(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 114385815 A
(43)申请公布日 2022.04.22

(21)申请号 202210032540.2

(22)申请日 2022.01.12

(71)申请人 平安普惠企业管理有限公司
地址 518000 广东省深圳市前海深港合作区前湾一路1号A栋201室

(72)发明人 刘锴靖

(74)专利代理机构 深圳市沃德知识产权代理事务所(普通合伙) 44347
代理人 高杰 于志光

(51)Int.Cl.
G06F 16/35(2019.01)
G06F 40/284(2020.01)

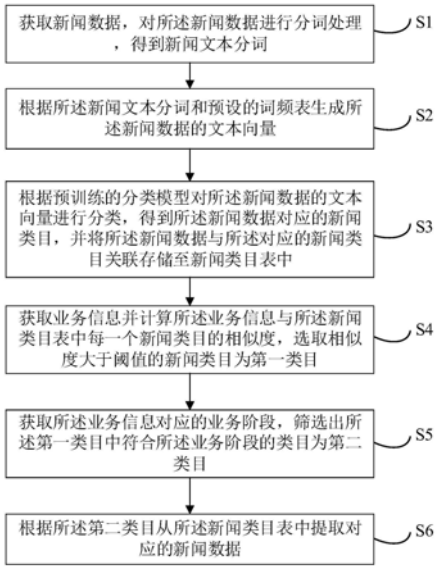
权利要求书2页 说明书10页 附图3页

(54)发明名称

基于业务需求的新闻筛选方法、装置、设备及存储介质

(57)摘要

本发明涉及人工智能技术，揭露一种基于业务需求的新闻筛选方法，包括：获取新闻数据并进行分词处理；根据分词的结果和词频表生成新闻数据的文本向量；根据分类模型对文本向量进行分类，得到对应的新闻类目，将新闻数据与对应的新闻类目关联存储至新闻类目表中；计算获取的业务信息与新闻类目表中的新闻类目的相似度，选取相似度大于阈值的新闻类目为第一类目；筛选出第一类目中符合业务信息对应的业务阶段的类目为第二类目；根据第二类目从新闻类目表中提取对应的新闻数据。此外，本发明还涉及区块链技术，新闻数据可存储于区块链的节点。本发明还提出一种基于业务需求的新闻筛选装置、设备以及介质。本发明可以提高获取符合业务需求新闻的效率。



CN 114385815 A

1.一种基于业务需求的新闻筛选方法,其特征在于,所述方法包括:

获取新闻数据,对所述新闻数据进行分词处理,得到新闻文本分词;

根据所述新闻文本分词和预设的词频表生成所述新闻数据的文本向量;

根据预训练的分类模型对所述新闻数据的文本向量进行分类,得到所述新闻数据对应的新闻类目,并将所述新闻数据与所述对应的新闻类目关联存储至新闻类目表中;

获取业务信息并计算所述业务信息与所述新闻类目表中每一个新闻类目的相似度,选取相似度大于阈值的新闻类目为第一类目;

获取所述业务信息对应的业务阶段,筛选出所述第一类目中符合所述业务阶段的类目为第二类目;

根据所述第二类目从所述新闻类目表中提取对应的新闻数据。

2.如权利要求1所述的基于业务需求的新闻筛选方法,其特征在于,所述对所述新闻数据进行分词处理,得到新闻文本分词,包括:

从所述新闻数据中提取标题以及摘要作为标准文本;

利用分词器对所述标准文本进行分词处理,得到第一分词;

根据预设的词性表和停用词表删除所述文本分词的特定分词,得到第二分词;

删除所述第二分词中的标点符号,得到新闻文本分词。

3.如权利要求2所述的基于业务需求的新闻筛选方法,其特征在于,所述根据预设的词性表和停用词表删除所述文本分词的特定分词,得到第二分词,包括:

获取需删除的词性标签,并根据所述需删除的词性标签提取所述词性表中对应的分词;

提取所述停用词表中对应的分词;

从所述第一分词中删除与所述词性表中对应的分词及所述停用词表中对应的分词相同的分词,得到所述第二分词。

4.如权利要求1所述的基于业务需求的新闻筛选方法,其特征在于,所述根据所述新闻文本分词和预设的词频表生成所述新闻数据的文本向量,包括:

在所述词频表中提取所述新闻文本分词对应的编号,根据所述编号生成编号向量;

判断所述编号向量的长度是否超过预设长度;

若所述编号向量的长度超过预设长度,则从所述编号向量中截取所述预设长度的向量作为所述文本向量;

若所述编号向量的长度未超过预设长度,则对所述编号向量补零,直至所述编号向量的长度达到预设长度,将补零后的编号向量作为文本向量。

5.如权利要求1所述的基于业务需求的新闻筛选方法,其特征在于,所述根据预训练的分类模型对所述新闻数据的文本向量进行分类,得到所述新闻数据对应的新闻类目,包括:

将所述新闻数据的文本向量输入预设的分类模型所嵌入的word2vec得到向量矩阵;

通过所述分类模型对所述向量矩阵进行预设次数的卷积、池化和全连接,得到分类信息;

通过分类器计算所述分类信息属于每一个新闻类目的概率值;

选取概率值大于预设阈值的新闻类目作为所述新闻数据对应的新闻类目。

6.如权利要求1至5中任一项所述的基于业务需求的新闻筛选方法,其特征在于,所述

计算所述业务信息与所述新闻类目表中每一个新闻类目的相似度，包括：

对所述业务信息的文本进行分词，得到业务文本分词；

根据所述业务文本分词在所述词频表中的词频提取关键词；

将所述关键词逐一与所述新闻类目表中的每一个新闻类目进行相似度计算。

7.如权利要求1所述的基于业务需求的新闻筛选方法，其特征在于，所述筛选出所述第一类目中符合所述业务阶段的类目为第二类目，包括：

获取所述业务阶段的业务标签，并逐一计算所述第一类目的每一个类目与所述业务标签的距离值；

从所述第一类目中选取所述距离值小于预设阈值的类目作为第二类目。

8.一种基于业务需求的新闻筛选装置，其特征在于，所述装置包括：

新闻文本分词生成模块，用于获取新闻数据，对所述新闻数据进行分词处理，得到新闻文本分词；

文本向量生成模块，用于根据所述新闻文本分词和预设的词频表生成所述新闻数据的文本向量；

新闻类目表生成模块，用于根据预训练的分类模型对所述新闻数据的文本向量进行分类，得到所述新闻数据对应的新闻类目，并将所述新闻数据与所述对应的新闻类目关联存储至新闻类目表中；

第一类目获取模块，用于获取业务信息并计算所述业务信息与所述新闻类目表中每一个新闻类目的相似度，选取相似度大于阈值的新闻类目为第一类目；

第二类目获取模块，用于获取所述业务信息对应的业务阶段，筛选出所述第一类目中符合所述业务阶段的类目为第二类目；

新闻数据获取模块，用于根据所述第二类目从所述新闻类目表中提取对应的新闻数据。

9.一种电子设备，其特征在于，所述电子设备包括：

至少一个处理器；以及，

与所述至少一个处理器通信连接的存储器；其中，

所述存储器存储有可被所述至少一个处理器执行的计算机程序，所述计算机程序被所述至少一个处理器执行，以使所述至少一个处理器能够执行如权利要求1至7中任意一项所述的基于业务需求的新闻筛选方法。

10.一种计算机可读存储介质，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中任意一项所述的基于业务需求的新闻筛选方法。

基于业务需求的新闻筛选方法、装置、设备及存储介质

技术领域

[0001] 本发明涉及人工智能技术领域，尤其涉及一种基于业务需求的新闻筛选方法、装置、电子设备及计算机可读存储介质。

背景技术

[0002] 在数字化浪潮的背景下，企业为保证稳步发展与市场地位，需要不断的向外学习，了解实时变动，因此企业从海量的新闻中获取符合自身发展的新闻数据变得尤为重要。现在大多数企业获取新闻数据的方法是通过企业关键词在互联网中进行搜索，此方法的工作量较大且获取的新闻没有进行分类，条理性差。

发明内容

[0003] 本发明提供一种基于业务需求的新闻筛选方法、装置及计算机可读存储介质，其主要目的在于解决获取符合业务需求的新闻效率低的问题。

[0004] 为实现上述目的，本发明提供的一种基于业务需求的新闻筛选方法，包括：

[0005] 获取新闻数据，对所述新闻数据进行分词处理，得到新闻文本分词；

[0006] 根据所述新闻文本分词和预设的词频表生成所述新闻数据的文本向量；

[0007] 根据预训练的分类模型对所述新闻数据的文本向量进行分类，得到所述新闻数据对应的新闻类目，并将所述新闻数据与所述对应的新闻类目关联存储至新闻类目表中；

[0008] 获取业务信息并计算所述业务信息与所述新闻类目表中每一个新闻类目的相似度，选取相似度大于阈值的新闻类目为第一类目；

[0009] 获取所述业务信息对应的业务阶段，筛选出所述第一类目中符合所述业务阶段的类目为第二类目；

[0010] 根据所述第二类目从所述新闻类目表中提取对应的新闻数据。

[0011] 可选地，所述对所述新闻数据进行分词处理，得到新闻文本分词，包括：

[0012] 从所述新闻数据中提取标题以及摘要作为标准文本；

[0013] 利用分词器对所述标准文本进行分词处理，得到第一分词；

[0014] 根据预设的词性表和停用词表删除所述文本分词的特定分词，得到第二分词；

[0015] 删除所述第二分词中的标点符号，得到新闻文本分词。

[0016] 可选地，所述根据预设的词性表和停用词表删除所述文本分词的特定分词，得到第二分词，包括：

[0017] 获取需删除的词性标签，并根据所述需删除的词性标签提取所述词性表中对应的分词；

[0018] 提取所述停用词表中对应的分词；

[0019] 从所述第一分词中删除与所述词性表中对应的分词及所述停用词表中对应的分词相同的分词，得到所述第二分词。

[0020] 可选地，所述根据所述新闻文本分词和预设的词频表生成所述新闻数据的文本向

量,包括:

[0021] 在所述词频表中提取所述新闻文本分词对应的编号,根据所述编号生成编号向量;

[0022] 判断所述编号向量的长度是否超过预设长度;

[0023] 若所述编号向量的长度超过预设长度,则从所述编号向量中截取所述预设长度的向量作为所述文本向量;

[0024] 若所述编号向量的长度未超过预设长度,则对所述编号向量补零,直至所述编号向量的长度达到预设长度,将补零后的编号向量作为文本向量。

[0025] 可选地,所述根据预训练的分类模型对所述新闻数据的文本向量进行分类,得到所述新闻数据对应的新闻类目,包括:

[0026] 将所述新闻数据的文本向量输入预设的分类模型所嵌入的word2vec得到向量矩阵;

[0027] 通过所述分类模型对所述向量矩阵进行预设次数的卷积、池化和全连接,得到分类信息;

[0028] 通过分类器计算所述分类信息属于每一个新闻类目的概率值;

[0029] 选取概率值大于预设阈值的新闻类目作为所述新闻数据对应的新闻类目。

[0030] 可选地,所述计算所述业务信息与所述新闻类目表中每一个新闻类目的相似度,包括:

[0031] 对所述业务信息的文本进行分词,得到业务文本分词;

[0032] 根据所述业务文本分词在所述词频表中的词频提取关键词;

[0033] 将所述关键词逐一与所述新闻类目表中的每一个新闻类目进行相似度计算。

[0034] 可选地,所述筛选出所述第一类目中符合所述业务阶段的类目为第二类目,包括:

[0035] 获取所述业务阶段的业务标签,并逐一计算所述第一类目的每一个类目与所述业务标签的距离值;

[0036] 从所述第一类目中选取所述距离值小于预设阈值的类目作为第二类目。为了解决上述问题,本发明还提供一种基于业务需求的新闻筛选装置,所述装置包括:

[0037] 新闻文本分词生成模块,用于获取新闻数据,对所述新闻数据进行分词处理,得到新闻文本分词;

[0038] 文本向量生成模块,用于根据所述新闻文本分词和预设的词频表生成所述新闻数据的文本向量;

[0039] 新闻类目表生成模块,用于根据预训练的分类模型对所述新闻数据的文本向量进行分类,得到所述新闻数据对应的新闻类目,并将所述新闻数据与所述对应的新闻类目关联存储至新闻类目表中;

[0040] 第一类目获取模块,用于获取业务信息并计算所述业务信息与所述新闻类目表中每一个新闻类目的相似度,选取相似度大于阈值的新闻类目为第一类目;

[0041] 第二类目获取模块,用于获取所述业务信息对应的业务阶段,筛选出所述第一类目中符合所述业务阶段的类目为第二类目;

[0042] 新闻数据获取模块,用于根据所述第二类目从所述新闻类目表中提取对应的新闻数据。

[0043] 为了解决上述问题，本发明还提供一种电子设备，所述电子设备包括：

[0044] 至少一个处理器；以及，

[0045] 与所述至少一个处理器通信连接的存储器；其中，

[0046] 所述存储器存储有可被所述至少一个处理器执行的计算机程序，所述计算机程序被所述至少一个处理器执行，以使所述至少一个处理器能够执行上述所述的基于业务需求的新闻筛选方法。

[0047] 为了解决上述问题，本发明还提供一种计算机可读存储介质，所述计算机可读存储介质中存储有至少一个计算机程序，所述至少一个计算机程序被电子设备中的处理器执行以实现上述所述的基于业务需求的新闻筛选方法。

[0048] 本发明实施例通过抓取新闻数据，进行处理以及分类，使新闻类目表中的新闻数据处于变动更新中，方便企业获取新闻数据，提高了获取新闻数据的效率；通过对业务信息及业务阶段与新闻类目表中的新闻类目计算，得到符合业务阶段需求的新闻类目，使业务需求更加直观，并且结合存储有预先抓取的新闻分类表，能够快速且直接的获取符合业务需求的新闻类目以及对应的新闻，并且使得新闻数据具有条理性。因此本发明提出的基于业务需求的新闻筛选方法、装置、电子设备及计算机可读存储介质，可以解决获取符合业务需求的新闻效率低的问题。

附图说明

[0049] 图1为本发明一实施例提供的基于业务需求的新闻筛选方法的流程示意图；

[0050] 图2为本发明一实施例提供的生成新闻文本分词的流程示意图；

[0051] 图3为本发明一实施例提供的得到所述新闻数据对应的新闻类目的流程示意图；

[0052] 图4为本发明一实施例提供的基于业务需求的新闻筛选装置的功能模块图；

[0053] 图5为本发明一实施例提供的实现所述基于业务需求的新闻筛选方法的电子设备的结构示意图。

[0054] 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

[0055] 应当理解，此处所描述的具体实施例仅仅用以解释本发明，并不用于限定本发明。

[0056] 本申请实施例提供一种基于业务需求的新闻筛选方法。所述基于业务需求的新闻筛选方法的执行主体包括但不限于服务端、终端等能够被配置为执行本申请实施例提供的该方法的电子设备中的至少一种。换言之，所述基于业务需求的新闻筛选方法可以由安装在终端设备或服务端设备的软件或硬件来执行，所述软件可以是区块链平台。所述服务端包括但不限于：单台服务器、服务器集群、云端服务器或云端服务器集群等。所述服务器可以是独立的服务器，也可以是提供云服务、云数据库、云计算、云函数、云存储、网络服务、云通信、中间件服务、域名服务、安全服务、内容分发网络(ContentDelivery Network，CDN)、以及大数据和人工智能平台等基础云计算服务的云服务器。

[0057] 参照图1所示，为本发明一实施例提供的基于业务需求的新闻筛选方法的流程示意图。在本实施例中，所述基于业务需求的新闻筛选方法包括：

[0058] S1、获取新闻数据，对所述新闻数据进行分词处理，得到新闻文本分词；

[0059] 本发明实施例中，所述新闻数据包括在发布在互联网上的实时或者历史新闻、报告、论文等。

[0060] 本发明实施例中，可通过爬虫技术从网络上抓取新闻数据，或者本发明实施例还可以利用具有数据抓取功能的python语句从用于存储所述新闻数据的区块链节点中抓取所述新闻数据，利用区块链对数据的高吞性，可提高获取新闻的效率。

[0061] 本发明实施例中，请参阅图2所示，在对上述新闻数据进行分词处理之前可通过UltraEdit将所述新闻数据转换为json格式。

[0062] 本发明实施例中，所述对所述新闻数据进行分词处理，得到新闻文本分词，包括：

[0063] S11、从所述新闻数据中提取标题以及摘要作为标准文本；

[0064] S12、利用分词器对所述标准文本进行分词处理，得到第一分词；

[0065] S13、根据预设的词性表和停用词表删除所述文本分词的特定分词，得到第二分词；

[0066] S14、删除所述第二分词中的标点符号，得到新闻文本分词。

[0067] 具体地，所述根据预设的词性表和停用词表删除所述文本分词的特定分词，得到第二分词，包括：

[0068] 获取需删除的词性标签，并根据所述需删除的词性标签提取所述词性表中对应的分词；

[0069] 提取所述停用词表中对应的分词；

[0070] 从所述第一分词中删除与所述词性表中对应的分词及所述停用词表中对应的分词相同的分词，得到所述第二分词。

[0071] 本发明实施例中，所述分词器包括但不限于结巴分词器；所述词性表中的词性包括形容词、副词、形语素、名形词等。

[0072] S2、根据所述新闻文本分词和预设的词频表生成所述新闻数据的文本向量；

[0073] 本发明实施例中，所述词频表是根据训练数据的词语频率而形成的，包括词语以及对应的编号，词频越大，对应的编号则越小。

[0074] 本发明实施例中，所述根据所述新闻文本分词和预设的词频表生成所述新闻数据的文本向量之前，还包括：

[0075] 获取训练数据，并对所述训练数据进行分词处理，得到训练数据文本分词；

[0076] 统计所述训练数据文本分词的词频，根据词频大小进行逆向编号；

[0077] 将所述编号与对应的训练数据文本分词关联存储至所述词频表中。

[0078] 本发明实施例中，所述对所述训练数据进行分词处理，得到训练数据文本分词的步骤与上述S1中对所述新闻数据进行分词处理，得到新闻文本分词的步骤相同，在此不过多赘述。

[0079] 例如，假设对所述训练数据进行分词处理，得到10000个文本分词；其中部分文本分词会出现重复，根据对重复的文本分词进行词频统计，得到不重复的文本分词以及对应的词频，再根据词频大小进行逆顺序编号，如：文本分词1、文本分词2、文本分词3对应的词频分别为10、4、50，则文本分词2的编号数字>文本分词1的编号数字>文本分词3的编号数字，最后将不重复的文本分词以及对应的编号关联存储至词频表中。

[0080] 本发明实施例中，所述根据所述新闻文本分词和预设的词频表生成所述新闻数据

的文本向量，包括：

[0081] 在所述词频表中提取所述新闻文本分词对应的编号，根据所述编号生成编号向量；

[0082] 判断所述编号向量的长度是否超过预设长度；

[0083] 若所述编号向量的长度超过预设长度，则从所述编号向量中截取所述预设长度的向量作为所述文本向量；

[0084] 若所述编号向量的长度未超过预设长度，则对所述编号向量补零，直至所述编号向量的长度达到预设长度，将补零后的编号向量作为文本向量。

[0085] 例如，假设一个训练数据对应的文本分词为[深圳市这些路段实施全面封闭维修]，其中存在七个文本分词，分别对应的编号为：77、15、95、54、46、152、101，则编号对应的编码向量为[7715955446152101]；预设的长度假设为20，编码向量的长度小于预设长度，则对实施编号向量补零，得到文本向量为[00007715955446152101]。

[0086] S3、根据预训练的分类模型对所述新闻数据的文本向量进行分类，得到所述新闻数据对应的新闻类目，并将所述新闻数据与所述对应的新闻类目关联存储至新闻类目表中；

[0087] 本发明实施例中，所述分类模型包括但不限于基于预训练word2vec模型的CNN模型、多项分布朴素贝叶斯。

[0088] 本发明一可选实施例中，所述预训练的分类模型的训练过程为：将训练数据的文本向量输入预设的分类模型中，进行预设次数的卷积、池化和全连接，再通过分类器输出得到每一个新闻类目的概率值，根据所述每一个新闻类目的概率值及所述训练数据对应的新闻类型计算损失值；根据所述损失值优化所述分类模型，当优化后的分类模型分类得到的所述训练数据对应的新闻类目的概率值达到预设条件时，则表示所述分类模型训练成功，即得到所述预训练的分类模型。

[0089] 本发明实施例中，请参阅图3所示，所述根据预训练的分类模型对所述新闻数据的文本向量进行分类，得到所述新闻数据对应的新闻类目，包括：

[0090] S31、将所述新闻数据的文本向量输入预设的分类模型所嵌入的word2vec中得到向量矩阵；

[0091] S32、通过所述分类模型对所述向量矩阵进行预设次数的卷积、池化和全连接，得到分类信息；

[0092] S33、通过分类器计算所述分类信息属于每一个新闻类目的概率值；

[0093] S34、选取概率值大于预设阈值的新闻类目作为所述新闻数据对应的新闻类目。

[0094] 例如，假设存在新闻类目有：IT、财经、体育、教育这四个新闻类目，新闻数据A的文本向量输入所述分类模型得到IT、财经、体育、教育这四个新闻类目概率分别为：0.2、0.5.、0.3、0.1，则确定所述新闻数据A对应的新闻类目为财经。

[0095] 本发明实施例中，所述将所述新闻数据与所述对应的新闻类目关联存储至新闻类目表中，包括：

[0096] 将所述新闻数据与所述对应的新闻类目形成映射关系；

[0097] 在所述新闻类目表中提取新闻类目对应的列表标签，将与所述新闻数据填入所述列表标签对应的列表内。

[0098] 本发明实施例中，所述新闻类目表可以存储于数据库、区块链节点、网络缓存中。

[0099] S4、获取业务信息并计算所述业务信息与所述新闻类目表中每一个新闻类目的相似度，选取相似度大于阈值的新闻类目为第一类目；

[0100] 本发明实施例中，所述业务信息包括企业简介、企业网络名片、企业的业务部门等，本发明实施例中可通过具有数据抓取功能的语句或应用从用于存储所述企业信息的区块链节点中抓取所述业务信息。

[0101] 本发明实施例中，所述计算所述业务信息与所述新闻类目表中每一个新闻类目的相似度，包括：

[0102] 对所述业务信息的文本进行分词，得到业务文本分词；

[0103] 根据所述业务文本分词在所述词频表中的词频提取关键词；

[0104] 将所述关键词逐一与所述新闻类目表中的每一个新闻类目进行相似度计算。

[0105] 进一步地，可通过如下公式对所述关键词与所述新闻类目表中的每一个新闻类目进行相似度计算：

[0106] $$\cos\theta = \frac{a \cdot b_i}{\|a\| \times \|b_i\|}$$

[0107] 其中，所述$\cos\theta$为相似度，a为所述关键词，b_i为所述新闻类目表中第i个新闻类目。

[0108] 本发明实施例中，相似度越大则说明所述关键性与该相似度对应的新闻类目越相似，当所述相似度大于预设阈值则可以确定改相似度对应的新闻类目为所述业务信息对应的新闻类目。

[0109] 例如，假设存在新闻类目有：IT、财经、体育、教育这四个新闻类目，业务信息B的关键词分别与新闻类目表中的上述四个新闻类目进行相似度计算，分别得到IT、财经、体育、教育的新闻类目的相似度为0.8、0.7、0.2、0.2，预设阈值为0.6，因此确定业务信息B所对应的第一类目为IT和财经。

[0110] S5、获取所述业务信息对应的业务阶段，筛选出所述第一类目中符合所述业务阶段的类目为第二类目；

[0111] 本发明实施例中，不同业务阶段所需要的信息可能是不同的，例如在产品上市业务的前期阶段，所需要的信息主要为市面上相关产品的功能、产品介绍等，在产品上市的后期阶段，所需要的信息更多为市场环境、产品行情等。

[0112] 本发明实施例中，所述筛选出所述第一类目中符合所述业务阶段的类目为第二类目，包括：

[0113] 获取所述业务阶段的业务标签，并逐一计算所述第一类目的每一个类目与所述业务标签的距离值；

[0114] 从所述第一类目中选取所述距离值小于预设阈值的类目作为第二类目。

[0115] 本发明实施例中，所述具体地，所述根据所述业务标签计算与所述第一类目的距离值，包括：

[0116] 利用如下距离值算法分别计算所述业务标签与所述第一类目的每一个类目之间的距离值：

[0117] $$D=\frac{\sqrt{a^2+{b_i}^2}}{[(a+b_i)(a-b_i)]^2}$$

[0118] 其中，D为所述距离值，a为所述业务标签，b_i为所述第一类目中第i个类目。

[0119] S6、根据所述第二类目从所述新闻类目表中提取对应的新闻数据。

[0120] 本发明实施例中，所述新闻类目表中的新闻数据可以定期抓取新闻数据进行分类和关联存储，在确定符合要求的第二类目后，可以直接从所述新闻类目表中提取该第二类目下的新闻数据。

[0121] 例如，假设目标类目为IT和财经，在新闻类目表中检索IT和财经的新闻类目的列表位置，检索到IT和财经在所述新闻类目表中的列表位置后，提取该列表位置对应的列表内的新闻数据。

[0122] 本发明实施例通过抓取新闻数据，进行处理以及分类，使新闻类目表中的新闻数据处于变动更新中，方便企业获取新闻数据，提高了获取新闻数据的效率；通过对业务信息及业务阶段与新闻类目表中的新闻类目计算，得到符合业务阶段需求的新闻类目，使业务需求更加直观，并且结合存储有预先抓取的新闻分类表，能够快速且直接的获取符合业务需求的新闻类目以及对应的新闻，并且使得新闻数据具有条理性。因此本发明提出的基于业务需求的新闻筛选方法，可以解决获取符合业务需求的新闻效率低的问题。

[0123] 如图4所示，是本发明一实施例提供的基于业务需求的新闻筛选装置的功能模块图。

[0124] 本发明所述基于业务需求的新闻筛选装置100可以安装于电子设备中。根据实现的功能，所述基于业务需求的新闻筛选装置100可以包括新闻文本分词生成模块101、文本向量生成模块102、新闻类目表生成模块103及新闻数据获取模块104。本发明所述模块也可以称之为单元，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。

[0125] 在本实施例中，关于各模块/单元的功能如下：

[0126] 所述新闻文本分词生成模块101，用于获取新闻数据，对所述新闻数据进行分词处理，得到新闻文本分词；

[0127] 所述文本向量生成模块102，用于根据所述新闻文本分词和预设的词频表生成所述新闻数据的文本向量；

[0128] 所述新闻类目表生成模块103，用于根据预训练的分类模型对所述新闻数据的文本向量进行分类，得到所述新闻数据对应的新闻类目，并将所述新闻数据与所述对应的新闻类目关联存储至新闻类目表中；

[0129] 所述第一类目获取模块104，用于获取业务信息并计算所述业务信息与所述新闻类目表中每一个新闻类目的相似度，选取相似度大于阈值的新闻类目为第一类目；

[0130] 所述第二类目获取模块105，用于获取所述业务信息对应的业务阶段，筛选出所述第一类目中符合所述业务阶段的类目为第二类目；

[0131] 所述新闻数据获取模块106，用于根据所述第二类目从所述新闻类目表中提取对应的新闻数据。

[0132] 所述详细地，本发明实施例中所述基于业务需求的新闻筛选装置100中所述的各

模块在使用时采用与上述图1至图3中所述的基于业务需求的新闻筛选方法一样的技术手段，并能够产生相同的技术效果，这里不再赘述。

[0133] 如图5所示，是本发明一实施例提供的实现基于业务需求的新闻筛选方法的电子设备的结构示意图。

[0134] 所述电子设备1可以包括处理器10、存储器11、通信总线12以及通信接口13，还可以包括存储在所述存储器11中并可在所述处理器10上运行的计算机程序，如基于业务需求的新闻筛选程序。

[0135] 其中，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(例如执行基于业务需求的新闻筛选程序等)，以及调用存储在所述存储器11内的数据，以执行电子设备的各种功能和处理数据。

[0136] 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的内部存储单元，例如该电子设备的移动硬盘。所述存储器11在另一些实施例中也可以是电子设备的外部存储设备，例如电子设备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的应用软件及各类数据，例如基于业务需求的新闻筛选程序的代码等，还可以用于暂时地存储已经输出或者将要输出的数据。

[0137] 所述通信总线12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。

[0138] 所述通信接口13用于上述电子设备与其他设备之间的通信，包括网络接口和用户接口。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与其他电子设备之间建立通信连接。所述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中处理的信息以及用于显示可视化的用户界面。

[0139] 图5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5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。

[0140] 例如，尽管未示出，所述电子设备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。

[0141] 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。

[0142] 所述电子设备1中的所述存储器11存储的基于业务需求的新闻筛选程序是多个指令的组合，在所述处理器10中运行时，可以实现：

[0143] 获取新闻数据，对所述新闻数据进行分词处理，得到新闻文本分词；

[0144] 根据所述新闻文本分词和预设的词频表生成所述新闻数据的文本向量；

[0145] 根据预训练的分类模型对所述新闻数据的文本向量进行分类，得到所述新闻数据对应的新闻类目，并将所述新闻数据与所述对应的新闻类目关联存储至新闻类目表中；

[0146] 获取业务信息并计算所述业务信息与所述新闻类目表中每一个新闻类目的相似度，选取相似度大于阈值的新闻类目为第一类目；

[0147] 获取所述业务信息对应的业务阶段，筛选出所述第一类目中符合所述业务阶段的类目为第二类目；

[0148] 根据所述第二类目从所述新闻类目表中提取对应的新闻数据。具体地，所述处理器10对上述指令的具体实现方法可参考附图对应实施例中相关步骤的描述，在此不赘述。

[0149] 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存储介质中。所述计算机可读存储介质可以是易失性的，也可以是非易失性的。例如，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)。

[0150] 本发明还提供一种计算机可读存储介质，所述可读存储介质存储有计算机程序，所述计算机程序在被电子设备的处理器所执行时，可以实现：

[0151] 获取新闻数据，对所述新闻数据进行分词处理，得到新闻文本分词；

[0152] 根据所述新闻文本分词和预设的词频表生成所述新闻数据的文本向量；

[0153] 根据预训练的分类模型对所述新闻数据的文本向量进行分类，得到所述新闻数据对应的新闻类目，并将所述新闻数据与所述对应的新闻类目关联存储至新闻类目表中；

[0154] 获取业务信息并计算所述业务信息与所述新闻类目表中每一个新闻类目的相似度，选取相似度大于阈值的新闻类目为第一类目；

[0155] 获取所述业务信息对应的业务阶段，筛选出所述第一类目中符合所述业务阶段的类目为第二类目；

[0156] 根据所述第二类目从所述新闻类目表中提取对应的新闻数据。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。

[0157] 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。

[0158] 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。

[0159] 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。

[0160] 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。

[0161] 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。

[0162] 本申请实施例可以基于人工智能技术对相关的数据进行获取和处理。其中，人工智能(Artificial Intelligence,AI)是利用数字计算机或者数字计算机控制的机器模拟、延伸和扩展人的智能，感知环境、获取知识并使用知识获得最佳结果的理论、方法、技术及应用系统。

[0163] 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。第一、第二等词语用来表示名称，而并不表示任何特定的顺序。

[0164] 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。

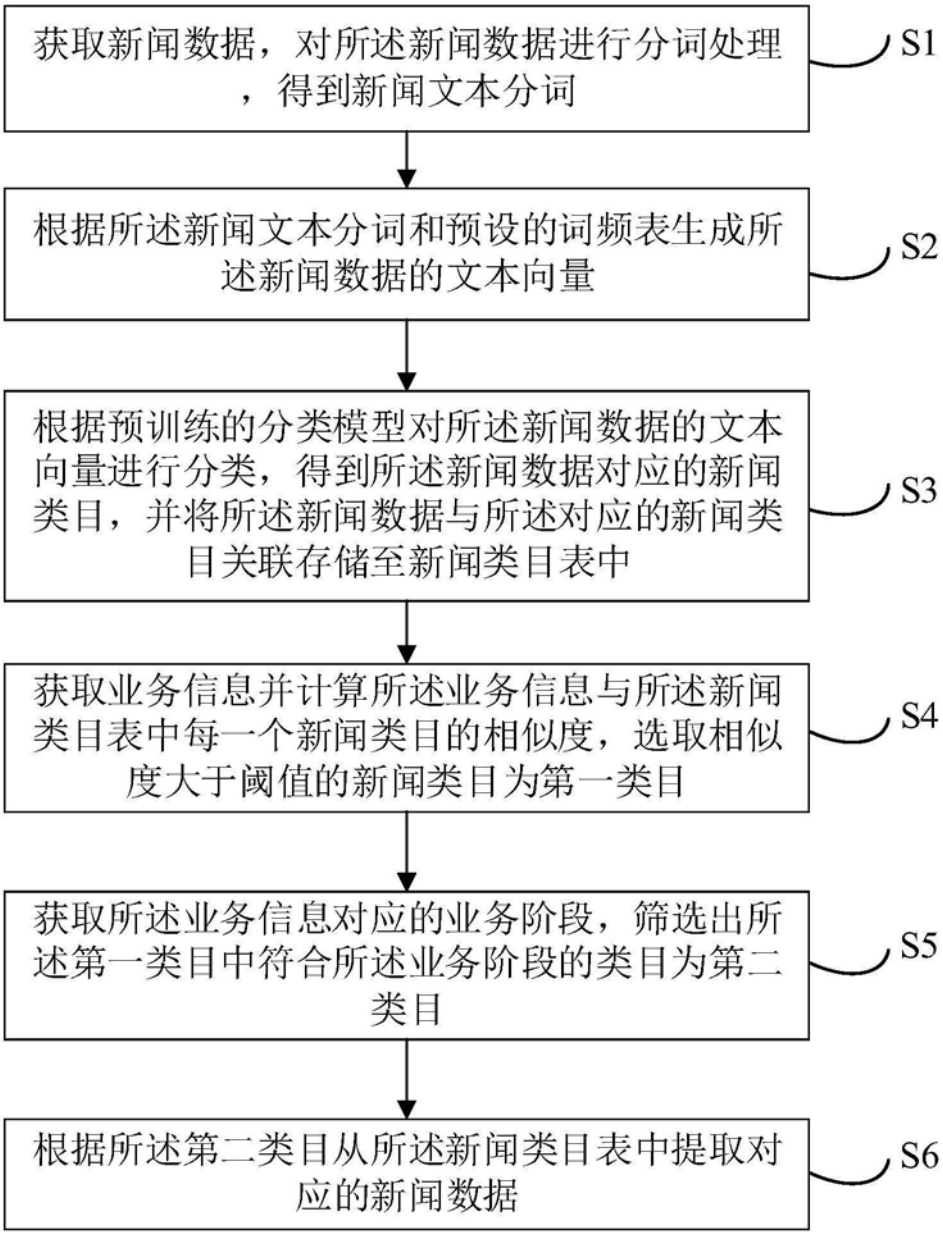

获取新闻数据，对所述新闻数据进行分词处理，得到新闻文本分词
S1
根据所述新闻文本分词和预设的词频表生成所述新闻数据的文本向量
S2
根据预训练的分类模型对所述新闻数据的文本向量进行分类，得到所述新闻数据对应的新闻类目，并将所述新闻数据与所述对应的新闻类目关联存储至新闻类目表中
S3
获取业务信息并计算所述业务信息与所述新闻类目表中每一个新闻类目的相似度，选取相似度大于阈值的新闻类目为第一类目
S4
获取所述业务信息对应的业务阶段，筛选出所述第一类目中符合所述业务阶段的类目为第二类目
S5
根据所述第二类目从所述新闻类目表中提取对应的新闻数据
S6

图1

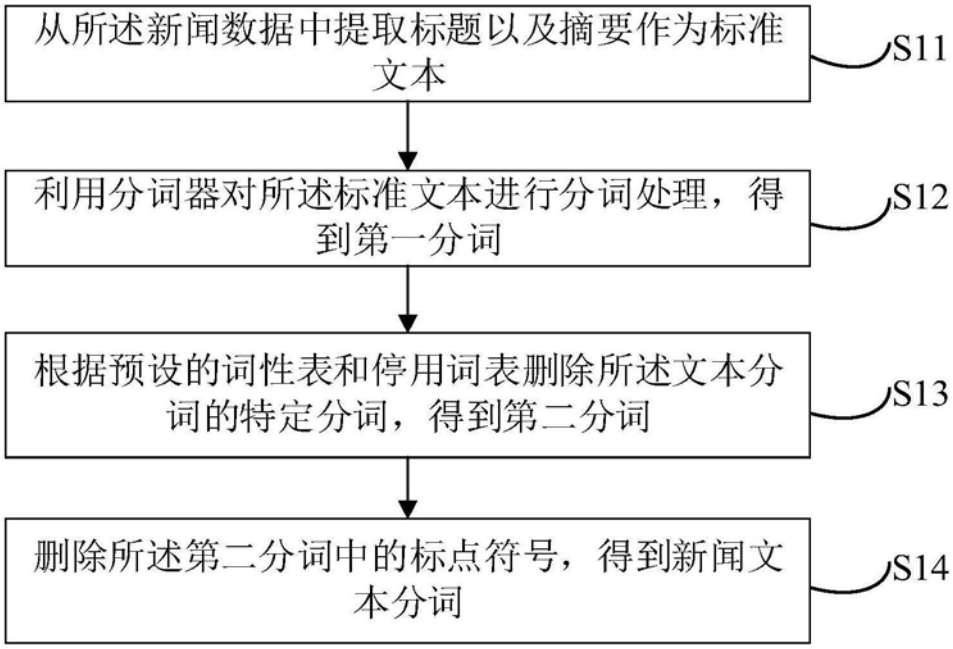

从所述新闻数据中提取标题以及摘要作为标准文本
S11
利用分词器对所述标准文本进行分词处理，得到第一分词
S12
根据预设的词性表和停用词表删除所述文本分词的特定分词，得到第二分词
S13
删除所述第二分词中的标点符号，得到新闻文本分词
S14

图2

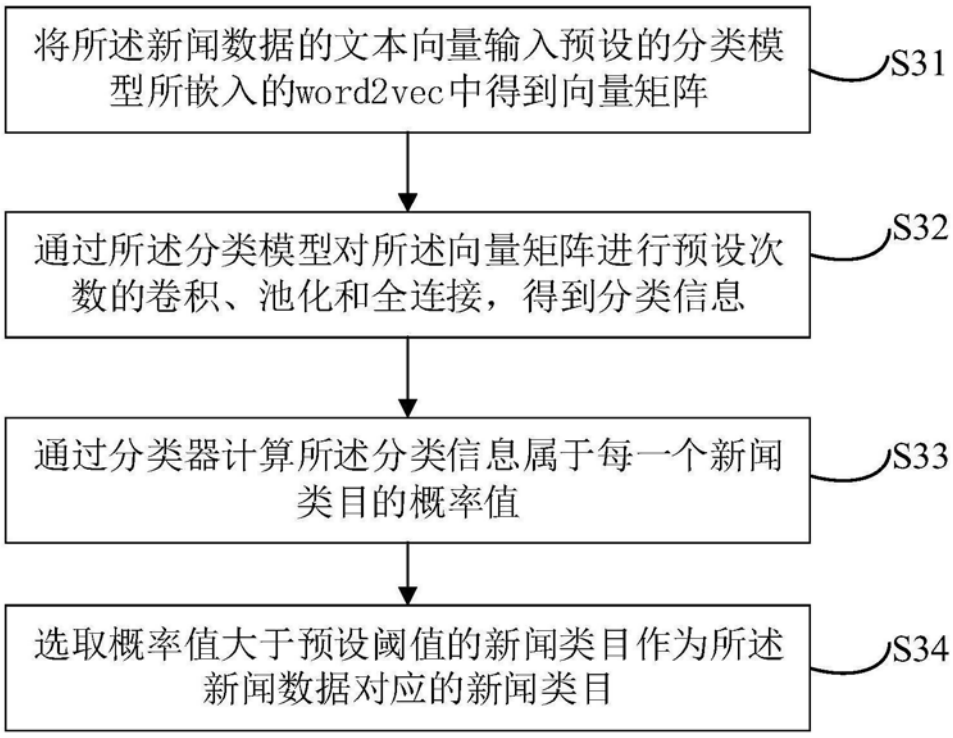

将所述新闻数据的文本向量输入预设的分类模型所嵌入的word2vec中得到向量矩阵
S31
通过所述分类模型对所述向量矩阵进行预设次数的卷积、池化和全连接，得到分类信息
S32
通过分类器计算所述分类信息属于每一个新闻类目的概率值
S33
选取概率值大于预设阈值的新闻类目作为所述新闻数据对应的新闻类目
S34

图3


基于业务需求的新闻筛选装置
100
新闻文本分词生成模块
101
文本向量生成模块
102
新闻类目表生成模块
103
第一类目获取模块
104
第二类目获取模块
105
新闻数据获取模块
106

图4

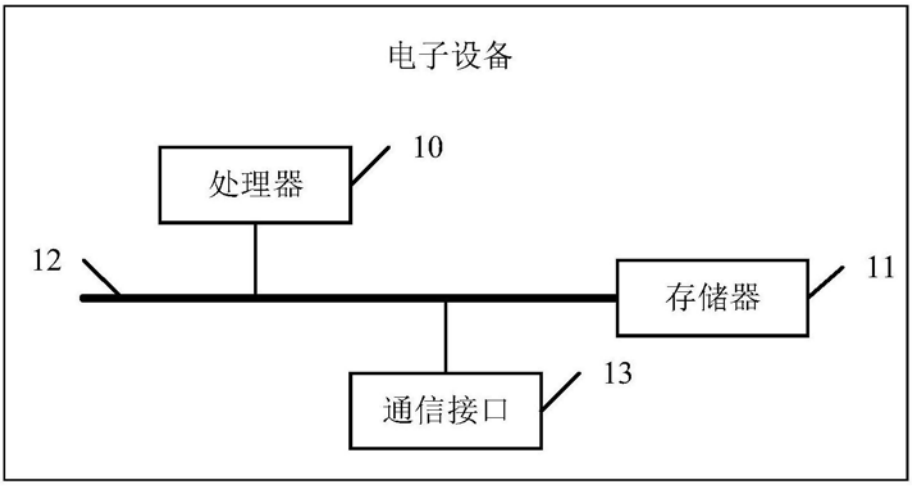

电子设备
处理器
10
12
存储器
11
通信接口
13

图5