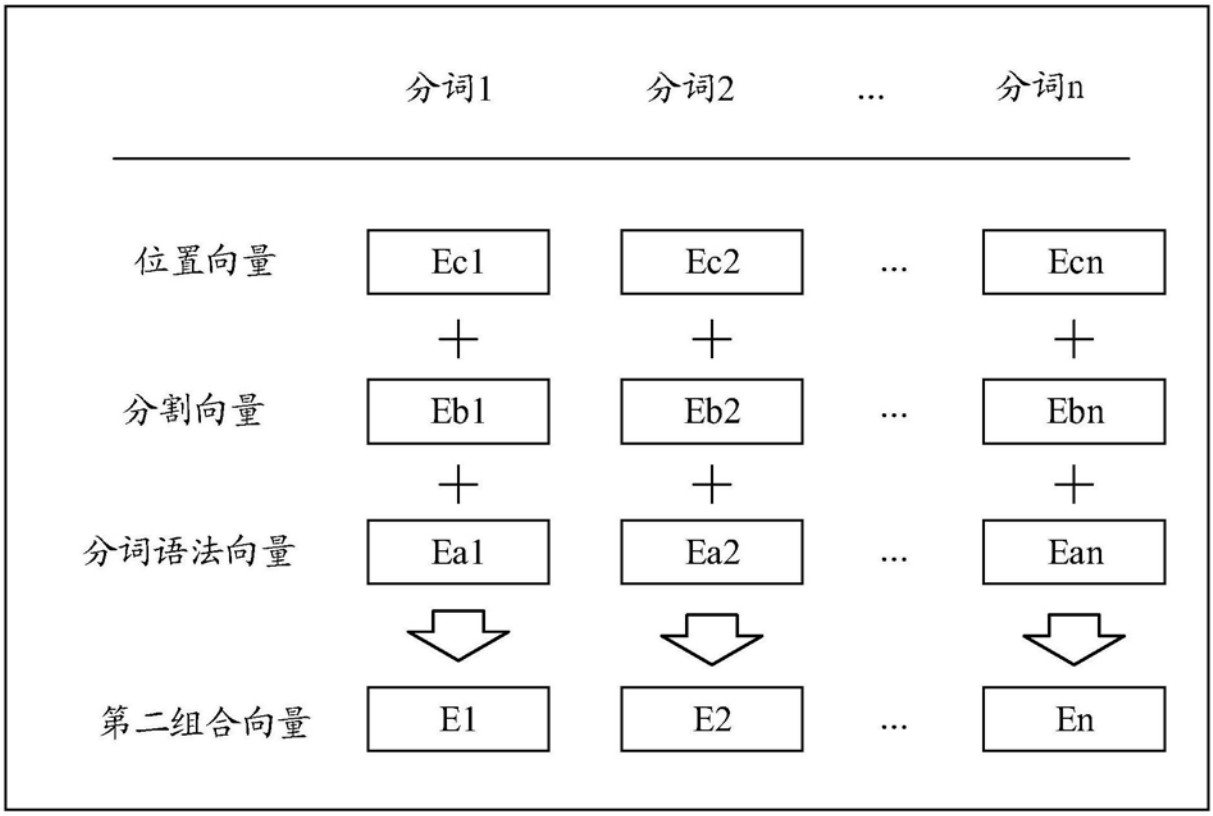

分词1
分词2
...
分词n
位置向量
Ec1
Ec2
...
Ecn
分割向量
Eb1
Eb2
...
Ebn
分词语法向量
Ea1
Ea2
...
Ean
第二组合向量
E1
E2
...
En

图8

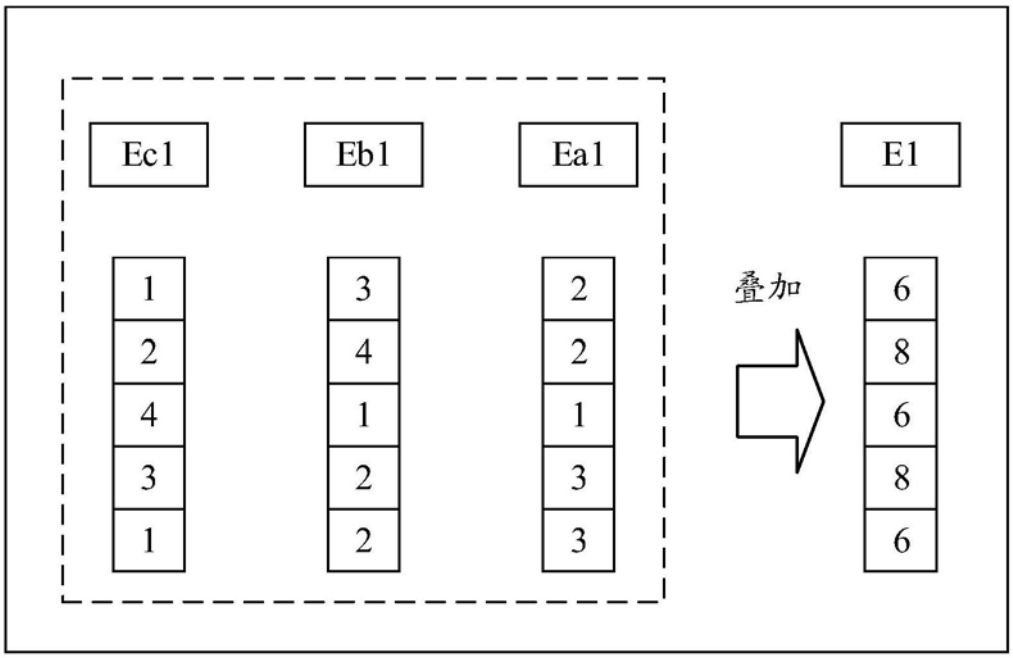

Ec1
Eb1
Ea1
E1
1
2
4
3
1
3
4
1
2
2
2
2
1
3
3
叠加
6
8
6
8
6

图9

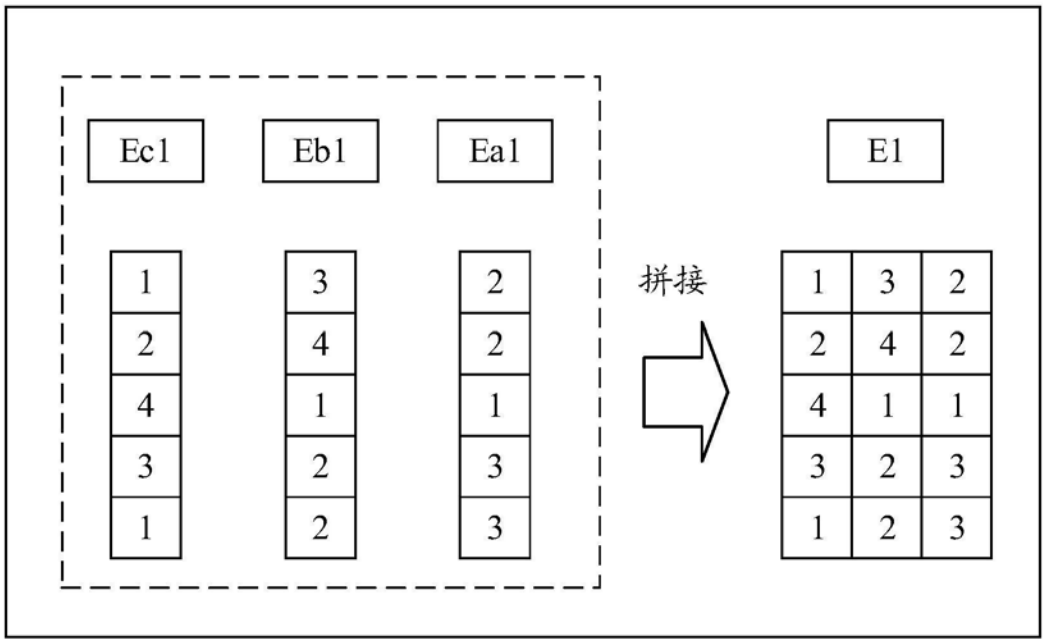

Ec1
Eb1
Ea1
E1
拼接
1 2 4 3 1
3 4 1 2 2
2 2 1 3 3
1 3 2
2 4 2
4 1 1
3 2 3
1 2 3

图10

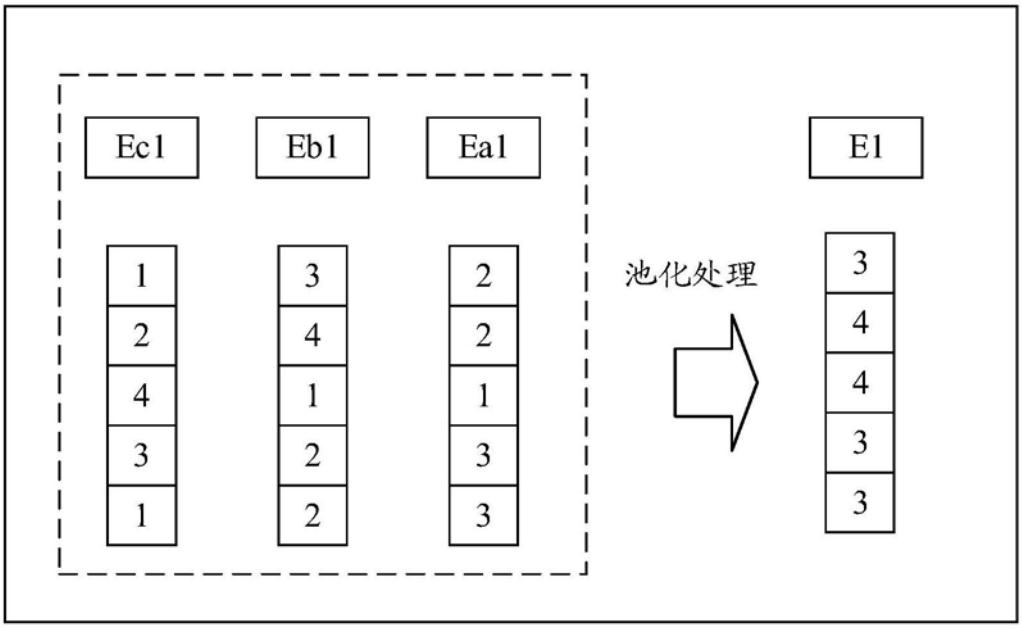

Ec1
Eb1
Ea1
E1
池化处理
1 2 4 3 1
3 4 1 2 2
2 2 1 3 3
3 4 4 3 3

图11

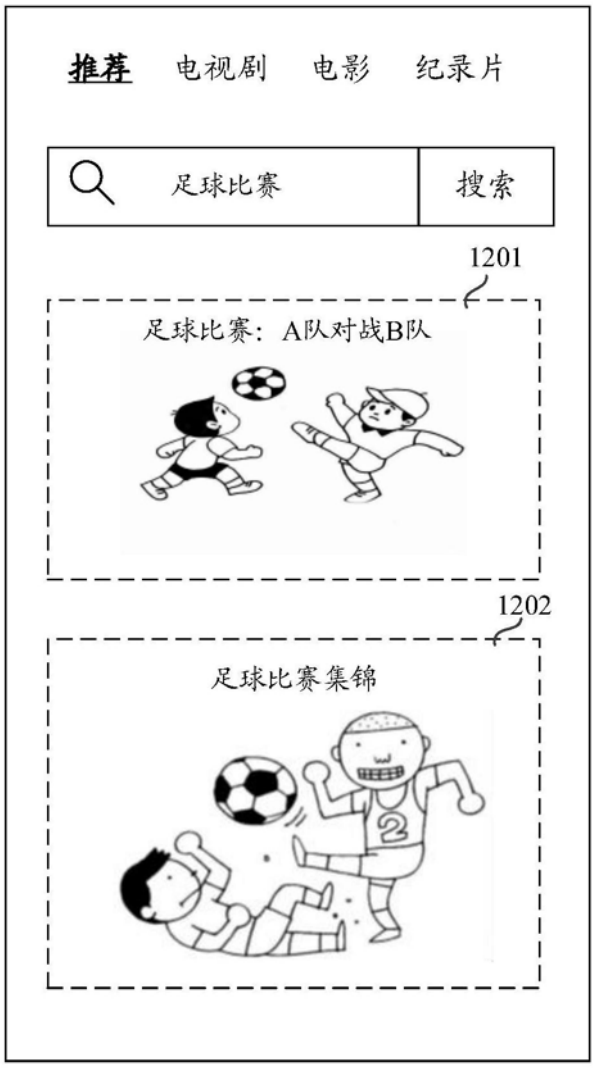


图12

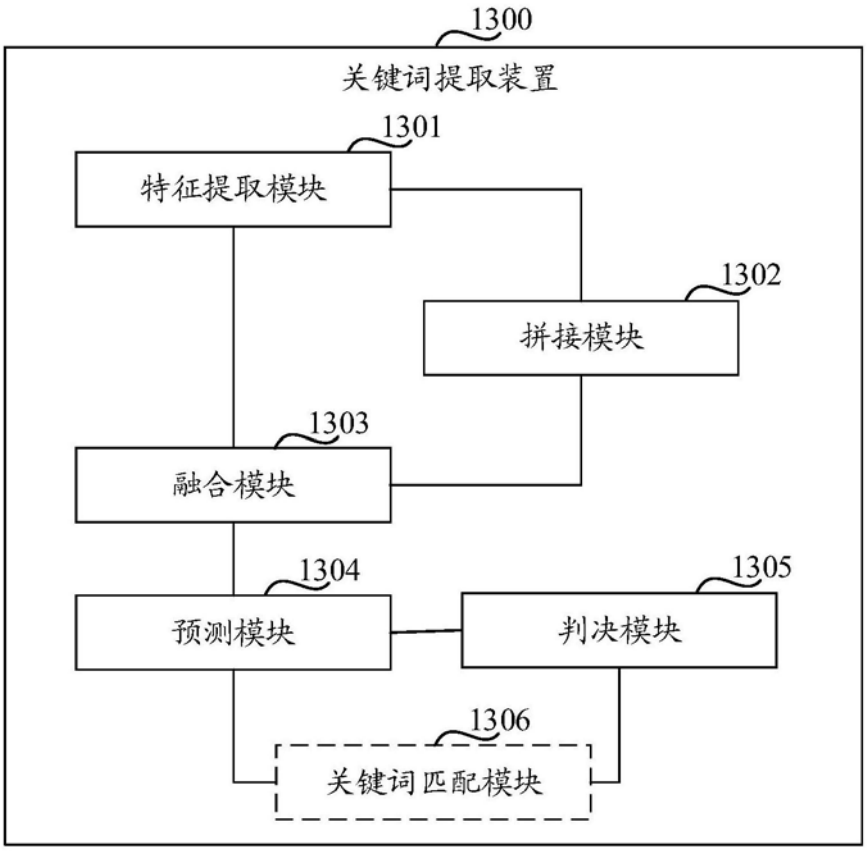

1300
关键词提取装置
1301
特征提取模块
1302
拼接模块
1303
融合模块
1304
预测模块
1305
判决模块
1306
关键词匹配模块

图13

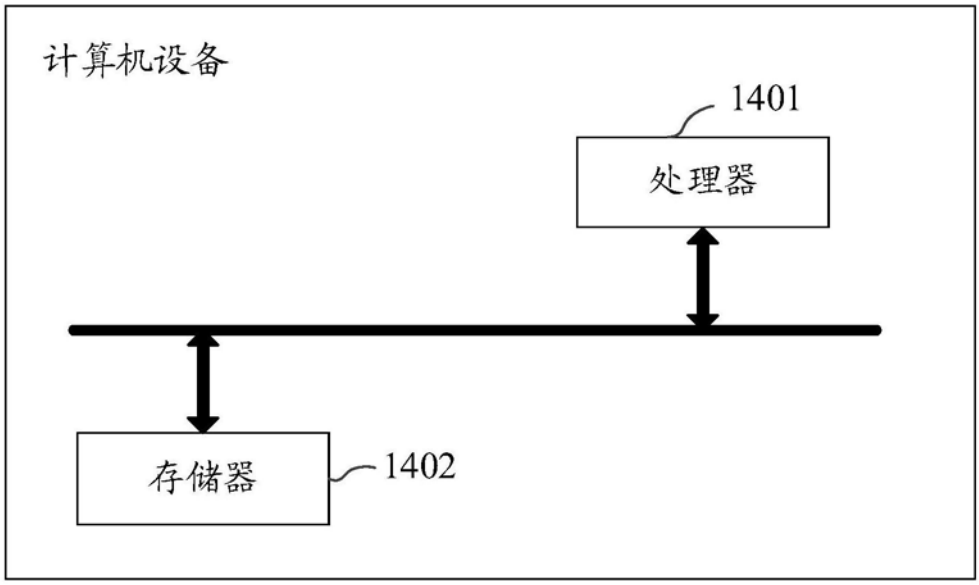

计算机设备
1401
处理器
存储器
1402

图14