(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 111985212 A
(43)申请公布日 2020.11.24

(21)申请号 202010910049.6

(22)申请日 2020.09.02

(71)申请人 深圳壹账通智能科技有限公司
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(入驻深圳市前海商务秘书有限公司)

(72)发明人 魏晓茹

(74)专利代理机构 北京英特普罗知识产权代理有限公司 11015
代理人 程超

(51)Int.Cl.
G06F 40/216(2020.01)
G06F 40/289(2020.01)

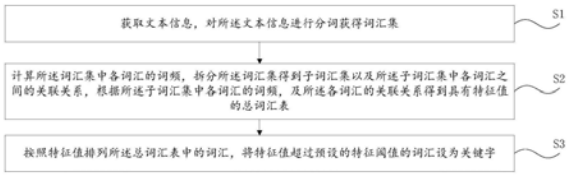
权利要求书2页 说明书10页 附图3页

(54)发明名称

文本关键字识别方法、装置、计算机设备及可读存储介质

(57)摘要

本发明涉及人工智能的智能决策技术领域，公开了一种文本关键字识别方法，包括：获取文本信息，对所述文本信息进行分词获得词汇集；计算所述词汇集中各词汇的词频，拆分所述词汇集得到子词汇集以及所述子词汇集中各词汇之间的关联关系，根据所述子词汇集中各词汇的词频，及所述各词汇的关联关系得到具有特征值的总词汇表；按照特征值排列所述总词汇表中的词汇，将特征值超过预设的特征阈值的词汇设为关键字。本发明还涉及区块链技术，信息可存储于区块链节点中。本发明从词汇集中各词汇的词频，以及所述词汇集中任一词汇被其他词汇所依赖的程度的两个维度，评价词汇的关键程度，提高了获得能够反映文本信息核心含义的关键字的准确度。



CN 111985212 A

1.一种文本关键字识别方法,其特征在于,包括:

获取文本信息,对所述文本信息进行分词获得词汇集;

计算所述词汇集中各词汇的词频,拆分所述词汇集得到子词汇集以及所述子词汇集中各词汇之间的关联关系,根据所述子词汇集中各词汇的词频,及所述各词汇的关联关系得到具有特征值的总词汇表,其中,所述特征值反映了词汇在文本信息中的关键程度;

按照特征值排列所述总词汇表中的词汇,将特征值超过预设的特征阈值的词汇设为关键字。

2.根据权利要求1所述的文本关键字识别方法,其特征在于,对所述文本信息进行分词获得词汇集包括:

通过自然语言技术对文本信息进行分词,得到至少具有一个词汇的词汇集。

3.根据权利要求1所述的文本关键字识别方法,其特征在于,计算所述词汇集中各词汇的词频,拆分所述词汇集得到子词汇集以及所述子词汇集中各词汇之间的关联关系的步骤,包括:

计算所述词汇集中各词汇在文本信息中出现的次数,以获得所述词汇的词频;

按照预设的分隔符拆分所述词汇集得到至少一个子词汇集,根据所述子词汇集中各词汇之间的关联关系制定关联特征;其中,所述词频反映了词汇在词汇集中出现的频率,所述关联特征是以特征向量的形式表达了子词汇集中任一词汇与其他词汇之间的关联关系。

4.根据权利要求3所述的文本关键字识别方法,其特征在于,根据所述子词汇集中各词汇的词频,及所述各词汇的关联关系得到具有特征值的总词汇表的步骤,包括:

将所述子词汇集的关联特征及其中各词汇的词频分别作为向量并对其进行运算,得到具有词汇及其特征值的子关键特征;

根据所述子关键特征的特征值排列所述子词汇集中的词汇得到对应有特征值的子词汇列表,汇总所述子词汇列表形成总词汇表;

汇总所述子词汇列表形成总词汇表之后,还包括:

将所述总词汇表上传至区块链中。

5.根据权利要求3所述的文本关键字识别方法,其特征在于,计算所述词汇集中各词汇在文本信息中出现的次数,以获得所述词汇的词频的步骤,包括:

计算所述词汇集中所有词汇的总数,及对所述词汇集中的词汇进行去重得到词汇表;

计算所述词汇表中各词汇在所述词汇集中出现的次数,将所述词汇的次数与所述总数相除得到所述词汇的词频。

6.根据权利要求3所述的文本关键字识别方法,其特征在于,按照预设的分隔符拆分所述词汇集得到至少一个子词汇集,根据所述子词汇集中各词汇之间的关联关系制定关联特征的步骤,包括:

以标点符号为分隔符划分所述文本信息形成子文本信息,汇总所述词汇集中与子文本信息对应的词汇得到所述子文本信息的子词汇集;

识别子词汇集中在其子文本信息上处于相邻位置的两个词汇,并认定所述两个词汇之间具有关联关系;

根据所述子词汇集中各词汇之间的关联关系,制定能够表达子词汇集中任一词汇与其他词汇之间关联关系特征向量,以得到所述子词汇集的关联特征。

7.根据权利要求4所述的文本关键字识别方法，其特征在于，将所述子词汇集的关联特征及其中各词汇的词频分别作为向量并对其进行运算，得到具有词汇及其特征值的子关键特征的步骤，包括：

汇总子词汇集中各词汇及其词频得到词频向量；

将所述子词汇集的关联特征与所述词频向量相乘得到具有特征值的得到子关键特征，其中，所述特征值为子关键特征的元素值，所述子词汇集中的词汇与所述元素值一一对应。

8.一种文本关键字识别装置，其特征在于，包括：

输入分词模块，用于获取文本信息，对所述文本信息进行分词获得词汇集；

词频关联模块，用于计算所述词汇集中各词汇的词频，拆分所述词汇集得到子词汇集以及所述子词汇集中各词汇之间的关联关系，根据所述子词汇集中各词汇的词频，及所述各词汇的关联关系得到具有特征值的总词汇表，其中，所述特征值反映了词汇在文本信息中的关键程度；

关键字识别模块，用于按照特征值排列所述总词汇表中的词汇，将特征值超过预设的特征阈值的词汇设为关键字。

9.一种计算机设备，其包括存储器、处理器以及存储在存储器上并可在处理器上运行的计算机程序，其特征在于，所述计算机设备的处理器执行所述计算机程序时实现权利要求1至7任一项所述文本关键字识别方法的步骤。

10.一种计算机可读存储介质，所述可读存储介质上存储有计算机程序，其特征在于，所述可读存储介质存储的所述计算机程序被处理器执行时实现权利要求1至7任一项所述文本关键字识别方法的步骤。

文本关键字识别方法、装置、计算机设备及可读存储介质

技术领域

[0001] 本发明涉及人工智能的智能决策技术领域，尤其涉及一种文本关键字识别方法、装置、计算机设备及可读存储介质。

背景技术

[0002] 针对于企业舆情信息，当前主流舆情供应商采用的方法主要是对词库进行匹配，实现对文本信息进行分词的效果。并对词语进行一个简单的数量排序，对于数量较多的词汇，则作为相应的关键词。

[0003] 然而发明人意识到当前的方法通常是以词汇出现的次数作为词汇关键程度的评价指标，那么往往会将诸如“的”，‘最’，‘不仅’，‘非常’这些连词、介词、量词等与文本信息含义无关的词汇作为关键词，那么上述方法获得的关键词将无法准确把握文本信息的核心含义。

发明内容

[0004] 本发明的目的是提供一种文本关键字识别方法、装置、计算机设备及可读存储介质，用于解决现有技术存在的获得的关键词将无法准确把握文本信息的核心含义的问题；本申请可应用于智慧政务场景中，从而推动智慧城市的建设。

[0005] 为实现上述目的，本发明提供一种文本关键字识别方法，包括：

[0006] 获取文本信息，对所述文本信息进行分词获得词汇集；

[0007] 计算所述词汇集中各词汇的词频，拆分所述词汇集得到子词汇集以及所述子词汇集中各词汇之间的关联关系，根据所述子词汇集中各词汇的词频，及所述各词汇的关联关系得到具有特征值的总词汇表，其中，所述特征值反映了词汇在文本信息中的关键程度；

[0008] 按照特征值排列所述总词汇表中的词汇，将特征值超过预设的特征阈值的词汇设为关键字。

[0009] 上述方案中，对所述文本信息进行分词获得词汇集包括：

[0010] 通过自然语言技术对文本信息进行分词，得到至少具有一个词汇的词汇集。

[0011] 上述方案中，计算所述词汇集中各词汇的词频，拆分所述词汇集得到子词汇集以及所述子词汇集中各词汇之间的关联关系的步骤，包括：

[0012] 计算所述词汇集中各词汇在文本信息中出现的次数，以获得所述词汇的词频；

[0013] 按照预设的分隔符拆分所述词汇集得到至少一个子词汇集，根据所述子词汇集中各词汇之间的关联关系制定关联特征；其中，所述词频反映了词汇在词汇集中出现的频率，所述关联特征是以特征向量的形式表达了子词汇集中任一词汇与其他词汇之间的关联关系。

[0014] 上述方案中，根据所述子词汇集中各词汇的词频，及所述各词汇的关联关系得到具有特征值的总词汇表的步骤，包括：

[0015] 将所述子词汇集的关联特征及其中各词汇的词频分别作为向量并对其进行运算，

得到具有词汇及其特征值的子关键特征；

[0016] 根据所述子关键特征的特征值排列所述子词汇集中的词汇得到对应有特征值的子词汇列表，汇总所述子词汇列表形成总词汇表；

[0017] 汇总所述子词汇列表形成总词汇表之后，还包括：

[0018] 将所述总词汇表上传至区块链中。

[0019] 上述方案中，计算所述词汇集中各词汇在文本信息中出现的次数，以获得所述词汇的词频的步骤，包括：

[0020] 计算所述词汇集中所有词汇的总数，及对所述词汇集中的词汇进行去重得到词汇表；

[0021] 计算所述词汇表中各词汇在所述词汇集中出现的次数，将所述词汇的次数与所述总数相除得到所述词汇的词频。

[0022] 上述方案中，按照预设的分隔符拆分所述词汇集得到至少一个子词汇集，根据所述子词汇集中各词汇之间的关联关系制定关联特征的步骤，包括：

[0023] 以标点符号为分隔符划分所述文本信息形成子文本信息，汇总所述词汇集中与子文本信息对应的词汇得到所述子文本信息的子词汇集；

[0024] 识别子词汇集中在其子文本信息上处于相邻位置的两个词汇，并认定所述两个词汇之间具有关联关系；

[0025] 根据所述子词汇集中各词汇之间的关联关系，制定能够表达子词汇集中任一词汇与其他词汇之间关联关系特征向量，以得到所述子词汇集的关联特征。

[0026] 上述方案中，将所述子词汇集的关联特征及其中各词汇的词频分别作为向量并对其进行运算，得到具有词汇及其特征值的子关键特征的步骤，包括：

[0027] 汇总子词汇集中各词汇及其词频得到词频向量；

[0028] 将所述子词汇集的关联特征与所述词频向量相乘得到具有特征值的得到子关键特征，其中，所述特征值为子关键特征的元素值，所述子词汇集中的词汇与所述元素值一一对应。

[0029] 为实现上述目的，本发明还提供一种文本关键字识别装置，包括：

[0030] 输入分词模块，用于获取文本信息，对所述文本信息进行分词获得词汇集；

[0031] 词频关联模块，用于计算所述词汇集中各词汇的词频，拆分所述词汇集得到子词汇集以及所述子词汇集中各词汇之间的关联关系，根据所述子词汇集中各词汇的词频，及所述各词汇的关联关系得到具有特征值的总词汇表，其中，所述特征值反映了词汇在文本信息中的关键程度；

[0032] 关键字识别模块，用于按照特征值排列所述总词汇表中的词汇，将特征值超过预设的特征阈值的词汇设为关键字。

[0033] 为实现上述目的，本发明还提供一种计算机设备，其包括存储器、处理器以及存储在存储器上并可在处理器上运行的计算机程序，所述计算机设备的处理器执行所述计算机程序时实现上述文本关键字识别方法的步骤。

[0034] 为实现上述目的，本发明还提供一种计算机可读存储介质，所述可读存储介质上存储有计算机程序，所述可读存储介质存储的所述计算机程序被处理器执行时实现上述文本关键字识别方法的步骤。

[0035] 本发明提供的文本关键字识别方法、装置、计算机设备及可读存储介质，通过计算所述词汇集中各词汇的词频，以从词汇出现次数的维度评价了词汇的重要性；通过拆分所述词汇集得到子词汇集，并根据所述子词汇集中各词汇之间的关联关系制定关联特征，该关联特征反映了子词汇集中任一词汇被其他词汇所依赖的程度，以从被依赖程度的维度评价词汇的重要性，因此，实现了获得词汇集中各词汇的词频，以及所述词汇集中任一词汇被其他词汇所依赖的程度的两个维度，评价词汇的关键程度的效果，提高获得能够反映文本信息核心含义的关键字的准确度。

附图说明

[0036] 图1为本发明文本关键字识别方法实施例一的流程图；

[0037] 图2为本发明文本关键字识别方法实施例一中计算所述词汇集中各词汇的词频，拆分所述词汇集得到子词汇集以及所述子词汇集中各词汇之间的关联关系的流程图；

[0038] 图3为本发明文本关键字识别方法实施例一中根据所述子词汇集中各词汇的词频，及所述各词汇的关联关系得到具有特征值的总词汇表的流程图；

[0039] 图4为本发明文本关键字识别方法实施例一中计算所述词汇集中各词汇在文本信息中出现的次数，以获得所述词汇的词频的流程图；

[0040] 图5为本发明文本关键字识别方法实施例一中按照预设的分隔符拆分所述词汇集得到至少一个子词汇集，根据所述子词汇集中各词汇之间的关联关系制定关联特征的流程图；

[0041] 图6为本发明文本关键字识别方法实施例一中将所述子词汇集的关联特征及其中各词汇的词频分别作为向量并对其进行运算，得到具有词汇及其特征值的子关键特征的流程图；

[0042] 图7为本发明文本关键字识别装置实施例二的程序模块示意图；

[0043] 图8为本发明计算机设备实施例三中计算机设备的硬件结构示意图。

具体实施方式

[0044] 为了使本发明的目的、技术方案及优点更加清楚明白，以下结合附图及实施例，对本发明进行进一步详细说明。应当理解，此处所描述的具体实施例仅用以解释本发明，并不用于限定本发明。基于本发明中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0045] 本发明提供的文本关键字识别方法、装置、计算机设备及可读存储介质，适用于人工智能的智能决策技术领域，为提供一种基于输入分词模块、词频关联模块和关键字识别模块的文本关键字识别方法。本发明通过获取文本信息，对所述文本信息进行分词获得词汇集；计算所述词汇集中各词汇的词频，拆分所述词汇集得到子词汇集以及所述子词汇集中各词汇之间的关联关系，根据所述子词汇集中各词汇的词频，及所述各词汇的关联关系得到具有特征值的总词汇表，按照特征值排列所述总词汇表中的词汇，将特征值超过预设的特征阈值的词汇设为关键字。

[0046] 实施例一：

[0047] 请参阅图1，本实施例的一种文本关键字识别方法，包括：

[0048] S1:获取文本信息,对所述文本信息进行分词获得词汇集。

[0049] 本步骤中,通过自然语言技术(NLP)对文本信息进行分词,得到至少具有一个词汇的词汇集;因此,相比于现有技术中使用词库对文本信息进行分词的方法,本步骤利用自然语言技术结合文本信息的上下文,更加准确的对文本信息进行了分词,为准确识别关键字,把握文本信息核心含义提供了可靠的分词前提。

[0050] 示例性地,获取的文本信息如:“一个高尚的人,一个纯粹的人,一个有道德的人,一个脱离了低级趣味的人,一个有益于人民的人”,那么通过NLP自然语言抽取技术对文本信息分词,得到:“一个/高尚/的/人,一个/纯粹/的/人,一个/有道德/的/人,一个/脱离了低级趣味/的/人,一个/有益于人民/的/人”;将分词得到的词汇进行汇总,得到词汇集:“一个、高尚、的、人、一个、纯粹、的、人、一个、有道德、的、人、一个、脱离了低级趣味、的、人、一个、有益于人民、的、人”。

[0051] 需要说明的是,自然语言技术是以一种智能与高效的方式,对文本数据进行系统化分析、理解与信息提取的过程。通过使用NLP以及它的组件,我们可以管理非常大块的文本数据,或者执行大量的自动化任务,并且解决各式各样的问题,如自动摘要,机器翻译,命名实体识别,关系提取,情感分析,语音识别,以及主题分割等等。

[0052] 由于本申请所解决的技术问题是如何准确识别文本信息中的关键字,因此,通过自然语言技术对文本信息进行分词的技术原理在此不做赘述。

[0053] S2:计算所述词汇集中各词汇的词频,拆分所述词汇集得到子词汇集以及所述子词汇集中各词汇之间的关联关系,根据所述子词汇集中各词汇的词频,及所述各词汇的关联关系得到具有特征值的总词汇表,其中,所述特征值反映了词汇在文本信息中的关键程度。

[0054] 在一个优选的实施例中,请参阅图2,计算所述词汇集中各词汇的词频,拆分所述词汇集得到子词汇集以及所述子词汇集中各词汇之间的关联关系的步骤,包括:

[0055] S21:计算所述词汇集中各词汇在文本信息中出现的次数,以获得所述词汇的词频;

[0056] S22:按照预设的分隔符拆分所述词汇集得到至少一个子词汇集,根据所述子词汇集中各词汇之间的关联关系制定关联特征;其中,所述词频反映了词汇在词汇集中出现的频率,所述关联特征是以特征向量的形式表达了子词汇集中任一词汇与其他词汇之间的关联关系。

[0057] 在一个优选的实施例中,请参阅图3,根据所述子词汇集中各词汇的词频,及所述各词汇的关联关系得到具有特征值的总词汇表的步骤,包括:

[0058] S23:将所述子词汇集的关联特征及其中各词汇的词频分别作为向量并对其进行运算,得到具有词汇及其特征值的子关键特征;

[0059] S24:根据所述子关键特征的特征值排列所述子词汇集中的词汇得到对应有特征值的子词汇列表,汇总所述子词汇列表形成总词汇表。

[0060] 在示例性的实施例中,为获得词汇集中各词汇的词频以及所述词汇集中任一词汇被其他词汇所依赖的程度,以从两个维度评价词汇的关键程度,本步骤通过TF-IDF算法计算所述词汇集中各词汇的词频,以得到各词汇在文本信息中出现的频率;按照预设的分隔符拆分所述词汇集得到至少一个子词汇集,其中,所述分隔符可为标点符号、空格、换行符

等;通过PageRank算法根据所述子词汇集中各词汇之间的关联关系制定关联特征,该关联特征反映了子词汇集中任一词汇被其他词汇所依赖的程度,并以特征向量的形式表达了子词汇集中任一词汇与其他词汇之间的关联关系。

[0061] 由于词频从词汇出现次数的维度评价了词汇的重要性,而关联特征从任一词汇被其他词汇依赖的维度评价了词汇的重要性,因此,为从词频和被依赖的程度两个维度综合评价子词汇集中各词汇的关键程度,以更加准确的提取能够反映子文本信息中核心含义的关键字,本步骤通过计算所述子词汇集的关联特征及其中各词汇的词频,降低了如:“最”,“的”,“一个”等介词或量词对关键字提取操作的干扰,提高了对虽然出现频次不高,但多次被其他词(如:形容词、介词、动词、副词)所修饰的词汇进行识别的概率,而该词汇在文本信息中往往处于能够反映文本信息核心含义的关键地位,进而实现了提高获得能够反映文本信息核心含义的关键字的准确度。

[0062] 根据所述子关键特征的特征值排列所述子词汇集中的词汇得到子词汇列表,实现了对子文本信息中各词汇的关键程度进行由高到低的评价,以便于服务器或用户根据该子词汇列表准确把握子文本信息的含义。

[0063] 需要说明的是,TF-IDF算法是一种统计方法,用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。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。其中,TF是词频(Term Frequency),IDF是逆文本频率指数(Inverse Document Frequency)。PageRank算法,又称网页排名、谷歌左侧排名,是一种由搜索引擎根据网页之间相互的超链接计算的技术,其是根据网页之间通过超链接形成的依赖关系,计算出各网页的PR值(即:权重值),得到所述网页的重要程度的计算机算法,本申请虽然将pagerank算法作为实现手段对词汇了计算,但是,本申请所解决的问题是如何得到子词汇集中各词汇之间的关联关系,并根据该关联关系得到关联特征,因此pagerank算法的技术原理在本申请中不做赘述。

[0064] 优选的,汇总所述子词汇列表形成总词汇表之后,还包括:

[0065] 将所述总词汇表上传至区块链中。

[0066] 需要说明的是,基于总词汇表得到对应的摘要信息,具体来说,摘要信息由总词汇表进行散列处理得到,比如利用sha256s算法处理得到。将摘要信息上传至区块链可保证其安全性和对用户的公正透明性。用户设备可以从区块链中下载得该摘要信息,以便查证总词汇表是否被篡改。本示例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。

[0067] 在一个优选的实施例中,请参阅图4,计算所述词汇集中各词汇在文本信息中出现的次数,以获得所述词汇的词频的步骤,包括:

[0068] S211:计算所述词汇集中所有词汇的总数,及对所述词汇集中的词汇进行去重得到词汇表。

[0069] S212:计算所述词汇表中各词汇在所述词汇集中出现的次数,将所述词汇的次数与所述总数相除得到所述词汇的词频。

[0070] 示例性地，基于上述举例，所述词汇集：“一个、高尚、的、人、一个、纯粹、的、人、一个、有道德、的、人、一个、脱离了低级趣味、的、人、一个、有益于人民、的、人”的词汇的总数为20；

[0071] 对所述词汇集中的词汇进行去重，得到词汇表：

[0072]

词汇	一个	的	人	高尚	纯粹	有道德	脱离了低级趣味	有益于人民

[0073] 计算词汇表中各词汇在词汇集中出现的次数：

[0074]

词汇	一个	的	人	高尚	纯粹	有道德	脱离了低级趣味	有益于人民
次数	5	5	5	1	1	1	1	1

[0075] 将所述词汇的次数与所述总数相除得到所述词汇的词频，如下表所示：

[0076]

词汇	一个	的	人	高尚	纯粹	有道德	脱离了低级趣味	有益于人民
词频	0.25	0.25	0.25	0.05	0.05	0.05	0.05	0.05

[0077] 在一个优选的实施例中，请参阅图5，按照预设的分隔符拆分所述词汇集得到至少一个子词汇集，根据所述子词汇集中各词汇之间的关联关系制定关联特征的步骤，包括：

[0078] S221：以标点符号为分隔符划分所述文本信息形成子文本信息，汇总所述词汇集中与子文本信息对应的词汇得到所述子文本信息的子词汇集。

[0079] S222：识别子词汇集中在其子文本信息上处于相邻位置的两个词汇，并认定所述两个词汇之间具有关联关系。

[0080] S223：根据所述子词汇集中各词汇之间的关联关系，制定能够表达子词汇集中任一词汇与其他词汇之间关联关系特征向量，以得到所述子词汇集的关联特征。

[0081] 示例性地，以标点符号为分隔符划分所述文本信息形成子文本信息，所述子文本信息包括：

[0082] “一个高尚的人”、“一个纯粹的人”、“一个有道德的人”、“一个脱离了低级趣味的人”、“一个有益于人民的人”

[0083] 接下来以子文本信息“一个高尚的人”举例，汇总所述词汇集中与子文本信息对应的词汇得到所述子文本信息的子词汇集，包括：“一个、高尚、的、人”。

[0084] 识别子词汇集中在其子文本信息上处于相邻位置的两个词汇，例如：“一个”与“高尚”相邻，“高尚”与“的”相邻，“的”与“人”相邻，因此，认定“一个”与“高尚”具有关联关系，“高尚”与“的”具有关联关系，“的”与“人”具有关联关系。

[0085] 根据上述关联关系，制定能够表达子词汇集中任一词汇与其他词汇之间关联关系特征向量。

[0086] 于本实施例中，PageRank算法总的来说就是预先给每个网页一个PR值(下面用PR值指代PageRank值)，由于PR值物理意义上为一个网页被访问概率，所以一般是1/N，其中N为网页总数。

[0087] 那么将所述pagerank算法应用到子词汇集中，用于评判其中各词汇被其他词汇依赖的程度，则将相互之间具有关联关系的元素值设为1，将相互之间不具有关联关系的元素值设为0，得到如下表所示的特征向量：

[0088]

	一个	高尚	的	人

一个	0	1	0	0
高尚	1	0	1	0
的	0	1	0	1
人	0	0	1	0

[0089] 进一步地，利用pagerank算法中的PR值原理，即：物理意义上为一个网页被访问概率，PR＝1/N，其中N为与所述网页产生连接的网页数量；将得到子词汇集中，任一词汇被其他词汇依赖的概率T，T＝1/M，其中M为与所述词汇产生关联关系的词汇的数量。

[0090] 如此将获得如下特征向量，即所述关联特征：

[0091]

	一个	高尚	的	人
一个	0	0.5	0	0
高尚	1	0	0.5	0
的	0	0.5	0	1
人	0	0	0.5	0

[0092] 在一个优选的实施例中，请参阅图6，将所述子词汇集的关联特征及其中各词汇的词频分别作为向量并对其进行运算，得到具有词汇及其特征值的子关键特征的步骤，包括：

[0093] S231：汇总子词汇集中各词汇及其词频得到词频向量。

[0094] S232：将所述子词汇集的关联特征与所述词频向量相乘得到具有特征值的得到子关键特征，其中，所述特征值为子关键特征的元素值，所述子词汇集中的词汇与所述元素值一一对应。

[0095] 示例性地，基于上述举例，子文本信息“一个高尚的人”的词频如下表所示：

[0096]

词汇	一个	高尚	的	人
词频	0.25	0.05	0.25	0.25

[0097] 汇总所述子词汇集中各词汇的词频得到词频向量如下所示：

[0098]

一个	0.25
高尚	0.05
的	0.25
人	0.25

[0099] 该文本信息的关联特征如下表所示：

[0100]

	一个	高尚	的	人
一个	0	0.5	0	0
高尚	1	0	0.5	0
的	0	0.5	0	1
人	0	0	0.5	0

[0101] 那么根据矩阵算法将所述关联特征与所述词频向量相乘，则代表着结合各词汇被依赖的程度及其出现的次数，综合评价词汇的重要程度，最终得到的子关键特征如下表所示：

[0102]

一个	0.025
高尚	0.375

的	0.275
人	0.125

[0103] 根据所述子关键特征的特征值排列所述子词汇集中的词汇得到如下的子词汇列表：

[0104]

词汇	特征值
高尚	0.375
的	0.275
人	0.125
一个	0.025

[0105] S3：按照特征值排列所述总词汇表中的词汇，将特征值超过预设的特征阈值的词汇设为关键字。

[0106] 为从词频和被依赖的程度两个维度综合评价词汇集中各词汇的关键程度，以更加准确的提取能够反映文本信息中核心含义的关键字，本步骤通过汇总所述子词汇列表形成总词汇表，并按照特征值排列所述总词汇表中的词汇。

[0107] 示例性地，按照上述方法将获得所述文本信息所对应的子词汇列表，如下：

[0108]

词汇	特征值
高尚	0.375
的	0.275
人	0.125
一个	0.025

[0109]

词汇	特征值
纯粹	0.375
的	0.275
人	0.125
一个	0.025

[0110]

词汇	特征值
有道德	0.375
的	0.275
人	0.125
一个	0.025

[0111]

词汇	特征值
脱离了低级趣味	0.375
的	0.275
人	0.125
一个	0.025

[0112]

词汇	特征值

有益于人民	0.375
的	0.275
人	0.125
一个	0.025

[0113] 汇总上述子词汇列表并按照特征值排列其中的词汇得到总词汇表如下：

[0114]

词汇	特征值
高尚	0.375
纯粹	0.375

[0115]

有道德	0.375
脱离了低级趣味	0.375
有益于人民	0.375
的	0.275
人	0.125
一个	0.025

[0116] 假设特征阈值为0.3，那么将得到：“高尚、纯粹、有道德、脱离了低级趣味、有益于人民”的关键字，而非将“一个、的、人”这些词频较高的词汇作为关键字，提高了关键字识别的准确度。

[0117] 本申请可应用于智慧政务场景中，从而推动智慧城市的建设。

[0118] 实施例二：

[0119] 请参阅图7，本实施例的一种文本关键字识别装置1，包括：

[0120] 输入分词模块11，用于获取文本信息，对所述文本信息进行分词获得词汇集；

[0121] 词频关联模块12，用于计算所述词汇集中各词汇的词频，拆分所述词汇集得到子词汇集以及所述子词汇集中各词汇之间的关联关系，根据所述子词汇集中各词汇的词频，及所述各词汇的关联关系得到具有特征值的总词汇表，其中，所述特征值反映了词汇在文本信息中的关键程度；

[0122] 关键字识别模块13，用于按照特征值排列所述总词汇表中的词汇，将特征值超过预设的特征阈值的词汇设为关键字。

[0123] 本技术方案应用于人工智能的智能决策技术领域，通过对文本信息进行分词获得词汇集；计算所述词汇集中各词汇的词频，拆分所述词汇集得到子词汇集以及所述子词汇集中各词汇之间的关联关系，根据所述子词汇集中各词汇的词频，及所述各词汇的关联关系得到具有特征值的总词汇表，以构建文本信息的检测模型，按照特征值排列所述总词汇表中的词汇，将特征值超过预设的特征阈值的词汇设为关键字，以实现关键字匹配的技术效果。

[0124] 实施例三：

[0125] 为实现上述目的，本发明还提供一种计算机设备2，实施例三的文本关键字识别装置1的组成部分可分散于不同的计算机设备中，计算机设备2可以是执行程序的智能手机、平板电脑、笔记本电脑、台式计算机、机架式服务器、刀片式服务器、塔式服务器或机柜式服务器(包括独立的服务器，或者多个应用服务器所组成的服务器集群)等。本实施例的计算

机设备至少包括但不限于:可通过系统总线相互通信连接的存储器21、处理器21,如图8所示。需要指出的是,图8仅示出了具有组件-的计算机设备,但是应理解的是,并不要求实施所有示出的组件,可以替代的实施更多或者更少的组件。

[0126] 本实施例中,存储器21(即可读存储介质)包括闪存、硬盘、多媒体卡、卡型存储器(例如,SD或DX存储器等)、随机访问存储器(RAM)、静态随机访问存储器(SRAM)、只读存储器(ROM)、电可擦除可编程只读存储器(EEPROM)、可编程只读存储器(PROM)、磁性存储器、磁盘、光盘等。在一些实施例中,存储器21可以是计算机设备的内部存储单元,例如该计算机设备的硬盘或内存。在另一些实施例中,存储器21也可以是计算机设备的外部存储设备,例如该计算机设备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。当然,存储器21还可以既包括计算机设备的内部存储单元也包括其外部存储设备。本实施例中,存储器21通常用于存储安装于计算机设备的操作系统和各类应用软件,例如实施例三的文本关键字识别装置的程序代码等。此外,存储器21还可以用于暂时地存储已经输出或者将要输出的各类数据。

[0127] 处理器21在一些实施例中可以是中央处理器(Central Processing Unit,CPU)、控制器、微控制器、微处理器、或其他数据处理芯片。该处理器21通常用于控制计算机设备的总体操作。本实施例中,处理器21用于运行存储器21中存储的程序代码或者处理数据,例如运行文本关键字识别装置,以实现实施例一的文本关键字识别方法。

[0128] 实施例四:

[0129] 为实现上述目的,本发明还提供一种计算机可读存储介质,如闪存、硬盘、多媒体卡、卡型存储器(例如,SD或DX存储器等)、随机访问存储器(RAM)、静态随机访问存储器(SRAM)、只读存储器(ROM)、电可擦除可编程只读存储器(EEPROM)、可编程只读存储器(PROM)、磁性存储器、磁盘、光盘、服务器、App应用商城等等,其上存储有计算机程序,程序被处理器21执行时实现相应功能。本实施例的计算机可读存储介质用于存储文本关键字识别装置,被处理器21执行时实现实施例一的文本关键字识别方法。

[0130] 上述本发明实施例序号仅仅为了描述,不代表实施例的优劣。

[0131] 通过以上的实施方式的描述,本领域的技术人员可以清楚地了解到上述实施例方法可借助软件加必需的通用硬件平台的方式来实现,当然也可以通过硬件,但很多情况下前者是更佳的实施方式。

[0132] 以上仅为本发明的优选实施例,并非因此限制本发明的专利范围,凡是利用本发明说明书及附图内容所作的等效结构或等效流程变换,或直接或间接运用在其他相关的技术领域,均同理包括在本发明的专利保护范围内。

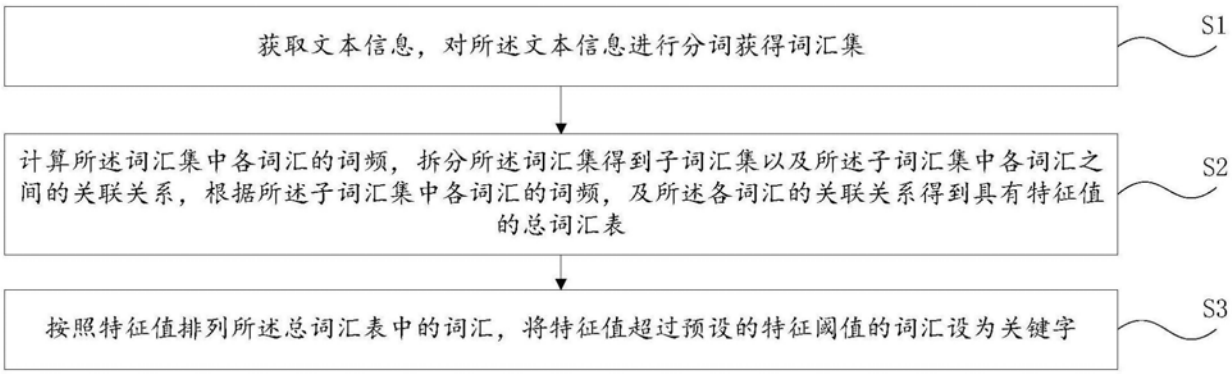

获取文本信息，对所述文本信息进行分词获得词汇集
S1
计算所述词汇集中各词汇的词频，拆分所述词汇集得到子词汇集以及所述子词汇集中各词汇之间的关联关系，根据所述子词汇集中各词汇的词频，及所述各词汇的关联关系得到具有特征值的总词汇表
S2
按照特征值排列所述总词汇表中的词汇，将特征值超过预设的特征阈值的词汇设为关键字
S3

图1

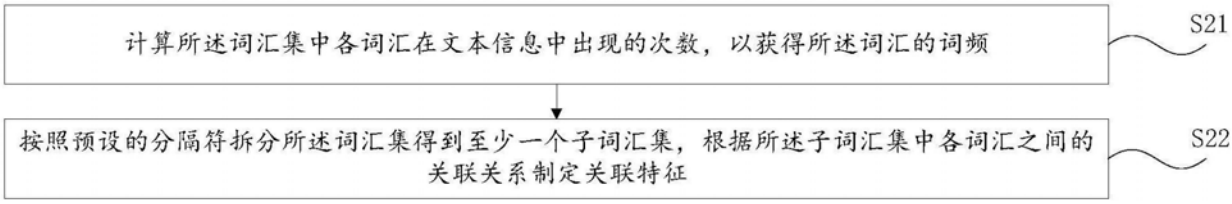

计算所述词汇集中各词汇在文本信息中出现的次数，以获得所述词汇的词频
S21
按照预设的分隔符拆分所述词汇集得到至少一个子词汇集，根据所述子词汇集中各词汇之间的关联关系制定关联特征
S22

图2

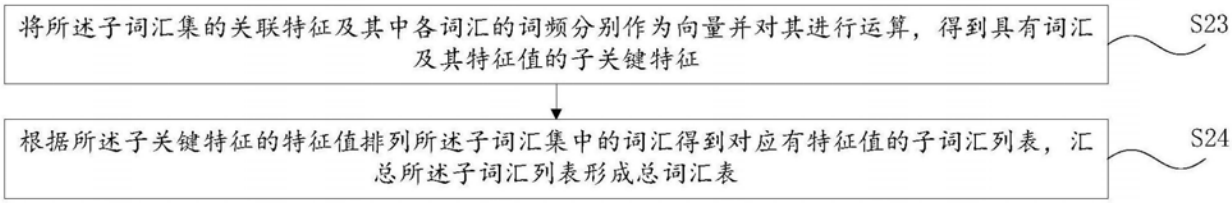

将所述子词汇集的关联特征及其中各词汇的词频分别作为向量并对其进行运算，得到具有词汇及其特征值的子关键特征
S23
根据所述子关键特征的特征值排列所述子词汇集中的词汇得到对应有特征值的子词汇列表，汇总所述子词汇列表形成总词汇表
S24

图3


计算所述词汇集中所有词汇的总数，及对所述词汇集中的词汇进行去重得到词汇表
S211
计算所述词汇表中各词汇在所述词汇集中出现的次数，将所述词汇的次数与所述总数相除得到所述词汇的词频
S212

图4

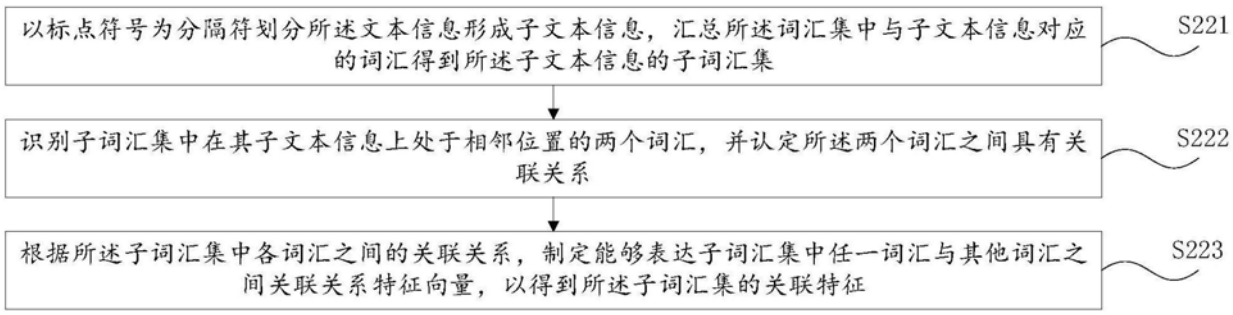

以标点符号为分隔符划分所述文本信息形成子文本信息，汇总所述词汇集中与子文本信息对应的词汇得到所述子文本信息的子词汇集
S221
识别子词汇集中在其子文本信息上处于相邻位置的两个词汇，并认定所述两个词汇之间具有关联关系
S222
根据所述子词汇集中各词汇之间的关联关系，制定能够表达子词汇集中任一词汇与其他词汇之间关联关系特征向量，以得到所述子词汇集的关联特征
S223

图5

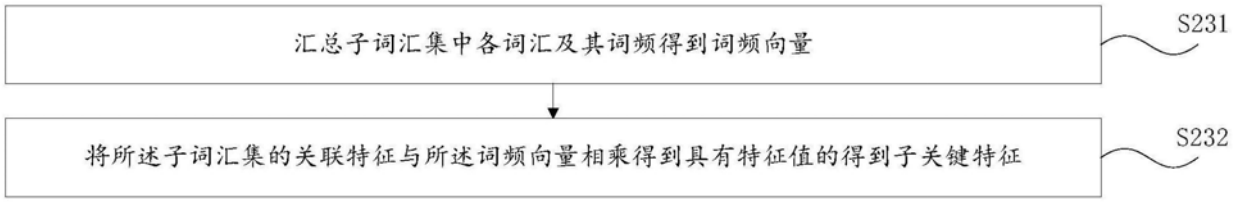

汇总子词汇集中各词汇及其词频得到词频向量
S231
将所述子词汇集的关联特征与所述词频向量相乘得到具有特征值的得到子关键特征
S232

图6

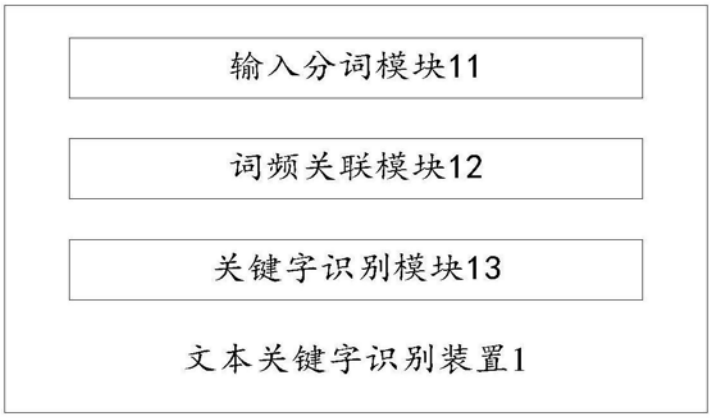

输入分词模块11
词频关联模块12
关键字识别模块13
文本关键字识别装置1

图7

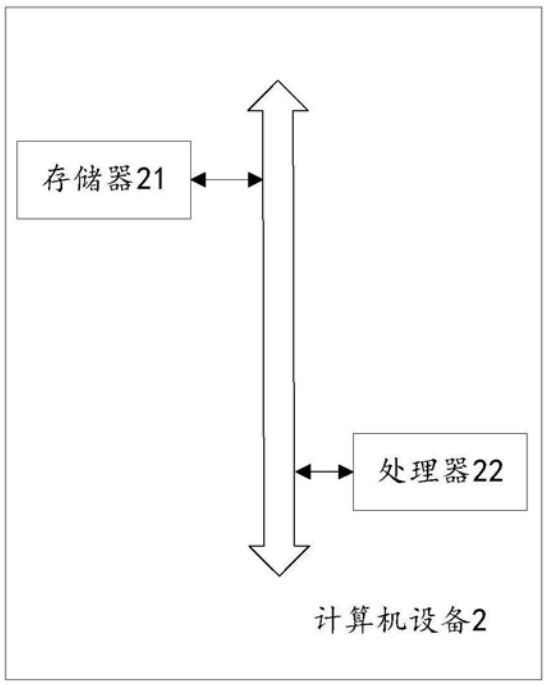


图8