

(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 7 月 23 日 (23.07.2020)



(10) 国际公布号
WO 2020/147393 A1

(51) 国际专利分类号:
G06F 16/35 (2019.01)

(21) 国际申请号: PCT/CN2019/117008

(22) 国际申请日: 2019 年 11 月 11 日 (11.11.2019)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201910042629.5 2019 年 1 月 17 日 (17.01.2019) 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 518033 (CN)。

(72) 发明人: 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 518033 (CN)。 金戈(JIN, Ge); 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 518033 (CN)。 肖京(XIAO, Jing); 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 518033 (CN)。

(74) 代理人: 北京市京大律师事务所(BEIJING JINGDA LAW FIRM); 中国北京市海淀区海淀中街 16 号 7 层 2 单元 703, Beijing 100080 (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: CONVOLUTIONAL NEURAL NETWORK-BASED TEXT CLASSIFICATION METHOD, AND RELATED DEVICE

(54) 发明名称: 基于卷积神经网络的文本分类方法及相关设备

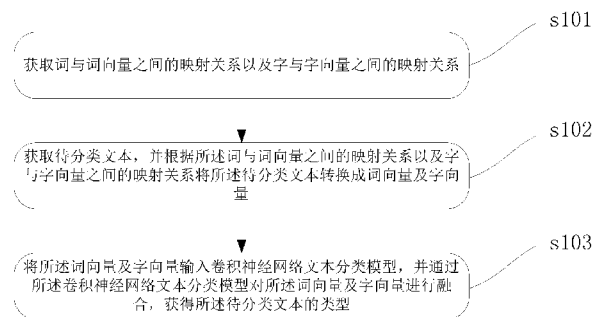


图 1

- S101 Acquire a mapping relationship between words and word vectors and a mapping relationship between characters and character vectors
- S102 Acquire a text to be classified, and according to the mapping relationship between words and word vectors and the mapping relationship between characters and character vectors, convert said text into word vectors and character vectors
- S103 Input the word vectors and the character vectors into a convolutional neural network text classification model, and fuse the word vectors and the character vectors by means of the convolutional neural network text classification model to obtain the type of said text

(57) Abstract: A convolutional neural network-based text classification method and a related device. The method comprises: acquiring a mapping relationship between words and word vectors and a mapping relationship between characters and character vectors (S101); acquiring a text to be classified, and according to the mapping relationship between words and word vectors and the mapping relationship between characters and character vectors, converting said text into word vectors and character vectors (S102); and inputting the word vectors and the character vectors into a convolutional neural network text classification model, and fusing the word vectors and the



WO 2020/147393 A1

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)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

character vectors by means of the convolutional neural network text classification model to obtain the type of said text (S103). The method extracts word vectors and character vectors from a text to be classified, and inputs the word vectors and the character vectors into a convolutional neural network for fusion, effectively improving the accuracy of text classification.

(57) 摘要: 一种基于卷积神经网络的文本分类方法及相关设备, 所述方法包括: 获取词与词向量之间的映射关系以及字与字向量之间的映射关系 (S101); 获取待分类文本, 并根据所述词与词向量之间的映射关系以及字与字向量之间的映射关系将所述待分类文本转换成词向量及字向量 (S102); 将所述词向量及字向量输入卷积神经网络文本分类模型, 并通过所述卷积神经网络文本分类模型对所述词向量及字向量进行融合, 获得所述待分类文本的类型 (S103)。上述方法通过将待分类文本提取词向量和字向量, 并将所述词向量和字向量输入卷积神经网络进行融合, 可以有效的提高文本分类的准确度。

基于卷积神经网络的文本分类方法及相关设备

本申请要求于 2019 年 01 月 17 日提交中国专利局、申请号为 201910042629.5、发明名称为“基于卷积神经网络的文本分类方法及相关设备”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

5 技术领域

本申请涉及人工智能领域，特别涉及一种基于卷积神经网络的文本分类方法及相关设备。

背景技术

文本分类是对大量非结构化的文字信息（文本文档、网页等）按照给定的
10 分类体系，根据文字信息内容分到指定的类别中去，是一种有指导的学习过程。词匹配法是最早被提出的分类算法。这种方法仅根据文档中是否出现了与类名相同的词来判断文档是否属于某个类别。很显然，这种过于简单机械的方法无法带来良好的分类效果。现如今，统计学习方法已经成为了文本分类领域绝对的主流。主要的原因在于其中的很多技术拥有坚实的理论基础，存在明确的评价标准，以及实际表现良好。统计分类算法将样本数据成功转化为向量表示之
15 后，计算机才算开始真正意义上的“学习”过程。常用的分类算法为：决策树，Rocchio，朴素贝叶斯，神经网络，支持向量机，线性最小平方拟合，kNN，遗传算法，最大熵，Generalized Instance Set 等。

现有神经网络文本分类模型主要基于词向量，尽管词向量在文本分类模型
20 中的应用优于字向量，但是字向量能够通过字符层面表示文本语义，对词向量的应用是很好的补充。目前对文本分类的方法没有针对词向量和字向量的，而少了字向量，会显著降低文本分类的准确度，不利于对文本的分析。

发明内容

本申请的目的在于针对现有技术的不足，提供一种基于卷积神经网络的文
25 本分类方法及相关设备，通过将待分类文本提取词向量和字向量，并将所述词向量和字向量输入卷积神经网络进行融合，可以有效的提高文本分类的准确度。

为达到上述目的，本申请的技术方案提供一种基于卷积神经网络的文本分类方法及相关设备。

本申请公开了一种基于卷积神经网络的文本分类方法，包括以下步骤：

获取词与词向量之间的映射关系以及字与字向量之间的映射关系；

获取待分类文本，并根据所述词与词向量之间的映射关系以及字与字向量之间的映射关系将所述待分类文本转换成词向量及字向量；

5 将所述词向量及字向量输入卷积神经网络文本分类模型，并通过所述卷积神经网络文本分类模型对所述词向量及字向量进行融合，获得所述待分类文本的类型。

本申请还公开了一种基于卷积神经网络的文本分类装置，所述装置包括：

10 向量映射模块：设置为获取词与词向量之间的映射关系以及字与字向量之间的映射关系；

向量生成模块：设置为获取待分类文本，并根据所述词与词向量之间的映射关系以及字与字向量之间的映射关系将所述待分类文本转换成词向量及字向量；

15 文本分类模块：设置为将所述词向量及字向量输入卷积神经网络文本分类模型，并通过所述卷积神经网络文本分类模型对所述词向量及字向量进行融合，获得所述待分类文本的类型。

本申请还公开了一种计算机设备，所述计算机设备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，所述计算机可读指令被一个或多个所述处理器执行时，使得一个或多个所述处理器执行以下步骤：

20 获取词与词向量之间的映射关系以及字与字向量之间的映射关系；

获取待分类文本，并根据所述词与词向量之间的映射关系以及字与字向量之间的映射关系将所述待分类文本转换成词向量及字向量；

25 将所述词向量及字向量输入卷积神经网络文本分类模型，并通过所述卷积神经网络文本分类模型对所述词向量及字向量进行融合，获得所述待分类文本的类型。

本申请还公开了一种存储介质，所述存储介质可被处理器读写，所述存储介质存储有计算机指令，所述计算机可读指令被一个或多个处理器执行时，使得一个或多个处理器执行以下步骤：

获取词与词向量之间的映射关系以及字与字向量之间的映射关系；

获取待分类文本，并根据所述词与词向量之间的映射关系以及字与字向量之间的映射关系将所述待分类文本转换成词向量及字向量；

5 将所述词向量及字向量输入卷积神经网络文本分类模型，并通过所述卷积神经网络文本分类模型对所述词向量及字向量进行融合，获得所述待分类文本的类型。

本申请的有益效果是：本申请通过将待分类文本提取词向量和字向量，并将所述词向量和字向量输入卷积神经网络进行融合，可以有效地提高文本分类的准确度。

10 附图说明

图 1 为本申请实施例的一种基于卷积神经网络的文本分类方法的流程示意图；

图 2 为本申请实施例的一种基于卷积神经网络的文本分类方法的流程示意图；

15 图 3 为本申请实施例的一种基于卷积神经网络的文本分类方法的流程示意图；

图 4 为本申请实施例的一种基于卷积神经网络的文本分类方法的流程示意图；

20 图 5 为本申请实施例的一种基于卷积神经网络的文本分类方法的流程示意图；

图 6 为本申请实施例的一种基于卷积神经网络的文本分类方法的流程示意图；

图 7 为本申请实施例的一种基于卷积神经网络的文本分类方法的流程示意图；

25 图 8 为本申请实施例的一种基于卷积神经网络的文本分类装置结构示意图。

具体实施方式

为了使本申请的目的、技术方案及优点更加清楚明白，以下结合附图及实施例，对本申请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

仅用以解释本申请，并不用于限定本申请。

本技术领域技术人员可以理解，除非特意声明，这里使用的单数形式“一”、“一个”、“所述”和“该”也可包括复数形式。应该进一步理解的是，本申请的说明书中使用的措辞“包括”是指存在所述特征、整数、步骤、操作、元件和/或组件，
5 但是并不排除存在或添加一个或多个其他特征、整数、步骤、操作、元件、组件和/或它们的组。

本申请实施例的一种基于卷积神经网络的文本分类方法流程如图 1 所示，本实施例包括以下步骤：

步骤 s101，获取词与词向量之间的映射关系以及字与字向量之间的映射关系；
10

具体的，文本的分类是基于词向量和字向量，而待分类文本可看成是由词和字组成的文本，因此在对待分类文本进行向量转换之前，可预先设定词与词向量之间的映射关系及字与字向量之间的映射关系。

步骤 s102，获取待分类文本，并根据所述词与词向量之间的映射关系以及
15 字与字向量之间的映射关系将所述待分类文本转换成词向量及字向量；

具体的，当获取一篇待分类文本后，可先将所述待分类文本进行分词和分字，由于所述待分类文本也是有词和字组成，因此可将所述待分类文本分割成词和字，分别获得词数据和字数据，然后根据词与词向量之间的映射关系将所述词数据转换成词向量，根据字与字向量之间的映射关系将所述字数据转换成
20 字向量。

步骤 s103，将所述词向量及字向量输入卷积神经网络文本分类模型，并通过所述卷积神经网络文本分类模型对所述词向量及字向量进行融合，获得所述待分类文本的类型。

具体的，当获取到所述待分类文本的词向量及字向量后，可将所述词向量和字向量同时输入卷积神经网络文本分类模型，所述卷积神经网络文本分类模型包含卷积层及全连接层，所述卷积层对所述词向量和字向量进行卷积运算，分别提取所述词向量和字向量的特征，然后将所述词向量和字向量的特征输入全连接层进行融合，当通过所述全连接层对所述词向量和字向量的特征信息进
25

行融合后，输入输出层，获取所述待分类文本的类型。

本实施例中，通过将待分类文本提取词向量和字向量，并将所述词向量和字向量输入卷积神经网络进行融合，可以有效的提高文本分类的准确度。

图 2 为本申请实施例的一种基于卷积神经网络的文本分类方法流程示意图，
5 如图所示，所述步骤 s101，获取词与词向量之间的映射关系以及字与字向量之间的映射关系，包括：

步骤 s201，获取文本训练数据，对所述文本训练数据进行分词，获得词数据；

具体的，所述文本训练数据可以采用中文维基百科作为训练语料，当获取
10 到所述训练数据后，可通过 Python 中的 jieba 模块对所述训练数据进行分词，即对训练文本通过 jieba 分词工具划分成一组词数据。

步骤 s202，对所述文本训练数据进行分字，获得字数据；

具体的，通过分词工具 jieba 对所述训练文本进行分词后，可将训练文本中的每个字提取出来，获得一组字数据。

15 步骤 s203，对所述词数据和字数据通过 word2vec 模型进行转换，获得词向量和字向量，并分别建立词与词向量之间的映射关系及字与字向量之间的映射关系。

具体的，可先将所述词数据载入 gensim 库中的 word2vec 模块，将所述词数据转换成词向量，并将所述词与词向量之间的映射关系保存下来；然后将所述
20 字数据也载入 gensim 库中的 word2vec 模块，将所述字数据也转换成字向量，并将所述字与字向量之间的映射关系保存下来。

本实施例中，通过 jieba 模块和 word2vec 模块对训练文本向量的转换，可以有效获取词与词向量之间的映射关系及字与字向量之间的映射关系。

图 3 为本申请实施例的一种基于卷积神经网络的文本分类方法流程示意图，
25 如图所示，所述步骤 s102，获取待分类文本，并根据所述词与词向量之间的映射关系以及字与字向量之间的映射关系将所述待分类文本转换成词向量及字向量，包括：

步骤 s301，获取待分类文本，将所述待分类文本进行分词，获得词数据，

并根据所述词与词向量之间的映射关系将所述词数据转换为词向量；

具体的，所述待分类文本可以是一篇文档，也可以是一个网页，当获取到所述待分类文本后，可通过分词工具，如 jieba 分词工具，对所述待分类文本进行分词，获得词数据，然后将所述词数据根据步骤 s101 中词与词向量之间的映射关系转换为词向量。

步骤 s302，对所述待分类文本进行分字，获得字数据，并根据所述字与字向量之间的映射关系将所述字数据转换为字向量。

具体的，可先将所述待分类文本分割成一个一个字，获得一组字数据，然后将所述字数据根据步骤 s101 中字与字向量之间的映射关系转换为字向量。

本实施例中，通过词与词向量之间的映射关系及字与字向量之间的映射关系，可将所述待分类文本转换成词向量及字向量。

图 4 为本申请实施例的一种基于卷积神经网络的文本分类方法流程示意图，如图所示，所述步骤 s103，将所述词向量及字向量输入卷积神经网络文本分类模型，并通过所述卷积神经网络文本分类模型对所述词向量及字向量进行融合，获得所述待分类文本的类型，包括：

步骤 s401，将所述词向量及字向量输入卷积神经网络文本分类模型的卷积层，通过所述卷积层对所述词向量及字向量进行卷积运算分别获得所述词向量及字向量的特征，并发送给全连接层；

具体的，可先将所述词向量及字向量输入卷积神经网络文本分类模型的卷积层，在所述卷积层中可先建立尺度为 1、3、5 的一维卷积核，分别对所述词向量和字向量进行卷积运算以提取所述词向量和字向量的特征，每种尺度的一维卷积核的通道数为 128，卷积运算后的结果可通过激活函数 ReLU 激活，并输入到池化层进行数据压缩，然后将所述词向量和字向量的特征信息发送给全连接层。

步骤 s402，通过所述全连接层对所述词向量及字向量的特征进行融合，获得词向量和字向量的融合信息，根据所述词向量和字向量的融合信息获得所述待分类文本的类型。

具体的，所述全连接层连接所有的卷积通道，所述卷积通道包含两路信息，

即词向量信息和字向量信息，当所述全连接层获取到所述词向量信息和字向量信息后，可对所述词向量信息和字向量信息进行融合，即将所述词向量信息和字向量信息转换为文本类型信息，然后根据所述文本类型信息计算出属于各个文本类型的概率，并从中选取最大的那个概率对应的文本类型作为待分类文本的类型。

本实施例中，通过卷积层成运算和全连接层的信息融合，可以有效获取文本类型，提高文本分类准确度。

图5为本申请实施例的一种基于卷积神经网络的文本分类方法流程示意图，如图所示，所述步骤s401，将所述词向量及字向量输入卷积神经网络文本分类模型的卷积层，通过所述卷积层对所述词向量及字向量进行卷积运算分别获得所述词向量及字向量的特征，并发送给全连接层，包括：

步骤s501，将所述词向量及字向量输入卷积神经网络文本分类模型的卷积层，通过所述卷积层的卷积运算分别获得所述词向量及字向量的特征，并发送给注意力层；

具体的，可先将所述词向量及字向量输入卷积神经网络文本分类模型的卷积层，在所述卷积层中可先建立尺度为1、3、5的一维卷积核，分别对所述词向量和字向量进行卷积运算以提取所述词向量和字向量的特征，每种尺度的一维卷积核的通道数为128，卷积运算后的结果可通过ReLU激活函数激活，并输入到池化层进行数据压缩，然后将所述词向量和字向量的特征信息发送给注意力层。

步骤s502，通过所述注意力层分别对所述词向量和字向量进行权重分配后发送给全连接层。

具体的，当注意力层获取到所述词向量和字向量的两路特征信息后，可分别对词向量通道信息和字向量通道信息进行权重分配，所述注意力层是平行于全连接层的一个全连接结构，连接卷积输出，并通过softmax函数输出，所述softmax函数用于对通道进行权重分配，以词向量通道为例，如词向量通道有128个通道，每个通道对应词向量的特征，那么通过softmax函数可对所述128个通道进行权重分配，包含重要特征信息的通道会分配较大的权重，这样可以

过滤不必要的词组信息，当分配好权重以后，对每条通道的特征信息乘以权重然后相加即为词向量通道总的特征信息；同样的，对字向量通道进行权重分配并加权计算后，将所述词向量信息和字向量信息发送给全连接层进行信息融合。

本实施例中，通过注意力层进行权重分配，可以过滤词向量和字向量中不重要的特征信息，提高文本分类的效率。

图6为本申请实施例的一种基于卷积神经网络的文本分类方法流程示意图，如图所示，所述步骤s103，将所述词向量及字向量输入卷积神经网络文本分类模型，并通过所述卷积神经网络文本分类模型对所述词向量及字向量进行融合，获得所述待分类文本的类型，包括：

10 步骤s601，将所述词向量及字向量输入卷积神经网络文本分类模型的第一卷积层，通过所述第一卷积层对所述词向量及字向量进行卷积运算后发送至第一全连接层；

具体的，可先将所述词向量及字向量输入卷积神经网络文本分类模型的第一个卷积层，在第一个卷积层中可先建立尺度为1、3、5的一维卷积核，分别
15 对所述词向量和字向量进行卷积运算以提取所述词向量和字向量的特征，每种尺度的一维卷积核的通道数为128，卷积运算后的结果可通过ReLU激活函数激活，并输入到池化层进行数据压缩，然后将所述词向量和字向量的特征信息发送给第一个全连接层。

步骤s602，通过所述第一全连接层对所述词向量和字向量进行融合后，获得
20 第一融合信息，并将所述第一融合信息发送至第二卷积层；

具体的，当第一个全连接层收到词向量和字向量的特征信息后，对词向量和字向量的两路特征信息进行融合，获得第一个融合信息，并将第一个融合信息发送至第二个卷积层。

步骤s603，通过所述第二卷积层对所述第一融合信息进行卷积运算后发送
25 至第二全连接层，通过所述第二全连接层进行融合后获得第二融合信息，并根据所述第二融合信息获得所述待分类文本的类型。

具体的，当第二个卷积层收到第一个融合信息后，对第一个融合信息建立通道并再次进行卷积运算，提取第一个融合信息中的特征信息，并发送给第二

个全连接层，通过第二个全连接层对卷积通道的输出信息进行再次融合，获得第二个融合信息，然后根据第二个融合信息计算出属于各个文本类型的概率，并从中选取最大的那个概率对应的文本类型作为待分类文本的类型。

本实施例中，通过两个卷积层的卷积运算和两个全连接层的信息融合，可以
5 以提高文本分类的准确度。

图 7 为本申请实施例的一种基于卷积神经网络的文本分类方法流程示意图，如图所示，所述步骤 s603，通过所述第二全连接层进行融合后获得第二融合信息，并根据所述第二融合信息获得所述待分类文本的类型，包括：

步骤 s701，通过所述第二全连接层进行融合后获得第二融合信息，并将所
10 述第二融合信息发送给输出层；

具体的，通过第二全连接层对卷积运算后的第一个融合信息进行再次融合后，获得第二个融合信息，并将第二个融合信息发送给输出层。

步骤 s702，通过所述输出层的 softmax 函数根据所述第二融合信息获取每个文本类型的概率，在所述概率中获取最大的概率，并将所述最大的概率对应
15 的文本类型作为所述待分类文本的类型进行输出。

具体的，输出层收到第二个融合信息后，所述第二个融合信息可看做文本特征在各个文本类型上的分布，例如，本次文本分类的类型是体育或者财经，那么第二融合信息就是指有多少信息是包含体育特征的，有多少信息是包含财经特征的，然后通过输出层的 softmax 函数计算属于体育类型的概率及属于财经类型的概率，然后选择概率最大的类型作为输出类型，例如，属于体育类型的
20 的概率为 0.8，属于财经类型的概率为 0.2，那么待分类文本的类型为体育类型。

本实施例中，通过输出层对融合信息的分析计算，可以有效获取待分类文本的类型。

本申请实施例的一种基于卷积神经网络的文本分类装置结构如图 8 所示，
25 包括：

向量映射模块 801、向量生成模块 802 及文本分类模块 803；其中，向量映射模块 801 与向量生成模块 802 相连，向量生成模块 802 与文本分类模块 803 相连；向量映射模块 801 设置为获取词与词向量之间的映射关系以及字与字向

量之间的映射关系；向量生成模块 802 设置为获取待分类文本，并根据所述词与词向量之间的映射关系以及字与字向量之间的映射关系将所述待分类文本转换成词向量及字向量；文本分类模块 803 设置为将所述词向量及字向量输入卷积神经网络文本分类模型，并通过所述卷积神经网络文本分类模型对所述词向量及字向量进行融合，获得所述待分类文本的类型。

在一个实施例中，向量映射模块，包括：

分词单元：设置为获取文本训练数据，对所述文本训练数据进行分词，获得词数据；

分字单元：设置为对所述文本训练数据进行分字，获得字数据；

10 转换单元：设置为对所述词数据和字数据通过 word2vec 模型进行转换，获得词向量和字向量，并分别建立词与词向量之间的映射关系及字与字向量之间的映射关系。

在一个实施例中，向量生成模块，包括：

15 第一向量转换单元：设置为获取待分类文本，将所述待分类文本进行分词，获得词数据，并根据所述词与词向量之间的映射关系将所述词数据转换为词向量；

第二向量转换单元：设置为对所述待分类文本进行分字，获得字数据，并根据所述字与字向量之间的映射关系将所述字数据转换为字向量。

在一个实施例中，文本分类模块，包括：

20 第一卷积单元：设置为将所述词向量及字向量输入卷积神经网络文本分类模型的卷积层，通过所述卷积层对所述词向量及字向量进行卷积运算分别获得所述词向量及字向量的特征，并发送给全连接层；

第一融合单元：设置为通过所述全连接层对所述词向量及字向量的特征进行融合，获得词向量和字向量的融合信息，根据所述词向量和字向量的融合信息获得所述待分类文本的类型。

在一个实施例中，文本分类模块，包括：

特征获取单元：设置为将所述词向量及字向量输入卷积神经网络文本分类模型的卷积层，通过所述卷积层的卷积运算分别获得所述词向量及字向量的特

征，并发送给注意力层；

权重分配单元：设置为通过所述注意力层分别对所述词向量和字向量进行权重分配后发送给全连接层。

在一个实施例中，文本分类模块，包括：

5 第二卷积单元：设置为将所述词向量及字向量输入卷积神经网络文本分类模型的第一卷积层，通过所述第一卷积层对所述词向量及字向量进行卷积运算后发送至第一全连接层；

第二融合单元：设置为通过所述第一全连接层对所述词向量和字向量进行融合后，获得第一融合信息，并将所述第一融合信息发送至第二卷积层；

10 分类单元：设置为通过所述第二卷积层对所述第一融合信息进行卷积运算后发送至第二全连接层，通过所述第二全连接层进行融合后获得第二融合信息，并根据所述第二融合信息获得所述待分类文本的类型。

在一个实施例中，文本分类模块，包括：

15 发送单元：设置为通过所述第二全连接层进行融合后获得第二融合信息，并将所述第二融合信息发送给输出层；

输出单元：设置为通过所述输出层的 softmax 函数根据所述第二融合信息获取每个文本类型的概率，在所述概率中获取最大的概率，并将所述最大的概率对应的文本类型作为所述待分类文本的类型进行输出。

20 本申请实施例还公开了一种计算机设备，所述计算机设备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，所述计算机可读指令被一个或多个所述处理器执行时，使得一个或多个所述处理器执行上述各实施例中所述文本分类方法中的步骤。

25 本申请实施例还公开了一种存储介质，所述存储介质可被处理器读写，所述存储器存储有计算机可读指令，所述计算机可读指令被一个或多个处理器执行时，使得一个或多个处理器执行上述各实施例中所述文本分类方法中的步骤。

本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程，是可以通过计算机程序来指令相关的硬件来完成，该计算机程序可存储于一计算机可读取存储介质中，该程序在执行时，可包括如上述各方法的实施例的流

程。其中，前述的存储介质可为磁碟、光盘、只读存储记忆体（Read-Only Memory, ROM）等非易失性存储介质，或随机存储记忆体（Random Access Memory, RAM）等。

5 以上所述实施例的各技术特征可以进行任意的组合，为使描述简洁，未对上述实施例中的各个技术特征所有可能的组合都进行描述，然而，只要这些技术特征的组合不存在矛盾，都应当认为是本说明书记载的范围。

10 以上所述实施例仅表达了本申请的几种实施方式，其描述较为具体和详细，但并不能因此而理解为对本申请专利范围的限制。应当指出的是，对于本领域的普通技术人员来说，在不脱离本申请构思的前提下，还可以做出若干变形和改进，这些都属于本申请的保护范围。因此，本申请专利的保护范围应以所附权利要求为准。

权 利 要 求 书

1、一种基于卷积神经网络的文本分类方法，包括以下步骤：

获取词与词向量之间的映射关系以及字与字向量之间的映射关系；

获取待分类文本，并根据所述词与词向量之间的映射关系以及字与字向量之间的映射关系将所述待分类文本转换成词向量及字向量；

5 将所述词向量及字向量输入卷积神经网络文本分类模型，并通过所述卷积神经网络文本分类模型对所述词向量及字向量进行融合，获得所述待分类文本的类型。

2、如权利要求 1 所述的基于卷积神经网络的文本分类方法，其中，所述获取词与词向量之间的映射关系以及字与字向量之间的映射关系，包括：

10 获取文本训练数据，对所述文本训练数据进行分词，获得词数据；

对所述文本训练数据进行分字，获得字数据；

对所述词数据和字数据通过 word2vec 模型进行转换，获得词向量和字向量，并分别建立词与词向量之间的映射关系及字与字向量之间的映射关系。

3、如权利要求 1 所述的基于卷积神经网络的文本分类方法，其中，所述获取待分类文本，并根据所述词与词向量之间的映射关系以及字与字向量之间的映射关系将所述待分类文本转换成词向量及字向量，包括：

获取待分类文本，将所述待分类文本进行分词，获得词数据，并根据所述词与词向量之间的映射关系将所述词数据转换为词向量；

20 对所述待分类文本进行分字，获得字数据，并根据所述字与字向量之间的映射关系将所述字数据转换为字向量。

4、如权利要求 1 所述的基于卷积神经网络的文本分类方法，其中，所述将所述词向量及字向量输入卷积神经网络文本分类模型，并通过所述卷积神经网络文本分类模型对所述词向量及字向量进行融合，获得所述待分类文本的类型包括：

25 将所述词向量及字向量输入卷积神经网络文本分类模型的卷积层，通过所述卷积层对所述词向量及字向量进行卷积运算分别获得所述词向量及字向量的特征，并发送给全连接层；

通过所述全连接层对所述词向量及字向量的特征进行融合，获得词向量和字向量的融合信息，根据所述词向量和字向量的融合信息获得所述待分类文本的类型。

5 5、如权利要求4所述的基于卷积神经网络的文本分类方法，其中，所述将所述词向量及字向量输入卷积神经网络文本分类模型的卷积层，通过所述卷积层对所述词向量及字向量进行卷积运算分别获得所述词向量及字向量的特征，并发送给全连接层，包括：

将所述词向量及字向量输入卷积神经网络文本分类模型的卷积层，通过所述卷积层的卷积运算分别获得所述词向量及字向量的特征，并发送给注意力层；

10 通过所述注意力层分别对所述词向量和字向量进行权重分配后发送给全连接层。

6、如权利要求1所述的基于卷积神经网络的文本分类方法，其中，所述将所述词向量及字向量输入卷积神经网络文本分类模型，并通过所述卷积神经网络文本分类模型对所述词向量及字向量进行融合，获得所述待分类文本的类型，
15 包括：

将所述词向量及字向量输入卷积神经网络文本分类模型的第一卷积层，通过所述第一卷积层对所述词向量及字向量进行卷积运算后发送至第一全连接层；

20 通过所述第一全连接层对所述词向量和字向量进行融合后，获得第一融合信息，并将所述第一融合信息发送至第二卷积层；

通过所述第二卷积层对所述第一融合信息进行卷积运算后发送至第二全连接层，通过所述第二全连接层进行融合后获得第二融合信息，并根据所述第二融合信息获得所述待分类文本的类型。

7、如权利要求6所述的基于卷积神经网络的文本分类方法，其中，所述通
25 过所述第二全连接层进行融合后获得第二融合信息，并根据所述第二融合信息获得所述待分类文本的类型，包括：

通过所述第二全连接层进行融合后获得第二融合信息，并将所述第二融合信息发送给输出层；

通过所述输出层的 softmax 函数根据所述第二融合信息获取每个文本类型的概率，在所述概率中获取最大的概率，并将所述最大的概率对应的文本类型作为所述待分类文本的类型进行输出。

8、一种基于卷积神经网络的文本分类装置，所述装置包括：

5 向量映射模块：设置为获取词与词向量之间的映射关系以及字与字向量之间的映射关系；

向量生成模块：设置为获取待分类文本，并根据所述词与词向量之间的映射关系以及字与字向量之间的映射关系将所述待分类文本转换成词向量及字向量；

10 文本分类模块：设置为将所述词向量及字向量输入卷积神经网络文本分类模型，并通过所述卷积神经网络文本分类模型对所述词向量及字向量进行融合，获得所述待分类文本的类型。

9、根据权利要求 8 所述的基于卷积神经网络的文本分类装置，其中，所述向量映射模块，包括：

15 分词单元：设置为获取文本训练数据，对所述文本训练数据进行分词，获得词数据；

分字单元：设置为对所述文本训练数据进行分字，获得字数据；

转换单元：设置为对所述词数据和字数据通过 word2vec 模型进行转换，获得词向量和字向量，并分别建立词与词向量之间的映射关系及字与字向量之间的映射关系。

20 10、根据权利要求 8 所述的基于卷积神经网络的文本分类装置，其中，所述向量生成模块，包括：

第一向量转换单元：设置为获取待分类文本，将所述待分类文本进行分词，获得词数据，并根据所述词与词向量之间的映射关系将所述词数据转换为词向量；

第二向量转换单元：设置为对所述待分类文本进行分字，获得字数据，并根据所述字与字向量之间的映射关系将所述字数据转换为字向量。

11、根据权利要求 8 所述的基于卷积神经网络的文本分类装置，其中，所

述文本分类模块，包括：

第一卷积单元：设置为将所述词向量及字向量输入卷积神经网络文本分类模型的卷积层，通过所述卷积层对所述词向量及字向量进行卷积运算分别获得所述词向量及字向量的特征，并发送给全连接层；

5 第一融合单元：设置为通过所述全连接层对所述词向量及字向量的特征进行融合，获得词向量和字向量的融合信息，根据所述词向量和字向量的融合信息获得所述待分类文本的类型。

12、根据权利要求 11 所述的基于卷积神经网络的文本分类装置，其中，所述文本分类模块，包括：

10 特征获取单元：设置为将所述词向量及字向量输入卷积神经网络文本分类模型的卷积层，通过所述卷积层的卷积运算分别获得所述词向量及字向量的特征，并发送给注意力层；

权重分配单元：设置为通过所述注意力层分别对所述词向量和字向量进行权重分配后发送给全连接层。

15 13、根据权利要求 8 所述的基于卷积神经网络的文本分类装置，其中，所述文本分类模块，包括：

第二卷积单元：设置为将所述词向量及字向量输入卷积神经网络文本分类模型的第一卷积层，通过所述第一卷积层对所述词向量及字向量进行卷积运算后发送至第一全连接层；

20 第二融合单元：设置为通过所述第一全连接层对所述词向量和字向量进行融合后，获得第一融合信息，并将所述第一融合信息发送至第二卷积层；

分类单元：设置为通过所述第二卷积层对所述第一融合信息进行卷积运算后发送至第二全连接层，通过所述第二全连接层进行融合后获得第二融合信息，并根据所述第二融合信息获得所述待分类文本的类型。

25 14、根据权利要求 13 所述的基于卷积神经网络的文本分类装置，其中，所述文本分类模块，包括：

发送单元：设置为通过所述第二全连接层进行融合后获得第二融合信息，并将所述第二融合信息发送给输出层；

输出单元：设置为通过所述输出层的 softmax 函数根据所述第二融合信息获取每个文本类型的概率，在所述概率中获取最大的概率，并将所述最大的概率对应的文本类型作为所述待分类文本的类型进行输出。

15、一种计算机设备，所述计算机设备包括存储器和处理器，所述存储器
5 中存储有计算机可读指令，所述计算机可读指令被一个或多个所述处理器执行时，使得一个或多个所述处理器执行以下步骤：

获取词与词向量之间的映射关系以及字与字向量之间的映射关系；

获取待分类文本，并根据所述词与词向量之间的映射关系以及字与字向量之间的映射关系将所述待分类文本转换成词向量及字向量；

10 将所述词向量及字向量输入卷积神经网络文本分类模型，并通过所述卷积神经网络文本分类模型对所述词向量及字向量进行融合，获得所述待分类文本的类型。

16、根据权利要求 15 所述的计算机设备，其中，所述将所述词向量及字向量输入卷积神经网络文本分类模型，并通过所述卷积神经网络文本分类模型对
15 所述词向量及字向量进行融合，获得所述待分类文本的类型时，使得所述处理器执行以下步骤：

将所述词向量及字向量输入卷积神经网络文本分类模型的卷积层，通过所述卷积层对所述词向量及字向量进行卷积运算分别获得所述词向量及字向量的特征，并发送给全连接层；

20 通过所述全连接层对所述词向量及字向量的特征进行融合，获得词向量和字向量的融合信息，根据所述词向量和字向量的融合信息获得所述待分类文本的类型。

17、根据权利要求 15 所述的计算机设备，其中，所述将所述词向量及字向量输入卷积神经网络文本分类模型，并通过所述卷积神经网络文本分类模型对
25 所述词向量及字向量进行融合，获得所述待分类文本的类型时，使得所述处理器执行以下步骤：

将所述词向量及字向量输入卷积神经网络文本分类模型的第一卷积层，通

过所述第一卷积层对所述词向量及字向量进行卷积运算后发送至第一全连接层；

通过所述第一全连接层对所述词向量和字向量进行融合后，获得第一融合信息，并将所述第一融合信息发送至第二卷积层；

5 通过所述第二卷积层对所述第一融合信息进行卷积运算后发送至第二全连接层，通过所述第二全连接层进行融合后获得第二融合信息，并根据所述第二融合信息获得所述待分类文本的类型。

18、一种存储介质，所述存储介质可被处理器读写，所述存储介质存储有计算机指令，所述计算机可读指令被一个或多个处理器执行时，使得一个或多个处理器执行以下步骤：

获取词与词向量之间的映射关系以及字与字向量之间的映射关系；

获取待分类文本，并根据所述词与词向量之间的映射关系以及字与字向量之间的映射关系将所述待分类文本转换成词向量及字向量；

15 将所述词向量及字向量输入卷积神经网络文本分类模型，并通过所述卷积神经网络文本分类模型对所述词向量及字向量进行融合，获得所述待分类文本的类型。

19、根据权利要求 18 所述的存储介质，其中，所述将所述词向量及字向量输入卷积神经网络文本分类模型，并通过所述卷积神经网络文本分类模型对所述词向量及字向量进行融合，获得所述待分类文本的类型时，使得一个或多个所述处理器执行以下步骤：

将所述词向量及字向量输入卷积神经网络文本分类模型的卷积层，通过所述卷积层对所述词向量及字向量进行卷积运算分别获得所述词向量及字向量的特征，并发送给全连接层；

25 通过所述全连接层对所述词向量及字向量的特征进行融合，获得词向量和字向量的融合信息，根据所述词向量和字向量的融合信息获得所述待分类文本的类型。

20、根据权利要求 18 所述的存储介质，其中，所述将所述词向量及字向量输入卷积神经网络文本分类模型，并通过所述卷积神经网络文本分类模型对所

述词向量及字向量进行融合，获得所述待分类文本的类型时，使得一个或多个所述处理器执行以下步骤：

5 将所述词向量及字向量输入卷积神经网络文本分类模型的第一卷积层，通过所述第一卷积层对所述词向量及字向量进行卷积运算后发送至第一全连接层；

通过所述第一全连接层对所述词向量和字向量进行融合后，获得第一融合信息，并将所述第一融合信息发送至第二卷积层；

10 通过所述第二卷积层对所述第一融合信息进行卷积运算后发送至第二全连接层，通过所述第二全连接层进行融合后获得第二融合信息，并根据所述第二融合信息获得所述待分类文本的类型。

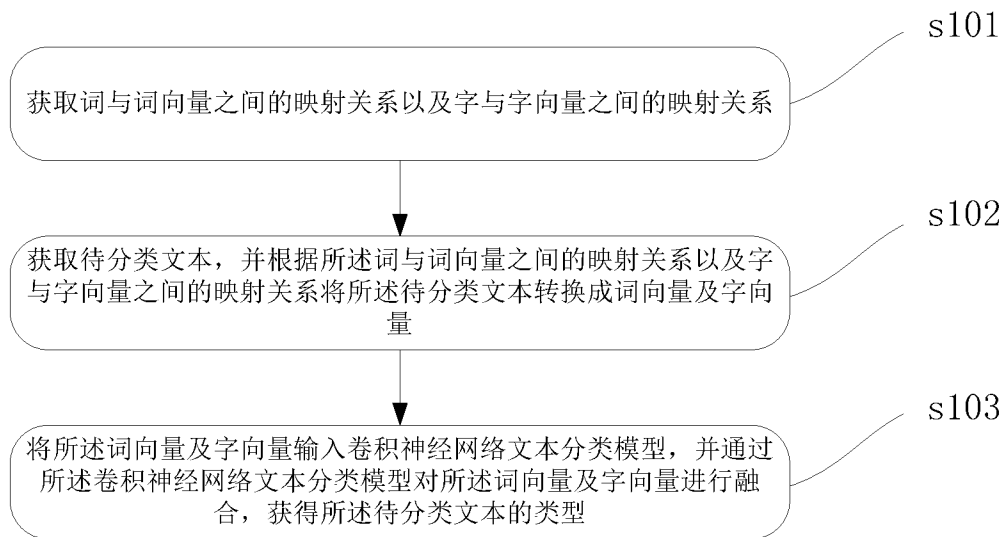


图 1

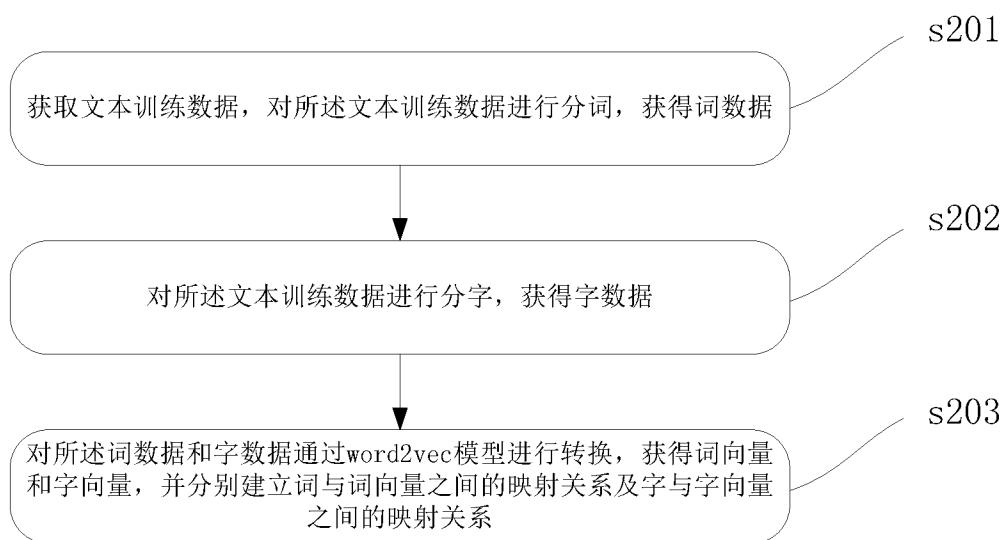


图 2

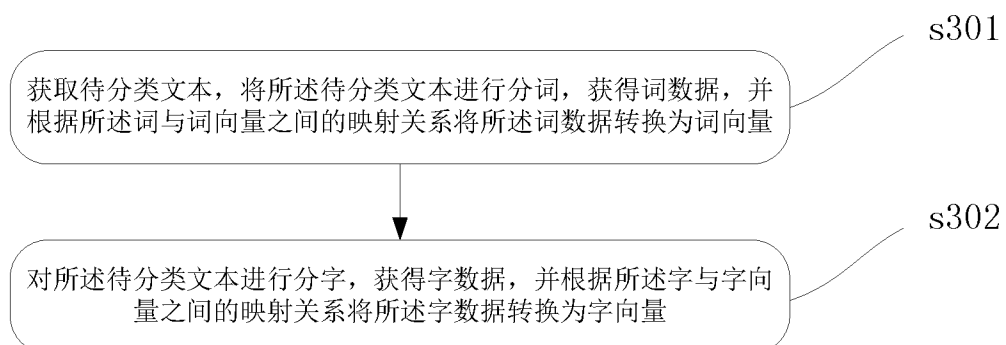


图 3

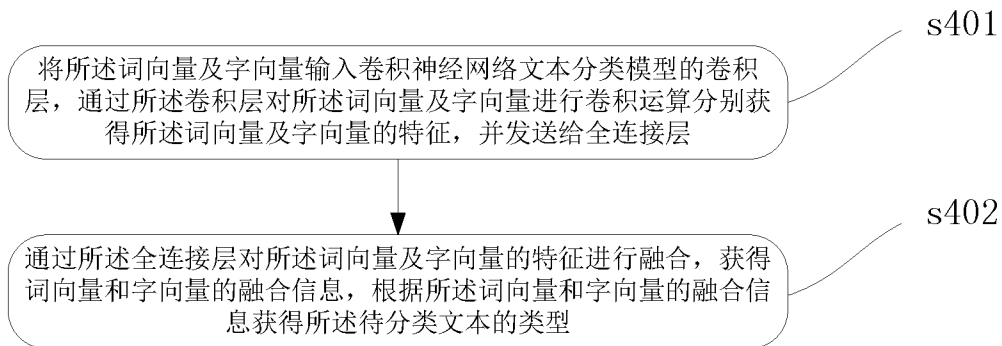


图 4

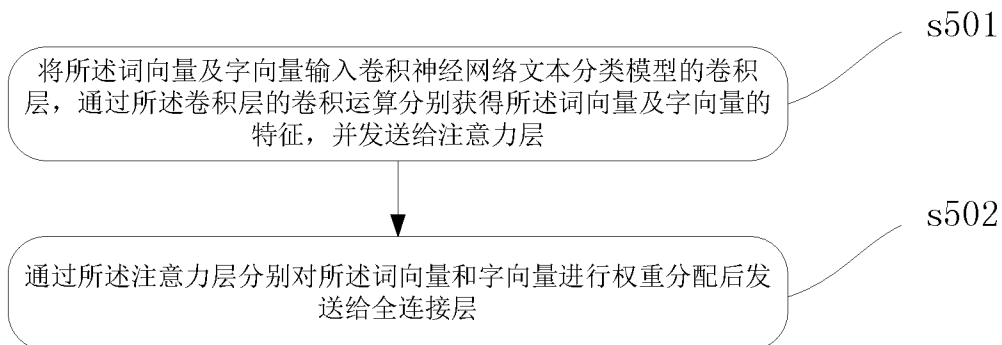


图 5

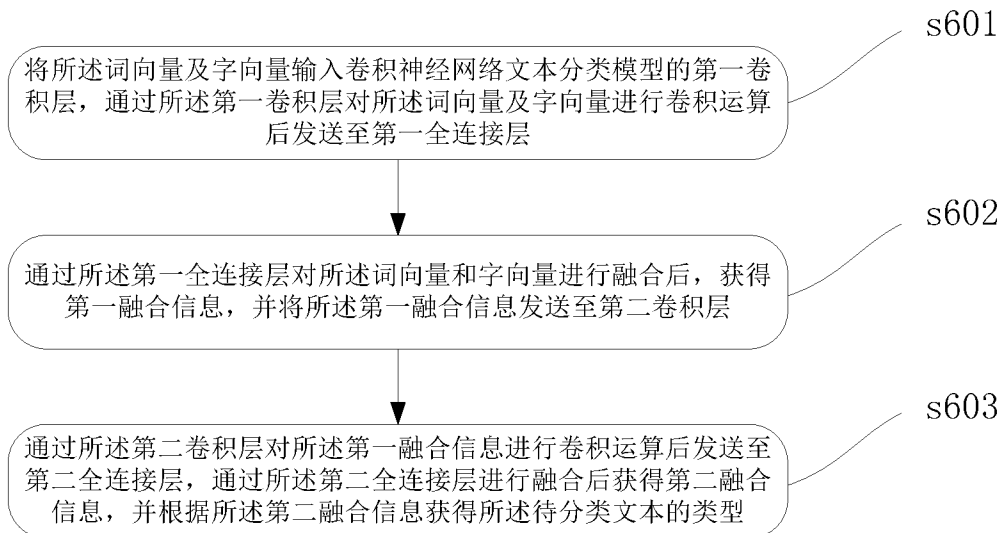


图 6

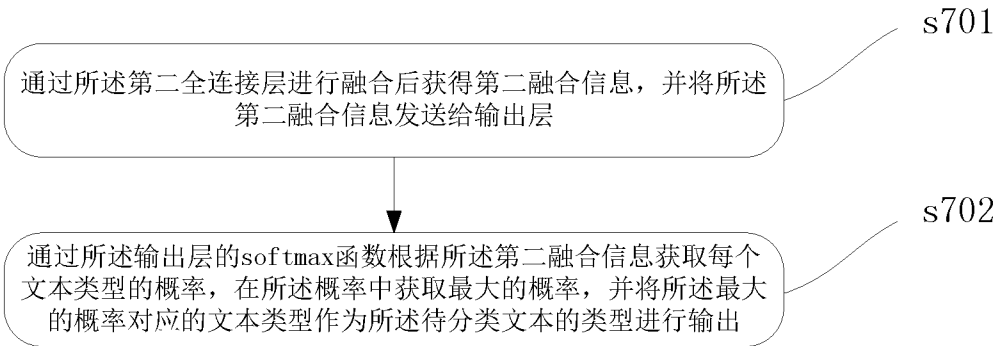


图 7

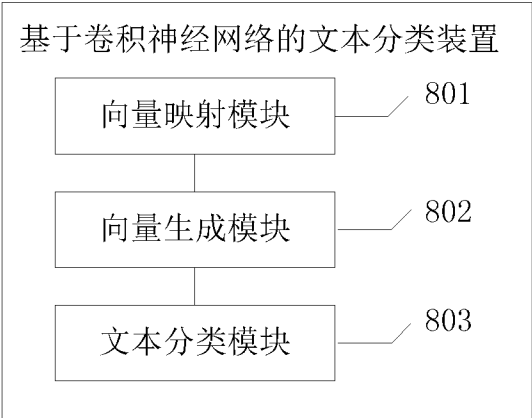


图 8

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2019/117008

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/35(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)

CNKI, CNPAT, WPI, EPODOC: 字, 词, 向量, 分类, 文本, 特征, 融合, 卷积神经, word, vector, classify, text, feature, CNN

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 109918500 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 21 June 2019 (2019-06-21) claims 1-10, and description, paragraphs [0048]-[0099]	1-20
X	CN 107656990 A (SUN YAT-SEN UNIVERSITY) 02 February 2018 (2018-02-02) claims 1-6, and description, paragraphs [0031]-[0068]	1-20
X	CN 108875034 A (HUNAN DANIEL INTELLIGENT TECH CO., LTD.) 23 November 2018 (2018-11-23) description, paragraphs [0031]-[0063]	1-20
X	CN 107301225 A (WACAI NETWORK TECH CO., LTD.) 27 October 2017 (2017-10-27) description, paragraphs [0059]-[0067]	1-20



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

08 January 2020

Date of mailing of the international search report

14 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/CN2019/117008

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN	109918500	A	21 June 2019	None	
CN	107656990	A	02 February 2018	None	
CN	108875034	A	23 November 2018	None	
CN	107301225	A	27 October 2017	None	

国际检索报告

国际申请号

PCT/CN2019/117008

A. 主题的分类 G06F 16/35 (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)) CNKI, CNPAT, WPI, EPDOC: 字, 词, 向量, 分类, 文本, 特征, 融合, 卷积神经, word, vector, classify, text, feature, CNN																	
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 109918500 A (平安科技深圳有限公司) 2019年 6月 21日 (2019 - 06 - 21) 权利要求1-10, 说明书第[0048]-[0099]段</td> <td>1-20</td> </tr> <tr> <td>X</td> <td>CN 107656990 A (中山大学) 2018年 2月 2日 (2018 - 02 - 02) 权利要求1-6, 说明书第[0031]-[0068]段</td> <td>1-20</td> </tr> <tr> <td>X</td> <td>CN 108875034 A (湖南丹尼尔智能科技有限公司) 2018年 11月 23日 (2018 - 11 - 23) 说明书第[0031]-[0063]段</td> <td>1-20</td> </tr> <tr> <td>X</td> <td>CN 107301225 A (挖财网络技术有限公司) 2017年 10月 27日 (2017 - 10 - 27) 说明书第[0059]-[0067]段</td> <td>1-20</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 109918500 A (平安科技深圳有限公司) 2019年 6月 21日 (2019 - 06 - 21) 权利要求1-10, 说明书第[0048]-[0099]段	1-20	X	CN 107656990 A (中山大学) 2018年 2月 2日 (2018 - 02 - 02) 权利要求1-6, 说明书第[0031]-[0068]段	1-20	X	CN 108875034 A (湖南丹尼尔智能科技有限公司) 2018年 11月 23日 (2018 - 11 - 23) 说明书第[0031]-[0063]段	1-20	X	CN 107301225 A (挖财网络技术有限公司) 2017年 10月 27日 (2017 - 10 - 27) 说明书第[0059]-[0067]段	1-20
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求															
PX	CN 109918500 A (平安科技深圳有限公司) 2019年 6月 21日 (2019 - 06 - 21) 权利要求1-10, 说明书第[0048]-[0099]段	1-20															
X	CN 107656990 A (中山大学) 2018年 2月 2日 (2018 - 02 - 02) 权利要求1-6, 说明书第[0031]-[0068]段	1-20															
X	CN 108875034 A (湖南丹尼尔智能科技有限公司) 2018年 11月 23日 (2018 - 11 - 23) 说明书第[0031]-[0063]段	1-20															
X	CN 107301225 A (挖财网络技术有限公司) 2017年 10月 27日 (2017 - 10 - 27) 说明书第[0059]-[0067]段	1-20															
☐ 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 1月 8日		国际检索报告邮寄日期 2020年 2月 14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 刘宇儒 电话号码 86-(10)-53961294															

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2019/117008

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	109918500	A	2019年 6月 21日	无	
CN	107656990	A	2018年 2月 2日	无	
CN	108875034	A	2018年 11月 23日	无	
CN	107301225	A	2017年 10月 27日	无	