

(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 6 月 4 日 (04.06.2020)



(10) 国际公布号
WO 2020/107835 A1

- (51) 国际专利分类号:
G06F 16/35 (2019.01)
- (21) 国际申请号: PCT/CN2019/088803
- (22) 国际申请日: 2019 年 5 月 28 日 (28.05.2019)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201811421160.8 2018 年 11 月 26 日 (26.11.2018) 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: 周涛涛(ZHOU, Taotao); 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)。周宝(ZHOU, Bao); 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)。陈远旭(CHEN, Yuanxu); 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)。王健宗(WANG, Jianzong); 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)。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 518000 (CN)。

(74) 代理人: 广州三环专利商标代理有限公司(SCIHEAD IP LAW FIRM); 中国广东省广州市越秀区先烈中路 80 号汇华商贸大厦 1508 室, Guangdong 510070 (CN)。

(54) Title: SAMPLE DATA PROCESSING METHOD AND DEVICE

(54) 发明名称: 一种样本数据处理方法及装置

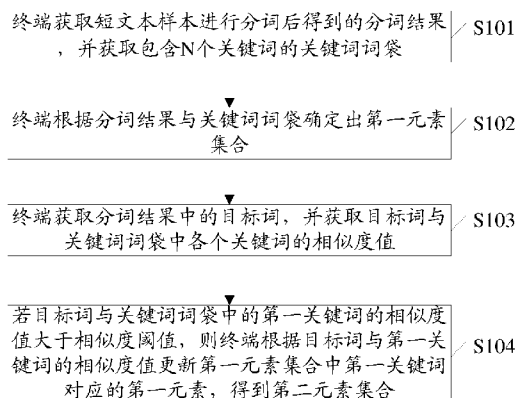


图 1

(57) Abstract: Disclosed are a sample data processing method and device. The method is suitable for training of a machine learning model for short text classification. The method comprises: obtaining a word segmentation result obtained by performing word segmentation on a short text sample, and obtaining a bag of keywords comprising N keywords; determining a first element set according to the word segmentation result and the bag of keywords; obtaining a target word in the word segmentation result, and obtaining a similarity value between the target word and each keyword in the bag of keywords; and if the similarity value between the target word and a first keyword in the bag of keywords is greater than a similarity threshold, updating a first element in the first element set according to the similarity value between the target word and the first keyword to obtain a second element set, each element in the second element set being used for constructing the machine learning model for short text classification. By adoption of embodiments of the present application, the performance of the machine learning model constructed using the element set can be improved.

(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, 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))。

(57) 摘要: 本申请实施例公开了一种样本数据处理方法及装置, 该方法适用于短文本分类的机器学习模型训练, 该方法包括: 获取短文本样本进行分词后的分词结果, 并获取包含N个关键词的关键词词袋, 根据该分词结果与该关键词词袋确定出第一元素集合, 获取该分词结果中的目标词, 并获取该目标词与该关键词词袋中各个关键词的相似度值, 当该目标词与该关键词词袋中的第一关键词的相似度值大于相似度阈值, 则根据该目标词与该第一关键词的相似度值更新该第一元素集合中的第一元素, 得到第二元素集合, 该第二元素集合中的各个元素用于构建用于短文本分类的机器学习模型。采用本申请实施例, 可以提高利用元素集合构建的机器学习模型的性能。

一种样本数据处理方法及装置

本申请要求于 2018 年 11 月 26 日提交中国专利局、申请号为 2018114211608、申请名称为“一种样本数据处理方法及装置”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请涉及计算机技术领域，尤其涉及一种样本数据处理方法及装置。

背景技术

随着微博、社交网站和热线电话等应用的发展，越来越多的信息开始以短文本的形式呈现，并且呈爆炸式增长。文本分类可以帮助人们快速有效的从海量数据中获取关键信息，而文本分类的准确性依赖于机器学习模型的性能，机器学习模型的性能又依赖于样本数据。

现有的文本样本数据处理方法大多采用基于关键词词袋（Bag of Words）模型的方法，该方法用在长文本中通常能取得较好的效果，但是用在短文本中常常效果不佳、质量低。主要原因是，相较于长文本，短文本具有特征稀疏、主题不明确的特点。首先，由于短文本长度的限制，其特征词很少，而用关键词词袋模型生成的样本数据维度大，这增加了文本处理的难度。其次，在长文本中，跟主题相关的词通常会大量出现，可以由此来判断整篇文章的主要内容；而在短文本中则不能根据词频来判断主要内容。比如短文本“咨询羽毛球主题的餐厅中”，“羽毛球”和“餐厅”的词频相同，但显然该文本的主题是“餐厅”，在文本分类时应被分到“餐饮”这一类别而不是“运动”类别。由此可见，现有的样本数据处理方法不能很好的表示短文本。

综上，因为短文本具有上述的特征稀疏和主题不明确的特点，所以，利用现有的样本数据处理方法得到的样本数据构建用于文本分类的机器学习模型的性能差，文本分类准确性低。

发明内容

本申请实施例提供一种样本数据处理方法及装置，可以增加短文本样本数据的信息量，提高利用该短文本样本数据构建的机器学习模型的性能，进而提高短文本分类的精确度。

第一方面，本申请实施例提供了一种样本数据处理方法，该方法包括：

获取短文本样本进行分词后得到的分词结果，并获取包含 N 个关键词的关键词词袋，该分词结果中包括至少一个词；

根据该分词结果与该关键词词袋确定出第一元素集合，该第一元素集合中包括 N 个元素，该第一元素集合中各个元素的值为该关键词词袋中各个关键词在该分词结果中出现的次数；

获取该分词结果中的目标词，并获取该目标词与该关键词词袋中各个关键词的相似度值，该目标词包括在该分词结果中存在且在该关键词词袋中不存在的词；

若该目标词与该关键词词袋中的第一关键词的相似度值大于相似度阈值，则根据该目标词与该第一关键词的相似度值更新该第一元素集合中该第一关键词对应的第一元素，得

到第二元素集合；

其中，该第二元素集合中的各个元素用于构建用于短文本分类的机器学习模型。

第二方面，本申请实施例提供了一种样本数据处理装置，该装置包括：

第一获取模块，用于获取短文本样本进行分词后得到的分词结果，并获取包含 N 个关键词的关键词词袋，该分词结果中包括至少一个词；

确定模块，用于根据该分词结果与该关键词词袋确定出第一元素集合，该第一元素集合中包括 N 个元素，该第一元素集合中各个元素的值为该关键词词袋中各个关键词在该分词结果中出现的次数；

第二获取模块，用于获取该分词结果中的目标词，并获取该目标词与该关键词词袋中各个关键词的相似度值，该目标词包括在该分词结果中存在且在该关键词词袋中不存在的词；

更新模块，用于当该目标词与该关键词词袋中的第一关键词的相似度值大于相似度阈值时，根据该目标词与该第一关键词的相似度值更新该第一元素集合中该第一关键词对应的第一元素，得到第二元素集合；

其中，该第二元素集合中的各个元素用于构建用于短文本分类的机器学习模型。

第三方面，本申请实施例提供了一种终端，包括处理器和存储器，该处理器和存储器相互连接，其中，该存储器用于存储支持终端执行上述方法的计算机程序，该计算机程序包括程序指令，该处理器被配置用于调用该程序指令，执行上述第一方面的样本数据处理方法。

第四方面，本申请实施例提供了一种计算机可读存储介质，该计算机可读存储介质存储有计算机程序，该计算机程序包括程序指令，该程序指令当被处理器执行时使该处理器执行上述第一方面的样本数据处理方法。

本申请实施例通过更新短文本关键词的元素集合中的元素，可以增加短文本样本的元素集合中的信息量，提高利用该短文本样本的元素集合构建的机器学习模型的性能，进而提高短文本分类的精确度。

附图说明

图 1 是本申请实施例提供的样本数据处理方法的一示意图；

图 2 是本申请实施例提供的关键词词袋生成方法的一示意图；

图 3 是本申请实施例提供的关键词“篮球”返回的词条的示意图；

图 4 是本申请实施例提供的样本数据处理方法的另一示意图；

图 5 是本申请实施例提供的样本数据处理装置的一示意性框图；

图 6 是本申请实施例提供的终端的一示意性框图。

具体实施方式

下面将结合图 1 至图 6，对本申请实施例提供的样本数据处理方法及装置进行说明。

参见图 1，是本申请实施例提供的样本数据处理方法的一示意图。如图 1 所示，该样本数据处理方法可包括步骤：

S101，终端获取短文本样本进行分词后得到的分词结果，并获取包含 N 个关键词的关

键词词袋。

在一些可行的实施方式中，终端可以获取用户输入的短文本样本，并可以采用基于字符串匹配的分词方法或基于统计的机器学习算法等对该短文本样本进行分词，得到分词后的分词结果。该分词结果中可以包括至少一个词。终端可以从关键词词袋数据库中获取任意一个包含N个关键词的关键词词袋。N可以为大于1的整数。该关键词词袋数据库中包括预设的多个类别的关键词词袋。该关键词词袋数据库中的各个关键词词袋均包括N个关键词。该关键词词袋可以用于表示文本中忽略词序、语法和句法的词的集合。例如，短文本样本为“今天天气怎么样？”，经过分词后的分词结果为“今天/天气/怎么样”这三个词。

其中，短文本样本的信息长度都比较短，通常在200字以内，因此所包含的有效信息也非常少，通常具有稀疏性、实时性、不规则性等特点。基于字符串匹配的分词方法可以包括正向最大匹配法、逆向最大匹配法、双向匹配分词法等，基于统计的机器学习算法可以包括隐马尔可夫模型(hidden markov model, HMM)、条件随机场(conditional random fields, CRF)等。

在一些可行的实施方式中，终端获取关键词词袋的方式可以参见图2。图2为本申请实施例提供的关键词词袋生成方法的一示意图。如图2所示，本申请实施例的关键词词袋生成方法可包括步骤：

S1011，获取用于生成关键词词袋的训练样本集。

在一些可行的实施方式中，终端可以在百度知道上输入预先确定的不同类别的关键词进行搜索，并可以爬取该不同类别的关键词返回的标题，如搜索天气类关键词“天气”、“小雨”等；运动类关键词“篮球”、“棒球”、“跑步”等。终端可以将百度知道上爬取到的标题进行人工分类和标注，得到各个类别对应的训练样本。终端可以将各个类别对应的训练样本加入训练样本的集合中，形成训练样本集。其中，一个类别可以对应一个训练样本，该训练样本中可以包括同一类别的多个标题（样本）。

例如，如图3所示，是关键词“篮球”返回的词条的示意图。其中，篮球为运动类的关键词之一，在百度知道上输入关键词“篮球”进行搜索，终端爬取关键词“篮球”返回的前50条标题。如图3所示的标题“一个篮球的直径有多大”、“篮球的规格？”、“篮球场的标准尺寸是多少？”等等。

S1012，根据词频-逆文档频率算法从训练样本集中确定出N个关键词，根据N个关键词生成关键词词袋。

在一些可行的实施方式中，终端可以利用分词工具（如jieba分词、stanfordCoreNlp分词等）对上述训练样本集的每个训练样本中的各个样本（即标题）进行分词。针对每个类别的训练样本，终端可以利用词频-逆文档频率(term frequency-inverse document frequency, TF-IDF)算法计算该训练样本中各个样本进行分词后得到的各个词的TF-IDF值，并可以将计算出的该类别的TF-IDF值按从大到小排序，得到TF-IDF序列，并可以从该TF-IDF序列中提取出前N个TF-IDF值对应的N个词，再可以对该N个词进行人工增删得到包含N个关键词的关键词词袋。N为大于1的整数。终端可以预设一个关键词词袋数据库，该关键词词袋数据库可以用于存储至少一个类别的关键词词袋。该关键词词袋数据库中的各

个关键词词袋均包括 N 个关键词。

其中，TF-IDF 的值等于 TF_w 与 IDF_w 的乘积， TF_w 和 IDF_w 分别满足：

$$TF_w = \frac{\text{某一类中词条 } w \text{ 出现的次数}}{\text{该类中所有的词条数目}},$$

$$IDF_w = \log\left(\frac{\text{语料库的文档总数}}{\text{包含词条 } w \text{ 的文档数} + 1}\right),$$

词条 w 可以用于表示训练样本中各样本进行分词后得到的任一词，语料库可以用于表示所有训练样本的集合，即训练样本集，文档数可以用于表示训练样本的个数，文档数可以与训练样本的类别数相等。需要说明的是，本申请实施例中语料库的文档总数大于 2，即训练样本集中至少有 3 个类别的训练样本。

例如，假设爬取到“天气、信用卡、机票”三个类别的训练样本。N=100。由于每个类别的关键词词袋生成方式一致，故以生成“天气”类别的关键词词袋为例。首先终端利用 jieba 分词工具或 stanfordCoreNlp 分词工具将“天气”类别中的所有样本（标题）进行分词，再根据 TF-IDF 的公式计算“天气”类别的样本进行分词后得到的各个词的 TF-IDF 值，将计算出的“天气”类别的 TF-IDF 值按从大到小排序，并筛选出“天气”类别中 TF-IDF 值在前 100 的词语作为“天气”类别的关键词词袋。其中，TF-IDF 中的语料库由“天气”、“信用卡”和“机票”三个类别的训练样本组成，词条 w 为“天气”类别的训练样本进行分词后得到的任一词，文档为训练样本。此时 IDF_w 中语料库的文档总数为 3，分别为“天气”、“信用卡”和“机票”三个类别的训练样本，包含词条 w 的文档数为 1，即“天气”类别的训练样本。

由于传统关键词词袋中的关键词较多，比如一个关键词词袋中包括 1000 个关键词，而短文本长度较短（通常在 200 字以内），因此对于短文本来说，用传统的关键词词袋生成的向量维度过大。本申请实施例中的终端利用 TF-IDF 算法构建了较小的关键词词袋（包括 100 个关键词），极大的避免了在利用关键词词袋生成向量时维度过大的问题。

S102，终端根据分词结果与关键词词袋确定出第一元素集合。

在一些可行的实施方式中，终端可以统计上述关键词词袋中各个关键词在上述短文本样本的分词结果中出现的次数。终端可以将该关键词词袋中各个关键词在该分词结果中出现的次数作为第一元素集合中的元素。其中，该第一元素集合中可以包括 N 个元素。

在一些可行的实施方式中，上述第一元素集合可以为第一向量，该第一向量中可以包括 N 个元素。该第一向量满足：

$$V_1 = [C_{w_1}, C_{w_2}, \dots, C_{w_n}]$$

其中， V_1 表示第一向量， C_{w_n} 表示关键词词袋中的第 n 个词在短文本样本的分词结果中出现的次数，n 的取值范围为 1 到 N 的自然数。

例如，假设关键词词袋中有 5 个关键词，分别是“篮球/天气/足球/跑步/今天”，那么第一向量 V_1 中包括 5 个元素。假设关键词词袋中的第二个关键词“天气”和第五个关键词“今天”在短文本样本的分词结果中各出现一次，终端确定第一向量 V_1 中的第二个元素和第五个元素为 1。而关键词词袋中的其他关键词（即第一个关键词“篮球”、第三个关键词“足球”以及第四个关键词“跑步”）在短文本样本的分词结果中出现的次数为 0，终端确定第

一向量 V_1 中的第一个、第三个、第四个元素全为 0，此时第一向量 $V_1=[0,1,0,0,1]$ 。

S103，终端获取分词结果中的目标词，并获取目标词与关键词词袋中各个关键词的相似度值。

在一些可行的实施方式中，终端可以查找上述短文本样本的分词结果中是否存在目标词。若该分词结果中存在至少一个目标词，则终端可以获取该分词结果中的任一目标词。终端可以计算该目标词与上述关键词词袋中的各个关键词之间的相似度值，如余弦值、欧氏距离等。若该分词结果中不存在目标词，则终端可以将上述第一元素集合确定为第二元素集合。其中，该目标词可以用于表示在该分词结果中存在且在上述关键词词袋中不存在的词。

例如，假设短文本样本的分词结果中包括 5 个词，分别为 C1、C2、C3、C4 以及 C5。终端分别检测关键词词袋中是否存在与词 C1、C2、C3、C4、C5 相同的关键词。若关键词词袋中不存在与词 C2 和 C5 相同的关键词，则终端分别将词 C2 和 C5 确定为目标词。终端获取目标词 C2 和 C5 中任一目标词如 C2，终端计算目标词 C2 与关键词词袋中各个关键词之间的相似度值。

S104，若目标词与关键词词袋中的第一关键词的相似度值大于相似度阈值，则终端根据目标词与第一关键词的相似度值更新第一元素集合中第一关键词对应的第一元素，得到第二元素集合。

在一些可行的实施方式中，终端在获取到上述目标词与上述关键词词袋中各个关键词之间的相似度值之后，可以检测该目标词与该各个关键词之间的相似度值是否大于预设的相似度阈值。当检测到该目标词与该关键词袋中第一关键词之间的相似度值大于该相似度阈值时，说明该目标词与该第一关键词为同义词，则终端可以该目标词与该第一关键词的相似度值更新上述第一元素集合中该第一关键词对应的第一元素，得到第二元素集合。当检测到该目标词与该关键词袋中各个关键词之间的相似度值均小于或等于该相似度阈值时，说明该关键词词袋中没有该目标词的同义词，则终端可以获取上述分词结果中下一个目标词与该关键词词袋中各个关键词之间的相似度值。若上述分词结果中的各个目标词与该关键词词袋中各个关键词之间的相似度值均小于或等于该相似度阈值，说明该关键词词袋中没有任一目标词的同义词，则终端可以将上述第一元素集合确定为第二元素集合。该第二元素集合中的各个元素可以用于构建用于短文本分类的机器学习模型。

由于本申请实施例关键词词袋的关键词数量较少，从而导致在利用关键词词袋和短文本样本的分词结果确定第一元素集合时，短文本样本中的部分信息损失，即出现短文本样本的分词结果中存在的词不在关键词词袋中的情况，也即短文本样本中出现目标词的情况。又因为第一元素集合中的各元素为关键词词袋中各关键词在分词结果中出现的次数，那么分词结果中的目标词未在第一元素集合中体现。所以本申请实施例通过目标词与关键词词袋中各个关键词的相似度值，在关键词词袋中寻找目标词的同义词，更新目标词的同义词在第一元素集合中对应的元素，得到第二元素集合。从而增加了第一元素集合中的信息量，解决了因为关键词词袋的缩小而损失信息的问题，进而在利用第二元素集合构建用于短文本分类的传统机器学习模型时，可以得到性能较好的机器学习模型，使得短文本分类更准确。

在本申请实施例中，终端通过获取短文本样本进行分词后的分词结果，并获取包含 N 个关键词的关键词词袋，根据该分词结果与该关键词词袋确定出第一元素集合，获取该分词结果中的目标词，并获取该目标词与该关键词词袋中各个关键词的相似度值，当该目标词与该关键词词袋中的第一关键词的相似度值大于相似度阈值，则根据该目标词与该第一关键词的相似度值更新该第一元素集合中该第一关键词对应的第一元素，得到第二元素集合，该第二元素集合中的各个元素用于构建用于短文本分类的机器学习模型，可以增加短文本样本的元素集合中的信息量，提高利用该短文本样本的元素集合构建的机器学习模型的性能，进而提高短文本分类的精确度。

参见图 4，是本申请实施例提供的样本数据处理方法的另一示意图。如图 4 所示，该样本数据处理方法可包括步骤：

S401，终端获取短文本样本进行分词后得到的分词结果，并获取包含 N 个关键词的关键词词袋。

S402，终端根据分词结果与关键词词袋确定出第一元素集合。

在一些可行的实施方式中，本申请实施例中步骤 S401-步骤 S402 的实现方式可以参考图 1 所示实施例中的步骤 S101-步骤 S102 的实现方式，在此不再赘述。

S403，终端获取分词结果中的目标词。

S404，终端从词向量数据库中获取目标词的词向量，并从词向量数据库中获取关键词词袋中各个关键词的词向量。

S405，终端计算目标词的词向量与各个关键词的词向量之间的相似度值。

S406，若目标词的词向量与关键词词袋中第一关键词的词向量之间的相似度值大于相似度阈值，则终端计算相似度值与目标词在分词结果中出现的次数的乘积。

S407，终端将第一元素集合中第一关键词对应的第一元素更新为乘积与第一元素之和，得到第二元素集合。

在一些可行的实施方式中，终端可以查找上述短文本样本的分词结果中是否存在目标词。若该分词结果中存在至少一个目标词，则终端可以获取该分词结果中的任一目标词。终端可以从词向量数据库中获取该目标词的词向量，并可以从该词向量数据库中获取该关键词词袋中各个关键词的词向量。终端可以计算该目标词的词向量与该各个关键词的词向量之间的相似度值（如余弦值、欧氏距离等）。终端可以检测该目标词的词向量与该各个关键词的词向量之间的相似度值是否大于预设的相似度阈值。当检测到该目标词的词向量与该关键词袋中第一关键词的词向量之间的相似度值大于该相似度阈值时，说明该目标词与该第一关键词为同义词，则终端可以计算该相似度值与该目标词在上述短文本样本的分词结果中出现的次数的乘积。终端可以将上述第一元素集合中该第一关键词对应的第一元素更新为该乘积与该第一元素之和，得到第二元素集合。其中，该目标词可以用于表示在该分词结果中存在且在上述关键词词袋中不存在的词。

在一些可行的实施方式中，若该分词结果中不存在目标词，则终端可以将上述第一元素集合确定为第二元素集合。或者，当检测到该目标词的词向量与该关键词袋中各个关键词的词向量之间的相似度值均小于或等于该相似度阈值时，说明该关键词词袋中没有该目

标词的同义词,则终端可以获取上述分词结果中下一个目标词的词向量与该关键词词袋中各个关键词的词向量之间的相似度值。若上述分词结果中各个目标词的词向量与该关键词词袋中各个关键词的词向量之间的相似度值均小于或等于该相似度阈值,说明该关键词词袋中没有任一目标词的同义词,则终端可以将上述第一元素集合确定为第二元素集合。

在一些可行的实施方式中,词向量数据库的生成方法可以为:(1)终端从维基百科中爬取大规模无标注语料(大概在10个G左右),并对这些无标注语料进行分词,将分词后的这些语料输入连续关键词词袋模型(continuous bag of words, CBOW)中进行训练。(2)在CBOW模型训练结束后,终端可以获取CBOW模型输出的这些语料中所有词的词向量,并可以将这些语料中所有词的词向量存储到词向量数据库中。

在一些可行的实施方式中,上述第二元素集合可以为第二向量,该第二向量中可以包括N个元素。该第二向量满足:

$$V_2 = [C_{w_1}, C_{w_2}, \dots, C_{w_k} + C_{w_j} * \cos(w_k, w_j), \dots, C_{w_n}]$$

其中, V_2 表示第二向量, w_j 表示短文本样本的分词结果中的目标词, w_k 表示关键词词袋中的第k个关键词为第一关键词, $\cos(w_k, w_j)$ 表示目标词 w_j 的词向量与关键词词袋中第一关键词 w_k 的词向量之间的余弦值, $\cos(w_k, w_j)$ 大于预设的相似度阈值如0.7,即 w_j 与 w_k 是同义词。终端将第一向量 V_1 中的第一元素 C_{w_k} 更新为 $C_{w_k} + C_{w_j} * \cos(w_k, w_j)$ 。 C_{w_j} 表示目标词在短文本样本的分词结果中出现的次数, C_{w_k} 表示关键词词袋中的第一关键词在短文本样本的分词结果中出现的次数。

例如,假设短文本样本的分词结果中有2个目标词,分别为目标词 w_{j1} 和目标词 w_{j2} 。假设关键词词袋中有10个关键词,分别为 $w_1, w_2, w_3, \dots, w_{10}$ 。预设的相似度阈值为0.7。终端分别计算目标词 w_{j1} 的词向量与关键词词袋中10个关键词 $w_1, w_2, w_3, \dots, w_{10}$ 的词向量之间的余弦值。假设目标词 w_{j1} 的词向量与第四个关键词 w_4 的词向量之间的余弦值 $\cos(w_4, w_{j1}) = 0.8$ 大于预设的相似度阈值0.7,终端计算余弦值 $\cos(w_4, w_{j1})$ 与目标词 w_{j1} 在分词结果中出现的次数 $C_{w_{j1}}$ 的乘积 $\cos(w_4, w_{j1}) * C_{w_{j1}}$ 。终端将第一向量 V_1 中的第四个元素 C_{w_4} 更新为第四个元素 C_{w_4} 与乘积 $\cos(w_4, w_{j1}) * C_{w_{j1}}$ 之和,即 $C_{w_4} + \cos(w_4, w_{j1}) * C_{w_{j1}}$ 。终端分别计算目标词 w_{j2} 的词向量与关键词词袋中10个关键词 $w_1, w_2, w_3, \dots, w_{10}$ 的词向量之间的余弦值。假设目标词 w_{j2} 的词向量与第三个关键词 w_3 、第六个关键词 w_6 、第七个关键词 w_7 的词向量之间的余弦值均大于相似度阈值0.7,即 $\cos(w_3, w_{j2})$ 、 $\cos(w_6, w_{j2})$ 和 $\cos(w_7, w_{j2})$ 的值均大于0.7。终端分别更新第一向量 V_1 中的第三个元素 C_{w_3} 、第六个元素 C_{w_6} 和第七个元素 C_{w_7} 。

S408,终端获取第二元素集合中M个0元素对应的M个第二关键词。

在一些可行的实施方式中,终端可以检测上述第二元素集合中是否存在0元素。若该第二元素集合中存在0元素,则终端可以获取上述第二元素集合中的M个0元素,并可以将该第二元素集合中各个0元素对应的关键词确定为第二关键词,从而得到M个第二关键词。若该第二元素集合中不存在任何0元素,则终端可以将上述第二元素集合确定为第三元素集合。其中,M可以为大于或等于1的整数。第二元素集合中的0元素可以用于表示上述关键词词袋中的某个关键词在上述分词结果中出现的次数为0且该关键词的同义词也在上述分词结果中出现的次数为0,即第二关键词未在该分词结果中出现过且该第二关键词的同义词也未在该分词结果中出现过。

在一些可行的实施方式中,终端可以从上述关键词词袋中任取一个关键词,并可以检测该关键词词袋中的关键词是否取完。若该检测该关键词词袋中的关键词未取完,检测该关键词在上述分词结果中出现的次数是否为 0,当该关键词在上述分词结果中出现的次数为 0 时,计算该关键词与该分词结果中各个词之间的相似度值,当该关键词与该分词结果中各个词之间的相似度值均小于或等于预设的相似度阈值时,说明该分词结果中没有该关键词的同义词,则可以将该关键词确定为第二关键词。若该关键词词袋中不存在第二关键词,则可以将上述第二元素集合确定为第三元素集合。

S409,终端获取关键词词袋中的至少一个第三关键词。

S410,根据 M 个第二关键词中各个第二关键词与关键词词袋中各个第三关键词之间的相似度值替换第二元素集合中的 0 元素,得到第三元素集合。

在一些可行的实施方式中,终端可以从上述关键词词袋中获取至少一个第三关键词,该第三关键词在上述分词结果中出现的次数大于或等于 1。终端可以计算上述 M 个第二关键词中各个第二关键词与上述关键词词袋中各个第三关键词之间的相似度值(如余弦值、欧氏距离等),并可以根据该相似度值替换上述第二元素中的 0 元素,得到第三元素集合。该第三元素集合中的各个元素用于构建用于短文本分类的机器学习模型。因为机器学习模型的主要原理是在元素集合中的每个元素上乘以一个权重,最后输出该元素集合属于不同类别的概率。若元素集合中存在 0 元素,那么 0 元素与任意值的权重相乘,仍会得到 0,这时机器学习模型输出的概率就可能为 0,从而导致存在 0 元素的元素集合构建的用于短文本分类的机器学习模型精确度不高。所以,本申请实施例通过替换掉第二元素集合中的 0 元素,即第三元素集合中不存在 0 元素,可以在保证元素集合信息量的同时提高机器学习模型的准确度,从而提高短文本分类的精确度。

在一些可行的实施方式中,终端在获取到至少一个第三关键词之后,可以从上述词向量数据库中获取上述 M 个第二关键词的各个第二关键词的词向量,并可以从该词向量数据库中获取上述关键词词袋中至少一个第三关键词的各个第三关键词的词向量。针对该 M 个第二关键词中任一第二关键词 i,终端可以获取该第二关键词 i 的词向量与该至少一个第三关键词中各个第三关键词的词向量之间的相似度值(如余弦值、欧氏距离等)。若该第二关键词 i 的词向量与该至少一个第三关键词中的第三关键词 m 之间的相似度值最大,则终端可以获取该第三关键词 m 在上述分词结果中出现的次数。终端可以计算该第二关键词 i 的词向量与该第三关键词 m 之间的相似度值与该次数的乘积,并可以将上述第二元素集合中该第二关键词 i 对应的 0 元素替换为该乘积。终端将该第二元素集合中该 M 个第二关键词对应的 M 个 0 元素全部替换后可以得到第三元素集合。该第三元素集合中的各个元素用于构建用于短文本分类的机器学习模型。

在一些可行的实施方式中,上述第三元素集合可以为第三向量,该第三向量中可以包括 N 个元素。该第三向量满足:

$$V_3 = [C_{w_1}, C_{w_2}, \dots, [\cos(w_m, w_i)]_{\max} * C_{w_m}, \dots, C_{w_n}]$$

其中, V_3 表示第三向量, w_m 表示关键词词袋中的第 m 个关键词为第三关键词, C_{w_m} 表示关键词词袋中第 m 个关键词在短文本样本的分词结果中出现的次数, $C_{w_m} \geq 1$, 所述 w_i 表示关键词词袋中的第 i 个关键词为第二关键词, $[\cos(w_m, w_i)]_{\max}$ 表示关键词词袋中的第 m

个关键词与关键词词袋中的第 i 个关键词的相似度值最大。

例如, 假设第二元素集合中的第 5 个元素为 0, 第二元素集合中第 5 个元素对应关键词为 w_5 (即第二关键词)。假设关键词词袋中存在 3 个第三关键词, 分别为 w_3 、 w_6 以及 w_{12} 。终端分别计算第二关键词 w_5 的词向量与各个第三关键词 w_3 、 w_6 以及 w_{12} 的词向量之间的余弦值。假设 $\cos(w_3, w_5)=0.3$ 、 $\cos(w_6, w_5)=0.7$ 、 $\cos(w_{12}, w_5)=0.5$, 由于第二关键词 w_5 与第三关键词 w_6 之间的余弦值最大, 则终端获取第三关键词在分词结果中出现的次数 C_{w_6} , 其中 $C_{w_6} \geq 1$ 。终端计算余弦值 $\cos(w_6, w_5)$ 与该次数 C_{w_6} 的乘积 $\cos(w_6, w_5) * C_{w_6}$ 。终端将第二元素集合中的第 5 个元素 $C_{w_5}=0$ 替换为该乘积 $\cos(w_6, w_5) * C_{w_6}$ 。

在本申请实施例中, 终端通过计算短文本样本的分词结果中目标词的词向量与关键词词袋中各个关键词的词向量之间的相似度值, 来对第一元素集合进行更新, 得到第二元素集合。再根据关键词词袋中第二关键词的词向量与关键词词袋中第三关键词的词向量之间的相似度值来替换第二元素集合中的 0 元素, 得到第三元素集合。该第三元素集合中的各个元素用于构建用于短文本分类的机器学习模型。可以在保证元素集合信息量的同时提高机器学习模型的准确度, 从而提高短文本分类的精确度。

参见图 5, 是本申请实施例提供的样本数据处理装置的一示意性框图。如图 5 所示, 本申请实施例中的样本数据处理装置包括:

第一获取模块 10, 用于获取短文本样本进行分词后得到的分词结果, 并获取包含 N 个关键词的关键词词袋, 该分词结果中包括至少一个词;

确定模块 20, 用于根据该分词结果与该关键词词袋确定出第一元素集合, 该第一元素集合中包括 N 个元素, 该第一元素集合中各个元素的值为该关键词词袋中各个关键词在该分词结果中出现的次数;

第二获取模块 30, 用于获取该分词结果中的目标词, 并获取该目标词与该关键词词袋中各个关键词的相似度值, 该目标词包括在该分词结果中存在且在该关键词词袋中不存在的词;

更新模块 40, 用于当该目标词与该关键词词袋中的第一关键词的相似度值大于相似度阈值时, 根据该目标词与该第一关键词的相似度值更新该第一元素集合中该第一关键词对应的第一元素, 得到第二元素集合;

其中, 该第二元素集合中的各个元素用于构建用于短文本分类的机器学习模型。

在一些可行的实施方式中, 该第一获取模块 10 中包括第一获取单元 101 和第二获取单元 102。该第一获取单元 101 用于获取短文本样本进行分词后得到的分词结果, 该分词结果中包括至少一个词; 该第二获取单元 102 用于获取包含 N 个关键词的关键词词袋。

其中, 该第二获取单元 102 具体用于: 获取用于生成关键词词袋的训练样本集; 根据词频-逆文档频率算法从该训练样本集中确定出 N 个关键词, 根据该 N 个关键词生成该关键词词袋。

在一些可行的实施方式中, 该第一元素集合为第一向量, 该第一向量中包括 N 个元素, 该第一向量满足:

$$V_1 = [C_{w_1}, C_{w_2}, \dots, C_{w_n}];$$

其中, 该 V_1 表示该第一向量, 该 C_{w_n} 表示关键词词袋中的第 n 个关键词在该分词结果中出现的次数, 该 n 的取值范围为 1 到 N 的自然数。

在一些可行的实施方式中, 该第二获取模块 30 包括第三获取单元 301 和第四获取单元 302。该第三获取单元 301 用于获取该分词结果中的目标词, 该目标词包括在该分词结果中存在且在该关键词词袋中不存在的词; 该第四获取单元 302 用于获取该目标词与该关键词词袋中各个关键词的相似度值。

其中, 该第四获取单元 302 具体用于: 从词向量数据库中获取该目标词的词向量, 并从该词向量数据库中获取该关键词词袋中各个关键词的词向量; 计算该目标词的词向量与该各个关键词的词向量之间的相似度值。该更新模块 40 具体用于: 若该目标词的词向量与该关键词词袋中第一关键词的词向量之间的相似度值大于相似度阈值, 则计算该相似度值与该目标词在该分词结果中出现的次数的乘积; 将该第一元素集合中该第一关键词对应的第一元素更新为该乘积与该第一元素之和, 得到第二元素集合。

在一些可行的实施方式中, 该装置还包括第三获取模块 50、第四获取模块 60 以及替换模块 70。

该第三获取模块 50, 用于获取该第二元素集合中 M 个 0 元素对应的 M 个第二关键词; 该第四获取模块 60, 用于获取该关键词词袋中的至少一个第三关键词, 该第三关键词在该分词结果中出现的次数大于或等于 1; 该替换模块 70, 用于根据该 M 个第二关键词中各个第二关键词与该关键词词袋中各个第三关键词之间的相似度值替换该第二元素集合中的 0 元素, 得到第三元素集合。其中, 该第三元素集合中的各个元素用于构建用于短文本分类的机器学习模型。

在一些可行的实施方式中, 该替换模块 70 具体用于:

从词向量数据库中获取该 M 个第二关键词中各个第二关键词的词向量, 并从该词向量数据库中获取该关键词词袋中至少一个第三关键词的各个第三关键词的词向量; 获取该 M 个第二关键词中任一第二关键词 i 的词向量与该至少一个第三关键词中各个第三关键词的词向量之间的相似度值; 若该第二关键词 i 的词向量与该至少一个第三关键词中的第三关键词 m 的词向量之间的相似度值最大, 则将该第二元素集合中该第二关键词 i 对应的 0 元素替换为该第二关键词 i 与该第三关键词 m 之间的相似度值与该第三关键词 m 在该分词结果中出现的次数的乘积, 得到第三元素集合。

在一些可行的实施方式中, 该第三元素集合为第三向量, 该第三向量中包括 N 个元素, 该第三向量满足:

$$V_3 = [C_{w_1}, C_{w_2}, \dots, [\cos(w_m, w_i)]_{\max} * C_{w_m}, \dots, C_{w_n}];$$

其中, 该 V_3 表示该第三向量, 该 w_m 表示关键词词袋中的第 m 个关键词为第三关键词, C_{w_m} 表示关键词词袋中第 m 个关键词在该分词结果中的出现次数, $C_{w_m} \geq 1$, 该 w_i 表示关键词词袋中的第 i 个关键词为第二关键词, $[\cos(w_m, w_i)]_{\max}$ 表示关键词词袋中的第 m 个关键词与关键词词袋中的第 i 个关键词的相似度值最大。

具体实现中, 上述样本数据处理装置可通过上述各个模块执行上述图 1、图 2 或图 4 所提供的实现方式中各个步骤所提供的实现方式, 实现上述各实施例中所实现的功能, 具体可参见上述图 1、图 2 或图 4 所示的方法实施例各个步骤提供的相应描述, 在此不再

赘述。

在本申请实施例中，样本数据处理装置通过获取短文本样本进行分词后的分词结果，并获取包含 N 个关键词的关键词词袋，根据该分词结果与该关键词词袋确定出第一元素集合，获取该分词结果中的目标词，并获取该目标词与该关键词词袋中各个关键词的相似度值，当该目标词与该关键词词袋中的第一关键词的相似度值大于相似度阈值，则根据该目标词与该第一关键词的相似度值更新该第一元素集合中该第一关键词对应的第一元素，得到第二元素集合，该第二元素集合中的各个元素用于构建用于短文本分类的机器学习模型，可以增加短文本样本的元素集合中的信息量，提高利用该短文本样本的元素集合构建的机器学习模型的性能，进而提高短文本分类的精确度。

参见图 6，是本申请实施例提供的终端的一示意性框图。如图 6 所示，本申请实施例中的终端可以包括：一个或多个处理器 601 和存储器 602。上述处理器 601 和存储器 602 通过总线 603 连接。存储器 602 用于存储计算机程序，所述计算机程序包括程序指令，处理器 601 用于执行存储器 602 存储的程序指令。其中，处理器 601 被配置用于调用所述程序指令执行：

获取短文本样本进行分词后得到的分词结果，并获取包含 N 个关键词的关键词词袋，该分词结果中包括至少一个词；

根据该分词结果与该关键词词袋确定出第一元素集合，该第一元素集合中包括 N 个元素，该第一元素集合中各个元素的值为该关键词词袋中各个关键词在该分词结果中出现的次数；

获取该分词结果中的目标词，并获取该目标词与该关键词词袋中各个关键词的相似度值，该目标词包括在该分词结果中存在且在该关键词词袋中不存在的词；

若该目标词与该关键词词袋中的第一关键词的相似度值大于相似度阈值，则根据该目标词与该第一关键词的相似度值更新该第一元素集合中该第一关键词对应的第一元素，得到第二元素集合；

其中，该第二元素集合中的各个元素用于构建用于短文本分类的机器学习模型。

应当理解，在本申请实施例中，所称处理器 601 可以是中央处理单元 (Central Processing Unit, CPU)，该处理器还可以是其他通用处理器、数字信号处理器 (Digital Signal Processor, DSP)、专用集成电路 (Application Specific Integrated Circuit, ASIC)、现成可编程门阵列 (Field-Programmable Gate Array, FPGA) 或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等。

该存储器 602 可以包括只读存储器和随机存取存储器，并向处理器 601 提供指令和数据。存储器 602 的一部分还可以包括非易失性随机存取存储器。例如，存储器 602 还可以存储设备类型的信息。

具体实现中，本申请实施例中所描述的处理器 601 可执行本申请实施例提供的样本数据处理方法的实现方式，也可执行本申请实施例提供的关键词词袋生成方法的实现方式，还可执行本申请实施例所描述的样本数据处理装置的实现方式，在此不再赘述。

本申请实施例还提供一种计算机可读存储介质，该计算机可读存储介质存储有计算机程序，该计算机程序包括程序指令，该程序指令被处理器执行时实现图 1、图 4 所示的样本数据处理方法或图 2 所示的关键词词袋生成方法，具体细节请参照图 1、图 2 或图 4 所示实施例的描述，在此不再赘述。

上述计算机可读存储介质可以是前述任一实施例所述的样本数据处理装置或终端的内部存储单元，例如终端的硬盘或内存。该计算机可读存储介质也可以是该终端的外部存储设备，例如该终端上配备的插接式硬盘，智能存储卡（smart media card, SMC），安全数字（secure digital, SD）卡，闪存卡（flash card）等。进一步地，该计算机可读存储介质还可以既包括该终端的内部存储单元也包括外部存储设备。该计算机可读存储介质用于存储该计算机程序以及该电子设备所需的其他程序和数据。该计算机可读存储介质还可以用于暂时地存储已经输出或者将要输出的数据。

以上所述，仅为本申请的具体实施方式，但本申请的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本申请揭露的技术范围内，可轻易想到变化或替换，都应涵盖在本申请的保护范围之内。因此，本申请的保护范围应以所述权利要求的保护范围为准。

权利要求

1、一种样本数据处理方法，其特征在于，包括：

获取短文本样本进行分词后得到的分词结果，并获取包含 N 个关键词的关键词词袋，所述分词结果中包括至少一个词；

根据所述分词结果与所述关键词词袋确定出第一元素集合，所述第一元素集合中包括 N 个元素，所述第一元素集合中各个元素的值为所述关键词词袋中各个关键词在所述分词结果中出现的次数；

获取所述分词结果中的目标词，并获取所述目标词与所述关键词词袋中各个关键词的相似度值，所述目标词包括在所述分词结果中存在且在所述关键词词袋中不存在的词；

若所述目标词与所述关键词词袋中的第一关键词的相似度值大于相似度阈值，则根据所述目标词与所述第一关键词的相似度值更新所述第一元素集合中所述第一关键词对应的第一元素，得到第二元素集合；

其中，所述第二元素集合中的各个元素用于构建用于短文本分类的机器学习模型。

2、根据权利要求 1 所述的方法，其特征在于，所述获取包含 N 个关键词的关键词词袋，包括：

获取用于生成关键词词袋的训练样本集；

根据词频-逆文档频率算法从所述训练样本集中确定出 N 个关键词，根据所述 N 个关键词生成所述关键词词袋。

3、根据权利要求 1 或 2 所述的方法，其特征在于，所述第一元素集合为第一向量，所述第一向量中包括 N 个元素，所述第一向量满足：

$$V_1 = [C_{w_1}, C_{w_2}, \dots, C_{w_n}];$$

其中，所述 V_1 表示所述第一向量，所述 C_{w_n} 表示关键词词袋中的第 n 个关键词在所述分词结果中出现的次数，所述 n 的取值范围为 1 到 N 的自然数。

4、根据权利要求 1-3 任意一项所述的方法，其特征在于，所述获取所述目标词与所述关键词词袋中各个关键词的相似度值，包括：

从词向量数据库中获取所述目标词的词向量，并从所述词向量数据库中获取所述关键词词袋中各个关键词的词向量；

计算所述目标词的词向量与所述各个关键词的词向量之间的相似度值；

所述若所述目标词与所述关键词词袋中的第一关键词的相似度值大于相似度阈值，则根据所述目标词与所述第一关键词的相似度值更新所述第一元素集合中所述第一关键词对应的第一元素，得到第二元素集合，包括：

若所述目标词的词向量与所述关键词词袋中第一关键词的词向量之间的相似度值大于相似度阈值，则计算所述相似度值与所述目标词在所述分词结果中出现的次数的乘积；

将所述第一元素集合中所述第一关键词对应的第一元素更新为所述乘积与所述第一元素之和，得到第二元素集合。

5、根据权利要求 1-4 任意一项所述的方法，其特征在于，所述在得到所述第二元素集合之后，所述方法还包括：

获取所述第二元素集合中 M 个 0 元素对应的 M 个第二关键词；

获取所述关键词词袋中的至少一个第三关键词，所述第三关键词在所述分词结果中出现的次数大于或等于 1；

根据所述 M 个第二关键词中各个第二关键词与所述关键词词袋中各个第三关键词之间的相似度值替换所述第二元素集合中的 0 元素，得到第三元素集合；

其中，所述第三元素集合中的各个元素用于构建用于短文本分类的机器学习模型。

6、根据权利要求 5 所述的方法，其特征在于，所述根据所述 M 个第二关键词中各个第二关键词与所述关键词词袋中各个第三关键词之间的相似度值替换所述第二元素集合中的 0 元素，得到第三元素集合，包括：

从词向量数据库中获取所述 M 个第二关键词中各个第二关键词的词向量，并从所述词向量数据库中获取所述关键词词袋中至少一个第三关键词的各个第三关键词的词向量；

获取所述 M 个第二关键词中任一第二关键词 i 的词向量与所述至少一个第三关键词中各个第三关键词的词向量之间的相似度值；

若所述第二关键词 i 的词向量与所述至少一个第三关键词中的第三关键词 m 的词向量之间的相似度值最大，则将所述第二元素集合中所述第二关键词 i 对应的 0 元素替换为所述第二关键词 i 与所述第三关键词 m 之间的相似度值与所述第三关键词 m 在所述分词结果中出现的次数的乘积，得到第三元素集合。

7、根据权利要求 6 所述的方法，其特征在于，所述第三元素集合为第三向量，所述第三向量中包括 N 个元素，所述第三向量满足：

$$V_3 = [C_{w_1}, C_{w_2}, \dots, [\cos(w_m, w_i)]_{\max} * C_{w_m}, \dots, C_{w_n}];$$

其中，所述 V_3 表示所述第三向量，所述 w_m 表示关键词词袋中的第 m 个关键词为第三关键词， C_{w_m} 表示关键词词袋中第 m 个关键词在所述分词结果中的出现次数， $C_{w_m} \geq 1$ ，所述 w_i 表示关键词词袋中的第 i 个关键词为第二关键词， $[\cos(w_m, w_i)]_{\max}$ 表示关键词词袋中的第 m 个关键词与关键词词袋中的第 i 个关键词的相似度值最大。

8、一种样本数据处理装置，其特征在于，包括：

第一获取模块，用于获取短文本样本进行分词后得到的分词结果，并获取包含 N 个关键词的关键词词袋，所述分词结果中包括至少一个词；

确定模块，用于根据所述分词结果与所述关键词词袋确定出第一元素集合，所述第一元素集合中包括 N 个元素，所述第一元素集合中各个元素的值为所述关键词词袋中各个关键词在所述分词结果中出现的次数；

第二获取模块，用于获取所述分词结果中的目标词，并获取所述目标词与所述关键词词袋中各个关键词的相似度值，所述目标词包括在所述分词结果中存在且在所述关键词词袋中不存在的词；

更新模块，用于当所述目标词与所述关键词词袋中的第一关键词的相似度值大于相似度阈值时，根据所述目标词与所述第一关键词的相似度值更新所述第一元素集合中所述第一关键词对应的第一元素，得到第二元素集合；

其中，所述第二元素集合中的各个元素用于构建用于短文本分类的机器学习模型。

9、根据权利要求 8 所述的装置，其特征在于，所述第一获取模块包括：

第一获取单元,用于获取短文本样本进行分词后得到的分词结果,所述分词结果中包括至少一个词;

第二获取单元,用于获取用于生成关键词词袋的训练样本集,并根据词频-逆文档频率算法从所述训练样本集中确定出N个关键词,根据所述N个关键词生成所述关键词词袋。

10、根据权利要求8或9所述的装置,其特征在于,所述第一元素集合为第一向量,所述第一向量中包括N个元素,所述第一向量满足:

$$V_1 = [C_{w_1}, C_{w_2}, \dots, C_{w_n}];$$

其中,所述 V_1 表示所述第一向量,所述 C_{w_n} 表示关键词词袋中的第n个关键词在所述分词结果中出现的次数,所述n的取值范围为1到N的自然数。

11、根据权利要求8-10任意一项所述的装置,其特征在于,所述第二获取模块包括:

第三获取单元,用于获取所述分词结果中的目标词,所述目标词包括在所述分词结果中存在且在所述关键词词袋中不存在的词;

第四获取单元,用于从词向量数据库中获取所述目标词的词向量,并从所述词向量数据库中获取所述关键词词袋中各个关键词的词向量,计算所述目标词的词向量与所述各个关键词的词向量之间的相似度值;

所述更新模块具体用于:

若所述目标词的词向量与所述关键词词袋中第一关键词的词向量之间的相似度值大于相似度阈值,则计算所述相似度值与所述目标词在所述分词结果中出现的次数的乘积;

将所述第一元素集合中所述第一关键词对应的第一元素更新为所述乘积与所述第一元素之和,得到第二元素集合。

12、根据权利要求8-11任意一项所述的装置,其特征在于,所述装置还包括:

第三获取模块,用于获取所述第二元素集合中M个0元素对应的M个第二关键词;

第四获取模块,用于获取所述关键词词袋中的至少一个第三关键词,所述第三关键词在所述分词结果中出现的次数大于或等于1;

替换模块,用于根据所述M个第二关键词中各个第二关键词与所述关键词词袋中各个第三关键词之间的相似度值替换所述第二元素集合中的0元素,得到第三元素集合;

其中,所述第三元素集合中的各个元素用于构建用于短文本分类的机器学习模型。

13、根据权利要求12所述的装置,其特征在于,所述替换模块具体用于:

从词向量数据库中获取所述M个第二关键词中各个第二关键词的词向量,并从所述词向量数据库中获取所述关键词词袋中至少一个第三关键词的各个第三关键词的词向量;

获取所述M个第二关键词中任一第二关键词i的词向量与所述至少一个第三关键词中各个第三关键词的词向量之间的相似度值;

若所述第二关键词i的词向量与所述至少一个第三关键词中的第三关键词m的词向量之间的相似度值最大,则将所述第二元素集合中所述第二关键词i对应的0元素替换为所述第二关键词i与所述第三关键词m之间的相似度值与所述第三关键词m在所述分词结果中出现的次数的乘积,得到第三元素集合。

14、根据权利要求13所述的装置,其特征在于,所述第三元素集合为第三向量,所述第三向量中包括N个元素,所述第三向量满足:

$$V_3 = [C_{w_1}, C_{w_2}, \dots, [\cos(w_m, w_i)]_{\max} * C_{w_m}, \dots, C_{w_n}];$$

其中, 所述 V_3 表示所述第三向量, 所述 w_m 表示关键词词袋中的第 m 个关键词为第三关键词, C_{w_m} 表示关键词词袋中第 m 个关键词在所述分词结果中的出现次数, $C_{w_m} \geq 1$, 所述 w_i 表示关键词词袋中的第 i 个关键词为第二关键词, $[\cos(w_m, w_i)]_{\max}$ 表示关键词词袋中的第 m 个关键词与关键词词袋中的第 i 个关键词的相似度值最大。

15、一种终端, 其特征不在于, 包括处理器和存储器, 所述处理器和存储器相互连接, 其中, 所述存储器用于存储计算机程序, 所述计算机程序包括程序指令, 所述处理器被配置用于调用所述程序指令执行:

获取短文本样本进行分词后得到的分词结果, 并获取包含 N 个关键词的关键词词袋, 所述分词结果中包括至少一个词;

根据所述分词结果与所述关键词词袋确定出第一元素集合, 所述第一元素集合中包括 N 个元素, 所述第一元素集合中各个元素的值为所述关键词词袋中各个关键词在所述分词结果中出现的次数;

获取所述分词结果中的目标词, 并获取所述目标词与所述关键词词袋中各个关键词的相似度值, 所述目标词包括在所述分词结果中存在且在所述关键词词袋中不存在的词;

若所述目标词与所述关键词词袋中的第一关键词的相似度值大于相似度阈值, 则根据所述目标词与所述第一关键词的相似度值更新所述第一元素集合中所述第一关键词对应的第一元素, 得到第二元素集合;

其中, 所述第二元素集合中的各个元素用于构建用于短文本分类的机器学习模型。

16、根据权利要求 15 所述的终端, 其特征不在于, 所述处理器具体用于:

获取用于生成关键词词袋的训练样本集;

根据词频-逆文档频率算法从所述训练样本集中确定出 N 个关键词, 根据所述 N 个关键词生成所述关键词词袋。

17、根据权利要求 15 或 16 所述的终端, 其特征不在于, 所述第一元素集合为第一向量, 所述第一向量中包括 N 个元素, 所述第一向量满足:

$$V_1 = [C_{w_1}, C_{w_2}, \dots, C_{w_n}];$$

其中, 所述 V_1 表示所述第一向量, 所述 C_{w_n} 表示关键词词袋中的第 n 个关键词在所述分词结果中出现的次数, 所述 n 的取值范围为 1 到 N 的自然数。

18、根据权利要求 15-17 任意一项所述的终端, 其特征不在于, 所述处理器具体用于:

从词向量数据库中获取所述目标词的词向量, 并从所述词向量数据库中获取所述关键词词袋中各个关键词的词向量;

计算所述目标词的词向量与所述各个关键词的词向量之间的相似度值;

若所述目标词的词向量与所述关键词词袋中第一关键词的词向量之间的相似度值大于相似度阈值, 则计算所述相似度值与所述目标词在所述分词结果中出现的次数的乘积;

将所述第一元素集合中所述第一关键词对应的第一元素更新为所述乘积与所述第一元素之和, 得到第二元素集合。

19、根据权利要求 15-18 任意一项所述的终端, 其特征不在于, 所述处理器还用于:

获取所述第二元素集合中 M 个 0 元素对应的 M 个第二关键词;

获取所述关键词词袋中的至少一个第三关键词，所述第三关键词在所述分词结果中出现的次数大于或等于 1；

根据所述 M 个第二关键词中各个第二关键词与所述关键词词袋中各个第三关键词之间的相似度值替换所述第二元素集合中的 0 元素，得到第三元素集合；

其中，所述第三元素集合中的各个元素用于构建用于短文本分类的机器学习模型。

20、一种计算机可读存储介质，其特征在于，所述计算机可读存储介质存储有计算机程序，所述计算机程序包括程序指令，所述程序指令当被处理器执行时使所述处理器执行如权利要求 1-7 任一项所述的方法。

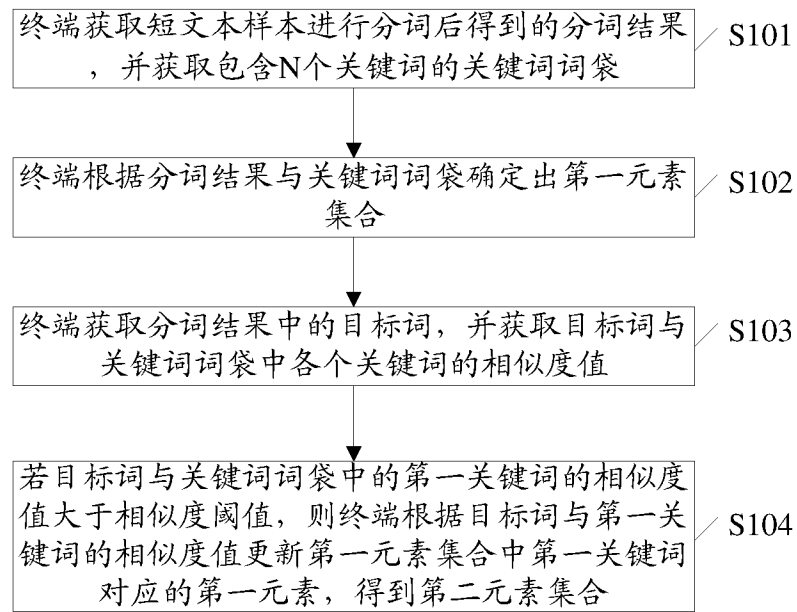


图 1

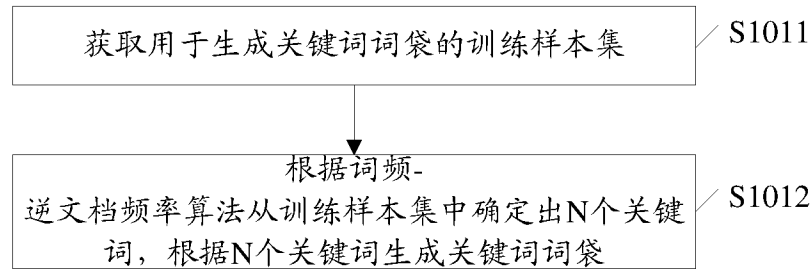


图 2



图 3

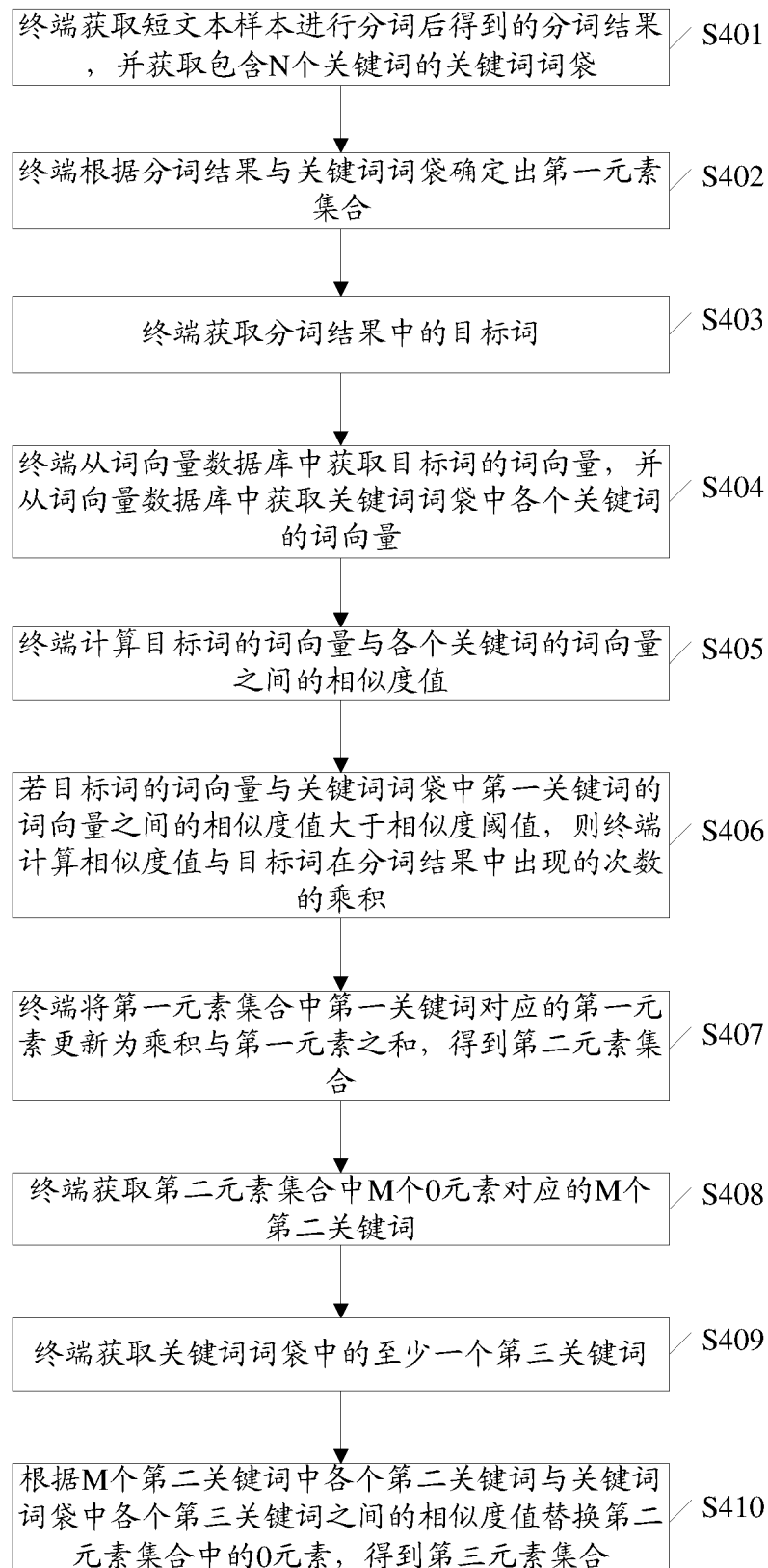


图 4

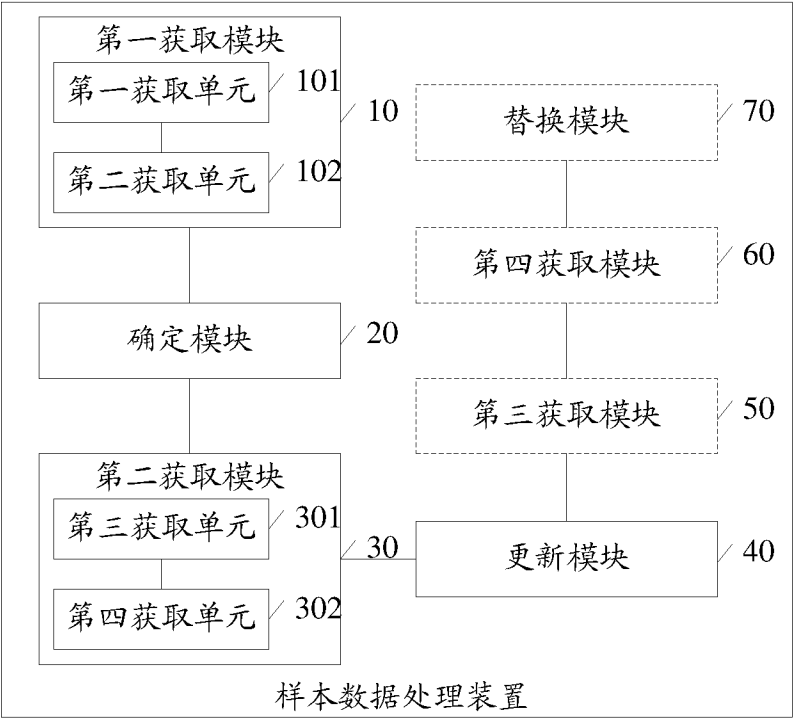


图 5

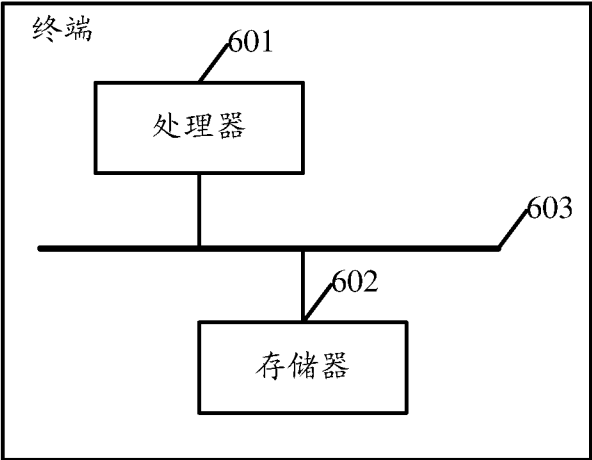


图 6

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2019/088803

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 G06N

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)

WPI, EPODOC, CNPAT, CNKI: 分词, 划分, 次数, 相似, 值, 阈值, 门限, 更新, 替换, 代替, 替代, 更换, 短文本, classif+, word, times, similar+, value, threshold, update, replace, short, text

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 109508378 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 22 March 2019 (2019-03-22) description, paragraphs [0006]-[0045]	1-20
X	CN 102622378 A (BEIJING QIANXIANG NETSCAPE TECHNOLOGY DEVELOPMENT CO., LTD.) 01 August 2012 (2012-08-01) description, paragraphs [0023]-[0082] and [0094]	1-20
X	CN 105045875 A (BEIJING FORESTRY UNIVERSITY) 11 November 2015 (2015-11-11) description, paragraphs [0050]-[0121], and figure 2a	1-20
A	CN 105488023 A (GUANGZHOU I9GAME INFORMATION TECHNOLOGY CO., LTD.) 13 April 2016 (2016-04-13) entire document	1-20
A	CN 107103012 A (ALIBABA GROUP HOLDING LIMITED) 29 August 2017 (2017-08-29) entire document	1-20
A	US 2016253597 A1 (XEROX CORPORATION) 01 September 2016 (2016-09-01) entire document	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

12 July 2019

Date of mailing of the international search report

26 July 2019

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/CN2019/088803

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN	109508378	A	22 March 2019	None	
CN	102622378	A	01 August 2012	None	
CN	105045875	A	11 November 2015	None	
CN	105488023	A	13 April 2016	None	
CN	107103012	A	29 August 2017	None	
US	2016253597	A1	01 September 2016	None	

国际检索报告

国际申请号

PCT/CN2019/088803

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 G06N 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)) WPI, EPDOC, CNPAT, CNKI: 分词, 划分, 次数, 相似, 值, 阈值, 门限, 更新, 替换, 代替, 替代, 更换, 短文本, classif+, word, times, similar+, value, threshold, update, replace, short, text																							
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 109508378 A (平安科技深圳有限公司) 2019年 3月 22日 (2019 - 03 - 22) 说明书第[0006]-[0045]段</td> <td>1-20</td> </tr> <tr> <td>X</td> <td>CN 102622378 A (北京千橡网景科技发展有限公司) 2012年 8月 1日 (2012 - 08 - 01) 说明书第[0023]-[0082]、[0094]段</td> <td>1-20</td> </tr> <tr> <td>X</td> <td>CN 105045875 A (北京林业大学) 2015年 11月 11日 (2015 - 11 - 11) 说明书第[0050]-[0121]段、图2a</td> <td>1-20</td> </tr> <tr> <td>A</td> <td>CN 105488023 A (广州爱九游信息技术有限公司) 2016年 4月 13日 (2016 - 04 - 13) 全文</td> <td>1-20</td> </tr> <tr> <td>A</td> <td>CN 107103012 A (阿里巴巴集团控股有限公司) 2017年 8月 29日 (2017 - 08 - 29) 全文</td> <td>1-20</td> </tr> <tr> <td>A</td> <td>US 2016253597 A1 (XEROX CORPORATION) 2016年 9月 1日 (2016 - 09 - 01) 全文</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 109508378 A (平安科技深圳有限公司) 2019年 3月 22日 (2019 - 03 - 22) 说明书第[0006]-[0045]段	1-20	X	CN 102622378 A (北京千橡网景科技发展有限公司) 2012年 8月 1日 (2012 - 08 - 01) 说明书第[0023]-[0082]、[0094]段	1-20	X	CN 105045875 A (北京林业大学) 2015年 11月 11日 (2015 - 11 - 11) 说明书第[0050]-[0121]段、图2a	1-20	A	CN 105488023 A (广州爱九游信息技术有限公司) 2016年 4月 13日 (2016 - 04 - 13) 全文	1-20	A	CN 107103012 A (阿里巴巴集团控股有限公司) 2017年 8月 29日 (2017 - 08 - 29) 全文	1-20	A	US 2016253597 A1 (XEROX CORPORATION) 2016年 9月 1日 (2016 - 09 - 01) 全文	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 109508378 A (平安科技深圳有限公司) 2019年 3月 22日 (2019 - 03 - 22) 说明书第[0006]-[0045]段	1-20																					
X	CN 102622378 A (北京千橡网景科技发展有限公司) 2012年 8月 1日 (2012 - 08 - 01) 说明书第[0023]-[0082]、[0094]段	1-20																					
X	CN 105045875 A (北京林业大学) 2015年 11月 11日 (2015 - 11 - 11) 说明书第[0050]-[0121]段、图2a	1-20																					
A	CN 105488023 A (广州爱九游信息技术有限公司) 2016年 4月 13日 (2016 - 04 - 13) 全文	1-20																					
A	CN 107103012 A (阿里巴巴集团控股有限公司) 2017年 8月 29日 (2017 - 08 - 29) 全文	1-20																					
A	US 2016253597 A1 (XEROX CORPORATION) 2016年 9月 1日 (2016 - 09 - 01) 全文	1-20																					
☐ 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		国际检索报告邮寄日期																					
2019年 7月 12日		2019年 7月 26日																					
ISA/CN的名称和邮寄地址		受权官员																					
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		王欣 电话号码 86-(10)-53961617																					

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2019/088803

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	109508378	A	2019年 3月 22日	无	
CN	102622378	A	2012年 8月 1日	无	
CN	105045875	A	2015年 11月 11日	无	
CN	105488023	A	2016年 4月 13日	无	
CN	107103012	A	2017年 8月 29日	无	
US	2016253597	A1	2016年 9月 1日	无	