(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 114706961 A
(43)申请公布日 2022.07.05

(21)申请号 202210064666.8

(22)申请日 2022.01.20

(71)申请人 平安国际智慧城市科技股份有限公司

地址 518000 广东省深圳市前海深港合作区妈湾兴海大道3048号前海自贸大厦1-34层

(72)发明人 陈芷昕

(74)专利代理机构 北京鸿元知识产权代理有限公司 11327

专利代理师 袁文婷 王迎

(51)Int.Cl.

G06F 16/332(2019.01)

G06K 9/62(2022.01)

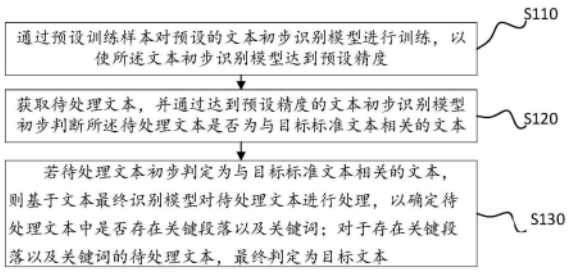
权利要求书2页 说明书12页 附图2页

(54)发明名称

目标文本识别方法、装置及存储介质

(57)摘要

本发明涉及数据处理技术领域，提供一种目标文本识别方法和电子设备，其中的方法包括：通过预设训练样本对文本初步识别模型进行训练，以使文本初步识别模型达到预设精度；获取待处理文本，并通过文本初步识别模型初步判断待处理文本是否为与目标标准文本相关的文本；若待处理文本初步判定为与目标标准文本相关的文本，则基于文本最终识别模型对待处理文本的正文进行处理，以确定待处理文本的正文中是否存在与目标标准文本相关的关键段落以及关键词；对于正文中存在与目标标准文本相关的关键段落以及关键词的待处理文本，最终判定为目标文本。本发明提供的技术方案既能够解决现有目标文本信息人工获取方式工作效率低的问题。



CN 114706961 A

1.一种目标文本识别方法,其特征在于,所述方法包括:

通过预设训练样本对预设的文本初步识别模型进行训练,以使所述文本初步识别模型达到预设精度;

通过训练后的所述文本初步识别模型对所获取的待处理文本进行初步识别处理,初步判断所述待处理文本是否为与目标标准文本相关的文本;其中,

若所述待处理文本为与所述目标标准文本相关的文本,则基于预设的文本最终识别模型对所述待处理文本的正文进行识别处理,以确定所述待处理文本的正文中是否存在与所述目标标准文本相关的关键段落以及关键词;

将正文中存在与所述目标标准文本相关的关键段落以及关键词的待处理文本,确定为目标文本。

2.根据权利要求1所述的目标文本识别方法,其特征在于,所述通过所述文本初步识别模型对所获取的待处理文本进行初步识别处理,初步判断所述待处理文本是否为与所述目标标准文本相关的文本的过程包括:

获取所述待处理文本的标题信息;

对所述待处理文本的标题信息进行分词处理,以将所述待处理文本的标题信息分成至少包括一个词条的实时词组排列;

将所述实时词组排列转换为实时数字向量;

基于所述实时数字向量初步判断所述待处理文本是否为与目标标准文本相关的文本。

3.根据权利要求2所述的目标文本识别方法,其特征在于,所述将所述实时词组排列转换为实时数字向量包括:

确定所述实时词组排列中的各词条的词频以及逆文档频率;

将各词条的词频与逆文档频率做相乘运算,得到各词条的词频与逆文档频率的频率乘积;

将所有词条的频率乘积组成的数字串排列为所述实时数字向量。

4.根据权利要求2所述的目标文本识别方法,其特征在于,所述基于所述实时数字向量判断所述待处理文本是否为与目标标准文本相关的文本包括:

将所述实时数字向量输入至预训练的支持向量机内,通过所述支持向量机的输出结果初步判断所述待处理文本是否为与目标标准文本相关的文本。

5.根据权利要求1所述的目标文本识别方法,其特征在于,所述文本最终识别模型包括长文本crf抽取模块和短文本crf抽取模块;并且,所述基于预设的文本最终识别模型对所述待处理文本的正文进行处理,以确定所述待处理文本的正文中是否存在与所述目标标准文本相关的关键段落以及关键词包括:

通过所述长文本crf抽取模块判断所述待处理文本的正文中是否存在与所述目标标准文本相关的所述关键段落;

通过所述短文本crf抽取模块判断所述关键段落中是否存在所述关键词。

6.根据权利要求5所述的目标文本识别方法,其特征在于,所述通过长文本crf抽取模块判断所述待处理文本的正文中是否存在与所述目标标准文本相关的所述关键段落包括:

采用预设的一级标注体系对所述待处理文本的正文中的各段落的所有字符进行一级标签标注;

若所述待处理文本的正文中的一个段落的所有一级标签中同时包含所有预设种类的一级实体标签，则判断所述待处理文本的正文中的该段落为所述关键段落。

7.根据权利要求5所述的目标文本识别方法，其特征在于，所述通过所述短文本crf抽取模块判断所述关键段落中是否存在所述关键词包括：

采用预设的二级标注体系对所述关键段落的所有字符进行二级标签标注；

若所述关键段落的所有二级标签中同时包含所有预设种类的二级实体标签，则判断所述关键段落中存在所述关键词。

8.一种目标文本识别装置，其特征在于，包括：

模型训练单元，用于通过预设训练样本对预设的文本初步识别模型进行训练，以使所述文本初步识别模型达到预设精度；

初步识别单元，用于通过训练后的所述文本初步识别模型对所获取的待处理文本进行初步识别处理，初步判断所述待处理文本是否为与目标标准文本相关的文本；

最终识别单元，用于若所述待处理文本为与所述目标标准文本相关的文本，则基于预设的文本最终识别模型对所述待处理文本的正文进行识别处理，以确定所述待处理文本的正文中是否存在与所述目标标准文本相关的关键段落以及关键词；并将正文中存在与所述目标标准文本相关的关键段落以及关键词的待处理文本，确定为目标文本。

9.一种电子设备，包括存储器和处理器，所述存储器存储有计算机程序，其特征在于，所述处理器执行所述计算机程序时实现权利要求1至7中任一项所述目标文本识别方法的步骤。

10.一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，所述计算机程序被处理器执行时实现如权利要求1至7中任一项所述的目标文本识别方法中的步骤。

目标文本识别方法、装置及存储介质

技术领域

[0001] 本发明涉及数据处理技术领域，尤其涉及一种目标文本识别方法、装置、电子设备及存储介质。

背景技术

[0002] 近些年，人们对各类文本信息的获取需求的日益增大，以产业扶持政策文本信息获取为例，为促进产业发展，出现了相应的产业扶持政策，产业扶持政策是指在制定区域发展计划或规划纲要时，针对地区经济发展的实际情况，采取重点倾斜、优先扶持某些产业或部门的措施，促使它们优先发展，快速发展，以期带动其他产业的共同发展，从而促进整个地区经济发展的政策和措施。

[0003] 对于扶持政策文本，通常会在正文中明确扶持对象的扶持手段，包括具体的扶持措施以及具体的扶持金额对象等等。为获取这类文本信息，相关业内人士需要通过人工阅读官网等平台上的所有的政策文本的方式才能从中获取下发的当前产业扶持政策。然而，这种人工阅读所有政策文本的方式，由于需要相关人员认真阅读所有的政策文本全文，因此无法做到快速、统一地对当前的扶策进行分类，从而从中获取到相应的产业扶持政策，进而无法高效地对各地方的产业发展进行研究，严重影响相关产业工作战略的布局。

[0004] 基于此，亟需一种能够快速从待处理文本中解析出目标文本(如产业扶持政策)的识别方法。

发明内容

[0005] 本发明提供一种目标文本识别方法、装置、电子设备以及存储介质，其主要目的在于解决现有目标文本信息人工获取方式工作效率低的问题。

[0006] 为实现上述目的，本发明提供一种目标文本识别方法，该方法包括如下步骤：

[0007] 通过预设训练样本对预设的文本初步识别模型进行训练，以使所述文本初步识别模型达到预设精度；

[0008] 获取待处理文本，并通过达到预设精度的所述文本初步识别模型初步判断所述待处理文本是否为与目标标准文本相关的文本；

[0009] 若所述待处理文本初步判定为与所述目标标准文本相关的文本，则基于预设的文本最终识别模型对所述待处理文本的正文进行处理，以确定所述待处理文本的正文中是否存在与所述目标标准文本相关的关键段落以及关键词；

[0010] 对于正文中存在与所述目标标准文本相关的关键段落以及关键词的所述待处理文本，最终判定为目标文本。

[0011] 优选地，所述通过达到预设精度的所述文本初步识别模型初步判断所述待处理文本是否为与所述目标标准文本相关的文本包括：

[0012] 获取所述待处理文本的标题信息；

[0013] 对所述待处理文本的标题信息进行分词，以将所述待处理文本的标题信息分成至

少包括一个词条的实时词组排列；

[0014] 将所述实时词组排列转换为实时数字向量；

[0015] 基于所述实时数字向量初步判断所述待处理文本是否为与目标标准文本相关的文本。

[0016] 优选地，所述将所述实时词组排列转换为实时数字向量包括：

[0017] 确定所述实时词组排列中的各词条的词频以及逆文档频率；

[0018] 将各词条的词频与逆文档频率做相乘运算，得到各词条的词频与逆文档频率的频率乘积；

[0019] 并将所有词条的频率乘积组成的数字串排列即为所述实时数字向量。

[0020] 优选地，所述基于所述实时数字向量判断所述待处理文本是否为与目标标准文本相关的文本包括：

[0021] 将所述实时数字向量输入至训练完毕的支持向量机内，通过训练完毕的所述支持向量机的输出结果初步判断所述待处理文本是否为与目标标准文本相关的文本。

[0022] 优选地，所述文本最终识别模型包括长文本crf抽取模块和短文本crf抽取模块；并且，所述基于预设的文本最终识别模型对所述待处理文本的正文进行处理，以确定所述待处理文本的正文中是否存在与所述目标标准文本相关的关键段落以及关键词包括：

[0023] 通过所述长文本crf抽取模块判断所述待处理文本的正文中是否存在与所述目标标准文本相关的所述关键段落；

[0024] 通过所述短文本crf抽取模块判断所述关键段落中是否存在所述关键词。

[0025] 优选地，所述通过长文本crf抽取模块判断所述待处理文本的正文中是否存在与所述目标标准文本相关的所述关键段落包括：

[0026] 采用预设的一级标注体系对所述待处理文本的正文中的各段落的所有字符进行一级标签标注；

[0027] 若所述待处理文本的正文中的一个段落的所有一级标签中同时包含预设的五种一级实体标签，则判断所述待处理文本的正文中的该段落为所述关键段落。

[0028] 优选地，所述通过所述短文本crf抽取模块判断所述关键段落中是否存在所述关键词包括：

[0029] 采用预设的二级标注体系对所述关键段落的所有字符进行二级标签标注；

[0030] 若所述关键段落的所有二级标签中同时包含预设的三种二级实体标签，则判断所述关键段落中存在所述关键词。

[0031] 另一方面，本发明还提供一种目标文本识别装置，包括：

[0032] 模型训练单元，用于通过预设训练样本对预设的文本初步识别模型进行训练，以使所述文本初步识别模型达到预设精度；

[0033] 初步识别单元，用于获取待处理文本，并通过达到预设精度的所述文本初步识别模型初步判断所述待处理文本是否为与目标标准文本相关的文本；

[0034] 最终识别单元，用于若所述待处理文本初步判定为与所述目标标准文本相关的文本，则基于预设的文本最终识别模型对所述待处理文本的正文进行处理，以确定所述待处理文本的正文中是否存在与所述目标标准文本相关的关键段落以及关键词；并对于正文中存在与所述目标标准文本相关的关键段落以及关键词的所述待处理文本，最终判定为目标

文本。

[0035] 另一方面，本发明还提供一种电子设备，包括存储器和处理器，所述存储器存储有计算机程序，所述处理器执行所述计算机程序时实现上述目标文本识别方法的步骤。

[0036] 另一方面，本发明还提供一种计算机可读存储介质，其上存储有计算机程序，所述计算机程序被处理器执行时实现上述的目标文本识别方法中的步骤。

[0037] 本发明提出的目标文本识别方法、装置、电子设备以及可读存储介质，能够基于历史工作经验自动深挖出目标文本(如产业扶持政策)的特性，总结出目标文本的主要特征(如产业扶持政策的政策分类和标题关键词)，从而高效准确的确定待处理文本中的目标文本。此外，通过文本初步识别模型，能够快速利用标题初步筛查出与目标标准文本相关的文本，而后利用文本最终识别模型，判断该待处理正文是否含有关键段落和关键词；若含有，则最终确定该待处理文本为可用的目标文本，从而显著提升目标文本的识别精度。

附图说明

[0038] 图1为根据本发明实施例的目标文本识别方法的较佳实施例流程图；

[0039] 图2为根据本发明实施例的目标文本识别装置的模块示意图；

[0040] 图3为根据本发明实施例的提供的实现目标文本识别方法的电子设备的内部结构示意图。

[0041] 本发明目的的实现、功能特点及优点将结合实施例，参照附图做进一步说明。

具体实施方式

[0042] 在下面的描述中，出于说明的目的，为了提供对一个或多个实施例的全面理解，阐述了许多具体细节。然而，很明显，也可以在没有这些具体细节的情况下实现这些实施例。

[0043] 以下将结合附图对本申请的具体实施例进行详细描述。

[0044] 实施例1

[0045] 为了说明本发明提供的目标文本识别方法，图1示出了根据本发明提供的目标文本识别方法的流程。

[0046] 如图1所示，本发明提供的目标文本识别方法，包括：

[0047] S110：通过预设训练样本对预设的文本初步识别模型进行训练，以使所述文本初步识别模型达到预设精度。

[0048] 需要说明的是，所述预设训练样本为目标文本标题样本，为预先通过人工确定为目标文本的某一领域、级别、地域等范围内的历史目标文本的标题。

[0049] 比如，以某一产业的扶持政策的目标文本为例，该预设训练样本则可以是为预先通过人工确定(人为通过经验确定)为该产业扶持政策的历史政策的标题。

[0050] 具体地，作为示例，在预先通过人工确定历史政策是否为产业扶持政策的过程中，需要通过人工阅读政策文本全文的方式挨个基于经验确定各历史政策是否为产业扶持政策。

[0051] 具体地，通过人工阅读的方式可以将现有的历史产业扶持政策分为：指导意见类、与资金分配有关的管理办法、补贴奖励类、税收类、资质认定类的五个类别。在确定了历史产业扶持政策的类别后，对于每类产业扶持政策，还需要工作人员基于历史经验，确定相应

的各类历史产业扶持政策中的标题信息中的标题关键词，例如：对于指导意见类历史产业扶持政策，标题关键词包括扶持、资助、支持、加快、加强、推进行业的意见/指导意见/实施意见/措施/实施方案。

[0052] 需要进一步说明的是，这些标题关键词是工作人员通过历史阅读工作经验总结出来的，当获取到新的标题关键词后即可填入上述关键词表格中，以扩充关键词库。

[0053] 还需要说明的是，所述目标文本初步识别模型后期用于基于待处理目标文本的标题信息来初步判断待处理目标文本是否为目标文本，需要通过目标文本标题样本(即预设训练样本，预先通过人工确定为目标文本的某一领域、级别、地域等范围内的历史目标文本的标题)对预设的目标文本初步识别模型进行训练，以使目标文本初步识别模型能够精准地模拟人工通过标题信息(主要是通过标题中的关键词)来判断历史目标文本是否为目标文本的过程，并使目标文本初步识别模型达到相应的识别精度。

[0054] 需要说明的是，文本初步识别模型其本质为一个二分类模型，用于初步判断待处理文本是否为目标文本，若初步识别模型输出为1，则初步判断待处理文本是目标文本，否则判定待处理文本不是目标文本。另外，还需要说明的是，该初步识别模型的精度是指文本初步识别模型判断待处理文本是否为目标文本的精度，通过预设训练样本对文本初步识别模型进行持续训练，文本初步识别模型的精度会不断提升，当文本初步识别模型的精度达到预设精度后(百分之九十五)后，即可停止对文本初步识别模型的训练。此处需要进一步说明的是，文本初步识别模型的精度可以通过测试的方式确定，例如，可以通过文本初步识别模型对1000个预设训练样本进行二分类，若分类结果中有950及以上个在与人工的分类结果比较后，确定为分类正确的结果，则认定文本初步识别模型达到预设精度，否则，继续通过预设训练样本对文本初步识别模型进行训练，直至文本初步识别模型达到预设精度。

[0055] 具体的，作为示例，在产业扶持政策应用中，可以将文本初步识别模型设定为扶持政策初步识别模型，后期用于基于待处理政策的标题信息来初步判断待处理政策是否为产业扶持政策。因此，在步骤S110中还需要通过扶持政策标题样本(预先通过人工确定为产业扶持政策的历史政策的标题)对预设的扶持政策初步识别模型进行训练(对应通过预设训练样本对预设的文本初步识别模型进行训练)，以使扶持政策初步识别模型能够精准地模拟人工通过标题信息(主要是通过标题中的关键字)来判断历史政策是否为产业扶持政策的过程，并使扶持政策初步识别模型达到相应的识别精度。

[0056] 具体地，为实现扶持政策初步识别模型的二分类，所述扶持政策初步识别模型进一步包括如下模块：

[0057] 分词模块，用于对扶持政策标题样本(在S110中的训练过程中)或者待处理政策的标题(在后续S120中的应用过程中)进行分词，已将扶持政策标题样本或者待处理政策的标题分成多个词组的实时词组排列(在实际分词过程中，可以利用现有的分词软件如：jieba)。

[0058] 向量转换模块，用于将分词后的实时词组排列转换为能够代表将扶持政策标题样本或者待处理政策(在后续S120中的应用过程中)的标题的实时数字向量。

[0059] 政策判断模块，用于基于扶持政策标题样本或者待处理政策的标题转换出的实时数字向量，初步判断与扶持政策标题样本或者待处理政策(在后续 S120中的应用过程中)的标题对应的政策是否为产业扶持政策。

[0060] 具体地,为提升转换后的数字向量对扶持政策标题样本或者待处理政策的标题的表征能力,向量转换模块可以采用F-IDF(Term Frequency-Inverse Document Frequency,词频-逆文件频率技术)实现扶持政策标题样本或者待处理政策的标题的数字向量的计算。

[0061] 需要说明的是,F-IDF(词频-逆文档频度技术),是一种用于资讯检索与文本挖掘的加权技术,可以用来评估一个词对于一个文档集或语料库中某个文档的重要程度。字词的重要性随着它在文件中出现的次数成正比增加,但同时会随着它在语料库中出现的频率成反比下降。如果某个词比较少见,但是它在这篇文章中多次出现,那么它很可能就反映了这篇文章的特性,正是我们所需要的关键词。由于本方案中的扶持政策标题样本是通过人员通过标题关键词来确定的,并且整个扶持政策初步识别模型也是模拟的该人工识别的过程,因此,使用F-IDF来实现扶持政策标题样本或者待处理政策的标题的数字向量的计算,能够显著提升扶持政策初步识别模型的精度。

[0062] 具体地,词频-逆文档频率,由词频(TF)和逆文档频率(IDF)两部分组成,其中,在实际使用过程中需要预先给定语料库D=｛dj｝,对于本方案来讲,语料库D需要设定为包含所有历史政策文本和待处理政策文本(包括已经通过人工判定为产业扶持政策的历史政策文本、判定为非产业扶持政策的历史政策文以及为确定是否为产业扶持政策的待处理政策文本,其中,非产业扶持政策的历史政策文用于防止述扶持政策初步识别模型出现过拟合)的文本库。

[0063] 其中的词频的计算公式如下:

[0064] $\mathrm{tf_{i,j}} = \frac{n_{i,j}}{\sum_k n_{k,j}}$,其中,$tf_{i,j}$为扶持政策标题样本或待处理政策的标题中的某一词条ti在相应的历史产业扶持政策dj或者待处理政策dj中出现的频率,$n_{i,j}$为该词条在相应的历史产业扶持政策dj或者待处理政策dj中出现的个数,$\Sigma_k\ n_{k,j}$.为该词条在相应的历史产业扶持政策dj或者待处理政策dj中的总词数。

[0065] 其中的逆文档频率的计算公式如下:

[0066] $$\mathrm{idf_i} = \lg \frac{|D|}{|\{j : t_i \in d_j\}|}$$

[0067] 其中.idf_i,为扶持政策标题样本或待处理政策的标题中的某一词条ti的逆文档频率,用于度量词条ti的重要性,|D|为语料库D中的总的文档数(文本数量).|｛j:t_i∈a_j｝|为语料库D包含该词条ti的文档数。

[0068] $tfidf_{i,j}$=$tf_{i,j}$×idf_i,即为计算后得到的能够代表将扶持政策标题样本或者待处理政策的标题的数字向量,并且,后续的每个待处理政策的标题经过tfidf方法计算后都会得到一个相应的包含文章语义的实时数字向量。

[0069] 需要再次强调的是,TF-IDF的主要思想是:如果某个词或短语在一篇文章中出现的频率TF高,并且在其他文章中很少出现,则认为此词或者短语具有很好的类别区分能力,适合用来分类。TFIDF实际上是:TF*IDF,TF为词频(Term Frequency),IDF为逆向文件频率(Inverse Document Frequency)。TF 表示某个关键词$n_{i,j}$在文档dj中出现的频率。IDF的主要思想是:如果包含关键词t的文档越少,也就是n越小,IDF越大,则说明词条ti具有很好的类别区分能力。

[0070] 在通过向量转换模块得到实时数字向量之后,则需要通过政策判断模块去判断待处理政策是否为产业扶持政策,需要说明的是,对于扶持政策标题样本形成的数字向量来讲,由于其对应的历史政策已经确定为产业扶持政策,因此,此处的扶持政策标题样本形成的数字向量则用于对政策判断模块的训练。

[0071] 具体地,政策判断模块可以选用具有二分类功能的svm(支持向量机)进行政策判断,支持向量机(support vector machines,SVM)是一种二分类模型文文,它的基本模型是定义在特征空间上的间隔最大的线性分类器,间隔最大使它有别于感知机;SVM还包括核技巧,这使它成为实质上的非线性分类器。SVM的学习策略就是间隔最大化,可形式化为一个求解凸二次规划的问题,也等价于正则化的合页损失函数的最小化问题,SVM的学习算法就是求解凸二次规划的最优化算。

[0072] 具体地,将上述步骤中生成的数字向量输入svm模型中进行训练,给定输入数据(对应上述的生成的数字向量组X＝{X1,X2,X3^……XN})(其中,xi 为第i个用于训练的数字向量)和学习目标y＝{y1,y2,}＝{0,1},0表示负类, 1表示正类,yi表示xi对应的y值,本例中采用的是硬边界SVM,该方法是在线性可分问题中求解最大边距超平面(maximum-marginhyperplane)的算法,约束条件是样本点到决策边界的距离大于等于1。给定超平面w·x+b＝0,硬边界SVM可以转化为一个等价的二次凸优化(quadratic convex optimization) 问题进行求解,约束条件公式如下:

[0073] $$\max_{w,b} \frac{2}{\|\boldsymbol{w}\|} \quad \text{s.t.} \quad y_i\left(\boldsymbol{w}^{\top}\boldsymbol{X}_i+b\right)\geq 1 \iff \min_{w,b} \frac{1}{2}\|\boldsymbol{w}\|^2 \quad \text{s.t.} \quad y_i\left(\boldsymbol{w}^{\top}\boldsymbol{X}_i+b\right)\geq 1$$,其中,w超平面的斜率,b为超平面的截距。

[0074] 若该模型的输出值为0,则代表该标题信息对应的政策与产业扶持有关 (初步),若该模型输出值为1,则代表该标题信息对应的政策与产业扶持无关(初步判断)。

[0075] 通过使用上述约束条件基于扶持政策标题样本的数字向量组进行训练,则可逐渐提升svm的分类精度,当svm的分类精度达到预设精度(百分之九十五时),停止训练,即可使用该扶持政策初步识别模型对待处理政策进行是否为产业扶持政策的初步判断(对应待处理文本是否为目标文本的初步判断)。

[0076] S120:获取待处理文本,并通过训练后的达到预设精度的所述文本初步识别模型初步判断所述待处理文本是否为与目标标准文本相关的文本。

[0077] 具体地,所述通过达到预设精度的所述文本初步识别模型初步判断所述待处理文本是否为与所述目标标准文本相关的文本包括:

[0078] 获取所述待处理文本的标题信息;

[0079] 对所述待处理文本的标题信息进行分词,以将所述待处理文本的标题信息分成至少包括一个词条的实时词组排列;

[0080] 将所述实时词组排列转换为实时数字向量;

[0081] 基于所述实时数字向量初步判断所述待处理文本是否为与目标标准文本相关的文本。

[0082] 具体地,所述将所述实时词组排列转换为实时数字向量包括:

[0083] 确定所述实时词组排列中的各词条的词频以及逆文档频率;

[0084] 将各词条的词频与逆文档频率做相乘运算，得到各词条的词频与逆文档频率的频率乘积；

[0085] 并将所有词条的频率乘积组成的数字串排列即为所述实时数字向量。

[0086] 进一步地，所述基于所述实时数字向量判断所述待处理文本是否为与目标标准文本相关的文本包括：

[0087] 将所述实时数字向量输入至训练完毕的支持向量机内，通过训练完毕的所述支持向量机的输出结果初步判断所述待处理文本是否为与目标标准文本相关的文本。

[0088] 具体地，作为示例，在产业扶持政策应用中，通过达到预设精度的扶持政策初步识别模型初步判断所述标题信息对应的待处理政策是否问与产业相关的扶持政策包括，通过分词模块对取待处理政策的标题进行分词，通过向量转换模块将将分词后的词组排列转换为能够代表待处理政策的标题的数字向量，最后通过政策判断模块初步判断与待处理政策的标题对应的政策是否为产业扶持政策。

[0089] 需要说明的是，上述过程与步骤S110中的训练过程相同，因此，对其相信过程不再赘述。

[0090] S130：若所述待处理文本初步判定为与所述目标标准文本相关的文本，则基于预设的文本最终识别模型对所述待处理文本的正文进行处理，以确定所述待处理文本的正文中是否存在与所述目标标准文本相关的关键段落以及关键词；对于正文中存在与所述目标标准文本相关的关键段落以及关键词的所述待处理文本，最终判定为目标文本。

[0091] 需要说明的是，文本最终识别模型的本质为一个关键词抽取模型，用于判断所述待处理文本的正文中是否存在与所述目标标准文本相关的关键段落以及关键词，具体可以通过长文本crf抽取模型和短文本crf抽取模型进行构建。

[0092] 具体地，所述文本最终识别模型包括长文本crf抽取模块和短文本crf抽取模块；并且，所述基于预设的文本最终识别模型对所述待处理文本的正文进行处理，以确定所述待处理文本的正文中是否存在与所述目标标准文本相关的关键段落以及关键词包括：

[0093] 通过所述长文本crf抽取模块判断所述待处理文本的正文中是否存在与所述目标标准文本相关的所述关键段落；

[0094] 通过所述短文本crf抽取模块判断所述关键段落中是否存在所述关键词。

[0095] 其中，所述通过长文本crf抽取模块判断所述待处理文本的正文中是否存在与所述目标标准文本相关的所述关键段落包括：

[0096] 采用预设的一级标注体系对所述待处理文本的正文中的各段落的所有字符进行一级标签标注；

[0097] 若所述待处理文本的正文中的一个段落的所有一级标签中同时包含预设的五种一级实体标签，则判断所述待处理文本的正文中的该段落为所述关键段落。

[0098] 其中，所述通过所述短文本crf抽取模块判断所述关键段落中是否存在所述关键词包括：

[0099] 采用预设的二级标注体系对所述关键段落的所有字符进行二级标签标注；

[0100] 若所述关键段落的所有二级标签中同时包含预设的三种二级实体标签，则判断所述关键段落中存在所述关键词。

[0101] 需要说明的是，长文本crf抽取模块和短文本crf抽取模块其本质均为一种实体标

注模型，用于对述待处理文本的正文进行标注，因此，这两个模块均需要预先进行训练，训练过程为：使用预先通过人工标注完成的文本对两个模块进行训练，以实现两个模块对人工标注过程的模拟，当两个模块均达到预设精度后，停止训练即可(该训练过程与上述的文本初步识别模型的训练过程原理相同，在此不再过多赘述)。另外，对于标注体系，是一种本领域常用的标注规范，常用的有BIO标注体系，其中，O表示与工作目标、扶持力度均不相关的字符，B表示工作目标起始的字符，I表示工作目标件相关字符；本申请为了区别长文本crf抽取模块和短文本crf抽取模块采用的是两个不同的标注体系，因此，分别定义为一级标注体系和二级标注体系。

[0102] 具体地，作为示例，在产业扶持政策应用中，通过标题信息初步判断出与产业扶持有关的政策并不全都是产业扶持政策。如“龙岗区工业和信息化局关于公开征求《A市龙岗区经济与科技发展专项资金支持招商引资工作实施细则》修订意见的通知”一例，从标题上看该政策属于“与资金分配有关的管理办法”一类的政策，但实际上该政策为征求意见稿，并不是真正的产业扶持政策。判断政策是否为真的产业扶持政策，需要通过判断该政策正文中是否含有扶持政策应有关键段落来确定。

[0103] 下面详细介绍通过预设的扶持政策最终识别模型对初步判断与产业扶持相关的政策的政策正文进行关键段落抽取的过程。

[0104] 具体地，扶持政策最终识别模型(对应文本最终识别模型)的具体识别过程如下：

[0105] 利用长文本crf对关键段落的相关信息进行抽取。

[0106] CRF算法，即条件随机场算法。设X与Y是随机变量，P(Y|X)是给定X 的条件下Y的条件概率分布，若随机变量Y构成一个由无向图G＝(V,E)表示的马尔科夫随机场。则称条件概率分布P(Y|X)为条件随机场。因为是在X条件下的马尔科夫随机场，所以叫条件随机场。具体过程如下：

[0107] 输入初步判断与产业扶持相关的政策的政策正文，利用预设的长文本crf 抽取模型抽取其中与工作目标及扶持力度相关的关键段落；

[0108] 模型的具体工作原理为：首先，采用BIO标注体系对政策正文中的每个字符进行标注，从而定位出与工作目标及扶持力度相关的关键段落。由于本方案最终只需要抽取出相关的申报条件，所以该步骤的实体标签会有五种(对应预设种类的一级实体标签)：O(与工作目标、扶持力度均不相关的字符)，B- tar(工作目标起始字符)，I-tar(工作目标件相关字符)、B-sup、I-sup。

[0109] 当某一段落中同时存在上述五种实体标签后，即可判断该段落为与工作目标及扶持力度相关的关键段落。

[0110] 然后，对抽取到的与工作目标及扶持力度相关的关键段落，利用预设的短文本crf抽取模型对该段落进行扶持产业提取。具体地，该步骤中需要抽取的只有产业实体，所以对应的字符标签只有三种(对应预设种类的二级实体标签)：O(与产业无关的字符)，B(产业起始字符)，I(产业相关字符)。

[0111] 具体地，由于产业扶持政策正文中应有的内容需要包括工作目标、扶持力度、扶持产业，因此，对于初步判断与产业扶持相关的政策，若通过摘要抽取的方式，能够在政策正文中抽取到工作目标、扶持力度、扶持产业相关的段落，则可以最终判断该政策为产业扶持政策。

[0112] 若已经抽取到的与工作目标及扶持力度相关段落中任然存在上述三种标签，则判定该段落中也存在相应的扶持产业(对应关键段落中的关键词)。即若一篇待处理政策文本中可抽出上述这三个信息，则判断所述政策正文中存在与产业扶持相关的关键段落，且关键段落中存在相应的关键词。对于正文中存在与所述目标标准文本相关的关键段落以及关键词的所述待处理政策文本，最终判定为扶持政策文本。

[0113] 需要说明的是，作为示例，在产业扶持政策应用中，在确定该待处理政策为产业扶持政策后，需要将该处理政策的关键段落和政策类型进行整理，后续通过该关键段落和政策类型即可进行相应地产业发展进行研究，实现即时的工作战略布局。

[0114] 另外，需要强调的是，为进一步保证上述目标文本识别方法中所用到的数据的私密和安全性，实时词组排列和实时数字向量均可以存储在区块链的节点中。

[0115] 通过上述具体实施例可知，本发明提出的目标文本识别方法，能够基于历史工作经验自动深挖出目标文本(如产业扶持政策)的特性，总结出目标文本的主要特征(如产业扶持政策的政策分类和标题关键词)，从而高效准确的确定待处理文本中的目标文本。此外，通过文本初步识别模型，能够快速利用标题初步筛查出与目标标准文本相关的文本，而后利用文本最终识别模型，判断该待处理正文是否含有关键段落和关键词；若含有，则最终确定该待处理文本为可用的目标文本，从而显著提升目标文本的识别精度。

[0116] 应理解，上述实施例中各步骤的序号的大小并不意味着执行顺序的先后，各过程的执行顺序应以其功能和内在逻辑确定，而不应对本申请实施例的实施过程构成任何限定。

[0117] 如图2所示，本发明还提供一种目标文本识别装置100，该装置可以安装于电子设备中。根据实现的功能，该目标文本识别装置100可以包括模型训练单元101、初步识别单元102、最终识别单元103。本发明提供的上述单元，是指一种能够被电子设备处理器所执行，并且能够完成固定功能的一系列计算机程序段，其存储在电子设备的存储器中。

[0118] 在本实施例中，关于各模块/单元的功能如下：

[0119] 模型训练单元101，于通过预设训练样本对预设的文本初步识别模型进行训练，以使所述文本初步识别模型达到预设精度；

[0120] 初步识别单元102，用于获取待处理文本，并通过达到预设精度的所述文本初步识别模型初步判断所述待处理文本是否为与目标标准文本相关的文本；

[0121] 最终识别单元103，用于若所述待处理文本初步判定为与所述目标标准文本相关的文本，则基于预设的文本最终识别模型对所述待处理文本的正文进行处理，以确定所述待处理文本的正文中是否存在与所述目标标准文本相关的关键段落以及关键词；并对于正文中存在与所述目标标准文本相关的关键段落以及关键词的所述待处理文本，最终判定为目标文本。

[0122] 其中，初步识别单元102进一步包括分词单元、向量转换单元以及第一判断单元，其中，分词单元用于对所述待处理文本的标题信息进行分词，以将所述待处理文本的标题信息分成至少包括一个词条的实时词组排列，向量转换单元用于将所述实时词组排列转换为实时数字向量；第一判断单元，用于基于所述实时数字向量初步判断所述待处理文本是否为与目标标准文本相关的文本。

[0123] 进一步地，向量转换单元至少包括词频确定单元、运算单元以及排列单元，其中，

词频确定单元用于确定所述实时词组排列中的各词条的词频以及逆文档频率;运算单元用于将各词条的词频与逆文档频率做相乘运算,得到各词条的词频与逆文档频率的频率乘积;排列单元用于将所有词条的频率乘积组成的数字串排列即为所述实时数字向量。

[0124] 此外,最终识别单元103包括长文本crf抽取单元和短文本crf抽取单元,其中,长文本crf抽取单元用于通过所述长文本crf抽取模块判断所述待处理文本的正文中是否存在与所述目标标准文本相关的所述关键段落;短文本crf 抽取单元用于通过所述短文本crf抽取模块判断所述关键段落中是否存在所述关键词。

[0125] 进一步地,长文本crf抽取单元包括一级标签标注单元和第二判断单元,其中,一级标签标注单元用于采用预设的一级标注体系对所述待处理文本的正文中的各段落的所有字符进行一级标签标注;第二判断单元用于若所述待处理文本的正文中的一个段落的所有一级标签中同时包含所有预设种类的一级实体标签,则判断所述待处理文本的正文中的该段落为所述关键段落。

[0126] 更进一步地,短文本crf抽取包括二级标签标注单元和第三判断单元,其中,二级标签标注单元用于采用预设的二级标注体系对所述关键段落的所有字符进行二级标签标注;第三判断单元用于若所述关键段落的所有二级标签中同时包含所有预设种类的二级实体标签,则判断所述关键段落中存在所述关键词。

[0127] 如图3所示,本发明还提供一种目标文本识别方法的电子设备1。

[0128] 该电子设备1可以包括处理器10、存储器11和总线,还可以包括存储在存储器11中并可在所述处理器10上运行的计算机程序,如目标文本识别程序12。

[0129] 其中,所述存储器11至少包括一种类型的可读存储介质,所述可读存储介质包括闪存、移动硬盘、多媒体卡、卡型存储器(例如:SD或DX存储器等)、磁性存储器、磁盘、光盘等。所述存储器11在一些实施例中可以是电子设备1的内部存储单元,例如该电子设备1的移动硬盘。所述存储器11在另一些实施例中也可以是电子设备1的外部存储设备,例如电子设备1上配备的插接式移动硬盘、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,所述存储器11还可以既包括电子设备1的内部存储单元也包括外部存储设备。所述存储器11不仅可以用于存储安装于电子设备1的应用软件及各类数据,例如目标文本识别程序的代码等,还可以用于暂时地存储已经输出或者将要输出的数据。

[0130] 所述处理器10在一些实施例中可以由集成电路组成,例如可以由单个封装的集成电路所组成,也可以是由多个相同功能或不同功能封装的集成电路所组成,包括一个或者多个中央处理器(Central Processing unit,CPU)、微处理器、数字处理芯片、图形处理器及各种控制芯片的组合等。所述处理器10 是所述电子设备的控制核心(Control Unit),利用各种接口和线路连接整个电子设备的各个部件,通过运行或执行存储在所述存储器11内的程序或者模块(例如目标文本识别程序等),以及调用存储在所述存储器11内的数据,以执行电子设备1的各种功能和处理数据。

[0131] 所述总线可以是外设部件互连标准(peripheral component interconnect,简称PCI)总线或扩展工业标准结构(extended industry standard architecture,简称EISA)总线等。该总线可以分为地址总线、数据总线、控制总线等。所述总线被设置为实现所述存储器11以及至少一个处理器10等之间的连接通信。

[0132] 图3仅示出了具有部件的电子设备,本领域技术人员可以理解的是,图3 示出的结构并不构成对所述电子设备1的限定,可以包括比图示更少或者更多的部件,或者组合某些部件,或者不同的部件布置。

[0133] 例如,尽管未示出,所述电子设备1还可以包括给各个部件供电的电源 (比如电池),优选地,电源可以通过电源管理装置与所述至少一个处理器10 逻辑相连,从而通过电源管理装置实现充电管理、放电管理、以及功耗管理等功能。电源还可以包括一个或一个以上的直流或交流电源、再充电装置、电源故障检测电路、电源转换器或者逆变器、电源状态指示器等任意组件。所述电子设备1还可以包括多种传感器、蓝牙模块、Wi-Fi模块等,在此不再赘述。

[0134] 进一步地,所述电子设备1还可以包括网络接口,可选地,所述网络接口可以包括有线接口和/或无线接口(如WI-FI接口、蓝牙接口等),通常用于在该电子设备1与其他电子设备之间建立通信连接。

[0135] 可选地,该电子设备1还可以包括用户接口,用户接口可以是显示器 (Display)、输入单元(比如键盘(Keyboard)),可选地,用户接口还可以是标准的有线接口、无线接口。可选地,在一些实施例中,显示器可以是LED 显示器、液晶显示器、触控式液晶显示器以及OLED(Organic Light-Emitting Diode,有机发光二极管)触摸器等。其中,显示器也可以适当的称为显示屏或显示单元,用于显示在电子设备1中处理的信息以及用于显示可视化的用户界面。

[0136] 应该了解,所述实施例仅为说明之用,在专利申请范围上并不受此结构的限制。

[0137] 所述电子设备1中的所述存储器11存储的目标文本识别程序12是多个指令的组合,在所述处理器10中运行时,可以实现:

[0138] 通过预设训练样本对预设的文本初步识别模型进行训练,以使所述文本初步识别模型达到预设精度;

[0139] 获取待处理文本,并通过达到预设精度的所述文本初步识别模型初步判断所述待处理文本是否为与目标标准文本相关的文本;

[0140] 若所述待处理文本初步判定为与所述目标标准文本相关的文本,则基于预设的文本最终识别模型对所述待处理文本的正文进行处理,以确定所述待处理文本的正文中是否存在与所述目标标准文本相关的关键段落以及关键词;

[0141] 对于正文中存在与所述目标标准文本相关的关键段落以及关键词的所述待处理文本,最终判定为目标文本。

[0142] 具体地,所述处理器10对上述指令的具体实现方法可参考图1对应实施例中相关步骤的描述,在此不赘述。需要强调的是,为进一步保证上述目标文本识别的私密和安全性,上述目标文本识别数据存储于本服务器集群所处区块链的节点中。

[0143] 进一步地,所述电子设备1集成的模块/单元如果以软件功能单元的形式实现并作为独立的产品销售或使用时,可以存储在一个计算机可读取存储介质中。所述计算机可读介质可以包括: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)。

[0144] 本发明实施例还提供一种计算机可读存储介质,所述存储介质可以是非易失性的,也可以是易失性的,所述存储介质存储有计算机程序,所述计算机程序被处理器执行时

实现：

[0145] 通过预设训练样本对预设的文本初步识别模型进行训练，以使所述文本初步识别模型达到预设精度；

[0146] 获取待处理文本，并通过达到预设精度的所述文本初步识别模型初步判断所述待处理文本是否为与目标标准文本相关的文本；

[0147] 若所述待处理文本初步判定为与所述目标标准文本相关的文本，则基于预设的文本最终识别模型对所述待处理文本的正文进行处理，以确定所述待处理文本的正文中是否存在与所述目标标准文本相关的关键段落以及关键词；

[0148] 对于正文中存在与所述目标标准文本相关的关键段落以及关键词的所述待处理文本，最终判定为目标文本。

[0149] 具体地，所述计算机程序被处理器执行时具体实现方法可参考实施例目标文本识别方法中相关步骤的描述，在此不赘述。

[0150] 在本发明所提供的几个实施例中，应该理解到，所揭露的设备，装置和方法，可以通过其它的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，例如，所述模块的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式。

[0151] 所述作为分离部件说明的模块可以是或者也可以不是物理上分开的，作为模块显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部模块来实现本实施例方案的目的。

[0152] 另外，在本发明各个实施例中的各功能模块可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现，也可以采用硬件加软件功能模块的形式实现。

[0153] 对于本领域技术人员而言，显然本发明不限于上述示范性实施例的细节，而且在不背离本发明的精神或基本特征的情况下，能够以其他的具体形式实现本发明。

[0154] 因此，无论从哪一点来看，均应将实施例看作是示范性的，而且是非限制性的，本发明的范围由所附权利要求而不是上述说明限定，因此旨在将落在权利要求的等同要件的含义和范围内的所有变化涵括在本发明内。不应将权利要求中的任何附关联图标记视为限制所涉及的权利要求。

[0155] 本发明所指区块链是分布式数据存储、点对点传输、共识机制、加密算法等计算机技术的新型应用模式。区块链(Blockchain)，本质上是一个去中心化的数据库，是一串使用密码学方法相关联产生的数据块，每一个数据块中包含了一批次网络交易的信息，用于验证其信息的有效性(防伪)和生成下一个区块。区块链可以包括区块链底层平台、平台产品服务层以及应用服务层等。

[0156] 此外，显然“包括”一词不排除其他单元或步骤，单数不排除复数。系统权利要求中陈述的多个单元或装置也可以由一个单元或装置通过软件或者硬件来实现。第二等词语用来表示名称，而并不表示任何特定的顺序。

[0157] 最后应说明的是，以上实施例仅用以说明本发明的技术方案而非限制，尽管参照较佳实施例对本发明进行了详细说明，本领域的普通技术人员应当理解，可以对本发明的技术方案进行修改或等同替换，而不脱离本发明技术方案的精神和范围。

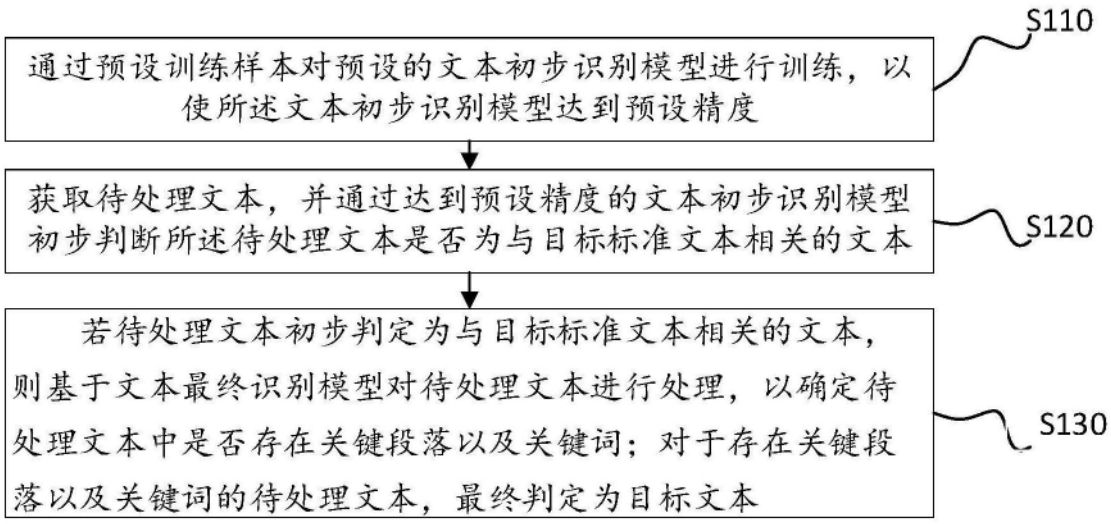

S110
通过预设训练样本对预设的文本初步识别模型进行训练，以使所述文本初步识别模型达到预设精度
S120
获取待处理文本，并通过达到预设精度的文本初步识别模型初步判断所述待处理文本是否为与目标标准文本相关的文本
S130
若待处理文本初步判定为与目标标准文本相关的文本，则基于文本最终识别模型对待处理文本进行处理，以确定待处理文本中是否存在关键段落以及关键词；对于存在关键段落以及关键词的待处理文本，最终判定为目标文本

图1

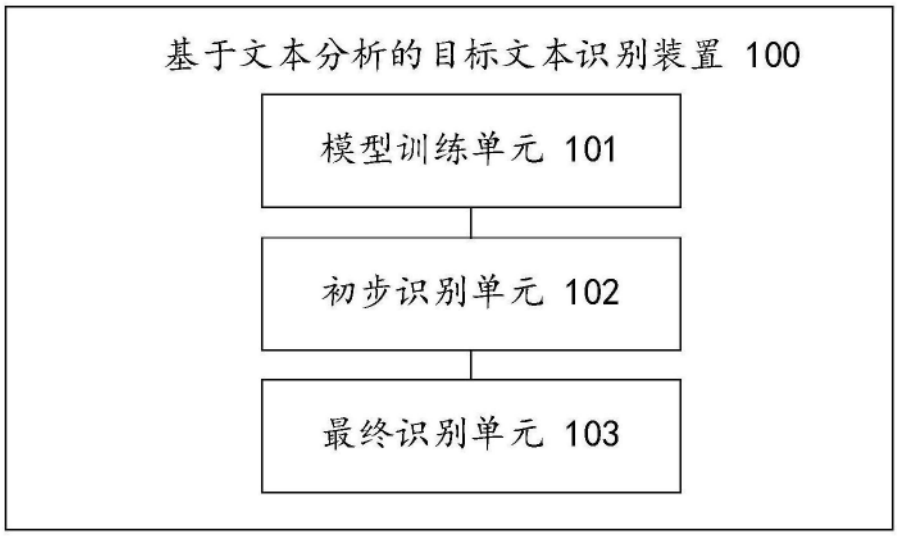

基于文本分析的目标文本识别装置 100
模型训练单元 101
初步识别单元 102
最终识别单元 103

图2

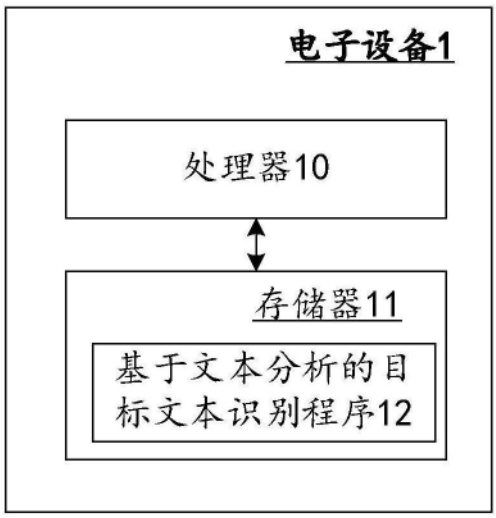


图3