

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2020 年 12 月 24 日 (24.12.2020)



(10) 国际公布号
WO 2020/253050 A1

(51) 国际专利分类号:
G06F 17/00 (2019.01)

(21) 国际申请号: PCT/CN2019/117725

(22) 国际申请日: 2019 年 11 月 13 日 (13.11.2019)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201910534810.8 2019年6月20日 (20.06.2019) CN

(71) 申请人: 平安科技(深圳)有限公司(PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) [CN/CN];
中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。

(72) 发明人: 郑立颖(ZHENG, Liying); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。 徐亮(XU, Liang); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。 阮晓雯(RUAN, Xiaowen); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。

(74) 代理人: 北京中强智尚知识产权代理有限公司 (BEIJING STRONG WISDOM INTELLECTUAL PROPERTY AGENCY LTD.); 中国北京市朝阳区北苑东路 19 号院 2 号楼 2608 室, Beijing 100012 (CN)。

(81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU,

(54) Title: WORD VECTOR CONFIGURATION METHOD AND APPARATUS, AND STORAGE MEDIUM AND ELECTRONIC DEVICE

(54) 发明名称: 词向量的配置方法、装置、存储介质、电子装置

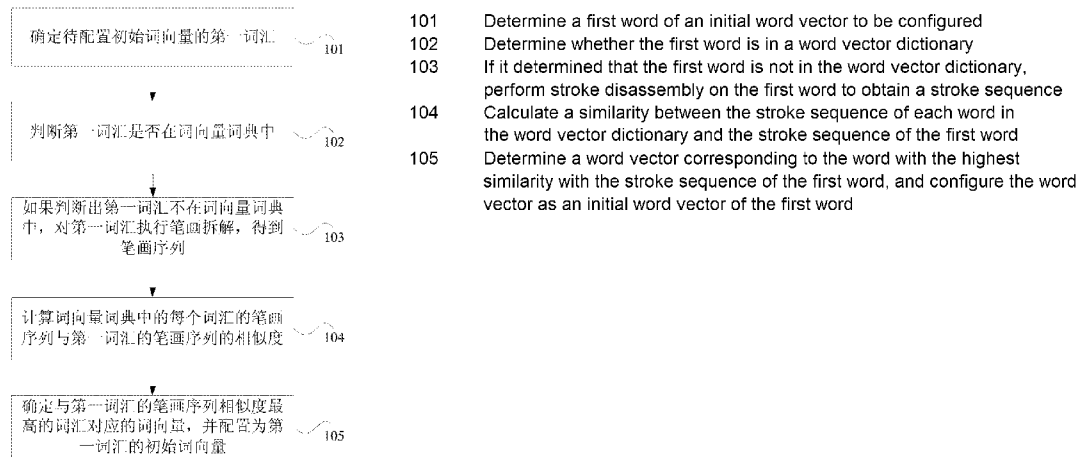


图 1

(57) Abstract: The present application provides a word vector configuration method and apparatus, and a storage medium and an electronic device. The word vector configuration method provided in the present application comprises: determining a first word of an initial word vector to be configured; determining whether the first word is in a word vector dictionary, wherein the word vector dictionary is used for storing a one-to-one correspondence between multiple words and multiple word vectors; if it determined that the first word is not in the word vector dictionary, performing stroke disassembly on the first word to obtain a stroke sequence; calculating a similarity between the stroke sequence of each word in the word vector dictionary and the stroke sequence of the first word; and determining a word vector corresponding to the word with the highest similarity with the stroke sequence of the first word, and configuring the word

CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国 (除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

根据细则4.17的声明:

- 关于申请人有权申请并被授予专利 (细则4.17(ii))
- 发明人资格 (细则4.17(iv))

本国际公布:

- 包括国际检索报告 (条约第21条(3))。

vector as an initial word vector of the first word. The present application solves the technical problem in the related art of decrease in the accuracy of a subsequent training task when the word vectors of unregistered words are configured by means of random allocation.

(57) 摘要: 本申请提供了一种词向量的配置方法、装置、存储介质、电子装置, 其中, 本申请提供的词向量的配置方法包括: 确定待配置初始词向量的第一词汇; 判断第一词汇是否在词向量词典中, 其中, 词向量词典用于存储多个词汇与多个词向量的一一对应关系; 如果判断出第一词汇不在词向量词典中, 对第一词汇执行笔画拆解, 得到笔画序列; 计算词向量词典中的每个词汇的笔画序列与第一词汇的笔画序列的相似度; 确定与第一词汇的笔画序列相似度最高的词汇对应的词向量, 并配置为第一词汇的初始词向量。通过本申请, 解决了相关技术中利用随机分配的方式配置未登录词的词向量时导致后续训练任务精度下降的技术问题。

词向量的配置方法、装置、存储介质、电子装置

本申请要求与 2019 年 6 月 20 日提交中国专利局、申请号为 2019105348108、申请名称为“词向量的配置方法、装置、存储介质、电子装置”的中国专利申请的优先权，其全部内容通过引用结合在申请中。

技术领域

本申请涉及神经网络领域，具体而言，涉及一种词向量的配置方法、装置、存储介质、电子装置。

背景技术

在处理文本类数据时，通常最基本的步骤就是分词并训练词向量（例如，采用 word2vec 方法进行训练），然后基于词向量进行后续的文本比对、分类等任务。发明人发现在实际处理过程中，经常会发生待处理的文本中包含不在词向量词典范围内的新词（未登录词），通常的处理方法是随机对未登录词随机的分配词向量，但是，随机分配的词向量并没有利用到新词的语义信息，造成后续任务精度下降。

针对相关技术中存在的上述问题，目前尚未发现有效的解决方案。

发明内容

本申请实施例提供了一种词向量的配置方法、装置、存储介质、电子装置，以至少解决现有技术中利用随机分配的方式配置未登录词的词向量时导致后续训练任务精度下降的技术问题。

根据本申请的一个实施例，提供了一种词向量的配置方法，包括：确定待配置初始词向量的第一词汇；判断第一词汇是否在词向量词典中，其中，词向量词典用于存储多个词汇与多个词向量的一一对应关系；如果判断出第一词汇不在词向量词典中，对第一词汇执行笔画拆解，得到笔画序列；计算词向量词典中的每个词汇的笔画序列与第一词汇的笔画序列的相似度；确定与第一词汇的笔画序列相似度最高的词汇对应的词向量，并配置为第一词汇的初始词向量。

通过本申请提供的词向量的配置方法，通过对未登录在词向量词典中的词汇进行拆解，在词向量词典中查找与其笔画最接近的已登录词，进而将笔画相近词的词向量配置为未登录词的初始词向量，解决了相关技术中利用随机分配的方式配置未登录词的词向量时导致后续训练任务精度下降的技术问题，利用中文的笔画中所携带的语义信息来为未登录词赋予初始词向量，能够减少后续训练任务的耗时、提高训练任务的精度。

根据本申请的另一个实施例，提供了一种词向量的配置装置，包括：第一确定模块，用于确定待配置初始词向量的第一词汇；判断模块，用于判断第一词汇是否在词向量词典中，其中，词向量词典用于存储多个词汇与多个词向量的一一对应关系；拆解模块，用于如果判断出第一词汇

不在词向量词典中，对第一词汇执行笔画拆解，得到笔画序列；计算模块，用于计算词向量词典中的每个词汇的笔画序列与第一词汇的笔画序列的相似度；第二确定模块，用于确定与第一词汇的笔画序列相似度最高的词汇对应的词向量，并配置为第一词汇的初始词向量。

通过本申请提供的词向量的配置装置，通过对未登录在词向量词典中的词汇进行拆解，在词向量词典中查找与其笔画最接近的已登录词，进而将笔画相近词的词向量配置为未登录词的初始词向量，解决了相关技术中利用随机分配的方式配置未登录词的词向量时导致后续训练任务精度下降的技术问题，利用中文的笔画中所携带的语义信息来为未登录词赋予初始词向量，能够减少后续训练任务的耗时、提高训练任务的精度。

根据本申请的又一个实施例，还提供了一种非易失性可读存储介质，所述存储介质中存储有计算机可读指令，其中，所述计算机可读指令被设置为运行时执行上述任一项方法实施例中的步骤。

根据本申请的又一个实施例，还提供了一种电子装置，包括存储器和处理器，所述存储器中存储有计算机可读指令，所述处理器被设置为运行所述计算机可读指令以执行上述任一项方法实施例中的步骤。

附图说明

此处所说明的附图用来提供对本申请的进一步理解，构成本申请的一部分，本申请的示意性实施例及其说明用于解释本申请，并不构成对本申请的不当限定。在附图中：

图 1 是根据本申请实施例的词向量的配置方法的流程图；

图 2 是根据本申请实施例的词向量的配置装置的示意图；

图 3 是本申请实施例的一种电子装置的硬件结构框图。

具体实施方式

为了使本技术领域的人员更好地理解本申请方案，下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本申请一部分的实施例，而不是全部的实施例，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。基于本申请中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都应当属于本申请保护的范围。

需要说明的是，本申请的说明书和权利要求书及上述附图中的术语“第一”、“第二”等是用于区别类似的对象，而不必用于描述特定的顺序或先后次序。应该理解这样使用的数据在适当情况下可以互换，以便这里描述的本申请的实施例能够以除了在这里图示或描述的那些以外的顺序实施。此外，术语“包括”和“具有”以及他们的任何变形，意图在于覆盖不排他的包含，例如，包含了一系列步骤或单元的过程、方法、系统、产品或设备不必限于清楚地列出的那些步骤或单元，而是可包括没有清楚地列出的或对于这些过程、方法、产品或设备固有的其它步骤或单元。

实施例 1

本实施例提供了一种词向量的配置方法，可以应用于移动终端、手持终端或类似的运算设备之中。运行在不同的运算设备仅是方案在执行主体上的差异，本领域人员可预见在不同运算设备中运行能够产生相同的技术效果。本实施例提供的词向量的配置方法，通过对未登录在词向量词典中的词汇进行拆解，在词向量词典中查找与其笔画最接近的已登录词，进而将笔画相近词的词向量配置为未登录词的初始词向量，解决了相关技术中利用随机分配的方式配置未登录词的词向量时导致后续训练任务精度下降的技术问题，利用中文的笔画中所携带的语义信息来为未登录词赋予初始词向量，能够减少后续训练任务的耗时、提高训练任务的精度。

如图 1 所示，本实施例提供的词向量的配置方法包括如下步骤：

步骤 101，确定待配置初始词向量的第一词汇。第一词汇可以是在待处理的语料中的任意一个分词词汇。本实施例应用于对词汇的词向量进行配置，由于在神经网络训练的过程中，机器不便于识别自然语言，需要将自然语言转变为机器语言，以便于机器识别不同的词汇，本实施例中是通过词向量来标识不同的词汇。在本实施例中，神经网络训练可以通过神经网络模型实现的，所述神经网络模型包括但不限于 BP 神经网络模型、CNN 神经网络模型等。在基于语料确定词汇时，首先对预料执行分词处理，然后确定当前需要配置初始词向量的词汇为第一词汇。具体而言，确定待配置初始词向量的第一词汇，包括如下步骤：

步骤 11，获取待分词的语料；

步骤 12，对语料进行分词，得到顺序的多个分词；

步骤 13，在多个分词中确定未配置初始词向量的首位分词，得到第一词汇。其中，分词处理可以采用现有的分词算法，本实施例在此不再赘述。

步骤 102，判断第一词汇是否在词向量词典中，其中，词向量词典用于存储多个词汇与多个词向量的一一对应关系。词向量词典用于存储已经训练、或已经人工标注过的词汇和词向量的关系对。在获取第一词汇之后，判断第一词汇是否在词向量词典中。如果判断出第一词汇在词向量词典中，则在词向量词典查找第一词汇对应的词向量。

步骤 103，如果判断出第一词汇不在词向量词典中，对第一词汇执行笔画拆解，得到笔画序列；在对第一词汇进行笔画拆解时，可以将词汇中的每个字拆解为中文最小单位的笔画，例如，横、竖、撇、捺等等，也可以依据预设的最小单位将词汇中的每个字拆解到预设的程度，例如，将日、月、木、目、口、田等常用的组成单元作为一个最小单位，以化简笔画拆解之后得到的笔画序列的长度。具体拆解到何种程度可以根据实际设定来确定，本实施例对此不作具体限制，在此不再赘述。笔画序列是对应词汇所包括的所有笔画的序列。可以通过数字、字母等字符的组合来标识不同的笔画，所有笔画的标识符组成的序列就是笔画序列。例如，如果横竖撇捺的标识符分别为 1、2、3、4，则“木”字的笔画序列为{1, 2, 3, 4}。又如，如果“木”的标识符为 1，目的标识符为 2，则“相”字的笔画序列为{1, 2}。

步骤 104, 计算词向量词典中的每个词汇的笔画序列与第一词汇的笔画序列的相似度; 该步骤可以是在每通过步骤 103 获取到一个词汇的笔画序列之后执行的, 也即, 每对一个语料中的一个词汇拆解完毕之后, 计算其与词向量词典中每个词汇的相似度。或者, 也可以是在对某一段语料中所有未登录词 (未在词向量词典中的词) 进行拆解完毕后分别执行的。此外, 在计算一个词汇与词向量词典中每个词汇的笔画序列相似度时, 也可以时同步执行或异步执行的, 同步执行即为分若干个进程, 每个进程计算该词汇与词向量词典中的一个词汇的相似度, 异步执行则为仅通过一个进程, 顺序执行计算该词汇与词向量词典中每一个词汇的相似度。

在计算词向量词典中的每个词汇的笔画序列与第一词汇的笔画序列的相似度时, 可以采用如下步骤:

步骤 21, 确定第二词汇的笔画序列与第一词汇的笔画序列的重合序列段的总长度, 其中, 重合序列段为两个词汇的笔画序列中笔画排列相同的序列段。例如, 第一词汇的笔画序列为{5, 3, 4, 9, 7, 1, 3, 13}; 第二词汇的笔画序列为{1, 3, 10, 5, 3, 4, 9, 11}, 则第一词汇和第二词汇的重合序列段有两段, 分别为 1, 3 和 5, 3, 4, 9, 长度分别为 2 和 4。

步骤 22, 基于第一词汇与第二词汇的重合序列段的总长度, 确定第一词汇的笔画序列与第二词汇的笔画序列的相似度。可选的, 在步骤 22 基于第一词汇与第二词汇的重合序列段的总长度, 确定第一词汇的笔画序列与第二词汇的笔画序列的相似度时, 可以采用如下公式:

$$S=2*p/(n+m)$$

其中, S 为第一词汇的笔画序列与第二词汇的笔画序列的相似度, p 为第一词汇与第二词汇的重合序列段的总长度, n 为第一词汇的笔画序列的长度, m 为第二词汇的笔画序列的长度。举例而言, 以一种可选的使用场景为例, 对步骤 103 进行详细说明如下:

步骤一: 将新词按笔画进行拆解, 形成笔画序列;

步骤二: 将已有词向量模型中的所有词, 按照笔画进行拆解, 每一个词得到一个笔画序列;

步骤三: 将步骤一中得到的序列值与步骤二中所有的序列集进行对比, 找出最相似的序列, 相似计算方法为: $2*(\text{共同存在的序列长度})/(\text{序列 1 的长度}+\text{序列 2 的长度})$ 。例如: 序列 1 为 abcd, 序列 2 为 bcde, 则共同存在的序列长度为 3 (bcd), 相似度为 $2*3/(4+4)=0.75$

步骤四: 找出最相似的序列后将该序列对应的词的向量赋予给新词。

步骤 105, 确定与第一词汇的笔画序列相似度最高的词汇对应的词向量, 并配置为第一词汇的初始词向量。也即, 如果判断出第一词汇不在词向量词典中, 依据第一词汇的笔画与词向量词典中的词汇的笔画进行比对, 寻找笔画序列相似度最高的词汇, 并将词向量词典中笔画序列相似度最高的词汇的词向量赋给第一词汇。本技术方案能够充分利用丰富的词向量词典, 利用中文本身的笔画构成信息来扩充其语义的特征, 利用已知词汇的词向量预测未知词汇的词向量, 弥补词向量未登录词随机赋值的缺陷, 并提升后续训练任务的精度, 减少训练梯度。需要说明的是, 在附图的流程图示出的步骤可以在诸如一组计算机可执行指令的计算机系统中执行, 并且, 虽然在流

程图中示出了逻辑顺序，但是在某些情况下，可以以不同于此处的顺序执行所示出或描述的步骤。

通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到根据上述实施例的方法可借助软件加必需的通用硬件平台的方式来实现，当然也可以通过硬件，但很多情况下前者是更佳的实施方式。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质（如 ROM/RAM、磁碟、光盘）中，包括若干指令用以使得一台终端设备（可以是手机，计算机，服务器，或者网络设备等）执行本申请各个实施例所述的方法。

实施例 2

在本实施例中还提供了一种词向量的配置装置，该装置用于实现上述实施例 1 及其优选实施方式，对于本实施例中未详述的术语或实现方式，可参见实施例 1 中的相关说明，已经进行过说明的不再赘述。如以下所使用的术语“模块”，是可以实现预定功能的软件和/或硬件的组合。尽管以下实施例所描述的装置较佳地以软件来实现，但是硬件，或者软件和硬件的组合的实现也是可以被构想的。

图 2 是根据本申请实施例的词向量的配置装置的示意图，如图 2 所示，该装置包括：第一确定模块 10，判断模块 20，拆解模块 30，计算模块 40 和第二确定模块 50。其中，第一确定模块用于确定待配置初始词向量的第一词汇；判断模块，用于判断第一词汇是否在词向量词典中，其中，词向量词典用于存储多个词汇与多个词向量的一一对应关系；拆解模块，用于如果判断出第一词汇不在词向量词典中，对第一词汇执行笔画拆解，得到笔画序列；计算模块，用于计算词向量词典中的每个词汇的笔画序列与第一词汇的笔画序列的相似度；第二确定模块，用于确定与第一词汇的笔画序列相似度最高的词汇对应的词向量，并配置为第一词汇的初始词向量。

可选的，词向量词典中包括第二词汇，计算模块包括：第一确定单元，用于确定第二词汇的笔画序列与第一词汇的笔画序列的重合序列段的总长度，其中，重合序列段为两个词汇的笔画序列中笔画排列相同的序列段；第二确定单元，用于基于第一词汇与第二词汇的重合序列段的总长度，确定第一词汇的笔画序列与第二词汇的笔画序列的相似度。可选的，第二确定单元基于第一词汇与第二词汇的重合序列段的总长度，确定第一词汇的笔画序列与第二词汇的笔画序列的相似度，采用如下公式： $S=2*p/(n+m)$ 其中，S 为第一词汇的笔画序列与第二词汇的笔画序列的相似度，p 为第一词汇与第二词汇的重合序列段的总长度，n 为第一词汇的笔画序列的长度，m 为第二词汇的笔画序列的长度。可选的，第一确定模块包括：获取单元，用于获取待分词的语料；分词单元，用于对语料进行分词，得到顺序的多个分词；第三确定单元，用于在多个分词中确定未配置初始词向量的首位分词，得到第一词汇。

本实施例通过对未登录在词向量词典中的词汇进行拆解，在词向量词典中查找与其笔画最接近的已登录词，进而将笔画相近词的词向量配置为未登录词的初始词向量，解决了相关技术中利用随机分配的方式配置未登录词的词向量时导致后续训练任务精度下降的技术问题，利用中文的

笔画中所携带的语义信息来为未登录词赋予初始词向量，能够减少后续训练任务的耗时、提高训练任务的精度。需要说明的是，上述各个模块是可以通过软件或硬件来实现的，对于后者，可以通过以下方式实现，但不限于此：上述模块均位于同一处理器中；或者，上述各个模块以任意组合的形式分别位于不同的处理器中。

显然，本领域的技术人员应该明白，上述的本申请各模块或各步骤可以用通用的计算装置来实现，它们可以集中在单个的计算装置上，或者分布在多个计算装置所组成的网络上，可选地，它们可以用计算装置可执行的程序代码来实现，从而，可以将它们存储在存储装置中由计算装置来执行，并且在某些情况下，可以以不同于此处的顺序执行所示出或描述的步骤，或者将它们分别制作成各个集成电路模块，或者将它们中的多个模块或步骤制作成单个集成电路模块来实现。这样，本申请不限制于任何特定的硬件和软件结合。

实施例 3

本申请的实施例还提供了一种非易失性可读存储介质，该存储介质中存储有计算机可读指令，其中，该计算机可读指令被设置为运行时执行上述任一项方法实施例中的步骤。本实施例通过对未登录在词向量词典中的词汇进行拆解，在词向量词典中查找与其笔画最接近的已登录词，进而将笔画相近词的词向量配置为未登录词的初始词向量，解决了相关技术中利用随机分配的方式配置未登录词的词向量时导致后续训练任务精度下降的技术问题，利用中文的笔画中所携带的语义信息来为未登录词赋予初始词向量，能够减少后续训练任务的耗时、提高训练任务的精度。可选地，在本实施例中，上述存储介质可以包括但不限于：U 盘、只读存储器（Read-Only Memory，简称为 ROM）、随机存取存储器（Random Access Memory，简称为 RAM）、移动硬盘、磁碟或者光盘等各种可以存储计算机程序的介质。

实施例 4

本申请的实施例还提供了一种电子装置，包括存储器和处理器，该存储器中存储有计算机可读指令，该处理器被设置为运行计算机可读指令以执行上述任一项方法实施例中的步骤。本实施例通过对未登录在词向量词典中的词汇进行拆解，在词向量词典中查找与其笔画最接近的已登录词，进而将笔画相近词的词向量配置为未登录词的初始词向量，解决了相关技术中利用随机分配的方式配置未登录词的词向量时导致后续训练任务精度下降的技术问题，利用中文的笔画中所携带的语义信息来为未登录词赋予初始词向量，能够减少后续训练任务的耗时、提高训练任务的精度。可选地，上述电子装置还可以包括传输设备以及输入输出设备，其中，该传输设备和上述处理器连接，该输入输出设备和上述处理器连接。以电子装置为电子装置为例，图 3 是本申请实施例的一种电子装置的硬件结构框图。如图 3 所示，电子装置可以包括一个或多个（图 3 中仅示出一个）处理器 302（处理器 302 可以包括但不限于微处理器 MCU 或可编程逻辑器件 FPGA 等的处理装置）和用于存储数据的存储器 304，可选地，上述电子装置还可以包括用于通信功能的传输设备 306 以及输入输出设备 308。本领域普通技术人员可以理解，图 3 所示的结构仅为示意，其并不对上述电子装置的结构造成限定。例如，电子装置还可包括比图 3 中所示更多或者更少的组件，

或者具有与图 3 所示不同的配置。存储器 304 可用于存储计算机可读指令，例如，应用程序的软件程序以及模块，如本申请实施例中的图像的识别方法对应的计算机可读指令，处理器 302 通过运行存储在存储器 304 内的计算机可读指令，从而执行各种功能应用以及数据处理，即实现上述的方法。存储器 304 可包括高速随机存储器，还可包括非易失性存储器，如一个或者多个磁性存储装置、闪存、或者其他非易失性固态存储器。在一些实例中，存储器 304 可进一步包括相对于处理器 302 远程设置的存储器，这些远程存储器可以通过网络连接至电子装置。上述网络的实例包括但不限于互联网、企业内部网、局域网、移动通信网及其组合。传输装置 306 用于经由一个网络接收或者发送数据。上述的网络具体实例可包括电子装置的通信供应商提供的无线网络。在一个实例中，传输装置 306 包括一个网络适配器（Network Interface Controller，简称为 NIC），其可通过基站与其他网络设备相连从而可与互联网进行通讯。在一个实例中，传输装置 306 可以为射频（Radio Frequency，简称为 RF）模块，其用于通过无线方式与互联网进行通讯。

以上所述仅为本申请的优选实施例而已，并不用于限制本申请，对于本领域的技术人员来说，本申请可以有各种更改和变化。凡在本申请的原则之内，所作的任何修改、等同替换、改进等，均应包含在本申请的保护范围之内。

权利要求

1.一种词向量的配置方法，所述方法包括：

确定待配置初始词向量的第一词汇；

判断所述第一词汇是否在词向量词典中，其中，所述词向量词典用于存储多个词汇与多个词向量的一一对应关系；

如果判断出所述第一词汇不在所述词向量词典中，对所述第一词汇执行笔画拆解，得到笔画序列；

计算所述词向量词典中的每个词汇的笔画序列与所述第一词汇的笔画序列的相似度；

确定与所述第一词汇的笔画序列相似度最高的词汇对应的词向量，并配置为所述第一词汇的初始词向量。

2.根据权利要求1所述的方法，所述词向量词典中包括第二词汇，所述计算所述词向量词典中的每个词汇的笔画序列与所述第一词汇的笔画序列的相似度，包括：

确定所述第二词汇的笔画序列与所述第一词汇的笔画序列的重合序列段的总长度，其中，所述重合序列段为两个词汇的笔画序列中笔画排列相同的序列段；

基于所述第一词汇与所述第二词汇的重合序列段的总长度，确定所述第一词汇的笔画序列与所述第二词汇的笔画序列的相似度。

3.根据权利要求2所述的方法，所述基于所述第一词汇与所述第二词汇的重合序列段的总长度，确定所述第一词汇的笔画序列与所述第二词汇的笔画序列的相似度，采用如下公式：

$$S=2*p/(n+m)$$

其中，S为所述第一词汇的笔画序列与所述第二词汇的笔画序列的相似度，p为所述第一词汇与所述第二词汇的重合序列段的总长度，n为所述第一词汇的笔画序列的长度，m为所述第二词汇的笔画序列的长度。

4.根据权利要求1所述的方法，所述确定待配置初始词向量的第一词汇，包括：获取待分词的语料；对所述语料进行分词，得到顺序的多个分词；在所述多个分词中确定未配置初始词向量的首位分词，得到所述第一词汇。

5.一种词向量的配置装置，包括：

第一确定模块，用于确定待配置初始词向量的第一词汇；

判断模块，用于判断所述第一词汇是否在词向量词典中，其中，所述词向量词典用于存储多个词汇与多个词向量的一一对应关系；

拆解模块，用于如果判断出所述第一词汇不在所述词向量词典中，对所述第一词汇执行笔画拆解，得到笔画序列；

计算模块,用于计算所述词向量词典中的每个词汇的笔画序列与所述第一词汇的笔画序列的相似度;

第二确定模块,用于确定与所述第一词汇的笔画序列相似度最高的词汇对应的词向量,并配置为所述第一词汇的初始词向量。

6.根据权利要求5所述的装置,所述词向量词典中包括第二词汇,所述计算模块包括:

第一确定单元,用于确定所述第二词汇的笔画序列与所述第一词汇的笔画序列的重合序列段的总长度,其中,所述重合序列段为两个词汇的笔画序列中笔画排列相同的序列段;

第二确定单元,用于基于所述第一词汇与所述第二词汇的重合序列段的总长度,确定所述第一词汇的笔画序列与所述第二词汇的笔画序列的相似度。

7.根据权利要求6所述的装置,所述第二确定单元基于所述第一词汇与所述第二词汇的重合序列段的总长度,确定所述第一词汇的笔画序列与所述第二词汇的笔画序列的相似度,采用如下公式:

$$S=2*p/(n+m)$$

其中, S 为所述第一词汇的笔画序列与所述第二词汇的笔画序列的相似度, p 为所述第一词汇与所述第二词汇的重合序列段的总长度, n 为所述第一词汇的笔画序列的长度, m 为所述第二词汇的笔画序列的长度。

8.根据权利要求5所述的装置,所述第一确定模块包括:

获取单元,用于获取待分词的语料;

分词单元,用于对所述语料进行分词,得到顺序的多个分词;

第三确定单元,用于在所述多个分词中确定未配置初始词向量的首位分词,得到所述第一词汇。

9.一种非易失性可读存储介质,所述存储介质中存储有计算机可读指令,其中,所述计算机可读指令被设置为运行时执行词向量的配置方法,包括:确定待配置初始词向量的第一词汇;判断所述第一词汇是否在词向量词典中,其中,所述词向量词典用于存储多个词汇与多个词向量的——对应关系;如果判断出所述第一词汇不在所述词向量词典中,对所述第一词汇执行笔画拆解,得到笔画序列;计算所述词向量词典中的每个词汇的笔画序列与所述第一词汇的笔画序列的相似度;确定与所述第一词汇的笔画序列相似度最高的词汇对应的词向量,并配置为所述第一词汇的初始词向量。

10.根据权利要求9所述的存储介质,所述词向量词典中包括第二词汇,所述计算机可读指令被设置为运行时执行所述计算所述词向量词典中的每个词汇的笔画序列与所述第一词汇的笔画序列的相似度,包括:确定所述第二词汇的笔画序列与所述第一词汇的笔画序列的重合序列段的总长度,其中,所述重合序列段为两个词汇的笔画序列中笔画排列相同的序列段;基于所述第一词汇与所述第二词汇的重合序列段的总长度,确定所述第一词汇的笔画序列与所述第二词汇的笔画序列的相似度。

11.根据权利要求10所述的存储介质,所述计算机可读指令被设置为运行时执行所述基于所述第一词汇与所述第二词汇的重合序列段的总长度,确定所述第一词汇的笔画序列与所述第二词汇的笔画序列的相似度,采用如下公式:

$$S=2*p/(n+m)$$

其中，S 为所述第一词汇的笔画序列与所述第二词汇的笔画序列的相似度，p 为所述第一词汇与所述第二词汇的重合序列段的总长度，n 为所述第一词汇的笔画序列的长度，m 为所述第二词汇的笔画序列的长度。

12. 根据权利要求 9 所述的存储介质，所述计算机可读指令被设置为运行时执行所述确定待配置初始词向量的第一词汇，包括：获取待分词的语料；对所述语料进行分词，得到顺序的多个分词；在所述多个分词中确定未配置初始词向量的首位分词，得到所述第一词汇。

13. 一种电子装置，包括存储器和处理器，所述存储器中存储有计算机可读指令，所述处理器被设置为运行所述计算机可读指令以执行词向量的配置方法，包括：确定待配置初始词向量的第一词汇；判断所述第一词汇是否在词向量词典中，其中，所述词向量词典用于存储多个词汇与多个词向量的一一对应关系；如果判断出所述第一词汇不在所述词向量词典中，对所述第一词汇执行笔画拆解，得到笔画序列；计算所述词向量词典中的每个词汇的笔画序列与所述第一词汇的笔画序列的相似度；确定与所述第一词汇的笔画序列相似度最高的词汇对应的词向量，并配置为所述第一词汇的初始词向量。

14. 根据权利要求 13 所述的电子装置，所述词向量词典中包括第二词汇，所述计算机可读指令被设置为运行时执行所述计算所述词向量词典中的每个词汇的笔画序列与所述第一词汇的笔画序列的相似度，包括：确定所述第二词汇的笔画序列与所述第一词汇的笔画序列的重合序列段的总长度，其中，所述重合序列段为两个词汇的笔画序列中笔画排列相同的序列段；基于所述第一词汇与所述第二词汇的重合序列段的总长度，确定所述第一词汇的笔画序列与所述第二词汇的笔画序列的相似度。

15. 根据权利要求 14 所述的电子装置，所述计算机可读指令被设置为运行时执行所述基于所述第一词汇与所述第二词汇的重合序列段的总长度，确定所述第一词汇的笔画序列与所述第二词汇的笔画序列的相似度，采用如下公式：

$$S=2*p/(n+m)$$

其中，S 为所述第一词汇的笔画序列与所述第二词汇的笔画序列的相似度，p 为所述第一词汇与所述第二词汇的重合序列段的总长度，n 为所述第一词汇的笔画序列的长度，m 为所述第二词汇的笔画序列的长度。

16. 根据权利要求 13 所述的电子装置，所述计算机可读指令被设置为运行时执行所述确定待配置初始词向量的第一词汇，包括：获取待分词的语料；对所述语料进行分词，得到顺序的多个分词；在所述多个分词中确定未配置初始词向量的首位分词，得到所述第一词汇。

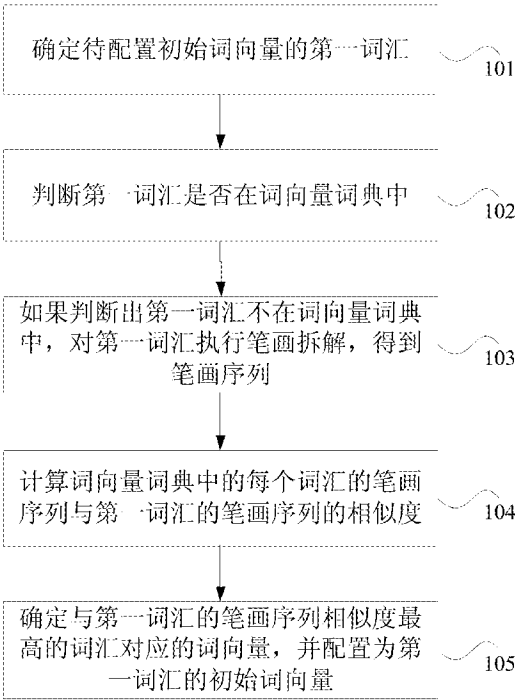


图 1

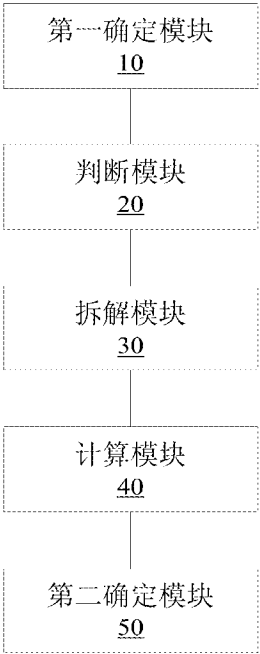


图 2

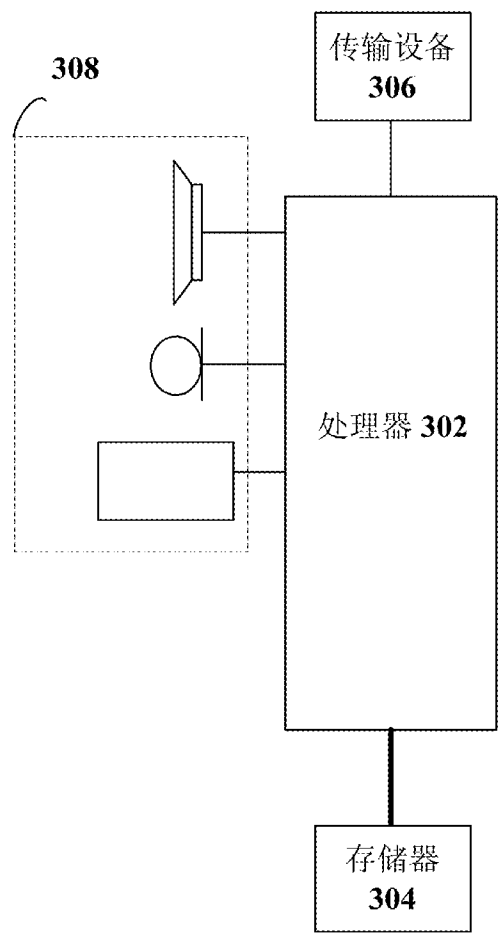


图 3

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2019/117725

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/00(2019.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNPAT; CNKI; WPI; EPODOC: 字, 词, 向量, 词典, 字典, 笔画, 相似, 重合, word, vector, dictionary, stroke, similar, resemble, overlap

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 110413990 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 05 November 2019 (2019-11-05) description, paragraphs [0026]-[0082]	1-16
X	CN 108959250 A (ZHONG'AN INFORMATION TECHNOLOGY SERVICES CO., LTD.) 07 December 2018 (2018-12-07) description, paragraphs [0053]-[0116]	1-16
A	CN 109299269 A (ALIBABA GROUP HOLDING LIMITED) 01 February 2019 (2019-02-01) entire document	1-16
A	US 2015178631 A1 (NEURAL ID LLC) 25 June 2015 (2015-06-25) entire document	1-16



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

19 February 2020

Date of mailing of the international search report

26 February 2020

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/
CN)
No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing
100088
China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2019/117725

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	110413990	A	05 November 2019	None			
CN	108959250	A	07 December 2018	None			
CN	109299269	A	01 February 2019	None			
US	2015178631	A1	25 June 2015	WO	2015034759	A1	12 March 2015
				EP	3042342	A1	13 July 2016

国际检索报告

国际申请号

PCT/CN2019/117725

A. 主题的分类

G06F 17/00 (2019.01) i

按照国际专利分类 (IPC) 或者同时按照国家分类和 IPC 两种分类

B. 检索领域

检索的最低限度文献 (标明分类系统和分类号)

G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库 (数据库的名称, 和使用的检索词 (如使用))

CNPAT; CNKI; WPI; EPDOC: 字, 词, 向量, 词典, 字典, 笔画, 相似, 重合, word, vector, dictionary, stroke, similar, resemble, overlap

C. 相关文件

类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
PX	CN 110413990 A (平安科技深圳有限公司) 2019年 11月 5日 (2019 - 11 - 05) 说明书第[0026]-[0082]段	1-16
X	CN 108959250 A (众安信息技术服务有限公司) 2018年 12月 7日 (2018 - 12 - 07) 说明书第[0053]-[0116]段	1-16
A	CN 109299269 A (阿里巴巴集团控股有限公司) 2019年 2月 1日 (2019 - 02 - 01) 全文	1-16
A	US 2015178631 A1 (NEURAL ID LLC) 2015年 6月 25日 (2015 - 06 - 25) 全文	1-16

☐ 其余文件在C栏的续页中列出。☒ 见同族专利附件。

* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期

2020年 2月 19日

国际检索报告邮寄日期

2020年 2月 26日

ISA/CN的名称和邮寄地址

中国国家知识产权局 (ISA/CN)

中国北京市海淀区蓟门桥西土城路6号 100088

传真号 (86-10)62019451

授权官员

陈希元

电话号码 86-(10)-53961594

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2019/117725

检索报告引用的专利文件			公布日 (年/月/日)	同族专利			公布日 (年/月/日)
CN	110413990	A	2019年 11月 5日	无			
CN	108959250	A	2018年 12月 7日	无			
CN	109299269	A	2019年 2月 1日	无			
US	2015178631	A1	2015年 6月 25日	WO	2015034759	A1	2015年 3月 12日
				EP	3042342	A1	2016年 7月 13日