(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 115757760 A
(43)申请公布日 2023.03.07

(21)申请号 202111031997.3

(22)申请日 2021.09.03

(71)申请人 北京中关村科金技术有限公司
地址 100080 北京市海淀区上地四街一号院5号楼一层130

(72)发明人 刘光辉 周健

(74)专利代理机构 北京万思博知识产权代理有限公司 11694
专利代理师 柴国伟

(51)Int.Cl.
G06F 16/34(2019.01)
G06F 40/216(2020.01)
G06F 40/289(2020.01)

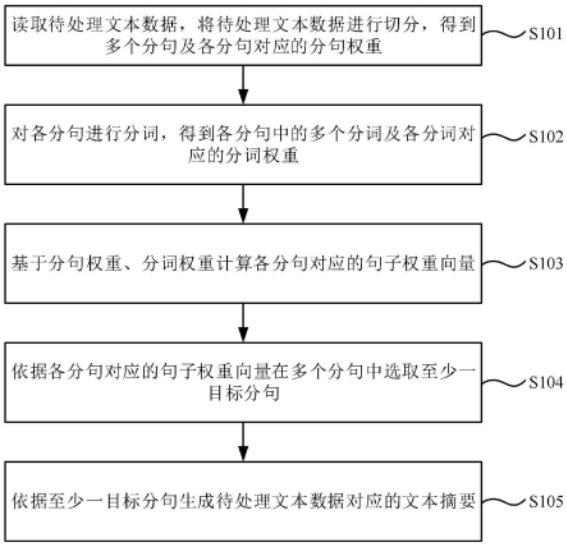
权利要求书2页 说明书9页 附图4页

(54)发明名称

文本摘要提取方法及系统、计算设备、存储介质

(57)摘要

本申请提供了一种文本摘要提取方法及系统、计算设备、存储介质，在本申请提供的方法中，先读取待处理文本数据进行切分，得到多个分句及各分句对应的分句权重；再对各分句进行分词，得到各分句中的多个分词及各分词对应的分词权重；然后基于分句权重、分词权重计算各分句对应的句子权重向量；最后依据各分句对应的句子权重向量在多个分句中选取至少一目标分句，并生成待处理文本数据对应的文本摘要。基于本申请提供的方案，基于关键词词频与位置和句子位置自动生成的文本摘要可以有效改过待处理文本数据的基本含义，并且还能够解决基于TF-IDF文本摘要未考虑词频位置信息和TextRank未考虑句子位置信息导致信息提取不足的问题。



CN 115757760 A

1.一种文本摘要提取方法,包括:

读取待处理文本数据,将所述待处理文本数据进行切分,得到多个分句及各所述分句对应的分句权重;

对各所述分句进行分词,得到各所述分句中的多个分词及各分词对应的分词权重;

基于所述分句权重、所述分词权重计算各所述分句对应的句子权重向量;

依据各所述分句对应的句子权重向量在所述多个分句中选取至少一目标分句;

依据所述至少一目标分句生成所述待处理文本数据对应的文本摘要。

2.根据权利要求1所述的方法,其特征在于,所述将所述待处理文本数据进行切分,得到多个分句及各所述分句对应的分句权重,包括:

按照字符分割方式对所述待处理文本数据,进行切分,得到多个分句;其中,所述字符包括标点符号和/或换行符;

将所述待处理文本数据划分为连续的多个段落组成部分,并对各所述段落组成部分分别赋予不同的权重值;

依据各所述分句在所述多个段落组成部分中所属的段落组成部分对应的权重值,确定各所述分句的分句权重。

3.根据权利要求2所述的方法,其特征在于,所述将所述待处理文本数据划分为连续的多个段落组成部分,包括:

按照所述待处理文本数据的文本长度结构,将所述待处理文本数据划分为连续的首部段落组成部分、中部段落组成部分和尾部段落组成部分,作为所述多个段落组成部分。

4.根据权利要求1所述的方法,其特征在于,所述对各所述分句进行分词,得到各所述分句中的多个分词及各分词对应的分词权重,包括:

对于任一分句,对所述分句进行分词处理,得到所述分句对应的多个词语;

去除所述分句中所述多个词语中的停用词后得到词语,作为所述分句中的多个分词;

依次判断各所述分词对应的词语属性,并基于所述词语属性为各所述分词分配对应的分词权重;所述词语属性包括关键词和普通词。

5.根据权利要求1所述的方法,其特征在于,所述基于所述分句权重、所述分词权重计算各所述分句对应的句子权重向量,包括:

利用Glove词嵌入生成各所述分词对应的分词向量,并将每个所述分句中的各分词对应的分词向量相加得到各分句对应的分句向量;

将同一分句对应的所述分句向量、所述分句权重和所述分词权重相乘,得到各所述分句对应的句子权重向量。

6.根据权利要求1所述的方法,其特征在于,所述依据各所述分句对应的句子权重向量在所述多个分句中选取至少一目标分句,包括:

利用余弦相似度对各所述句子对应的句子权重向量进行计算得到相似度矩阵;

将所述相似度矩阵转化为图形,所述图形中的节点表示句子,边表示句子之间的相似度得分;

利用PageRank网页排名算法对各分句进行排序,根据排序结果在所述多个分句中选取至少一目标分句。

7.根据权利要求6所述的方法,其特征在于,所述依据所述至少一目标分句生成所述待

处理文本数据对应的文本摘要，包括：

依据所述目标分句在所述待处理文本数据中的分布顺序进行拼接，生成所述待处理文本数据对应的文本摘要。

8.一种文本摘要提取系统，包括：

分句切分模块，其配置成读取待处理文本数据，将所述待处理文本数据进行切分，得到多个分句及各所述分句对应的分句权重；

分词获取模块，其配置成对各所述分句进行分词，得到各所述分句中的多个分词及各分词对应的分词权重；

句子权重向量计算模块，其配置成基于所述分句权重、所述分词权重计算各所述分句对应的句子权重向量；

目标分句选取模块，其配置成依据各所述分句对应的句子权重向量在所述多个分句中选取至少一目标分句；

文本摘要生成模块，其配置成依据所述至少一目标分句生成所述待处理文本数据对应的文本摘要。

9.一种计算设备，包括存储器、处理器和存储在所述存储器内并能由所述处理器运行的计算机程序，其中，所述处理器执行所述计算机程序时实现如权利要求1-7中任一项所述的方法。

10.一种计算机可读存储介质，优选为非易失性可读存储介质，其内存储有计算机程序，所述计算机程序在由处理器执行时实现如权利要求1-7中任一项所述的方法。

文本摘要提取方法及系统、计算设备、存储介质

技术领域

[0001] 本申请涉及数据处理技术领域，特别是涉及一种文本摘要提取方法及系统、计算设备、存储介质。

背景技术

[0002] 文本摘要技术是人工智能领域的重要技术。摘要是一段简短的文本，它准确地捕获和传达人们想要摘要的文档中包含的最重要和最相关的信息。对于人类来说，阅读一段长文本，并提炼其核心摘要内容，是一种天生的能力。但对于计算机来说，却代表了人工智能领域最具挑战性技术的进展和突破。

[0003] 自动文本摘要早前就引起了人们的注意。一篇题为《文学文摘的自动创作》，利用词频和短语频等特征，从文本中提取重要句子进行总结。另一项重要的研究利用线索词的出现、出现在文章标题中的词以及句子的位置等方法，提取出有意义的句子进行文本总结。

[0004] 常见的文本摘要技术包括基于词频逆文档频率(英文全称:Term Frequency-Inverse Document Frequency，英文简称TF-IDF)的抽取式摘要和基于TextRank(一种基于图的用于关键词抽取和文本摘要的排序算法)的抽取式摘要。基于TF-IDF的抽取式摘要依据“词频”作为衡量一个词的重要性，但无法体现句子的位置信息；而基于TextRank的抽取式摘要把GloVe词嵌入作为词的向量表示，这种方法得到的词向量也是基于词频并且受限与窗口大小，没有考虑逆文档频率，也同样无法体现句子的位置信息，出现位置不同的句子都被认为同样的重要性，显然是不正确的。

发明内容

[0005] 本申请的目的在于克服上述问题或者至少部分地解决或缓减解决上述问题。

[0006] 根据本申请的一个方面，提供了一种文本摘要提取方法，包括：

[0007] 读取待处理文本数据，将所述待处理文本数据进行切分，得到多个分句及各所述分句对应的分句权重；

[0008] 对各所述分句进行分词，得到各所述分句中的多个分词及各分词对应的分词权重；

[0009] 基于所述分句权重、所述分词权重计算各所述分句对应的句子权重向量；

[0010] 依据各所述分句对应的句子权重向量在所述多个分句中选取至少一目标分句；

[0011] 依据所述至少一目标分句生成所述待处理文本数据对应的文本摘要。

[0012] 可选地，所述将所述待处理文本数据进行切分，得到多个分句及各所述分句对应的分句权重，包括：

[0013] 按照字符分割方式对所述待处理文本数据，进行切分，得到多个分句；其中，所述字符包括标点符号和/或换行符；

[0014] 将所述待处理文本数据划分为连续的多个段落组成部分，并对各所述段落组成部分分别赋予不同的权重值；

[0015] 依据各所述分句在所述多个段落组成部分中所属的段落组成部分对应的权重值，确定各所述分句的分句权重。

[0016] 可选地，所述将所述待处理文本数据划分为连续的多个段落组成部分，包括：

[0017] 按照所述待处理文本数据的文本长度结构，将所述待处理文本数据划分为连续的首部段落组成部分、中部段落组成部分和尾部段落组成部分，作为所述多个段落组成部分。

[0018] 可选地，所述对各所述分句进行分词，得到各所述分句中的多个分词及各分词对应的分词权重，包括：

[0019] 对于任一分句，对所述分句进行分词处理，得到所述分句对应的多个词语；

[0020] 去除所述分句中所述多个词语中的停用词后得到词语，作为所述分句中的多个分词；

[0021] 依次判断各所述分词对应的词语属性，并基于所述词语属性为各所述分词分配对应的分词权重；所述词语属性包括关键词和普通词。

[0022] 可选地，所述基于所述分句权重、所述分词权重计算各所述分句对应的句子权重向量，包括：

[0023] 利用Glove词嵌入生成各所述分词对应的分词向量，并将每个所述分句中的各分词对应的分词向量相加得到各分句对应的分句向量；

[0024] 将同一分句对应的所述分句向量、所述分句权重和所述分词权重相乘，得到各所述分句对应的句子权重向量。

[0025] 可选地，所述依据各所述分句对应的句子权重向量在所述多个分句中选取至少一目标分句，包括：

[0026] 利用余弦相似度对各所述句子对应的句子权重向量进行计算得到相似度矩阵；

[0027] 将所述相似度矩阵转化为图形，所述图形中的节点表示句子，边表示句子之间的相似度得分；

[0028] 利用PageRank网页排名算法对各分句进行排序，根据排序结果在所述多个分句中选取至少一目标分句。

[0029] 可选地，所述依据所述至少一目标分句生成所述待处理文本数据对应的文本摘要，包括：

[0030] 依据所述目标分句在所述待处理文本数据中的分布顺序进行拼接，生成所述待处理文本数据对应的文本摘要。

[0031] 根据本申请的另一个方面，提供了一种文本摘要提取系统，包括：

[0032] 分句切分模块，其配置成读取待处理文本数据，将所述待处理文本数据进行切分，得到多个分句及各所述分句对应的分句权重；

[0033] 分词获取模块，其配置成对各所述分句进行分词，得到各所述分句中的多个分词及各分词对应的分词权重；

[0034] 句子权重向量计算模块，其配置成基于所述分句权重、所述分词权重计算各所述分句对应的句子权重向量；

[0035] 目标分句选取模块，其配置成依据各所述分句对应的句子权重向量在所述多个分句中选取至少一目标分句；

[0036] 文本摘要生成模块，其配置成依据所述至少一目标分句生成所述待处理文本数据

对应的文本摘要。

[0037] 可选地，所述分句切分模块，其还可以配置成：

[0038] 按照字符分割方式对所述待处理文本数据，进行切分，得到多个分句；其中，所述字符包括标点符号和/或换行符；

[0039] 将所述待处理文本数据划分为连续的多个段落组成部分，并对各所述段落组成部分分别赋予不同的权重值；

[0040] 依据各所述分句在所述多个段落组成部分中所属的段落组成部分对应的权重值，确定各所述分句的分句权重。

[0041] 可选地，所述分句切分模块，其还可以配置成：

[0042] 按照所述待处理文本数据的文本长度结构，将所述待处理文本数据划分为连续的首部段落组成部分、中部段落组成部分和尾部段落组成部分，作为所述多个段落组成部分。

[0043] 可选地，所述分词获取模块，其还可以配置成：

[0044] 对于任一分句，对所述分句进行分词处理，得到所述分句对应的多个词语；

[0045] 去除所述分句中所述多个词语中的停用词后得到词语，作为所述分句中的多个分词；

[0046] 依次判断各所述分词对应的词语属性，并基于所述词语属性为各所述分词分配对应的分词权重；所述词语属性包括关键词和普通词。

[0047] 可选地，所述句子权重向量计算模块，其还可以配置成：

[0048] 利用Glove词嵌入生成各所述分词对应的分词向量，并将每个所述分句中的各分词对应的分词向量相加得到各分句对应的分句向量；

[0049] 将同一分句对应的所述分句向量、所述分句权重和所述分词权重相乘，得到各所述分句对应的句子权重向量。

[0050] 可选地，所述目标分句选取模块，其还可以配置成：

[0051] 利用余弦相似度对各所述句子对应的句子权重向量进行计算得到相似度矩阵；

[0052] 将所述相似度矩阵转化为图形，所述图形中的节点表示句子，边表示句子之间的相似度得分；

[0053] 利用PageRank网页排名算法对各分句进行排序，根据排序结果在所述多个分句中选取至少一目标分句。

[0054] 可选地，所述文本摘要生成模块，其还可以配置成：

[0055] 依据所述目标分句在所述待处理文本数据中的分布顺序进行拼接，生成所述待处理文本数据对应的文本摘要。

[0056] 根据本发明的另一方面，还提供了一种计算设备，包括存储器、处理器和存储在所述存储器内并能由所述处理器运行的计算机程序，其中，所述处理器执行所述计算机程序时实现如上述任一项所述的文本摘要提取方法。

[0057] 根据本发明的另一方面，还提供了一种计算机可读存储介质，优选为非易失性可读存储介质，其内存储有计算机程序，所述计算机程序在由处理器执行时实现如上述任一项所述的文本摘要提取方法。

[0058] 本申请提供了一种文本摘要提取方法及系统、计算设备、存储介质，在本申请提供的方法中，先读取待处理文本数据进行切分，得到多个分句及各分句对应的分句权重；再对

各分句进行分词,得到各分句中的多个分词及各分词对应的分词权重;然后基于分句权重、分词权重计算各分句对应的句子权重向量;最后依据各分句对应的句子权重向量在多个分句中选取至少一目标分句,并生成待处理文本数据对应的文本摘要。基于本申请提供的文本摘要提取方法及系统,基于关键词词频与位置和句子位置进行自动文本摘要,解决基于TF-IDF文本摘要未考虑词频位置信息和TextRank未考虑句子位置信息导致信息提取不足的问题,同时结合分句权重、分词权重选取待处理文本数据中的有代表性的目标分句,进而准确生成待处理文本数据对应的文本摘要。

[0059] 根据下文结合附图对本申请的具体实施例的详细描述,本领域技术人员将会更加明了本申请的上述以及其他目的、优点和特征。

附图说明

[0060] 后文将参照附图以示例性而非限制性的方式详细描述本申请的一些具体实施例。附图中相同的附图标记标示了相同或类似的部件或部分。本领域技术人员应该理解,这些附图未必是按比例绘制的。附图中:

[0061] 图1是根据本申请实施例的文本摘要提取方法流程示意图;

[0062] 图2是根据本申请实施例的文本摘要提取整体流程框图;

[0063] 图3是根据本申请实施例的文本摘要提取系统结构示意图;

[0064] 图4是根据本申请实施例的计算设备结构示意图;

[0065] 图5是根据本申请实施例的计算机可读存储介质示意图。

具体实施方式

[0066] 根据下文结合附图对本申请的具体实施例的详细描述,本领域技术人员将会更加明了本申请的上述以及其他目的、优点和特征。

[0067] 图1是根据本申请实施例的文本摘要提取方法流程示意图。参见图1所知,本申请实施例提供的文本摘要提取方法至少可以包括以下步骤S101～S105。

[0068] 步骤S101:读取待处理文本数据,将待处理文本数据进行切分,得到多个分句及各分句对应的分句权重;

[0069] 步骤S102:对各分句进行分词,得到各分句中的多个分词及各分词对应的分词权重;

[0070] 步骤S103:基于分句权重、分词权重计算各分句对应的句子权重向量;

[0071] 步骤S104:依据各分句对应的句子权重向量在多个分句中选取至少一目标分句;

[0072] 步骤S105:依据至少一目标分句生成待处理文本数据对应的文本摘要。

[0073] 本申请提供了一种文本摘要提取方法,在本申请提供的方法中,先读取待处理文本数据进行切分,得到多个分句及各分句对应的分句权重;再对各分句进行分词,得到各分句中的多个分词及各分词对应的分词权重;然后基于分句权重、分词权重计算各分句对应的句子权重向量;最后依据各分句对应的句子权重向量在多个分句中选取至少一目标分句,并生成待处理文本数据对应的文本摘要。基于本申请提供的文本摘要提取方法,通过分句位置与分词权重的结合,充分考虑分句位置和分词权重和分词距离对文本摘要的影响,得到最终的权重分句向量,以选取待处理文本数据中的有代表性的目标分句,进而准确生

成待处理文本数据对应的文本摘要。本实施例提供的方法有效解决了通过TF-IDF方法未考虑分词顺序和分词距离和句子位置的缺陷问题，同时也解决了TextRank算法不能考虑句子位置的缺陷问题。

[0074] 下面分别对上述实施例提及的文本摘要提取方法进行详细说明。

[0075] 首先，如步骤S101所述，读取待处理文本数据，将待处理文本数据进行切分，得到多个分句及各所述分句对应的分句权重。

[0076] 待处理文本数据即需要提取摘要的文本数据，其可以是一片文章，也可以是一篇时评，待处理文本数据的字数本实施例对此不做限定。

[0077] 可选地，上述步骤S101得到多个分句及各所述分句对应的分句权重可以进一步包括：按照字符分割方式进行切分，得到多个分句；其中，字符包括标点符号和/或换行符。在实际应用中，一篇文章是由句子组成的，短句之间可能被任意标点符号隔开，段落组成部分之间被换行符隔开，而若想从文章中提取句子，则需要按照标点符号、换行符等字符方式切分句子，进而得到多个分句，以对各个分句赋予不同的权重。

[0078] 步骤S101-2，将待处理文本数据划分为连续的多个段落组成部分，并对各段落组成部分分别赋予不同的权重值。具体地，可以按照待处理文本数据的文本长度结构，将待处理文本数据划分为连续的首部段落组成部分、中部段落组成部分和尾部段落组成部分，作为多个段落组成部分。例如，可以按照待处理文本数据的总字数进行划分，或者是按照待处理文本数据的段落分布进行划分。

[0079] 步骤S101-3，依据各分句在多个段落组成部分中所属的段落组成部分对应的权重值，确定各分句的分句权重。每个段落组成部分按照在文本中位置的不同赋予的权重值也不同，段落组成部分中的分句赋予的权重值也就不同。以新闻文章为例，通常先介绍事件背景、中间介绍事件经过，最后介绍事件结果，因此，一般来讲，首段段落组成部分对应的权重值较大，而中部段落组成部分对应的权重值较小。

[0080] 进一步地，得到待处理文本数据中的分句及分句对应的分句权重后，执行步骤S102，对各分句进行分词，得到各分句中的多个分词及各分词对应的分词权重。

[0081] 具体来讲，对于任一分句，得到对应的分词权重可以包括：

[0082] S102-1，对该分句进行分词处理，得到该分句对应的多个词语。分词是自然语言处理的基础，分词准确度直接决定了后面的词性标注、句法分析、词向量以及文本分析的质量。对于中文文本来讲，可以采用基于词典分词算法或者是基于统计的机器学习算法实现对任一分句的分词。

[0083] S102-2，去除分句中多个词语中的停用词后得到词语，作为分句中的多个分词。

[0084] 停用词是指在信息检索中，为节省存储空间和提高搜索效率，在处理自然语言数据(或文本)之前或之后会自动过滤掉某些字或词，这些字或词即被称为停用词。通常意义上，停用词大致分为两类：一类是人类语言中包含的功能词，这些功能词极其普遍，与其他词相比，功能词没有什么实际含义，如“在”、“的”等；另一类是词包括词汇词，由于其本身相对于文本整体来讲不具备实质性意义，因此，可以将其删除。去除分句中的停用词，主要是对待处理文本数据做一些基本的文本清理以尽可能避免文本数据的噪音对摘要提取的影响。

[0085] S102-3，依次判断各分词对应的词语属性，并基于词语属性为各分词分配对应的

分词权重;本实施例中的词语属性包括关键词和普通词,而且关键词的权重大于普通词的权重。对于属于不同段落组成部分中,关键词和普通词的权重可以相同,也可以不同;亦或是,关键词的权重不同,而普通词的权重相同,具体可以根据不同的需求进行设置。本实施例中的关键词可以是出现频率较高的分词,普通词则是出现频率较低的词。亦或是,根据分词的词性进行确定,对于分词中的实体名词可以作为关键词,动词为普通词等等。

[0086] 本申请实施例中,可以通过TF-IDF判断各分词对应的词语属性从而获得各分词的分词权重。TF-IDF是一种用于信息检索与数据挖掘的常用加权技术,也是一种统计方法,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。其中,TF是词频,IDF是逆文本频率指数。

[0087] 词语权重的计算公式如下:

[0088] TF-IDF=TF*IDF

[0089] 其中:

[0090] TF=词条t出现次数/文档中总词条数量

[0091] IDF=log(语料库中文档总数/(包含词条t的文档数+1))

[0092] 而在一个句子中,基于词语权重的句子权重的计算公式如下:

[0093] $Score=\sum weight(word_1+word2+word3+\cdots word_n)$

[0094] 其中:

[0095] Score表示句子的权重;

[0096] (word_1,word2,word_3,..word_n)表示句子中句子分词得到的词语序列。

[0097] 获取到各分句及分句权重、分句中各分词及分词权重后,接下来执行步骤S103,基于分句权重、分词权重计算各分句对应的句子权重向量。

[0098] 参见上述步骤S103,得到分句权重和分词权重后,可以基于所述分句权重、所述分词权重计算各所述分句对应的句子权重向量。在本申请一可选实施例中,计算各分句对应的句子权重向量包括:利用Glove词嵌入生成各分词对应的分词向量,并将每个分句中的各分词对应的分词向量相加得到各分句对应的分句向量;再将同一分句对应的分句向量、分句权重和分词权重相乘,得到各分句对应的句子权重向量。

[0099] Glove(英文全称,Global Vectors for Word Representation),它是一个基于全局词频统计的词表征(英文全称word representation)工具,是一种新的词矩阵的生成的方法,综合运用词的全局统计信息和局部统计信息来生成语言模型和词向量。Glove可以无监督地学习词向量表示,本质上为以加权最小二乘为目标的对数双线性模型。

[0100] 参见步骤S104,依据各分句对应的句子权重向量在多个分句中选取至少一目标分句。

[0101] 进一步地,先利用余弦相似度对各句子对应的句子权重向量进行计算得到相似度矩阵;再将相似度矩阵转化为图形,图形中的节点表示句子,边表示句子之间的相似度得分;然后利用PageRank网页排名算法对各分句进行排序,根据排序结果在所述多个分句中选取至少一目标分句。

[0102] 也就是说,先创建一个空的相似度矩阵,基于各句子对应的句子权重向量,使用余弦相似法计算各句子之间的余弦相似性,进而填充相似度矩阵;然后将相似矩阵转换为图形,并利用PageRank网页排名算法在此图形中对各句子进行排名,再截取排名前N名的句子

作为目标分句,但对于目标分句的数量,本申请不作限定。

[0103] PageRank算法是图的链接分析的代表性算法,属于图数据上的无监督学习方法。PageRank算法最初作为互联网网页重要度的计算方法,早先用于谷歌搜索引擎的网页排序。事实上,PageRank可以定义在任意有向图上,后来被应用到社会影响力分析、文本摘要等多个问题。

[0104] 最后,执行步骤S105,依据至少一目标分句生成待处理文本数据对应的文本摘要。

[0105] 在本申请一可选实施例中,依据目标分句在待处理文本数据中的分布顺序进行拼接,生成待处理文本数据对应的文本摘要。即生成文本摘要结果时,要考虑原有句子的顺序。

[0106] 举例来讲,如图2所述,对需要提取摘要的文本数据,可按以下步骤来进行提取:

[0107] 第一步,分句初步得到分句权重。将文本数据读取,按照标点符号、换行符等进行分句得到每个句子,按照文本长度整个文本分为三个部分,分别为首段、中段、尾段;对每个分句出现的位置进行首、中、尾段的划分,并赋予不同的权重,首段为文章的前1/4,权重为0.85,中段为文章中间1/2,权重为0.58,尾段为文章后1/4,权重为0.68;

[0108] 第二步,分词并根据TF-IDF初步得到分词权重。对分句进行分词,去除停用词,判断分词是否是关键词,如果是,得到关键词权重;否则得到普通词权重。首段中,关键词权重为0.82,普通词权重为0.32;中段中,关键词权重为0.92,普通词权重为0.25;尾段中,关键词权重为0.62,普通词权重为0.25。

[0109] 第三步,利用Glove词嵌入,得到句子权重向量。对词向量进行相加得到句子向量,句子向量乘以句子权重值再乘上分词权重得到最终的句子权重向量。

[0110] 第四步,排序得到文本摘要。使用余弦相似度对句子权重向量进行计算得到相似度矩阵,将相似矩阵转换为图形。图形中的节点表示句子,边表示句子之间的相似度得分。在这个图中,使用PageRank算法得到句子的排名。选择TopK个句子并排序作为文本摘要。

[0111] 基于同一发明构思,本申请实施例还提供了文本摘要提取系统,如图3所示,本申请实施例提供的文本摘要提取系统可以包括:

[0112] 分句切分模块310,其配置成读取待处理文本数据,将待处理文本数据进行切分,得到多个分句及各分句对应的分句权重;

[0113] 分词获取模块320,其配置成对各分句进行分词,得到各分句中的多个分词及各分词对应的分词权重;

[0114] 句子权重向量计算模块330,其配置成基于分句权重、分词权重计算各分句对应的句子权重向量;

[0115] 目标分句选取模块340,其配置成依据各分句对应的句子权重向量在多个分句中选取至少一目标分句;

[0116] 文本摘要生成模块350,其配置成依据至少一目标分句生成待处理文本数据对应的文本摘要。

[0117] 本申请一可选实施例中,分句切分模块310,其还可以配置成:

[0118] 按照字符分割方式对待处理文本数据,进行切分,得到多个分句;其中,字符包括标点符号和/或换行符;

[0119] 将待处理文本数据划分为连续的多个段落组成部分,并对各段落组成部分分别赋

予不同的权重值；

[0120] 依据各分句在多个段落组成部分中所属的段落组成部分对应的权重值，确定各分句的分句权重。

[0121] 本申请一可选实施例中，分句切分模块310，其还可以配置成：

[0122] 按照待处理文本数据的文本长度结构，将所述待处理文本数据划分为连续的首部段落组成部分、中部段落组成部分和尾部段落组成部分，作为多个段落组成部分。

[0123] 本申请一可选实施例中，分词获取模块320，其还可以配置成：

[0124] 对于任一分句，对分句进行分词处理，得到分句对应的多个词语；

[0125] 去除分句中多个词语中的停用词后得到词语，作为分句中的多个分词；

[0126] 依次判断各分词对应的词语属性，并基于词语属性为各分词分配对应的分词权重；词语属性包括关键词和普通词。

[0127] 本申请一可选实施例中，句子权重向量计算模块330，其还可以配置成：

[0128] 利用Glove词嵌入生成各分词对应的分词向量，并将每个分句中的各分词对应的分词向量相加得到各分句对应的分句向量；

[0129] 将同一分句对应的分句向量、分句权重和分词权重相乘，得到各分句对应的句子权重向量。

[0130] 本申请一可选实施例中，目标分句选取模块340，其还可以配置成：

[0131] 利用余弦相似度对各句子对应的句子权重向量进行计算得到相似度矩阵；

[0132] 将相似度矩阵转化为图形，图形中的节点表示句子，边表示句子之间的相似度得分；

[0133] 利用PageRank网页排名算法对各分句进行排序，根据排序结果在多个分句中选取至少一目标分句。

[0134] 本申请一可选实施例中，文本摘要生成模块350，其还可以配置成：

[0135] 依据目标分句在待处理文本数据中的分布顺序进行拼接，生成待处理文本数据对应的文本摘要。

[0136] 本申请实施例还提供了一种计算设备，包括存储器、处理器和存储在存储器内并能由处理器运行的计算机程序，其中，处理器执行所述计算机程序时实现如上述任一项所述的文本摘要提取方法。

[0137] 本申请实施例还提供了一种计算机可读存储介质，优选为非易失性可读存储介质，其内存储有计算机程序，计算机程序在由处理器执行时实现如上述任一项所述的文本摘要提取方法。

[0138] 本申请提供了一种文本摘要提取方法及系统、计算设备、存储介质，在本申请提供的方法中，先读取待处理文本数据进行切分，得到多个分句及各分句对应的分句权重；再对各分句进行分词，得到各分句中的多个分词及各分词对应的分词权重；然后基于分句权重、分词权重计算各分句对应的句子权重向量；最后依据各分句对应的句子权重向量在多个分句中选取至少一目标分句，并生成待处理文本数据对应的文本摘要。基于本申请提供的一种文本摘要提取方法及系统，基于关键词词频与位置和句子位置自动生成的文本摘要可以有效改过待处理文本数据的基本含义，并且还能够解决基于TF-IDF文本摘要未考虑词频位置信息和TextRank未考虑句子位置信息导致信息提取不足的问题。

[0139] 在上述实施例中，可以全部或部分地通过软件、硬件、固件或者其任意组合来实现。当使用软件实现时，可以全部或部分地以计算机程序产品的形式实现。所述计算机程序产品包括一个或多个计算机指令。在计算机加载和执行所述计算机程序指令时，全部或部分地产生按照本申请实施例所述的流程或功能。所述计算机可以是通用计算机、专用计算机、计算机网络、获取其他可编程装置。所述计算机指令可以存储在计算机可读存储介质中，或者从一个计算机可读存储介质向另一个计算机可读存储介质传输，例如，所述计算机指令可以从一个网站站点、计算机、服务器或数据中心通过有线(例如同轴电缆、光纤、数字用户线(DSL))或无线(例如红外、无线、微波等)方式向另一个网站站点、计算机、服务器或数据中心进行传输。所述计算机可读存储介质可以是计算机能够存取的任何可用介质或者是包含一个或多个可用介质集成的服务器、数据中心等数据存储设备。所述可用介质可以是磁性介质，(例如，软盘、硬盘、磁带)、光介质(例如，DVD)、或者半导体介质(例如固态硬盘Solid State Disk(SSD))等。

[0140] 专业人员应该还可以进一步意识到，结合本文中所公开的实施例描述的各示例的单元及算法步骤，能够以电子硬件、计算机软件或者二者的结合来实现，为了清楚地说明硬件和软件的可互换性，在上述说明中已经按照功能一般性地描述了各示例的组成及步骤。这些功能究竟以硬件还是软件方式来执行，取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能，但是这种实现不应认为超出本申请的范围。

[0141] 本领域普通技术人员可以理解实现上述实施例方法中的全部或部分步骤是可以通过程序来指令处理器完成，所述的程序可以存储于计算机可读存储介质中，所述存储介质是非短暂性(英文:non-transitory)介质，例如随机存取存储器，只读存储器，快闪存储器，硬盘，固态硬盘，磁带(英文:magnetic tape)，软盘(英文:floppy disk)，光盘(英文:optical disc)及其任意组合。

[0142] 以上所述，仅为本申请较佳的具体实施方式，但本申请的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本申请揭露的技术范围内，可轻易想到的变化或替换，都应涵盖在本申请的保护范围之内。因此，本申请的保护范围应该以权利要求的保护范围为准。

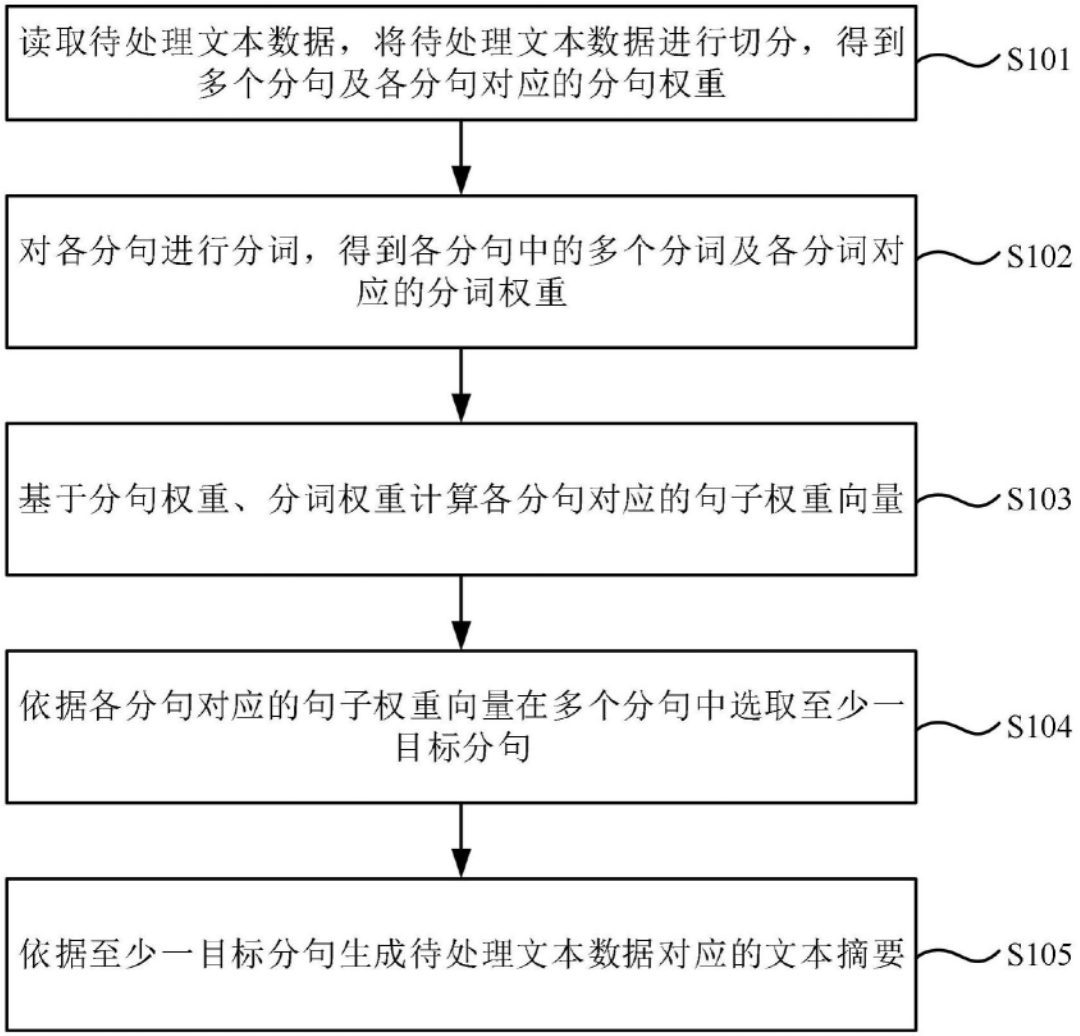

读取待处理文本数据，将待处理文本数据进行切分，得到多个分句及各分句对应的分句权重
S101
对各分句进行分词，得到各分句中的多个分词及各分词对应的分词权重
S102
基于分句权重、分词权重计算各分句对应的句子权重向量
S103
依据各分句对应的句子权重向量在多个分句中选取至少一目标分句
S104
依据至少一目标分句生成待处理文本数据对应的文本摘要
S105

图1

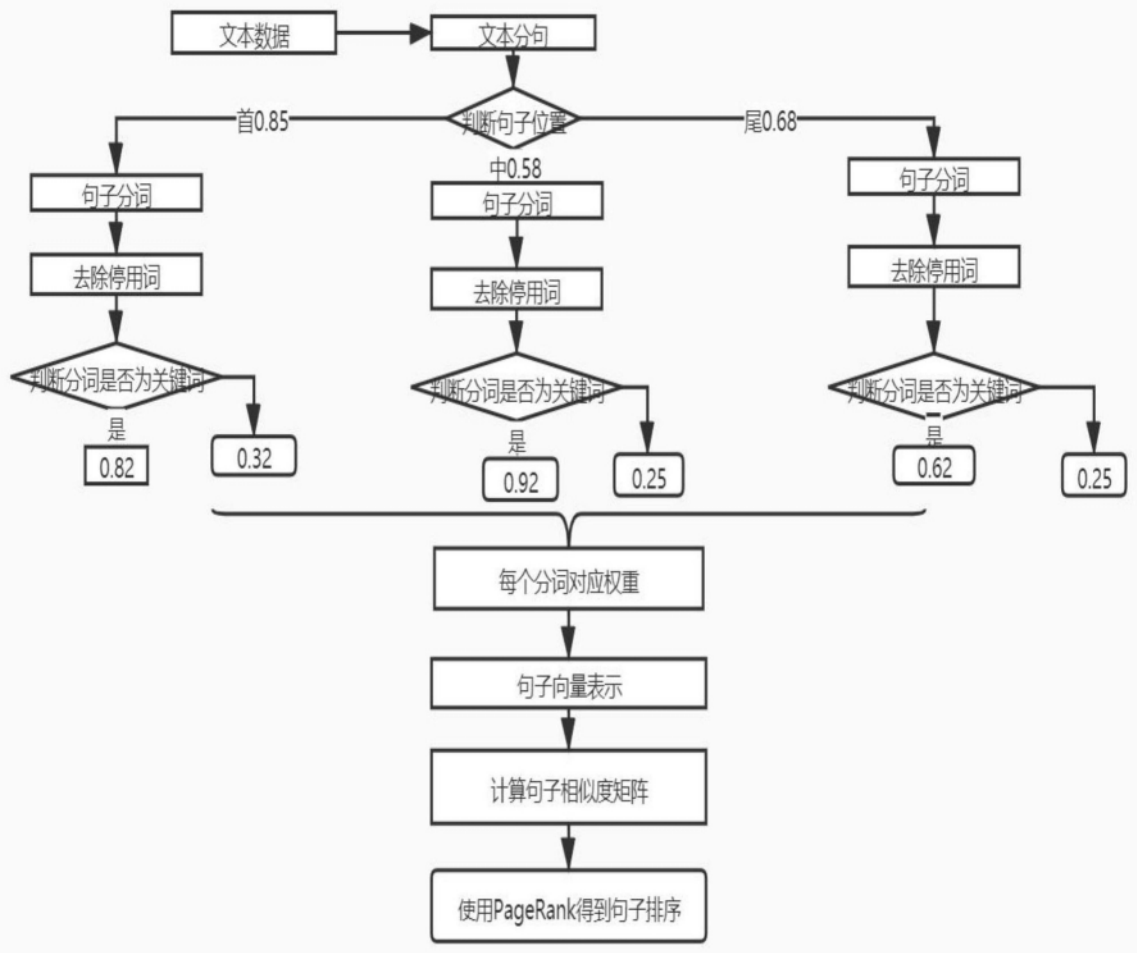

文本数据
文本分句
判断句子位置
首0.85
中0.58
尾0.68
句子分词
去除停用词
判断分词是否为关键词
是
0.82
0.32
句子分词
去除停用词
判断分词是否为关键词
是
0.92
0.25
句子分词
去除停用词
判断分词是否为关键词
是
0.62
0.25
每个分词对应权重
句子向量表示
计算句子相似度矩阵
使用PageRank得到句子排序

图2

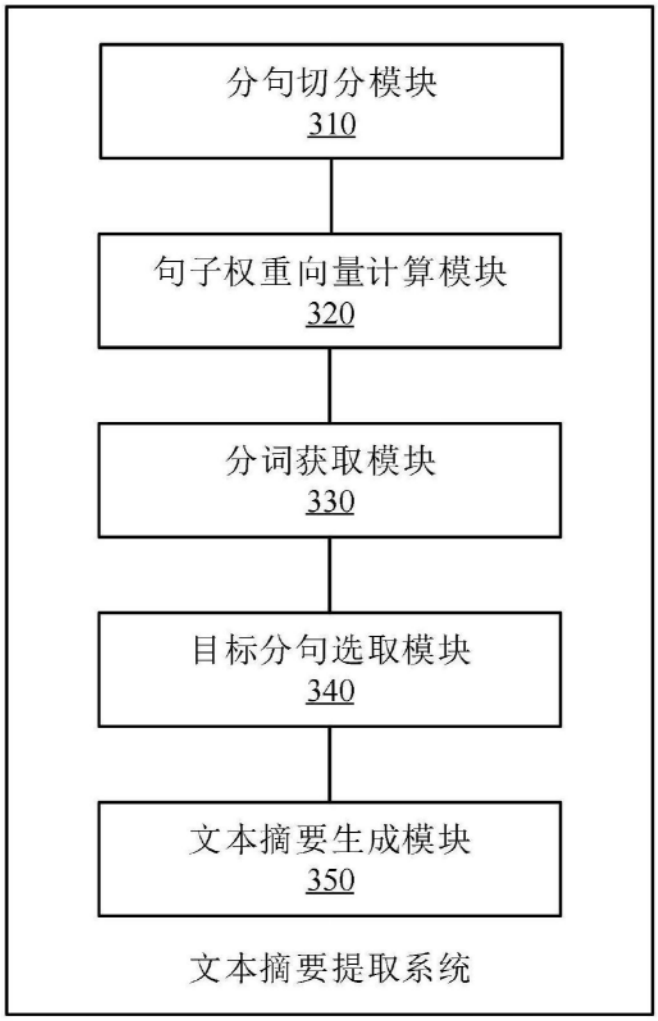

分句切分模块
310
句子权重向量计算模块
320
分词获取模块
330
目标分句选取模块
340
文本摘要生成模块
350
文本摘要提取系统

图3

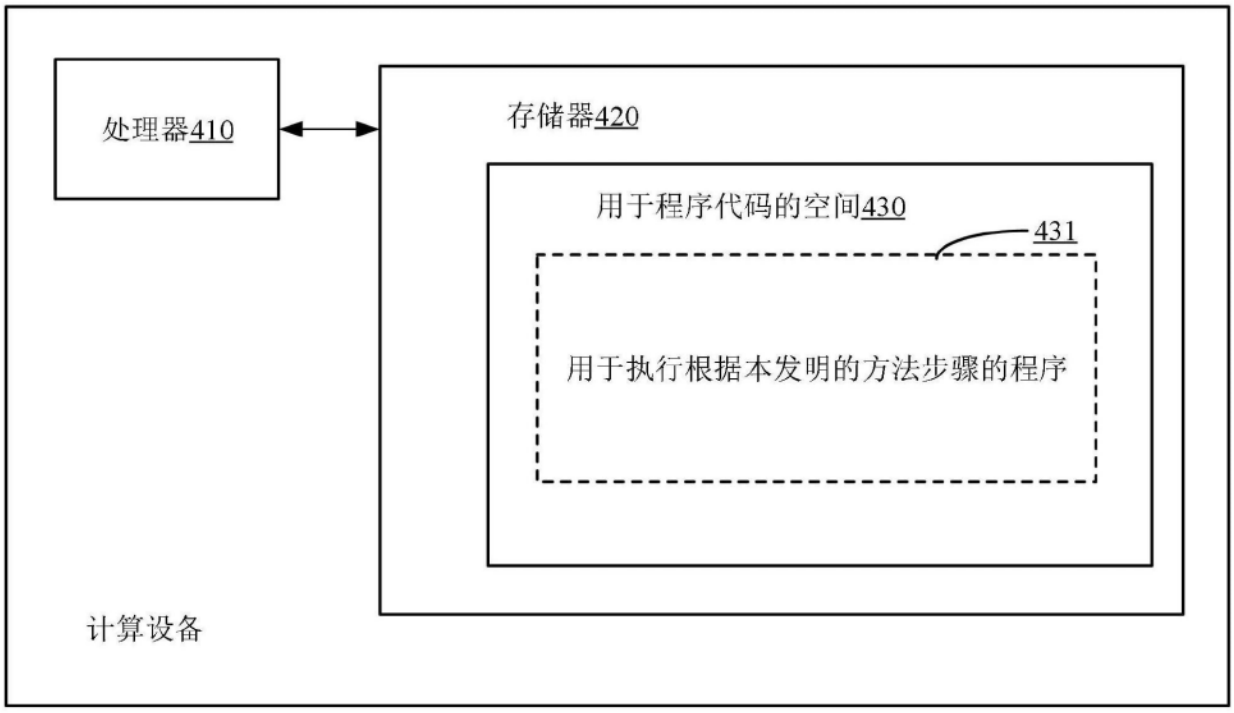

处理器410
存储器420
用于程序代码的空间430
431
用于执行根据本发明的方法步骤的程序
计算设备

图4

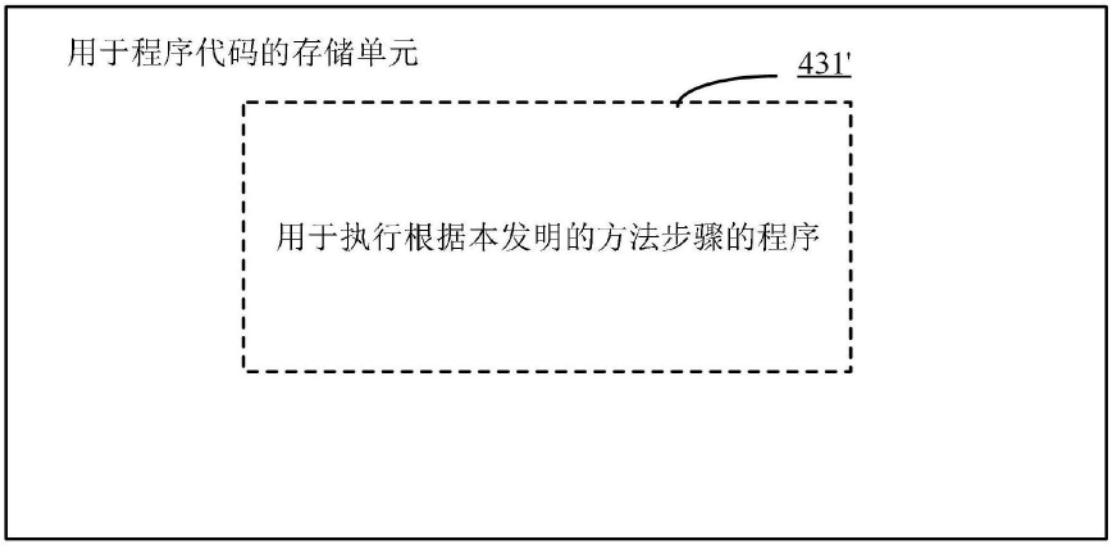

用于程序代码的存储单元
431'
用于执行根据本发明的方法步骤的程序

图5