

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2022 年 12 月 8 日 (08.12.2022)



(10) 国际公布号
WO 2022/253138 A1

- (51) 国际专利分类号:
G06F 16/332 (2019.01) *G06F 40/289* (2020.01)
G06F 40/216 (2020.01) *G06N 3/04* (2006.01)
- (21) 国际申请号: PCT/CN2022/095648
- (22) 国际申请日: 2022 年 5 月 27 日 (27.05.2022)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
202110603987.6 2021年5月31日 (31.05.2021) CN
- (71) 申请人: 维沃移动通信有限公司 (VIVO MOBILE COMMUNICATION CO., LTD.) [CN/CN];
中国广东省东莞市长安镇维沃路1号,
Guangdong 523863 (CN)。
- (72) 发明人: 蒋建光(JIANG, Jianguang); 中国广东省东莞市长安镇维沃路1号, Guangdong 523863 (CN)。
时从斌(SHI, Congbin); 中国广东省东莞市长安镇维沃路1号, Guangdong 523863 (CN)。
- (74) 代理人: 北京远志博慧知识产权代理事务所(普通合伙) (BOHUI INTELLECTUAL PROPERTY); 中国北京市海淀区交大东路31号东区10号楼等17幢31幢108, Beijing 100044 (CN)。
- (81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN,

(54) Title: TEXT PROCESSING METHOD AND APPARATUS, AND ELECTRONIC DEVICE

(54) 发明名称: 文本处理方法、装置和电子设备

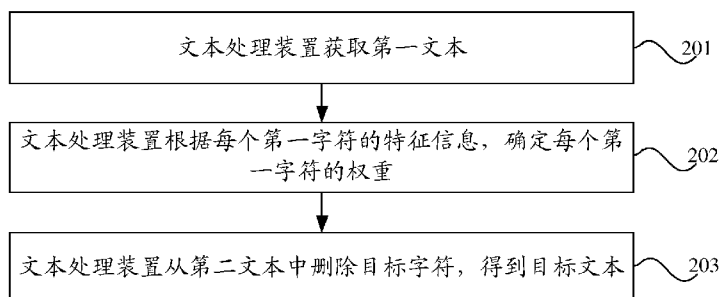


图 1

- 201 A text processing apparatus acquires first text
202 The text processing apparatus determines the weight of each first character according to feature information of each first character
203 The text processing apparatus deletes a target character from second text, so as to obtain target text

(57) Abstract: Provided are a text processing method and apparatus, and an electronic device, which belong to the technical field of communications. The method comprises: acquiring first text, wherein the first text comprises M first characters, M being a positive integer; according to feature information of each first character, determining the weight of each first character, wherein the weight of any first character is used for representing the degree of importance of any first character in the first text; and deleting a target character from second text, so as to obtain target text, wherein the second text is all or part of the text of the first text, and the target character is a first character having a weight that meets a first condition.

(57) 摘要: 提供了一种文本处理方法、装置和电子设备, 属于通信技术领域。方法包括: 获取第一文本, 第一文本包括M个第一字符, M为正整数; 根据每个第一字符的特征信息, 确定每个第一字符的权重, 任一第一字符的权重用于表征任一第一字符在第一文本中的重要程度; 从第二文本中删除目标字符, 得到目标文本, 第二文本为第一文本的全部或部分文本, 目标字符为权重满足第一条件的第一字符。

MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE,
PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE,
SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ,
UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW。

(84) 指定国 (除另有指明, 要求每一种可提供的地区
保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ,
NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM,
AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG,
CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU,
IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT,
RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI,
CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告 (条约第21条(3))。

文本处理方法、装置和电子设备

相关申请的交叉引用

本申请主张在 2021 年 05 月 31 日在中国提交的中国专利申请号
5 202110603987.6 的优先权，其全部内容通过引用包含于此。

技术领域

本申请属于通信技术领域，具体涉及一种文本处理方法、装置和电子设备。

背景技术

随着电子设备具有的搜索功能越来越强大，用户可以使用电子设备的搜索
10 功能对长文本进行搜索。但是由于长文本内容过多，可能会导致电子设备搜索过程耗时过长、搜索效率低。

目前，为了提高搜索效率，减少搜索时间等，用户在使用电子设备的搜索功能对长文本进行搜索的过程中，通常是通过截断长文本来实现的。具体地，电子设备可以仅提取长本文的前 N 个字符，然后，电子设备可以只对该 N 个
15 字符进行搜索，如此，达到减少搜索时长，提高搜索效率的目的。

然而，由于未被搜索的文本中可能包括用户需求搜索的信息，因此，通过上述的搜索方案可能会导致电子设备无法搜索到用户需求的信息，从而使得电子设备在对长文本进行搜索的过程中搜索正确率低。

发明内容

20 本申请实施例的目的是提供一种文本处理方法、装置和电子设备，能够解决在电子设备对长文本进行搜索的过程中，能够在提高搜索效率的同时提高搜索正确率的问题。

第一方面，本申请实施例提供了一种文本处理方法，该方法包括：获取第一文本，该第一文本包括 M 个第一字符，M 为正整数；根据每个第一字符的特征信息，确定每个第一字符的权重，任一第一字符的权重用于表征该任一第一字符在第一文本中的重要程度；从第二文本中删除目标字符，得到目标文本，
25 该第二文本为该第一文本的全部或部分文本，该目标字符为权重满足第一条件的第一字符。

第二方面，本申请实施例提供了一种文本处理装置，该装置包括：获取模块、确定模块和执行模块，其中：获取模块，用于获取第一文本，该第一文本包括 M 个第一字符， M 为正整数；确定模块，用于根据每个第一字符的特征信息，确定每个第一字符的权重，任一第一字符的权重用于表征该任一第一字符在第一文本中的重要程度；执行模块，用于从第二文本中删除目标字符，得到目标文本，该第二文本为第一文本的全部或部分文本，该目标字符为权重满足第一条件的第一字符。

第三方面，本申请实施例提供了一种电子设备，该电子设备包括处理器、存储器及存储在上述存储器上并可在所述处理器上运行的程序或指令，所述程序或指令被所述处理器执行时实现如第一方面所述的方法的步骤。

第四方面，本申请实施例提供了一种可读存储介质，所述可读存储介质上存储程序或指令，所述程序或指令被处理器执行时实现如第一方面所述的方法的步骤。

第五方面，本申请实施例提供了一种芯片，所述芯片包括处理器和通信接口，所述通信接口和所述处理器耦合，所述处理器用于运行程序或指令，实现如第一方面所述的方法。

在本申请实施例中，文本处理装置在获取包括 M 个第一字符的第一文本之后，可以根据每个第一字符的特征信息，确定每个第一字符的权重。然后，文本处理装置可以从第二文本中删除目标字符，得到目标文本。其中，任一第一字符的权重用于表征该任一第一字符在第一文本中的重要程度，该第二文本为该第一文本的全部或部分文本，该目标字符为权重满足第一条件的第一字符。通过上述方案，在电子设备对长文本（如第一文本）进行搜索的情况下，文本处理装置可以根据第一文本中每个第一字符的特征信息，确定每个第一字符的权重。然后，文本处理装置可以删除权重满足第一条件的第一字符，从而可以以词为单位精细地简化第一文本，保留第一文本中的重要字符。如此，使得电子设备在对长文本进行搜索的过程中，可以仅搜索第一文本中的重要字符，进而能够在提高搜索效率的同时提高搜索正确率。

附图说明

图 1 为本申请实施例提供的一种文本处理方法流程示意图；

图 2 为本申请实施例提供的一种文本处理方法应用的界面示意图；

图 3 为本申请实施例提供的一种文本处理装置的结构示意图；

图 4 为本申请实施例提供的一种电子设备的结构示意图之一；

图 5 为本申请实施例提供的一种电子设备的结构示意图之二。

具体实施方式

下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚描述，显然，所描述的实施例是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员获得的所有其他实施例，都属于本申请保护的范围。

本申请的说明书和权利要求书中的术语“第一”、“第二”等是用于区别类似的对象，而不适用于描述特定的顺序或先后次序。应该理解这样使用的数据在适当情况下可以互换，以便本申请的实施例能够以除了在这里图示或描述的那些以外的顺序实施。且“第一”、“第二”等所区分的对象通常为一类，并不限定对象的个数，例如第一对象可以是一个，也可以是多个。此外，说明书以及权利要求中“和/或”表示所连接对象的至少其中之一，字符“/”，一般表示前后关联对象是一种“或”的关系。

下面结合附图，通过具体的实施例及其应用场景对本申请实施例提供的文本处理方法进行详细地说明。

图 1 为本申请实施例提供的一种文本处理方法流程示意图，包括步骤 201 至步骤 203：

步骤 201：文本处理装置获取第一文本。

其中，上述第一文本包括 M 个第一字符， M 为正整数。

在本申请实施例中，本申请中的字符可以包括以下至少一项：文字，数字，字母，符号。

在本申请实施例中，第一文本可以是用户通过文本处理装置输入的，也可以是文本处理装置下载的，还可以是文本处理装置通过任意可能的方式获取的，本申请实施例对此不作限定。

在本申请实施例中，文本处理装置可以主动获取第一文本，也可以被动获取第一文本，本申请实施例对此不作限定。

在一种示例中，用户可以在搜索区域中输入第一文本，此时，文本处理装置可以获取该第一文本，即主动获取。

在另一种示例中，用户可以在搜索区域中输入第一文本，然后，文本处理装置可以在接收到用户的第一输入之后，获取该第一文本，即被动获取。

示例性地，上述的第一输入可以为：用户对目标控件的点击输入，或者为用户输入的语音指令，或者为用户输入的特定手势，具体的可以根据实际使用需求确定，本申请实施例对此不作限定。其中，该目标控件可以为原有控件，也可以为新增控件，本申请实施例对此不作限定。

- 5 本申请实施例中的特定手势可以为单击手势、滑动手势、拖动手势、压力识别手势、长按手势、面积变化手势、双按手势、双击手势中的任意一种；本申请实施例中的点击输入可以为单击输入、双击输入或任意次数的点击输入等，还可以为长按输入或短按输入。

10 步骤 202：文本处理装置根据每个第一字符的特征信息，确定每个第一字符的权重。

其中，任一第一字符的权重用于表征该任一第一字符在第一文本中的重要程度。

在本申请实施例中，上述的特征信息可以包括一个信息，也可以包括多个信息，本申请实施例对此不作限定。

- 15 可选地，在本申请实施例中，上述的特征信息可以包括以下至少一项：文本信息，历史使用信息，上下文信息，语言概率信息。

示例性地，上述的文本信息可以包括以下至少一项：任一第一字符字符数量（或称为长度），该任一第一字符所在的第一字符串的字符数量，该任一第一字符在所在的第一字符串中的偏移量。

- 20 需要说明的是，本申请中的本文信息包括但不限于上述的三种信息。

示例性地，上述的历史使用信息可以包括以下至少一项：任一第一字符的历史点击次数，该任一第一字符在第一文本出现的次数。

需要说明的是，本申请中的历史使用信息包括但不限于上述的两种信息。

- 25 示例性地，上述的上下文信息可以包括以下任一项：任一第一字符所在的第一字符串中的全部或部分字符，第一文本中的全部或部分字符。

需要说明的是，本申请中的上下文信息包括但不限于上述的两种信息。

- 30 示例性的，上述的语言概率信息包括以下至少一项：任一第一字符所在的第一字符串的语言模型概率，将该任一第一字符随机替换为其他字符后的所在的第一字符串的语言模型概率，第一文本的语言模型概率，将该任一第一字符随机替换为其他字符后的第一文本的语言模型概率。

需要说明的是，本申请中的语言概率信息包括但不限于上述的两种信息。

示例性地，文本处理装置根据每个第一字符的特征信息，构建每个第一字

符的特征向量，然后，文本处理装置可以采用词权重模型对每个第一字符的特征向量进行计算，得到每个第一字符的权重。

具体的构建字符的特征向量的过程可以如下：

示例性地，文本处理装置可以根据句号和逗号等标点符号对第一文本进行分句。例如，第一文本为“苹果手机，苹果很好吃”，分句后的结果为“苹果手机”和“苹果很好吃”。然后，文本处理装置可以对分句后的各个句子进行分词。例如，文本处理装置可以将句子“苹果手机”分成“苹果”和“手机”。在完成分词之后，文本处理装置可以获取如下四种特征：

第一种特征，根据文本信息获取第一字符的文本特征。

例如，“苹果很好吃”中的“苹果”一词的文本信息为文本长度 4，词的长度 2，词在文本中的偏移量 0，即可以构建向量 (4, 2, 0)，然后，经过一个线性变换层（输出为 1 维）将其变成一维的数据，再经过一个激活函数，将最后的结果作为文本特征，可以记为 v1。

第二种特征，根据历史使用信息获取第一字符的统计特征。

在一种示例中，文本处理装置可以将词频逆文本频率指数（term frequency-inverse document frequency, TF-IDF）的值作为统计特征。

其中，传统的 TF-IDF 公式可以表示为：

$$\text{TF-IDF} = \frac{tf}{n_{doc}} * \log\left(\frac{m}{df+1}\right) \quad (1)$$

示例性地，本申请实施例在传统的 TF-IDF 公式的基础上进行了点击率加权。具体地，本申请实施例中的进行点击率加权的 TF-IDF 公式可以表示为：

$$\text{TF-IDF} = \frac{tf}{n_{doc}} * \log\left(\frac{m}{df+1}\right) * \log(1.1 + \text{CTR} * 100) \quad (2)$$

其中， tf 为词在文本中出现的次数， n_{doc} 为文档的总词数， m 为语料库的总文档数， df 为包含该词的文档数，CTR 为该词的点击率。

需要说明的是，传统的 TF-IDF 计算可以参考现有技术，此处不再赘述。

本申请实施例中从收集的文本数据和用户历史点击数据中统计得到的 TF-IDF 值，可以记为 v2。

第三种特征，根据上下文信息获取第一字符的词嵌入（词向量）特征。

示例性地，文本处理装置可以先获得词的上下文无关的向量表示，文本处理装置在词向量表中查寻，获得分词后每个词的词的上下文无关的向量表示。

例如，查询得到“苹果”的向量为 (0.1, 0.2, 0.6)，“手机”的向量为 (0.2, 0.2, 0.6)。需要说明的是，此时“苹果手机”和“苹果很好吃”中的“苹果”

的向量均为 (0.1, 0.2, 0.6), 均与上下文无关。然后, 文本处理装置可以获得词的上下文相关的向量表示。文本处理装置可以将上述词的上下文无关的向量表示输入上下文相关的编码器 (例如, transformer 编码器), 获得词的上下文相关的向量表示。

5 举例说明, 假设一段文本包含 n 个词, $s=(w_1, w_2, \dots, w_n)$, 其中 w_i 表示第 i 个词, v_i 表示第 i 个词的上下文无关的向量表示。而词 w_i 的上下文相关的向量表示就是在计算词 w_i 的表示时考虑上下文所有词对词 w_i 的影响。具体地, 上下文所有词 w_j 对 w_i 的影响程度为点乘相似性 $a_{ij}=v_i \cdot v_j$, 则 w_i 的上下文相关的向量表示可以为 $\sum_{j=1}^n a_{ij} * v_j$ 。此时“苹果手机”中的“苹果”的向量变为 (0.2, 0.3, 0.7), 而“苹果很好吃”中的“苹果”的向量变为 (0.1, 0.4, 0.6)。如此, 10 两个“苹果”的向量表示由于上下文不同而变得不同了, 即为上下文相关的向量表示。

示例性地, 文本处理装置可以将上述词的上下文相关的向量表示经过一个线性变换层 (输出为 1 维) 变成一维特征, 记为 v_3 , 可以表示为: $v_3 = \sum_{j=1}^m a_j w_j + b$, 15 其中 w_j 为词向量的第 j 维分量, m 为词向量的维数, a_j 为相应的权重, b 为偏置项。

第四种特征, 根据语言概率信息获取第一字符的语言模型特征。

示例性地, 语言模型用于评估文本的合理性, 例如, 文本处理装置可以采用 n -gram 语言模型或神经网络语言模型对文本的合理性进行评估。本申请实 20 施例中的词 w 的语言模型特征可以记为 v_4 , 定义为 $v_4 = \text{输入文本的语言模型概率} / \text{输入文本将该词随机替换为别的词后的语言模型概率}$, 即可以表示为:

$v_4 = P(w_0, w, w_2, \dots, w, w_n) / P(w_0, \text{random}, w_2, \dots, \text{random}, w_n)$, 其中, random 表示随机替换的 w 以外的词。

在一种示例中, 采用语言模型求一段文本序列的概率的过程: 首先获得词的上下文相关的向量表示, 然后对句子中的所有词的上下文相关的向量求平均 25 得到句子的向量表示, 然后对句子的向量表示做一个线性变换变成一维数据, 再经过一个非线性激活函数 (比如 sigmoid) 得到一个概率值。

示例性地, 在文本处理装置获得上述四类特征之后, 可以拼接获得词的特征向量表示, 记为 $V = (v_1, v_2, v_3, v_4)$ 。如此, 文本处理装置可以获得第一 30 文本中每个第一字符的特征向量。

示例性地, 在文本处理装置获得第一文本中每个第一字符的特征向量之后, 可以采用词权重模型获得每个第一字符的权重得分。具体地, 以上述获得的词的特征向量 V 作为输入为例, 先经过一个线性变换层 (输出为 4 维) 将数

据变成 4 维 $Y = (y_1, y_2, y_3, y_4)$, 该线性变换层的公式可以表示为:

$$Y = W * V + B \quad (3)$$

其中, W 为权重矩阵, Y 为经过线性变换和激活函数的 4 维向量 (对应四类分值), 再经过一个激活函数给数据增加一些非线性, 该激活函数可以表示为:

$$Y = \text{sigmoid}(Y) \quad (4)$$

最后经过一个 softmax 分类器得到预测词权重得分, 公式如下:

$$\text{score}(W) = \arg \max \left(\frac{\exp(y_j)}{\sum_0^3 \exp(y_i)} \right) \quad (5)$$

其中, y_j 为 Y 的第 j 个维度, 对应取第 j 类分值的可能性大小, 然后经过
 10 $\frac{\exp(y_j)}{\sum_0^3 \exp(y_i)}$ 得到总和为 1 的概率分布, 取其中概率最大的维度对应的分值作为预测得分, 假设最后得到的概率分布为 (0.1, 0.0, 0.1, 0.8), 该概率分布表示取 0 分的概率为 0.1, 取 1 分的概率为 0, 取 2 分的概率为 0.1, 取 3 分的概率为 0.8, 因此, 文本处理装置可以预测该词为 3 分。

需要说明的是, 本申请实施例将词权重预测问题看做一个 0-3 的四分类问题, (1, 0, 0, 0) 表示 1 分, (0, 0, 0, 1) 表示 3 分, (1, 0, 0, 0) 和 (0, 0, 0, 1) 称为真实分布, 而 (0.1, 0.0, 0.1, 0.8) 称为预测分布, 取预测分布中概率最大的维度对应的分值作为最终的预测得分, 此处即为 3 分。

示例性地, 上述的词权重模型可以选择逻辑回归或树模型 (如 LightGBM)。

示例性地, 词权重模型的训练为离线训练, 训练时的标签为 0-3 分的 4 档
 20 人工标注的词权重得分, 训练的目标为使得该词权重模型基于输入特征预测的词权重分布与人工评分尽可能的接近。

需要说明的是, 用户可以通过收集到的一定数量的文本数据及用户点击数据, 人工主观判断文本中每个词的重要性并结合用户点击数据进行打分。具体的: 主观判定重要或比较重要且高点击, 得分为 3 (即非常重要); 主观判定重要但点击率一般, 得分为 2 (即比较重要); 主观判定不重要且点击率低或无点击, 得分为 0 (即不重要); 其余的得分为 1 (即比较不重要)。

步骤 203: 文本处理装置从第二文本中删除目标字符, 得到目标文本。

其中, 上述第二文本为第一文本中的全部或部分文本, 上述目标字符为权重满足第一条件的第一字符。

30 在本申请实施例中, 上述的第一条件包括以下任一项: 权重小于第一阈值,

权重处于第一阈值区间，权重大于或等于该第一阈值。

示例性地，在文本处理装置获得第一文本中每个第一字符的权重之后，可以丢弃掉权重小于第一阈值（比如 1 分）的第一字符，获得目标文本。该目标文本不但简洁且保留了重要的第一字符，更加适合长文本搜索的需求。

- 5 本申请实施例提供的文本处理方法，文本处理装置在获取包括 M 个第一字符的第一文本之后，可以根据每个第一字符的特征信息，确定每个第一字符的权重。然后，文本处理装置可以从第二文本中删除目标字符，得到目标文本。其中，任一第一字符的权重用于表征该任一第一字符在第一文本中的重要程度，该第二文本为该第一文本的全部或部分文本，该目标字符为权重满足第一条件的第一字符。通过上述方案，在电子设备对长文本（如第一文本）进行搜索的情况下，文本处理装置可以根据第一文本中每个第一字符的特征信息，确定每个第一字符的权重。然后，文本处理装置可以删除权重满足第一条件的第一字符，从而可以以词为单位精细地简化第一文本，保留第一文本中的重要字符。如此，使得电子设备在对长文本进行搜索的过程中，可以仅搜索第一文本中的重要字符，进而能够在提高搜索效率的同时提高搜索正确率。
- 10 可选地，在本申请实施例中，电子设备在以字符为单位精细化的对第一文本进行去冗余之前，可以先以字符串（即句子）为单位对第一文本进行去冗余。

- 示例性地，在第一文本包括至少一个第一字符串，该至少一个第一字符串包括上述 M 个第一字符，一个第一字符串包括至少一个第一字符的情况下，
- 20 在上述的步骤 203 之前，该方法还可以包括如下步骤 203a 和步骤 203b:

 步骤 203a: 文本处理装置根据上述每个第一字符的权重，确定每个第一字符串的权重。

 其中，任一第一字符串的权重用于表征该任一第一字符串在第一文本中的重要程度。

- 25 步骤 203b: 文本处理装置从第一文本中删除目标字符串，得到第二文本。其中，上述目标字符串为权重满足第二条件的第一字符串。

 示例性地，上述的第二条件包括以下任一项：权重小于第二阈值，权重处于第二阈值区间，权重大于或等于该第二阈值。

- 可选地，在本申请实施例中，上述的步骤 203a 具体可以包括如下步骤
- 30 203a1:

 步骤 203a1: 文本处理装置将每个第一字符串对应的每个第一字符的权重进行线性变换和函数变换，确定每个第一字符串的权重。

示例性地，文本处理装置可以采用一个冗余句子剔除器对各个句子进行打分，具体的，文本处理装置可以将每个第一字符的权重先经过一个线性变换层将输出为 1 维向量，该线性变换层的公式可以表示为：

$$v_s = \sum_{i=1}^n \alpha_i w_i + b \quad (6)$$

5 其中， $\text{score}(s)$ 为句子得分， $\text{score}(w_i)$ 为句子中第 i 个词的词权重得分， α_i 为相应的权重。再经过 sigmoid 激活函数将数据变成 0-1 之间，并以 0-1 之间的数据作为句子的权重得分，该激活函数的公式可以表示为：

$$\text{score}(s) = \text{sigmoid}(\text{score}(s)) \quad (7)$$

10 最后，文本处理装置可以保留权重得分大于或等于第二阈值（如 0.5）的句子，删除其余的句子得到第二文本。

本申请实施例提供的文本处理方法可以应用于提高去冗余效果的场景中，用户可以先删除第一文本中不重要的第一字符串（即句子），然后，再从删除了不重要的句子中的第二文本中再删除第一字符（即词），如此可以提高去冗余的效果，从而进一步提高搜索效率。

15 可选地，在本申请实施例中，在上述的步骤 201 之前，该方法还可以包括如下步骤 201a 至步骤 201f：

步骤 201a：文本处理装置对待处理的第三文本执行分句处理，得到至少一个第二字符串。

20 示例性地，文本处理装置可以根据句号和逗号等标点符号对第三文本进行分句。

示例性的，第三文本可以是用户通过文本处理装置输入的，也可以是文本处理装置下载的，还可以是文本处理装置通过任意可能的方式获取的，本申请实施例对此不作限定。

25 示例性的，文本处理装置可以主动获取第三文本，也可以被动获取第三文本，本申请实施例对此不作限定。

步骤 201b：文本处理装置对上述至少一个第二字符串中的每个第二字符串执行分词处理，得到至少一个第二字符。

可以理解，上述第三文本包括至少一个第二字符串，一个第二字符串包括至少一个第二字符。

30 需要说明的是，上述的分句处理和分词处理的具体过程可以参考相关技术，此处不再赘述。

步骤 201c：文本处理装置获取上述至少一个第二字符串中的每个第二字符的

向量。

示例性的，上述每个第二字符的向量可以理解为每个第二字符的上下文相关向量。

需要说明的是，文本处理装置确定每个第二字符的上下文相关的向量的过程，具体可以参考本申请实施例中对根据每个第一字符的上下文信息，确定每个第一字符的上下文相关的向量过程的描述，此处不再赘述。

步骤 201d: 文本处理装置根据每个第二字符串对应的每个第二字符的向量，确定每个第二字符串的向量。

示例性地，针对任一第二字符串，在文本处理装置确定该任一第二字符串中的每个第二字符的上下文相关的向量之后，可以将所有第二字符的上下文相关的向量输入一个句子编码器，句子编码器从该任一第二字符串中各个第二字符的向量表示学习到整个第二字符串的向量表示。例如，可以采用求和、求平均或者加权求和的方式。其中，求和的公式可以表示为： $v_s = \sum_1^n w_i$ ；求平均的公式可以表示为： $v_s = \sum_1^n w_i / n$ ；加权求和的公式可以表示为： $v_s = \sum_1^n \alpha_i w_i + b$ 。其中， v_s 为句子的向量表示， w_i 为句子中第 i 个词的上下文相关的向量表示， α_i 为权重， b 为偏置项。

步骤 201e: 文本处理装置将每个第二字符串的向量进行线性变换和函数变换，确定每个第二字符串的权重。

其中，任一第二字符串的权重用于表征该任一第二字符串在第三文本中的重要程度。

示例性地，文本处理装置可以将任一第二字符串的向量表示输入一个句子分类器，其中，该句子分类器由一个线性变换层（输出为 1 维）和一个激活函数组成。具体的，文本处理装置可以把句子的向量送入线性变换层进行线性变换，得到一维的数值，该线性变换层的公式可以表示为：

$$y_s = \sum_1^n \alpha_j w_j + b \quad (8)$$

其中， y_s 为线性变换层的输出， w_j 为句子向量表示的第 j 维分量， α_j 为相应的权重。然后，文本处理装置可以把一维的数值送入激活函数进行变换，获得一个 0-1 的数值作为句子的得分 score。该激活函数的公式可以表示为：

$$\text{score} = \text{sigmoid}(y_s) \quad (9)$$

其中，score 为句子的得分。

步骤 201f: 文本处理装置从第三文本中删除权重满足第三条件的第二字符串, 得到第一文本。

示例性地, 上述的第三条件包括以下任一项: 权重小于第三阈值, 权重处于第三阈值区间, 权重大于或等于该第三阈值。

- 5 在一种示例中, 在第二字符串的得分 (score) 在 0-1 之间时, 当 score 大于或等于第三阈值 (如 0.5) 时预测为标签 1, 否则预测为标签 0, 其中, 1 表示属于摘要, 0 表示不属于摘要, 因此, 文本处理装置可以从第三文本中删除标签为 0 的句子, 选择标签为 1 的句子组成第一文本, 即对第三文本进行初步去冗余, 然后, 文本处理装置可以在第一文本的基础上, 再根据每个第一字符
- 10 的文本信息、历史使用信息、上下文信息和语言概率信息确定的每个第一字符的权重进行精细化的去冗余。如此, 可以提高对文本去冗余的效果, 从而进一步提高文本搜索的效率。

- 举例说明, 如图 2 中的 (a) 所示, 手机屏幕显示有新闻应用程序的主界面 31, 该主界面 31 的搜索框中显示有文本 1 “乙国是一个伟大的国家, 要问
- 15 为何如此伟大? 首先地大物博, 拥有 960 万平方公里的土地。其次历史悠久, 上下五千年。”。当用户想要搜索该长文本时, 用户可以点击“搜索”控件, 此时, 手机可以对该文本进行分句, 分为句子 1 “乙国是一个伟大的国家。”、句子 2 “要问为何如此伟大?”、句子 3 “首先地大物博,”、句子 4 “拥有 960 万平方公里的土地。”、句 5 “其次历史悠久,” 和句子 6 “上下五千年。”。

- 20 然后, 手机可以对句子 1 至句子 6 中的每个句子进行分词, 并确定句子 1 至句子 6 中每个词的上下文相关的向量表示。接着, 手机可以对句子 1 中的所有词的上下文相关的向量表示加权求和得到该句子 1 的向量表示, 手机可以对句子 2 中的所有词的上下文相关的向量表示加权求和得到该句子 2 的向量表示, 直至手机可以对句子 6 中的所有词的上下文相关的向量表示加权求和得到
- 25 该句子 6 的向量表示。然后, 手机可以根据句子的向量表示确定句子 1、句子 3 和句子 5 的权重得分大于 0.5, 标记为标签 1, 句子 2、句子 4 和句子 6 的权重得分小于 0.5, 标记为标签 0。

- 接着, 如图 2 中 (b) 所示, 手机可以从文本 1 中删除标签为 0 的句子 2、句子 4 和句子 6, 仅显示句子 1、句子 3 和句子 5, 得到句子级去冗余后的文本
- 30 1。然后, 手机可以对句子 1、句子 3 和句子 5 进行分词, 句子 1 可以分成词“乙国”、“是”、“一个”、“伟大”、“的”和“国家”。然后, 手机可以根据句子 1 中的每个词的文本信息、历史使用信息、上下文信息和语言概率信息联合确定

每个词的权重得分，其中，词“是”、“一个”和“的”权重得分小于1分，词“Z国”、“伟大”和“国家”权重得分大于1分，此时，手机可以删除句子1中的词“是”、“一个”和“的”，保留词“Z国”、“伟大”和“国家”。

最后，以此类推，对句子2和句子3进行处理后，如图2中的(c)所示，
5 手机显示词“Z国”、“伟大”、“国家”、“地大物博”和“历史悠久”。

需要说明的是，手机显示句子1、句子3和句子5以及手机显示词“Z国”、“伟大”、“国家”、“地大物博”和“历史悠久”，只是为了方便理解本申请，在实际去冗余的过程中，手机可以不执行显示句子1、句子3和句子5以及显示词“Z国”、“伟大”、“国家”、“地大物博”和“历史悠久”的步骤。

10 在一种示例中，文本处理装置可以在上述的特征信息可以包括文本信息、历史使用信息、上下文信息和语言概率信息的情况下，执行上步骤201a至步骤201f。

本申请实施例提供的文本处理方法可以应用于提高去冗余效果的场景中，文本处理装置可以根据第三文本中的每个第二字符的向量，确定每个第二字符串的向量，从而根据每个第二字符串的向量确定每个第二字符串的权重，并从
15 第三文本中删除权重满足第三条件的第二字符串，得到第一文本，从而实现对第三文本进行句子级的初步去冗余，然后，文本处理装置可以在对初步去冗余得到的第一文本进行词级去冗余，如此，可以提高对文本去冗余的效果。

需要说明的是，本申请实施例提供的文本处理方法，执行主体可以为文本
20 处理装置，或者该文本处理装置中的用于执行文本处理方法的控制模块。本申请实施例中以文本处理装置执行文本处理方法为例，说明本申请实施例提供的文本处理装置。

图3为实现本申请实施例提供的一种文本处理装置的可能的结构示意图，如图3所示，文本处理装置400包括：获取模块401、确定模块402和执行模块403，其中：获取模块401，用于获取第一文本，该第一文本包括M个第一
25 字符，M为正整数；确定模块402，用于根据每个第一字符的特征信息，确定每个第一字符的权重，任一第一字符的权重用于表征该任一第一字符在第一文本中的重要程度；执行模块403，用于从第二文本中删除目标字符，得到目标文本，该第二文本为第一文本的全部或部分文本，该目标字符为权重满足第一
30 条件的第一字符。

可选地，上述第一文本包括至少一个第一字符串，该至少一个第一字符串包括上述M个第一字符，一个第一字符串包括至少一个第一字符；确定模块

402, 还用于根据每个第一字符的权重, 确定每个第一字符串的权重, 任一第一字符串的权重用于表征该任一第一字符串在第一文本中的重要程度; 执行模块 403, 还用于从第一文本中删除目标字符串, 得到第二文本, 该目标字符串为权重满足第二条件的第一字符串。

5 可选地, 确定模块 402, 具体用于将每个第一字符串对应的每个第一字符的权重进行线性变换和函数变换, 确定每个第一字符串的权重。

可选地, 执行模块 403, 还用于对待处理的第三文本执行分句处理, 得到至少一个第二字符串; 以及对至少一个第二字符串中的每个第二字符串执行分词处理, 得到至少一个第二字符; 获取模块 401, 还用于获取至少一个第二字符串中的每个第二字符的向量; 确定模块 402, 还用于根据获取模块 401 获取的每个第二字符串对应的每个第二字符的向量, 确定每个第二字符串的向量; 以及将每个第二字符串的向量进行线性变换和函数变换, 确定每个第二字符串的权重, 任一第二字符串的权重用于表征该任一第二字符串在第三文本中的重要程度; 执行模块 403, 还用于从第三文本中删除权重满足第三条件的第二字符串, 得到第一文本。

15 可选地, 特征信息包括以下至少一项: 文本信息, 历史使用信息, 上下文信息, 语言概率信息。

本申请实施例提供的文本处理装置, 文本处理装置在获取包括 M 个第一字符的第一文本之后, 可以根据每个第一字符的特征信息, 确定每个第一字符的权重。然后, 文本处理装置可以从第二文本中删除目标字符, 得到目标文本。其中, 任一第一字符的权重用于表征该任一第一字符在第一文本中的重要程度, 该第二文本为该第一文本的全部或部分文本, 该目标字符为权重满足第一条件的第一字符。通过上述方案, 在电子设备对长文本(如第一文本)进行搜索的情况下, 文本处理装置可以根据第一文本中每个第一字符的特征信息, 确定每个第一字符的权重。然后, 文本处理装置可以删除权重满足第一条件的第一字符, 从而可以以词为单位精细地简化第一文本, 保留第一文本中的重要字符。如此, 使得电子设备在对长文本进行搜索的过程中, 可以仅搜索第一文本中的重要字符, 进而能够在提高搜索效率的同时提高搜索正确率。

25 本实施例中各种实现方式具有的有益效果具体可以参见上述方法实施例中相应实现方式所具有的有益效果, 为避免重复, 此处不再赘述。

本申请实施例中的文本处理装置可以是装置, 也可以是终端中的部件、集成电路、或芯片。该装置可以是移动电子设备, 也可以为非移动电子设备。示

例性地，移动电子设备可以为手机、平板电脑、笔记本电脑、掌上电脑、车载电子设备、可穿戴设备、超级移动个人计算机（ultra-mobile personal computer, UMPC）、上网本或者个人数字助理（personal digital assistant, PDA）等，非移动电子设备可以为服务器、网络附属存储器（Network Attached Storage, NAS）、个人计算机（personal computer, PC）、电视机（television, TV）、柜员机或者自助机等，本申请实施例不作具体限定。

本申请实施例中的文本处理装置可以为具有操作系统的装置。该操作系统可以为安卓（Android）操作系统，可以为 ios 操作系统，还可以为其他可能的操作系统，本申请实施例不作具体限定。

本申请实施例提供的文本处理装置能够实现图 1 和图 2 的方法实施例实现的各个过程，为避免重复，这里不再赘述。

可选的，如图 4 所示，本申请实施例还提供一种电子设备 500，包括处理器 501，存储器 502，存储在存储器 502 上并可在所述处理器 501 上运行的程序或指令，该程序或指令被处理器 501 执行时实现上述文本处理方法实施例的各个过程，且能达到相同的技术效果，为避免重复，这里不再赘述。

需要注意的是，本申请实施例中的电子设备包括上述所述的移动电子设备和非移动电子设备。

图 5 为实现本申请实施例的一种电子设备的硬件结构示意图。

该电子设备 100 包括但不限于：射频单元 101、网络模块 102、音频输出单元 103、输入单元 104、传感器 105、显示单元 106、用户输入单元 107、接口单元 108、存储器 109、以及处理器 110 等部件。

本领域技术人员可以理解，电子设备 100 还可以包括给各个部件供电的电源（比如电池），电源可以通过电源管理系统与处理器 110 逻辑相连，从而通过电源管理系统实现管理充电、放电、以及功耗管理等功能。图 5 中示出的电子设备结构并不构成对电子设备的限定，电子设备可以包括比图示更多或更少的部件，或者组合某些部件，或者不同的部件布置，在此不再赘述。

其中，处理器 110，用于获取第一文本，该第一文本包括 M 个第一字符， M 为正整数；并根据每个第一字符的特征信息，确定每个第一字符的权重，任一第一字符的权重用于表征该任一第一字符在第一文本中的重要程度；以及从第二文本中删除目标字符，得到目标文本，该第二文本为第一文本的全部或部分文本，该目标字符为权重满足第一条件的第一字符。

可选地，上述第一文本包括至少一个第一字符串，该至少一个第一字符串

包括上述 M 个第一字符，一个第一字符串包括至少一个第一字符；处理器 110，还用于根据每个第一字符的权重，确定每个第一字符串的权重，任一第一字符串的权重用于表征该任一第一字符串在第一文本中的重要程度；以及从第一文本中删除目标字符串，得到第二文本，该目标字符串为权重满足第二条件的第一字符串。

可选地，处理器 110，具体用于将每个第一字符串对应的每个第一字符的权重进行线性变换和函数变换，确定每个第一字符串的权重。

可选地，处理器 110，还用于对待处理的第三文本执行分句处理，得到至少一个第二字符串；对至少一个第二字符串中的每个第二字符串执行分词处理，得到至少一个第二字符；获取至少一个第二字符中的每个第二字符的向量；根据每个第二字符串对应的每个第二字符的向量，确定每个第二字符串的向量；将每个第二字符串的向量进行线性变换和函数变换，确定每个第二字符串的权重，任一第二字符串的权重用于表征该任一第二字符串在第三文本中的重要程度；以及从第三文本中删除权重满足第三条件的第二字符串，得到第一文本。

可选地，特征信息包括以下至少一项：文本信息，历史使用信息，上下文信息，语言概率信息。

本申请实施例提供的电子设备，电子设备在获取包括 M 个第一字符的第一文本之后，可以根据每个第一字符的特征信息，确定每个第一字符的权重。然后，电子设备可以从第二文本中删除目标字符，得到目标文本。其中，任一第一字符的权重用于表征该任一第一字符在第一文本中的重要程度，该第二文本为该第一文本的全部或部分文本，该目标字符为权重满足第一条件的第一字符。通过上述方案，在电子设备对长文本（如第一文本）进行搜索的情况下，电子设备可以根据第一文本中每个第一字符的特征信息，确定每个第一字符的权重。然后，电子设备可以删除权重满足第一条件的第一字符，从而可以以词为单位精细地简化第一文本，保留第一文本中的重要字符。如此，使得电子设备在对长文本进行搜索的过程中，可以仅搜索第一文本中的重要字符，进而能够在提高搜索效率的同时提高搜索正确率。

本实施例中各种实现方式具有的有益效果具体可以参见上述方法实施例中相应实现方式所具有的有益效果，为避免重复，此处不再赘述。

应理解的是，本申请实施例中，输入单元 104 可以包括图形处理器（Graphics Processing Unit, GPU）1041 和麦克风 1042，图形处理器 1041 对在

视频捕获模式或图像捕获模式中由图像捕获装置（如摄像头）获得的静态图片或视频的图像数据进行处理。显示单元 106 可包括显示面板 1061，可以采用液晶显示器、有机发光二极管等形式来配置显示面板 1061。用户输入单元 107 包括触控面板 1071 以及其他输入设备 1072。触控面板 1071，也称为触摸屏。

- 5 触控面板 1071 可包括触摸检测装置和触摸控制器两个部分。其他输入设备 1072 可以包括但不限于物理键盘、功能键（比如音量控制按键、开关按键等）、轨迹球、鼠标、操作杆，在此不再赘述。存储器 109 可用于存储软件程序以及各种数据，包括但不限于应用程序和操作系统。处理器 110 可集成应用处理器和调制解调处理器，其中，应用处理器主要处理操作系统、用户界面和应用程序等，调制解调处理器主要处理无线通信。可以理解的是，上述调制解调处理器也可以不集成到处理器 110 中。

本申请实施例还提供一种可读存储介质，所述可读存储介质上存储有程序或指令，该程序或指令被处理器执行时实现上述文本处理方法实施例的各个过程，且能达到相同的技术效果，为避免重复，这里不再赘述。

- 15 其中，所述处理器为上述实施例中所述的电子设备中的处理器。所述可读存储介质，包括计算机可读存储介质，如计算机只读存储器（Read-Only Memory, ROM）、随机存取存储器（Random Access Memory, RAM）、磁碟或者光盘等。

- 20 本申请实施例另提供了一种芯片，所述芯片包括处理器和通信接口，所述通信接口和所述处理器耦合，所述处理器用于运行程序或指令，实现上述文本处理方法实施例的各个过程，且能达到相同的技术效果，为避免重复，这里不再赘述。

应理解，本申请实施例提到的芯片还可以称为系统级芯片、系统芯片、芯片系统或片上系统芯片等。

- 25 需要说明的是，在本文中，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、物品或者装置不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、物品或者装置所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括该要素的过程、方法、物品或者装置中还存在另外的相同要素。此外，需要指出的是，本申请实施方式中的方法和装置的范围不限按示出或讨论的顺序来执行功能，还可包括根据所涉及的功能按基本同时的方式或按相反的顺序来执行功能，例如，可以按不同
- 30

于所描述的次序来执行所描述的方法，并且还可以添加、省去、或组合各种步骤。另外，参照某些示例所描述的特征可在其他示例中被组合。

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。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分可以以计算机软件产品的形式体现出来，该计算机软件产品存储在一个存储介质（如 ROM/RAM、磁碟、光盘）中，包括若干指令用以使得一台终端（可以是手机，计算机，服务器，或者网络设备）执行本申请各个实施例所述的方法。

上面结合附图对本申请的实施例进行了描述，但是本申请并不局限于上述的具体实施方式，上述的具体实施方式仅仅是示意性的，而不是限制性的，本领域的普通技术人员在本申请的启示下，在不脱离本申请宗旨和权利要求所保护的范围情况下，还可做出很多形式，均属于本申请的保护之内。

权 利 要 求 书

1. 一种文本处理方法，所述方法包括：

获取第一文本，所述第一文本包括 M 个第一字符， M 为正整数；

5 根据每个第一字符的特征信息，确定所述每个第一字符的权重，任一第一字符的权重用于表征所述任一第一字符在所述第一文本中的重要程度；

从第二文本中删除目标字符，得到目标文本，所述第二文本为所述第一文本的全部或部分文本，所述目标字符为权重满足第一条件的第一字符。

2. 根据权利要求 1 所述的方法，其中，所述第一文本包括至少一个第一字符串，所述至少一个第一字符串包括所述 M 个第一字符，一个第一字符串
10 包括至少一个第一字符；

在从第二文本中删除目标字符之前，所述方法还包括：

根据所述每个第一字符的权重，确定所述每个第一字符串的权重，任一第一字符串的权重用于表征所述任一第一字符串在所述第一文本中的重要程度；

15 从所述第一文本中删除目标字符串，得到所述第二文本，所述目标字符串为权重满足第二条件的第一字符串。

3. 根据权利要求 2 所述的方法，其中，所述根据所述每个第一字符的权重，确定所述每个第一字符串的权重，包括：

将每个第一字符串对应的每个第一字符的权重进行线性变换和函数变换，确定所述每个第一字符串的权重。

20 4. 根据权利要求 1 至 3 任一项所述的方法，其中，所述获取第一文本之前，所述方法还包括：

对待处理的第三文本执行分句处理，得到至少一个第二字符串；

对所述至少一个第二字符串中的每个第二字符串执行分词处理，得到至少一个第二字符；

25 获取所述至少一个第二字符串中的每个第二字符的向量；

根据所述每个第二字符串对应的每个第二字符的向量，确定所述每个第二字符串的向量；

30 将所述每个第二字符串的向量进行线性变换和函数变换，确定所述每个第二字符串的权重，任一第二字符串的权重用于表征所述任一第二字符串在所述第三文本中的重要程度；

从所述第三文本中删除权重满足第三条件的第二字符串，得到所述第一文本。

5. 根据权利要求 1 所述的方法, 其中, 所述特征信息包括以下至少一项: 文本信息, 历史使用信息, 上下文信息, 语言概率信息。

6. 一种文本处理装置, 所述文本处理装置包括: 获取模块、确定模块和执行模块;

5 所述获取模块, 用于获取第一文本, 所述第一文本包括 M 个第一字符, M 为正整数;

所述确定模块, 用于根据每个第一字符的特征信息, 确定所述每个第一字符的权重, 任一第一字符的权重用于表征所述任一第一字符在所述第一文本中的重要程度;

10 所述执行模块, 用于从第二文本中删除目标字符, 得到目标文本, 所述第二文本为所述第一文本的全部或部分文本, 所述目标字符为权重满足第一条件的第一字符。

7. 根据权利要求 6 所述的文本处理装置, 其中, 所述第一文本包括至少一个第一字符串, 所述至少一个第一字符串包括所述 M 个第一字符, 一个第一字符串包括至少一个第一字符;

所述确定模块, 还用于根据所述每个第一字符的权重, 确定所述每个第一字符串的权重, 任一第一字符串的权重用于表征所述任一第一字符串在所述第一文本中的重要程度;

20 所述执行模块, 还用于从所述第一文本中删除目标字符串, 得到所述第二文本, 所述目标字符串为权重满足第二条件的第一字符串。

8. 根据权利要求 7 所述的文本处理装置, 其中, 所述确定模块, 具体用于将每个第一字符串对应的每个第一字符的权重进行线性变换和函数变换, 确定所述每个第一字符串的权重。

9. 根据权利要求 6 至 8 任一项所述的文本处理装置, 其中,

25 所述执行模块, 还用于对待处理的第三文本执行分句处理, 得到至少一个第二字符串; 以及对所述至少一个第二字符串中的每个第二字符串执行分词处理, 得到至少一个第二字符;

所述获取模块, 还用于获取所述至少一个第二字符串中的每个第二字符的向量;

30 所述确定模块, 还用于根据所述获取模块获取的所述每个第二字符串对应的每个第二字符的向量, 确定所述每个第二字符串的向量; 以及将所述每个第二字符串的向量进行线性变换和函数变换, 确定所述每个第二字符串的权重,

任一第二字符串的权重用于表征所述任一第二字符串在所述第三文本中的重要程度；

所述执行模块，还用于从所述第三文本中删除权重满足第三条件的第二字符串，得到所述第一文本。

- 5 10. 根据权利要求 6 所述的文本处理装置，其中，所述特征信息包括以下至少一项：文本信息、历史使用信息、上下文信息和语言概率信息。
11. 一种电子设备，包括处理器，存储器及存储在所述存储器上并可在所述处理器上运行的程序或指令，所述程序或指令被所述处理器执行时实现如权利要求 1 至 5 中任一项所述的文本处理方法的步骤。
- 10 12. 一种可读存储介质，所述可读存储介质上存储程序或指令，所述程序或指令被处理器执行时实现如权利要求 1 至 5 中任一项所述的文本处理方法的步骤。
13. 一种计算机程序产品，所述计算机程序产品被至少一个处理器执行以实现如权利要求 1-5 任一项所述的文本处理方法的步骤。
- 15 14. 一种芯片，所述芯片包括处理器和通信接口，所述通信接口和所述处理器耦合，所述处理器用于运行程序或指令，实现如权利要求 1-5 任一项所述的文本处理方法的步骤。
15. 一种电子设备，所述电子设备被配置成用于执行如权利要求 1-5 任一项所述的文本处理方法的步骤。

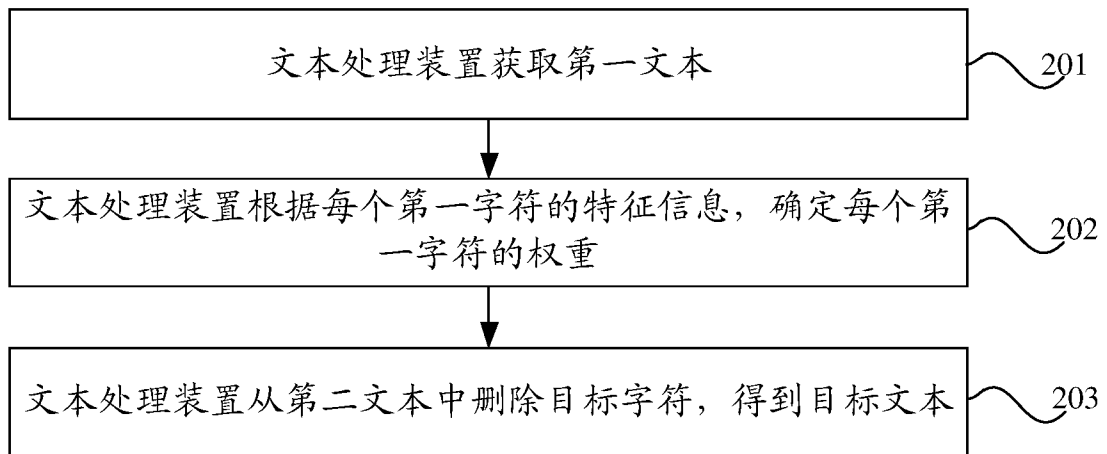


图 1

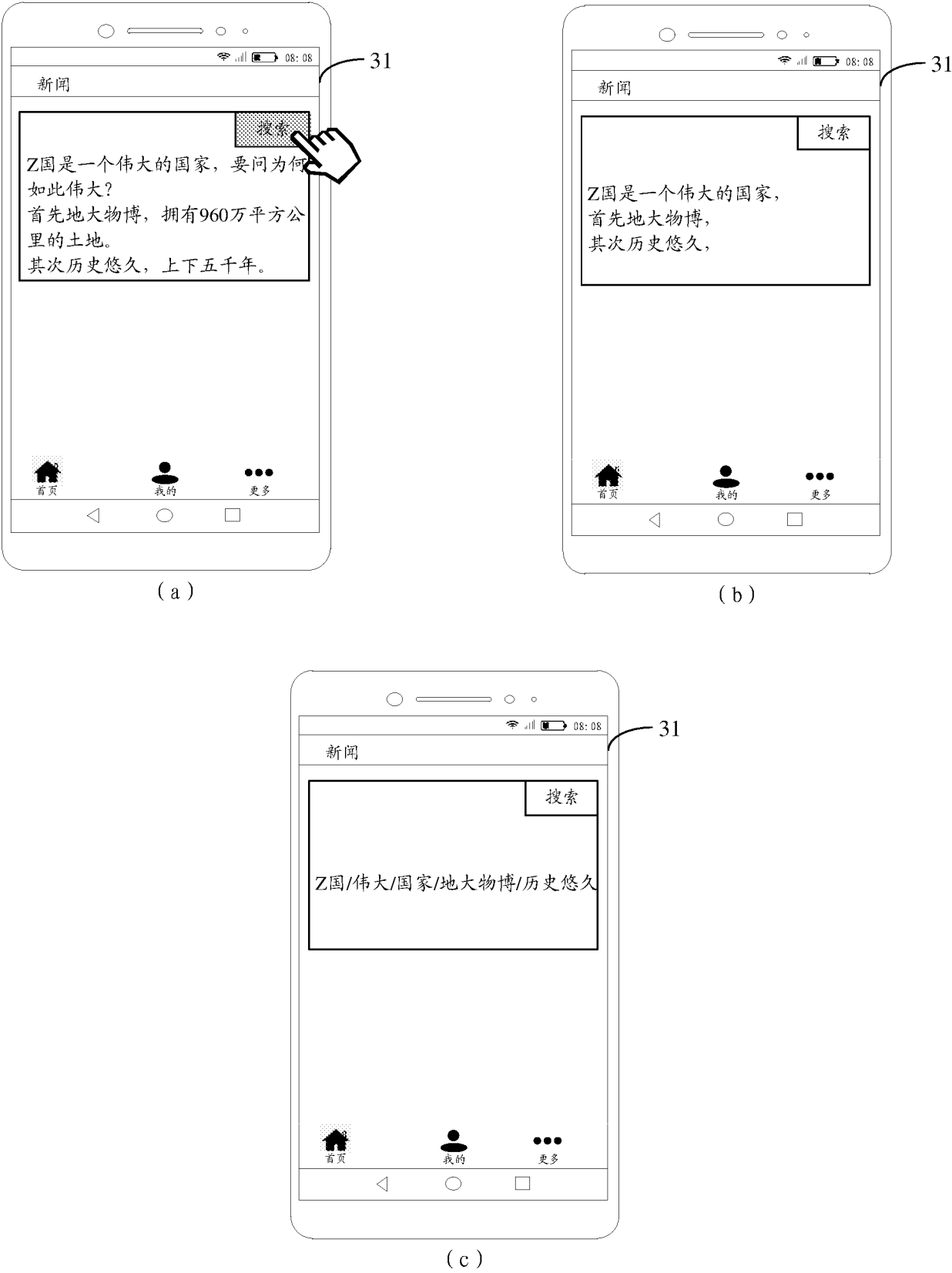


图 2

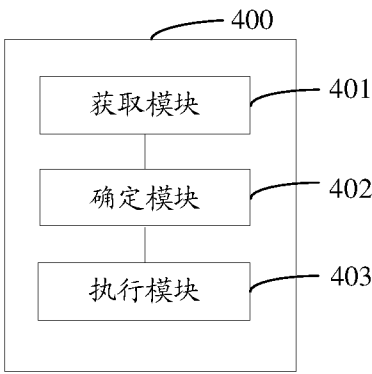


图 3

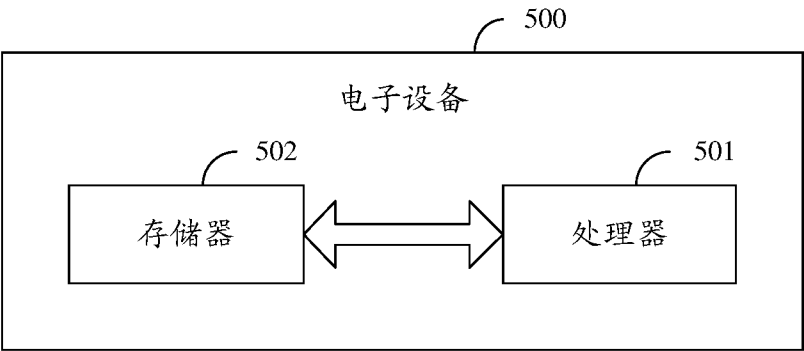


图 4

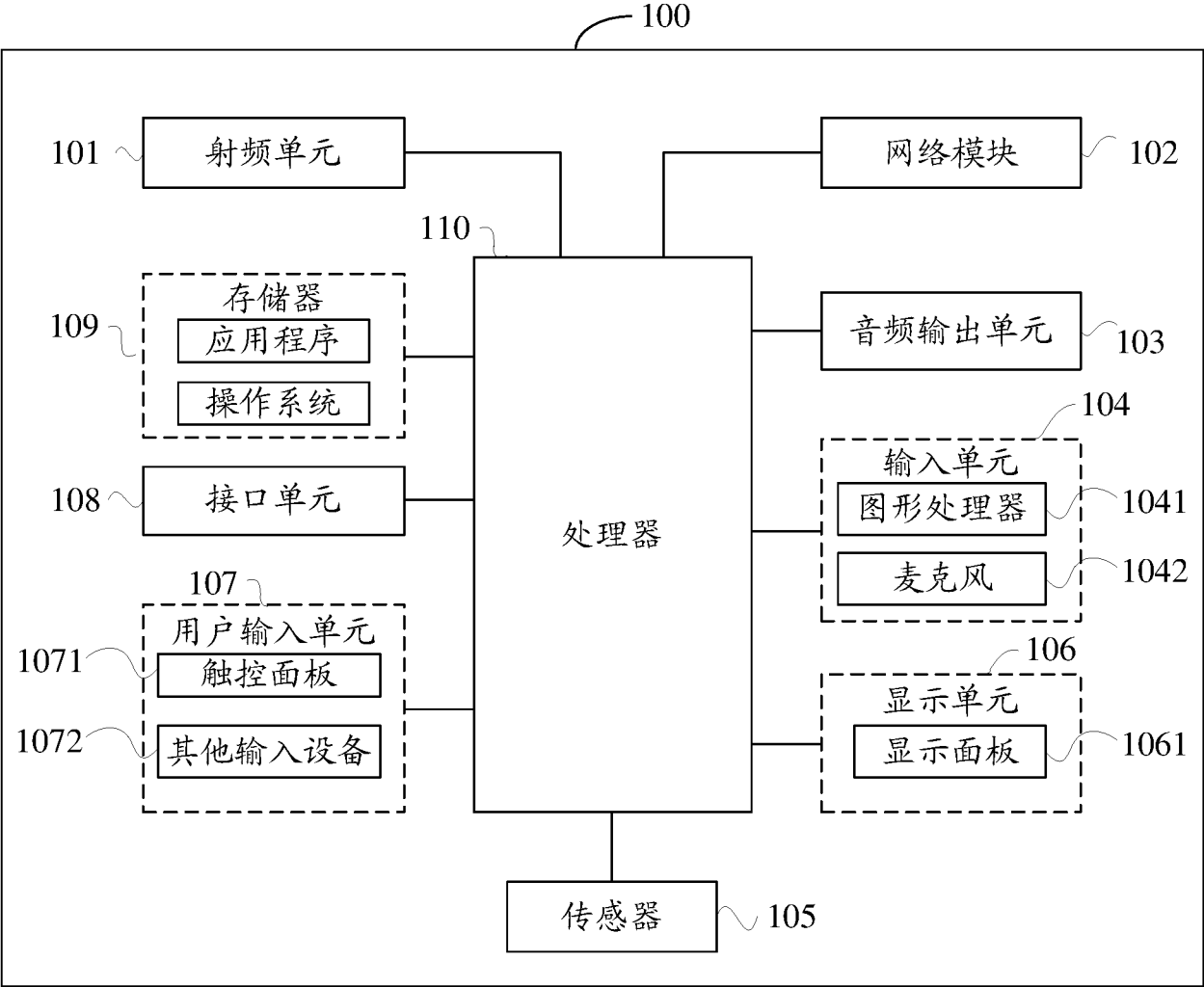


图 5

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2022/095648

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/332(2019.01)i; G06F 40/216(2020.01)n; G06F 40/289(2020.01)n; G06N 3/04(2006.01)n

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F; G06N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNTXT; WPABSC; WPABS; ENTXT; DWPI; ENTXTC; VEN; CNKI; IEEE; 冗余, 文本, 删除, 删减, 过滤, 滤除, 权重, 重要度, 重要性, 重要程度, 权值, 分词, 搜索, 检索, 词权重模型, 语句, 字符, 词语, 长文本, long, text, search, word, weight, importance, keyword

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 113360613 A (VIVO COMMUNICATION TECHNOLOGY CO., LTD.) 07 September 2021 (2021-09-07) claims 1-12, and description, paragraphs 20-153	1-15
X	CN 112711695 A (NETEASE MEDIA TECHNOLOGY (BEIJING) CO., LTD.) 27 April 2021 (2021-04-27) claims 1-10, and description, paragraphs 76-181	1-15
A	CN 110232156 A (TRICORN (BEIJING) TECHNOLOGY CO., LTD.) 13 September 2019 (2019-09-13) entire document	1-15
A	CN 111680510 A (TENCENT TECHNOLOGY SHENZHEN CO., LTD.) 18 September 2020 (2020-09-18) entire document	1-15
A	CN 112784009 A (BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS et al.) 11 May 2021 (2021-05-11) entire document	1-15



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

15 July 2022

Date of mailing of the international search report

04 August 2022

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/
CN)
No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing
100088, China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2022/095648**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 110704621 A (BEIJING DA MI TECHNOLOGY CO., LTD.) 17 January 2020 (2020-01-17) entire document	1-15
A	US 2020050638 A1 (HANCOCK PARKER DOUGLAS) 13 February 2020 (2020-02-13) entire document	1-15

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2022/095648

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN	113360613	A	07 September 2021	None	
CN	112711695	A	27 April 2021	None	
CN	110232156	A	13 September 2019	None	
CN	111680510	A	18 September 2020	CN 111680510 B	15 October 2021
CN	112784009	A	11 May 2021	None	
CN	110704621	A	17 January 2020	None	
US	2020050638	A1	13 February 2020	None	

国际检索报告

国际申请号

PCT/CN2022/095648

A. 主题的分类

G06F 16/332(2019.01)i; G06F 40/216(2020.01)n; G06F 40/289(2020.01)n; G06N 3/04(2006.01)n

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06F; G06N

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用))

CNTXT;WPABSC;WPABS;ENTXT;DWPI;ENTXTC;VEN;CNKI;IEEE:冗余, 文本, 删除, 删减, 过滤, 滤除, 权重, 重要度, 重要性, 重要程度, 权值, 分词, 搜索, 检索, 词权重模型, 语句, 字符, 词语, 长文本, long, text, search, word, weight, importance, keyword

C. 相关文件

类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
PX	CN 113360613 A (维沃移动通信有限公司) 2021年9月7日 (2021 - 09 - 07) 权利要求1-12, 说明书第20-153段	1-15
X	CN 112711695 A (网易传媒科技北京有限公司) 2021年4月27日 (2021 - 04 - 27) 权利要求1-10, 说明书第76-181段	1-15
A	CN 110232156 A (三角兽北京科技有限公司) 2019年9月13日 (2019 - 09 - 13) 全文	1-15
A	CN 111680510 A (腾讯科技深圳有限公司) 2020年9月18日 (2020 - 09 - 18) 全文	1-15
A	CN 112784009 A (北京邮电大学等) 2021年5月11日 (2021 - 05 - 11) 全文	1-15
A	CN 110704621 A (北京大米科技有限公司) 2020年1月17日 (2020 - 01 - 17) 全文	1-15
A	US 2020050638 A1 (HANC-I HANCOCK P D) 2020年2月13日 (2020 - 02 - 13) 全文	1-15

☐ 其余文件在C栏的续页中列出。☒ 见同族专利附件。

* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件(如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期

2022年7月15日

国际检索报告邮寄日期

2022年8月4日

ISA/CN的名称和邮寄地址

中国国家知识产权局(ISA/CN)

中国北京市海淀区蓟门桥西土城路6号 100088

传真号 (86-10)62019451

受权官员

陈响

电话号码 (86-27)59371802

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2022/095648

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	113360613	A	2021年9月7日	无	
CN	112711695	A	2021年4月27日	无	
CN	110232156	A	2019年9月13日	无	
CN	111680510	A	2020年9月18日	CN 111680510	B 2021年10月15日
CN	112784009	A	2021年5月11日	无	
CN	110704621	A	2020年1月17日	无	
US	2020050638	A1	2020年2月13日	无	