(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 110276023 B
(45)授权公告日 2021.04.02

(21)申请号 201910537388.1

(22)申请日 2019.06.20

(65)同一申请的已公布的文献号
申请公布号 CN 110276023 A

(43)申请公布日 2019.09.24

(73)专利权人 北京百度网讯科技有限公司
地址 100085 北京市海淀区上地十街10号百度大厦2层

(72)发明人 潘禄 梁海金 陈玉光 彭卫华 罗雨 刘远圳 韩翠云 施茜

(74)专利代理机构 北京品源专利代理有限公司 11332
代理人 孟金喆

(51)Int.Cl.
G06F 16/9537(2019.01)
G06F 40/279(2020.01)

(56)对比文件
CN 108647582 A,2018.10.12
CN 105260361 A,2016.01.20
US 2017031895 A1,2017.02.02
王红斌等.基于word2vec和依存分析的事件识别研究.《软件》.2017,第38卷(第6期),62-65.

审查员 李萌

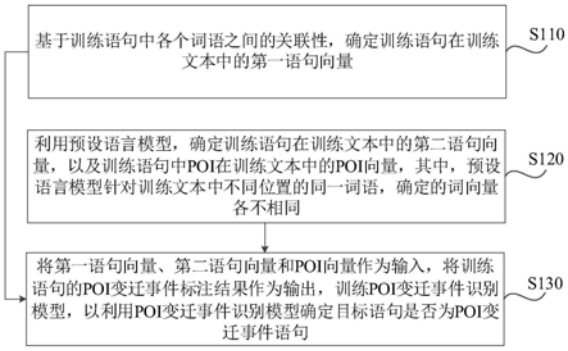
权利要求书3页 说明书9页 附图4页

(54)发明名称

POI变迁事件发现方法、装置、计算设备和介质

(57)摘要

本发明实施例公开了一种POI变迁事件发现方法、装置、计算设备和介质，其中，该方法包括：基于训练语句中各个词语之间的关联性，确定训练语句在训练文本中的第一语句向量；利用预设语言模型，确定训练语句在训练文本中的第二语句向量，以及训练语句中POI在训练文本中的POI向量；将第一语句向量、第二语句向量和POI向量作为输入，将训练语句的POI变迁事件标注结果作为输出，训练POI变迁事件识别模型，以利用POI变迁事件识别模型确定目标语句是否为POI变迁事件语句。本发明实施例可以从海量网络信息中准确与高效地发现POI变迁事件，提高关于确定POI变迁事件的召回率，从而为下游业务提供准确的POI信息。



CN 110276023 B

1.一种POI变迁事件发现方法,其特征在于,包括:

基于训练语句中各个词语之间的关联性,确定所述训练语句在训练文本中的第一语句向量;

利用预设语言模型,确定所述训练语句在所述训练文本中的第二语句向量,以及所述训练语句中POI在所述训练文本中的POI向量,其中,所述预设语言模型针对所述训练文本中不同位置的同一词语,确定的词向量各不相同;

将所述第一语句向量、所述第二语句向量和所述POI向量作为输入,将所述训练语句的POI变迁事件标注结果作为输出,训练POI变迁事件识别模型,以利用所述POI变迁事件识别模型确定目标语句是否为POI变迁事件语句。

2.根据权利要求1所述的方法,其特征在于,所述利用预设语言模型,确定所述训练语句在所述训练文本中的第二语句向量:

利用所述预设语言模型,确定所述训练语句中每个词语在所述训练文本中的词向量,将所述每个词语的词向量进行合并,作为所述训练语句在所述训练文本中的第二语句向量;

或者

在所述训练语句的特定位置添加标识词语,利用所述预设语言模型,确定所述标识词语在所述训练文本中的词向量,将所述标识词语的词向量作为所述训练语句在所述训练文本中的第二语句向量。

3.根据权利要求1所述的方法,其特征在于,所述利用预设语言模型,确定所述训练语句中POI在所述训练文本中的POI向量,包括:

如果所述训练语句中不包含POI,则将预设替代向量作为所述训练语句的POI向量;

如果所述训练语句中包含POI,则提取所述训练语句中的至少一个POI;

利用所述预设语言模型对所述至少一个POI进行编码,得到所述至少一个POI各自在所述训练文本中的POI向量,其中,不同POI的POI向量维度相同。

4.根据权利要求1所述的方法,其特征在于,所述基于训练语句中各个词语之间的关联性,确定所述训练语句在训练文本中的第一语句向量包括:

对所述训练语句进行分词,并利用词向量分析模型确定经分词得到的每个词语在所述训练文本中的词向量、位置向量和词性向量;

基于所述词向量、位置向量和词性向量,通过考虑各个词语在所述训练语句中的关联性,确定所述训练语句在所述训练文本中的第一语句向量。

5.根据权利要求4所述的方法,其特征在于,基于所述词向量、位置向量和词性向量,通过考虑各个词语在所述训练语句中的关联性,确定所述训练语句在所述训练文本中的第一语句向量,包括:

基于所述词向量、位置向量和词性向量,在卷积层中采用预设数量的卷积核进行卷积计算,提取所述训练语句在所述训练文本中的局部特征;

对提取的局部特征进行池化,并对池化结果进行非线性变换,得到所述训练语句在所述训练文本中的第一语句向量。

6.根据权利要求1所述的方法,其特征在于,所述目标语句包括网络媒体文本中的语句。

7.一种POI变迁事件发现装置,其特征在于,包括:

第一向量确定模块,用于基于训练语句中各个词语之间的关联性,确定所述训练语句在训练文本中的第一语句向量;

第二向量确定模块,用于利用预设语言模型,确定所述训练语句在所述训练文本中的第二语句向量,以及所述训练语句中POI在所述训练文本中的POI向量,其中,所述预设语言模型针对所述训练文本中不同位置的同一词语,确定的词向量各不相同;

模型训练模块,用于将所述第一语句向量、所述第二语句向量和所述POI向量作为输入,将所述训练语句的POI变迁事件标注结果作为输出,训练POI变迁事件识别模型,以利用所述POI变迁事件识别模型确定目标语句是否为POI变迁事件语句。

8.根据权利要求7所述的装置,其特征在于,所述第二向量确定模块包括语句向量确定单元,所述语句向量确定单元用于:

利用所述预设语言模型,确定所述训练语句中每个词语在所述训练文本中的词向量,将所述每个词语的词向量进行合并,作为所述训练语句在所述训练文本中的第二语句向量;

或者

在所述训练语句的特定位置添加标识词语,利用所述预设语言模型,确定所述标识词语在所述训练文本中的词向量,将所述标识词语的词向量作为所述训练语句在所述训练文本中的第二语句向量。

9.根据权利要求7所述的装置,其特征在于,所述第二向量确定模块包括POI向量确定单元,所述POI向量确定单元用于:

如果所述训练语句中不包含POI,则将预设替代向量作为所述训练语句的POI向量;

如果所述训练语句中包含POI,则提取所述训练语句中的至少一个POI;

利用所述预设语言模型对所述至少一个POI进行编码,得到所述至少一个POI各自在所述训练文本中的POI向量,其中,不同POI的POI向量维度相同。

10.根据权利要求7所述的装置,其特征在于,所述第一向量确定模块包括:

分词单元,用于对所述训练语句进行分词,并利用词向量分析模型确定经分词得到的每个词语在所述训练文本中的词向量、位置向量和词性向量;

关联单元,用于基于所述词向量、位置向量和词性向量,通过考虑各个词语在所述训练语句中的关联性,确定所述训练语句在所述训练文本中的第一语句向量。

11.根据权利要求10所述的装置,其特征在于,所述关联单元包括:

卷积计算子单元,用于基于所述词向量、位置向量和词性向量,在卷积层中采用预设数量的卷积核进行卷积计算,提取所述训练语句在所述训练文本中的局部特征;

池化与非线性变换子单元,用于对提取的局部特征进行池化,并对池化结果进行非线性变换,得到所述训练语句在所述训练文本中的第一语句向量。

12.根据权利要求7所述的装置,其特征在于,所述目标语句包括网络媒体文本中的语句。

13.一种计算设备,其特征在于,包括:

一个或多个处理器;

存储装置,用于存储一个或多个程序,

当所述一个或多个程序被所述一个或多个处理器执行，使得所述一个或多个处理器实现如权利要求1-6中任一所述的POI变迁事件发现方法。

14.一种计算机可读存储介质，其上存储有计算机程序，其特征在于，该程序被处理器执行时实现如权利要求1-6中任一所述的POI变迁事件发现方法。

POI变迁事件发现方法、装置、计算设备和介质

技术领域

[0001] 本发明实施例涉及互联网信息处理技术领域，尤其涉及一种POI变迁事件发现方法、装置、计算设备和介质。

背景技术

[0002] 当前网络社交内容(例如微博、网页和公众号等平台上发布的各类信息)中包含了一部分POI(Point of Interest，兴趣点)变迁事件信息，如商场“搬迁”、“暂停营业”或者“营业时间调整”等，这类信息在地图中有着非常重要的作用。例如，用户在检索POI时，如“xx博物馆”，对于已经暂停营业或营业时间调整等信息，如果能够通过强样式提醒用户，便可以减少用户无效出行，同时，也能够提升地图用户的使用体验。但是，社交内容中包含有用的POI变迁事件信息是非常稀少的，存在大量的噪声信息，因此，需要对获取的社交内容进行数据处理，准确提取其中的POI变迁事件信息。

[0003] 现有方法是通过POI抽取工具提取句子中的POI和触发词(表示具体事件的词语，通常为动词)，然后利用语言学工具判断POI和触发词之间是否存在联系，如果存在关联，则确定当前句子为POI变迁事件语句。其中，为了去掉POI与触发词之间无联系的句子，通过语言学工具对POI与触发词进行关联时，需要人工总结各个环节的规则(即人工干预的成分比较多)，可能导致POI与触发词之间的关联出错，并且该类方法不具备泛化能力，人工干预也导致现有方法耗时耗力；此外，由于语言的复杂性以及变化性，抽取工具本身也存在对POI和触发词的抽取错误，进一步导致现有方法对POI变迁事件的判断准确性较低，召回率低。

发明内容

[0004] 本发明实施例提供一种POI变迁事件发现方法、装置、计算设备和介质，以实现从海量网络信息中准确与高效地发现POI变迁事件，提高关于确定POI变迁事件的召回率。

[0005] 第一方面，本发明实施例提供了一种POI变迁事件发现方法，该方法包括：

[0006] 基于训练语句中各个词语之间的关联性，确定所述训练语句在训练文本中的第一语句向量；

[0007] 利用预设语言模型，确定所述训练语句在所述训练文本中的第二语句向量，以及所述训练语句中POI在所述训练文本中的POI向量，其中，所述预设语言模型针对所述训练文本中不同位置的同一词语，确定的词向量各不相同；

[0008] 将所述第一语句向量、所述第二语句向量和所述POI向量作为输入，将所述训练语句的POI变迁事件标注结果作为输出，训练POI变迁事件识别模型，以利用所述POI变迁事件识别模型确定目标语句是否为POI变迁事件语句。

[0009] 第二方面，本发明实施例还提供了一种POI变迁事件发现装置，该装置包括：

[0010] 第一向量确定模块，用于基于训练语句中各个词语之间的关联性，确定所述训练语句在训练文本中的第一语句向量；

[0011] 第二向量确定模块，用于利用预设语言模型，确定所述训练语句在所述训练文本

中的第二语句向量，以及所述训练语句中POI在所述训练文本中的POI向量，其中，所述预设语言模型针对所述训练文本中不同位置的同一词语，确定的词向量各不相同；

[0012] 模型训练模块，用于将所述第一语句向量、所述第二语句向量和所述POI向量作为输入，将所述训练语句的POI变迁事件标注结果作为输出，训练POI变迁事件识别模型，以利用所述POI变迁事件识别模型确定目标语句是否为POI变迁事件语句。

[0013] 第三方面，本发明实施例还提供了一种计算设备，包括：

[0014] 一个或多个处理器；

[0015] 存储装置，用于存储一个或多个程序，

[0016] 当所述一个或多个程序被所述一个或多个处理器执行，使得所述一个或多个处理器实现如本发明任一实施例所述的POI变迁事件发现方法。

[0017] 第四方面，本发明实施例还提供了一种计算机可读存储介质，其上存储有计算机程序，该程序被处理器执行时实现如本发明任一实施例所述的POI变迁事件发现方法。

[0018] 本发明实施例通过利用两种语句向量确定方式确定训练语句的向量表示，即基于分词得到的各个词语之间的关联性确定语句向量的方式与利用基于模型的神经网络语言模型(即预设语言模型)确定语句向量的方式相结合，保证了基于深度学习思想训练POI变迁事件识别模型的语句特征的完整性，然后结合训练语句中POI的向量表示，进一步强化训练语句中POI特征，保证了模型训练的准确性，解决了现有技术中对POI变迁事件的判断准确性较低的问题，实现了从海量网络信息中准确与高效地发现POI变迁事件，提高了关于确定POI变迁事件的召回率，从而为下游业务提供准确的POI信息。

附图说明

[0019] 图1是本发明实施例一提供的POI变迁事件发现方法的流程图；

[0020] 图2是本发明实施例二提供的POI变迁事件发现方法的流程图；

[0021] 图3是本发明实施例二提供的POI变迁事件识别模型的训练过程示意图；

[0022] 图4是本发明实施例三提供的POI变迁事件发现装置的结构示意图；

[0023] 图5是本发明实施例四提供的一种计算设备的结构示意图。

具体实施方式

[0024] 下面结合附图和实施例对本发明作进一步的详细说明。可以理解的是，此处所描述的具体实施例仅仅用于解释本发明，而非对本发明的限定。另外还需要说明的是，为了便于描述，附图中仅示出了与本发明相关的部分而非全部结构。

[0025] 实施例一

[0026] 图1是本发明实施例一提供的POI变迁事件发现方法的流程图，本实施例可适用于通过对海量网络信息进行挖掘，从中确认出属于描述POI变迁事件的信息的情况。该方法可以由POI变迁事件发现装置来执行，该装置可以采用软件和/或硬件的方式实现，并可集成在任意的计算设备上，包括但不限于服务器。

[0027] 如图1所示，本实施例提供的POI变迁事件发现方法可以包括：

[0028] S110、基于训练语句中各个词语之间的关联性，确定训练语句在训练文本中的第一语句向量。

[0029] 在基于深度学习思想训练模型之前，需要预先准备训练文本，训练文本可以是任意的社交媒体文本，例如微博、网页和公众号等平台上发布的各类信息文本，对每一个训练文本进行语句拆分，然后人工标注拆分得到的每一个训练语句中是否包含POI变迁事件，即确认训练语句中是否包含POI名称和相关联的触发词，采用人工标注的方式可以保证标注结果的准确性。如果训练语句中包含POI变迁事件，则该训练语句属于描述POI变迁事件的语句(或称为正样本)，反之，训练语句不属于描述POI变迁事件的语句(或称为负样本)。

[0030] 针对每个训练语句，可以通过分词技术得到句子中包括的词语，然后考虑各个词语在语句中的语义关联性，确定每个训练语句在训练文本中的第一语句向量，例如可以使用word2vector等传统语言模型确定。需要说明的是，确定第一语句向量使用的传统语言模型，针对训练文本中不同位置的同一词语，确定的词向量表示是相同的，不同于下文中所使用的预设语言模型。

[0031] S120、利用预设语言模型，确定训练语句在训练文本中的第二语句向量，以及训练语句中POI在训练文本中的POI向量，其中，预设语言模型针对训练文本中不同位置的同一词语，确定的词向量各不相同。

[0032] 其中，预设语言模型包括但不限于BERT语言模型(Bidirectional Encoder Representations from Transfoemers，用于语言理解的深度双向预训练转换器)、ELMO语言模型(Embeddings from Language Models，属于多个双向语言模型biLM的多层表示)和ERNIE语言模型(Enhanced Representation from kNowledge IntEgration，知识增强语义表示模型)等基于模型的神经网络语言模型，这类语言模型针对同一训练文本不同位置的同一词语，可以结合具体语句给出不同的向量表示，即实现每个词向量的动态表示。本实施例中所说的词语包括至少一个语言要素，例如对于中文而言，一个词语可以是单个字组成。此外，操作S110和操作S120之间没有严格的执行顺序限定。

[0033] 可选的，利用预设语言模型，确定训练语句在训练文本中的第二语句向量：

[0034] 利用预设语言模型，确定训练语句中每个词语在训练文本中的词向量，将每个词语的词向量进行合并，作为训练语句在训练文本中的第二语句向量；

[0035] 或者

[0036] 在训练语句的特定位置添加标识词语，利用预设语言模型，确定标识词语在训练文本中的词向量，将标识词语的词向量作为训练语句在训练文本中的第二语句向量。

[0037] 其中，训练语句的特定位置包括语句的开头或者结尾(在特定位置添加标识词语不能破坏训练语句本身的语义完整性)，标识词语可以是预先定义的能够用于区分不同句子的任意词语，例如可以是[SEP]。示例性的，在每个训练语句的开头添加标识词语[SEP]，然后将各个训练语句输入预设语言模型中，得到各个训练语句中每个词语的多层向量表示，例如对于BERT语言模型而言，Transformer有12层，合并多层向量或者使用最后一层向量都可以用于表示每个词语当前的特征向量，可以取“[SEP]”位置的词向量作为整个训练语句的编码向量，即第二语句向量。

[0038] 可选的，利用预设语言模型，确定训练语句中POI在训练文本中的POI向量，包括：

[0039] 如果训练语句中不包含POI，则将预设替代向量作为训练语句的POI向量；

[0040] 如果训练语句中包含POI，则提取训练语句中的至少一个POI；

[0041] 利用预设语言模型对至少一个POI进行编码，得到至少一个POI各自在训练文本中

的POI向量，其中，不同POI的POI向量维度相同。

[0042] 每个训练文本中的训练语句包括两种：包含POI的训练语句和不包含POI的训练语句。如果训练语句中不包含POI，则可以利用预设替代向量作为当前训练语句的POI向量，其中，预设替代向量与训练语句中的其他词向量具有相同的维度，为了保证模型计算的可行性，用来替代POI向量，实质上并不表示任何POI，并且其具体向量表示本实施例不作限定；如果训练语句中包含POI，则利用POI提取工具提取训练语句中的至少一个POI，并将训练语句输入预设语言模型确定POI的向量表示，其中，在能够实现准确提取语句中POI的基础上，POI提取工具可以是现有技术中任意可用的技术。本实施例中，POI是POI变迁事件语句中的重要特征，将POI向量作为模型训练输入的一部分，可以发挥强化训练语句中POI特征的作用，保证模型训练的准确性。

[0043] S130、将第一语句向量、第二语句向量和POI向量作为输入，将训练语句的POI变迁事件标注结果作为输出，训练POI变迁事件识别模型，以利用POI变迁事件识别模型确定目标语句是否为POI变迁事件语句。

[0044] 针对每个训练语句，均采用两种语句向量确定方式确定其语句向量，实现语句向量特征的互补，保证了基于深度学习思想训练POI变迁事件识别模型的语句特征的完整性。目标语句包括社交媒体文本中的语句，可以通过对从互联网上抓取的社交媒体文本进行语句拆分得到。将目标语句输入POI变迁事件识别模型中，确认目标语句是否为用于描述POI变迁事件的语句，如果是，则可以将识别出的POI变迁事件语句作为处理对象，进一步提取出POI变迁事件，为下游业务提供准确的POI信息。例如，在地图制作过程或者POI信息搜索过程中，可以将识别到的POI变迁事件及时用于POI数据的更新，为用户提供准确的POI信息，使得用户可以实时掌握POI的状态变化信息，减少用户的无效出行，提升用户的地图使用体验或者搜索体验。通过利用POI变迁事件识别模型对社交媒体文本中的语句进行识别分类，可以实现对社交媒体文本的有效筛选，减少针对海量网络信息的数据处理量，提高数据挖掘效率。

[0045] 本实施例的技术方案通过利用两种语句向量确定方式确定训练语句的向量表示，即基于分词得到的各个词语之间的关联性确定语句向量的方式与利用基于模型的神经网络语言模型(即预设语言模型)确定语句向量的方式相结合，保证了基于深度学习思想训练POI变迁事件识别模型的语句特征的完整性，然后结合训练语句中POI的向量表示，进一步强化训练语句中POI特征，保证了模型训练的准确性，解决了现有技术中对POI变迁事件的判断准确性较低的问题，实现了从海量网络信息中准确与高效地发现POI变迁事件，提高了关于确定POI变迁事件的召回率，从而为下游业务提供准确的POI信息，并且，本实施例方案具有较高的泛化能力，能够适用于对任意类型的社交媒体文本中的语句识别，识别过程不需要人为参与。

[0046] 实施例二

[0047] 图2是本发明实施例二提供的POI变迁事件发现方法的流程图，本实施例是在上述实施例的基础上进一步进行优化。如图2所示，该方法可以包括：

[0048] S210、对训练语句进行分词，并利用词向量分析模型确定经分词得到的每个词语在训练文本中的词向量、位置向量和词性向量。

[0049] 本实施例中，训练语句经过分词得到的每个词语的向量表示，由三部分向量拼接

而成：词向量(Word Embeddings)、位置向量(Position Embedding)和词性向量(POS Embedding)。其中，词向量可以利用预先训练的无监督模型得到，例如word2vector模型等，该无监督模型可以是基于已有的开源词向量或者自行构建的训练语料训练得到，训练语料包括网络社交媒体文本中的标题和正文；位置向量表示每个词语在训练文本的位置，可以是当前词语与潜在POI事件主体(包括潜在实体和潜在事件触发词)相对位置的向量表示，例如当前词语是训练语句中的第4个词，训练语句中的POI实体在句子中的位置是7，当前词语相对于该POI实体的位置是-4，然后将-4映射到一个固定维度的正态分布向量上，从而得到当前词语的位置向量，不同的数字映射为不同的向量；词性向量指将每个词语的词性映射为一个多维向量，相同的词性使用相同的向量初始化。

[0050] S220、基于词向量、位置向量和词性向量，通过考虑各个词语在训练语句中的关联性，确定训练语句在训练文本中的第一语句向量。

[0051] 通过考虑各个词语之间的关联性，可以保证训练语句的语义正确性。

[0052] 可选的，基于词向量、位置向量和词性向量，通过考虑各个词语在训练语句中的关联性，确定训练语句在训练文本中的第一语句向量，包括：

[0053] 基于词向量、位置向量和词性向量，在卷积层中采用预设数量的卷积核进行卷积计算，提取训练语句在训练文本中的局部特征；

[0054] 对提取的局部特征进行池化，并对池化结果进行非线性变换，得到训练语句在训练文本中的第一语句向量。

[0055] 图3以卷积神经网络为例，示出了本实施例提供的POI变迁事件识别模型的训练过程的一种示意图，如图3所示，在输入层中输入训练语句中每个词语的词向量、位置向量和词性向量；在卷积层中通过多个卷积核(Feature Map)提取局部特征，同时避免网络中参数过多，本实施例中可以使用卷积窗口为3的卷积层提取特征，提取的特征数量与预先定义的参数有关，并且，本实施例可以使用等长卷积，卷积结果与输入的宽度一致；继续对卷积特征(即提取的局部特征)进行池化，池化的目的是找出相同位置处最重要的特征信息，本实施例可以使用最大池化操作，即相同维度取最大值，然后输出池化后的结果；在全连接层中，对池化后的结果做非线性变换得到训练语句在训练文本中的第一语句向量，该第一语句向量中考虑了各个词语在训练语句中的语义关联性，也可以称为语句上下文向量(该特征表示了整个语句的上下文特征)，其中，非线性变换包括但不限于利用tanh等激活函数进行非线性变换。

[0056] S230、利用预设语言模型，确定训练语句在训练文本中的第二语句向量，以及训练语句中POI在训练文本中的POI向量，其中，预设语言模型针对训练文本中不同位置的同一词语，确定的词向量各不相同。

[0057] 继续如图3所示，利用预设语言模型，确定训练语句在训练文本中的第二语句向量，并使用POI抽取工具从训练语句中抽取出POI，然后对基于预设语言模型得到的POI编码向量进行池化，得到最终的POI向量表示。POI是判断训练语句是否包含POI事件的重要信息。

[0058] S240、将第一语句向量、第二语句向量和POI向量作为输入，将训练语句的POI变迁事件标注结果作为输出，训练POI变迁事件识别模型，以利用POI变迁事件识别模型确定目标语句是否为POI变迁事件语句。

[0059] 最后，将第一语句向量、第二语句向量和POI向量拼接在一起，形成多维向量，作为全连接层的输入，输出层的输出是预先定义的语句类别：POI变迁事件语句和非POI变迁事件语句。

[0060] 本实施例的技术方案通过利用两种语句向量确定方式确定训练语句的向量表示，即基于分词得到的各个词语之间的关联性确定语句向量的方式与利用基于模型的神经网络语言模型(即预设语言模型)确定语句向量的方式相结合，保证了基于深度学习思想训练POI变迁事件识别模型的语句特征的完整性，然后结合训练语句中POI的向量表示，进一步强化训练语句中POI特征，保证了模型训练的准确性，解决了现有技术中对POI变迁事件的判断准确性较低的问题，实现了从海量网络信息中准确与高效地发现POI变迁事件，提高了关于确定POI变迁事件的召回率，从而为下游业务提供准确的POI信息。

[0061] 实施例三

[0062] 图4是本发明实施例三提供的POI变迁事件发现装置的结构示意图，本实施例可适用于通过对海量网络信息进行挖掘，从中确认出属于描述POI变迁事件的信息的情况。该装置可以采用软件和/或硬件的方式实现，并可集成在任意的计算设备上，包括但不限于服务器。

[0063] 如图4所示，本实施例提供的POI变迁事件发现装置可以包括第一向量确定模块310、第二向量确定模块320和模型训练模块330，其中：

[0064] 第一向量确定模块310，用于基于训练语句中各个词语之间的关联性，确定训练语句在训练文本中的第一语句向量；

[0065] 第二向量确定模块320，用于利用预设语言模型，确定训练语句在训练文本中的第二语句向量，以及训练语句中POI在训练文本中的POI向量，其中，预设语言模型针对训练文本中不同位置的同一词语，确定的词向量各不相同；

[0066] 模型训练模块330，用于将第一语句向量、第二语句向量和POI向量作为输入，将训练语句的POI变迁事件标注结果作为输出，训练POI变迁事件识别模型，以利用POI变迁事件识别模型确定目标语句是否为POI变迁事件语句。

[0067] 可选的，第二向量确定模块320包括语句向量确定单元，语句向量确定单元用于：

[0068] 利用预设语言模型，确定训练语句中每个词语在训练文本中的词向量，将每个词语的词向量进行合并，作为训练语句在训练文本中的第二语句向量；

[0069] 或者

[0070] 在训练语句的特定位置添加标识词语，利用预设语言模型，确定标识词语在训练文本中的词向量，将标识词语的词向量作为训练语句在训练文本中的第二语句向量。

[0071] 可选的，第二向量确定模块320包括POI向量确定单元，POI向量确定单元用于：

[0072] 如果训练语句中不包含POI，则将预设替代向量作为训练语句的POI向量；

[0073] 如果训练语句中包含POI，则提取训练语句中的至少一个POI；

[0074] 利用预设语言模型对至少一个POI进行编码，得到至少一个POI各自在训练文本中的POI向量，其中，不同POI的POI向量维度相同。

[0075] 可选的，第一向量确定模块310包括：

[0076] 分词单元，用于对训练语句进行分词，并利用词向量分析模型确定经分词得到的每个词语在训练文本中的词向量、位置向量和词性向量；

[0077] 关联单元，用于基于词向量、位置向量和词性向量，通过考虑各个词语在训练语句中的关联性，确定训练语句在训练文本中的第一语句向量。

[0078] 可选的，关联单元包括：

[0079] 卷积计算子单元，用于基于词向量、位置向量和词性向量，在卷积层中采用预设数量的卷积核进行卷积计算，提取训练语句在训练文本中的局部特征；

[0080] 池化与非线性变换子单元，用于对提取的局部特征进行池化，并对池化结果进行非线性变换，得到训练语句在训练文本中的第一语句向量。

[0081] 可选的，模型训练模块330中的目标语句包括网络媒体文本中的语句。

[0082] 本发明实施例所提供的POI变迁事件发现装置可执行本发明任意实施例所提供的POI变迁事件发现方法，具备执行方法相应的功能模块和有益效果。本实施例中未详尽描述的内容可以参考本发明任意方法实施例中的描述。

[0083] 实施例四

[0084] 图5是本发明实施例四提供的一种计算设备的结构示意图。图5示出了适于用来实现本发明实施方式的示例性计算设备412的框图。图5显示的计算设备412仅仅是一个示例，不应对本发明实施例的功能和使用范围带来任何限制。计算设备412可以是任意的具有计算能力的设备，包括但不限于服务器。

[0085] 如图5所示，计算设备412以通用计算设备的形式表现。计算设备412的组件可以包括但不限于：一个或者多个处理器416，存储装置428，连接不同系统组件(包括存储装置428和处理器416)的总线418。

[0086] 总线418表示几类总线结构中的一种或多种，包括存储装置总线或者存储装置控制器，外围总线，图形加速端口，处理器或者使用多种总线结构中的任意总线结构的局域总线。举例来说，这些体系结构包括但不限于工业标准体系结构(Industry Subversive Alliance，ISA)总线，微通道体系结构(Micro Channel Architecture，MAC)总线，增强型ISA总线、视频电子标准协会(Video Electronics Standards Association，VESA)局域总线以及外围组件互连(Peripheral Component Interconnect，PCI)总线。

[0087] 计算设备412典型地包括多种计算机系统可读介质。这些介质可以是任何能够被计算设备412访问的可用介质，包括易失性和非易失性介质，可移动的和不可移动的介质。

[0088] 存储装置428可以包括易失性存储器形式的计算机系统可读介质，例如随机存取存储器(Random Access Memory，RAM)430和/或高速缓存存储器432。计算设备412可以进一步包括其它可移动/不可移动的、易失性/非易失性计算机系统存储介质。仅作为举例，存储系统434可以用于读写不可移动的、非易失性磁介质(图5未显示，通常称为“硬盘驱动器”)。尽管图5中未示出，可以提供用于对可移动非易失性磁盘(例如“软盘”)读写的磁盘驱动器，以及对可移动非易失性光盘，例如只读光盘(Compact Disc Read-Only Memory，CD-ROM)，数字视盘(Digital Video Disc-Read Only Memory，DVD-ROM)或者其它光介质)读写的光盘驱动器。在这些情况下，每个驱动器可以通过一个或者多个数据介质接口与总线418相连。存储装置428可以包括至少一个程序产品，该程序产品具有一组(例如至少一个)程序模块，这些程序模块被配置以执行本发明各实施例的功能。

[0089] 具有一组(至少一个)程序模块442的程序/实用工具440，可以存储在例如存储装置428中，这样的程序模块442包括但不限于操作系统、一个或者多个应用程序、其它程序模

块以及程序数据，这些示例中的每一个或某种组合中可能包括网络环境的实现。程序模块442通常执行本发明所描述的实施例中的功能和/或方法。

[0090] 计算设备412也可以与一个或多个外部设备414(例如键盘、指向终端、显示器424等)通信，还可与一个或者多个使得用户能与该计算设备412交互的终端通信，和/或与使得该计算设备412能与一个或多个其它计算终端进行通信的任何终端(例如网卡，调制解调器等等)通信。这种通信可以通过输入/输出(I/O)接口422进行。并且，计算设备412还可以通过网络适配器420与一个或者多个网络(例如局域网(Local Area Network，LAN)，广域网(Wide Area Network，WAN)和/或公共网络，例如因特网)通信。如图5所示，网络适配器420通过总线418与计算设备412的其它模块通信。应当明白，尽管图中未示出，可以结合计算设备412使用其它硬件和/或软件模块，包括但不限于：微代码、设备驱动器、冗余处理器、外部磁盘驱动阵列、磁盘阵列(Redundant Arrays of Independent Disks，RAID)系统、磁带驱动器以及数据备份存储系统等。

[0091] 处理器416通过运行存储在存储装置428中的程序，从而执行各种功能应用以及数据处理，例如实现本发明任意实施例所提供的POI变迁事件发现方法，该方法可以包括：

[0092] 基于训练语句中各个词语之间的关联性，确定训练语句在训练文本中的第一语句向量；

[0093] 利用预设语言模型，确定训练语句在训练文本中的第二语句向量，以及训练语句中POI在训练文本中的POI向量，其中，预设语言模型针对训练文本中不同位置的同一词语，确定的词向量各不相同；

[0094] 将第一语句向量、第二语句向量和POI向量作为输入，将训练语句的POI变迁事件标注结果作为输出，训练POI变迁事件识别模型，以利用POI变迁事件识别模型确定目标语句是否为POI变迁事件语句。

[0095] 实施例五

[0096] 本发明实施例五还提供了一种计算机可读存储介质，其上存储有计算机程序，该程序被处理器执行时实现如本发明任意实施例所提供的POI变迁事件发现方法，该方法可以包括：

[0097] 基于训练语句中各个词语之间的关联性，确定训练语句在训练文本中的第一语句向量；

[0098] 利用预设语言模型，确定训练语句在训练文本中的第二语句向量，以及训练语句中POI在训练文本中的POI向量，其中，预设语言模型针对训练文本中不同位置的同一词语，确定的词向量各不相同；

[0099] 将第一语句向量、第二语句向量和POI向量作为输入，将训练语句的POI变迁事件标注结果作为输出，训练POI变迁事件识别模型，以利用POI变迁事件识别模型确定目标语句是否为POI变迁事件语句。

[0100] 本发明实施例的计算机存储介质，可以采用一个或多个计算机可读的介质的任意组合。计算机可读介质可以是计算机可读信号介质或者计算机可读存储介质。计算机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。计算机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、便携式计算机磁盘、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、便携式紧凑磁盘只读存储器(CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。在本文件中,计算机可读存储介质可以是任何包含或存储程序的有形介质,该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。

[0101] 计算机可读的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,其中承载了计算机可读的程序代码。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。计算机可读的信号介质还可以是计算机可读存储介质以外的任何计算机可读介质,该计算机可读介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。

[0102] 计算机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、电线、光缆、RF等等,或者上述的任意合适的组合。

[0103] 可以以一种或多种程序设计语言或其组合来编写用于执行本发明操作的计算机程序代码,所述程序设计语言包括面向对象的程序设计语言—诸如Java、Smalltalk、C++,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机上执行、部分地在用户计算机上执行、作为一个独立的软件包执行、部分在用户计算机上部分在远程计算机上执行、或者完全在远程计算机或终端上执行。在涉及远程计算机的情形中,远程计算机可以通过任意种类的网络——包括局域网(LAN)或广域网(WAN)—连接到用户计算机,或者,可以连接到外部计算机(例如利用因特网服务提供商来通过因特网连接)。

[0104] 注意,上述仅为本发明的较佳实施例及所运用技术原理。本领域技术人员会理解,本发明不限于这里所述的特定实施例,对本领域技术人员来说能够进行各种明显的变化、重新调整和替代而不会脱离本发明的保护范围。因此,虽然通过以上实施例对本发明进行了较为详细的说明,但是本发明不仅仅限于以上实施例,在不脱离本发明构思的情况下,还可以包括更多其他等效实施例,而本发明的范围由所附的权利要求范围决定。

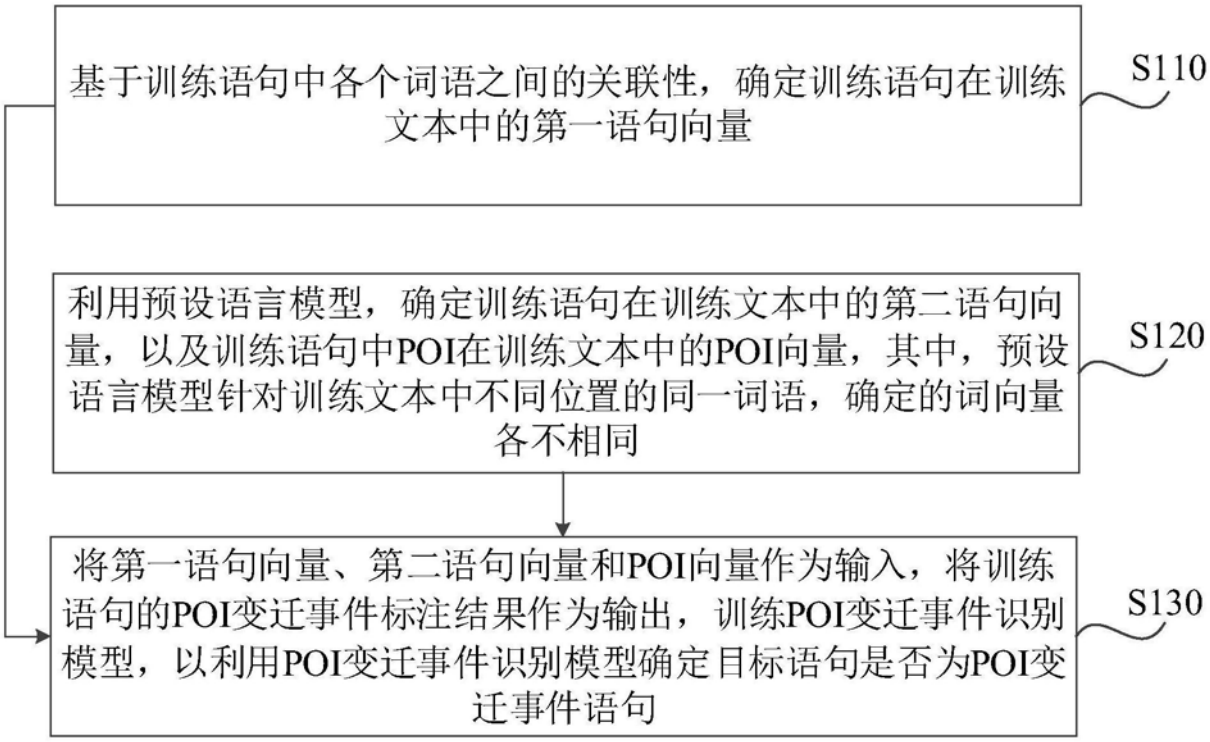

基于训练语句中各个词语之间的关联性，确定训练语句在训练文本中的第一语句向量
S110
利用预设语言模型，确定训练语句在训练文本中的第二语句向量，以及训练语句中POI在训练文本中的POI向量，其中，预设语言模型针对训练文本中不同位置的同一词语，确定的词向量各不相同
S120
将第一语句向量、第二语句向量和POI向量作为输入，将训练语句的POI变迁事件标注结果作为输出，训练POI变迁事件识别模型，以利用POI变迁事件识别模型确定目标语句是否为POI变迁事件语句
S130

图1

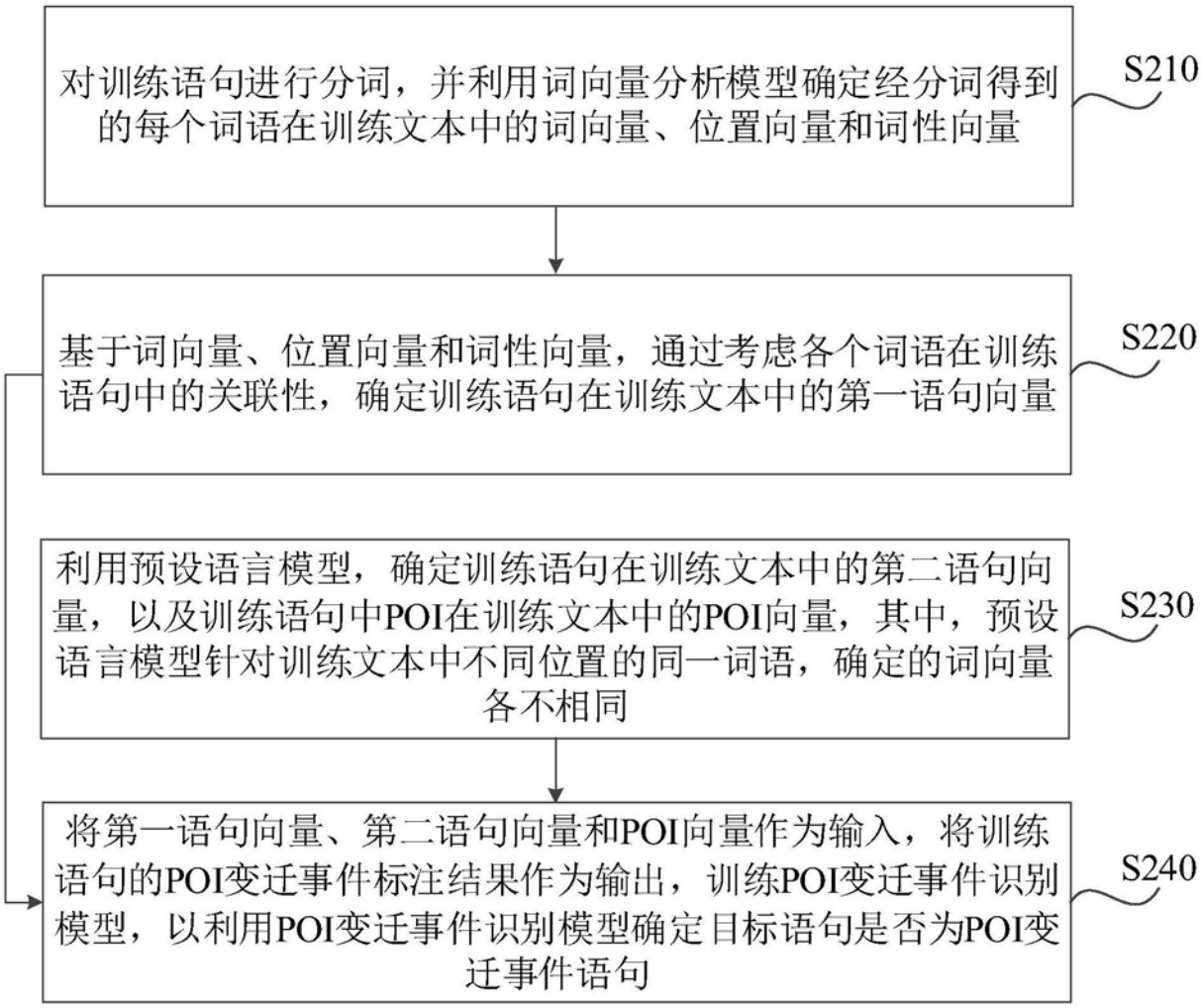

对训练语句进行分词，并利用词向量分析模型确定经分词得到的每个词语在训练文本中的词向量、位置向量和词性向量
S210
基于词向量、位置向量和词性向量，通过考虑各个词语在训练语句中的关联性，确定训练语句在训练文本中的第一语句向量
S220
利用预设语言模型，确定训练语句在训练文本中的第二语句向量，以及训练语句中POI在训练文本中的POI向量，其中，预设语言模型针对训练文本中不同位置的同一词语，确定的词向量各不相同
S230
将第一语句向量、第二语句向量和POI向量作为输入，将训练语句的POI变迁事件标注结果作为输出，训练POI变迁事件识别模型，以利用POI变迁事件识别模型确定目标语句是否为POI变迁事件语句
S240

图2

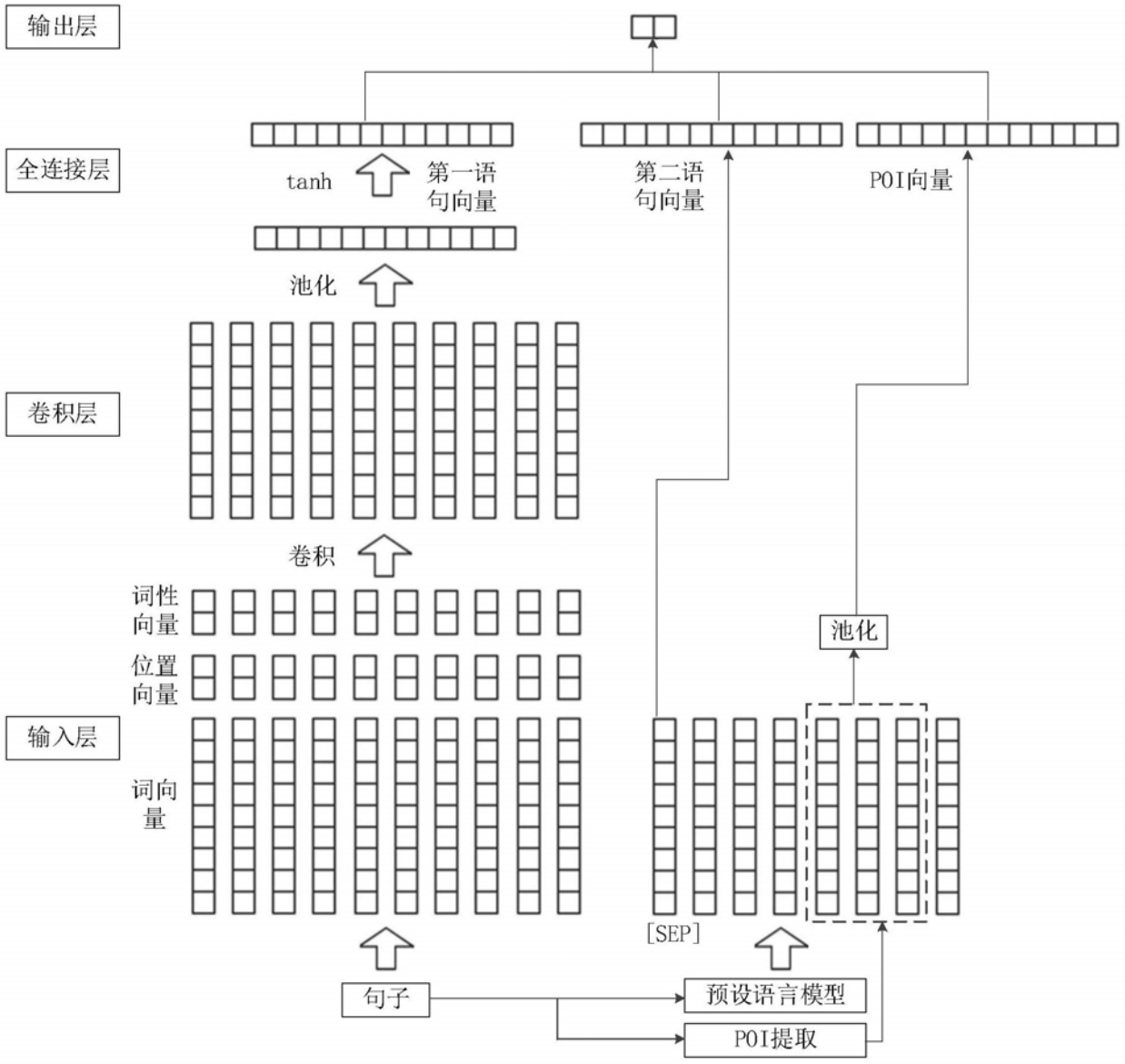

输出层
全连接层
tanh
第一语句向量
第二语句向量
POI向量
池化
卷积层
卷积
词性向量
位置向量
输入层
词向量
池化
[SEP]
句子
预设语言模型
POI提取

图3

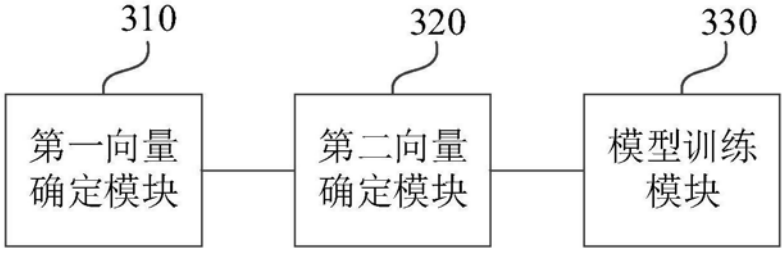

310
320
330
第一向量确定模块
第二向量确定模块
模型训练模块

图4

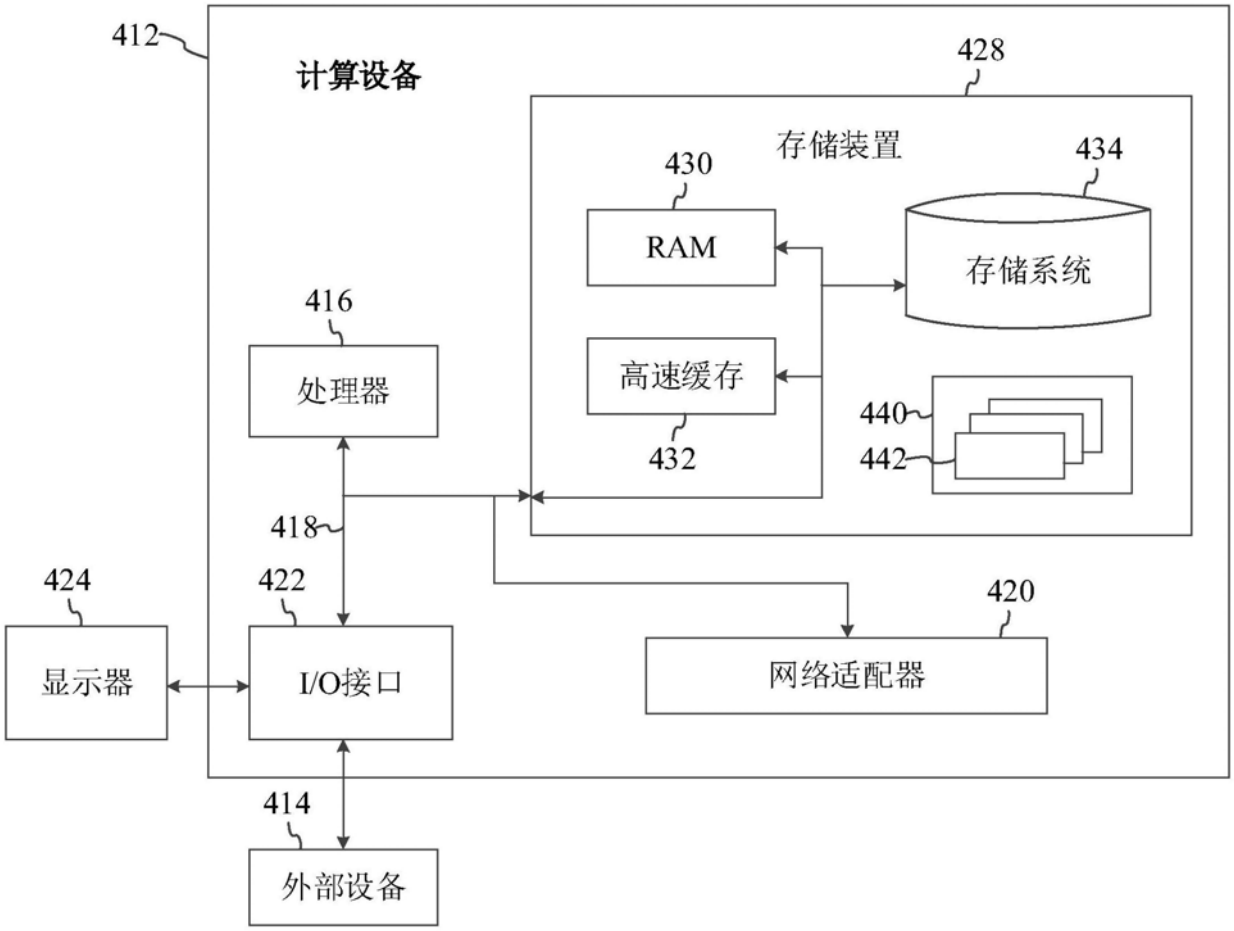

412
计算设备
428
存储装置
430
RAM
434
存储系统
416
处理器
高速缓存
432
440
442
418
424
422
420
显示器
I/O接口
网络适配器
414
外部设备

图5