

(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
2019 年 12 月 5 日 (05.12.2019)



(10) 国际公布号
WO 2019/228014 A1

(51) 国际专利分类号:
G06K 9/62 (2006.01)

(21) 国际申请号: PCT/CN2019/077438

(22) 国际申请日: 2019 年 3 月 8 日 (08.03.2019)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
20181055518.X 2018年6月1日 (01.06.2018) CN

(71) 申请人: 阿里巴巴集团控股有限公司 (ALIBABA GROUP HOLDING LIMITED) [—/CN]; 开曼群岛
大开曼资本大厦一座四层 847 号 邮箱,
Grand Cayman (KY)。

(72) 发明人: 林建滨(LIN, Jianbin); 中国浙江省杭州市
余杭区文一西路969号3号楼5楼阿里巴巴集团法
务部, Zhejiang 311121 (CN)。 周俊(ZHOU, Jun);
中国浙江省杭州市余杭区文一西路969号3号楼5楼
阿里巴巴集团法务部, Zhejiang 311121 (CN)。

(74) 代理人: 北京博思佳知识产权代理
有限公司 (BEIJING BESTIPR INTELLECTUAL
PROPERTY LAW CORPORATION); 中国北京市
海淀区上地三街9号嘉华大厦B座409,
Beijing 100085 (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,
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,

(54) Title: METHOD AND APPARATUS FOR PERFORMING, FOR TRAINING CORPUS, NEGATIVE SAMPLING IN WORD FREQUENCY TABLE

(54) 发明名称: 针对训练语料从词频表中进行负例采样的方法及装置

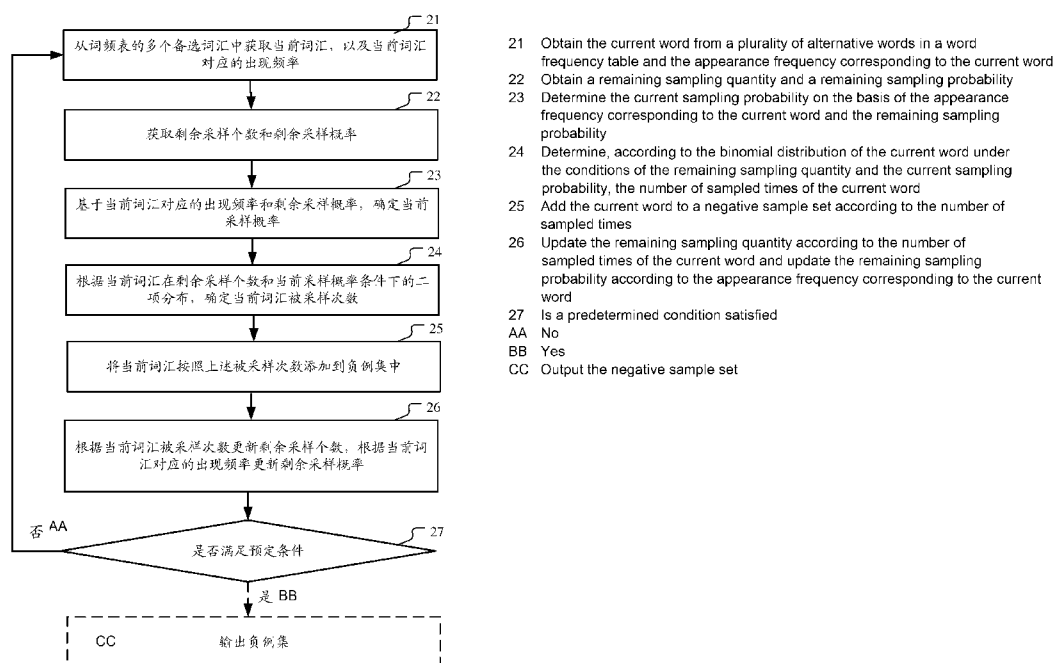


图 2

(57) Abstract: Embodiments of the description provide a method and apparatus for performing, for a training corpus, negative sampling in a word frequency table. The method comprises: first obtaining, from a word frequency table, an alternative word as the current word, and obtaining a remaining sampling quantity, a remaining sampling probability, and the current sampling probability; determining, on the basis of the binomial distribution of the current word under the conditions of the remaining sampling quantity and the current sampling probability, the number of sampled times of the current word; and then adding the current word to a negative sample set

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)。

本国际公布:

— 包括国际检索报告 (条约第21条(3))。

according to the number of sampled times. Because in the case that the steps above are executed to the current word, the current word can be added to a negative sample set for the number of sampled times, such that the number of total negative sampling times is reduced, thereby reducing the negative sampling time and implementing quick and effective negative sampling.

(57) 摘要: 本说明书实施例提供一种针对训练语料从词频表中进行负例采样的方法和装置, 根据该方法, 首先从词频表获取一个备选词汇作为当前词汇, 并获取剩余采样个数, 剩余采样概率和当前采样概率, 基于当前词汇在剩余采样个数和当前采样概率条件下的二项分布, 确定当前词汇被采样次数, 然后将当前词汇按照被采样次数添加到负例集中。由于对一个当前词汇执行以上步骤的情况下, 可以添加被采样次数的当前词汇到负例集, 使总体的负例采样次数减少, 从而减少负例采样的时间, 进而能够快速有效地进行负例采样。

针对训练语料从词频表中进行负例采样的方法及装置

技术领域

[01]本说明书一个或多个实施例涉及计算机技术领域，尤其涉及通过计算机针对训练语料从词频表中进行负例采样的方法和装置。

5 背景技术

[02]噪声对比估计 NCE (Noise Contrastive Estimation) 是 Word2Vec、Node2Vec 等无监督算法中通常用到的损失函数。应用该损失函数时，需要先对训练语料中每一个词汇及其上下文生成一定数量的负例。其中，对训练预料中的某个词汇而言，该词及其上下文词汇之外的任一个词都可以是一个负例。一般地，上述负例根据训练语料的词汇分布随机采样，该词汇分布例如采用词频表近似表示。

[03]常规技术中，在使用负例时对相应词汇生成负例。具体地，将词汇分布（如词频表）映射到一个区间上，生成区间内的数值，从而查找相对应的词汇作为负例。在训练预料中的字典中词汇较多（例如数亿级别），需要负例数量较大的情况下，希望能有改进的方案，减少采样的时间，从而能够快速有效地进行负例采样。

15 发明内容

[04]本说明书一个或多个实施例描述了一种方法和装置，在训练预料中的字典中词汇较多（例如数亿级别），需要负例数量较大的情况下，可以减少采样的时间，从而能够快速有效地进行负例采样。

[05]根据第一方面，提供了一种针对训练语料从词频表中进行负例采样的方法，所述词频表包括多个备选词汇和各个备选词汇在所述训练语料中的出现频率，所述方法包括：

[06]从所述多个备选词汇的未采样词汇集合中获取当前词汇，以及当前词汇对应的出现频率；

[07]获取针对所述未采样词汇集合确定的剩余采样个数和剩余采样概率；

[08]基于所述当前词汇对应的出现频率和所述剩余采样概率，确定当前采样概率；

[09]根据所述当前词汇在所述剩余采样个数和所述当前采样概率条件下的二项分布，确

定所述当前词汇被采样次数;

[10]将所述当前词汇按照所述被采样次数添加到负例集中;

[11]根据所述当前词汇被采样次数更新所述剩余采样个数,并根据所述当前词汇对应的出现频率更新所述剩余采样概率,用于对所述词频表中的其他备选词汇进行采样,直到
5 检测到预定条件得到满足。

[12]在一个实施例中,基于所述当前词汇对应的出现频率和所述剩余采样概率,确定当前采样概率包括:将所述当前采样概率确定为,所述当前词汇对应的出现频率与所述剩余采样概率的比值。

[13]根据一个实施例,所述确定所述当前词汇被采样次数包括:

10 [14]模拟执行所述剩余采样个数次的采样操作,其中,在各次采样操作中,所述当前词汇被采样到的概率为所述当前采样概率;

[15]确定所述被采样次数为,所述当前词汇在所述剩余采样个数次的采样操作中被采样到的次数。

15 [16]在一个实施方式中,根据所述当前词汇被采样次数更新所述剩余采样个数包括:将剩余采样个数更新为,所述剩余采样个数与所述被采样次数的差。

[17]进一步地,在一个实施例中,所述预定条件包括:所述负例集中的负例个数达到预设数目;或者更新后的剩余采样个数为零;或者所述未采样词汇集合为空。

20 [18]在一个可能的实施例中,所述根据所述当前词汇对应的出现频率更新所述剩余采样概率包括:将剩余采样概率更新为,所述剩余采样概率与所述当前词汇对应的出现频率的差。

[19]根据一种可能的设计,所述方法还包括:在所述负例集中的负例个数满足预定条件的情况下,输出所述负例集。

[20]在一些可能的实施例中,所述方法还包括:针对训练语料中的训练词汇,从所述负例集中选取负例。

25 [21]进一步地,在一些实施例中,从所述负例集中选取负例包括:生成预定区间上的随机数,其中,所述预定区间上的各个数值分别与所述负例集中的各个负例相对应,所述随机数取自所述各个数值;从所述负例集中获取与所述随机数相对应的负例。

[22]根据一种实施方式,所述从所述负例集中获取与所述随机数相对应的负例包括:

[23]比较所获取的负例与所述训练词汇是否一致；在一致的情况下，重新执行所述生成预定区间上的随机数的步骤。

[24]根据一种可能的设计，所述方法还包括：检测所述负例集的更新条件是否满足；在所述更新条件满足的情况下，重新生成负例集。

5 [25]根据第二方面，提供一种针对训练语料从词频表中进行负例采样的装置，所述词频表包括多个备选词汇和各个备选词汇在所述训练语料中的出现频率，所述装置包括：

[26]第一获取单元，配置为从所述多个备选词汇的未采样词汇集合中获取当前词汇，以及当前词汇对应的出现频率；

10 [27]第二获取单元，配置为获取针对所述未采样词汇集合确定的剩余采样个数和剩余采样概率；

[28]第一确定单元，配置为基于所述当前词汇对应的出现频率和所述剩余采样概率，确定当前词汇对应的当前采样概率；

[29]第二确定单元，配置为根据所述当前词汇在所述剩余采样个数和所述当前采样概率条件下的二项分布，确定所述当前词汇被采样次数；

15 [30]添加单元，配置为将所述当前词汇按照所述被采样次数添加到负例集中；

[31]更新单元，配置为根据所述当前词汇被采样次数更新所述剩余采样个数，根据所述当前词汇对应的出现频率更新所述剩余采样概率，用于对所述词频表中的其他备选词汇进行采样，直到检测到所述负例集中的负例个数满足预定条件。

20 [32]根据第三方面，提供了一种计算机可读存储介质，其上存储有计算机程序，当所述计算机程序在计算机中执行时，令计算机执行第一方面的方法。

[33]根据第四方面，提供了一种计算设备，包括存储器和处理器，其特征在于，所述存储器中存储有可执行代码，所述处理器执行所述可执行代码时，实现第一方面的方法。

[34]通过本说明书实施例提供的方法和装置，针对训练语料从词频表中进行负例采样时，对于一个从词频表获取一个备选词汇作为当前词汇，并获取剩余采样个数和剩余采样概率，基于当前词汇在剩余采样个数和当前采样概率条件下的二项分布，确定当前词汇被采样次数，然后将当前词汇按照被采样次数添加到负例集中。由于对一个当前词汇执行以上步骤的情况下，可以添加被采样次数的当前词汇到负例集，使总体的负例采样次数减少，从而减少负例采样的时间，进而能够快速有效地进行负例采样。

附图说明

[35]为了更清楚地说明本发明实施例的技术方案，下面将对实施例描述中所需要使用的附图作简单地介绍，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，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其它的附图。

[36]图 1 示出本说明书披露的一个实施例的实施场景示意图；

[37]图 2 示出根据一个实施例的针对训练语料从词频表中进行负例采样的方法流程图；

[38]图 3 示出从负例集中选取负例的一个具体例子；

[39]图 4 示出从负例集中选取负例的另一个具体例子；

10 [40]图 5 示出根据一个实施例的用于针对训练语料从词频表中进行负例采样的装置的示意性框图。

具体实施方式

[41]下面结合附图，对本说明书提供的方案进行描述。

15 [42]图 1 为本说明书披露的一个实施例的实施场景示意图。在一个无监督模型（例如 Word2Vec、Node2Vec）训练过程中，损失函数可以为噪声对比估计 NCE，表达式如下：

$$J_{\theta} = \sum_{w_i \in V} [\log P(y = 1 | w_i, c_i) + \sum_{j=1}^k \log P(y = 0 | w_{ij}, c_j)];$$

[43]其中：V 表示字典； w_i 表示第 i 个训练词汇； c_i 表示与第 i 个词汇相邻的上下文词汇；k 表示 w_i 对应的负例个数； w_{ij} 表示 w_i 的第 j 个负例； c_j 表示与该第 j 个负例相邻的上下文词汇。

20 [44]由以上公式可知，在语料训练过程中，对每个训练词汇 w_i ，都需要从其在字典的概率分布中进行 k 次随机取样，获取 k 个负例。

[45]字典中的多个词汇和各个词汇在训练语料中的出现频率通常通过词频表来表示。往往将字典 V 对应的词频表投影到一个区间[0, 1]上，区间中各段的长度和相应词汇的出现频率成正比。进一步地，在一种负例采样方式中，将各个词汇对应的区间段按照最小频率单位划分为多个“格子”，并记录各个格子的编号作为索引。一个词汇的出现频率越大，对应的区间段越长，包含的格子数目越多。每次进行负例采样时，生成索引中的

随机数，并将索引为该随机数的相应词汇作为负例。实际使用中，索引数量越多，对字典词频表的模拟越准确。例如，由于每个索引对应一个“格子”，为了保证每个词汇都有对应的索引，出现频率最小的词汇对应的索引至少为 1 个，其他词汇对应的频率可能有多个，如词汇 1 的出现频率为 0.03，词汇 2 的出现频率为词汇 0.001...，则可以使词汇 2 对应 1 个索引，而词汇 1 对应 30 个索引。当字典 V 中词汇量较多（如以亿计）时，索引数量更多。需要较大的存储空间，甚至存储到远程服务器，每次获取负例花费额外通信时间。

[46]如图 1 所示，本说明书实施例提供了一种方案，先通过词频表进行负例的预采样，将采样的词汇加入到负例集中。在预采样过程中，进行批量采样，对词频表中的各个词汇只采样一次，而采样的数量可以是多个，并保证最终各个词汇的采样数量与词频表中的出现频率一致。如图 1 中，对词频表中的词汇 w1 采样 s1 次，对词汇 w2 采样 s2 次，对词汇 w3 采样 s3 次，等等。如此，既减少了预采样时的采样次数，并保证负例集中各个词汇的采样数量与词频表中的出现频率一致。

[47]在词汇训练过程中，如果需要负例，则从负例集中随机获取相应数量的负例即可。

如图 1 中，针对训练词汇 u1 从负例集中随机取出 k1 个负例，对训练词汇 u2 从负例集中随机取出 k2 个负例，对训练词汇 u3 从负例集中随机取出 k3 个负例；等等。由于负例集中是预采样的负例，其中各个词汇的采样数量与词频表中的出现频率一致，因此使用时只需随机取出相应数量负例即可，而无需考虑词汇在词汇表中的出现频率，就可以保证对各个负例的取样概率与词频表中相应词汇的出现频率一致。如此，运算复杂度大大降低。同时，预采样的负例集可以多次使用，进一步提高模型训练中负例采样的有效性。

[48]可以理解，图 1 中的计算平台可以是具有一定运算能力的各种装置、设备，例如台式计算机、服务器等等。可以理解，计算平台还可以是上述装置、设备组成的设备集群。在该计算平台是多个设备或装置的情况下，根据一种实施方式，可以由其中一些设备或装置完成负例采样操作，生成负例集，另一些设备或装置获取该负例集，在训练词汇时从该负例集中随机取出负例。

[49]下面描述针对训练语料从词频表中进行负例采样的具体执行过程。

[50]图 2 示出了本说明书一个实施例的针对训练语料从词频表中进行负例采样的方法流程图。该方法的执行主体例如是图 1 的计算平台。如图 2 所示，该方法包括以下步骤：步骤 21，从词频表的未采样词汇集合中获取当前词汇，以及当前词汇对应的出现频率；

步骤 22, 获取针对所述未采样词汇集合确定的剩余采样个数和剩余采样概率; 步骤 23, 基于当前词汇对应的出现频率和所述剩余采样概率, 确定当前词汇对应的当前采样概率; 步骤 24, 根据当前词汇在剩余采样个数和当前采样概率条件下的二项分布, 确定当前词汇被采样次数; 步骤 25, 将当前词汇按照上述被采样次数添加到负例集中; 步骤 26, 根据当前词汇被采样次数更新上述剩余采样个数, 以及根据当前词汇对应的出现频率更新上述剩余采样概率, 用于对词频表中的其他备选词汇进行采样, 直到检测到预定条件得到满足。下面描述以上各个步骤的具体执行过程。

[51] 首先, 在步骤 21, 从词频表的未采样词汇集合中获取当前词汇, 以及当前词汇对应的出现频率。可以理解, 词频表可以包括多个备选词汇, 以及各个备选词汇在训练预料中的出现频率。这多个备选词汇可以包括训练预料中出现的所有词汇。词频表可以是表格、向量、数组、键值对 (key-value) 等各种形式, 本说明书对此不作限定。

[52] 各个备选词汇在训练语料中的出现次数各不相同, 如此, 词频表还可以通过出现频率来衡量各个词汇在训练语料中的比重。一个备选词汇的出现频率可以包括, 该备选词汇在训练语料中的总出现次数与训练语料中的总词汇数量的比值。其中, 该总词汇数量计算时不合并重复的词汇, 即: 统计时, 每个词汇每出现一次, 总词汇数量增加 1。

[53] 如前所述, 根据本说明实施例的方法, 对于词频表中的各个备选词汇, 可以依次对其进行批量采样。因此, 词频表可以划分为已采样词汇集合和未采样词汇集合, 其各自包括已采样的备选词汇和未采样的备选词汇。在步骤 21, 从上述未采样词汇集合中获取当前词汇, 以及当前词汇对应的出现频率, 以便用于接下来对当前词汇进行采样。

[54] 在一个实施例中, 可以按照词频表中各个词汇的存储地址依次获取一个备选词汇作为当前词汇, 按这样的顺序取词可以保证不会重复取词重复采样, 即每次都是从未采样的词汇集合中获取当前词汇。例如, 获取词频表的存储地址, 根据各个备选词汇相对于词频表的存储地址的偏移量来获取一个备选词汇。此时, 词频表的存储地址加上偏移量就是各个备选词汇的存储地址。如上述偏移量在[0000-FFFF]之间, 可以首先将偏移量为 0000 的备选词汇获取为当前词汇, 下一轮执行该流程时会获取偏移量为 0001 的备选词汇, 以此类推。可选地, 备选词汇及其出现频率可以存储在同一个存储地址对应的存储单元中, 此时, 可以同时获取当前词汇及其出现频率。另一种情况下, 备选词汇及其出现频率可以存储在不同的存储单元, 此时, 可以按照备选词汇的存储地址获取相关联的出现频率。

[55] 在另一个实施例中, 也可以按照词频表中各个备选词汇的排列顺序获取一个备选词

汇作为当前词汇，如此保证每次都是从未采样词汇集合中获取当前词汇。例如在词频表是表格的情况下，按照表格中的各行顺序获取备选词汇，如第一轮执行该流程时获取第一行的备选词汇，第二轮获取第二行的备选词汇，等等。在表格有多个列的情况下，还可以按照第一列第一行、第一列第二行……这样的顺序这样获取一个备选词汇。

5 [56]在步骤 22 中，获取针对未采样词汇集合确定的剩余采样个数 s 和剩余采样概率 r 。

[57]剩余采样个数 s 可以是负例集中还需要的负例个数，也是未采样词汇集合中所有未采样词汇需要被采样的总次数。

[58]初始地，剩余采样个数 s 是整个负例集所需要的总的负例个数 S_0 。在一个实施例中，整个负例集所需要的负例个数 S_0 可以根据训练语料中的词汇数计算得到的，也可以是人工设定的，本申请对此不作限定。例如，在前述损失函数中，针对每个训练词汇需要取 k 个负例，假定训练预料中包含 n 个词汇，那么可以将负例个数 S_0 设定为 $S_0=n*k$ 。
10 在另一实施例中，还可以将该初始需要的负例个数 S_0 设定为训练预料中词汇数目的预定比例，等等。

[59]在初始设置之后，每对一个备选词汇采样完毕，会对剩余采样次数进行更新，即剩
15 余采样次数减小相应次数。例如，由人工设定了负例集需要 10000 个负例，对备选词汇 w_0 采样了 5 次，则剩下的词总共需要的采样次数为 $10000-5=9995$ 。

[60]剩余采样概率 r ，可以是所有未采样词汇在生成负例集的负例采样过程中的总采样概率。作为示例，假定词频表中的备选词汇包括 w_0 、 w_1 、 w_2 ……，对应出现频率 p_0 、 p_1 、 p_2 ……，剩余采样概率 r 表示，未采样词汇的总采样概率。初始地，所有备选词汇
20 均未被采样，此时剩余采样概率 r 即为词频表中所有备选词汇在生成负例集的负例采样过程中的总采样概率，因此， r 的初始值为 1。

[61]可以理解，为了保证最终负例集中的各个负例在负例集中的占比与相应的备选词汇的出现频率一致，每对一个备选词汇采样完毕，会对剩余采样概率也进行更新。例如，在第一个备选词汇 w_0 采样完毕的情况下，剩余采样概率会被更新为 $r' = r - p_0 = 1 - p_0$ ，以此类推，在第二个备选词汇 w_1 采样完毕的情况下，剩余采样概率会被更新为 $r' - p_1 = 1 - p_0 - p_1$ ……。
25

[62]因此，如果当前词汇 w_i 为词频表中的第一个词汇，那么在步骤 22，分别获取负例集中需要的负例个数的初始值 S_0 为所述剩余采样个数 s ，获取初始值 $r=1$ 为所述剩余采样概率 r 。如果当前词汇 w_i 不是词汇表的第一个词汇，那么在步骤 22，分别读取在对前

一词汇 w_{i-1} 采样之后更新得到的剩余采样个数 s 和剩余采样概率 r 。

[63] 步骤 23, 基于当前词汇对应的出现频率 p_i 和所述剩余采样概率 r , 确定当前词汇对应的当前采样概率 P 。当前采样概率 P , 可以是当前词汇在整个未采样集合中的采样概率。

5 [64] 可以理解, 由于本实施例是针对各个备选词汇批量采样, 换言之, 要一次采集够相应数量的某个备选词汇。那么, 当一个备选词汇采样完毕后, 其会被加入已采样词汇集合, 后续被采样到的概率是 0。这样, 之后的采样过程不需要考虑采样过的备选词汇, 而是在未采样词汇集合中进行。其中, 由于当前词汇还未进行采样, 所以上述未采样词汇集合包括当前词汇在内。

10 [65] 仍然参考以上示例, 容易理解, 备选词汇 w_0 、 w_1 、 w_2 ... 的出现频率分别为 p_0 、 p_1 、 p_2 ...。在对第一个备选词汇 w_0 采样的情况下, 采样概率为 p_0 , 剩余备选词汇 (未采样词汇集合) 的总采样概率为 $r=1-p_0=p_1+p_2+\dots$ 。第二个备选词汇 w_1 的出现频率为 p_1 , 则其在剩余备选词汇 (未采样词汇集合) 中的采样概率为 $p_1/(p_1+p_2+\dots)=p_1/1-p_0$ 。以此类推, 对于当前词汇 w_i , 当前采样概率也可以表示为: $P=p_i/r$, 即: 当前词汇对应的出现频率 p_i 和剩余采样概率 r 的比值。

15 [66] 步骤 24, 根据当前词汇 w_i 在剩余采样个数 s 和当前采样概率 P 条件下的二项分布, 确定当前词汇 w_i 的被采样次数 b 。可以理解, 词频表中的备选词汇都对应一个被采样次数, 例如图 1 所示的对词汇 w_1 采样 s_1 次, 对词汇 w_2 采样 s_2 次, 对词汇 w_3 采样 s_3 次, 等等, 以完成对备选词汇批量采样。可选地, 当一个备选词汇的出现频率较小时, 其被采样次数可能为 0。

[67] 在一个实施例中, 利用二项分布确定上述被采样次数。二项分布是若干次独立的伯努利试验中成功的次数的离散概率分布。在每次试验中, 只能出现两个可能结果中的一个结果, 并且各次实验的实验结果相互独立。每个结果发生的概率在每一次独立试验中都保持不变, 当试验次数为 1 时, 二项分布服从 0-1 分布, 亦即, 对其中的一个结果而言, 要么发生 (成功), 要么不发生。

[68] 用 ξ 表示随机试验的结果。如果某个事件发生的概率是 p , 则不发生的概率 $q=1-p$, 用 P 表示 n 次独立重复试验中该事件发生 k 次的概率为:

$$P(\xi=k)=C(n, k) \times p^k \times (1-p)^{(n-k)};$$

[69] 其中: $C(n, k)=n!/(k! \times (n-k)!)$ 。

[70]这个就是该事件在次数 n 和概率 p 条件下的二项分布概率。

[71]具体到该步骤 24 中，在一个实施例，调用二项分布函数 $\text{Binomial}(s, P)$ 来确定当前词汇的被采样次数 b 。可以看到，该二项分布函数的参数为剩余采样个数 s 和当前采样概率 P ，表示在 s 次采样试验，每次采样到当前词汇 w_i 的概率为 P 的条件下， w_i 被采样到的次数。

[72]上述二项分布函数的执行可以包括，模拟执行剩余采样个数 s 次的采样操作（伯努利试验），相当于采样试验，其中，这些采样操作是针对剩余的备选词汇执行的。在各次采样操作中，确保当前词汇 w_i 被采样到（试验成功）的概率为当前采样概率 P 。统计当前词汇被采样到的次数，并确定当前词汇的被采样次数 b 为，当前词汇在剩余采样个数 s 次的采样操作中被采样到的次数。

[73]在另一个实施例，还可以从满足二项分布采样条件的数值中随机获取一个值，作为当前词汇被采样次数。可以理解，根据二项分布的含义，假设当前词汇“财富”最终被采样 b 次，则数值 b 满足的条件可以为：与剩余采样个数的比值应与当前采样概率一致。例如，剩余采样个数 s 为 8000，当前采样概率 P 为 0.03，则可能当 b 在 200-272 范围内时 $b/8000$ 都可以四舍五入为 0.03。如此，可以在 200-272 之间取一个随机数，作为当前词汇“财富”的被采样次数。

[74]在步骤 25，将当前词汇 w_i 按照上述被采样次数 b 添加到负例集中。这里，步骤 24 确定的被采样次数 b 是多少，就添加多少个当前词汇到负例集。如以上示例中 b 的取值为 232，则添加 232 个当前词汇“财富”到负例集中。

[75]步骤 26，根据当前词汇被采样次数 b 更新上述剩余采样个数 s ，根据当前词汇对应的出现频率 p_i 更新上述剩余采样概率 r 。更新后的剩余采样次数 s 和剩余采样概率 r 可以用于对词频表中的其他备选词汇进行采样。例如对下一个备选词汇而言，步骤 22 中所获取的剩余采样个数和剩余采样概率是本步骤更新后的剩余采样个数和剩余采样概率。

[76]可以理解，对每个备选词汇进行采样后，可以将其从未采样词汇集合移动到已采样词汇集合中。如此，针对未采样词汇集合设置的剩余采样个数 s 会减少相应数量，剩余采样概率 r 也会发生改变。换句话说，对于下一个备选词汇而言，采样条件发生变化。例如，负例集需要 10000 个负例，则初始剩余负例个数为 10000，初始剩余采样频率为 1，当对一个出现频率为 0.03 的备选词汇 w_0 采样 200 次后，对下一个出现频率为 0.05

的备选词汇而言,是在剩余负例个数为 9800、剩余采样概率为 0.97 的情况下进行采样。

[77]在一个实施例中,对当前词汇采样完成后,可以将剩余采样个数 s 更新为,原剩余采样个数与当前词汇被采样次数 b 的差。如实现逻辑为: $s=s-b$; 其中, s 为剩余采样个数, b 为当前词汇 w_i 的被采样次数。

- 5 [78]在一个实施例中,将剩余采样概率 r 更新为,原剩余采样概率与当前词汇对应的出现频率 p_i 的差。如实现逻辑为: $r=r-p_i$; 其中, r 为剩余采样概率, p_i 为当前词汇 w_i 的出现频率。

- 10 [79]值得说明的是,由于负例集中需要的负例个数是有限的,因此,还可以预先设置与负例集中的负例个数相关的预定条件,当满足该条件时,停止负例采样,否则继续针对词频表的其他备选词汇执行以上采样流程。该检测步骤可以在更新步骤 26 之后进行,也可以和步骤 26 并列执行。其可以是步骤 26 的一部分,也可以是步骤 26 的后续步骤 27。下面以后续步骤 27 的方式详细说明该检测步骤的具体实现。

- 15 [80]在该步骤 27 中,检测预定条件是否得到满足,在满足预定条件的情况下,该负例采样流程结束,在不满足预定条件的情况下,根据更新后的剩余采样个数和剩余采样概率对词频表中的其他备选词汇采样。

[81]在一个实施例中,预定条件可以包括负例集中的负例总个数达到初始剩余采样个数,如人工设置的负例个数 10000 等。

[82]在另一个实施例中,预定条件可以包括更新后的剩余采样个数为 0。此时,代表着已经不需要采集其他备选词汇作为负例了。

- 20 [83]在另一实施例中,预定条件可以包括,未采样词汇集合为空。此时,已经对词频表中所有词汇进行采样。

[84]根据另一方面的实施例,在上述预定条件得到满足的情况下,还可以输出该负例集。该负例集可以输出到本地,也可以输出到其他设备。负例集中的各个词汇可以按照采样顺序排列,也可以随机打乱顺序排列,本申请对此不作限定。

- 25 [85]在进一步的实施例中,针对训练语料中的训练词汇,可以从该负例集中选取负例。例如,针对训练语料中的训练词汇 U_i 需要 k 个负例的情况下,可以直接从该负例集中取出 k 个词汇。

[86]根据一方面的实施例,负例集中的词汇可以和预定区间上的各个数值相对应,如图

3 所示, 将负例集 31 中的各个备选负例与数值区间 32 中的数值一一对应。假如负例集 31 中有 10000 个预采样到的负例词汇, 可以选择区间[1, 10000]上的正整数, 每个数值对应一个负例词汇。对一个训练词汇选取负例时, 生成这个预定区间上的随机数, 例如数值区间 32 上的随机数 5, 则可以选择负例集 31 中与数值 5 相对应的负例词汇 w_1 。实践中, 需要多少个负例, 则生成多少个随机数。可以一次生成一个随机数, 获取相应负例, 也可以一次生成多个随机数, 批量获取相应负例, 本申请对此不作限定。

[87]可以理解, 在极小的概率下, 所获取的负例还可能与训练词汇本身或其关联词汇一致, 上述关联词汇例如是, 在上下文预测模型中训练词汇的上下文, 在同义词预测模型中训练词汇的同义词等。在这样的情况下, 从负例集中选取的词汇将不能作为该训练词汇的负例。因此, 针对训练词汇从负例集选取负例时, 在选取的词汇与训练词汇本身或其关联词汇一致的情况下, 重新执行生成预定区间上的随机数的步骤, 生成新的随机数, 获取新的随机数对应的负例词汇。

[88]根据另一方面的实施例, 负例集中的各个词汇随机打乱顺序排列的情况下, 还可以从一个选定位置开始按顺序选定 k 个词汇作为负例。该选定位置可以按照一定规则确定, 也可以将生成的随机数对应的位置确定为选定位置。例如: 查找到第一个和训练词汇相同的词汇, 将下一个词汇的位置作为选定位置。再例如: 上述预定区间的例子中, 生成一个 1-10000 之间的随机数。这种情况下只需生成一个随机数, 计算量较小。如图 4 所示, 对于负例集 41 而言, 对一个训练词汇需要取出 7 个负例的情况下, 可以生成一个数值区间 42 上的随机数, 例如是数值 5, 则可以将数值 5 对应的位置作为选定位置, 从该选定位置开始获取负例集 41 中, 区间 43 上的 7 个备选负例 w_3 、 w_9 、 w_3 、 w_7 、 w_6 、 w_4 、 w_8 , 作为该训练词汇的负例。

[89]如此, 针对训练语料的训练词汇获取负例的过程就简化很多, 获取速度也得以提高。

[90]在一些可能的设计中, 图 2 所示的流程还可以包括以下步骤: 检测负例集的更新条件是否满足; 在更新条件满足的情况下, 重新执行针对训练语料从词频表中进行负例采样的方法, 以重新生成负例集。可以理解, 当需要的负例集中词汇数量较多时, 例如几亿, 计算量也非常大, 因此可以一次生成一个较小的负例集, 例如 1 千万, 然后设定负例集更新条件 (例如使用次数为 1 千万等), 对负例集进行更新。由于上述方法执行过程中, 针对每个备选词汇, 获取被采样次数时, 会模拟执行剩余采样个数 s 次的采样操作 (伯努利试验), 或者从满足条件的数值中随机获取一个值, 等等, 因此, 每次重新执行针对训练语料从词频表中进行负例采样的方法生成的负例集都可能不同。

[91]回顾以上过程，一方面，由于负例集中是预采样的负例，使用时只需随机取出相应数量负例即可，而无需考虑词汇在词汇表中的出现频率，运算复杂度大大降低。另一方面，在预采样过程中，进行批量采样，对词频表中的各个词汇只采样一次，而采样的数量可以是多个，从而减少负例采样的时间，进而能够快速有效地进行负例采样。总之，

5 图 2 示出的流程可以提高负例采样的有效性。

[92]根据另一方面的实施例，还提供一种针对训练语料从词频表中进行负例采样的装置。图 5 示出根据一个实施例的用于针对训练语料从词频表中进行负例采样的装置的示意性框图。如图 5 所示，用于针对训练语料从词频表中进行负例采样的装置 500 包括：第一
10 获取单元 51，配置为从词频表的未采样词汇集合中获取当前词汇，以及当前词汇对应的出现频率；第二获取单元 52，配置为获取针对未采样词汇集合确定的剩余采样个数和剩余采样概率；第一确定单元 53，配置为基于当前词汇对应的出现频率和剩余采样概率，确定当前采样概率；第二确定单元 54，配置为根据当前词汇在剩余采样个数和当前采样概率条件下的二项分布，确定当前词汇被采样次数；添加单元 55，配置为将所述当前词汇按照被采样次数添加到负例集中；更新单元 56，配置为根据当前词汇被采样次数更新
15 剩余采样个数，根据当前词汇对应的出现频率更新剩余采样概率，用于对词频表中的其他备选词汇进行采样，直到检测到预定条件得到满足。

[93]第一获取单元 51 首先可以从词频表的多个备选词汇的未采样词汇集合中，获取一个备选词汇作为当前词汇，并获取该当前词汇对应的出现频率。其中，该出现频率可以是该当前词汇在训练语料中的出现频率。

20 [94]第二获取单元 52，配置为获取针对未采样词汇集合确定的剩余采样个数和剩余采样概率。剩余采样个数可以是负例集中还需要的负例个数。换句话说，就是未采样词汇在生成负例集的负例采样过程中的总采样次数。剩余采样概率，可以是未采样词汇在生成负例集的负例采样过程中的总采样概率。剩余采样概率为 r 的初始值一般为 1。

[95]第一确定单元 53 可以基于当前词汇对应的出现频率和剩余采样概率，确定当前词汇
25 对应的当前采样概率。当前采样概率，可以是当前词汇在未采样词汇集合中的采样概率。在一个可选的实施例中，当前采样概率可以是当前词汇对应的出现频率和剩余采样概率的比值。

[96]第二确定单元 54 可以根据当前词汇在剩余采样个数和当前采样概率条件下的二项分布，确定当前词汇被采样次数。二项分布是若干次独立的伯努利试验中成功的次数的
30 离散概率分布。具体到某个实施例，对于进行的待采样负例个数次试验，对于每次试验，

当前词汇被采样到的概率为当前采样概率。第二确定单元 54 的主要作用在于：确定在 s 次试验中，第 i 个词汇成功被采样的次数 b 。

5 [97]根据另一方面的实施例，第二确定单元 54 可以模拟执行剩余采样个数次的采样操作，在各次采样操作中，确保当前词汇被采样到的概率为当前采样概率。统计当前词汇被采样到的次数，并确定当前词汇的被采样次数为，当前词汇被采样到的次数。

[98]根据另一方面的实施例，第二确定单元 54 还可以从满足条件的数值中随机获取一个值，作为当前词汇被采样次数。这里的数值满足的条件可以为：与剩余采样个数的比值应与当前采样概率一致。

10 [99]添加单元 55，可以按照第二确定单元 54 确定的被采样次数将当前词汇添加到负例集中。被采样次数是多少，就添加多少个当前词汇到负例集。

[100] 更新单元 56，根据当前词汇被采样次数更新上述剩余采样个数，根据当前词汇对应的出现频率更新上述剩余采样概率。可以理解，对每个备选词汇进行采样后，剩余采样个数会减少相应数量，剩余采样概率也会发生改变。换句话说，对于下一个备选词汇而言，采样条件发生变化。在一些可能的设计中，更新单元 56 可以将剩余采样个数更新为，原剩余采样个数与当前词汇被采样次数的差。更新单元 56 将剩余采样概率更新为，原剩余采样概率与当前词汇对应的出现频率的差。

[101] 另一方面，由于负例集中需要的负例个数是有限的，因此，还可以预先设置预定条件，当满足该条件时，停止负例采样，否则继续针对词频表的其他备选词汇执行采样流程。该检测功能可以由更新单元 26 实现，也可以由一个独立的检测单元实现。由此，在一些实施例中，装置 500 还包括检测单元 57，配置为在更新单元 26 更新完剩余采样个数和剩余采样概率后，检测预定条件是否得到满足，在不满足预定条件的情况下，根据更新后的剩余采样个数和剩余采样概率对词频表中的其他备选词汇采样。这里，预定条件可以包括负例集中的负例总个数达到初始剩余采样个数，也可以包括更新后的剩余采样个数为 0，还可以包括未采样词汇集合为空。

25 [102] 在一些可能的设计中，装置 500 还可以包括：

[103] 输出模块（未示出），配置为在所述负例集中的负例个数满足预定条件的情况下，输出所述负例集。该负例集可以输出到本地，也可以输出到其他设备。在进一步的实施例中，装置 500 还可以包括选择单元（未示出），配置为针对训练语料中的训练词汇，可以从该负例集中选取负例。

[104] 根据一方面的实施例，负例集中的词汇可以和预定区间上的各个数值相对应，选择单元进一步可以包括：生成模块，配置为生成预定区间上的随机数，其中，所生成随机数取自前述各个数值；获取模块，配置为从负例集中获取与上述随机数相对应的负例。

5 [105] 在一些实现中，所获取的负例还可能与训练词汇或其上下文词汇一致，此时，该负例词汇将不是该训练词汇的负例。因此，获取模块进一步可以配置为：比较所获取的负例与训练词汇是否一致；在一致的情况下，由上述生成模块重新生成预定区间上的随机数。

10 [106] 根据一种可能的设计，装置 500 还可以包括：检测单元（未示出），配置为检测负例集的更新条件是否满足；以使装置 500 在更新条件满足的情况下，重新生成负例集，从而对负例集进行更新。

[107] 通过以上装置，一方面，可以生成预采样的负例集，由于负例集中是预采样的负例，使用时只需随机取出相应数量负例即可，而无需考虑词汇在词汇表中的出现频率，运算复杂度大大降低。另一方面，可以在预采样过程中，进行批量采样，对词频表中的
15 各个词汇只采样一次，而采样的数量可以是多个，从而减少负例采样的时间，进而能够快速有效地进行负例采样。总之，图 5 示出的装置 500 可以提高负例采样的而有效性。

[108] 根据另一方面的实施例，还提供一种计算机可读存储介质，其上存储有计算机程序，当所述计算机程序在计算机中执行时，令计算机执行结合图 2 所描述的方法。

20 [109] 根据再一方面的实施例，还提供一种计算设备，包括存储器和处理器，所述存储器中存储有可执行代码，所述处理器执行所述可执行代码时，实现结合图 2 所述的方法。

[110] 本领域技术人员应该可以意识到，在上述一个或多个示例中，本发明所描述的功能可以用硬件、软件、固件或它们的任意组合来实现。当使用软件实现时，可以将这些功能存储在计算机可读介质中或者作为计算机可读介质上的一个或多个指令或代码
25 进行传输。

[111] 以上所述的具体实施方式，对本发明的目的、技术方案和有益效果进行了进一步详细说明，所应理解的是，以上所述仅为本发明的具体实施方式而已，并不用于限定本发明的保护范围，凡在本发明的技术方案的基础之上，所做的任何修改、等同替换、改进等，均应包括在本发明的保护范围之内。

权利要求书

1.一种针对训练语料从词频表中进行负例采样的方法，所述词频表包括多个备选词汇和各个备选词汇在所述训练语料中的出现频率，所述方法包括：

从所述多个备选词汇中的未采样词汇集合中获取当前词汇，以及当前词汇对应的出现频率 p_i ；

获取针对所述未采样词汇集合确定的剩余采样个数 s 和剩余采样概率 r ；

基于所述当前词汇对应的出现频率 p_i 和所述剩余采样概率 r ，确定所述当前词汇对应的当前采样概率 P ；

根据所述当前词汇在所述剩余采样个数 s 和所述当前采样概率 P 条件下的二项分布，确定所述当前词汇被采样次数 b ；

将所述当前词汇按照所述被采样次数 b 添加到负例集中；

根据所述被采样次数 b 更新所述剩余采样个数 s ，并根据所述当前词汇对应的出现频率 p_i 更新所述剩余采样概率 r ，用于对所述词频表中的其他备选词汇进行采样，直到检测到预定条件得到满足。

2. 根据权利要求 1 所述的方法，其中，基于所述当前词汇对应的出现频率 p_i 和所述剩余采样概率 r ，确定当前采样概率 P 包括：

将所述当前采样概率 P 确定为，所述当前词汇对应的出现频率 p_i 与所述剩余采样概率 r 的比值。

3. 根据权利要求 1 所述的方法，其中，确定所述当前词汇被采样次数 b 包括：

模拟执行所述剩余采样个数 s 次的采样操作，其中，在各次采样操作中，所述当前词汇被采样到的概率为所述当前采样概率 P ；

确定所述被采样次数 b 为，所述当前词汇在所述剩余采样个数 s 次的采样操作中被采样到的次数。

4. 根据权利要求 1 所述的方法，其中，根据所述当前词汇被采样次数 b 更新所述剩余采样个数 s 包括：

将剩余采样个数 s 更新为，所述剩余采样个数 s 与所述被采样次数 b 的差。

5. 根据权利要求 1 所述的方法，其中，所述预定条件包括：

所述负例集中的负例个数达到预设数目；或者

更新后的剩余采样个数 s 为零；或者

所述未采样词汇集合为空。

6. 根据权利要求 1 所述的方法，其中，所述根据所述当前词汇对应的出现频率 p_i

更新所述剩余采样概率 r 包括:

将剩余采样概率 r 更新为, 所述剩余采样概率 r 与所述当前词汇对应的出现频率 p_i 的差。

7. 根据权利要求 1 所述的方法, 其中, 所述方法还包括:

5 在所述预定条件得到满足的情况下, 输出所述负例集。

8. 根据权利要求 7 所述的方法, 其中, 所述方法还包括:

针对训练语料中的训练词汇, 从所述负例集中选取负例。

9. 根据权利要求 8 所述的方法, 其中从所述负例集中选取负例包括:

10 生成预定区间上的随机数, 其中, 所述预定区间包括多个数值, 各个数值分别与所述负例集中的各个负例相对应, 所述随机数取自所述多个数值;

从所述负例集中获取与所述随机数相对应的负例。

10. 根据权利要求 9 所述的方法, 其中, 所述从所述负例集中获取与所述随机数相对应的负例包括:

比较所获取的负例与所述训练词汇是否一致;

15 在一致的情况下, 重新执行所述生成预定区间上的随机数的步骤。

11. 根据权利要求 1-10 中任一所述的方法, 其中, 所述方法还包括:

检测所述负例集的更新条件是否满足;

在所述更新条件满足的情况下, 重新生成负例集。

20 12. 一种针对训练语料从词频表中进行负例采样的装置, 所述词频表包括多个备选词汇和各个备选词汇在所述训练语料中的出现频率, 所述装置包括:

第一获取单元, 配置为从所述多个备选词汇中的未采样词汇集合中获取当前词汇, 以及当前词汇对应的出现频率 p_i ;

第二获取单元, 配置为获取针对所述未采样词汇集合确定的剩余采样个数 s 和剩余采样概率 r ;

25 第一确定单元, 配置为基于所述当前词汇对应的出现频率 p_i 和所述剩余采样概率 r , 确定所述当前词汇对应的当前采样概率 P ;

第二确定单元, 配置为根据所述当前词汇在所述剩余采样个数 s 和所述当前采样概率 P 条件下的二项分布, 确定所述当前词汇被采样次数 b ;

添加单元, 配置为将所述当前词汇按照所述被采样次数 b 添加到负例集中;

30 更新单元, 配置为根据所述当前词汇被采样次数 b 更新所述剩余采样个数 s , 根据所述当前词汇对应的出现频率 p_i 更新所述剩余采样概率 r , 用于对所述词频表中的其他

备选词汇进行采样，直到检测到预定条件得到满足。

13. 根据权利要求 12 所述的装置，其中，所述第一确定单元进一步配置为：

将所述当前采样概率 P 确定为，所述当前词汇对应的出现频率 p_i 与剩余采样概率 r 的比值。

5 14. 根据权利要求 12 所述的装置，其中，所述第二确定单元包括：

试验模块，配置为模拟执行所述剩余采样个数 s 次的采样操作，其中，在各次采样操作中，所述当前词汇被采样到的概率为所述当前采样概率 P ；

确定模块，配置为确定所述被采样次数 b 为，所述当前词汇在所述剩余采样个数 s 次的采样操作中被采样到的次数。

10 15. 根据权利要求 12 所述的装置，其中，所述更新单元进一步配置为：

将剩余采样个数 s 更新为，所述剩余采样个数 s 与所述被采样次数 b 的差。

16. 根据权利要求 11 所述的装置，其中，所述预定条件包括：

所述负例集中的负例个数达到预设数目；或者

更新后的剩余采样个数为零；或者

15 所述未采样词汇集合为空。

17. 根据权利要求 12 所述的装置，其中，所述更新单元还进一步配置为：

将剩余采样概率 r 更新为，所述剩余采样概率 r 与所述当前词汇对应的出现频率 p_i 的差。

18. 根据权利要求 12 所述的装置，其中，所述装置还包括：

20 输出模块，配置为在预定条件得到满足的情况下，输出所述负例集。

19. 根据权利要求 18 所述的装置，其中，所述装置还包括：

选择单元，配置为针对训练语料中的训练词汇，从所述负例集中选取负例。

20. 根据权利要求 19 所述的装置，其中所述选择单元包括：

生成模块，配置为生成预定区间上的随机数，其中，所述预定区间包括多个数值，

25 各个数值分别与所述负例集中的各个负例相对应，所述随机数取自所述多个数值；

获取模块，配置为从所述负例集中获取与所述随机数相对应的负例。

21. 根据权利要求 20 所述的装置，其中，所述获取模块进一步配置为：

比较所获取的负例与所述训练词汇是否一致；

在一致的情况下，重新通过所述生成模块执行所述生成预定区间上的随机数的步骤。

30 22. 根据权利要求 12-21 中任一所述的装置，其中，所述装置还包括：

检测单元，配置为检测负例集的更新条件是否满足；

以使所述装置在所述更新条件满足的情况下，重新生成负例集。

23.一种计算机可读存储介质，其上存储有计算机程序，当所述计算机程序在计算机中执行时，令计算机执行权利要求 1-11 中任一项的所述的方法。

24. 一种计算设备，包括存储器和处理器，其特征在于，所述存储器中存储有可执行代码，所述处理器执行所述可执行代码时，实现权利要求 1-11 中任一项所述的方法。

5

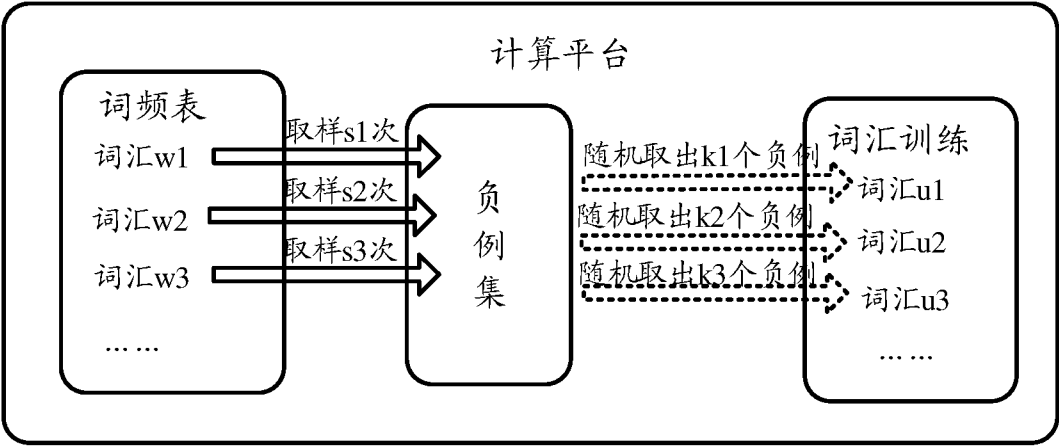


图 1

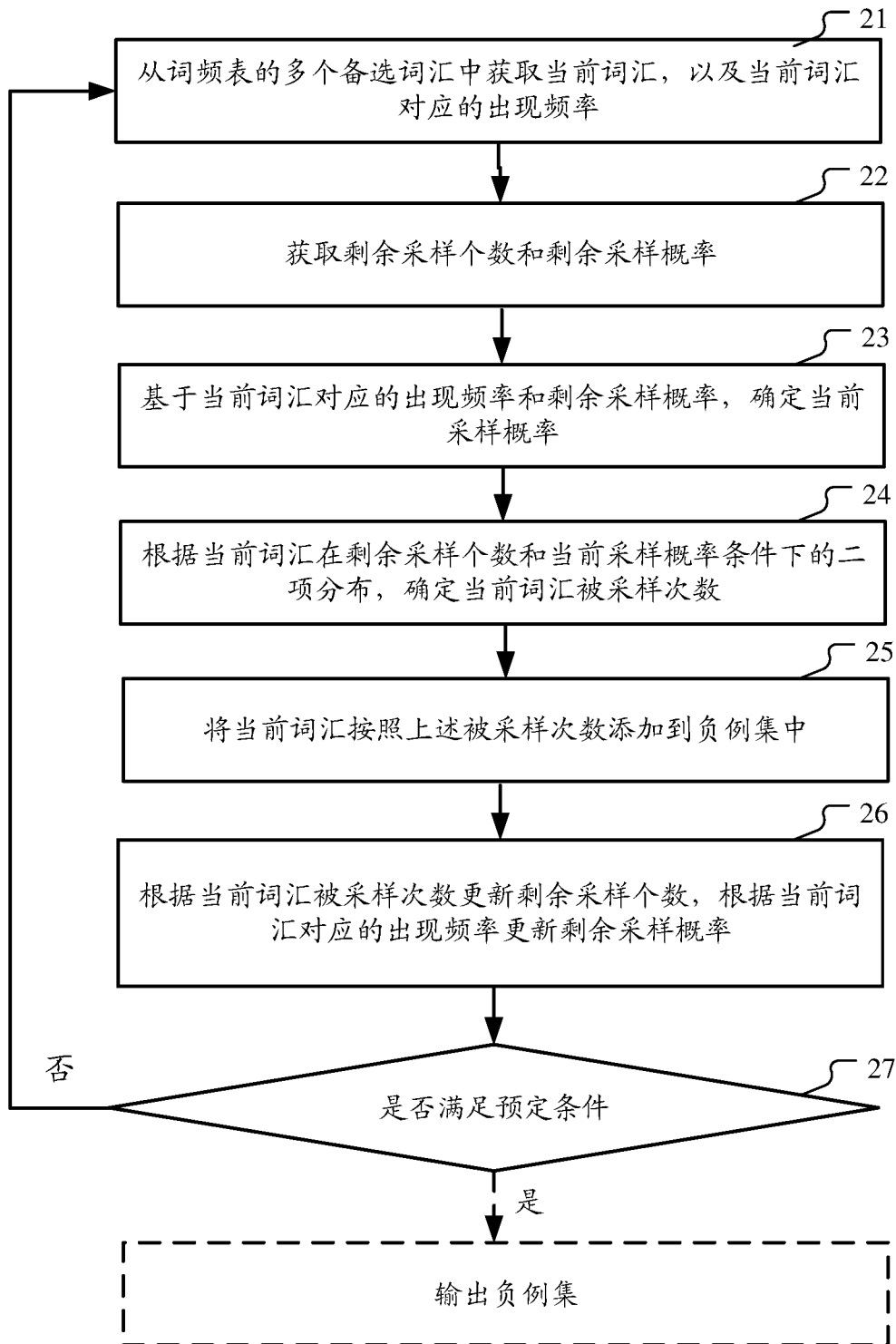


图 2

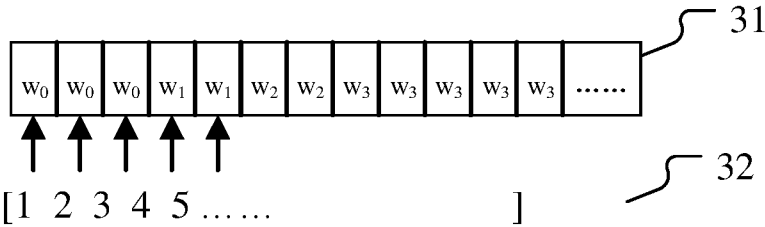


图 3

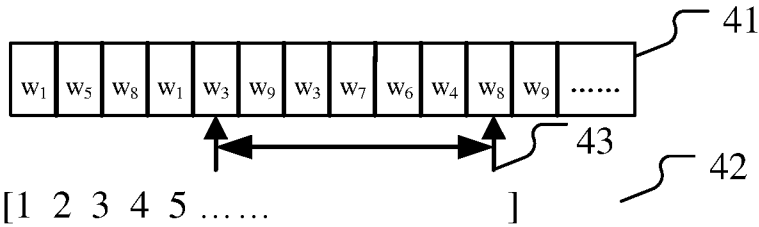


图 4

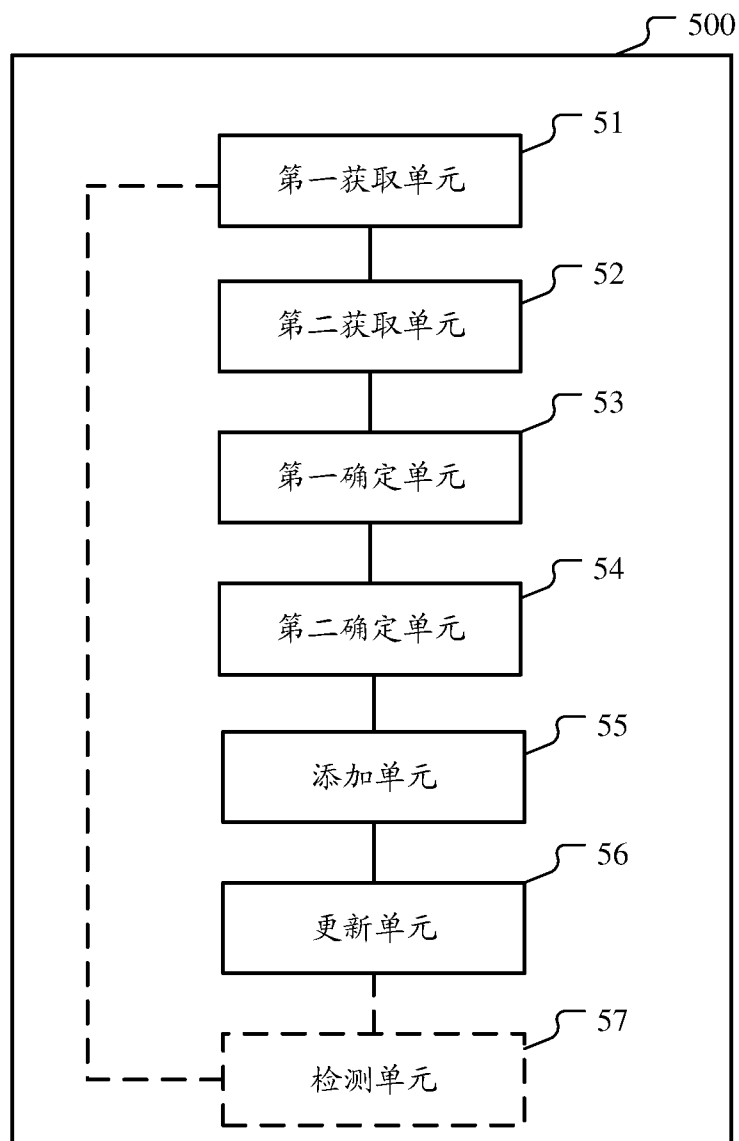


图 5

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2019/077438

A. CLASSIFICATION OF SUBJECT MATTER

G06K 9/62(2006.01)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)

G06K

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)

CNABS; SIPOABS; DWPI; CNTXT; CNKI; IEEE: 训练, 语料, 词频表, 负例, 采样, 词汇, 频率, 次数, 集合, 概率, 剩余, 条件, 二项分布, training, corpus, word frequency table, negative example, sampling, vocabulary, frequency, number of times, set, probability, residual, condition, binomial distribution

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 108875810 A (ALIBABA GROUP HOLDING LIMITED) 23 November 2018 (2018-11-23) claims 1-24, and figures 1-5	1-24
A	CN 108021934 A (ALIBABA GROUP HOLDING LIMITED) 11 May 2018 (2018-05-11) entire document	1-24
A	CN 107220233 A (BEIJING INSTITUTE OF TECHNOLOGY) 29 September 2017 (2017-09-29) entire document	1-24
A	CN 106257441 A (UNIVERSITY OF ELECTRONIC SCIENCE AND TECHNOLOGY OF CHINA) 28 December 2016 (2016-12-28) entire document	1-24
A	WO 2004095421 A1 (INTEL CORPORATION) 04 November 2004 (2004-11-04) entire document	1-24



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

14 May 2019

Date of mailing of the international search report

21 May 2019

Name and mailing address of the ISA/CN

State Intellectual Property Office of the P. R. China (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/CN2019/077438

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	108875810	A	23 November 2018	None			
CN	108021934	A	11 May 2018	None			
CN	107220233	A	29 September 2017	None			
CN	106257441	A	28 December 2016	CN	106257441	B	15 March 2019
WO	2004095421	A1	04 November 2004	CN	1761997	A	19 April 2006
				US	2004186718	A1	23 September 2004
				US	7454342	B2	18 November 2008
				GB	2411507	A	31 August 2005

国际检索报告

国际申请号

PCT/CN2019/077438

A. 主题的分类 G06K 9/62 (2006.01) 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) G06K 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)) CNABS; SIPOABS; DWPI; CNTXT; CNKI; IEEE: 训练, 语料, 词频表, 负例, 采样, 词汇, 频率, 次数, 集合, 概率, 剩余, 条件, 二项分布, training, corpus, word frequency table, negative example, sampling, vocabulary, frequency, number of times, set, probability, residual, condition, binomial distribution																				
C. 相关文件 <table border="1"> <thead> <tr> <th>类 型*</th> <th>引用文件, 必要时, 指明相关段落</th> <th>相关的权利要求</th> </tr> </thead> <tbody> <tr> <td>PX</td> <td>CN 108875810 A (阿里巴巴集团控股有限公司) 2018年 11月 23日 (2018 - 11 - 23) 权利要求1-24、图1-5</td> <td>1-24</td> </tr> <tr> <td>A</td> <td>CN 108021934 A (阿里巴巴集团控股有限公司) 2018年 5月 11日 (2018 - 05 - 11) 全文</td> <td>1-24</td> </tr> <tr> <td>A</td> <td>CN 107220233 A (北京理工大学) 2017年 9月 29日 (2017 - 09 - 29) 全文</td> <td>1-24</td> </tr> <tr> <td>A</td> <td>CN 106257441 A (电子科技大学) 2016年 12月 28日 (2016 - 12 - 28) 全文</td> <td>1-24</td> </tr> <tr> <td>A</td> <td>W0 2004095421 A1 (英特尔公司) 2004年 11月 4日 (2004 - 11 - 04) 全文</td> <td>1-24</td> </tr> </tbody> </table>			类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求	PX	CN 108875810 A (阿里巴巴集团控股有限公司) 2018年 11月 23日 (2018 - 11 - 23) 权利要求1-24、图1-5	1-24	A	CN 108021934 A (阿里巴巴集团控股有限公司) 2018年 5月 11日 (2018 - 05 - 11) 全文	1-24	A	CN 107220233 A (北京理工大学) 2017年 9月 29日 (2017 - 09 - 29) 全文	1-24	A	CN 106257441 A (电子科技大学) 2016年 12月 28日 (2016 - 12 - 28) 全文	1-24	A	W0 2004095421 A1 (英特尔公司) 2004年 11月 4日 (2004 - 11 - 04) 全文	1-24
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求																		
PX	CN 108875810 A (阿里巴巴集团控股有限公司) 2018年 11月 23日 (2018 - 11 - 23) 权利要求1-24、图1-5	1-24																		
A	CN 108021934 A (阿里巴巴集团控股有限公司) 2018年 5月 11日 (2018 - 05 - 11) 全文	1-24																		
A	CN 107220233 A (北京理工大学) 2017年 9月 29日 (2017 - 09 - 29) 全文	1-24																		
A	CN 106257441 A (电子科技大学) 2016年 12月 28日 (2016 - 12 - 28) 全文	1-24																		
A	W0 2004095421 A1 (英特尔公司) 2004年 11月 4日 (2004 - 11 - 04) 全文	1-24																		
☐ 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 2019年 5月 14日		国际检索报告邮寄日期 2019年 5月 21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 李昕宇 电话号码 (86-10)62089924																		

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2019/077438

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	108875810	A	2018年 11月 23日	无	
CN	108021934	A	2018年 5月 11日	无	
CN	107220233	A	2017年 9月 29日	无	
CN	106257441	A	2016年 12月 28日	CN	106257441 B 2019年 3月 15日
WO	2004095421	A1	2004年 11月 4日	CN	1761997 A 2006年 4月 19日
				US	2004186718 A1 2004年 9月 23日
				US	7454342 B2 2008年 11月 18日
				GB	2411507 A 2005年 8月 31日