

(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
2017 年 12 月 7 日 (07.12.2017)



(10) 国际公布号
WO 2017/206492 A1

(51) 国际专利分类号:
G06F 17/30 (2006.01)

(21) 国际申请号: PCT/CN2016/111571

(22) 国际申请日: 2016 年 12 月 22 日 (22.12.2016)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201610379719.X 2016年5月31日 (31.05.2016) CN

(71) 申请人: 北京百度网讯科技有限公司 (BEIJING Baidu NETCOM SCIENCE AND TECHNOLOGY CO., LTD.) [CN/CN]; 中国北京市海淀区上地十街 10 号百度大厦 2 层, Beijing 100085 (CN)。

(72) 发明人: 周坤胜(ZHOU, Kunsheng); 中国北京市海淀区上地十街 10 号百度大厦 2 层, Beijing 100085 (CN)。 何径舟(HE, Jingzhou); 中国北京市海淀区上地十街 10 号百度大厦 2 层, Beijing 100085 (CN)。 石磊(SHI, Lei); 中国北京市海淀区上地十街 10 号百度大厦 2 层, Beijing 100085 (CN)。 冯仕堃(FENG, Shikun); 中国北京市海淀区上地十街 10 号百度大厦 2 层, Beijing 100085 (CN)。

(74) 代理人: 北京清亦华知识产权代理事务所(普通合伙) (TSINGYIHUA INTELLECTUAL PROPERTY LLC); 中国北京市海淀区清华园清华大学照澜院商业楼 301 室, Beijing 100084 (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,

(54) Title: BINARY FEATURE DICTIONARY CONSTRUCTION METHOD AND APPARATUS

(54) 发明名称: 二元特征词典的构建方法和装置

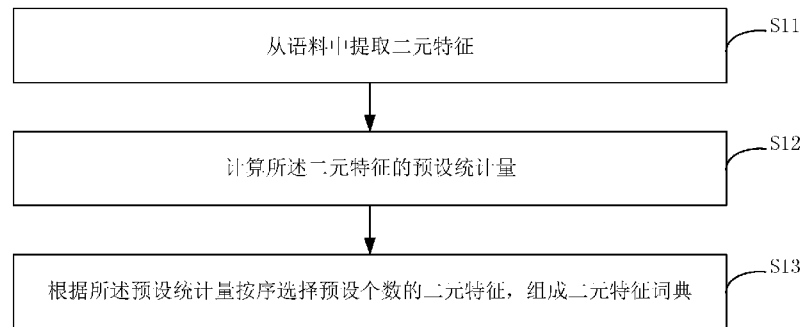


图 1

S11 Extracting binary features from language materials
S12 Calculating a pre-set statistical magnitude of the binary features
S13 Selecting, according to the pre-set statistical magnitude, a pre-set number of binary features in sequence, so as to construct a binary feature dictionary

(57) Abstract: Disclosed are a binary feature dictionary construction method and apparatus. The binary feature dictionary construction method comprises: extracting binary features from language materials (S11); calculating a pre-set statistical magnitude of the binary features (S12); and selecting, according to the pre-set statistical magnitude, a pre-set number of binary features in sequence, so as to construct a binary feature dictionary (S13). By using the method, the binary feature dictionary can be constructed, whereby new features can be introduced to training data in a semantic similarity model.

(57) 摘要: 一种二元特征词典的构建方法和装置, 该二元特征词典的构建方法包括: 从语料中提取二元特征 (S11); 计算所述二元特征的预设统计量 (S12); 根据所述预设统计量按序选择预设个数的二元特征, 组成二元特征词典 (S13)。该方法能够构建出二元特征词典, 从而可以在语义相似度模型的训练数据中引入新的特征。

GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, 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))。

二元特征词典的构建方法和装置

相关申请的交叉引用

本申请要求北京百度网讯科技有限公司于 2016 年 5 月 31 日提交的、发明名称为“二元特征词典的构建方法和装置”的、中国专利申请号“201610379719.X”的优先权。

技术领域

本申请涉及自然语言处理技术领域，尤其涉及一种二元特征词典的构建方法和装置。

背景技术

搜索引擎是一种检索系统，能够向用户反馈检索结果。搜索引擎的一种实现是基于语义相似度模型，通过语义相似度模型可以计算用户检索的检索词（query）与候选的标题（title）之间的相似度，再根据相似度进行排序，以返回检索结果。语义相似度模型可以采用神经网络训练构建。

相关技术中，采用神经网络构建语义相似度模型时，训练数据通常是基础（basic）粒度的信息或短语（phrase）粒度的信息，其中，basic 粒度的信息是对句子语料进行切词后，得到的最小粒度的分词，如“百度”，phrase 粒度的信息

包括多个最小粒度的分词，如“巴西葡语”。为了提高语义相似度模型的精准度，还可以在训练数据中引入新的粒度信息。当需要引入新的粒度信息时，需要解决如何提取新的粒度信息的问题。

发明内容

本申请旨在至少在一定程度上解决相关技术中的技术问题之一。

为此，本申请的一个目的在于提出一种二元特征词典的构建方法，该方法可以构建出二元特征词典，从而可以采用该二元特征词典从模型语料中提取出二元特征，进一步的，该二元特征可以作为语义相似度模型的训练数据中引入的新特征。

本申请的另一个目的在于提出一种二元特征词典的构建装置。

为达到上述目的，本申请第一方面实施例提出的二元特征词典的构建方法，包括：从语料中提取二元特征；计算所述二元特征的预设统计量；根据所述预设统计量按序选择预设个数的二元特征，组成二元特征词典。

本申请第一方面实施例提出的二元特征词典的构建方法，通过计算二元特征的预设统计量，根据预设统计量选择二元特征组成二元特征词典，可以实现二元特征词典的构建，从而可以采用该二元特征词典从模型语料中提取出二元特征，进一步的，该二元特征可以作为语义相似度模型的训练数据中引入的新特征。

为达到上述目的，本申请第二方面实施例提出的二元特征词典的构建装置，

包括：提取模块，用于从语料中提取二元特征；计算模块，用于计算所述二元特征的预设统计量；生成模块，用于根据所述预设统计量按序选择预设个数的二元特征，组成二元特征词典。

本申请第二方面实施例提出的二元特征词典的构建装置，通过计算二元特征的预设统计量，根据预设统计量选择二元特征组成二元特征词典，可以实现二元特征词典的构建，从而可以采用该二元特征词典从模型语料中提取出二元特征，进一步的，该二元特征可以作为语义相似度模型的训练数据中引入的新特征。

本申请实施例提出一种终端，包括：处理器；用于存储处理器可执行指令的存储器；其中，所述处理器被配置为：从语料中提取二元特征；计算所述二元特征的预设统计量；根据所述预设统计量按序选择预设个数的二元特征，组成二元特征词典。

本申请实施例提出一种非临时性计算机可读存储介质，当所述存储介质中的指令由终端的处理器执行时，使得终端能够执行一种方法，所述方法包括：从语料中提取二元特征；计算所述二元特征的预设统计量；根据所述预设统计量按序选择预设个数的二元特征，组成二元特征词典。

本申请实施例提出一种计算机程序产品，当所述计算机程序产品中的指令被处理器执行时，执行一种方法，所述方法包括：从语料中提取二元特征；计算所述二元特征的预设统计量；根据所述预设统计量按序选择预设个数的二元

特征，组成二元特征词典。

本申请附加的方面和优点将在下面的描述中部分给出，部分将从下面的描述中变得明显，或通过本申请的实践了解到。

附图说明

本申请上述的和/或附加的方面和优点从下面结合附图对实施例的描述中将变得明显和容易理解，其中：

图 1 是本申请一个实施例提出的二元特征词典的构建方法的流程示意图；

图 2 是本申请另一个实施例提出的二元特征词典的构建方法的流程示意图；

图 3 是本申请另一个实施例提出的二元特征词典的构建方法的流程示意图；

图 4 是本申请实施例采用的神经网络的示意图；

图 5 是本发明一个实施例提出的二元特征词典的构建装置的结构示意图；

图 6 是本发明一个实施例提出的二元特征词典的构建装置的结构示意图。

具体实施方式

下面详细描述本申请的实施例，所述实施例的示例在附图中示出，其中自始至终相同或类似的标号表示相同或类似的模块或具有相同或类似功能的模块。

下面通过参考附图描述的实施例是示例性的，仅用于解释本申请，而不能理解为对本申请的限制。相反，本申请的实施例包括落入所附加权利要求书的精神和内涵范围内的所有变化、修改和等同物。

在获取语义相似模型的训练数据时，可以先对句子语料进行切词，之后根据分词获取需要粒度的信息，如 basic 粒度的信息或 phrase 粒度的信息。

如上所示，在构建语义相似度模型时，除了上述的 basic 粒度的信息或 phrase 粒度的信息之外，还可以引入新粒度的信息。该引入的新粒度的信息可以称为二元特征。每个二元特征由两个 basic 粒度的信息或两个 phrase 粒度的信息组成。

当需要引入二元特征时，可以根据二元特征词典获取对应的二元特征。因此，需要解决二元特征词典如何建立的问题。

图 1 是本申请一个实施例提出的二元特征词典的构建方法的流程示意图。

如图 1 所示，本实施例的流程包括：

S11：从语料中提取二元特征。

其中，在构建二元特征词典时，可以先收集语料，与语义相似度模型的构建类似，构建二元特征词典时收集的语料也可以具体是句子语料，即，收集大量的句子作为语料。

在语料中提取二元特征时，可以具体是将语料中任意两个相邻的词条，作为一个二元特征。

具体的，对应每个句子语料，可以对句子语料进行切分，得到词条(term)。

其中，词条可以是 basic 粒度，或者，也可以是 phrase 粒度，但是，上述的两个词条需要具有同种粒度，例如，均是 basic 粒度，或者，均是 phrase

粒度。将句子切分成 basic 粒度的词条或 phrase 粒度的词条可以采用各种已有技术，在此不再详细说明。

在将句子切分成词条后，可以将任意两个相邻的词条作为一个二元特征，因此，可以对应每个句子，假设该句子包括 n 个词条，则可以提取出 $n-1$ 个二元特征。将语料中的所有句子对应的二元特征相加后，可以得到多个二元特征。

S12：计算所述二元特征的预设统计量。

本实施例中，以预设统计量是 t 统计量为例。

t 统计量是对模型参数的单个假设进行检验的一种统计量。

具体的，一个二元特征的 t 统计量的计算公式是：

$$t = \frac{f - \mu}{\sqrt{s^2 / N}}$$

其中， f 是该二元特征在语料中的出现概率，例如，语料中共有 M 个句子，其中 m 个句子中存在该二元特征，即该二元特征中的两个词条相邻且位置关系不变的出现在 m 个句子中，则 $f = m/M$ ；

$\mu = p_1 \times p_2$ ， p_1 是该二元特征中一个词条在语料中的出现概率， p_2 是该二元特征中另一个词条在语料中的出现概率；

本实施例中，假设 $s^2 = f$ ；

N 是语料中所有二元特征的总数， $N = \sum N_i$ ， N 是单个句子语料对应的二元特征的数量，假设某个句子语料包括 n 个词条，则该句子语料对应的二元特征的数量是 $n-1$ 。

因此，通过上述计算公式，可以计算出每个二元特征的 t 统计量。

S13：根据所述预设统计量按序选择预设个数的二元特征，组成二元特征词典。

例如，在得到每个二元特征的 t 统计量后，可以根据 t 统计量从大到小的顺序选择预设个数的二元特征，将选择的这些二元特征组成二元特征词典。

进一步的，如上所示，一个二元特征可以是由两个 basic 粒度的词条组成，或者，一个二元特征也可以是由两个 phrase 粒度的词条组成。在上述排序时，是对这些二元特征（包括根据 basic 粒度的词条得到的二元特征以及根据 phrase 粒度的词条得到的二元特征）进行统一排序的，从而在二元特征词典中可以包括根据 basic 粒度的词条得到的二元特征和/或根据 phrase 粒度的词条得到的二元特征。

通过上述的流程，可以实现二元特征词典的构建。

在构建出二元特征词典后，可以使用该词典生成二元特征，以用于语义相似度模型的训练。

相应的，一些实施例中，参见图 2，本实施例的方法还可以包括：

S14：采用所述二元特征词典，从语义相似度模型的分词中提取二元特征，作为语义相似度模型的训练数据。

其中，可以将二元特征词典中包括的且可以由相邻的两个分词组成的二元特征，作为提取的二元特征。

另外，上述的分词可以是 basic 粒度的分词或 phrase 粒度的分词。

例如，语义相似度模块的一个语料是“百度巴西葡语”，经过切词后，以 basic 粒度的分词为例，各分词是“百度”、“巴西”、“葡语”，构建的二元特征词典中的一个二元特征是“巴西葡语”，则提取的一个二元特征是“巴西葡语”。

进一步的，如何二元特征词典中不包括任意两个相邻分词组成的二元特征，则直接将分词作为提取的二元特征。依然以上述的语料为例，假设构建的二元特征词典中不包括“巴西葡语”，也不包括“百度巴西”，则提取的二元特征就是原始的分词，如“百度”、“巴西”、“葡语”。

因此，通过二元特征词典，可以实现二元特征的提取，从而在训练语义相似度模型时可以引入新的特征。

一些实施例中，参见图 3，本实施例的方法还可以包括：

S15：根据所述训练数据进行神经网络训练，生成语义相似度模型。

其中，神经网络可以如图 4 所示，与通常的语义相似度模型不同的是，本实施例的输入可以包括二元特征。

因此，可以通过引入二元特征训练语义相似度模型。

本实施例中，通过计算二元特征的预设统计量，根据预设统计量选择二元特征组成二元特征词典，可以实现二元特征词典的构建，从而可以采用该二元特征词典从模型语料中提取出二元特征，进一步的，该二元特征可以作为语义相似度模型的训练数据中引入的新特征。

图 5 是本发明一个实施例提出的二元特征词典的构建装置的结构示意图。

参见图 5，本实施例的装置包括：提取模块 51、计算模块 52 和生成模块 53。

提取模块 51，用于从语料中提取二元特征；

计算模块 52，用于计算所述二元特征的预设统计量；

生成模块 53，用于根据所述预设统计量按序选择预设个数的二元特征，组成二元特征词典。

一些实施例中，所述提取模块 51 具体用于：

将语料中任意两个相邻的词条，作为一个二元特征。

一些实施例中，所述计算模块 52 计算的所述预设统计量是 t 统计量。

一些实施例中，参见图 6，本实施例的装置还可以包括：

获取模块 54，用于采用所述二元特征词典，从语义相似度模型的分词中提取二元特征，作为语义相似度模型的训练数据。

一些实施例中，参见图 6，本实施例的装置还可以包括：

训练模块 55，用于根据所述训练数据进行神经网络训练，生成语义相似度模型。

可以理解的是，本实施例的装置与方法实施例对应，具体内容可以参见方法实施例中的相关描述，在此不再详细说明。

本实施例中，通过计算二元特征的预设统计量，根据预设统计量选择二元

特征组成二元特征词典，可以实现二元特征词典的构建，从而可以采用该二元特征词典从模型语料中提取出二元特征，进一步的，该二元特征可以作为语义相似度模型的训练数据中引入的新特征。

可以理解的是，上述各实施例中相同或相似部分可以相互参考，在一些实施例中未详细说明的内容可以参见其他实施例中相同或相似的内容。

本申请实施例提出一种终端，包括：处理器；用于存储处理器可执行指令的存储器；其中，所述处理器被配置为：从语料中提取二元特征；计算所述二元特征的预设统计量；根据所述预设统计量按序选择预设个数的二元特征，组成二元特征词典。

本申请实施例提出一种非临时性计算机可读存储介质，当所述存储介质中的指令由终端的处理器执行时，使得终端能够执行一种方法，所述方法包括：从语料中提取二元特征；计算所述二元特征的预设统计量；根据所述预设统计量按序选择预设个数的二元特征，组成二元特征词典。

本申请实施例提出一种计算机程序产品，当所述计算机程序产品中的指令被处理器执行时，执行一种方法，所述方法包括：从语料中提取二元特征；计算所述二元特征的预设统计量；根据所述预设统计量按序选择预设个数的二元特征，组成二元特征词典。

需要说明的是，在本申请的描述中，术语“第一”、“第二”等仅用于描述目的，而不能理解为指示或暗示相对重要性。此外，在本申请的描述中，除

非另有说明，“多个”的含义是指至少两个。

流程图中或在此以其他方式描述的任何过程或方法描述可以被理解为，表示包括一个或更多个用于实现特定逻辑功能或过程的步骤的可执行指令的代码的模块、片段或部分，并且本申请的优选实施方式的范围包括另外的实现，其中可以不按所示出或讨论的顺序，包括根据所涉及的功能按基本同时的方式或按相反的顺序，来执行功能，这应被本申请的实施例所属技术领域的技术人员所理解。

应当理解，本申请的各部分可以用硬件、软件、固件或它们的组合来实现。在上述实施方式中，多个步骤或方法可以用存储在存储器中且由合适的指令执行系统执行的软件或固件来实现。例如，如果用硬件来实现，和在另一实施方式中一样，可用本领域公知的下列技术中的任一项或他们的组合来实现：具有用于对数据信号实现逻辑功能的逻辑门电路的离散逻辑电路，具有合适的组合逻辑门电路的专用集成电路，可编程门阵列(PGA)，现场可编程门阵列(FPGA)等。

本技术领域的普通技术人员可以理解实现上述实施例方法携带的全部或部分步骤是可以通程序来指令相关的硬件完成，所述的程序可以存储于一种计算机可读存储介质中，该程序在执行时，包括方法实施例的步骤之一或其组合。

此外，在本申请各个实施例中的各功能单元可以集成在一个处理模块中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个模块

中。上述集成的模块既可以采用硬件的形式实现，也可以采用软件功能模块的形式实现。所述集成的模块如果以软件功能模块的形式实现并作为独立的产品销售或使用时，也可以存储在一个计算机可读取存储介质中。

上述提到的存储介质可以是只读存储器，磁盘或光盘等。

在本说明书的描述中，参考术语“一个实施例”、“一些实施例”、“示例”、“具体示例”或“一些示例”等的描述意指结合该实施例或示例描述的具体特征、结构、材料或者特点包含于本申请的至少一个实施例或示例中。在本说明书中，对上述术语的示意性表述不一定指的是相同的实施例或示例。而且，描述的具体特征、结构、材料或者特点可以在任何的一个或多个实施例或示例中以合适的方式结合。

尽管上面已经示出和描述了本申请的实施例，可以理解的是，上述实施例是示例性的，不能理解为对本申请的限制，本领域的普通技术人员在本申请的范围内可以对上述实施例进行变化、修改、替换和变型。

权利要求书

1、一种二元特征词典的构建方法，其特征在于，包括：

从语料中提取二元特征；

计算所述二元特征的预设统计量；

根据所述预设统计量按序选择预设个数的二元特征，组成二元特征词典。

2、根据权利要求 1 所述的方法，其特征在于，所述从语料中提取二元特征，包括：

将语料中任意两个相邻的词条，作为一个二元特征。

3、根据权利要求 1-2 任一项所述的方法，其特征在于，所述预设统计量是 t 统计量。

4、根据权利要求 1-3 任一项所述的方法，其特征在于，还包括：

采用所述二元特征词典，从语义相似度模型的分词中提取二元特征，作为语义相似度模型的训练数据。

5、根据权利要求 4 所述的方法，其特征在于，还包括：

根据所述训练数据进行神经网络训练，生成语义相似度模型。

6、一种二元特征词典的构建装置，其特征在于，包括：

提取模块，用于从语料中提取二元特征；

计算模块，用于计算所述二元特征的预设统计量；

生成模块，用于根据所述预设统计量按序选择预设个数的二元特征，组成二元特征词典。

7、根据权利要求 6 所述的装置，其特征不在于，所述提取模块具体用于：
将语料中任意两个相邻的词条，作为一个二元特征。

8、根据权利要求 6-7 任一项所述的装置，其特征不在于，所述计算模块计算的所述预设统计量是 t 统计量。

9、根据权利要求 6-8 任一项所述的装置，其特征不在于，还包括：
获取模块，用于采用所述二元特征词典，从语义相似度模型的分词中提取二元特征，作为语义相似度模型的训练数据。

10、根据权利要求 9 所述的装置，其特征不在于，还包括：
训练模块，用于根据所述训练数据进行神经网络训练，生成语义相似度模型。

11、一种终端，其特征不在于，包括：处理器；用于存储处理器可执行指令的存储器；其中，所述处理器被配置为：

从语料中提取二元特征；

计算所述二元特征的预设统计量；

根据所述预设统计量按序选择预设个数的二元特征，组成二元特征词典。

12、一种非临时性计算机可读存储介质，其特征不在于，当所述存储介质中的指令由终端的处理器执行时，使得终端能够执行一种方法，所述方法包括：

从语料中提取二元特征；

计算所述二元特征的预设统计量；

根据所述预设统计量按序选择预设个数的二元特征，组成二元特征词典。

13、一种计算机程序产品，其特征在于，当所述计算机程序产品中的指令被处理器执行时，执行一种方法，所述方法包括：

从语料中提取二元特征；

计算所述二元特征的预设统计量；

根据所述预设统计量按序选择预设个数的二元特征，组成二元特征词典。

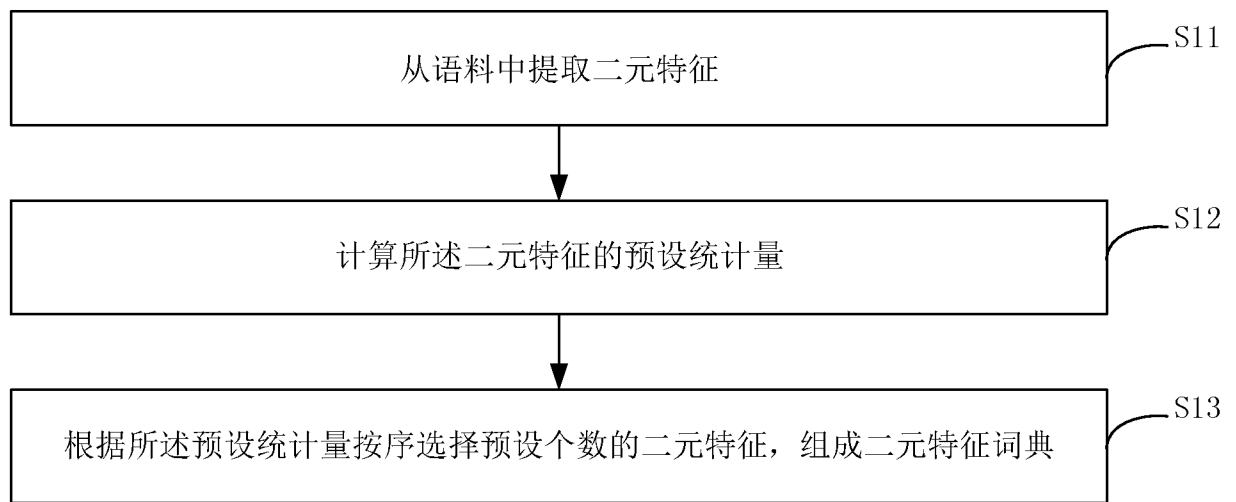


图 1

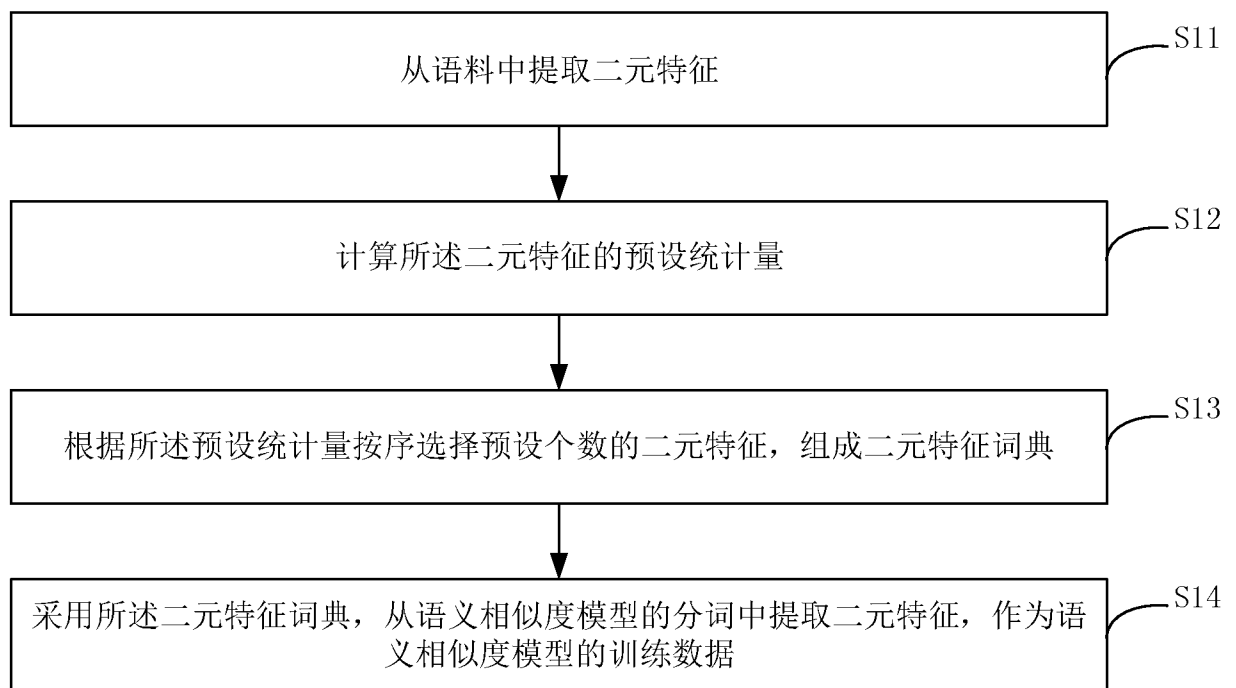


图 2

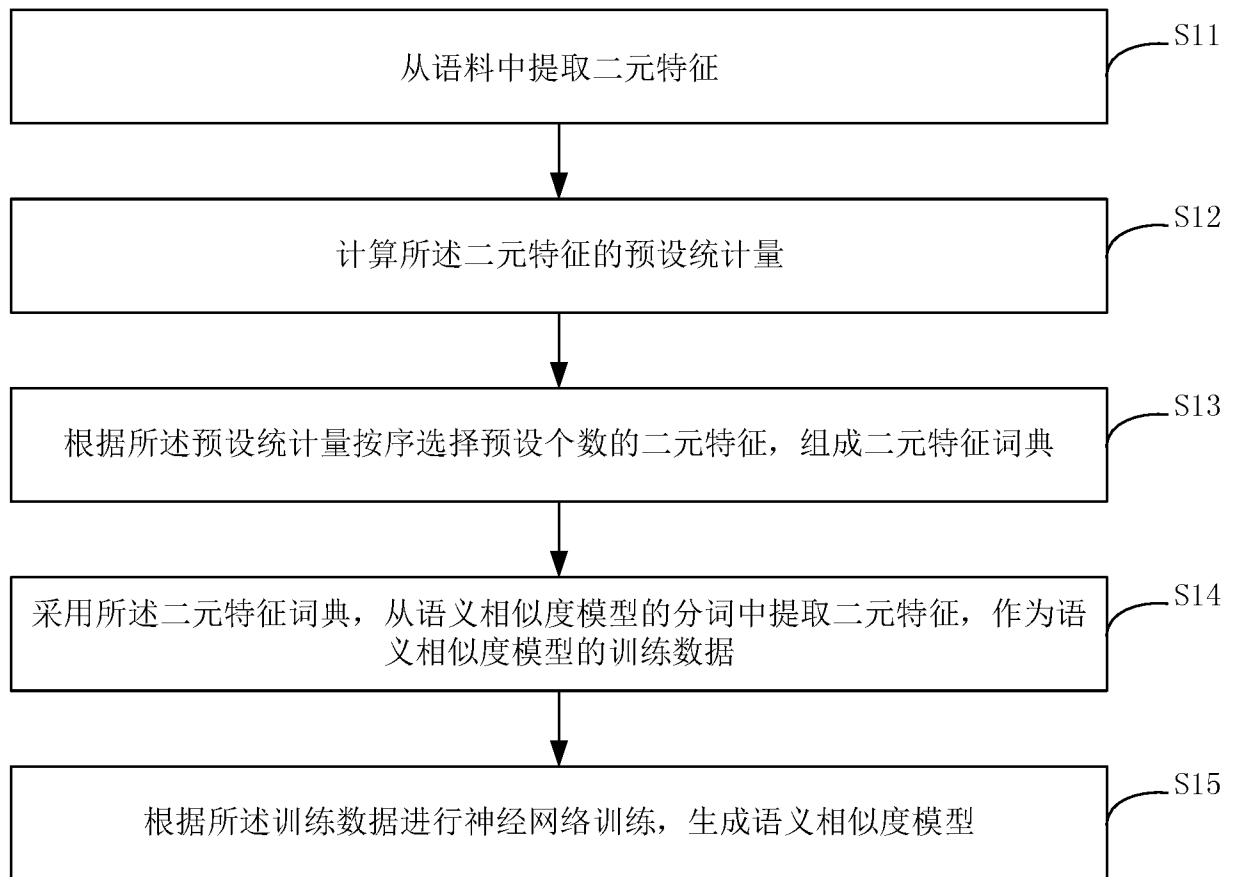


图 3

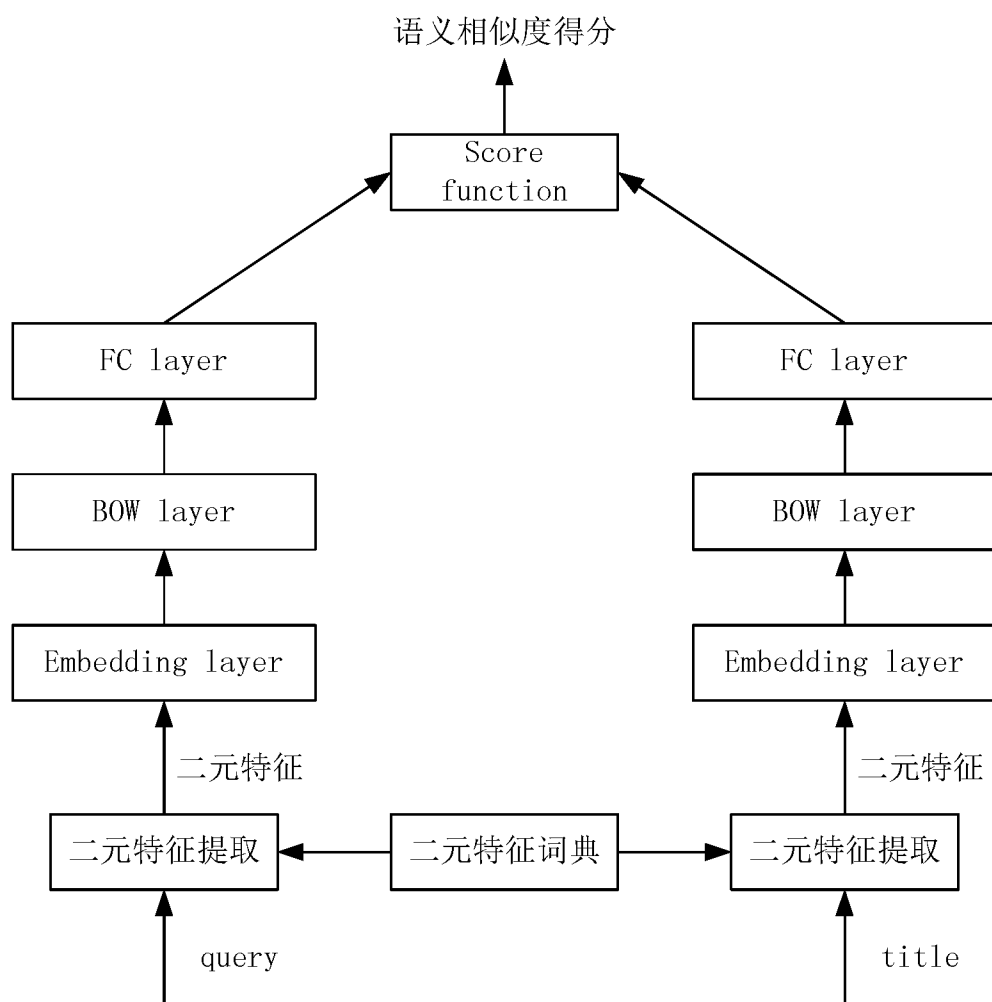


图 4

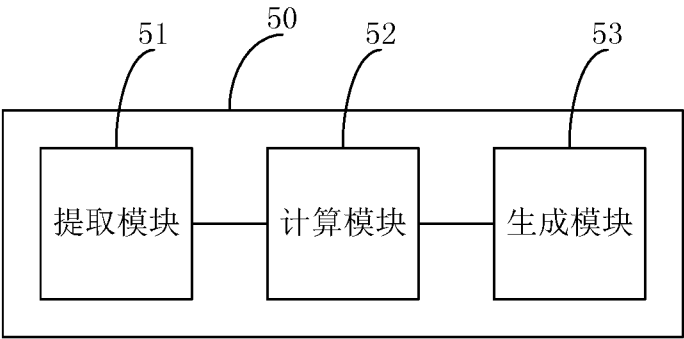


图 5

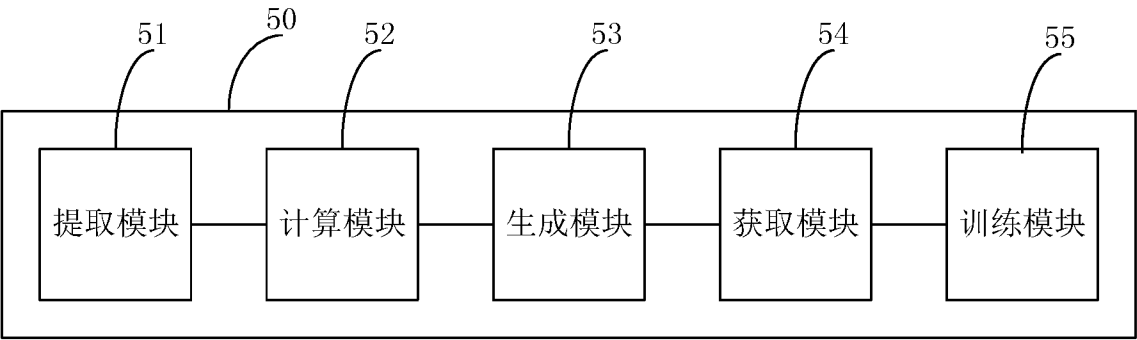


图 6

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2016/111571

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/30(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)

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, IEEE: word frequency, word segmentation, frequency, neural network, binary, feature, two, statistic, module, similar+, adjacent, train, neural, network

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 106021572 A (BEIJING BAIDU NETCOM SCIENCE AND TECHNOLOGY CO., LTD.), 12 October 2016 (12.10.2016), description, paragraphs [0025]-[0079]	1-13
Y	CN 104899190 A (BAIDU ONLINE NETWORK TECHNOLOGY (BEIJING) CO., LTD.), 09 September 2015 (09.09.2015), description, paragraphs [0035]-[0106]	1-13
Y	CN 104375989 A (STATE GRID CORPORATION OF CHINA et al.), 25 February 2015 (25.02.2015), description, paragraphs [0032]-[0062]	1-13
Y	CN 104391963 A (BEIJING ZHONGKE CHUANGYI TECHNOLOGY CO., LTD.), 04 March 2015 (04.03.2015), description, paragraphs [0031]-[0064]	1-13
A	US 2014289238 A1 (KABUSHIKI KAISHA TOSHIBA), 25 September 2014 (25.09.2014), the whole document	1-13

☐ Further documents are listed in the continuation of Box C. ☒ See patent family annex.

* Special categories of cited documents:	“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
“A” document defining the general state of the art which is not considered to be of particular relevance	“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
“E” earlier application or patent but published on or after the international filing date	“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
“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)	“&” document member of the same patent family
“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	

Date of the actual completion of the international search 24February 2017 (24.02.2017)	Date of mailing of the international search report 31March 2017 (31.03.2017)
Name and mailing address of the ISA/CN: State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No.: (86-10) 62019451	Authorized officer LIN, Shaohua Telephone No.:(86-10) 62414053

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2016/111571

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 106021572 A	12 October 2016	None	
CN 104899190 A	09 September 2015	None	
CN 104375989 A	25 February 2015	None	
CN 104391963 A	04 March 2015	None	
US 2014289238 A1	25 September 2014	CN 104077346 A	01 October 2014
		JP 2014186395 A	02 October 2014

A. 主题的分类 G06F 17/30 (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) 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, EPODOC, IEEE: 统计, 相邻, 词频, 分词, 特征, 频率, 神经网络, 相似, binary, feature, two, statistic+, module, similar+, adjacent, train, neural, network																				
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 106021572 A (北京百度网讯科技有限公司) 2016年 10月 12日 (2016 - 10 - 12) 说明书第[0025]-[0079]段</td> <td>1-13</td> </tr> <tr> <td>Y</td> <td>CN 104899190 A (百度在线网络技术北京有限公司) 2015年 9月 9日 (2015 - 09 - 09) 说明书第[0035]-[0106]段</td> <td>1-13</td> </tr> <tr> <td>Y</td> <td>CN 104375989 A (国家电网公司 等) 2015年 2月 25日 (2015 - 02 - 25) 说明书第[0032]-[0062]段</td> <td>1-13</td> </tr> <tr> <td>Y</td> <td>CN 104391963 A (北京中科益创科技有限公司) 2015年 3月 4日 (2015 - 03 - 04) 说明书第[0031]-[0064]段</td> <td>1-13</td> </tr> <tr> <td>A</td> <td>US 2014289238 A1 (KABUSHIKI KAISHA TOSHIBA) 2014年 9月 25日 (2014 - 09 - 25) 全文</td> <td>1-13</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 106021572 A (北京百度网讯科技有限公司) 2016年 10月 12日 (2016 - 10 - 12) 说明书第[0025]-[0079]段	1-13	Y	CN 104899190 A (百度在线网络技术北京有限公司) 2015年 9月 9日 (2015 - 09 - 09) 说明书第[0035]-[0106]段	1-13	Y	CN 104375989 A (国家电网公司 等) 2015年 2月 25日 (2015 - 02 - 25) 说明书第[0032]-[0062]段	1-13	Y	CN 104391963 A (北京中科益创科技有限公司) 2015年 3月 4日 (2015 - 03 - 04) 说明书第[0031]-[0064]段	1-13	A	US 2014289238 A1 (KABUSHIKI KAISHA TOSHIBA) 2014年 9月 25日 (2014 - 09 - 25) 全文	1-13
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求																		
PX	CN 106021572 A (北京百度网讯科技有限公司) 2016年 10月 12日 (2016 - 10 - 12) 说明书第[0025]-[0079]段	1-13																		
Y	CN 104899190 A (百度在线网络技术北京有限公司) 2015年 9月 9日 (2015 - 09 - 09) 说明书第[0035]-[0106]段	1-13																		
Y	CN 104375989 A (国家电网公司 等) 2015年 2月 25日 (2015 - 02 - 25) 说明书第[0032]-[0062]段	1-13																		
Y	CN 104391963 A (北京中科益创科技有限公司) 2015年 3月 4日 (2015 - 03 - 04) 说明书第[0031]-[0064]段	1-13																		
A	US 2014289238 A1 (KABUSHIKI KAISHA TOSHIBA) 2014年 9月 25日 (2014 - 09 - 25) 全文	1-13																		
☐ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。 <table border="0"> <tr> <td> * 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 </td> <td> “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 </td> </tr> </table>			* 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																
* 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 2017年 2月 24日		国际检索报告邮寄日期 2017年 3月 31日																		
ISA/CN的名称和邮寄地址 中华人民共和国国家知识产权局 (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 林少华 电话号码 (86-10) 62414053																		

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2016/111571

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	106021572	A	2016年 10月 12日	无	
CN	104899190	A	2015年 9月 9日	无	
CN	104375989	A	2015年 2月 25日	无	
CN	104391963	A	2015年 3月 4日	无	
US	2014289238	A1	2014年 9月 25日	CN 104077346 A	2014年 10月 1日
				JP 2014186395 A	2014年 10月 2日