(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 113239190 B
(45)授权公告日 2024.02.20

(21)申请号 202110462274.2

(22)申请日 2021.04.27

(65)同一申请的已公布的文献号
申请公布号 CN 113239190 A

(43)申请公布日 2021.08.10

(73)专利权人 天九共享网络科技集团有限公司
地址 100012 北京市朝阳区安立路30号仰山公园8号楼

(72)发明人 喻银根

(74)专利代理机构 北京英创嘉友知识产权代理事务所(普通合伙) 11447
专利代理师 李柯莹

(51)Int.Cl.
*G06F 16/35*(2019.01)
*G06F 40/30*(2020.01)
*G06F 18/214*(2023.01)
*G06N 3/0464*(2023.01)
*G06N 3/044*(2023.01)

(56)对比文件
CN 111414336 A,2020.07.14
CN 111753060 A,2020.10.09
CN 110427609 A,2019.11.08
CN 109582794 A,2019.04.05
CN 109791594 A,2019.05.21
CN 110134786 A,2019.08.16
CN 110263323 A,2019.09.20
CN 110598191 A,2019.12.20
CN 110717042 A,2020.01.21
CN 110825848 A,2020.02.21
CN 111488556 A,2020.08.04
CN 111507099 A,2020.08.07
CN 112231645 A,2021.01.15
CN 112463933 A,2021.03.09
CN 112597312 A,2021.04.02
CN 110413783 A,2019.11.05
CN 107665248 A,2018.02.06
卢毓海等.基于颜色聚类的计算机桌面图像压缩算法.《 计算机工程》.2012,第221-225页,第236页.

审查员 唐文俊

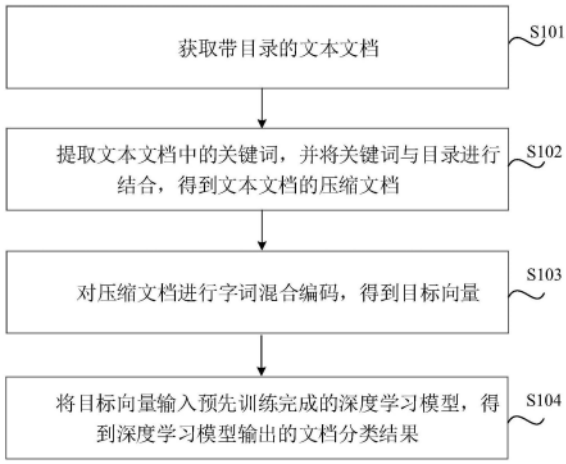
权利要求书2页 说明书6页 附图2页

(54)发明名称

# 文档分类方法、装置、存储介质及电子设备

(57)摘要

本公开涉及一种文档分类方法、装置、存储介质及电子设备。该方法包括:获取带目录的文本文档;提取文本文档中的关键词,并将关键词与所述目录进行结合,得到文本文档的压缩文档;对压缩文档进行字词混合编码,得到目标向量;将目标向量输入预先训练完成的深度学习模型,得到深度学习模型输出的文档分类结果。本公开实施例通过提取关键词,并将关键词与目录进行结合的方式,实现了在保留文本核心内容的基础上,对文本进行了压缩,降低了深度学习模型因文本过长而对文本进行截断,导致大部分文本核心信息丢失的概率。




CN 113239190 B

1.一种文档分类方法,其特征在于,所述方法包括:

获取带目录的文本文档;

提取所述文本文档中的关键词,并将所述关键词与所述目录进行结合,得到所述文本文档的压缩文档;

对所述压缩文档进行字词混合编码,得到目标向量;

将所述目标向量输入预先训练完成的深度学习模型,得到所述深度学习模型输出的文档分类结果;

其中,所述将所述关键词与所述目录进行结合,包括:

在结合所述关键词与所述目录时,将所述关键词放在所述目录之前,使得所述关键词能够优先于所述目录进行编码;

所述对所述压缩文档进行字词混合编码,得到目标向量,包括:

根据词向量映射模型,将所述压缩文档中的每一词映射为词向量,以及将所述压缩文档中的每一字随机初始化为字向量,其中,所述词向量与所述字向量的维度相同,所述词向量映射模型是预先基于无监督方法训练形成的word2vec的词向量映射模型;

通过冗余方式将所述词向量和所述字向量进行混合,得到所述目标向量。

2.根据权利要求1所述的方法,其特征在于,所述深度学习模型包括bert模型层、双向LSTM模型层、卷积层以及softmax模型层;

其中,所述bert模型层与所述双向LSTM模型层相结合能够提取所述压缩文档的语义特征;

所述bert模型层与所述卷积层相结合能够提取所述压缩文档的深度特征,并结合最大池化方式和平均池化方式对提取到的所述深度特征进行池化;

所述softmax模型层用于输出所述文档分类结果。

3.根据权利要求2所述的方法,其特征在于,所述bert模型层的参数在所述深度学习模型的训练过程中保持冻结,和/或,所述卷积层包括多层,且每一层具有不同的卷积核。

4.一种文档分类装置,其特征在于,所述装置包括:

获取模块,用于获取带目录的文本文档;

结合模块,用于提取所述文本文档中的关键词,并将所述关键词与所述目录进行结合,得到所述文本文档的压缩文档;

编码模块,用于对所述压缩文档进行字词混合编码,得到目标向量;

生成模块,用于将所述目标向量输入预先训练完成的深度学习模型,得到所述深度学习模型输出的文档分类结果;

其中,所述将所述关键词与所述目录进行结合,包括:

在结合所述关键词与所述目录时,将所述关键词放在所述目录之前,使得所述关键词能够优先于所述目录进行编码;

所述编码模块包括:

映射子模块,用于根据词向量映射模型,将所述压缩文档中的每一词映射为词向量,以及将所述压缩文档中的每一字随机初始化为字向量;

混合子模块,用于通过冗余方式将所述词向量和所述字向量进行混合,得到所述目标向量。

5.根据权利要求4所述的装置,其特征在于,所述深度学习模型包括bert模型层、双向LSTM模型层、卷积层以及softmax模型层;

其中,所述bert模型层与所述双向LSTM模型层相结合能够提取所述压缩文档的语义特征;

所述bert模型层与所述卷积层相结合能够提取所述压缩文档的深度特征,并结合最大池化方式和平均池化方式对提取到的所述深度特征进行池化;

所述softmax模型层用于输出所述文档分类结果。

6.一种计算机可读存储介质,其上存储有计算机程序,其特征在于,该程序被处理器执行时实现权利要求1-3中任一项所述方法的步骤。

7.一种电子设备,其特征在于,包括:

存储器,其上存储有计算机程序;

处理器,用于执行所述存储器中的所述计算机程序,以实现权利要求1-3中任一项所述方法的步骤。

# 文档分类方法、装置、存储介质及电子设备

## 技术领域

[0001] 本公开涉及自然语言处理技术领域,具体地,涉及一种文档分类方法、装置、存储介质及电子设备。

## 背景技术

[0002] 深度学习是学习样本数据的内在规律和表示层次,这些学习过程中获得的信息对诸如文字,图像和声音等数据的解释有很大的帮助。它的最终目标是让机器能够像人一样具有分析学习能力,能够识别文字、图像和声音等数据。深度学习使机器模仿视听和思考等人类的活动,解决了很多复杂的模式识别难题,使得人工智能相关技术取得了很大进步。如今的深度学习技术还有一个问题,它需要提取大量的数据作为训练基础,而训练所得的结果却难以应用到其他问题上。

[0003] 现阶段的文本分类是通过利用文本字词词频等特征输入模型中,利用机器学习方法进行分类,但是会存在着提取的文本特征较浅的问题,而且很难学习到文本的语义信息;还有一种方法是对文本的字词向量化后作为输入,利用深度神经网络的方法进行分类,但是这种方法也存在着训练的硬件成本和时间成本较高的问题,当文档内容过长时,其输入存在对文本的截断,会丢失掉部分信息。

## 发明内容

[0004] 本公开的目的是提供一种文档分类方法、装置、存储介质及电子设备,以解决上述问题。

[0005] 为了实现上述目的,第一方面,本公开实施例提供一种文档分类方法,包括:

[0006] 获取带目录的文本文档;

[0007] 提取所述文本文档中的关键词,并将所述关键词与所述目录进行结合,得到所述文本文档的压缩文档;

[0008] 对所述压缩文档进行字词混合编码,得到目标向量;

[0009] 将所述目标向量输入预先训练完成的深度学习模型,得到所述深度学习模型输出的文档分类结果。

[0010] 可选地,所述对所述压缩文档进行字词混合编码,得到目标向量,包括:

[0011] 根据词向量映射模型,将所述压缩文档中的每一词映射为词向量,以及将所述压缩文档中的每一字随机初始化为字向量,其中,所述词向量与所述字向量的维度相同,所述词向量映射模型是预先基于无监督方法训练形成的word2vec的词向量映射模型;

[0012] 通过冗余方式将所述词向量和所述字向量进行混合,得到所述目标向量。

[0013] 可选地,所述深度学习模型包括bert模型层、双向LSTM模型层、卷积层以及softmax模型层;

[0014] 其中,所述bert模型层与所述双向LSTM模型层相结合能够提取所述压缩文档的语义特征;

[0015] 所述bert模型层与所述卷积层相结合能够提取所述压缩文档的深度特征,并结合最大池化方式和平均池化方式对提取到的所述深度特征进行池化;

[0016] 所述softmax模型层用于输出所述文档分类结果。

[0017] 可选地,所述bert模型层的参数在所述深度学习模型的训练过程中保持冻结,和/或,所述卷积层包括多层,且每一层具有不同的卷积核。

[0018] 可选地,所述将所述关键词与所述目录进行结合,包括:

[0019] 在结合所述关键词与所述目录时,将所述关键词放在所述目录之前,使得所述关键词能够优先于所述目录进行编码。

[0020] 第二方面,本公开实施例提供一种文档分类装置,所述装置包括:

[0021] 获取模块,用于获取带目录的文本文档;

[0022] 结合模块,用于提取所述文本文档中的关键词,并将所述关键词与所述目录进行结合,得到所述文本文档的压缩文档;

[0023] 编码模块,用于对所述压缩文档进行字词混合编码,得到目标向量;

[0024] 生成模块,用于将所述目标向量输入预先训练完成的深度学习模型,得到所述深度学习模型输出的文档分类结果。

[0025] 可选地,所述编码模块包括:

[0026] 映射子模块,用于根据词向量映射模型,将所述压缩文档中的每一词映射为词向量,以及将所述压缩文档中的每一字随机初始化为字向量;

[0027] 混合子模块,用于通过冗余方式将所述词向量和所述字向量进行混合,得到所述目标向量。

[0028] 可选地,所述深度学习模型包括bert模型层、双向LSTM模型层、卷积层以及softmax模型层;

[0029] 其中,所述bert模型层与所述双向LSTM模型层相结合能够提取所述压缩文档的语义特征;

[0030] 所述bert模型层与所述卷积层相结合能够提取所述压缩文档的深度特征,并结合最大池化方式和平均池化方式对提取到的所述深度特征进行池化;

[0031] 所述softmax模型层用于输出所述文档分类结果。

[0032] 第三方面,本公开实施例提供一种计算机可读存储介质,其上存储有计算机程序,该程序被处理器执行时实现本公开第一方面所述方法的步骤。

[0033] 第四方面,本公开实施例提供一种电子设备,包括:

[0034] 存储器,其上存储有计算机程序;

[0035] 处理器,用于执行所述存储器中的所述计算机程序,以实现本公开第一方面所述方法的步骤。

[0036] 本公开的实施例提供的技术方案可以包括以下有益效果:

[0037] 采用上述技术方案,在获取带目录的文本文档后,首先提取所述文本文档中的关键词并将所述关键词与所述目录进行结合得到所述文本文档的压缩文档,之后对所述压缩文档进行字词混合编码,将所述目标向量输入预先训练完成的深度学习模型,最后得到所述深度学习模型输出的文档分类结果。本公开实施例通过提取关键词,并将关键词与目录进行结合的方式,实现了在保留文本核心内容的基础上,对文本进行了压缩,降低了深度学

习模型因文本过长而对文本进行截断,导致大部分文本核心信息丢失的概率。

[0038] 本公开的其他特征和优点将在随后的具体实施方式部分予以详细说明。

## 附图说明

[0039] 附图是用来提供对本公开的进一步理解,并且构成说明书的一部分,与下面的具体实施方式一起用于解释本公开,但并不构成对本公开的限制。在附图中:

[0040] 图1是根据一示例性实施例示出的一种文档分类方法的流程图。

[0041] 图2是根据一示例性实施例示出的一种文档分类装置的框图。

[0042] 图3是根据一示例性实施例示出的一种电子设备的框图。

## 具体实施方式

[0043] 以下结合附图对本公开的具体实施方式进行详细说明。应当理解的是,此处所描述的具体实施方式仅用于说明和解释本公开,并不用于限制本公开。

[0044] 本公开实施例提供一种文档分类的方法,如图1所示,该方法包括:

[0045] 在步骤S101中,获取带目录的文本文档。

[0046] 在本公开实施例中,带目录的文本文档为从原始文档中提取出的带有相关目录和正文的文本文档,其中,对原始文档提取的格式例如可以是word格式或PDF格式,本公开实施对此不做限定。进一步地,可以将word格式或PDF格式的原始文档转为文本文档,并且还可以将文本文档的目录与正文进行分离,以便后续步骤能够直接对目录进行使用。

[0047] 在步骤S102中,提取文本文档中的关键词,并将关键词与目录进行结合,得到所述文本文档的压缩文档。

[0048] 其中,关键词例如可以包括正文中的关键词,还可以包括正文中的关键词以及目录中的关键词。

[0049] 由于,目录中也会包含着文本的关键信息和一定的语义信息,因此,将提取到的关键词以及目录进行结合,可以在不损失文本的核心内容的情况下,实现对文本的压缩。

[0050] 在本公开实施例中,可以通过tf-idf方法但不限于通过此方法来提取文本文档中的关键词,例如还可以通过改进方法tf-iwf,或者潜在语义分析等方法来实现关键词提取,本公开实施例并不限制提取关键词的方法。此外,在具体实施时,为了保证压缩文档的长度可控,可以设定关键词数量阈值,这样,在提取关键词时,可以只提取满足关键词数量阈值要求的关键词,例如,提取文本文档中top30的关键词。

[0051] 在步骤S103中,对压缩文档进行字词混合编码,得到目标向量。

[0052] 示例地,字词混合编码作为得到目标向量的方式是根据训练好的词向量模型将每个词映射768维的词向量,将每个字随机初始化为768维的字向量,再通过将词向量和字向量进行混合得到上述目标向量。

[0053] 在步骤S104中,将目标向量输入预先训练完成的深度学习模型,得到深度学习模型输出的文档分类结果。

[0054] 采用上述方法,在获取带目录的文本文档后,首先提取所述文本文档中的关键词并将所述关键词与所述目录进行结合得到所述文本文档的压缩文档,之后对所述压缩文档进行字词混合编码,将所述目标向量输入预先训练完成的深度学习模型,最后得到所述深

度学习模型输出的文档分类结果。本公开实施例通过提取关键词,并将关键词与目录进行结合的方式,实现了在保留文本核心内容的基础上,对文本进行了压缩,降低了深度学习模型因文本过长而对文本进行截断,导致大部分文本核心信息丢失的概率。

[0055] 在一种可能的实施方式中,所述将所述关键词与所述目录进行结合,包括:

[0056] 在结合所述关键词与所述目录时,将所述关键词放在所述目录之前,使得所述关键词能够优先于所述目录进行编码。这样,即便在压缩文档的长度依然过长的情况下,也能够减少压缩文档对关键词的截断,从而实现最大程度上的保留文本的核心内容。

[0057] 在一种可能的实施方式中,所述对所述压缩文档进行字词混合编码,得到目标向量,包括:

[0058] 根据词向量映射模型,将所述压缩文档中的每一词映射为词向量,以及将所述压缩文档中的每一字随机初始化为字向量,其中,所述词向量与所述字向量的维度相同,所述词向量映射模型是预先基于无监督方法训练形成的word2vec的词向量映射模型;

[0059] 通过冗余方式将所述词向量和所述字向量进行混合,得到所述目标向量。

[0060] 示例地,词向量映射模型可以将输入的压缩文档中的每一词映射为768维的词向量,并将压缩文档中的每个字随机初始化为768维的字向量,再通过将词向量和字向量通过冗余方式进行混合得到上述目标向量。从而能够最大程度的保留文档中的语义信息。

[0061] 在一种可能的实现方式中,所述深度学习模型包括bert模型层、双向LSTM模型层、卷积层以及softmax模型层;其中,所述bert模型层与所述双向LSTM模型层相结合能够提取所述压缩文档的语义特征;所述bert模型层与所述卷积层相结合能够提取所述压缩文档的深度特征,并结合最大池化方式和平均池化方式对提取到的所述深度特征进行池化;所述softmax模型层用于输出所述文档分类结果。

[0062] 在该种实现方式中,所述bert模型层的参数在所述深度学习模型的训练过程中保持冻结,也即在对深度学习模型的训练过程中,无需对bert模型层的参数进行更新,从而能够提升模型的训练效率。和/或,所述卷积层包括多层,且每一层具有不同的卷积核。例如,三层卷积层,每一层的卷积核的尺寸分别可以为3×3,4×4,5×5。

[0063] 下面对深度学习模型的训练过程进行说明深度学习模型的损失函数可以为交叉熵损失函数,并通过adam优化器对所述深度学习模型进行参数优化,直到得到满足模型精度要求的深度学习模型。此外,由于深度学习模型在训练过程中采用字向量和词向量混合编码得到的向量作为输入,而字向量是词向量映射模型通过随机化方式生成的,因此,为了尽可能的提取到压缩文档的深度特征,可以在深度学习模型在训练过程中,词向量映射模型可以在保持词向量不变的情况下,对字向量不断进行更新,并将更新后的字向量重新与词向量进行混合编码,以得到更多的混合编码的向量对深度学习模型进行训练,提高了深度学习模型的准确率。

[0064] 图2是根据本公开一示例性实施例示出一种文档分类装置的框图,所述装置200包括:

[0065] 获取模块201,用于获取带目录的文本文档;

[0066] 结合模块202,用于提取所述文本文档中的关键词,并将所述关键词与所述目录进行结合,得到所述文本文档的压缩文档;

[0067] 编码模块203,用于对所述压缩文档进行字词混合编码,得到目标向量;

[0068] 生成模块204,用于将所述目标向量输入预先训练完成的深度学习模型,得到所述深度学习模型输出的文档分类结果。

[0069] 采用上述装置,该装置在获取带目录的文本文档后,首先提取所述文本文档中的关键词并将所述关键词与所述目录进行结合得到所述文本文档的压缩文档,之后对所述压缩文档进行字词混合编码,将所述目标向量输入预先训练完成的深度学习模型,最后得到所述深度学习模型输出的文档分类结果。本公开实施例通过提取关键词,并将关键词与目录进行结合的方式,实现了在保留文本核心内容的基础上,对文本进行了压缩,降低了深度学习模型因文本过长而对文本进行截断,导致大部分文本核心信息丢失的概率。

[0070] 可选地,所述编码模块包括:

[0071] 映射子模块,用于根据词向量映射模型,将所述压缩文档中的每一词映射为词向量,以及将所述压缩文档中的每一字随机初始化为字向量;

[0072] 混合子模块,用于通过冗余方式将所述词向量和所述字向量进行混合,得到所述目标向量。

[0073] 可选地,所述深度学习模型包括bert模型层、双向LSTM模型层、卷积层以及softmax模型层;

[0074] 其中,所述bert模型层与所述双向LSTM模型层相结合能够提取所述压缩文档的语义特征;

[0075] 所述bert模型层与所述卷积层相结合能够提取所述压缩文档的深度特征,并结合最大池化方式和平均池化方式对提取到的所述深度特征进行池化;

[0076] 所述softmax模型层用于输出所述文档分类结果。

[0077] 可选地,所述bert模型层的参数在所述深度学习模型的训练过程中保持冻结,和/或,所述卷积层包括多层,且每一层具有不同的卷积核。

[0078] 可选地,所述结合模块具体用于:在结合所述关键词与所述目录时,将所述关键词放在所述目录之前,使得所述关键词能够优先于所述目录进行编码。

[0079] 关于上述实施例中的装置,其中各个模块执行操作的具体方式已经在有关该方法的实施例中进行了详细描述,此处将不做详细阐述说明。

[0080] 本公开实施例还提供一种计算机可读存储介质,其上存储有计算机程序,该程序被处理器执行时实现上述方法实施例提供的方法的步骤。

[0081] 本公开实施例还提供一种电子设备,包括:

[0082] 存储器,其上存储有计算机程序;

[0083] 处理器,用于执行所述存储器中的所述计算机程序,以实现上述方法实施例提供的方法的步骤。

[0084] 图3是根据一示例性实施例示出的一种电子设备1900的框图。例如,电子设备1900可以被提供为一服务器。参照图3,电子设备1900包括处理器1922,其数量可以为一个或多个,以及存储器1932,用于存储可由处理器1922执行的计算机程序。存储器1932中存储的计算机程序可以包括一个或一个以上的每一个对应于一组指令的模块。此外,处理器1922可以被配置为执行该计算机程序,以执行上述的文档分类方法。

[0085] 另外,电子设备1900还可以包括电源组件1926和通信组件1950,该电源组件1926可以被配置为执行电子设备1900的电源管理,该通信组件1950可以被配置为实现电子设备

1900的通信,例如,有线或无线通信。此外,该电子设备1900还可以包括输入/输出(I/O)接口1958。电子设备1900可以操作基于存储在存储器1932的操作系统,例如Windows Server™,Mac OS X™,Unix™,Linux™等等。

[0086] 在另一示例性实施例中,还提供了一种包括程序指令的计算机可读存储介质,该程序指令被处理器执行时实现上述的文档分类方法的步骤。例如,该计算机可读存储介质可以为上述包括程序指令的存储器1932,上述程序指令可由电子设备1900的处理器1922执行以完成上述的文档分类方法。

[0087] 在另一示例性实施例中,还提供一种计算机程序产品,该计算机程序产品包含能够由可编程的装置执行的计算机程序,该计算机程序具有当由该可编程的装置执行时用于执行上述的文档分类方法的代码部分。

[0088] 此外,本公开的各种不同的实施方式之间也可以进行任意组合,只要其不违背本公开的思想,其同样应当视为本公开所公开的内容。

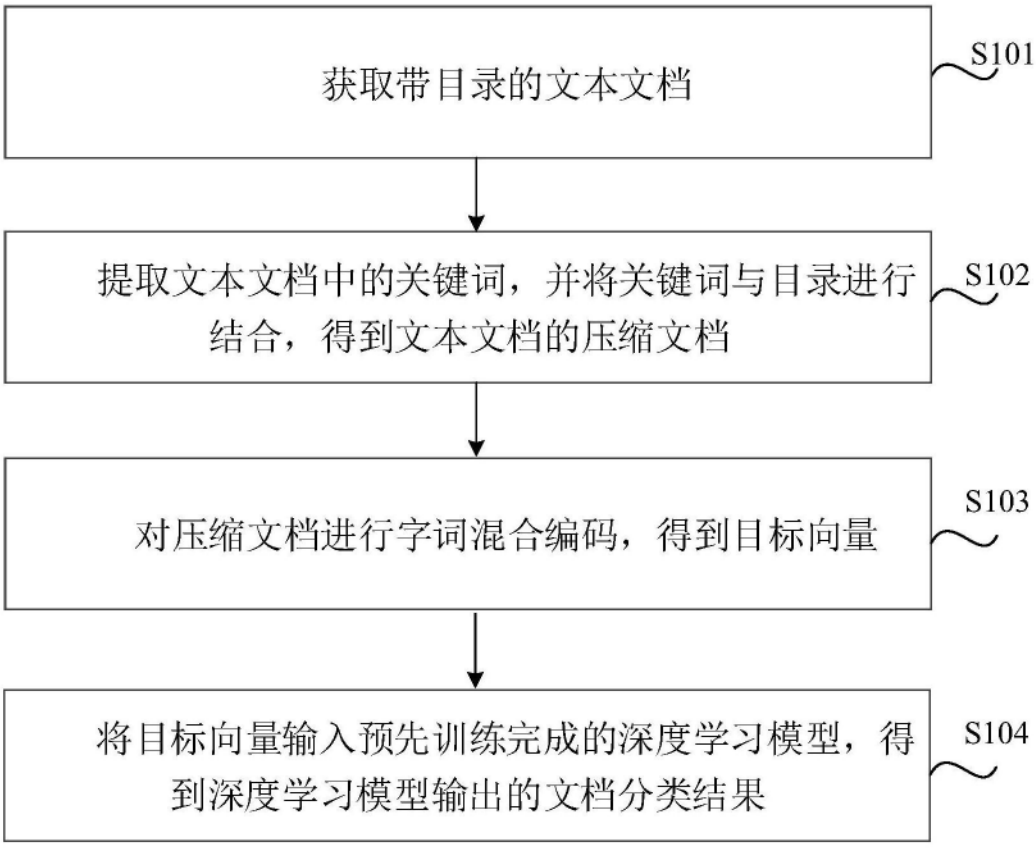

获取带目录的文本文档
S101
提取文本文档中的关键词，并将关键词与目录进行结合，得到文本文档的压缩文档
S102
对压缩文档进行字词混合编码，得到目标向量
S103
将目标向量输入预先训练完成的深度学习模型，得到深度学习模型输出的文档分类结果
S104

图1

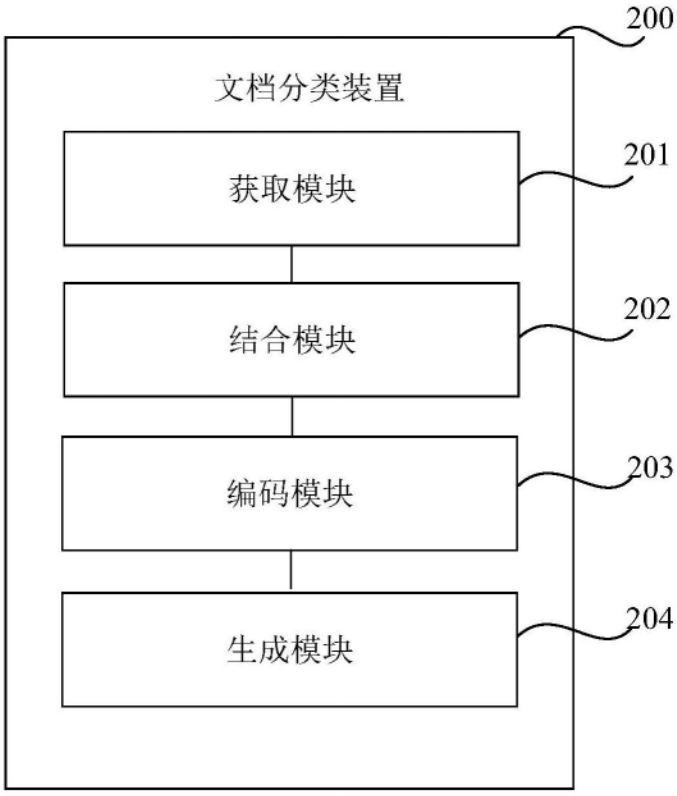

200
文档分类装置
201
获取模块
202
结合模块
203
编码模块
204
生成模块

图2

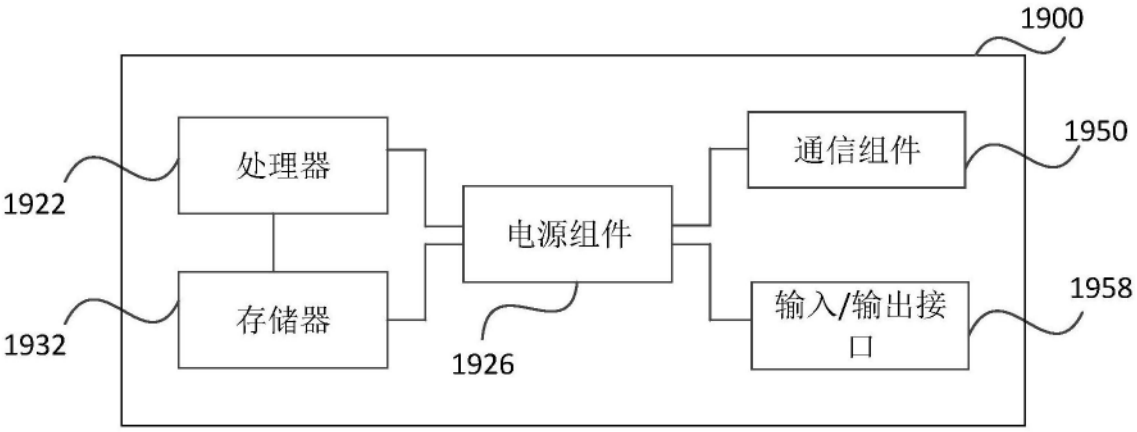

1900
处理器
1922
存储器
1932
电源组件
1926
通信组件
1950
输入/输出接口
1958

图3