(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 114896404 A
(43)申请公布日 2022.08.12

(21)申请号 202210576341.8

(22)申请日 2022.05.25

(71)申请人 北京金山数字娱乐科技有限公司
地址 100085 北京市海淀区西二旗中路33号院5号楼11层002号
申请人 成都金山互动娱乐科技有限公司

(72)发明人 王得贤 李长亮

(74)专利代理机构 北京智信禾专利代理有限公司 11637
专利代理师 金鹏

(51)Int.Cl.
G06F 16/35(2019.01)
G06V 30/148(2022.01)
G06V 30/19(2022.01)

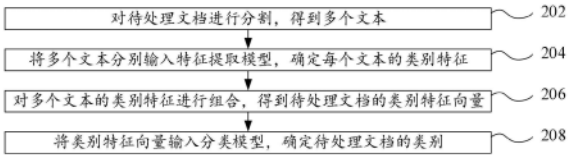
权利要求书2页 说明书24页 附图5页

(54)发明名称

文档分类方法及装置

(57)摘要

本申请提供文档分类方法及装置,其中所述文档分类方法包括:对待处理文档进行分割,得到多个文本;将多个文本分别输入特征提取模型,确定每个文本的类别特征;对多个文本的类别特征进行组合,得到待处理文档的类别特征向量;将所述类别特征向量输入分类模型,确定所述待处理文档的类别。该方法不仅能够适用于长文档处理,而且能够得到融合了待处理文档全文类别信息的类别特征向量,该类别特征向量不仅能够体现待处理文档中各部分内容的类别特征,还能够体现待处理文档中各部分内容之间的关联,因此将该类别特征向量输入分类模型进行分类,能够给分类模型提供更多的信息,使得分类模型的分类结果更加准确,提高了文档分类的准确率。



CN 114896404 A

1.一种文档分类方法,其特征在于,所述方法包括:

对待处理文档进行分割,得到多个文本;

将所述多个文本分别输入特征提取模型,确定每个文本的类别特征;

对所述多个文本的类别特征进行组合,得到所述待处理文档的类别特征向量;

将所述类别特征向量输入分类模型,确定所述待处理文档的类别。

2.如权利要求1所述的方法,其特征在于,所述特征提取模型包括输入层、嵌入层和输出层,将所述多个文本分别输入特征提取模型,确定每个文本的类别特征,包括:

通过所述输入层,对每个文本进行分词处理,得到每个文本的词单元;

通过所述嵌入层,对每个文本的词单元分别进行词嵌入处理,得到每个文本中词单元的词嵌入向量;

针对任一文本,通过所述输出层,基于该文本中词单元的词嵌入向量,确定该文本的类别特征。

3.如权利要求2所述的方法,其特征在于,所述输出层包括词级注意力层和全连接层,针对任一文本,通过所述输出层,基于该文本中词单元的词嵌入向量,确定该文本的类别特征,包括:

针对任一文本,通过所述词级注意力层,将该文本的第一词单元的词嵌入向量与该文本中每个词单元的词嵌入向量进行注意力计算,确定该文本的特征向量,其中,所述第一词单元是该文本中的任一词单元;

通过所述全连接层,基于该文本的特征向量确定该文本的类别特征。

4.如权利要求3所述的方法,其特征在于,所述输出层还包括文本级注意力层,通过所述全连接层,基于该文本的特征向量确定该文本的类别特征之前,还包括:

通过所述文本级注意力层,将该文本的特征向量与多个文本中每个文本的特征向量进行注意力计算,确定该文本的增强特征向量;

通过所述全连接层,基于该文本的特征向量确定该文本的类别特征,包括:

通过所述全连接层,基于该文本的增强特征向量确定该文本的类别特征。

5.如权利要求1所述的方法,其特征在于,对所述多个文本的类别特征进行组合,得到所述待处理文档的类别特征向量,包括:

按照所述多个文本在所述待处理文档中的先后顺序,对所述多个文本的类别特征进行拼接,得到所述待处理文档的类别特征向量。

6.如权利要求1-5任一项所述的方法,其特征在于,所述特征提取模型包括BERT模型。

7.如权利要求1所述的方法,其特征在于,所述分类模型通过如下方式训练得到:

获取多个样本文档,其中,每个样本文档对应一个类别特征向量;

基于多个类别特征向量构建第一决策树,并基于所述第一决策树确定每个样本文档的预测概率;

基于每个样本文档的预测概率和多个类别特征向量构建第二决策树,并基于第二决策树确定每个样本文档的预测概率,以此类推,直到达到停止条件,将构建的多个决策树确定为分类模型。

8.如权利要求1或7所述的方法,其特征在于,所述分类模型包括光梯度增压机Lightgbm模型,且所述分类模型的损失函数是对数损失函数。

9.如权利要求1所述的方法,其特征在于,对待处理文档进行分割,得到多个文本之前,包括:

基于字符识别算法对所述待处理文档的内容进行识别,获取所述待处理文档的字符内容,其中,所述字符识别算法用于识别文档中的字符内容;

对待处理文档进行分割,得到多个文本,包括:

按照预设分割策略,对所述字符内容进行分割,得到所述多个文本。

10.一种文档分类装置,其特征在于,包括:

分割模块,被配置为对待处理文档进行分割,得到多个文本;

第一确定模块,被配置为将所述多个文本分别输入特征提取模型,确定每个文本的类别特征;

组合模块,被配置为对所述多个文本的类别特征进行组合,得到所述待处理文档的类别特征向量;

第二确定模块,被配置为将所述类别特征向量输入分类模型,确定所述待处理文档的类别。

11.一种计算设备,其特征在于,包括:

存储器和处理器;

所述存储器用于存储计算机可执行指令,所述处理器用于执行所述计算机可执行指令实现权利要求1至9任意一项所述文档分类方法的步骤。

12.一种计算机可读存储介质,其存储有计算机指令,其特征在于,该指令被处理器执行时实现权利要求1至9任意一项所述文档分类方法的步骤。

文档分类方法及装置

技术领域

[0001] 本说明书涉及数据处理技术领域，特别涉及文档分类方法及装置。

背景技术

[0002] 文档分类是对文档进行智能识别，从而确定文档的类别，判断该文档是否是目标类别。现有技术中，通常采用基于文本截取的深度学习方法进行文档分类，如对于较长的文档，如3000字以上的文档，因此现有技术一般从文档的前面部分或者中间部分截取部分文本，通过LSTM(Long Short-Term Memory，长短期记忆网络)、CNN(Convolutional Neural Networks，卷积神经网络)等神经网络模型对截取的部分文本进行分类，以确定输入文档的类别。

[0003] 但由于文档较长，无法全部输入神经网络模型，而从文档中截取的部分文本会造成文本信息缺失，影响文档分类的准确性。因此亟需一种文档分类方法以解决上述问题。

发明内容

[0004] 有鉴于此，本申请实施例提供了一种文档分类方法，以解决现有技术中存在的技术缺陷。本申请实施例同时提供了一种文档分类装置，一种计算设备，以及一种计算机可读存储介质。

[0005] 根据本申请实施例的第一方面，提供了一种文档分类方法，包括：

[0006] 对待处理文档进行分割，得到多个文本；

[0007] 将所述多个文本分别输入特征提取模型，确定每个文本的类别特征；

[0008] 对所述多个文本的类别特征进行组合，得到所述待处理文档的类别特征向量；

[0009] 将所述类别特征向量输入分类模型，确定所述待处理文档的类别

[0010] 根据本申请实施例的第二方面，提供了一种文档分类装置，包括：

[0011] 分割模块，被配置为对待处理文档进行分割，得到多个文本；

[0012] 第一确定模块，被配置为将所述多个文本分别输入特征提取模型，确定每个文本的类别特征；

[0013] 组合模块，被配置为对所述多个文本的类别特征进行组合，得到所述待处理文档的类别特征向量；

[0014] 第二确定模块，被配置为将所述类别特征向量输入分类模型，确定所述待处理文档的类别。

[0015] 根据本申请实施例的第三方面，提供了一种计算设备，包括：

[0016] 存储器和处理器；

[0017] 所述存储器用于存储计算机可执行指令，所述处理器执行所述计算机可执行指令时实现所述文档分类方法的步骤。

[0018] 根据本申请实施例的第四方面，提供了一种计算机可读存储介质，其存储有计算机可执行指令，该指令被处理器执行时实现所述文档分类方法的步骤。

[0019] 根据本申请实施例的第五方面，提供了一种芯片，其存储有计算机程序，该计算机程序被芯片执行时实现所述文档分类方法的步骤。

[0020] 本申请提供的文档分类方法，对待处理文档进行分割，得到多个文本；将所述多个文本分别输入特征提取模型，确定每个文本的类别特征；对所述多个文本的类别特征进行组合，得到所述待处理文档的类别特征向量；将所述类别特征向量输入分类模型，确定所述待处理文档的类别。上述方法先将待处理文档分割成比较短的文本，适用于长文档处理，并且先确定每个文本的类别特征，然后将多个文本的类别特征组合得到待处理文档的类别特征向量，则可以认为该类别特征向量融合了待处理文档全文的类别信息，即该类别特征向量不仅能够体现待处理文档中各部分内容的类别特征，还能够体现待处理文档中各部分内容之间的关联，因此将该类别特征向量输入分类模型进行分类，能够给分类模型提供更多的信息，使得分类模型的分类结果更加准确，即提高了文档分类的准确率。

附图说明

[0021] 图1是本申请实施例提供的一种执行文档分类方法的系统的系统架构图；

[0022] 图2是本申请一实施例提供的一种文档分类方法的流程图；

[0023] 图3是本申请一实施例提供的一种分类模型的训练方法的流程图；

[0024] 图4是本申请一实施例提供的一种确定文本的类别特征的方法的流程图；

[0025] 图5是本申请一实施例提供的另一种确定文本的类别特征的方法的流程图；

[0026] 图6是本申请一实施例提供的又一种确定文本的类别特征的方法的流程图；

[0027] 图7是本申请一实施例提供的一种确定待处理文档的类别特征向量的方法的流程图；

[0028] 图8是本申请一实施例提供的一种分割待处理文档的方法的流程图；

[0029] 图9是本申请一实施例提供的一种应用于合同文档识别的文档分类方法的处理流程图；

[0030] 图10是本申请一实施例提供的一种文档分类方法的处理过程示意图；

[0031] 图11是本申请一实施例提供的一种文档分类装置的结构示意图；

[0032] 图12是本申请一实施例提供的一种计算设备的结构框图。

具体实施方式

[0033] 在下面的描述中阐述了很多具体细节以便于充分理解本申请。但是本申请能够以很多不同于在此描述的其它方式来实施，本领域技术人员可以在不违背本申请内涵的情况下做类似推广，因此本申请不受下面公开的具体实施的限制。

[0034] 在本申请一个或多个实施例中使用的术语是仅仅出于描述特定实施例的目的，而非旨在限制本申请一个或多个实施例。在本申请一个或多个实施例和所附权利要求书中所使用的单数形式的“一种”、“所述”和“该”也旨在包括多数形式，除非上下文清楚地表示其他含义。还应当理解，本申请一个或多个实施例中使用的术语“和/或”是指并包含一个或多个相关联的列出项目的任何或所有可能组合。

[0035] 应当理解，尽管在本申请一个或多个实施例中可能采用术语第一、第二等来描述各种信息，但这些信息不应限于这些术语。这些术语仅用来将同一类型的信息彼此区分开。

例如,在不脱离本申请一个或多个实施例范围的情况下,第一也可以被称为第二,类似地,第二也可以被称为第一。

[0036] 首先,对本申请一个或多个实施例涉及的名词术语进行解释。

[0037] 特征提取模型:用于对输入的文本进行特征提取,得到输入文本的类别特征。

[0038] 类别特征:用于表征文本所属类别的特征。

[0039] 分类模型:用于对输入的文档进行分类,确定文档所属的类别。

[0040] 类别特征向量:可以用来确定文档所属类别的特征向量,不仅能够表征文档中各部分内容的类别特征,还能够表征文档中各部分内容之间的关联。

[0041] 词单元:对输入文本做任何实际处理前,都需要将其分割成诸如字、标点符号、数字或字母等语言单元,这些语言单元被称为词单元。对于英文文本,词单元可以是一个单词、一个标点符号、一个数字等;对于中文文本,最小的词单元可以是一个字、一个标点符号、一个数字等。

[0042] Word Embedding Layer(嵌入层):用于对输入的文本进行嵌入式编码处理的层,可以通过一个映射或者一个函数生成文本在新的空间上的表达,该表达可以是文本的词嵌入向量。

[0043] 词嵌入:是指把一个维数为所有词的数量的高维空间嵌入到一个维数低得多的连续向量空间中,每个单词或词组被映射为实数域上的向量的处理过程。

[0044] 词嵌入向量:对词单元进行词嵌入处理后得到的向量。

[0045] WordAttention Layer(词级注意力层):可以包括注意力机制,以词单元为单位进行注意力计算。

[0046] word2vec:进行词嵌入处理的一种方法,是Mikolov在Bengio NNLM(Neural Network Language Model,神经网络语言模型)的基础上构建的一种高效的词向量训练方法。即通过使用该方法可以对文本进行词嵌入处理,得到文本的词嵌入向量。

[0047] 注意力机制:在认知科学中,由于信息处理的瓶颈,人类会选择性地关注所有信息的一部分,同时忽略其他可见的信息,上述机制通常被称为注意力机制。在神经网络模型中,注意力机制通过允许模型动态地关注有助于当前任务的输入的某些部分,可以提高对任务处理的效率。

[0048] 注意力计算:对于某个时刻的输出y,它在输入x上各个部分的注意力,这里的注意力也就是权重,即输入x的各个部分对某时刻输出y贡献的权重。

[0049] 特征向量:融合文本中词单元的词嵌入向量后得到的向量,第一词单元的特征向量中融入了第一词单元与文本中词单元之间的关系,融合了该文本全文的语义信息。

[0050] 增强特征向量:融合文本与其他文本的特征向量后得到的向量,文本的增强特征向量中融入了该文本与自身及其他文本之间的关系,融合了文档全文的语义信息。

[0051] BERT(Bidirectional Encoder Representations from Transformer,基于转换器的双向编码表征)模型:是一种动态词向量技术,采用双向Transformer模型,对无标记数据集进行训练,综合考虑前后文特征信息,可以更好地解决一词多义等问题。

[0052] Lightgbm模型:是一种梯度提升框架,它使用决策树作为基学习器,支持高效率的并行训练,并且具有更快的训练速度、更低的内存消耗、更好的准确率、支持分布式可以快速处理海量数据等优点。

[0053] 对数损失:即Log-likelihood Loss(对数似然损失),也称Logistic Loss(逻辑斯谛回归损失)或cross-entropy Loss(交叉熵损失),是在概率估计上定义的。它常用于multi-nominal(多项)逻辑斯谛回归和神经网络,以及一些期望极大算法的变体,可用于评估分类器的概率输出。

[0054] TF-IDF(Term Frequency-Inverse Document Frequency,词频-逆文档频率):是一种用于资讯检索与文本挖掘的常用加权技术,可以用来评估一个词单元对于一个文本集或文本库中某个文本的重要程度。词单元的重要性随着它在文本中出现的次数成正比增加,但同时会随着它在文本库中出现的频率成反比下降。

[0055] N-gram:是一种统计语言模型,用来根据前(n-1)个item来预测第n个item。

[0056] 接下来对本申请提供的文档分类方法的应用场景进行说明。

[0057] 文档分类是对文档进行智能化识别,确定文档所属的预设分类类别。应用于合同文档分类的场景下,合同文档分类即是对文档进行智能识别,从而判断该文档是否属于合同。当前合同文档分类的方法主要有三种。第一种,基于规则的方法:人工设计分类规则,通过规则匹配确定文档是否是合同,以实现合同文档的识别;第二种,基于传统机器学习的方法:人工构建词库和文档的特征,如TF-IDF、N-gram、关键词等,通过机器学习模型(如SVM(support vector machines,支持向量机)、XGBoost(extreme gradientboosting,极限梯度提升)、LR(Logistic Regression,逻辑回归)等)实现合同文档识别;第三种,基于文本截取的深度学习方法:对于字符内容比较多的文档,如3000字的长文档,一般从文档的前面部分或者中间部分截取部分文本,通过LSTM、CNN等神经网络模型对所截取的部分文本进行处理,以确定文档的类别。

[0058] 但基于规则的方法,需要用户设计大量规则,构建成本较大且费事费力;基于传统机器学习的方法需要人工构建词库和特征,特征工程复杂且难以完全构建,影响识别准确性;基于深度学习的方法,由于文本较长,无法全部输入神经网络模型,而截取文档中部分文本会造成文本信息缺失,影响识别准确性。

[0059] 基于此,本申请提供了一种文档分类方法,无需复杂的规则设计与复杂的特征提取,方便快捷,能够有效提高文档的识别准确率。该文档分类方法的具体实现可以参见下述各个实施例的相关描述。

[0060] 参见图1,图1是本申请一实施例提供的一种执行文档分类方法的系统的系统架构图。

[0061] 该系统可以包括执行文档分类方法的服务端101、训练特征提取模型的第一训练端102和训练分类模型的第二训练端103。并且,该服务端、第一训练端和第二训练端可以集成于同一个计算设备中,也可以在相互独立的不同计算设备中。示例性地,服务端、第一训练端和第二训练端分别是三个相互独立的计算设备;或者,第一训练端和第二训练端集成在同一个计算设备中,服务端在另一个计算设备中;或者,服务端和第一训练端集成在同一个计算设备中,第二训练端在另一个计算设备中;或者,服务端、第一训练端和第二训练端集成在同一个计算设备中,本申请实施例对此不作限定。

[0062] 并且,上述计算设备可以是终端,也可以是服务器,该终端可以是任何一种可与用户进行人机交互的电子产品,该服务器可以是一台服务器,也可以是由多台服务器组成的服务器集群,或者是一个云计算服务中心,本申请实施例对此不做限定。

[0063] 以服务端、第一训练端和第二训练端集成在同一个计算设备中为例，对本申请实施例提供的文档分类方法进行简单介绍。

[0064] 第一训练端通过样本文档训练特征提取模型，且能够通过特征提取模型输出样本文档的类别特征向量，然后将该样本文档的类别特征向量发送至第二训练端，然后第二训练端通过样本文档的类别特征向量训练分类模型。

[0065] 服务端对待处理文档分割得到多个文本，然后将多个文本发送至第一训练端，通过第一训练端的特征提取模型确定每个文本的类别特征，再将多个文本的类别特征发送至服务端，由服务端对多个文本的类别特征进行组合，得到待处理文档的类别特征向量，再将该待处理文档的类别特征向量发送至第二训练端，通过第二训练端的分类模型确定待处理文档的类别。

[0066] 本申请实施例提供的文档分类方法，先将待处理文档分割成比较短的文本，适用于长文档处理，并且先确定每个文本的类别特征，然后将多个文本的类别特征组合得到待处理文档的类别特征向量，则可以认为该类别特征向量融合了待处理文档全文的类别信息，即该类别特征向量不仅能够体现待处理文档中各部分内容的类别特征，还能够体现待处理文档中各部分内容之间的关联，因此将该类别特征向量输入分类模型进行分类，能够给分类模型提供更多的信息，使得分类模型的分类结果更加准确，即提高了文档分类的准确率。

[0067] 在本申请中，提供了一种文档分类方法。本申请同时涉及一种文档分类装置、一种计算设备，以及一种计算机可读存储介质，在下面的实施例中逐一进行详细说明。

[0068] 图2示出了根据本申请一实施例提供的一种文档分类方法的流程图，具体包括以下步骤：

[0069] 步骤202：对待处理文档进行分割，得到多个文本。

[0070] 其中，待处理文档是需要进行分类以确定类别的文档，或者说是需要进行识别以确定是否属于目标类别的文档。该目标类别是用户想要获取的文档所属的类别，例如该目标类别可以是合同、专利文件、简历等等。

[0071] 在一些实施例中，若待处理文档是长文档，即待处理文档包括的字符内容比较多，待处理文档的数据量比较大，则无法直接将待处理文档输入特征提取模型进行处理，因此，需要将待处理文档划分为包括字符内容较少的多个文本。

[0072] 作为一种示例，待处理文档可以是图片格式的文档，如待处理文档的格式是PDF(Portable Document Format，可携带文档格式)，或者待处理文档可以是doc、docx、txt等可编辑格式的文档，本申请实施例对待处理文档的格式不进行限定。

[0073] 在一些实施例，对待处理文档进行分割之前，可以先获取待处理文档的字符内容，然后按照分割策略对字符内容进行分割，可以得到多个文本。针对不同格式的待处理文档可以采用对应的字符识别方法识别字符内容并进行获取，如对于图片格式的文档，可以采用OCR(光学字符识别)技术识别待处理文档中的字符内容，本申请实施例对字符识别方法不进行限定。

[0074] 作为一种示例，分割策略可以包括按段分割、按句分割、按章节分割、按页分割等，本申请实施例对分割策略不做限定。并且，在实际使用时，由于按页分割可能会出现将一句完整的内容分在两个文本的情况，因此，可以将按页分割与其他分割方式结合使用，以确保

分割得到的每个文本的内容都是完整。也即是，使用该分割策略对待处理文档进行分割，能够确保分割得到的每个文本的内容均是完整的。

[0075] 示例性地，在按页分割的情况下，可以通过判断每页的最后一个字符是否是结束符号，如句号、感叹号等，以确定如何对字符内容进行分割。例如，若当前页的最后一个字符是句号，则将当前页的字符内容确定为一个文本，若当前页的最后一个字符不是结束符号，则从当前页的下一页中查找结束符号，将下一页中第一个结束符号之前的字符内容划分到当前页中，即将当前页的字符内容和下一页中该第一个结束符号之前的字符内容确定为一个文本；或者将下一页中第一段的字符内容划分到当前页中，即将当前页的字符内容和下一页中第一段的字符内容确定为一个文本。

[0076] 在本申请实施例中，不是对整个待处理文档进行处理，而是将待处理文档划分为多个短文本，以便于模型处理，解决了长文档处理困难的问题。

[0077] 步骤204：将多个文本分别输入特征提取模型，确定每个文本的类别特征。

[0078] 其中，特征提取模型用于对输入的文本进行特征提取，类别特征用于表征文本的类别。

[0079] 在一些实施例中，特征提取模型可以包括输入层、嵌入层和输出层，输出层还进一步包括词级注意力层和全连接层。其中，输入层用于对文本进行分词处理，得到词单元；嵌入层用于对输入的词单元进行词嵌入处理，得到词单元的词嵌入向量；词级注意力层用于对同一个文本中词单元的词嵌入向量进行注意力计算，得到融合该文本上下文语义信息的特征向量；全连接层用于基于每个文本的特征向量或增强特征向量确定每个文本的类别特征，即确定每个文本所属的类别。

[0080] 作为一种示例，将待处理文档进行分割后的多个文本分别输入特征提取模型，针对任一文本，可以先通过输入层对该文本进行分词处理，得到该文本的多个词单元，然后将该文本的多个词单元输入嵌入层，得到每个词单元的词嵌入向量，然后将多个词单元的词嵌入向量输入词级注意力层，可以得到该文本中每个词单元的特征向量，该特征向量融合了该文本中词单元的语义信息，然后将该文本中多个词单元的特征向量进行拼接，能够得到该文本的特征向量。即经过词级注意力层之后，能够得到每个文本的特征向量，将每个文本的特征向量输入全连接层，可以确定每个文本的分类结果，该分类结果可以称为该文本的类别特征。

[0081] 在另一些实施例中，为了加强整个待处理文档中各部分文本之间的关联性，该特征提取模型的输出层还可以包括文本级注意力层，该文本级注意力层用于对多个文本的特征向量进行注意力计算，得到融合了待处理文档的上下文语义信息的增强特征向量。

[0082] 作为一种示例，将待处理文档进行分割后的多个文本分别输入特征提取模型，针对任一文本，可以先通过输入层对该文本进行分词处理，得到该文本的多个词单元，然后将该文本的多个词单元输入嵌入层，得到每个词单元的词嵌入向量，然后将多个词单元的词嵌入向量输入词级注意力层，得到每个文本融合自身各个词单元语义特征的特征向量后，可以将多个文本的特征向量输入文本级注意力层，能够得到每个文本融合了自身及其他文本语义信息的增强特征向量，将每个文本的增强特征向量输入全连接层进行处理，可以得到每个文本的类别特征。

[0083] 作为一种示例，特征提取模型可以包括BERT模型，由于BERT模型能够提取到文本

融合全文语义信息后的特征向量，因此基于BERT模型提取文本的类别特征能够得到更加准确的结果。

[0084] 作为另一种示例，特征提取模型还可以是BERT模型的变形，如Roberta、Tinybert、Albert、ERNIE(Enhanced Language Representation with Informative Entities，使用信息实体增强语言表示)等，这些模型的结构和训练方式有差异，针对不同任务效果不同，但都可以用来对文本进行特征提取。

[0085] 另外，本申请实施例使用的特征提取模型可以通过如下方式训练得到：

[0086] 获取样本文档集，该样本文档集中每个样本文档携带类别标签，标签1表示目标类别，标签2表示非目标类别。将每个文档分割成多个文本，将每个文本和该文本所属的文档的类别标签作为一条训练数据，然后对样本文档集中多个文档进行同样处理后得到多条训练数据，将多条训练数据输入特征提取模型进行训练，针对每条训练数据，特征提取模型可以预测一个分类结果，该分类结果用于表示特征提取模型预测的该条训练数据中文本的类别，将该类别转换为向量的形式，得到类别特征，且将该条训练数据中的类别标签转换为向量的形式，通过损失函数确定预测的类别和类别标签的损失值，若该损失值小于损失阈值，则停止训练，认为该特征提取模型已经训练完成，若该损失值大于或等于损失阈值，基于损失值对特征提取模型的参数进行调整并继续训练，直至损失值小于损失阈值。

[0087] 本申请实施例中，将多个文本分别输入特征提取模型，得到每个文本的类别特征，能够对文档中各部分的内容先进行初步的类别划分，且使用BERT模型能够得到融合文本各部分语义信息的增强特征向量，提高了确定文本的类别特征的准确率。

[0088] 步骤206：对多个文本的类别特征进行组合，得到待处理文档的类别特征向量。

[0089] 在一些实施例中，对多个文本的类别特征进行组合，可以是对多个类别特征进行拼接，得到待处理文档的类别特征向量；也可以对多个类别特征进行注意力计算，得到每个文本的增强类别特征，将增强类别特征进行拼接，得到待处理文档的类别特征向量。

[0090] 作为一种示例，若每个文本的类别特征是一维向量，则类别特征向量的维度与文档分割得到的文本的数量相同，若每个文本的类别特征均是多维向量，可以先将多个类别特征调整至相同维度，然后将多个类别特征拼接得到类别特征向量。

[0091] 本申请实施例中，待处理文档的类别特征向量是根据文本的类别特征通过注意力计算或者拼接得到的，由于类别特征能够反映待处理文档中文本的类别信息，注意力计算或拼接还可以体现文档中文本之间的关联，因此该类别特征向量能够为后续文档分类提供更多的分类依据，进而提高文档分类的准确性。

[0092] 步骤208：将类别特征向量输入分类模型，确定待处理文档的类别。

[0093] 在一些实施例中，将待处理文档的类别特征向量输入训练完成的分类模型，通过构建决策树可以确定出该待处理文档的类别。

[0094] 作为一种示例，分类模型可以包括Lightgbm模型，并且该分类模型的损失函数可以是对数损失函数。示例性地，该对数损失函数可以是二元对数损失函数，该二元对数损失函数用于对Lightgbm模型的参数进行优化。

[0095] 作为一种示例，该分类模型的损失函数也可以是交叉熵损失。本申请实施例对分类模型的损失函数不进行限定。

[0096] 示例性地，分类模型可以包括多个决策树，将待处理文档的类别特征向量输入每

个决策树，基于每个决策树可以确定一个预测概率，将该多个预测概率相加并进行归一化处理，可以得到该待处理文档对应的类别概率，基于该类别概率确定该待处理文档的类别。例如，在该分类模型中，概率越接近1，表示文档是合同的可能性越大，概率越接近0，表示文档是非合同的可能性越大。假设确定的待处理文档对应的类别概率是0.9，则可以确定该待处理文档是合同。

[0097] 本申请提供的文档分类方法，对待处理文档进行分割，得到多个文本；将所述多个文本分别输入特征提取模型，确定每个文本的类别特征；对所述多个文本的类别特征进行组合，得到所述待处理文档的类别特征向量；将所述类别特征向量输入分类模型，确定所述待处理文档的类别。上述方法先将待处理文档分割成比较短的文本，适用于长文档处理，并且先确定每个文本的类别特征，然后将多个文本的类别特征组合得到待处理文档的类别特征向量，则可以认为该类别特征向量融合了待处理文档全文的类别信息，即该类别特征向量不仅能够体现待处理文档中各部分内容的类别特征，还能够体现待处理文档中各部分内容之间的关联，因此将该类别特征向量输入分类模型进行分类，能够给分类模型提供更多的信息，使得分类模型的分类结果更加准确，即提高了文档分类的准确率。另外，上述方法避免复杂的人工构建词库和特征工程，且并不是根据文档中一部分文本确定文档的类别，解决了文本信息缺失的问题，进而减少了对分类准确性的影响。

[0098] 图3是本申请一实施例提供的一种分类模型的训练方法的流程图。具体包括以下步骤：

[0099] 步骤302：获取多个样本文档，其中，每个样本文档对应一个类别特征向量。

[0100] 步骤304：基于多个类别特征向量构建第一决策树，并基于第一决策树确定每个样本文档的预测概率。

[0101] 步骤306：基于每个样本文档的预测概率和多个类别特征向量构建第二决策树，并基于第二决策树确定每个样本文档的预测概率，以此类推，直到达到停止条件，将构建的多个决策树确定为分类模型。

[0102] 也就是说，分类模型可以认为是多个决策树组成的模型，且每个决策树可以认为是一个计算公式，多个计算公式结合作为分类模型的参数，用于对分类模型的输入进行分类。实际上，构建多个决策树是通过调整决策树的参数使得基于多个决策树确定的最终的预测类别与样本文档的类别标签无限接近甚至相同。

[0103] 作为一种示例，每个样本文档对应的类别特征向量是用于表示该样本文档类别的特征，该类别特征向量中每一维表示一个类别特征。

[0104] 另外，在构建决策树之前，可以设置待构建的决策树的预设数量，以及设置每个决策树的预设深度，该预设深度是决策树包括的层数。并且，预设数量和预设深度均可以由用户根据实际需求设置，也可以由设备默认设置，还可以根据实际情况调整，本申请实施例对此不做限定。

[0105] 作为一种示例，假设包括M个样本文档，且每个样本文档的类别特征向量是N维，每一维表示一种类别特征，并且对于每一维类别特征来说，其包括0和1两种取值，其中0表示不是合同，1表示是合同，每个样本文档的初始预测概率可以设置为0.5，表示每个样本文档是合同和非合同的概率相同。在构建第一决策树时，对于每个类别特征向量中的X1特征，可以得到以下几种划分方式：X1<2，X1<1，X1<0，通过损失函数，基于M个样本文档的初始预测

概率确定每种划分方式的增益Gain，同理，对于每个类别特征向量中的X2特征重复上述步骤，一直到所有的类别特征遍历完，从计算得到的所有增益中选择增益最大的划分方式作为分裂点，将M个样本文档按照增益最大的划分方式划分，然后重复上述步骤，直到达到预设深度。并且，在构建决策树过程中，若叶子节点只有一个样本文档，则可以计算该叶子节点的节点值。

[0106] 其中，X1<2，X1<1，X1<0仅是一种举例，表示类别特征向量中的X1特征可以按照这几种方式进行划分，且0、1、2可以表示不同的含义，可以根据用户需求设置或者由设备默认设置，并不对本申请实施例中X1特征的划分方式进行限定。

[0107] 实际上，可以根据特征的类型不同采用不同的划分方式对特征进行划分。例如，假设X1特征表示的是样本文档的格式，且样本文档的格式包括单栏、双栏和混合栏，则对X1特征的划分可以包括以下几种划分方式：X1是否小于2，若是，表示样本文档的格式是单栏，若否，表示样本文档的格式是非单栏，即按照样本文档的格式是否是单栏进行划分；X1是否小于1，若是，表示样本文档的格式是双栏，若否，表示样本文档的格式是非双栏，即按照样本文档的格式是否是双栏进行划分；X1是否小于0，若是，表示样本文档的格式是混合栏，若否，表示样本文档的格式是非混合栏，即按照样本文档的格式是否是混合栏进行划分。其中选取的0、1、2与样本文档的格式之间的关系可以根据实际需求设置，本申请实施例对此不做限定。

[0108] 以计算X1<1这种划分方式的增益为例，先确定按照X1<1划分之后，“X1<1”这一分支包括的样本文档集A，和“X1≥1”这一分支包括的样本文档集B，通过损失函数根据样本文档集A中样本文档的初始预测概率确定样本文档集A的损失值，且通过损失函数根据样本文档集B中样本文档的初始预测概率确定样本文档集B的损失值，且通过损失函数根据M个样本文档的初始预测概率确定M个样本文档的损失值，基于这三种损失值确定按照X1<1的方式划分时的增益。

[0109] 作为一种示例，构建完第一决策树后，每个样本文档都被划分至一个叶子节点，每个叶子结点对应有一个节点值，基于该叶子节点的节点值可以确定该叶子节点对应的样本文档的预测概率，然后基于每个样本文档的预测概率，按照上述构建第一决策树的方式构建第二决策树，并基于第二决策树确定每个样本文档的预测概率。以此类推，直至构建的决策树的数量大于或等于预设数量；或者，每个样本文档携带有类别标签，基于当前构建的决策树确定每个样本文档的预测概率，基于该预测概率确定的每个样本文档的预测类别与类别标签相同，则停止构建决策树，将当前构建的多个决策树确定为训练完成的分类模型。

[0110] 需要说明的是，本申请实施例中提到的分类模型均采用上述步骤302-步骤306的方式训练得到。

[0111] 本申请实施例通过样本文档的类别特征向量构建多个决策树，并根据决策树确定每个样本文档的预测概率，在根据预测概率和预测标签确定满足决策树构建停止条件时，停止构建决策树，将构建的所有决策树确定为分类模型，能够得到可以预测文档类别的分类模型，便于对待处理文档进行分类。

[0112] 图4示出了根据本申请一实施例提供的一种确定文本的类别特征的方法的流程图，具体包括以下步骤：

[0113] 步骤402：通过输入层，对每个文本进行分词处理，得到每个文本的词单元。

[0114] 本申请实施例中，特征提取模型包括输入层、嵌入层和输出层，输入层用于对输入进行分词处理。

[0115] 在一些实施例中，对文本进行分词处理的过程中，若文本是中文文本，可以将一个字划分为一个词单元，将一个短语划分为一个词单元，将一个标点符号划分为一个词单元；若文本是外文文本，可以将一个单词划分为一个词单元，或者，将一个短语划分为一个词单元，或者，将一个外文字符划分为一个词单元；若文本中有数字，可以将数字单独划分为一个词单元。

[0116] 在本申请实施例中，可以采用基于词典、基于词频度统计、基于规则等任意一种分词方法对每个文本进行分词处理。在一些实施例中，基于词典的分词方法可以包括正向最大匹配，逆向最大匹配，最少词切分法和双向匹配法。基于规则的分词方法可以包括基于HMM(Hidden Markov Model，隐马尔科夫模型)的分词方法。或者，在本申请实施例中，若文本是中文文本，还可以将每个字划分为一个词单元。

[0117] 以正向最大匹配的分词方法为例，对于任一文本，按照文本的阅读顺序正向获取该文本的m个字符作为匹配字段，将该匹配字段与词典中的词进行匹配，若词典中存在与该匹配字段相同的词，则认为匹配成功，将该匹配字段作为一个词单元切分出来。若词典中不存在与该匹配字段相同的词，则认为匹配失败，将该匹配字段的最后一个字符去掉，剩下的字符作为新的匹配字段，进而再次匹配，直到剩余字符串的长度为零，可以认为完成了一轮匹配，然后从文本中取出下一组m个字符作为匹配字段进行匹配处理，直到该文本中所有字符均被切分完为止。

[0118] 其中，m可以是词典中最长的词包含字符的数量，也可以根据经验预设的，本申请实施例对此不作限定。

[0119] 以逆向最大匹配的分词方法为例，对于任一文本，按照文本的阅读顺序逆向获取该文本的m个字符作为匹配字段，将该匹配字段与逆序词典中的词进行匹配，若逆序词典中存在与该匹配字段相同的词，则匹配成功，将该匹配字段作为一个词单元切分出来。若逆序词典中不存在与该匹配字段相同的词，则认为匹配失败，去掉匹配字段最前面的一个字符，将剩下的字符作为新的匹配字段，继续匹配，直到剩余字符的长度为零，可以认为完成了一轮匹配，然后从文本中取出下一组m个字符作为匹配字段进行匹配处理，直到该文本中所有字符均被切分完为止。其中，逆序词典中每个词按照逆序方式存放。

[0120] 作为一种示例，可以先将文本进行倒排处理，生成逆序文本，然后根据逆序词典对逆序文本用正向最大匹配的分词法处理，也可以实现分词效果。

[0121] 例如，以匹配字段是“履行义务”为例，通过上述分词方法，可以得到词单元“履行”、“义务”。

[0122] 本申请实施例中，通过对文本进行分词处理，能够得到便于特征提取模型处理的词单元，以便于特征提取模型进行后续处理。

[0123] 步骤404：通过嵌入层，对每个文本的词单元分别进行词嵌入处理，得到每个文本中词单元的词嵌入向量。

[0124] 在一些实施例中，可以通过特征提取模型的嵌入层对词单元进行词嵌入处理，即将多个文本的词单元输入词嵌入层，得到每个文本中词单元的词嵌入向量。

[0125] 作为一种示例，可以对每个文本的词单元进行随机初始化处理，得到每个词单元

的词嵌入向量;或者,可以通过one-hot(独热)编码的方式对每个文本的词单元进行词嵌入处理,得到每个词单元的词嵌入向量;或者,可以通过word2vec编码的方式对每个文本的词单元进行词嵌入处理,得到每个词单元的词嵌入向量。

[0126] 步骤406:针对任一文本,通过输出层,基于该文本中词单元的词嵌入向量,确定该文本的类别特征。

[0127] 在本申请实施例中,确定每个文本中词单元的词嵌入向量后,可以将同一个文本中多个词单元的词嵌入向量进行组合,得到该文本的词嵌入向量,基于该文本的词嵌入向量可以确定该文本的类别特征。

[0128] 在一些实施例中,针对任一文本,可以按照多个词单元在该文本中的顺序对多个词嵌入向量进行拼接,得到该文本的词嵌入向量;或者,可以将多个词单元的词嵌入向量相加,得到该文本的词嵌入向量。

[0129] 例如,以文本是“双方必须按照合约履行义务”为例,通过上述两个步骤分别确定了词单元“双方”、“必须”、“按照”、“合约”、“履行”、“义务”的词嵌入向量,假设“双方”的词嵌入向量是001,“必须”的词嵌入向量是000,“按照”的词嵌入向量是001,“合约”的词嵌入向量是010,“履行”的词嵌入向量是100,“义务”的词嵌入向量是110,可以按照该多个词单元在文本中的顺序对该多个词嵌入向量进行拼接,得到该文本的词嵌入向量。按照拼接方式不同,可以得到两种文本的词嵌入向量,一种是001000001010100110,另一种是6×3的矩阵$\begin{bmatrix}0&0&1\\0&0&0\\0&0&1\\0&1&0\\1&0&0\\1&1&0\end{bmatrix}$。

[0130] 在一些实施例中,该输出层可以包括全连接层,该全连接层可以称为Fully Connected Layer,且该全连接层中包括激活函数。示例性地,该激活函数可以是Sigmoid函数。该Sigmoid函数可以对输入进行归一化处理,以将输入的变量映射到0,1之间。

[0131] 作为一种示例,对于任一文本,可以将该文本的词嵌入向量输入全连接层,通过全连接层的参数对词嵌入向量进行转换,得到该文本与每种类别的相关性分值,再通过激活函数对相关性分值进行处理,可以确定该文本属于每种类别的概率,将最大概率对应的类别确定为该文本所属的类别,将该类别转换为向量表示得到该文本的类别特征。

[0132] 示例性地,假设文本的类别包括合同和非合同两种,合同用1表示,非合同用0表示,则该文本的类别特征可以是一维向量0或1;或者,该文本的类别特征可以是n(n≥2)维向量,若该文本是合同,其类别特征可以是00.....01,其中包括n-1个0,若该文本是非合同,其类别特征可以是00.....00,其中包括n个0。

[0133] 例如,假设文本的类别包括合同和非合同两种,合同用1表示,非合同用2表示,则该文本的类别特征可以是1或2。

[0134] 需要说明的是,在本申请实施例中,对于每个文本执行的操作相同,为了便于描述,仅以任一文本为例对确定文本的类别特征的过程进行描述。

[0135] 需要说明的是,步骤402-步骤406是上述步骤204的一种具体实现方式。

[0136] 本申请实施例中，通过特征提取模型对文本中的词单元进行词嵌入处理，且根据词单元的词嵌入向量确定文本的词嵌入向量，能够得到可以准确反映文本的语义特征的词嵌入向量，用更加准确的词嵌入向量确定文本的类别，可以提高模型确定文本类别的准确率。

[0137] 图5示出了根据本申请一实施例提供的另一种确定文本的类别特征的方法的流程图，具体包括以下步骤：

[0138] 步骤502：针对任一文本，通过词级注意力层，将该文本的第一词单元的词嵌入向量与该文本中每个词单元的词嵌入向量进行注意力计算，确定该文本的特征向量。

[0139] 其中，第一词单元是该文本中的任一词单元。特征提取模型的输出层包括词级注意力层和全连接层。

[0140] 在本申请实施例中，词单元的词嵌入向量是仅针对单个词单元的，单纯的向量拼接虽然可以得到表征文本语义的特征向量，但得到的该特征向量忽略了文本中词单元之间的联系，因此，可以通过词级注意力层对多个词嵌入向量进行进一步处理，得到考虑了文本中词单元之间关联关系的特征向量作为文本的特征向量。

[0141] 在一些实施例中，将文本中多个词单元的词嵌入向量输入词级注意力层，将每个第一词单元的词嵌入向量与该文本中包括第一词单元的每个词单元的词嵌入向量进行注意力计算，可以得到注意力矩阵，该注意力矩阵中的元素是第一词单元与文本中词单元的相关性取值。然后基于注意力矩阵确定每个词单元对应的多个权重值，根据每个词单元对应的多个权重值与每个词单元的词嵌入向量，确定该文本的特征向量。

[0142] 作为一种示例，对第一词单元和该文本中每个词单元进行注意力计算可以是确定第一词单元和该文本中每个词单元之间的相似度。

[0143] 示例性地，假设文本包括4个词单元甲、乙、丙、丁，则可以将词单元甲的词嵌入向量和自身(即词单元甲的词嵌入向量)进行注意力计算，得到A11，将词单元甲的词嵌入向量和词单元乙的词嵌入向量进行注意力计算，得到词单元甲和词单元乙的相关性取值A12作为注意力矩阵第一行第二列的元素，将词单元甲的词嵌入向量和词单元丙的词嵌入向量进行注意力计算，得到词单元甲和词单元丙的相关性取值A13作为注意力矩阵第一行第三列的元素，将词单元甲的词嵌入向量和词单元丁的词嵌入向量进行注意力计算，得到词单元甲和词单元丁的相关性取值A14作为注意力矩阵第一行第四列的元素，以此类推，对其他词单元进行相同处理，可以得到注意力矩阵。例如，假设注意力矩阵是

$$\begin{bmatrix} A11 & A12 & A13 & A14 \\ A21 & A22 & A23 & A24 \\ A31 & A32 & A33 & A34 \\ A41 & A42 & A43 & A44 \end{bmatrix}。$$

[0144] 其中，该注意力矩阵的行数和列数相同，且均等于该文本中词单元的数量。并且，该注意力矩阵中第i行第j列的元素Aij表示的是文本中第i个词单元和第j个词单元之间的相关性取值，其中，i与j均是大于0的整数。

[0145] 作为一种示例，基于注意力矩阵确定每个词单元对应的多个权重值的具体实现可以包括：

[0146] 按行对注意力矩阵中的相关性取值进行归一化处理，得到归一化相关性取值，则

第i行第j列的相关性取值是第i个词单元相对于第j个词单元的权重值，可以得到该文本中每个词单元对应的多个权重值；

[0147] 或者，按列对注意力矩阵中的相关性取值进行归一化处理，得到归一化相关性取值，则第i列第j行的相关性取值是第i个词单元相对于第j个词单元的权重值，可以得到该文本中每个词单元对应的多个权重值。

[0148] 继续上述举例，按行对上述注意力矩阵中的相关性取值进行归一化处理。例如，假设对第一行的相关性取值进行归一化处理后，可以得到a11，a12，a13和a14，则可以确定a11是词单元“甲”相对于自身的权重值，a12是词单元“甲”相对于词单元“乙”的权重值，a13是词单元“甲”相对于词单元“丙”的权重值，a14是词单元“甲”相对于词单元“丁”的权重值，以此类推，可以确定文本中每个词单元相对于自身及其他词单元的权重值，即每个词单元对应的多个权重值。

[0149] 同理，按列进行归一化处理，也可以得到每个词单元对应的多个权重值。

[0150] 作为一种示例，根据每个词单元对应的多个权重值与每个词单元的词嵌入向量，确定该文本的特征向量的具体实现可以包括：根据每个词单元对应的多个权重值与每个词单元的词嵌入向量，确定每个词单元的特征向量；基于每个词单元的特征向量和预设权重矩阵，确定文本的特征向量。

[0151] 其中，预设权重矩阵是词级注意力层中已有的通用矩阵，通过对特征提取模型进行训练可以确定。

[0152] 在一种实现方式中，可以基于多个词单元的词嵌入向量组成词嵌入向量矩阵；基于每个词单元对应的多个权重值组成该词单元对应的第一权重矩阵，基于词嵌入向量矩阵和每个词单元对应的第一权重矩阵确定该词单元的特征向量。

[0153] 继续上述举例，针对词单元甲、乙、丙、丁，每个词单元对应4个权重值，且每个权重值与该4个词单元中的一个词单元对应，即在该4个权重值中，有1个权重值与该词单元自身对应，其他3个权重值与除该词单元之外的其余词单元对应。例如，词单元“甲”对应4个权重值，分别是权重值a11、权重值a12、权重值a13和权重值a14，且权重值a11与词单元“甲”对应，权重值a12与词单元“乙”对应，权重值a13与词单元“丙”对应，权重值a14与词单元“丁”对应。假设文本中每个词单元的词嵌入向量是M维向量，则基于该4个词单元的词嵌入向量可以得到4×M的词嵌入向量矩阵。针对词单元“甲”，其对应的4个权重值可以组成一个4×1的第一权重矩阵，可以将该4×1的第一权重矩阵的转置与该4×M的词嵌入向量矩阵相乘，则可以得到一个1×M的矩阵，该1×M的矩阵是词单元“甲”的特征向量。同理，可以分别确定词单元“乙”、“丙”、“丁”的特征向量。

[0154] 在另一种实现方式中，针对第一词单元，可以将该第一词单元对应的每个权重值与该权重值对应的词单元的词嵌入向量进行加权融合，得到该第一词单元的特征向量。

[0155] 继续上述举例，针对词单元甲、乙、丙、丁，每个词单元对应有4个权重值，且每个权重值与该4个词单元中的一个词单元对应，即在该4个权重值中，有1个权重值与该词单元自身对应，其他3个权重值与除该词单元之外的其余词单元对应。针对词单元“甲”，其对应4个权重值分别为a11，a12，a13和a14，可以将a11与“甲”的词嵌入向量相乘，将a12与“乙”的词嵌入向量相乘，将a13与“丙”的词嵌入向量相乘，将a14与“丁”的词嵌入向量相乘，将4个乘积相加作为词单元“甲”的特征向量，如此可以分别确定词单元“乙”、“丙”、“丁”的特征向

量。

[0156] 示例性地，确定每个词单元的特征向量后，基于多个词单元的特征向量组成特征向量矩阵，基于特征向量矩阵和预设权重矩阵，确定文本的特征向量。

[0157] 继续上述举例，假设确定了4个词单元的特征向量，将该4个词单元的特征向量组成4×M的特征向量矩阵，将4×1的预设权重矩阵的转置与该4×M的特征向量矩阵相乘，则可以得到1×M的矩阵，则该矩阵是文本的特征向量，融合了该文本中所有词单元的语义特征。

[0158] 步骤504：通过全连接层，基于该文本的特征向量确定该文本的类别特征。

[0159] 在一些实施例中，可以将文本的特征向量输入全连接层，该全连接层可以称为Fully Connected Layer，且该全连接层中包括激活函数。示例性地，该激活函数可以是Sigmoid函数。该Sigmoid函数可以对输入进行归一化处理，以将输入的变量映射到0，1之间。

[0160] 作为一种示例，对于任一文本，将该文本的特征向量输入全连接层，通过全连接层的参数对特征向量进行转换，得到该文本与每种类别的相关性分值，再通过激活函数对相关性分值进行处理，可以确定该文本属于每种类别的概率，将最大概率对应的类别确定为该文本所属的类别，将该类别转换为向量表示得到该文本的类别特征。

[0161] 需要说明的是，步骤502-步骤504是上述步骤406的一种具体实现方式。

[0162] 本申请实施例中，通过词级注意力层对文本中第一词单元与该文本中每个词单元进行注意力计算，考虑到了文本中词单元之间的关系，能够得到可以准确反映文本的上下文语义关系以及文本中词单元语义的特征向量，用更加准确的特征向量表征文本，进而确定文本的类别，可以提高模型确定文本类别的准确率。

[0163] 图6示出了根据本申请一实施例提供的又一种确定文本的类别特征的方法的流程图，具体包括以下步骤：

[0164] 步骤602：通过文本级注意力层，将该文本的特征向量与多个文本中每个文本的特征向量进行注意力计算，确定该文本的增强特征向量。

[0165] 在本申请实施例中，特征提取模型的输出层还包括文本级注意力层。

[0166] 在本申请实施例中，文本的特征向量是仅针对单个文本的，单纯的拼接虽然可以得到表征文档语义的特征向量，但得到的该特征向量忽略了文档中文本之间的联系，因此，可以通过文本级注意力层对多个特征向量进行进一步处理，得到考虑了文档中文本之间关联关系的特征向量作为文本的增强特征向量。

[0167] 在一些实施例中，将多个文本的特征向量输入文本级注意力层，将每个文本的特征向量与多个文本中包括该文本自身的每个文本的特征向量进行注意力计算，可以得到注意力矩阵，该注意力矩阵中的元素是该文本与多个文本中每个文本的相关性取值。然后基于注意力矩阵确定每个文本对应的多个权重值，根据每个文本对应的多个权重值与每个文本的特征向量，确定每个文本的增强特征向量。

[0168] 作为一种示例，对该文本和多个文本中包括该文本的每个文本进行注意力计算可以是确定该文本和多个文本中每个文本之间的相似度。

[0169] 示例性地，假设待处理文档被分割成了4个文本X、Y、Z和W，则可以将文本X的特征向量和自身(即文本X的特征向量)进行注意力计算，得到B11，文本X的特征向量和文本Y的

特征向量进行注意力计算，得到文本X和文本Y的相关性取值B12作为注意力矩阵第一行第二列的元素，将文本X的特征向量和文本Z的特征向量进行注意力计算，得到文本X和文本Z的相关性取值B13作为注意力矩阵第一行第三列的元素，将文本X的特征向量和文本W的特征向量进行注意力计算，得到文本X和文本W的相关性取值B14作为注意力矩阵第一行第四列的元素，以此类推，对其他文本进行相同处理，可以得到注意力矩阵。例如，假设注意力矩阵是$\begin{bmatrix} B11 & B12 & B13 & B14 \\ B21 & B22 & B23 & B24 \\ B31 & B32 & B33 & B34 \\ B41 & B42 & B43 & B44 \end{bmatrix}$。

[0170] 其中，该注意力矩阵的行数和列数相同，且均等于待处理文档分割得到的文本的数量。并且，该注意力矩阵中第i行第j列的元素Bij表示的是文档中第i个文本和第j个文本之间的相关性取值，其中，i与j均是大于0的整数。

[0171] 作为一种示例，基于注意力矩阵确定每个文本对应的多个权重值的具体实现可以包括：

[0172] 按行对注意力矩阵中的相关性取值进行归一化处理，得到归一化相关性取值，则第i行第j列的相关性取值是第i个文本相对于第j个文本的权重值，可以得到每个文本对应的多个权重值；

[0173] 或者，按列对注意力矩阵中的相关性取值进行归一化处理，得到归一化相关性取值，则第i列第j行的相关性取值是第i个文本相对于第j个文本的权重值，可以得到每个文本对应的多个权重值。

[0174] 继续上述举例，按行对上述注意力矩阵中的相关性取值进行归一化处理。例如，假设对第一行的相关性取值进行归一化处理后，可以得到b11，b12，b13和b14，则可以确定b11是文本X相对于自身的权重值，b12是文本X相对于文本Y的权重值，b13是文本X相对于文本Z的权重值，b14是文本X相对于文本W的权重值，以此类推，可以确定每个文本对应的多个权重值。

[0175] 在一种实现方式中，根据每个文本对应的多个权重值与每个文本的特征向量，确定每个文本的增强特征向量的具体实现可以包括：基于多个文本的特征向量组成特征向量矩阵；基于每个文本对应的多个权重值组成该文本对应的第二权重矩阵，基于特征向量矩阵和每个文本对应的第二权重矩阵确定该文本的增强特征向量。

[0176] 继续上述举例，针对文本X、Y、Z、W，每个文本对应有4个权重值，且每个权重值与该4个文本的一个文本对应，即在该4个权重值中，有1个权重值与该文本自身对应，其他3个权重值与除该文本之外的其余文本对应。例如，文本X对应有4个权重值，分别是权重值b11、权重值b12、权重值b13和权重值b14，且权重值b11与文本X对应，权重值b12与文本Y对应，权重值b13与文本Z对应，权重值b14与文本W对应。假设每个文本的特征向量是M维向量，则基于该4个文本的特征向量可以得到一个4×M的特征向量矩阵。针对文本X，其对应的4个权重值可以组成一个4×1的第二权重矩阵，可以将该4×1的第二权重矩阵的转置与该4×M的特征向量矩阵相乘，则可以得到一个1×M的矩阵，该1×M的矩阵是文本X的增强特征向量。同理，可以分别确定文本Y、文本Z、文本W的增强特征向量。

[0177] 在另一种实现方式中，根据每个文本对应的多个权重值与每个文本的特征向量，

确定每个文本的增强特征向量的具体实现可以包括:针对第一文本,将该第一文本对应的每个权重值与该权重值对应的文本的特征向量进行加权融合,得到该第一文本的增强特征向量。

[0178] 继续上述举例,针对文本X、Y、Z、W,每个文本对应有4个权重值,且每个权重值与该4个文本的一个文本对应,即在该4个权重值中,有1个权重值与该文本自身对应,其他3个权重值与除该文本之外的其余文本对应。针对文本X,其对应的4个权重值分别为b11,b12,b13和b14,可以将b11与文本X的特征向量相乘,将b12与文本Y的特征向量相乘,将b13与文本Z的特征向量相乘,将b14与文本W的特征向量相乘,将4个乘积相加作为文本X的增强特征向量。如此可以分别确定文本Y、文本Z、文本W的增强特征向量。

[0179] 步骤604:通过全连接层,基于该文本的增强特征向量确定该文本的类别特征。

[0180] 在一些实施例中,可以将文本的增强特征向量输入全连接层,该全连接层中包括激活函数。作为一种示例,对于任一文本,将该文本的增强特征向量输入全连接层,通过全连接层的参数对增强特征向量进行转换,得到该文本与每种类别的相关性分值,再通过激活函数对相关性分值进行处理,可以确定该文本属于每种类别的概率,将最大概率对应的类别确定为该文本所属的类别,将该类别转换为向量表示得到该文本的类别特征。

[0181] 需要说明的是,步骤602-步骤604是上述步骤504的一种具体实现方式。

[0182] 本申请实施例中,在确定文本的能够准确反映文本的上下文语义关系以及文本中词单元语义的特征向量后,可以通过文本级注意力层对文本的特征向量进行处理,得到每个文本融合了自身及其他文本特征后的增强特征向量,该增强特征向量考虑了文本之间的关联关系,则不仅能够准确地表征文本,还能够表征文本之间的关联关系,基于该增强特征向量确定文本的类别,能够让文本类别的确定考虑到整个文档内容的关联关系,因此可以提高确定文本类别的准确率。

[0183] 图7示出了根据本申请一实施例提供的一种确定待处理文档的类别特征向量的方法的流程图,具体包括以下步骤:

[0184] 步骤702:按照多个文本在待处理文档中的先后顺序,对多个文本的类别特征进行拼接,得到待处理文档的类别特征向量。

[0185] 在一种可能的实现方式中,对待处理文档进行分割得到的多个文本在待处理文档中必然存在先后顺序,为了使得得到的类别特征向量能够更加准确地表征待处理文档,可以按照多个文本在待处理文档中的先后顺序,对多个文本的类别特征进行拼接,得到待处理文档的类别特征向量。

[0186] 作为一种示例,可以通过增加特征维度的方式对类别特征进行拼接。示例性地,假设待处理文档分割得到3个文本,文本1的类别特征是00,文本2的类别特征是01,文本3的类别特征是01,则可以拼接得到待处理文档的类别特征向量是000101。或者,假设文本1的类别特征是1,文本2的类别特征是2,文本3的类别特征是1,则可以拼接得到待处理文档的类别特征向量是121。

[0187] 作为另一种示例,可以按照矩阵的形式对类别特征进行拼接。示例性地,假设待处理文档分割得到3个文本,文本1的类别特征是00,文本2的类别特征是01,文本3的类别特征

是01，则可以拼接得到待处理文档的类别特征向量是$\begin{bmatrix}0 & 0\\0 & 1\\0 & 1\end{bmatrix}$。

[0188] 需要说明的是，上述仅是将类别特征拼接得到待处理文档的类别特征向量的示例，在实际应用中，可以设置类别特征向量的标准维度，在拼接得到的类别特征向量的维度不足标准维度时，用0补充以使得类别特征向量的维度达到标准维度。例如，假设设置的类别特征向量的维度是20，合同类的文本的类别特征用1表示，非合同类的文本的类别特征用2表示，假设待处理文档划分得到18个文本，且该18个文本的类别特征按照顺序分别是111111111112222211，则拼接得到该待处理文档的类别特征向量是11111111112222211100。

[0189] 在另一种可能的实现方式中，可以将每个文本的类别特征与自身的类别特征以及除自身之外的其他文本的类别特征进行注意力计算，可以得到注意力矩阵，该注意力矩阵中的元素是该文本与自身及其他文本的相关性取值。然后基于注意力矩阵确定每个文本对应的多个权重值，根据每个文本对应的多个权重值与每个文本的类别特征，确定待处理文档的类别特征向量。

[0190] 示例性地，假设待处理文档被分割成了4个文本X、Y、Z和W，则可以将文本X的类别特征和自身(即文本X的类别特征)进行注意力计算，得到C11，文本X的类别特征和文本Y的类别特征进行注意力计算，得到文本X和文本Y的相关性取值C12作为注意力矩阵第一行第二列的元素，将文本X的类别特征和文本Z的类别特征进行注意力计算，得到文本X和文本Z的相关性取值C13作为注意力矩阵第一行第三列的元素，将文本X的类别特征和文本W的类别特征进行注意力计算，得到文本X和文本W的相关性取值C14作为注意力矩阵第一行第四列的元素，以此类推，对其他文本进行相同处理，可以得到注意力矩阵。例如，假设注意力矩阵是$\begin{bmatrix}C11 & C12 & C13 & C14\\C21 & C22 & C23 & C24\\C31 & C32 & C33 & C34\\C41 & C42 & C43 & C44\end{bmatrix}$。

[0191] 其中，该注意力矩阵的行数和列数相同，且均等于待处理文档分割得到的文本的数量。并且，该注意力矩阵中第i行第j列的元素Cij表示的是文档中第i个文本和第j个文本之间的相关性取值，其中，i与j均是大于0的正整数。

[0192] 作为一种示例，基于注意力矩阵确定每个文本对应的多个权重值的具体实现可以包括：

[0193] 按行对注意力矩阵中的相关性取值进行归一化处理，得到归一化相关性取值，则第i行第j列的相关性取值是第i个文本相对于第j个文本的权重值，可以得到每个文本对应的多个权重值；

[0194] 或者，按列对注意力矩阵中的相关性取值进行归一化处理，得到归一化相关性取值，则第i列第j行的相关性取值是第i个文本相对于第j个文本的权重值，可以得到每个文本对应的多个权重值。

[0195] 继续上述举例，按行对上述注意力矩阵中的相关性取值进行归一化处理。例如，假设对第一行的相关性取值进行归一化处理后，可以得到c11，c12，c13和c14，则可以确定c11

是文本X相对于自身的权重值，c12是文本X相对于文本Y的权重值，c13是文本X相对于文本Z的权重值，c14是文本X相对于文本W的权重值，以此类推，可以确定每个文本相对于自身以及其他文本的权重值，即每个文本对应的多个权重值。

[0196] 同理，按列进行归一化处理，也可以得到每个文本对应的多个权重值。

[0197] 作为一种示例，根据每个文本对应的多个权重值与每个文本的类别特征，确定待处理文档的类别特征向量的具体实现可以包括：根据每个文本对应的多个权重值与每个文本的类别特征，确定每个文本的增强类别特征；基于每个文本的增强特征向量和预设权重矩阵，确定待处理文档的类别特征向量。

[0198] 在一种实现方式中，可以基于多个文本的类别特征组成类别特征矩阵；基于每个文本对应的多个权重值组成该文本对应的第三权重矩阵，基于类别特征矩阵和每个文本对应的第三权重矩阵确定该文本的增强类别特征。

[0199] 继续上述举例，针对文本A、B、C、D，每个文本对应4个权重值，且每个权重值与除该文本之外的一个文本对应，假设每个文本的类别特征是M维向量，则基于该4个文本的类别特征可以得到4×M的类别特征矩阵。针对文本A，其对应的4个权重值可以组成一个4×1的第三权重矩阵，可以将该4×1的第三权重矩阵的转置与该4×M的类别特征矩阵相乘，则可以得到一个1×M的矩阵，该1×M的矩阵是文本A的类别特征向量。同理，可以分别确定文本B、文本C和文本D的增强类别特征。

[0200] 在另一种实现方式中，针对第一文本，将该第一文本对应的每个权重值与该权重值对应的文本的类别特征进行加权融合，得到该第一文本的增强类别特征。

[0201] 继续上述举例，针对文本X、Y、Z、W，每个文本对应有4个权重值，且每个权重值与该4个文本的一个文本对应，即在该4个权重值中，有1个权重值与该文本自身对应，其他3个权重值与除该文本之外的其余文本对应。例如，文本X对应有4个权重值，分别是权重值c11、权重值c12、权重值c13和权重值c14，且权重值c11与文本X对应，权重值c12与文本Y对应，权重值c13与文本Z对应，权重值c14与文本W对应。针对文本X，其对应的4个权重值分别为c11，c12，c13和c14，可以将c11与文本X的类别特征相乘，将c12与文本Y的类别特征相乘，将c13与文本Z的类别特征相乘，将c14与文本W的类别特征相乘，将4个乘积相加作为文本X的增强类别特征。如此可以分别确定文本Y、文本Z、文本W的增强类别特征。

[0202] 示例性地，确定每个文本的增强类别特征后，基于多个文本的增强类别特征组成增强类别特征矩阵，基于增强类别特征矩阵和预设权重矩阵，确定待处理文档的类别特征向量。

[0203] 继续上述举例，假设确定了4个文本的增强类别特征，将该4个文本的增强类别特征组成4×M的增强类别特征矩阵，将4×1的预设权重矩阵的转置与该4×M的增强类别特征矩阵相乘，可以得到1×M的矩阵，则该矩阵是待处理文档的类别特征向量，融合了待处理文档中所有文本的类别特征。

[0204] 需要说明的是，步骤702是步骤206的一种具体实现方式。

[0205] 步骤704：将类别特征向量输入分类模型，确定待处理文档的类别。

[0206] 需要说明的是，步骤704的具体实现可以参见步骤208的相关描述，本实施例在此不再赘述。

[0207] 本申请实施例中，按照多个文本在待处理文档中的先后顺序，对多个文本的类别

特征进行拼接，能够得到符合待处理文档的行文逻辑，能够表征待处理文档中文本之间关联关系的类别特征向量，则基于该类别特征确定待处理文档的类别，不仅能够考虑到待处理文档的整体语义，还能够考虑到待处理文档上下文的关系，确定的类别会更加准确。

[0208] 图8示出了根据本申请一实施例提供的一种分割待处理文档的方法的流程图，具体包括以下步骤：

[0209] 步骤802：基于字符识别算法对待处理文档的内容进行识别，获取待处理文档的字符内容。

[0210] 其中，字符识别算法用于识别文档中的字符内容。例如，该字符识别算法可以是OCR算法或者PDF解析工具。

[0211] 在一些实施例中，可以通过PDF解析工具对待处理文档进行解析；或者，可以通过OCR算法对待处理文档进行字符识别；或者，可以将PDF解析工具和OCR算法融合起来对待处理文档进行字符识别，通过这几种方式均能够确定待处理文档的字符内容。

[0212] 作为一种示例，虽然基于OCR算法的字符识别整体效果比较好，但OCR算法可能存在特殊字符识别错误或者将一些符号错误识别为字等类似的问题。而PDF解析工具虽然识别字符比较准确，但是单独使用无法还原待处理文档的版式信息。因此，可以将这两种方式融合使用，即可解决OCR算法对特殊字符识别错误的问题，又可解决PDF解析工具无法还原待处理版式信息的问题，提升了字符内容识别的效果。

[0213] 步骤804：按照预设分割策略，对字符内容进行分割，得到多个文本。

[0214] 其中，预设分割策略可以是人工按照经验设置的用于将字符内容划分为多个文本的策略。

[0215] 在一些实施例中，预设分割策略可以是按照待处理文档的章节对待处理文档进行分割；或者，预设分割策略可以是按照待处理文档的段落对待处理文档进行分割；或者，预设分割策略可以是按照特定字符数量对字符内容进行划分，且在划分时需要保证文本内容的完整性；或者，可以将按照章节、按照段落和按照特定字符数量这三种方式相结合待处理文档进行分割。

[0216] 作为一种示例，可以按照章节编号，将第一章划分为一个文本，第二章划分为一个文本，以此类推；或者，可以按照段落，将第一段划分为一个文本，第二段划分为一个文本，以此类推；或者，先按照章节划分得到H个文本，再在每个文本中按照段落进行划分；或者，先按照段落划分得到K个文本，再在每个文本中按照特定字符数量进行划分；或者，先按照章节划分得到S个文本，再在每个文本中按照段落划分得到子文本，再在每个子文本中按照特定字符数量进行划分。

[0217] 需要说明的是，步骤802-步骤804是步骤202的一种具体实现方式。

[0218] 本申请实施例中，在对待处理文档进行分类之前，先通过OCR算法和PDF解析工具对待处理文档进行字符识别，得到待处理文档的字符内容，然后将待处理文档的字符内容按照预设分割策略进行分割，得到多个文本，解决了长文档无法直接输入模型进行分类处理的问题。

[0219] 下述结合附图9，以本申请提供的文档分类方法在合同文档识别问题上的应用为例，对所述文档分类方法进行进一步说明。其中，图9示出了本申请一实施例提供的一种应用于合同文档识别的文档分类方法的处理流程图，具体包括以下步骤：

[0220] 步骤902:基于字符识别算法对待处理文档的内容进行识别,获取待处理文档的字符内容。

[0221] 其中,该字符识别算法用于识别文档中的字符内容。

[0222] 以待处理文档是PDF文档为例,可以先通过OCR算法和PDF解析工具结合的方法对待处理文档的内容进行识别,即提取出待处理文档中的字符内容。

[0223] 步骤904:按照预设分割策略,对字符内容进行分割,得到多个文本。

[0224] 继续上述举例,可以按照BERT模型能够处理的最大文本长度分割,且要保证句子的完整性。例如,最大长度是510,则可以将每510个字符划分为一个文本,但若到达第510个字符时是半句话,则从不足510个字符的该句话的结尾处分割。

[0225] 例如,参见图10,图10是本申请一实施例提供的一种文档分类方法的处理过程示意图。在图10中,输入BERT模型的是N个文本。

[0226] 步骤906:将多个文本分别输入特征提取模型,对每个文本进行分词处理,得到每个文本的词单元。

[0227] 例如,特征提取模型可以是BERT模型。

[0228] 步骤908:对每个文本的词单元分别进行词嵌入处理,得到每个文本中词单元的词嵌入向量。

[0229] 步骤910:针对任一文本,通过特征提取模型的词级注意力层,将该文本的第一词单元的词嵌入向量与该文本中每个词单元的词嵌入向量进行注意力计算,确定该文本的特征向量。

[0230] 步骤912:将该文本的特征向量与多个文本中每个文本的特征向量进行注意力计算,确定该文本的增强特征向量。

[0231] 步骤914:将该文本的增强特征向量输入特征提取模型的全连接层,确定该文本的类别特征。

[0232] 步骤916:按照多个文本在待处理文档中的先后顺序,对多个文本的类别特征进行拼接,得到待处理文档的类别特征向量。

[0233] 例如,参见图10,通过BERT模型进行处理后,能够得到N个类别特征,将该N个类别特征按顺序拼接,可以得到待处理文档的类别特征向量。

[0234] 步骤918:将待处理文档的类别特征向量输入分类模型,确定待处理文档的类别。

[0235] 例如,分类模型可以是Lightgbm,参见图10,将类别特征向量输入该Lightgbm模型,可以得到该待处理文档是合同的概率和非合同的概率,若是合同的概率大于非合同的概率,则确定该待处理文档的类别是合同,若是合同的概率小于非合同的概率,则确定该待处理文档的类别不是合同,若是合同的概率和非合同的概率相同,则需要重新确定该待处理文档的类别。

[0236] 本申请提供的文档分类方法,对待处理文档进行分割,得到多个文本;将多个文本分别输入特征提取模型,确定每个文本的类别特征;对多个文本的类别特征进行组合,得到待处理文档的类别特征向量;将类别特征向量输入分类模型,可以确定该待处理文档是否是合同。上述方法先将待处理文档分割成比较短的文本,适用于长文档处理,并且先确定每个文本的类别特征,然后将多个文本的类别特征组合得到待处理文档的类别特征向量,则可以认为该类别特征向量融合了待处理文档全文的类别信息,即该类别特征向量不仅能够

体现待处理文档中各部分内容的类别特征，还能够体现待处理文档中各部分内容之间的关联，因此将该类别特征向量输入分类模型进行分类，能够给分类模型提供更多的信息，使得分类模型的分类结果更加准确，即提高了识别合同文档的准确率。另外，提高了识别准确率，进而会提高用户的使用体验，用户对该方法的使用频率就会提高，进而会提高合同审核等后续任务的转化率。

[0237] 与上述方法实施例相对应，本申请还提供了文档分类装置实施例，图11示出了本申请一实施例提供的一种文档分类装置的结构示意图。如图11所示，该装置包括：

[0238] 分割模块1102，被配置为对待处理文档进行分割，得到多个文本；

[0239] 第一确定模块1104，被配置为将所述多个文本分别输入特征提取模型，确定每个文本的类别特征；

[0240] 组合模块1106，被配置为对所述多个文本的类别特征进行组合，得到所述待处理文档的类别特征向量；

[0241] 第二确定模块1108，被配置为将所述类别特征向量输入分类模型，确定所述待处理文档的类别。

[0242] 在本申请一种可能的实现方式中，所述特征提取模型包括输入层、嵌入层和输出层，第一确定模块1104进一步被配置为：

[0243] 通过所述输入层，对所述每个文本进行分词处理，得到每个文本的词单元；

[0244] 通过所述嵌入层，对每个文本的词单元分别进行词嵌入处理，得到每个文本中词单元的词嵌入向量；

[0245] 针对任一文本，基于该文本中词单元的词嵌入向量，确定该文本的类别特征。

[0246] 在本申请一种可能的实现方式中，所述输出层包括词级注意力层和全连接层，第一确定模块1104进一步被配置为：

[0247] 针对任一文本，通过所述词级注意力层，将该文本的第一词单元的词嵌入向量与该文本中每个词单元的词嵌入向量进行注意力计算，确定该文本的特征向量，其中，所述第一词单元是该文本中的任一词单元；

[0248] 通过所述全连接层，基于该文本的特征向量确定该文本的类别特征。

[0249] 在本申请一种可能的实现方式中，所述输出层还包括文本级注意力层，第一确定模块1104进一步被配置为：

[0250] 通过所述文本级注意力层，将该文本的特征向量与多个文本中每个文本的特征向量进行注意力计算，确定该文本的增强特征向量；

[0251] 通过所述全连接层，基于该文本的特征向量确定该文本的类别特征，包括：

[0252] 通过所述全连接层，基于该文本的增强特征向量确定该文本的类别特征。

[0253] 在本申请一种可能的实现方式中，组合模块1106进一步被配置为：

[0254] 按照所述多个文本在所述待处理文档中的先后顺序，对所述多个文本的类别特征进行拼接，得到所述待处理文档的类别特征向量。

[0255] 在本申请一种可能的实现方式中，所述特征提取模型包括BERT模型。

[0256] 在本申请一种可能的实现方式中，所述装置还包括分类模型训练模块，所述分类模型训练模块被配置为：

[0257] 获取多个样本文档，其中，每个样本文档对应一个类别特征向量；

[0258] 基于多个类别特征向量构建第一决策树，并基于所述第一决策树确定每个样本文档的预测概率；

[0259] 基于每个样本文档的预测概率和多个类别特征向量构建第二决策树，并基于第二决策树确定每个样本文档的预测概率，以此类推，直到达到停止条件，将构建的多个决策树确定为分类模型。

[0260] 在本申请一种可能的实现方式中，所述分类模型包括Lightgbm模型，且所述分类模型的损失函数是对数损失函数。

[0261] 在本申请一种可能的实现方式中，分割模块1102进一步被配置为：

[0262] 基于字符识别算法对所述待处理文档的内容进行识别，获取所述待处理文档的字符内容，其中，所述字符识别算法用于识别文档中的字符内容；

[0263] 按照预设分割策略，对所述字符内容进行分割，得到所述多个文本。

[0264] 本申请提供的文档分类装置，对待处理文档进行分割，得到多个文本；将所述多个文本分别输入特征提取模型，确定每个文本的类别特征；对所述多个文本的类别特征进行组合，得到所述待处理文档的类别特征向量；将所述类别特征向量输入分类模型，确定所述待处理文档的类别。如此，先将待处理文档分割成比较短的文本，适用于长文档处理，并且先确定每个文本的类别特征，然后将多个文本的类别特征组合得到待处理文档的类别特征向量，则可以认为该类别特征向量融合了待处理文档全文的类别信息，即该类别特征向量不仅能够体现待处理文档中各部分内容的类别特征，还能够体现待处理文档中各部分内容之间的关联，因此将该类别特征向量输入分类模型进行分类，能够给分类模型提供更多的信息，使得分类模型的分类结果更加准确，即提高了文档分类的准确率。

[0265] 上述为本实施例的一种文档分类装置的示意性方案。需要说明的是，该文档分类装置的技术方案与上述的文档分类方法的技术方案属于同一构思，文档分类装置的技术方案未详细描述的细节内容，均可以参见上述文档分类方法的技术方案的描述。此外，装置实施例中的各组成部分应当理解为实现该程序流程各步骤或该方法各步骤所必须建立的功能模块，各个功能模块并非实际的功能分割或者分离限定。由这样一组功能模块限定的装置权利要求应当理解为主要通过说明书记载的计算机程序实现该解决方案的功能模块构架，而不应当理解为主要通过硬件方式实现该解决方案的实体装置。

[0266] 图12示出了根据本申请一实施例提供的一种计算设备1200的结构框图。该计算设备1200的部件包括但不限于存储器1210和处理器1220。处理器1220与存储器1210通过总线1230相连接，数据库1250用于保存数据。

[0267] 计算设备1200还包括接入设备1240，接入设备1240使得计算设备1200能够经由一个或多个网络1260通信。这些网络的示例包括公用交换电话网(PSTN，Public Switched Telephone Network)、局域网(LAN，LocalAreaNetwork)、广域网(WAN，Wide Area Network)、个域网(PAN，Personal Area Network)或诸如因特网的通信网络的组合。接入设备120可以包括有线或无线的任何类型的网络接口(例如，网络接口卡(NIC，Network Interface Controller))中的一个或多个，诸如IEEE802.11无线局域网(WLAN，Wireless Local Area Network)无线接口、全球微波互联接入(Wi-MAX)接口、以太网接口、通用串行总线(USB，Universal Serial Bus)接口、蜂窝网络接口、蓝牙接口、近场通信(NFC，Near Field Communication)接口，等等。

[0268] 在本申请的一个实施例中，计算设备1200的上述部件以及图12中未示出的其他部件也可以彼此相连接，例如通过总线。应当理解，图12所示的计算设备结构框图仅仅是出于示例的目的，而不是对本申请范围的限制。本领域技术人员可以根据需要，增添或替换其他部件。

[0269] 计算设备1200可以是任何类型的静止或移动计算设备，包括移动计算机或移动计算设备(例如，平板计算机、个人数字助理、膝上型计算机、笔记本计算机、上网本等)、移动电话(例如，智能手机)、可佩戴的计算设备(例如，智能手表、智能眼镜等)或其他类型的移动设备，或者诸如台式计算机或PC(Personal Computer)的静止计算设备。计算设备1200还可以是移动式或静止式的服务器。

[0270] 其中，处理器1220用于执行所述文档分类方法的计算机可执行指令。

[0271] 上述为本实施例的一种计算设备的示意性方案。需要说明的是，该计算设备的技术方案与上述的文档分类方法的技术方案属于同一构思，计算设备的技术方案未详细描述的细节内容，均可以参见上述文档分类方法的技术方案的描述。

[0272] 本申请一实施例还提供一种计算机可读存储介质，其存储有计算机指令，该指令被处理器执行时以用于文档分类方法。

[0273] 上述为本实施例的一种计算机可读存储介质的示意性方案。需要说明的是，该存储介质的技术方案与上述的文档分类方法的技术方案属于同一构思，存储介质的技术方案未详细描述的细节内容，均可以参见上述文档分类方法的技术方案的描述。

[0274] 本申请一实施例还提供一种芯片，其存储有计算机程序，该计算机程序被芯片执行时实现所述文档分类方法的步骤。

[0275] 上述对本申请特定实施例进行了描述。其它实施例在所附权利要求书的范围内。在一些情况下，在权利要求书中记载的动作或步骤可以按照不同于实施例中的顺序来执行并且仍然可以实现期望的结果。另外，在附图中描绘的过程不一定要求示出的特定顺序或者连续顺序才能实现期望的结果。在某些实施方式中，多任务处理和并行处理也是可以的或者可能是有利的。

[0276] 所述计算机指令包括计算机程序代码，所述计算机程序代码可以为源代码形式、对象代码形式、可执行文件或某些中间形式等。所述计算机可读存储介质可以包括：能够携带所述计算机程序代码的任何实体或装置、记录介质、U盘、移动硬盘、磁碟、光盘、计算机存储器、只读存储器(ROM，Read-Only Memory)、随机存取存储器(RAM，Random Access Memory)、电载波信号、电信信号以及软件分发介质等。

[0277] 需要说明的是，对于前述的各方法实施例，为了简便描述，故将其都表述为一系列的动作组合，但是本领域技术人员应该知悉，本申请并不受所描述的动作顺序的限制，因为依据本申请，某些步骤可以采用其它顺序或者同时进行。其次，本领域技术人员也应该知悉，说明书中所描述的实施例均属于优选实施例，所涉及的动作和模块并不一定都是本申请所必须的。

[0278] 在上述实施例中，对各个实施例的描述都各有侧重，某个实施例中没有详述的部分，可以参见其它实施例的相关描述。

[0279] 以上公开的本申请优选实施例只是用于帮助阐述本申请。可选实施例并没有详尽叙述所有的细节，也不限制该发明仅为所述的具体实施方式。显然，根据本申请的内容，可

作很多的修改和变化。本申请选取并具体描述这些实施例，是为了更好地解释本申请的原理和实际应用，从而使所属技术领域技术人员能很好地理解和利用本申请。本申请仅受权利要求书及其全部范围和等效物的限制。

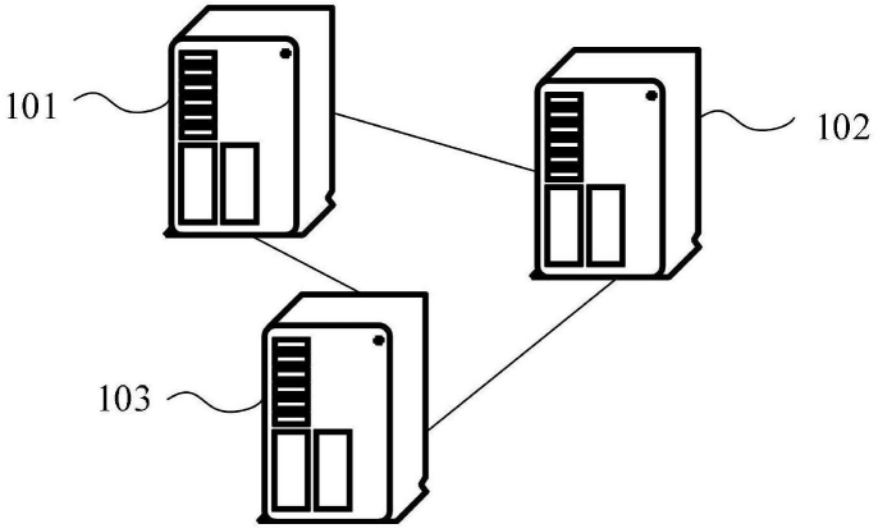

101
102
103

图1

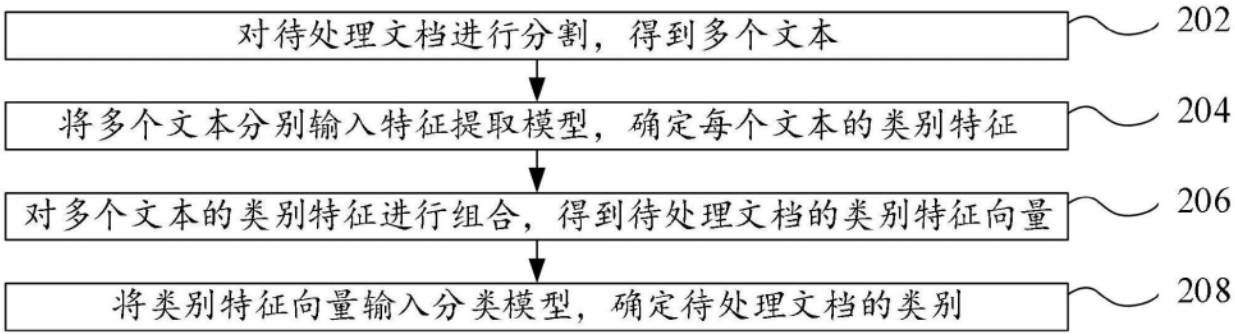

对待处理文档进行分割，得到多个文本
202
将多个文本分别输入特征提取模型，确定每个文本的类别特征
204
对多个文本的类别特征进行组合，得到待处理文档的类别特征向量
206
将类别特征向量输入分类模型，确定待处理文档的类别
208

图2

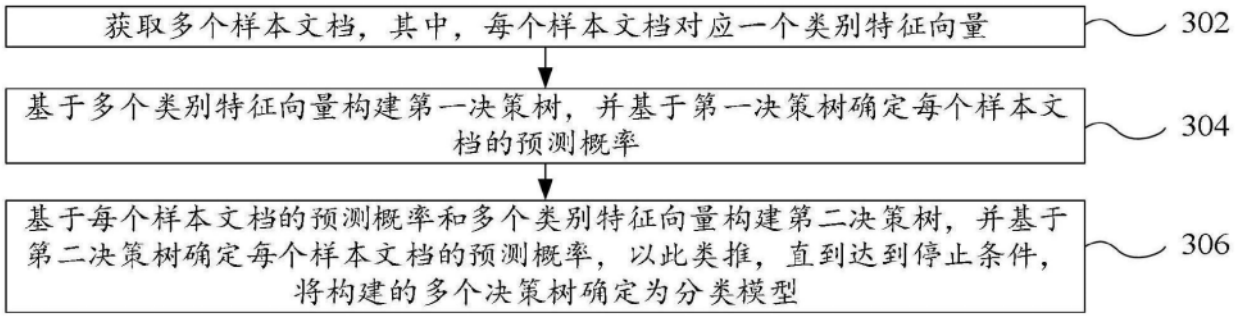

获取多个样本文档，其中，每个样本文档对应一个类别特征向量
302
基于多个类别特征向量构建第一决策树，并基于第一决策树确定每个样本文档的预测概率
304
基于每个样本文档的预测概率和多个类别特征向量构建第二决策树，并基于第二决策树确定每个样本文档的预测概率，以此类推，直到达到停止条件，将构建的多个决策树确定为分类模型
306

图3

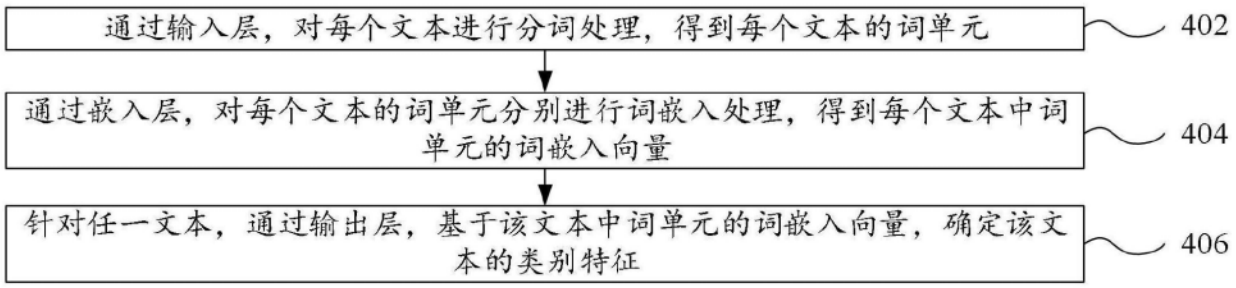

通过输入层，对每个文本进行分词处理，得到每个文本的词单元
402
通过嵌入层，对每个文本的词单元分别进行词嵌入处理，得到每个文本中词单元的词嵌入向量
404
针对任一文本，通过输出层，基于该文本中词单元的词嵌入向量，确定该文本的类别特征
406

图4

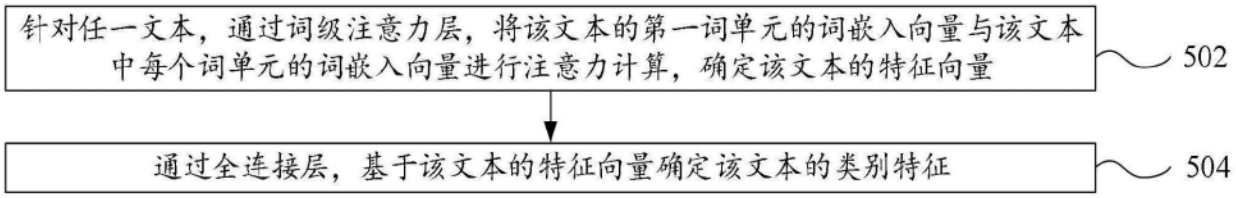

针对任一文本，通过词级注意力层，将该文本的第一词单元的词嵌入向量与该文本中每个词单元的词嵌入向量进行注意力计算，确定该文本的特征向量
502
通过全连接层，基于该文本的特征向量确定该文本的类别特征
504

图5

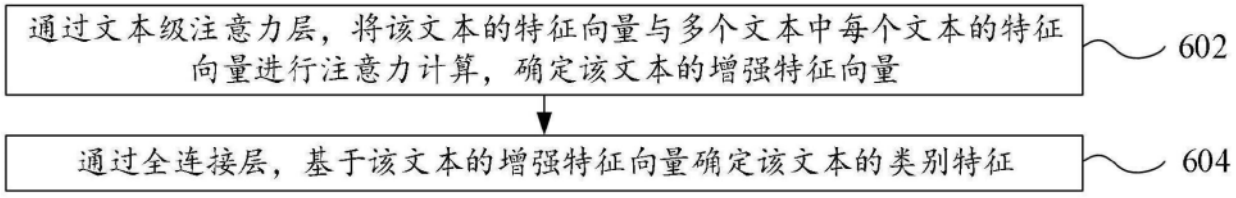

通过文本级注意力层，将该文本的特征向量与多个文本中每个文本的特征向量进行注意力计算，确定该文本的增强特征向量
602
通过全连接层，基于该文本的增强特征向量确定该文本的类别特征
604

图6

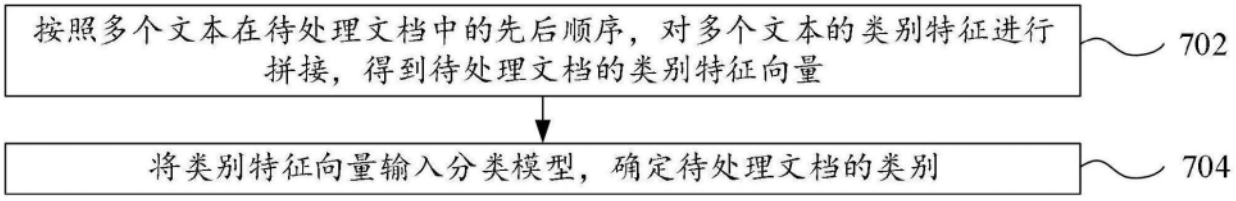

按照多个文本在待处理文档中的先后顺序，对多个文本的类别特征进行拼接，得到待处理文档的类别特征向量
702
将类别特征向量输入分类模型，确定待处理文档的类别
704

图7

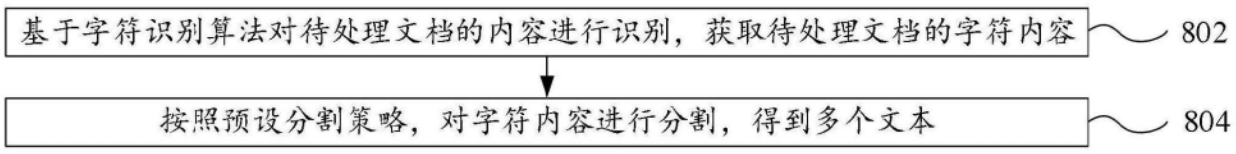

基于字符识别算法对待处理文档的内容进行识别，获取待处理文档的字符内容
802
按照预设分割策略，对字符内容进行分割，得到多个文本
804

图8

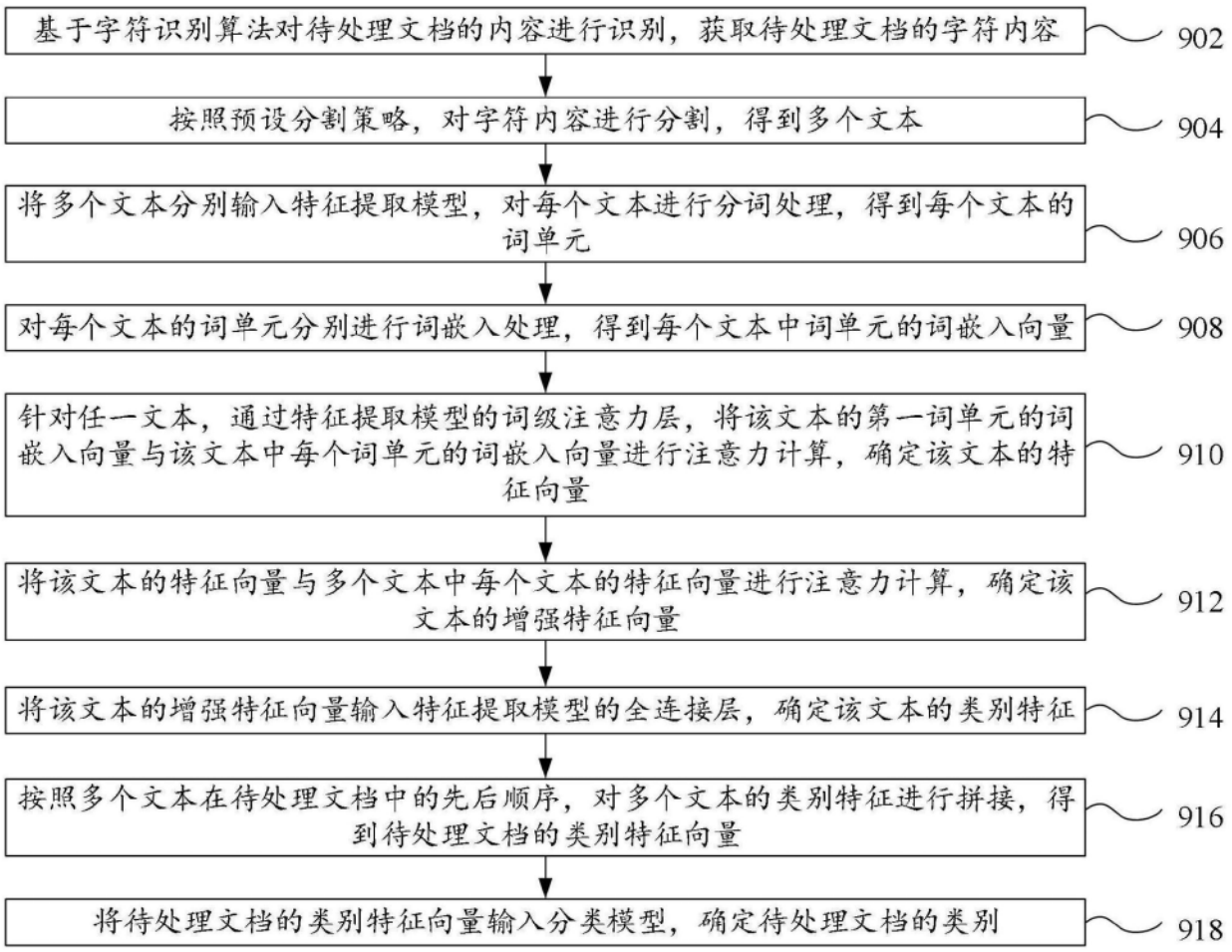

基于字符识别算法对待处理文档的内容进行识别，获取待处理文档的字符内容
902
按照预设分割策略，对字符内容进行分割，得到多个文本
904
将多个文本分别输入特征提取模型，对每个文本进行分词处理，得到每个文本的词单元
906
对每个文本的词单元分别进行词嵌入处理，得到每个文本中词单元的词嵌入向量
908
针对任一文本，通过特征提取模型的词级注意力层，将该文本的第一词单元的词嵌入向量与该文本中每个词单元的词嵌入向量进行注意力计算，确定该文本的特征向量
910
将该文本的特征向量与多个文本中每个文本的特征向量进行注意力计算，确定该文本的增强特征向量
912
将该文本的增强特征向量输入特征提取模型的全连接层，确定该文本的类别特征
914
按照多个文本在待处理文档中的先后顺序，对多个文本的类别特征进行拼接，得到待处理文档的类别特征向量
916
将待处理文档的类别特征向量输入分类模型，确定待处理文档的类别
918

图9

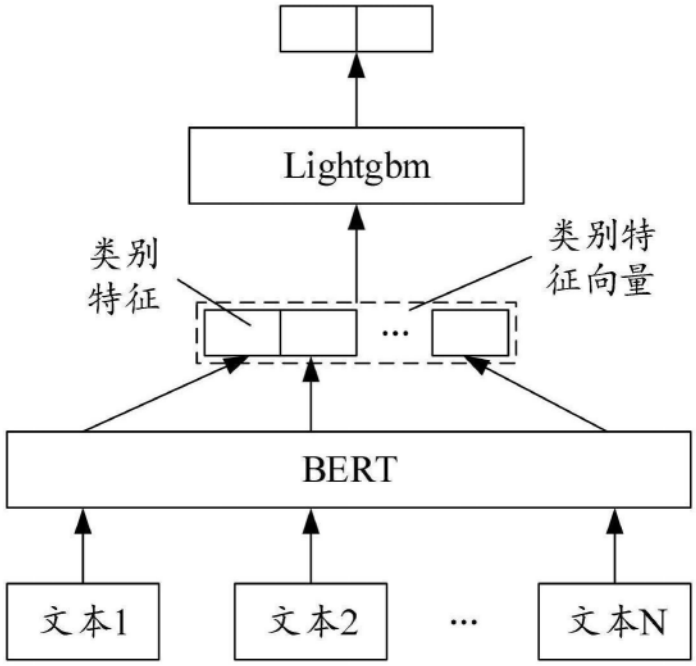

Lightgbm
类别
特征
类别特
征向量
BERT
文本1
文本2
文本N

图10

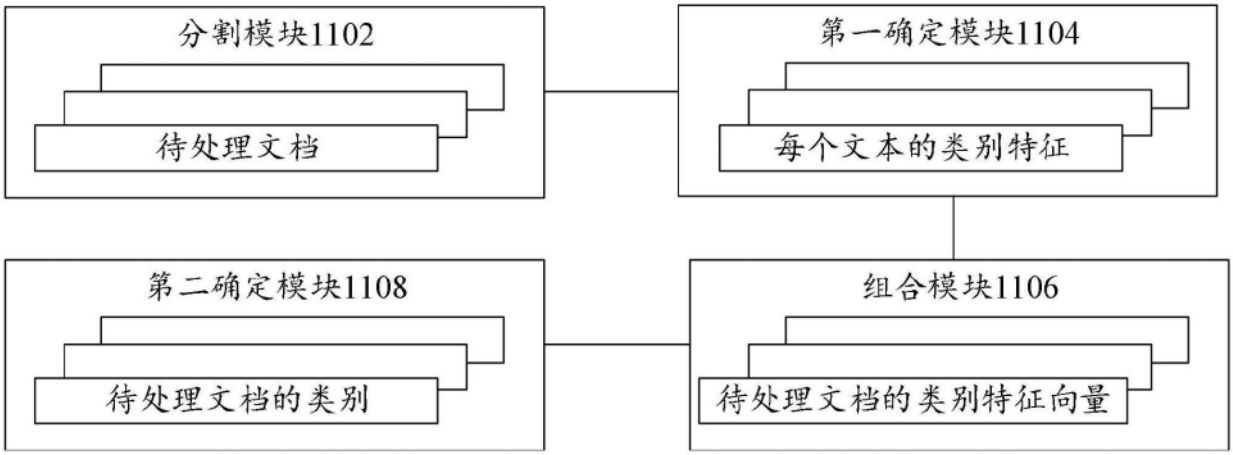

分割模块1102
待处理文档
第一确定模块1104
每个文本的类别特征
第二确定模块1108
待处理文档的类别
组合模块1106
待处理文档的类别特征向量

图11

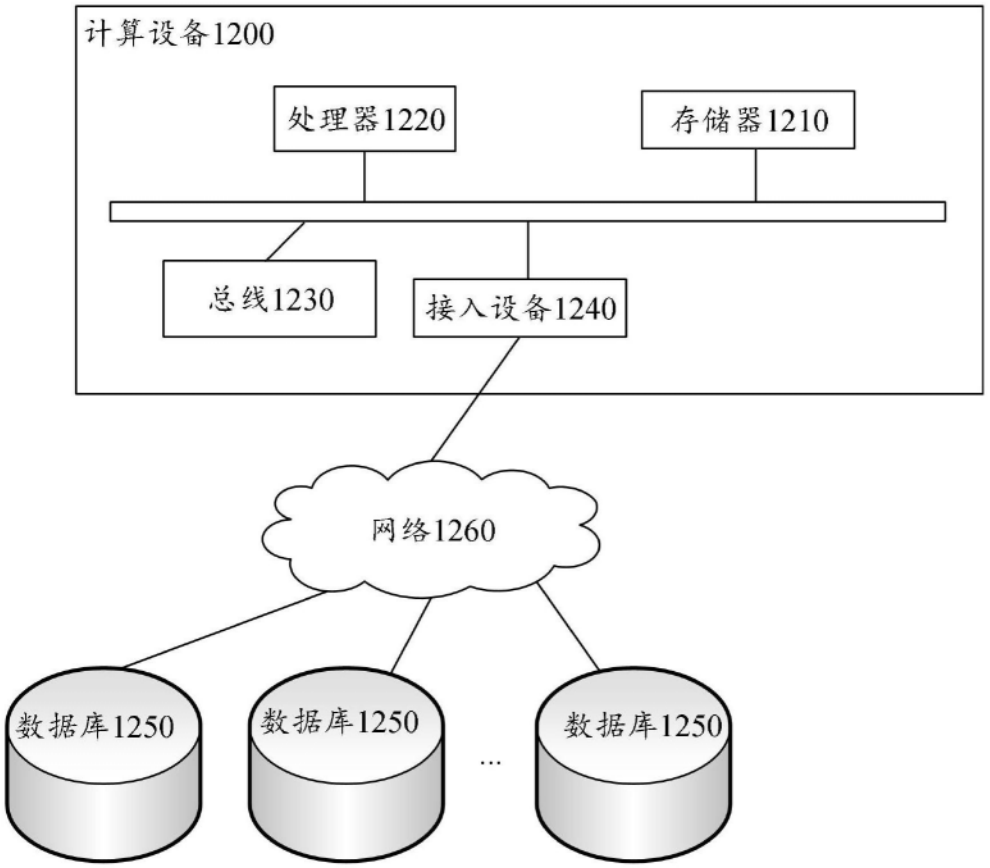

计算设备1200
处理器1220
存储器1210
总线1230
接入设备1240
网络1260
数据库1250
数据库1250
...
数据库1250

图12