

(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
2018 年 1 月 18 日 (18.01.2018)



(10) 国际公布号
WO 2018/010579 A1

(51) 国际专利分类号:
G06F 17/27 (2006.01)

(21) 国际申请号: PCT/CN2017/091783

(22) 国际申请日: 2017 年 7 月 5 日 (05.07.2017)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201610552115.0 2016 年 7 月 13 日 (13.07.2016) CN

(71) 申请人: 阿里巴巴集团控股有限公司 (ALIBABA GROUP HOLDING LIMITED) [—/CN]; 开曼群岛
大开曼资本大厦一座四层 847 号邮箱,
Grand Cayman (KY)。

(72) 发明人: 及

(71) 申请人 (仅对 US): 张增明 (ZHANG, Zengming)
[CN/CN]; 中国浙江省杭州市余杭区文一
西路 969 号 3 号楼 5 楼阿里巴巴集团法务
部, Zhejiang 311121 (CN)。

(74) 代理人: 北京国昊天诚知识产权代理有限公司 (CO-HORIZON INTELLECTUAL PROPERTY INC.); 中
国北京市朝阳区小关北里甲 2 号渔阳置业
大厦 B 座 605, Beijing 100029 (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, IR, IS,
JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK,
LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX,

(54) Title: CHARACTER STRING SEGMENTATION METHOD, APPARATUS AND DEVICE

(54) 发明名称: 字符串的分词方法、装置及设备

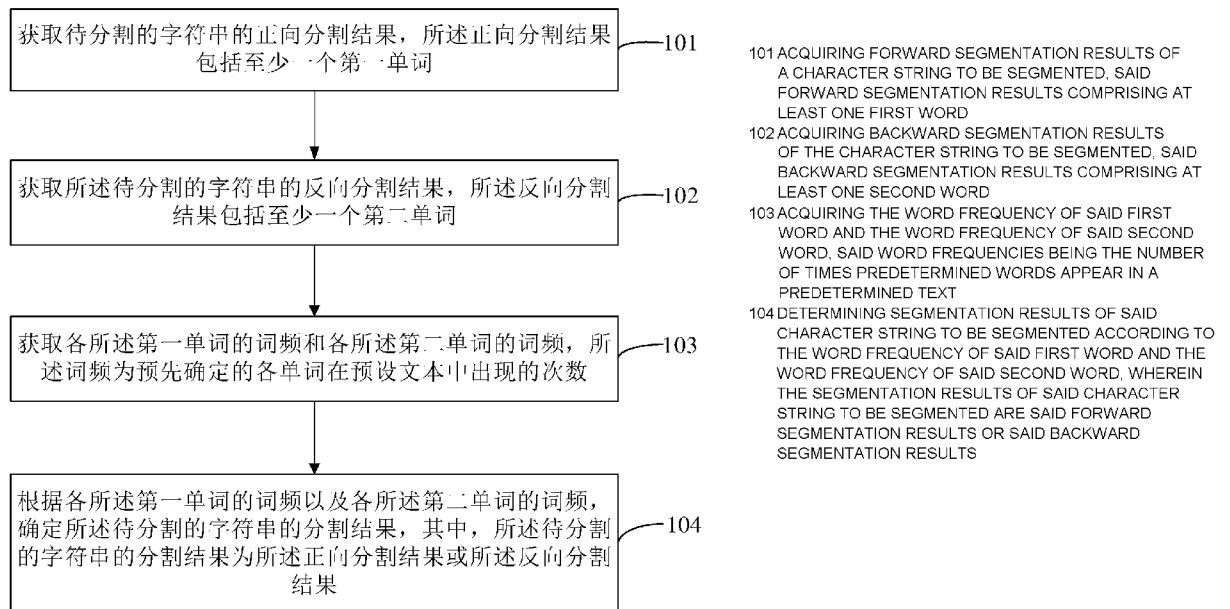


图 2

(57) Abstract: A character string segmentation method, apparatus and device, said method comprising: acquiring forward segmentation results of a character string to be segmented, said forward segmentation results comprising at least one first word (101); acquiring backward segmentation results of the character string to be segmented, said backward segmentation results comprising at least one second word (102); acquiring a word frequency of said first word and a word frequency of said second word, said word frequencies being the number of times predetermined words appear in a predetermined text (103); determining segmentation results of said character string to be segmented according to the word frequency of said first word and the word frequency of said second word, wherein the

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,
SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG,
US, UZ, VC, VN, 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))。

segmentation results of said character string to be segmented are said forward segmentation results or said backward segmentation results (104). Not only does the present invention improve the success rate of segmentation, but also increases the probability of the semantics of words being accurate within the segmentation results.

(57) 摘要: 一种字符串的分词方法、装置及设备, 该方法包括: 获取待分割的字符串的正向分割结果, 所述正向分割结果包括至少一个第一单词 (101); 获取所述待分割的字符串的反向分割结果, 所述反向分割结果包括至少一个第二单词 (102); 获取各所述第一单词的词频和各所述第二单词的词频, 所述词频为预先确定的各单词在预设文本中出现的次数 (103); 根据各所述第一单词的词频以及各所述第二单词的词频, 确定所述待分割的字符串的分割结果, 其中, 所述待分割的字符串的分割结果为所述正向分割结果或所述反向分割结果 (104)。不仅提高了分割成功率, 还提高了分割结果中的各单词语义正确的概率。

字符串的分词方法、装置及设备

技术领域

本申请涉及计算机技术，尤其涉及一种字符串的分词方法、装置及设备。

5

背景技术

自然语言处理是运用计算机对自然语言进行分析和理解，从而使计算机在某种程度上具有人的语言能力。在对英文文本进行自然语言处理时，经常会遇到不符合自然语言规则的脏数据，导致自然语言处理效果大打折扣。因此，需要先对英文文本进行分词预处理，得到包含多个英文单词的正常自然语言，然后再使用自然语言模型进行处理。

现有技术中的脏数据主要包括因空格字符缺失造成多个单词连在一起形成的字符串、掺杂有干扰字符的字符串等。现有技术对英文文本进行分词的具体过程如下：按顺序依次读取待分割的字符串的一个字母，添加到已经取得的字母们的后面，组成一个子字符串，然后检查此子字符串是否能在预先获取的英文词典中查到。如果能查到，则说明该子字符串是一个单词，先将其从原字符串中分割出来。然后对剩下的字符串重复使用这种方法，最终完成单词分割，或者剩下的字符串没法分割直接输出。

然而，现有技术对英文文本进行分词的方法，在待分割的字符串中前一单词与后一单词的前缀组成单词、或掺杂有干扰字符等情况下，会出现分割不当导致语义错误、甚至无法分割的现象。

发明内容

本发明提供一种字符串的分词方法、装置及设备，不仅提高了分割成功率，还提高了分割结果中的各单词语义正确的概率。

第一方面，本发明提供一种字符串的分词方法，包括：

获取待分割的字符串的正向分割结果，所述正向分割结果包括至少一个第一单词；

获取所述待分割的字符串的反向分割结果，所述反向分割结果包括至少

一个第二单词;

获取各所述第一单词的词频和各所述第二单词的词频, 所述词频为预先确定的各单词在预设文本中出现的次数;

根据各所述第一单词的词频以及各所述第二单词的词频, 确定所述待分割的字符串的分割结果, 其中, 所述待分割的字符串的分割结果为所述正向分割结果或所述反向分割结果。

作为一种可实现的方式, 所述获取待分割的字符串的正向分割结果, 包括:

对所述待分割的字符串进行正向分割的操作, 判断是否获取到第一单词;

若是, 将除去所述第一单词的待分割的字符串作为新的待分割的字符串, 并返回执行对待分割的字符串进行正向分割的操作;

若否, 对所述待分割的字符串的正向的首字符进行删除处理, 得到处理后的待分割的字符串, 将处理后的待分割的字符串作为新的待分割的字符串, 并返回执行对待分割的字符串进行正向分割的操作;

重复执行对所述待分割的字符串进行正向分割的操作, 直至对所述待分割的字符串分割结束, 得到正向分割结果。

本实施例提供的正向分割方法, 为一层一层的正向递进式分割方式, 经过一层一层的尝试, 克服了干扰字符, 最终得到了正向分割结果。

作为一种可实现的方式, 所述获取待分割的字符串反向分割结果, 包括:

对所述待分割的字符串进行反向分割的操作, 判断是否获取到第二单词; 若是, 将除去所述第二单词的待分割的字符串作为新的待分割的字符串, 并返回执行对待分割的字符串进行反向分割的操作;

若否, 对所述待分割的字符串的反向的首字符进行删除处理, 得到处理后的待分割的字符串, 将处理后的待分割的字符串作为新的待分割的字符串, 并返回执行对待分割的字符串进行反向分割的操作;

重复执行对所述待分割的字符串进行正向分割的操作, 直至对所述待分割的字符串分割结束, 得到反向分割结果。

本实施例提供的反向分割方法, 为一层一层的反向递进式分割方式, 经过一层一层的尝试, 克服了干扰字符, 最终得到了反向分割结果。

作为一种可实现的方式, 还包括:

获取待分割的文本，对所述待分割的文本进行符号删除操作，得到所述待分割的字符串。

作为一种可实现的方式，还包括：

构建正向字典树和反向字典树；

5 所述对所述待分割的字符串进行正向分割的操作，包括：

根据所述正向字典树，对所述待分割的字符串进行正向分割的操作；

所述对所述待分割的字符串进行反向分割的操作，包括：

根据所述反向字典树，对所述待分割的字符串进行反向分割的操作。

本实施例根据字典树来对字符串进行正向分割或反向分割，由于公共查找
10 路径的存在，可以在读取到的子字符串增加一个字符后，基于该字符增加前的
查找路径继续向下一级节点查找，从而可以避免重复查找，最大限度地减少无
谓的字符串比较，减少查询时间，提高查找效率。

作为一种可实现的方式，所述正向字典树的每个第一节点中存储有所述
第一节点对应的单词的词频，所述反向字典树的每个第二节点中存储有所述
15 第二节点对应的单词的词频；

所述获取各所述第一单词的词频和各所述第二单词的词频，包括：

从所述第一单词对应的第一节点中获取所述第一单词的词频；

从所述第二单词对应的第二节点中获取所述第二单词的词频。

作为一种可实现的方式，所述构建正向字典树和反向字典树之前，还包
20 括：

构建语料库，所述语料库包括单词库和所述单词库中的单词的词频；

所述构建正向字典树和反向字典树，包括：

根据所述语料库，构建正向字典树和反向字典树，并将各单词的词频存
储至对应的第一节点和第二节点。

25 作为一种可实现的方式，所述预设文本包括：满足预设使用条件的文本
以及待分割的文本；所述构建语料库，包括：

根据满足预设使用条件的词典，得到单词库；

确定所述单词库中的单词在所述满足预设使用条件的文本以及所述待分
割的文本中出现的次数；

30 根据所述单词库、所述单词库中的单词在所述满足预设使用条件的文本

以及所述待分割的文本中出现的次数，构建所述语料库。

作为一种可实现的方式，所述确定单词库中的单词在所述待分割的文本中出现的次数，包括：

根据所述待分割的文本中的空格符，获取至少一个第一字符串；

- 5 将所述至少一个第一字符串与所述单词库中的单词进行匹配，得到与所述单词库中的单词匹配的至少一个第二字符串；

根据各所述第二字符串在所述待分割的文本中出现的次数，确定单词库中的单词在所述待分割的文本中出现的次数。

- 10 本实施例构建的语料库，该语料库中的单词的词频是通过待分割的文本进行修正的，与待分割的文本具有一定的相关性，使得语料库中的单词的词频更接近待分割的文本的应用情况，从而可以使得分割结果的语义与待分割的文本表达的语义更接近，提高了字符串分割的正确性。

作为一种可实现的方式，所述根据各所述第一单词的词频以及各所述第二单词的词频，确定所述待分割的字符串的分割结果，包括：

- 15 对所有所述第一单词的词频进行求和处理，得到第一词频和值；

对所有所述第二单词的词频进行求和处理，得到第二词频和值；

若所述第一词频和值大于所述第二词频和值，则确定所述待分割的字符串的分割结果为正向分割结果；

- 20 若所述第二词频和值大于所述第一词频和值，则确定所述待分割的字符串的分割结果为反向分割结果。

作为一种可实现的方式，所述正向分割和所述反向分割均采用最长单词分割方式。

第二方面，本发明提供一种字符串的分词方法，包括：

- 25 向云端服务器发送用户输入的待分割的文本，以使所述云端服务器获取待分割的字符串，并根据正向分割结果中的各第一单词的词频以及反向分割结果中的各第二单词的词频确定分割结果；

接收所述云端服务器反馈的所述待分割的字符串的分割结果信息，所述分割结果信息包括所述待分割的字符串的分割结果；其中，所述待分割的字符串的分割结果为所述正向分割结果或所述反向分割结果；

- 30 向用户输出所述分割结果。

本实施例提供的字符串的分词方法，通过向云端服务器发送用户输入的待分割的文本，以使云端服务器获取待分割的字符串，并根据正向分割结果中的各第一单词的词频以及反向分割结果中的各第二单词的词频确定分割结果；由于云端服务器通过双向分割字符串，可以识别字符串头或字符串尾的干扰字符，提高了分割成功率，基于词频来确定最终的分割结果，提高了分割结果中的各单词语义正确的概率，接收云端服务器反馈的待分割的字符串的分割结果信息，分割结果信息包括待分割的字符串的分割结果；向用户输出分割结果，用户可以获知分割结果，使得用户可以获知最终的查询结果对应的查询单词，提高了用户的体验。

10 作为一种可实现的方式，所述向用户输出所述分割结果，包括：
 在显示界面上显示所述分割结果。

 作为一种可实现的方式，所述分割结果信息中还包括所述分割结果对应的分割类型，所述分割类型为正向分割或反向分割；

 所述在显示界面上显示所述分割结果，包括：
15 在显示界面上显示所述分割结果以及所述分割结果的分割类型。

 作为一种可实现的方式，若所述分割结果为正向分割结果，则所述分割信息中还包括反向分割结果；或者

 若所述分割结果为反向分割结果，则所述分割信息中还包括正向分割结果；

20 所述在显示界面上显示所述分割结果，包括：

 在所述显示界面上显示所述正向分割结果和所述反向分割结果，并标注所述待分割字符串对应的分割结果。

 作为一种可实现的方式，所述分割信息中还包括所述正向分割结果中的各所述第一单词的词频和所述反向分割结果中的各所述第二单词的词频；

25 在所述显示界面上显示所述正向分割结果和所述反向分割结果，并标注所述待分割字符串对应的分割结果之后，还包括：

 获取所述用户操作所述显示界面触发的词频显示指令；

 根据所述词频显示指令，显示各所述第一单词的词频和/或各所述第二单词的词频；

30 或者

在所述显示界面上显示所述正向分割结果和所述反向分割结果，包括：

在所述显示界面上显示所述正向分割结果、所述正向分割结果中的第一单词的词频，以及所述反向分割结果、所述反向分割结果中的第二单词的词频。

- 5 作为一种可实现的方式，所述分割信息中还包括所述正向分割结果中的各所述第一单词对应的第一词频和值以及所述反向分割结果中的各所述第二单词对应的第二词频和值；

在所述显示界面上显示所述正向分割结果和所述反向分割结果，并标注所述待分割字符串对应的分割结果之后，还包括：

- 10 获取所述用户操作所述显示界面触发的词频显示指令；
根据所述词频显示指令，显示所述第一词频和值和/或所述第二词频和值；

或者

在所述显示界面上显示所述正向分割结果和所述反向分割结果，包括：

- 15 在所述显示界面上显示所述正向分割结果、所述第一词频和值，以及所述反向分割结果、所述第二词频和值。

作为一种可实现的方式，所述在所述显示界面上显示所述正向分割结果和所述反向分割结果之后，还包括：

- 20 获取所述用户对所述显示界面上的所述正向分割结果或反向分割结果的操作信息，

根据所述操作信息确定待处理的分割结果；

向所述云端服务器发送所述待处理的分割结果，以使所述云端服务器对所述待处理的分割结果进行自然语言处理。

第三方面，本发明提供一种字符串的分词装置，包括：

- 25 第一分割模块，用于获取待分割的字符串的正向分割结果，所述正向分割结果包括至少一个第一单词；

第二分割模块，用于获取所述待分割的字符串的反向分割结果，所述反向分割结果包括至少一个第二单词；

- 30 词频获取模块，用于获取各所述第一单词的词频和各所述第二单词的词频，所述词频为预先确定的各单词在预设文本中出现的次数；

结果确定模块，用于根据各所述第一单词的词频以及各所述第二单词的词频，确定所述待分割的字符串的分割结果，其中，所述待分割的字符串的分割结果为所述正向分割结果或所述反向分割结果。

第四方面，本发明提供一种字符串的分词装置，包括：

- 5 发送模块，用于向云端服务器发送用户输入的待分割的文本，以使所述云端服务器获取待分割的字符串，并根据正向分割结果中的各所述第一单词的词频以及反向分割结果中的各所述第二单词的词频确定分割结果；

接收模块，用于接收所述云端服务器反馈的所述待分割的字符串的分割结果信息，所述分割结果信息包括所述待分割的字符串的分割结果；其中，
10 所述待分割的字符串的分割结果为所述正向分割结果或所述反向分割结果；

输出模块，用于向用户输出所述分割结果。

第五方面，本发明提供一种字符串的分词设备，包括：包括：

输入设备，用于获取待分割的文本；

- 处理器，耦合至所述输入设备，用于获取待分割的字符串的正向分割结果，所述正向分割结果包括至少一个第一单词，并获取所述待分割的字符串的反向分割结果，所述反向分割结果包括至少一个第二单词；获取各所述第一单词的词频和各所述第二单词的词频，所述词频为预先确定的各单词在预设文本中出现的次数；根据各所述第一单词的词频以及各所述第二单词的词频，确定所述待分割的字符串的分割结果，其中，所述待分割的字符串的分割结果为所述正向分割结果或所述反向分割结果。
15

第六方面，本发明提供一种云端服务器，包括：

输入设备，用于获取待分割的文本；

- 处理器，耦合至所述输入设备，用于获取待分割的字符串的正向分割结果，所述正向分割结果包括至少一个第一单词，并获取所述待分割的字符串的反向分割结果，所述反向分割结果包括至少一个第二单词；获取各所述第一单词的词频和各所述第二单词的词频，所述词频为预先确定的各单词在预设文本中出现的次数；根据各所述第一单词的词频以及各所述第二单词的词频，确定所述待分割的字符串的分割结果，其中，所述待分割的字符串的分割结果为所述正向分割结果或所述反向分割结果。
25

- 30 第七方面，本发明提供一种字符串的分词设备，包括：

输出设备，用于向云端服务器发送用户输入的待分割的文本，以使所述云端服务器获取待分割的字符串，并根据正向分割结果中的各第一单词的词频以及反向分割结果中的各第二单词的词频确定分割结果；

输入设备，用于接收所述云端服务器反馈的所述待分割的字符串的分割结果信息，所述分割结果信息包括所述待分割的字符串的分割结果；其中，所述待分割的字符串的分割结果为所述正向分割结果或所述反向分割结果；

处理器，耦合至所述输出设备和所述输入设备，用于根据所述分割结果信息，控制所述输入设备向用户输出所述分割结果。

第八方面，本发明提供一种用户设备，包括：

输出设备，用于向云端服务器发送用户输入的待分割的文本，以使所述云端服务器获取待分割的字符串，并根据正向分割结果中的各第一单词的词频以及反向分割结果中的各第二单词的词频确定分割结果；

输入设备，用于接收所述云端服务器反馈的所述待分割的字符串的分割结果信息，所述分割结果信息包括所述待分割的字符串的分割结果；其中，所述待分割的字符串的分割结果为所述正向分割结果或所述反向分割结果；

处理器，耦合至所述输出设备和所述输入设备，用于根据所述分割结果信息，控制所述输入设备向用户输出所述分割结果。。

本实施例通过获取包括至少一个第一单词的正向分割结果，并获取包括至少一个第二单词的反向分割结果，通过双向分割字符串，识别字符串头或字符串尾的干扰字符，提高了分割成功率，然后获取各第一单词的词频和各第二单词的词频，根据各第一单词的词频以及各第二单词的词频，确定待分割的字符串的分割结果，基于词频来确定最终的分割结果，提高了分割结果中的各单词语义正确的概率。

附图说明

为了更清楚地说明本发明实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作一简单地介绍，显而易见地，下面描述中的附图是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动性的前提下，还可以根据这些附图获得其他的附图。

图 1 为本发明一实施例提供的字符串的分词场景示意图；

- 图 2 为本发明一实施例提供的字符串的分词方法流程示意图；
- 图 3 为本发明一实施例提供的正向分割示意图；
- 图 4 为本发明一实施例提供的反向分割示意图；
- 图 5 为本发明一实施例提供的正向和反向分割示意图；
- 5 图 6 为本发明一实施例提供的正向分割示意图；
- 图 7 为本发明一实施例提供的反向分割示意图；
- 图 8 为本发明一实施例提供的正向字典树的示意图；
- 图 9 为本发明一实施例提供的反向字典树的示意图；
- 图 10 为本发明一实施例提供的字符串的分词方法流程示意图；
- 10 图 11 为本发明一实施例提供的字符串的分词方法流程示意图；
- 图 12 为本发明一实施例提供的字符串的分词方法的信令流程图；
- 图 13 为本发明一实施例提供的字符串的分词方法的显示界面示意图；
- 图 14 为本发明一实施例提供的字符串的分词方法的显示界面示意图；
- 图 15 为本发明一实施例提供的字符串的分词方法的显示界面示意图；
- 15 图 16 为本发明一实施例提供的字符串的分词方法的显示界面示意图；
- 图 17 为本发明一实施例提供的字符串的分词方法的显示界面示意图；
- 图 18 为本发明一实施例提供的字符串的分词方法的显示界面示意图；
- 图 19 为本发明一实施例提供的字符串的分词装置的结构示意图；
- 图 20 为本发明一实施例提供的字符串的分词装置的结构示意图；
- 20 图 21 为本发明一实施例提供的字符串的分词装置的结构示意图；
- 图 22 为本发明一实施例提供的字符串的分词装置的结构示意图；
- 图 23 为本发明一实施例提供的字符串的分词设备的硬件结构示意图；
- 图 24 为本发明一实施例提供的云端服务器的硬件结构示意图；
- 图 25 为本发明一实施例提供的字符串的分词设备的硬件结构示意图；
- 25 图 26 为本发明一实施例提供的用户设备的硬件结构示意图。

具体实施方式

这里将详细地对示例性实施例进行说明，其示例表示在附图中。下面的描述涉及附图时，除非另有表示，不同附图中的相同数字表示相同或相似

30 类似的要素。以下示例性实施例中所描述的实施方式并不代表与本发明相一

致的所有实施方式。相反，它们仅是与如所附权利要求书中所详述的、本发明的一些方面相一致的装置和方法的例子。

图 1 为本发明一实施例提供的字符串的分词场景示意图。如图 1 所示，用户通过用户设备 100 输入待分割的文本，对于用户而言，即用户输入的字符串，然后用户设备 100 将该待分割的文本发送给云端服务器 200。由于用户输入的字符串可能存在脏数据，因此，云端服务器 200 对用户输入的字符串进行分词处理。在一个具体的应用场景中，本实施例提供的字符串的分词方法可以应用到自然语言处理的处理过程中，该字符串的分词方法是对自然语言进行预处理，以得到包含多个语义正确的英文单词的自然语言，然后该自然语言作为自然语言模型的输入，由自然语言模型对该自然语言进行进一步的处理。例如，该自然语言模型可以为亮点词汇提取模型。在一个具体的应用场景中，用户设备 100 上可以安装电商平台对应的应用程序，也可以安装浏览器，用户可以通过该浏览器来浏览电商网站。当用户通过应用程序或电商网站来购买商品时，用户先查找商品，具体地，用户在电商平台对应的应用程序或电商网站的输入界面上输入字符串，然后用户设备 100 将该字符串发送给云端服务器 200。由于该字符串中可能存在脏数据，所以应用本发明提供的字符串的分词方法，云端服务器 200 对该字符串进行分词，得到多个英文单词，然后云端服务器 200 通过亮点词汇提取模型对该多个英文单词进行提取，获取商品的标题、属性等信息，即能够描述此商品的元素、风格等特征的亮点词汇，然后根据该亮点词汇向用户提供该用户需要的商品。可选地，云端服务器 200 在得到单词分割结果之后，还可以向用户设备反馈该单词分割结果，以使用户获知单词分割结果，从而得知云端服务器具体通过哪些单词来查找匹配商品。进一步还可以向用户设备反馈正向分割结果或者反向分割结果，由用户来选择单词分割结果，然后用户设备 100 向云端服务器 200 反馈用户选择的单词分割结果，云端服务器 200 根据用户选择的单词分割结果进行后续的处理。

本发明在此示出了一个具体的应用场景，在具体实现过程中，该字符串的分词方法还可以应用到网页搜索等场景中。或者，当用户设备，例如计算机、手机、平板等设备的处理功能比较强大时，还可以由用户设备来完成字符串的分词方法。对于本发明的字符串的分词方法的应用场景，本

实施例此处不做特别限制。下面首先采用详细的实施例，来说明上述云端服务器对字符串进行分词的方法。

图 2 为本发明一实施例提供的字符串的分词方法流程示意图。该字符串的分词方法可以由字符串的分词装置来实现。该装置可以通过软件和/或硬件来实现。该分词装置还可以被配置到云端服务器、计算机、手机、平板等设备中。该方法包括：

步骤 101、获取待分割的字符串的正向分割结果，所述正向分割结果包括至少一个第一单词；

步骤 102、获取所述待分割的字符串的反向分割结果，所述反向分割结果包括至少一个第二单词；

步骤 103、获取各所述第一单词的词频和各所述第二单词的词频，所述词频为预先确定的各单词在预设文本中出现的次数；

步骤 104、根据各所述第一单词的词频以及各所述第二单词的词频，确定所述待分割的字符串的分割结果，其中，所述待分割的字符串的分割结果为所述正向分割结果或所述反向分割结果。

在本实施例中，获取用户设备发送的待分割的文本，然后根据该待分割的文本获取待分割的字符串。然后对该待分割的字符串进行分词，本领域技术人员可以理解，该待分割的字符串为连续的不带任何符号的字符串。同时，用户输入的待分割的文本，若不带任何符号，则该待分割的文本即待分割的字符串。进一步地，若该待分割的文本中包括空格以及各种标点符号，则对待分割的文本进行符号删除操作，即删除空格和标点符号的操作，最终得到连续的待分割的字符串。

在获取到字符串之后，执行步骤 101 和步骤 102，通过对待分割的字符串分别进行正向分割和反向分割，获取正向分割结果和反向分割结果。本领域技术人员可以理解，本实施例中对待分割的字符串进行正向分割以获取正向分割结果，与对待分割的字符串进行反向分割以获取反向分割结果的过程，没有严格的时序关系。

下面以几个具体的例子，来说明对字符串进行正向分割和反向分割的过程。

一个具体的实施例，图 3 为本发明一实施例提供的正向分割示意图。

如图 3 所示, 本实施例对字符串 floorlengthsleevelessdressst 进行正向分割, 最终的正向分割结果为多个第一单词: floor length sleeveless dress。

具体的正向分割过程为: 从左到右取字符, 每取一次查一次词典, 来判断是否取到一个单词, 当取到 floor 时, 还会继续尝试 floorl、floorle、floorlen 直至取完整个字符串, 或者达到预设字符串长度, 该预设字符串长度为单词的最长长度, 然后在所有单词中, 取长度最长的单词作为分割结果, 由于后续没有单词, 则 floor 即为分割结果。

因此, 本领域技术人员可以理解, sleeveless 的长度为 10, 而 sleeve 的长度为 6, 则 sleeveless 为分割结果, 而 sleeve 以及 less 却不是最终的分割结果。本实施例采用单词最长的分割方式, 最能符合语义。一般情况下两个单词写在一起, 还是单词的例子不多, 但是如果写在一起还是单词, 则更符合语义。

需要说明的是, 在具体实现过程中, 本实施例的正向分割或者反向分割, 还可以采用现有技术中的其它分割方式, 本实施例此处不做特别限制。

然而, 在对该字符串 floorlengthsleevelessdressst 进行反向分割时, 由于存在干扰字符 st, 则反向分割结果为一个错误的第二单词。

另一个具体的例子, 图 4 为本发明一实施例提供的反向分割示意图。如图 4 所示, 本实施例对字符串 ssfloorlengthsleevelessdress 进行反向分割。

具体的反向分割过程为: 从右到左取字符, 每取一次查一次词典, 来判断是否取到一个单词, 具体的分割过程与正向分割过程类似, 本实施例此处不再赘述。最终的反向分割结果为多个第二单词: floor length sleeveless dress。

然而, 在对该字符串 ssfloorlengthsleevelessdress 进行正向分割时, 由于存在干扰字符 ss, 则正向分割结果为一个错误的第二单词。

又一个具体的例子, 对字符串 sleepshirt 进行正向分割, 正向分割结果为 sleeps hirt; 对字符串 sleepshirt 进行反向分割, 反向分割结果为 sleep shirt。

在步骤 103 中, 获取各第一单词的词频和各第二单词的词频。该词频

为预先确定的各单词在预设文本中出现的次数。该预设文本可以为英文文学全集或者英文教材等。

具体的，以上述各实施例为例来进行说明。在图 3 所示的实施例中，在对 floorlengthsleevelessdressst 进行正向分割时，得到的正向分割
5 结果为多个正确的第一单词：floor length sleeveless dress，而对该 floorlengthsleevelessdressst 进行反向分割时，则得到一个错误的第二单词。此时，则第二单词的词频为无限小。

在图 4 所示的实施例中，在对字符串 ssfloorlengthsleevelessdress
10 进行反向分割时，得到的反向分割结果为多个正确的第二单词：floor length sleeveless dress，进行正向分割时，则得到一个错误的第一单词。此时，则第一单词的词频为无限小。

在上述实施例中，在对 sleepshirt 进行正向分割或者反向分割时，可以得到两个正确的第一单词和两个正确的第二单词。图 5 为本发明一实施例提供的正向和反向分割示意图。如图 5 所示，正向分割结果为 sleeps
15 hirt，sleeps 的词频为 100，hirt 的词频为 10；反向分割结果为 sleep shirt，sleep 的词频为 10000，shirt 的词频为 9000。

本领域技术人员可以理解，如果在正向分割或反向分割过程中，如果该字符串本身即为一个正确的单词，则该单词的词频为无限大。

在步骤 104 中，根据各第一单词的词频以及各第二单词的词频，确定待
20 分割的字符串的分割结果。具体地，可以对所有第一单词的词频进行求和处理，得到第一词频和值；对所有第二单词的词频进行求和处理，得到第二词频和值；若第一词频和值大于第二词频和值，则确定待分割的字符串的分割结果为正向分割结果；若第二词频和值大于第一词频和值，则确定待分割的字符串的分割结果为反向分割结果。

25 以图 3 至图 5 所示的实施例为例，在图 3 所示的实施例中，无法得到反向分割结果，则第二单词的词频为无限小，则分割结果为正向分割结果。在图 4 所示的实施例中，无法得到正向分割结果，则第一单词的词频为无限小，则分割结果为反向分割结果。在图 5 所示的实施例中，第一词频和值为 110，而第二词频和值为 19000，则分割结果为反向分割结果。

30 本领域技术人员可以理解，在具体实现过程中，还可以设置词频阈值，

然后确定正向分割结果中大于该词频阈值的单词的数量，确定反向分割结果中大于该词频阈值的单词的数量，将数量大的正向分割结果或反向分割结果作为最终的分割结果。同时，还可以对该词频进行各种变形处理，然后确定分割结果。即只要根据各第一单词的词频和第二单词的词频，来确定分割结果中的各单词为比较常用的单词，从而保证语义正确的实现方式，都在本发明的保护范围之内。

本实施例通过获取包括至少一个第一单词的正向分割结果，并获取包括至少一个第二单词的反向分割结果，通过双向分割字符串，识别字符串头或字符串尾的干扰字符，提高了分割成功率，然后获取各第一单词的词频和各第二单词的词频，根据各第一单词的词频以及各第二单词的词频，确定待分割的字符串的分割结果，基于词频来确定最终的分割结果，提高了分割结果中的各单词语义正确的概率。

由上述实施例可知，在图 3 所示的实施例中，如果反向分割将无法得到正确的第二单词，在图 4 所示的实施例中，如果正向分割，将无法得到正确的第一单词。在本实施例中，对单词分割方法还做了进一步的改进，使得在存在干扰字符的情况下，图 3 实施例中的字符串也可以得到多个正确的第二单词，图 4 实施例中的字符串也可以得到多个正确的第一单词，下面结合图 6 和图 7 进行详细说明。

图 6 为本发明一实施例提供的正向分割示意图。如图 6 所示，对待分割的字符串 ssffloorlengthsleevelessdressst 进行正向分割，判断是否获取到第一单词，由于存在干扰字符 ss，因此无法获取到第一单词，则对待分割的字符串的正向的首字符进行删除处理，即删除正向的第一个字符 s，得到处理后的待分割的字符串。然后将处理后的待分割的字符串作为新的待分割的字符串，并继续执行对待分割的字符串进行正向分割的操作，由于存在干扰字符 s，则依然无法获取到第一单词，则删除处理后的待分割的字符串的正向的第一个字符 s。然后，将处理后的待分割的字符串作为新的待分割的字符串，并继续执行对待分割的字符串进行正向分割的操作，可以得到第一单词 floor，此时，将除去第一单词的待分割的字符串作为新的待分割的字符串，继续执行对待分割的字符串进行正向分割的操作，通过重复执行对待分割的字符串进行正向分割的操作，直至对待

分割的字符串分割结束，得到正向分割结果。

本领域技术人员可以理解，对于字符串中间位置存在的干扰字符，在除去已经分割出来的第一单词后，该中间位置的干扰字符就变成了剩余字符串的首字符，在进行正向分割没有得到正确的第一单词时，还可以将该中间位置的干扰字符删除，然后继续进行正向分割，直至对待分割的字符串分割结束，得到正向分割结果。最终，正向分割结果为 floor length sleeveless dress。

本实施例提供的正向分割方法，为一层一层的正向递进式分割方式，经过一层一层的尝试，克服了干扰字符，最终得到了正向分割结果。

图 7 为本发明一实施例提供的反向分割示意图。如图 7 所示，对待分割的字符串 ssfloorlengthsleevelessdressst 进行反向分割，判断是否获取到第二单词，由于存在干扰字符 st，因此无法获取到第二单词，则对待分割的字符串的反向的首字符进行删除处理，即删除反向的第一个字符 t，得到处理后的待分割的字符串。然后将处理后的待分割的字符串作为新的待分割的字符串，并继续执行对待分割的字符串进行反向分割的操作，由于存在干扰字符 s，则依然无法获取到第二单词，则删除处理后的待分割的字符串的反向的第一个字符 s。然后，将处理后的待分割的字符串作为新的待分割的字符串，并继续执行对待分割的字符串进行反向分割的操作，可以得到第二单词 dress，此时，将除去第二单词的待分割的字符串作为新的待分割的字符串，并继续执行对待分割的字符串进行反向分割的操作。通过重复执行对待分割的字符串进行正向分割的操作，直至对待分割的字符串分割结束，得到反向分割结果。

本领域技术人员可以理解，对于字符串中间位置存在的干扰字符，在除去已经分割出来的第二单词后，该中间位置的干扰字符就变成了剩余字符串的首字符，在进行反向分割没有得到正确的第二单词时，还可以将该中间位置的干扰字符删除，然后继续进行反向分割，直至对待分割的字符串分割结束，得到反向分割结果。最终，反向分割结果为 floor length sleeveless dress。

本实施例提供的反向分割方法，为一层一层的反向递进式分割方式，经过一层一层的尝试，克服了干扰字符，最终得到了反向分割结果。

进一步地，在上述实施例的基础上，为了提高查询单词的效率，本申请还可以在单词分割之前，即正向单词分割之前和反向单词分割之前，预先构建正向字典树和反向字典树，使得在分割单词时，可以根据正向字典树，对待分割的字符串进行正向分割，根据反向字典树，对待分割的字符串进行反向分割。

具体地，字典树是一种树形结构，是一种哈希树的变种。它的优点是：利用字符串的公共前缀来减少查询时间，最大限度地减少无谓的字符串比较，查询效率比哈希树高。它有 3 个基本性质：根节点不包含字符，除根节点外每一个节点都只包含一个字符；从根节点到某一节点，路径上经过的字符连接起来，为该节点对应的字符串；每个节点的所有子节点包含的字符都不相同。

另外，节点中可以存储一些数据，比如该单词的频率等。正向字典树的每个第一节点中存储有第一节点对应的单词的词频，反向字典树的每个第二节点中存储有第二节点对应的单词的词频。对应地，从第一单词对应的第一节点中获取第一单词的词频；从第二单词对应的第二节点中获取第二单词的词频。

图 8 为本发明一实施例提供的正向字典树的示意图。所谓正向字典树，即由根节点到各级子节点，按照单词中各个字符正向排列顺序建立的字典树。如图 8 所示，在正向字典树中，“expend（消费）”和“expense（费用）”两个单词具有相同前缀“expen”，通过正向字典树表示后，可以使这两个单词的查找路径具有公共部分（即由正向字典树中虚线连接的 5 个节点构成的一段路径）。

图 9 为本发明一实施例提供的反向字典树的示意图。所谓反向字典树，即由根节点到各级子节点，按照单词中各个字符反向排列顺序建立的字典树。如图 9 所示，具有相同后缀“less”的两个单词“endless”和“useless”也在反向字典树中存在公共查找路径（虚线连接），即通过反向字典树可以使得具有相同后缀的两个或两个以上的单词具有一段相同的查找路径。

本实施例根据字典树来对字符串进行正向分割或反向分割，由于公共查找路径的存在，可以在读取到的子字符串增加一个字符后，基于该字符增加前的查找路径继续向下一级节点查找，从而可以避免重复查找，最大

限度地减少无谓的字符串比较，减少查询时间，提高查找效率。

进一步地，在上述实施例的基础上，还可以预先构建语料库。语料库包括单词库和单词库中的单词的词频，然后根据语料库来构建正向字典树和反向字典树，并将各单词的词频存储至对应的第一节点和第二节点。下面采用结合图 10 来说明本发明构建预料库的具体实现过程。

图 10 为本发明一实施例提供的字符串的分词方法流程示意图。如图 10 所示，该方法包括：

步骤 201、根据满足预设使用条件的词典，得到单词库。

获取满足预设使用条件的词典，该满足预设使用条件的词典可以为词汇量超过预设值的词典，也可以为下载频率超过预设次数的词典等，提取该词典中的单词，所有的单词构成了单词库。

步骤 202、确定单词库中的单词在满足预设使用条件的文本以及待分割的文本中出现的次数。

其中，满足预设使用条件的文本可以为英文文学全集、英文教材、英文报纸等使用频率超过预设值的文本。确定单词库中的单词在这些文本中出现的次数。

在确定单词库中的单词在待分割文本中出现的次数时，根据待分割的文本中的空格符，获取至少一个第一字符串；将至少一个第一字符串与单词库中的单词进行匹配，得到与单词库中的单词匹配的至少一个第二字符串；根据各第二字符串在待分割的文本中出现的次数，确定单词库中的单词在待分割的文本中出现的次数。

具体地，将待分割的文本去掉标点符号，按照空格分出一个一个的第一字符串，对于这些第一字符串，如果不在字典库中，则舍弃，剩下的都是单词，即第二字符串，然后统计第二字符串在待分割文本中出现的次数，即单词库中的单词在待分割文本中出现的次数。

步骤 203、根据单词库、单词库中的单词在满足预设使用条件的文本以及待分割的文本中出现的次数，构建语料库。

其中，语料库中包括单词库以及单词库中的单词的词频。若同一单词即出现在满足预设使用条件的文本中，又出现在待分割的文本中，则该单词的词频为该单词在满足预设使用条件的文本中出现的次数与待分割的

文本中出现的次数的加和。

本实施例构建的语料库，该语料库中的单词的词频是通过待分割的文本进行修正的，与待分割的文本具有一定的相关性，使得语料库中的单词的词频更接近待分割的文本的应用情况，从而可以使得分割结果的语义与待分割的文本表达的语义更接近，提高了字符串分割的正确性。

当本实施例的字符串的分词方法由云端服务器来执行时，本实施例中的云端服务器还可以与用户设备进行交互，以使得用户可以获知分割结果。下面采用详细的实施例来进行详细说明。

图 11 为本发明一实施例提供的字符串的分词方法流程示意图。该字符串的分词方法可以由字符串的分词装置来实现。该装置可以通过软件和/或硬件来实现。该分词装置还可以被配置到用户设备中，例如计算机、手机、平板等设备。在本实施例中，以该分词装置被配置到用户设备为例，进行详细说明。如图 11 所示，该方法包括：

步骤 301、向云端服务器发送用户输入的待分割的文本，以使所述云端服务器获取待分割的字符串，并根据正向分割结果中的各第一单词的词频以及反向分割结果中的各第二单词的词频确定分割结果；

步骤 302、接收所述云端服务器反馈的所述待分割的字符串的分割结果信息，所述分割结果信息包括所述待分割的字符串的分割结果；其中，所述待分割的字符串的分割结果为所述正向分割结果或所述反向分割结果；

步骤 303、向用户输出所述分割结果。

在步骤 301 中，用户通过用户设备上安装的应用程序或者浏览器浏览电商平台时，当用户需要查找某一商品时，用户设备获取用户输入的待分割的文本，然后向云端服务器发送用户输入的待分割的文本。具体地，用户可以通过语音或者文字输入待分割的文本。

云端服务器在获取到待分割的文本之后，根据该待分割的文本，获取待分割的字符串，然后对该待分割的字符串进行分词处理，可以得到正向分割结果、正向分割结果中的各第一单词的词频、第一词频和值，反向分割结果、反向分割结果中的各第二单词的词频、第二词频和值，以及最终的分割结果。云端服务器对待分割的字符串进行分词处理的具体实现方

式，可参见上述图 2 至图 10 所示的实施例，本实施例此处不再赘述。

在步骤 302 中，云端服务器在得到分割结果之后，向用户设备反馈待分割的字符串的分割结果信息，该分割结果信息包括分割结果。

在步骤 303 中，用户设备在获取到分割结果之后，向用户输出分割结果。具体地，用户设备可以通过语音或文字的形式输出分割结果。

本实施例提供的字符串的分词方法，通过向云端服务器发送用户输入的待分割的文本，以使云端服务器获取待分割的字符串，并根据正向分割结果中的各第一单词的词频以及反向分割结果中的各第二单词的词频确定分割结果；由于云端服务器通过双向分割字符串，可以识别字符串头或字符串尾的干扰字符，提高了分割成功率，基于词频来确定最终的分割结果，提高了分割结果中的各单词语义正确的概率，接收云端服务器反馈的待分割的字符串的分割结果信息，分割结果信息包括待分割的字符串的分割结果；向用户输出分割结果，用户可以获知分割结果，使得用户可以获知最终的查询结果对应的查询单词，提高了用户的体验。

下面结合图 12，以一个具体的例子，来说明用户设备与云端服务器的交互过程。图 12 为本发明一实施例提供的字符串的分词方法的信令流程图。如图 12 所示，该方法包括：

步骤 401、用户设备获取用户输入的待分割的文本；

步骤 402、用户设备向云端服务器发送用户输入的待分割的文本；

步骤 403、云端服务器根据待分割的文本得到待分割的字符串，确定待分割的字符串的分割结果；

步骤 404、云端服务器向用户设备发送待分割的字符串的分割结果信息；

步骤 405、用户设备向用户输出分割结果信息；

步骤 401 至步骤 405 的具体实现方式，可参见上述图 11 所示的实施例。可选地，在步骤 405 之后，还可以执行步骤 406 至步骤 408。

步骤 406、用户设备获取用户确定的待处理的分割结果；

步骤 407、用户设备向云端服务器发送待处理的分割结果；

步骤 408、对待处理的分割结果进行自然语言处理。

本实施例通过用户设备与云端服务器的交互，使得用户不仅可以获知

分割结果信息，还可以确定待处理的分割结果，提高了用户体验。

下面采用具体的实施例，对本实施例中的用户设备获取用户输入的待分割的文本，以及用户设备向用户输出分割结果信息进行详细说明。在本实施例中，以通过电商平台进行购物为例，来进行详细说明。本领域技术人员可以理解，该场景仅为示意性的场景，该方法还可以应用到网页搜索等场景中，本实施例对具体的场景不做特别限制。

图 13 为本发明一实施例提供的字符串的分词方法的显示界面示意图。在本实施例中，用户可在用户设备的显示界面的搜索框中输入待查看的商品的类型。如图 13 所示，用户在显示界面的搜索框中输入了“sleep shirt”的文本，则用户设备将该文本发送给云端服务器。云端服务器在获取到待分割的文本后，对该待分割的文本进行处理，得到待分割的字符串“sleepshirt”。然后云端服务器对该待分割的字符串进行分割处理，具体的分割处理过程以及分割处理结果，可参见图 5 所示的实施例，本实施例此处不再赘述。

在本实施例中，当云端服务器获取到分割结果之后，云端服务器向用户设备返回分割结果信息。用户设备在接收到分割结果信息之后，根据该分割结果信息向用户输出分割结果。下面结合图 14 至图 18 来具体说明用户设备输出分割结果的实现过程。

图 14 为本发明一实施例提供的字符串的分词方法的显示界面示意图。在本实施例中，分割结果信息中包括待分割的字符串的分割结果，则对应地在用户设备的显示界面上显示该分割结果。如图 14 所示，在显示界面上显示有分割结果“sleep shirt”。

图 15 为本发明一实施例提供的字符串的分词方法的显示界面示意图。在本实施例中，分割结果信息中包括待分割的字符串的分割结果、分割结果对应的分割类型，分割类型为正向分割或反向分割。对应地，在用户设备的显示界面上显示该分割结果以及分割结果的分割类型。如图 15 所示，在显示界面上显示有分割结果“sleep shirt”，并显示有分割结果的分割类型“反向分割”。

图 16 为本发明一实施例提供的字符串的分词方法的显示界面示意图。在本实施例中，分割结果信息中包括正向分割结果、反向分割结果，以及

最终的分割结果。对应地，在用户设备的显示界面上显示正向分割结果和反向分割结果，并标注待分割字符串对应的分割结果。如图 16 所示，在显示界面上显示有反向分割结果“sleep shirt”以及正向分割结果“sleeps hirt”，并通过灰色背影标注待分割字符串对应的分割结果为反向分割结果。

图 17 为本发明一实施例提供的字符串的分词方法的显示界面示意图。本实施例在图 16 实施例的基础上，分割结果信息中还包括正向分割结果中的各第一单词的词频和反向分割结果中的各第二单词的词频。对应地，在图 17 所示的显示界面上，显示有反向分割结果以及反向分割结果中的各第二单词的词频，还显示有正向分割结果以及正向分割结果中的各第一单词的词频。在本实施例中，用户设备在获取到分割结果信息之后，可以直接在显示界面上显示图 17 所示的内容，也可以先在显示界面上显示如图 16 所示的内容，然后当获取到用户操作显示界面触发的词频显示指令后，根据词频显示指令，显示各第一单词的词频和/或各第二单词的词频。本领域技术人员可以理解，当根据词频显示指令显示各第一单词的词频和各第二单词的词频时，具体的显示内容可如图 17 所示。

图 18 为本发明一实施例提供的字符串的分词方法的显示界面示意图。本实施例在图 16 实施例的基础上，分割信息中还包括正向分割结果中的各第一单词对应的第一词频和值以及反向分割结果中的各第二单词对应的第二词频和值。对应地，在图 18 所示的显示界面上，显示有反向分割结果以及各第二单词对应的第二词频和值，还显示有正向分割结果以及各第一单词对应的第一词频和值。在本实施例中，用户设备在获取到分割结果信息之后，可以直接在显示界面上显示图 18 所示的内容，也可以先在显示界面上显示如图 16 所示的内容，然后当获取到用户操作显示界面触发的词频显示指令后，根据词频显示指令，显示所述第一词频和值和/或所述第二词频和值。本领域技术人员可以理解，当根据词频显示指令显示第一词频和值和第二词频和值时，具体的显示内容可如图 18 所示。

在上述图 16 至图 18 所示的实施例中，用户可以通过操作显示界面来决定云端服务器的待处理的分割结果。具体地，用户可以通过点击、滑动等操作方式来操作正向分割结果或反向分割结果。用户设备可以根据用户

对正向分割结果或反向分割结果的操作方式，来获取操作信息，根据操作信息来确定待处理的分割结果。在本实施例中，在用户执行点击“sleep shirt”输出框的操作时，用户设备根据该点击操作来获取操作信息，具体的操作信息为反向分割结果被用户选择，用户设备根据该操作信息确定待处理的分割结果为反向分割结果。然后，用户设备将待处理的分割结果反馈给云端服务器，由云端服务器对待处理的分割结果进行后续处理。

在本实施例中，由于同时在显示界面上显示了正向分割结果以及反向分割结果，则用户可以根据正向分割结果以及反向分割结果来确定自己需要查找或搜索的对象，提高了搜索的准确性和有效性。进一步地，本实施例还在显示界面上显示词频，用户在看到该词频后，能够快速做出更正确的判断，提高了用户体验。

以下将详细描述根据本申请的一个或多个实施例的字符串的分词装置。该字符串的分词装置可以被实现在各种设备上，例如，服务端设备、服务器、网络服务器等。本领域技术人员可以理解，该字符串的分词装置均可使用市售的硬件组件通过本方案所教导的步骤进行配置来构成。例如，下述实施例中的涉及控制功能、更新功能的模块可以使用来自德州仪器公司、英特尔公司、ARM公司等企业的单片机、微控制器、微处理器等组件实现。

下述为本申请装置实施例，可以用于执行本申请方法实施例。对于本申请装置实施例中未披露的细节，请参照本申请方法实施例。

图 19 为本发明一实施例提供的字符串的分词装置的结构示意图。如图 19 所示，该装置包括：

第一分割模块 10，用于获取待分割的字符串的正向分割结果，所述正向分割结果包括至少一个第一单词；

第二分割模块 11，用于获取所述待分割的字符串的反向分割结果，所述反向分割结果包括至少一个第二单词；

词频获取模块 12，用于获取各所述第一单词的词频和各所述第二单词的词频，所述词频为预先确定的各单词在预设文本中出现的次数；

结果确定模块 13，用于根据各所述第一单词的词频以及各所述第二单词的词频，确定所述待分割的字符串的分割结果，其中，所述待分割的字符串的分割结果为所述正向分割结果或所述反向分割结果。

本申请实施例提供的字符串的分词装置，可以执行上述方法实施例，其实现原理和技术效果类似，在此不再赘述。

图 20 为本发明一实施例提供的字符串的分词装置的结构示意图。本实施例在图 19 实施例的基础上实现，具体如下：

5 可选地，所述第一分割模块 10 具体用于，
 对所述待分割的字符串进行正向分割的操作，判断是否获取到第一单词；
 若是，将除去所述第一单词的待分割的字符串作为新的待分割的字符串，
 并返回执行对待分割的字符串进行正向分割的操作；

 若否，对所述待分割的字符串的正向的首字符进行删除处理，得到处理
10 后的待分割的字符串，将处理后的待分割的字符串作为新的待分割的字符串，
 并返回执行对待分割的字符串进行正向分割的操作；

 重复执行对所述待分割的字符串进行正向分割的操作，直至对所述待分割的字符串分割结束，得到正向分割结果。

 可选地，所述第二分割模块 11 具体用于，对所述待分割的字符串进行反
15 向分割的操作，判断是否获取到第二单词；

 若是，将除去所述第二单词的待分割的字符串作为新的待分割的字符串，
 并返回执行对待分割的字符串进行反向分割的操作；

 若否，对所述待分割的字符串的反向的首字符进行删除处理，得到处理
20 后的待分割的字符串，将处理后的待分割的字符串作为新的待分割的字符串，
 并返回执行对待分割的字符串进行反向分割的操作；

 重复执行对所述待分割的字符串进行正向分割的操作，直至对所述待分割的字符串分割结束，得到反向分割结果。

 可选地，还包括：文本获取模块 14，用于获取待分割的文本，对所述待分割的文本进行符号删除操作，得到所述待分割的字符串。

25 可选地，还包括：字典树构建模块 15，用于构建正向字典树和反向字典树；

 所述第一分割模块 10 具体用于，

 根据所述正向字典树，对所述待分割的字符串进行正向分割的操作；

 所述第二分割模块 11 具体用于，

30 根据所述反向字典树，对所述待分割的字符串进行反向分割的操作。

可选地，所述正向字典树的每个第一节点中存储有所述第一节点对应的单词的词频，所述反向字典树的每个第二节点中存储有所述第二节点对应的单词的词频；

所述词频获取模块 12 具体用于，

5 从所述第一单词对应的第一节点中获取所述第一单词的词频；

从所述第二单词对应的第二节点中获取所述第二单词的词频。

可选地，还包括：语料库构建模块 16，用于构建语料库，所述语料库包括单词库和所述单词库中的单词的词频；

10 所述字典树构建模块 15 具体用于，根据所述语料库，构建正向字典树和反向字典树，并将各单词的词频存储至对应的第一节点和第二节点。

可选地，所述预设文本包括：满足预设使用条件的文本以及待分割的文本；所述语料库构建模块 16 具体用于，

根据满足预设使用条件的词典，得到单词库；

15 确定所述单词库中的单词在所述满足预设使用条件的文本以及所述待分割的文本中出现的次数；

根据所述单词库、所述单词库中的单词在所述满足预设使用条件的文本以及所述待分割的文本中出现的次数，构建所述语料库。

可选地，所述语料库构建模块 16 具体用于，

根据所述待分割的文本中的空格符，获取至少一个第一字符串；

20 将所述至少一个第一字符串与所述单词库中的单词进行匹配，得到与所述单词库中的单词匹配的至少一个第二字符串；

根据各所述第二字符串在所述待分割的文本中出现的次数，确定单词库中的单词在所述待分割的文本中出现的次数。

可选地，所述结果确定模块 13 具体用于，

25 对所有所述第一单词的词频进行求和处理，得到第一词频和值；

对所有所述第二单词的词频进行求和处理，得到第二词频和值；

若所述第一词频和值大于所述第二词频和值，则确定所述待分割的字符串的分割结果为正向分割结果；

30 若所述第二词频和值大于所述第一词频和值，则确定所述待分割的字符串的分割结果为反向分割结果。

可选地，还包括：反馈模块 17；

所述文本获取模块 14 具体用于，获取用户设备发送的所述待分割的文本；

所述反馈模块 17 用于，向所述用户设备反馈所述待分割的字符串的分割结果信息，所述分割结果信息包括所述待分割的字符串的分割结果，以使所述用户设备向用户输出所述分割结果。

可选地，还包括：结果获取模块 18 和处理模块 19，

所述结果获取模块 18 用于，获取所述用户设备发送的待处理的分割结果；

所述处理模块 19 用于，对所述待处理的分割结果进行自然语言处理。

本申请实施例提供的字符串的分词装置，可以执行上述方法实施例，其实现原理和技术效果类似，在此不再赘述。

图 21 为本发明一实施例提供的字符串的分词装置的结构示意图。如图 21 所示，该装置包括：

发送模块 20，用于向云端服务器发送用户输入的待分割的文本，以使所述云端服务器获取待分割的字符串，并根据正向分割结果中的各第一单词的词频以及反向分割结果中的各第二单词的词频确定分割结果；

接收模块 21，用于接收所述云端服务器反馈的所述待分割的字符串的分割结果信息，所述分割结果信息包括所述待分割的字符串的分割结果；其中，所述待分割的字符串的分割结果为所述正向分割结果或所述反向分割结果；

输出模块 22，用于向用户输出所述分割结果。

本申请实施例提供的字符串的分词装置，可以执行上述方法实施例，其实现原理和技术效果类似，在此不再赘述。

图 22 为本发明一实施例提供的字符串的分词装置的结构示意图。如图 22 所示，本实施例在图 21 所示实施例的基础上实现，具体如下：

所述输出模块 22 具体用于，在显示界面上显示所述分割结果。

可选地，所述分割结果信息中还包括所述分割结果对应的分割类型，所述分割类型为正向分割或反向分割；

所述输出模块 22 具体用于，在显示界面上显示所述分割结果以及所述分割结果的分割类型。

可选地，若所述分割结果为正向分割结果，则所述分割信息中还包括反向分割结果；或者

若所述分割结果为反向分割结果，则所述分割信息中还包括正向分割结果；

- 5 所述输出模块 22 具体用于，在所述显示界面上显示所述正向分割结果和所述反向分割结果，并标注所述待分割字符串对应的分割结果。

可选地，所述分割信息中还包括所述正向分割结果中的各所述第一单词的词频和所述反向分割结果中的各所述第二单词的词频；

- 10 所述显示装置还包括：指令获取模块 23，用于获取所述用户操作所述显示界面触发的词频显示指令；

所述输出模块 22 还用于，根据所述词频显示指令，显示各所述第一单词的词频和/或各所述第二单词的词频；

或者

- 15 所述输出模块 22 具体用于，在所述显示界面上显示所述正向分割结果、所述正向分割结果中的第一单词的词频，以及所述反向分割结果、所述反向分割结果中的第二单词的词频。

可选地，所述分割信息中还包括所述正向分割结果中的各所述第一单词对应的第一词频和值以及所述反向分割结果中的各所述第二单词对应的第二词频和值；

- 20 所述显示装置还包括：指令获取模块 23，用于获取所述用户操作所述显示界面触发的词频显示指令；

所述输出模块 22 还用于，根据所述词频显示指令，显示所述第一词频和值和/或所述第二词频和值；

或者

- 25 所述输出模块 22 具体用于，在所述显示界面上显示所述正向分割结果、所述第一词频和值，以及所述反向分割结果、所述第二词频和值。

可选地，还包括：操作信息获取模块 24，用于获取所述用户对所述显示界面上的所述正向分割结果或反向分割结果的操作信息，

确定模块 25，用于根据所述操作信息确定待处理的分割结果；

- 30 所述发送模块 20 还用于，向所述云端服务器发送所述待处理的分割结

果，以使所述云端服务器对所述待处理的分割结果进行自然语言处理。

本申请实施例提供的字符串的分词装置，可以执行上述方法实施例，其实现原理和技术效果类似，在此不再赘述。

图 23 为本发明一实施例提供的字符串的分词设备的硬件结构示意图。

5 如图 23 所示，该字符串的分词设备可以包括输入设备 30、处理器 31、存储器 32 和至少一个通信总线 33 以及输出设备 34。通信总线 33 用于实现元件之间的通信连接。存储器 32 可能包含高速 RAM 存储器，也可能还包括非易失性存储 NVM，例如至少一个磁盘存储器，存储器 32 中可以存储各种程序，用于完成各种处理功能以及实现本实施例的方法步骤。

10 在本实施例中，输入设备 30，用于获取待分割的文本；

处理器 31，耦合至所述输入设备 30，用于获取待分割的字符串的正向分割结果，所述正向分割结果包括至少一个第一单词，并获取所述待分割的字符串的反向分割结果，所述反向分割结果包括至少一个第二单词；获取各所述第一单词的词频和各所述第二单词的词频，所述词频为预先确定的各单词
15 在预设文本中出现的次数；根据各所述第一单词的词频以及各所述第二单词的词频，确定所述待分割的字符串的分割结果，其中，所述待分割的字符串的分割结果为所述正向分割结果或所述反向分割结果。

输出设备 34，用于向用户设备反馈所述待分割的字符串的分割结果信息，所述分割结果信息包括所述待分割的字符串的分割结果，以使所述用户设备
20 向用户输出所述分割结果。

可选地，该处理器 31 还用于执行上述图 2 至图 10 所述的方法，输入设备 30 对应执行输入操作，输出设备 34 对应执行输出操作，具体实现过程，可参见上述实施例，本实施例此处不再赘述。

图 24 为本发明一实施例提供的云端服务器的硬件结构示意图。如图
25 24 所示，该云端服务器可以包括输入设备 40、处理器 41、存储器 42 和至少一个通信总线 43 以及输出设备 44。通信总线 43 用于实现元件之间的通信连接。存储器 42 可能包含高速 RAM 存储器，也可能还包括非易失性存储 NVM，例如至少一个磁盘存储器，存储器 42 中可以存储各种程序，用于完成各种处理功能以及实现本实施例的方法步骤。

30 在本实施例中，输入设备 40，用于获取待分割的文本；

处理器 41, 耦合至所述输入设备 40, 用于获取待分割的字符串的正向分割结果, 所述正向分割结果包括至少一个第一单词, 并获取所述待分割的字符串的反向分割结果, 所述反向分割结果包括至少一个第二单词; 获取各所述第一单词的词频和各所述第二单词的词频, 所述词频为预先确定的各单词在预设文本中出现的次数; 根据各所述第一单词的词频以及各所述第二单词的词频, 确定所述待分割的字符串的分割结果, 其中, 所述待分割的字符串的分割结果为所述正向分割结果或所述反向分割结果。

输出设备 44, 用于向用户设备反馈所述待分割的字符串的分割结果信息, 所述分割结果信息包括所述待分割的字符串的分割结果, 以使所述用户设备向用户输出所述分割结果。

可选地, 该处理器 41 还用于执行上述图 2 至图 10 所述的方法, 输入设备 40 对应执行输入操作, 输出设备 44 对应至少输出操作, 具体实现过程, 可参见上述实施例, 本实施例此处不再赘述。

图 25 为本发明一实施例提供的字符串的分词设备的硬件结构示意图。如图 25 所示, 该字符串的分词设备可以包括输入设备 50、处理器 51、存储器 52 和至少一个通信总线 53 以及输出设备 54。通信总线 53 用于实现元件之间的通信连接。存储器 52 可能包含高速 RAM 存储器, 也可能还包括非易失性存储 NVM, 例如至少一个磁盘存储器, 存储器 52 中可以存储各种程序, 用于完成各种处理功能以及实现本实施例的方法步骤。

其中, 输出设备 54, 用于向云端服务器发送用户输入的待分割的文本, 以使所述云端服务器获取待分割的字符串, 并根据正向分割结果中的各第一单词的词频以及反向分割结果中的各第二单词的词频确定分割结果;

输入设备 50, 用于接收所述云端服务器反馈的所述待分割的字符串的分割结果信息, 所述分割结果信息包括所述待分割的字符串的分割结果; 其中, 所述待分割的字符串的分割结果为所述正向分割结果或所述反向分割结果;

处理器 51, 用于根据所述分割结果信息, 控制所述输入设备向用户输出所述分割结果。

可选地, 该处理器 51 还用于执行上述图 11 至图 18 所示的方法, 输入设备 50 对应执行输入操作, 输出设备 54 对应至少输出操作, 具体实现过程, 可参见上述实施例, 本实施例此处不再赘述。

图 26 为本发明一实施例提供的用户设备的硬件结构示意图。如图 26 所示，该字符串的分词设备可以包括输入设备 60、处理器 61、存储器 62 和至少一个通信总线 63 以及输出设备 64。通信总线 63 用于实现元件之间的通信连接。存储器 62 可能包含高速 RAM 存储器，也可能还包括非易失性存储 NVM，例如至少一个磁盘存储器，存储器 62 中可以存储各种程序，用于完成各种处理功能以及实现本实施例的方法步骤。

其中，输出设备 64，用于向云端服务器发送用户输入的待分割的文本，以使所述云端服务器获取待分割的字符串，并根据正向分割结果中的各第一单词的词频以及反向分割结果中的各第二单词的词频确定分割结果；

输入设备 60，用于接收所述云端服务器反馈的所述待分割的字符串的分割结果信息，所述分割结果信息包括所述待分割的字符串的分割结果；其中，所述待分割的字符串的分割结果为所述正向分割结果或所述反向分割结果；

处理器 61，用于根据所述分割结果信息，控制所述输入设备向用户输出所述分割结果。

可选地，该处理器 61 还用于执行上述图 11 至图 18 所示的方法，输入设备 60 对应执行输入操作，输出设备 64 对应至少输出操作，具体实现过程，可参见上述实施例，本实施例此处不再赘述。

在上述图 23 至图 26 所示的实施例中，上述处理器例如可以为中央处理器 (Central Processing Unit, 简称 CPU)、应用专用集成电路 (ASIC)、数字信号处理器 (DSP)、数字信号处理设备 (DSPD)、可编程逻辑器件 (PLD)、现场可编程门阵列 (FPGA)、控制器、微控制器、微处理器或其他电子元件实现。

上述输入设备可以包括多种输入设备，例如可以包括面向用户的用户接口、面向设备的设备接口、软件的可编程接口、收发信机中的至少一个。可选的，该面向设备的设备接口可以是用于设备与设备之间进行数据传输的有线接口、还可以是用于设备与设备之间进行数据传输的硬件插入接口（例如 USB 接口、串口等）；可选的，该面向用户的用户接口例如可以是面向用户的控制按键、用于接收语音输入的语音输入设备以及用户接收用户触摸输入的触摸感知设备（例如具有触摸感应功能的触摸屏、触控板等）；可选的，上述软件的可编程接口例如可以是供用户编辑或者修改程序的入口，例如芯

片的输入引脚接口或者输入接口等；可选的，上述收发信机可以是具有通信功能的射频收发芯片、基带处理芯片以及收发天线等。

上述输出设备可以包括多种输出设备，例如可以包括面向用户的用户接口、面向设备的设备接口、软件的可编程接口、收发信机中的至少一个。可选的，该面向设备的设备接口可以是用于设备与设备之间进行数据传输的有线接口、还可以是用于设备与设备之间进行数据传输的硬件插入接口（例如 USB 接口、串口等）；可选的，该面向用户的用户接口例如可以是面向用户的显示设备或语音输出设备；可选的，上述软件的可编程接口例如可以是供用户编辑或者修改程序的入口，例如芯片的输入引脚接口或者输入接口等；可选的，上述收发信机可以是具有通信功能的射频收发芯片、基带处理芯片以及收发天线等。

在本发明实施例中使用的术语是仅仅出于描述特定实施例的目的，而非旨在限制本发明。在本发明实施例和所附权利要求书中所使用的单数形式的“一种”、“所述”和“该”也旨在包括多数形式，除非上下文清楚地表示其他含义。

应当理解，本文中使用的术语“和/或”仅仅是一种描述关联对象的关联关系，表示可以存在三种关系，例如，A 和/或 B，可以表示：单独存在 A，同时存在 A 和 B，单独存在 B 这三种情况。另外，本文中字符“/”，一般表示前后关联对象是一种“或”的关系。

应当理解，尽管在本发明实施例中可能采用术语第一、第二、第三等来描述 XXX，但这些 XXX 不应限于这些术语。这些术语仅用来将 XXX 彼此区分开。例如，在不脱离本发明实施例范围的情况下，第一 XXX 也可以被称为第二 XXX，类似地，第二 XXX 也可以被称为第一 XXX。

还需要说明的是，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商品或者系统不仅包括那些要素，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商品或者系统所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括所述要素的商品或者系统中还存在另外的相同要素。

上述可读存储介质可以由任何类型的易失性或非易失性存储

设备或者它们的组合实现，如静态随机存取存储器（SRAM），电可擦除可编程只读存储器（EEPROM），可擦除可编程只读存储器（EPROM），可编程只读存储器（PROM），只读存储器（ROM），磁存储器，快闪存储器，磁盘或光盘。

- 5 最后应说明的是：以上各实施例仅用以说明本申请的技术方案，而非对其限制；尽管参照前述各实施例对本申请进行了详细的说明，本领域的普通技术人员应当理解：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分或者全部技术特征进行等同替换；而这些修改或者替换，并不使相应技术方案的本质脱离本申请各实施例技术方案的范围。

10

15

20

25

30

权利要求书

1、一种字符串的分词方法，其特征在于，包括：

获取待分割的字符串的正向分割结果，所述正向分割结果包括至少一个第一单词；

5 获取所述待分割的字符串的反向分割结果，所述反向分割结果包括至少一个第二单词；

获取各所述第一单词的词频和各所述第二单词的词频，所述词频为预先确定的各单词在预设文本中出现的次数；

10 根据各所述第一单词的词频以及各所述第二单词的词频，确定所述待分割的字符串的分割结果，其中，所述待分割的字符串的分割结果为所述正向分割结果或所述反向分割结果。

2、根据权利要求1所述的方法，其特征在于，所述获取待分割的字符串的正向分割结果，包括：

对所述待分割的字符串进行正向分割的操作，判断是否获取到第一单词；

15 若是，将除去所述第一单词的待分割的字符串作为新的待分割的字符串，并返回执行对待分割的字符串进行正向分割的操作；

若否，对所述待分割的字符串的正向的首字符进行删除处理，得到处理后的待分割的字符串，将处理后的待分割的字符串作为新的待分割的字符串，并返回执行对待分割的字符串进行正向分割的操作；

20 重复执行对所述待分割的字符串进行正向分割的操作，直至对所述待分割的字符串分割结束，得到正向分割结果。

3、根据权利要求1所述的方法，其特征在于，所述获取待分割的字符串反向分割结果，包括：

对所述待分割的字符串进行反向分割的操作，判断是否获取到第二单词；

25 若是，将除去所述第二单词的待分割的字符串作为新的待分割的字符串，并返回执行对待分割的字符串进行反向分割的操作；

若否，对所述待分割的字符串的反向的首字符进行删除处理，得到处理后的待分割的字符串，将处理后的待分割的字符串作为新的待分割的字符串，并返回执行对待分割的字符串进行反向分割的操作；

30 重复执行对所述待分割的字符串进行正向分割的操作，直至对所述待分

割的字符串分割结束，得到反向分割结果。

4、根据权利要求 1 至 3 任一项所述的方法，其特征在于，还包括：

获取待分割的文本，对所述待分割的文本进行符号删除操作，得到所述待分割的字符串。

5 5、根据权利要求 2 或 3 所述的方法，其特征在于，还包括：

构建正向字典树和反向字典树；

所述对所述待分割的字符串进行正向分割的操作，包括：

根据所述正向字典树，对所述待分割的字符串进行正向分割的操作；

所述对所述待分割的字符串进行反向分割的操作，包括：

10 根据所述反向字典树，对所述待分割的字符串进行反向分割的操作。

6、根据权利要求 5 所述的方法，其特征在于，所述正向字典树的每个第一节点中存储有所述第一节点对应的单词的词频，所述反向字典树的每个第二节点中存储有所述第二节点对应的单词的词频；

所述获取各所述第一单词的词频和各所述第二单词的词频，包括：

15 从所述第一单词对应的第一节点中获取所述第一单词的词频；

从所述第二单词对应的第二节点中获取所述第二单词的词频。

7、根据权利要求 6 所述的方法，其特征在于，所述构建正向字典树和反向字典树之前，还包括：

构建语料库，所述语料库包括单词库和所述单词库中的单词的词频；

20 所述构建正向字典树和反向字典树，包括：

根据所述语料库，构建正向字典树和反向字典树，并将各单词的词频存储至对应的第一节点和第二节点。

8、根据权利要求 7 所述的方法，其特征在于，所述预设文本包括：满足预设使用条件的文本以及待分割的文本；所述构建语料库，包括：

25 根据满足预设使用条件的词典，得到单词库；

确定所述单词库中的单词在所述满足预设使用条件的文本以及所述待分割的文本中出现的次数；

根据所述单词库、所述单词库中的单词在所述满足预设使用条件的文本以及所述待分割的文本中出现的次数，构建所述语料库。

30 9、根据权利要求 8 所述的方法，其特征在于，所述确定单词库中的单词

在所述待分割的文本中出现的次数，包括：

根据所述待分割的文本中的空格符，获取至少一个第一字符串；

将所述至少一个第一字符串与所述单词库中的单词进行匹配，得到与所述单词库中的单词匹配的至少一个第二字符串；

- 5 根据各所述第二字符串在所述待分割的文本中出现的次数，确定单词库中的单词在所述待分割的文本中出现的次数。

10、根据权利要求 1 所述的方法，其特征在于，所述根据各所述第一单词的词频以及各所述第二单词的词频，确定所述待分割的字符串的分割结果，包括：

- 10 对所有所述第一单词的词频进行求和处理，得到第一词频和值；

对所有所述第二单词的词频进行求和处理，得到第二词频和值；

若所述第一词频和值大于所述第二词频和值，则确定所述待分割的字符串的分割结果为正向分割结果；

- 15 若所述第二词频和值大于所述第一词频和值，则确定所述待分割的字符串的分割结果为反向分割结果。

11、根据权利要求 1 至 3 任一项所述的方法，其特征在于，所述正向分割和所述反向分割均采用最长单词分割方式。

12、根据权利要求 4 所述的方法，其特征在于，所述获取待分割的文本，包括：

- 20 获取用户设备发送的所述待分割的文本；

所述确定所述待分割的字符串的分割结果之后，还包括：

向所述用户设备反馈所述待分割的字符串的分割结果信息，所述分割结果信息包括所述待分割的字符串的分割结果，以使所述用户设备向用户输出所述分割结果。

- 25 13、根据权利要求 12 所述的方法，其特征在于，所述分割结果信息中还包括所述分割结果对应的分割类型，所述分割类型为正向分割或反向分割。

14、根据权利要求 12 所述的方法，其特征在于，若所述分割结果为正向分割结果，则所述分割信息中还包括反向分割结果；或者

- 30 若所述分割结果为反向分割结果，则所述分割信息中还包括正向分割结果。

15、根据权利要求 14 所述的方法，其特征在于，所述分割信息中还包括所述正向分割结果中的各所述第一单词的词频和所述反向分割结果中的各所述第二单词的词频。

16、根据权利要求 14 所述的方法，其特征在于，所述分割信息中还包括
5 所述正向分割结果中的各所述第一单词对应的第一词频和值以及所述反向分割结果中的各所述第二单词对应的第二词频和值。

17、根据权利要求 12 所述的方法，其特征在于，所述向所述用户设备反馈所述待分割的字符串的分割结果信息之后，还包括：

获取所述用户设备发送的待处理的分割结果；

10 对所述待处理的分割结果进行自然语言处理。

18、一种字符串的分词方法，其特征在于，包括：

向云端服务器发送用户输入的待分割的文本，以使所述云端服务器获取待分割的字符串，并根据正向分割结果中的各第一单词的词频以及反向分割结果中的各第二单词的词频确定分割结果；

15 接收所述云端服务器反馈的所述待分割的字符串的分割结果信息，所述分割结果信息包括所述待分割的字符串的分割结果；其中，所述待分割的字符串的分割结果为所述正向分割结果或所述反向分割结果；

向用户输出所述分割结果。

19、根据权利要求 18 所述的方法，其特征在于，所述向用户输出所述分
20 割结果，包括：

在显示界面上显示所述分割结果。

20、根据权利要求 19 所述的方法，其特征在于，所述分割结果信息中还包括所述分割结果对应的分割类型，所述分割类型为正向分割或反向分割；

所述在显示界面上显示所述分割结果，包括：

25 在显示界面上显示所述分割结果以及所述分割结果的分割类型。

21、根据权利要求 19 所述的方法，其特征在于，若所述分割结果为正向分割结果，则所述分割信息中还包括反向分割结果；或者

若所述分割结果为反向分割结果，则所述分割信息中还包括正向分割结果；

30 所述在显示界面上显示所述分割结果，包括：

在所述显示界面上显示所述正向分割结果和所述反向分割结果，并标注所述待分割字符串对应的分割结果。

22、根据权利要求 21 所述的方法，其特征在于，所述分割信息中还包括所述正向分割结果中的各所述第一单词的词频和所述反向分割结果中的各所述第二单词的词频；

在所述显示界面上显示所述正向分割结果和所述反向分割结果，并标注所述待分割字符串对应的分割结果之后，还包括：

获取所述用户操作所述显示界面触发的词频显示指令；

根据所述词频显示指令，显示各所述第一单词的词频和/或各所述第二单词的词频；

或者

在所述显示界面上显示所述正向分割结果和所述反向分割结果，包括：

在所述显示界面上显示所述正向分割结果、所述正向分割结果中的第一单词的词频，以及所述反向分割结果、所述反向分割结果中的第二单词的词频。

23、根据权利要求 21 所述的方法，其特征在于，所述分割信息中还包括所述正向分割结果中的各所述第一单词对应的第一词频和值以及所述反向分割结果中的各所述第二单词对应的第二词频和值；

在所述显示界面上显示所述正向分割结果和所述反向分割结果，并标注所述待分割字符串对应的分割结果之后，还包括：

获取所述用户操作所述显示界面触发的词频显示指令；

根据所述词频显示指令，显示所述第一词频和值和/或所述第二词频和值；

或者

25 在所述显示界面上显示所述正向分割结果和所述反向分割结果，包括：

在所述显示界面上显示所述正向分割结果、所述第一词频和值，以及所述反向分割结果、所述第二词频和值。

24、根据权利要求 21 至 23 任一项所述的方法，其特征在于，所述在所述显示界面上显示所述正向分割结果和所述反向分割结果之后，还包括：

30 获取所述用户对所述显示界面上的所述正向分割结果或反向分割结果的

操作信息，

根据所述操作信息确定待处理的分割结果；

向所述云端服务器发送所述待处理的分割结果，以使所述云端服务器对所述待处理的分割结果进行自然语言处理。

5 25、一种字符串的分词装置，其特征在于，包括：

第一分割模块，用于获取待分割的字符串的正向分割结果，所述正向分割结果包括至少一个第一单词；

第二分割模块，用于获取所述待分割的字符串的反向分割结果，所述反向分割结果包括至少一个第二单词；

10 词频获取模块，用于获取各所述第一单词的词频和各所述第二单词的词频，所述词频为预先确定的各单词在预设文本中出现的次数；

结果确定模块，用于根据各所述第一单词的词频以及各所述第二单词的词频，确定所述待分割的字符串的分割结果，其中，所述待分割的字符串的分割结果为所述正向分割结果或所述反向分割结果。

15 26、根据权利要求 25 所述的装置，其特征在于，所述第一分割模块具体用于，

对所述待分割的字符串进行正向分割的操作，判断是否获取到第一单词；

若是，将除去所述第一单词的待分割的字符串作为新的待分割的字符串，并返回执行对待分割的字符串进行正向分割的操作；

20 若否，对所述待分割的字符串的正向的首字符进行删除处理，得到处理后的待分割的字符串，将处理后的待分割的字符串作为新的待分割的字符串，并返回执行对待分割的字符串进行正向分割的操作；

重复执行对所述待分割的字符串进行正向分割的操作，直至对所述待分割的字符串分割结束，得到正向分割结果。

25 27、根据权利要求 25 所述的装置，其特征在于，所述第二分割模块具体用于，对所述待分割的字符串进行反向分割的操作，判断是否获取到第二单词；

若是，将除去所述第二单词的待分割的字符串作为新的待分割的字符串，并返回执行对待分割的字符串进行反向分割的操作；

30 若否，对所述待分割的字符串的反向的首字符进行删除处理，得到处理

后的待分割的字符串，将处理后的待分割的字符串作为新的待分割的字符串，并返回执行对待分割的字符串进行反向分割的操作；

重复执行对所述待分割的字符串进行正向分割的操作，直至对所述待分割的字符串分割结束，得到反向分割结果。

- 5 28、根据权利要求 25 至 27 任一项所述的装置，其特征在于，还包括：文本获取模块，用于获取待分割的文本，对所述待分割的文本进行符号删除操作，得到所述待分割的字符串。

29、根据权利要求 25 至 27 任一项所述的装置，其特征在于，还包括：字典树构建模块，用于构建正向字典树和反向字典树；

- 10 所述第一分割模块具体用于，
根据所述正向字典树，对所述待分割的字符串进行正向分割的操作；
所述第二分割模块具体用于，
根据所述反向字典树，对所述待分割的字符串进行反向分割的操作。

- 15 30、根据权利要求 29 所述的装置，其特征在于，所述正向字典树的每个第一节点中存储有所述第一节点对应的单词的词频，所述反向字典树的每个第二节点中存储有所述第二节点对应的单词的词频；

所述词频获取模块具体用于，
从所述第一单词对应的第一节点中获取所述第一单词的词频；
从所述第二单词对应的第二节点中获取所述第二单词的词频。

- 20 31、根据权利要求 30 所述的装置，其特征在于，还包括：语料库构建模块，用于构建语料库，所述语料库包括单词库和所述单词库中的单词的词频；
所述字典树构建模块具体用于，根据所述语料库，构建正向字典树和反向字典树，并将各单词的词频存储至对应的第一节点和第二节点。

- 25 32、根据权利要求 31 所述的装置，其特征在于，所述预设文本包括：满足预设使用条件的文本以及待分割的文本；所述语料库构建模块具体用于，
根据满足预设使用条件的词典，得到单词库；

确定所述单词库中的单词在所述满足预设使用条件的文本以及所述待分割的文本中出现的次数；

- 30 根据所述单词库、所述单词库中的单词在所述满足预设使用条件的文本以及所述待分割的文本中出现的次数，构建所述语料库。

33、根据权利要求 32 所述的装置，其特征在于，所述语料库构建模块具体用于，

根据所述待分割的文本中的空格符，获取至少一个第一字符串；

5 将所述至少一个第一字符串与所述单词库中的单词进行匹配，得到与所述单词库中的单词匹配的至少一个第二字符串；

根据各所述第二字符串在所述待分割的文本中出现的次数，确定单词库中的单词在所述待分割的文本中出现的次数。

34、根据权利要求 28 所述的装置，其特征在于，还包括：反馈模块；

所述文本获取模块具体用于，获取用户设备发送的所述待分割的文本；

10 所述反馈模块用于，向所述用户设备反馈所述待分割的字符串的分割结果信息，所述分割结果信息包括所述待分割的字符串的分割结果，以使所述用户设备向用户输出所述分割结果。

35、根据权利要求 34 所述的装置，其特征在于，还包括：结果获取模块和处理模块，

15 所述结果获取模块用于，获取所述用户设备发送的待处理的分割结果；所述处理模块用于，对所述待处理的分割结果进行自然语言处理。

36、一种字符串的分词装置，其特征在于，包括：

20 发送模块，用于向云端服务器发送用户输入的待分割的文本，以使所述云端服务器获取待分割的字符串，并根据正向分割结果中的各第一单词的词频以及反向分割结果中的各第二单词的词频确定分割结果；

接收模块，用于接收所述云端服务器反馈的所述待分割的字符串的分割结果信息，所述分割结果信息包括所述待分割的字符串的分割结果；其中，所述待分割的字符串的分割结果为所述正向分割结果或所述反向分割结果；

输出模块，用于向用户输出所述分割结果。

25 37、根据权利要求 36 所述的分词装置，其特征在于，所述输出模块具体用于，在显示界面上显示所述分割结果。

38、根据权利要求 37 所述的分词装置，其特征在于，所述分割结果信息中还包括所述分割结果对应的分割类型，所述分割类型为正向分割或反向分割；

30 所述输出模块具体用于，在显示界面上显示所述分割结果以及所述分割

结果的分割类型。

39、根据权利要求 37 所述的分词装置，其特征在于，若所述分割结果为正向分割结果，则所述分割信息中还包括反向分割结果；或者

5 若所述分割结果为反向分割结果，则所述分割信息中还包括正向分割结果；

所述输出模块具体用于，在所述显示界面上显示所述正向分割结果和所述反向分割结果，并标注所述待分割字符串对应的分割结果。

40、根据权利要求 39 所述的分词装置，其特征在于，所述分割信息中还包括所述正向分割结果中的各所述第一单词的词频和所述反向分割结果中的
10 各所述第二单词的词频；

所述显示装置还包括：指令获取模块，用于获取所述用户操作所述显示界面触发的词频显示指令；

所述输出模块还用于，根据所述词频显示指令，显示各所述第一单词的词频和/或各所述第二单词的词频；

15 或者

所述输出模块具体用于，在所述显示界面上显示所述正向分割结果、所述正向分割结果中的第一单词的词频，以及所述反向分割结果、所述反向分割结果中的第二单词的词频。

41、根据权利要求 39 所述的分词装置，其特征在于，所述分割信息中还包括所述正向分割结果中的各所述第一单词对应的第一词频和值以及所述反向分割结果中的各所述第二单词对应的第二词频和值；
20

所述显示装置还包括：指令获取模块，用于获取所述用户操作所述显示界面触发的词频显示指令；

所述输出模块还用于，根据所述词频显示指令，显示所述第一词频和值和/或所述第二词频和值；
25

或者

所述输出模块具体用于，在所述显示界面上显示所述正向分割结果、所述第一词频和值，以及所述反向分割结果、所述第二词频和值。

42、根据权利要求 39 至 41 任一项所述的分词装置，其特征在于，还包括：操作信息获取模块，用于获取所述用户对所述显示界面上的所述正向分
30

割结果或反向分割结果的操作信息，

确定模块，用于根据所述操作信息确定待处理的分割结果；

所述发送模块还用于，向所述云端服务器发送所述待处理的分割结果，以使所述云端服务器对所述待处理的分割结果进行自然语言处理。

5 43、一种字符串的分词设备，其特征在于，包括：

输入设备，用于获取待分割的文本；

10 处理器，耦合至所述输入设备，用于获取待分割的字符串的正向分割结果，所述正向分割结果包括至少一个第一单词，并获取所述待分割的字符串的反向分割结果，所述反向分割结果包括至少一个第二单词；获取各所述第一单词的词频和各所述第二单词的词频，所述词频为预先确定的各单词在预设文本中出现的次数；根据各所述第一单词的词频以及各所述第二单词的词频，确定所述待分割的字符串的分割结果，其中，所述待分割的字符串的分割结果为所述正向分割结果或所述反向分割结果。

44、一种云端服务器，其特征在于，包括：

15 输入设备，用于获取待分割的文本；

20 处理器，耦合至所述输入设备，用于获取待分割的字符串的正向分割结果，所述正向分割结果包括至少一个第一单词，并获取所述待分割的字符串的反向分割结果，所述反向分割结果包括至少一个第二单词；获取各所述第一单词的词频和各所述第二单词的词频，所述词频为预先确定的各单词在预设文本中出现的次数；根据各所述第一单词的词频以及各所述第二单词的词频，确定所述待分割的字符串的分割结果，其中，所述待分割的字符串的分割结果为所述正向分割结果或所述反向分割结果。

45、一种字符串的分词设备，其特征在于，包括：

25 输出设备，用于向云端服务器发送用户输入的待分割的文本，以使所述云端服务器获取待分割的字符串，并根据正向分割结果中的各第一单词的词频以及反向分割结果中的各第二单词的词频确定分割结果；

输入设备，用于接收所述云端服务器反馈的所述待分割的字符串的分割结果信息，所述分割结果信息包括所述待分割的字符串的分割结果；其中，所述待分割的字符串的分割结果为所述正向分割结果或所述反向分割结果；

30 处理器，耦合至所述输出设备和所述输入设备，用于根据所述分割结果

信息，控制所述输入设备向用户输出所述分割结果。

46、一种用户设备，其特征在于，包括：

输出设备，用于向云端服务器发送用户输入的待分割的文本，以使所述云端服务器获取待分割的字符串，并根据正向分割结果中的各第一单词的词频以及反向分割结果中的各第二单词的词频确定分割结果；

输入设备，用于接收所述云端服务器反馈的所述待分割的字符串的分割结果信息，所述分割结果信息包括所述待分割的字符串的分割结果；其中，所述待分割的字符串的分割结果为所述正向分割结果或所述反向分割结果；

处理器，耦合至所述输出设备和所述输入设备，用于根据所述分割结果信息，控制所述输入设备向用户输出所述分割结果。

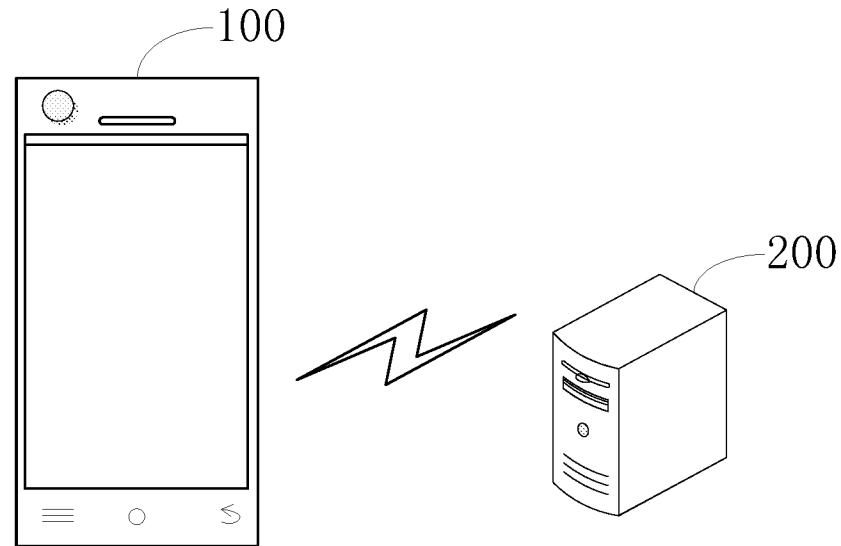


图 1

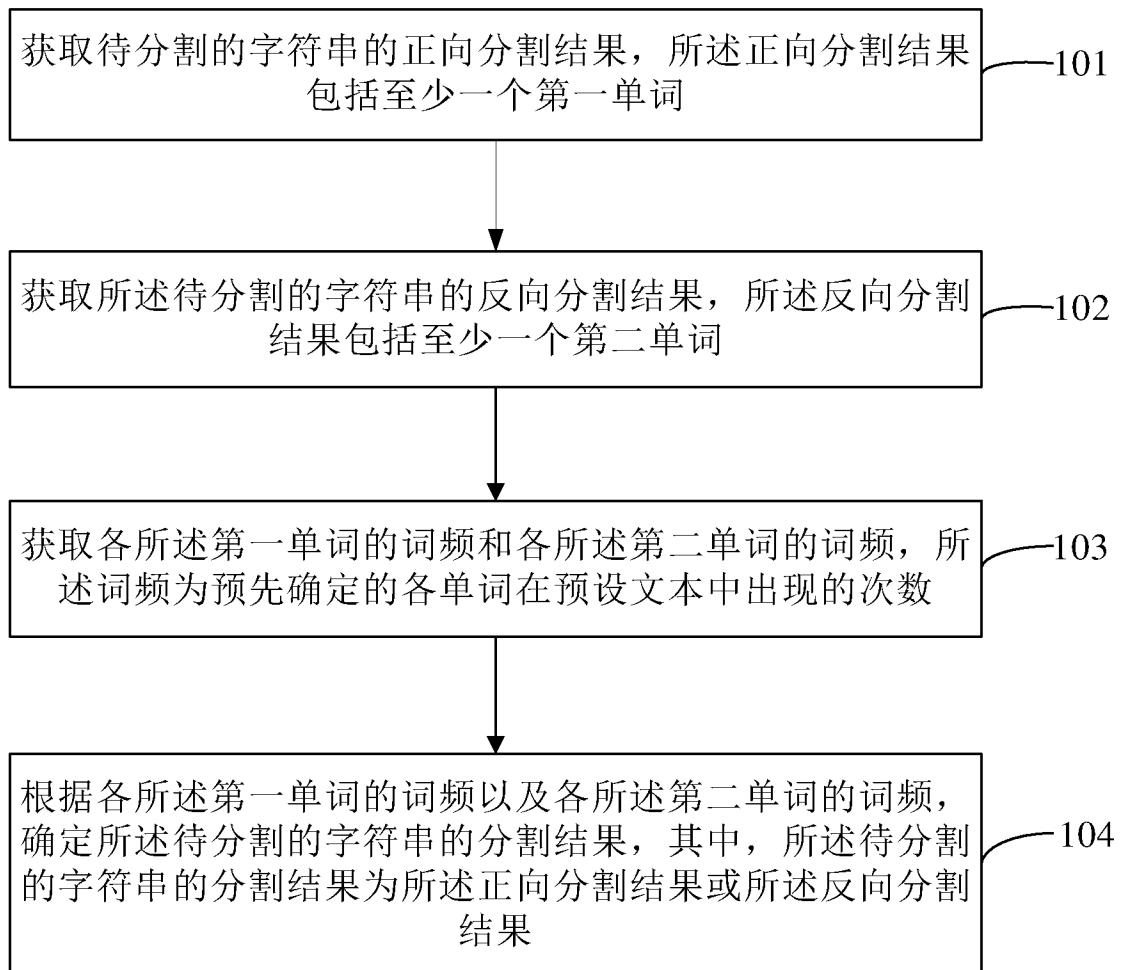


图 2

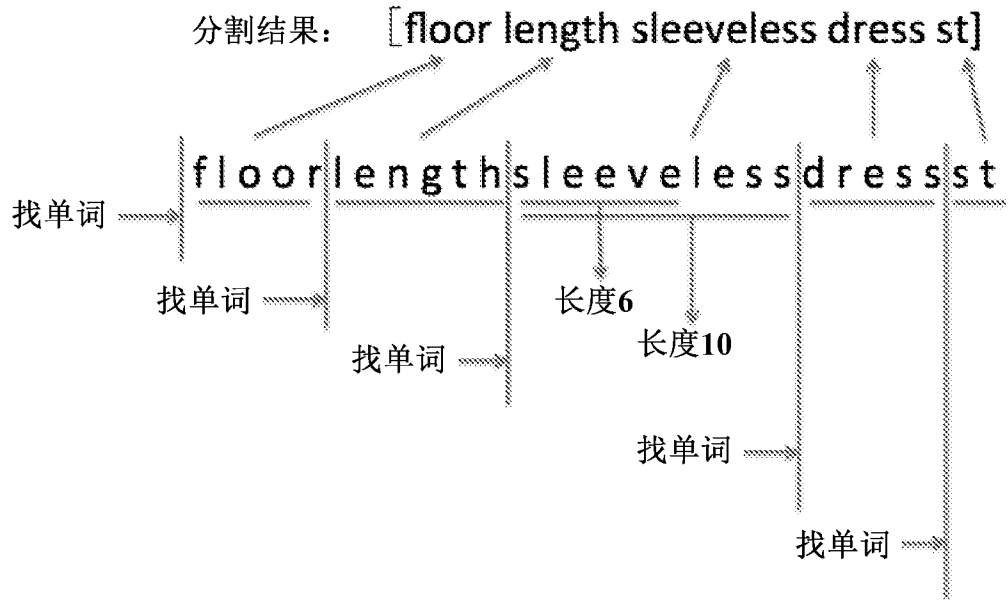


图 3

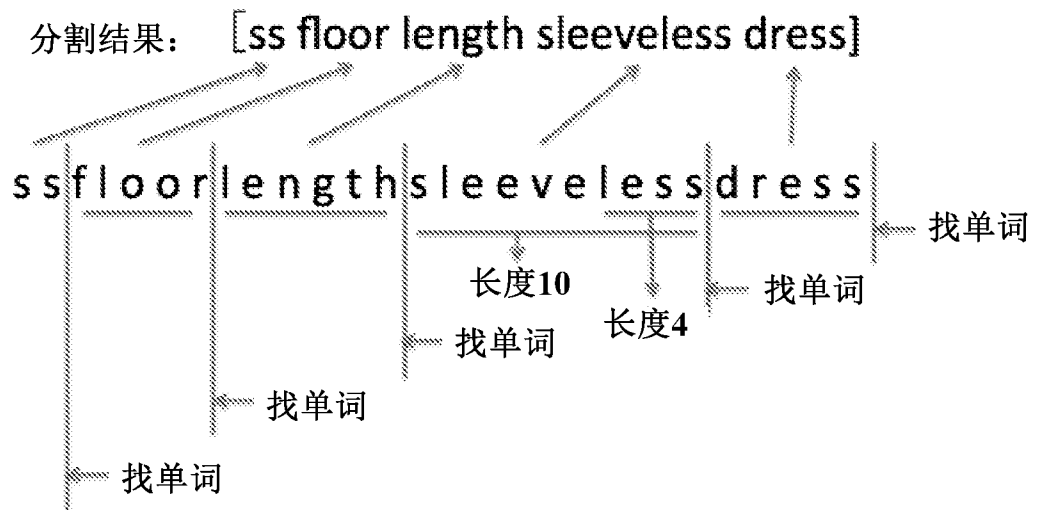


图 4



图 5

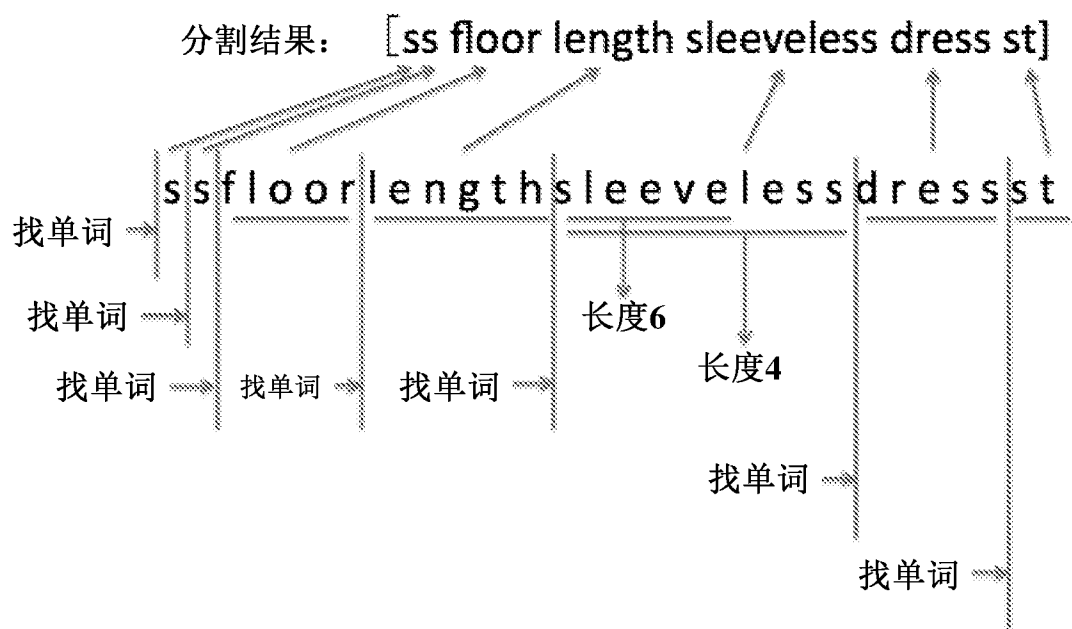


图 6

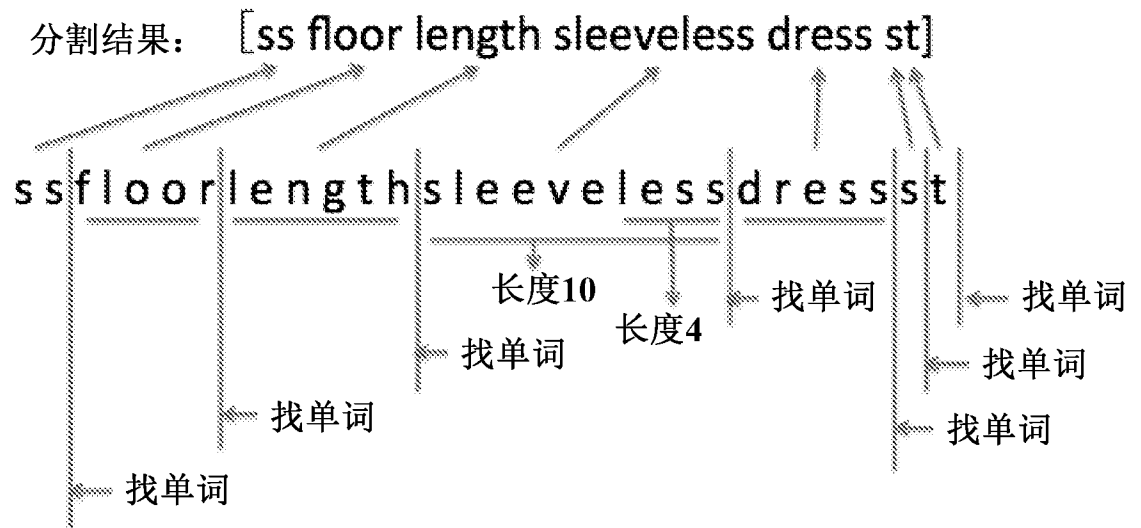


图 7

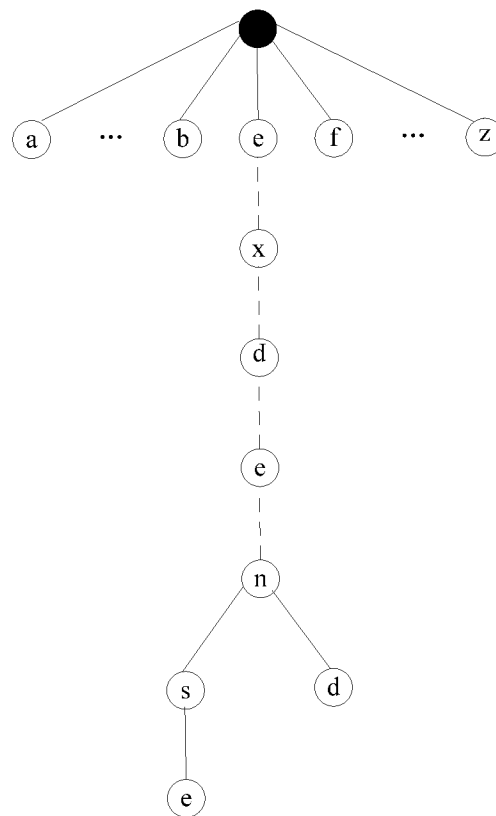


图 8

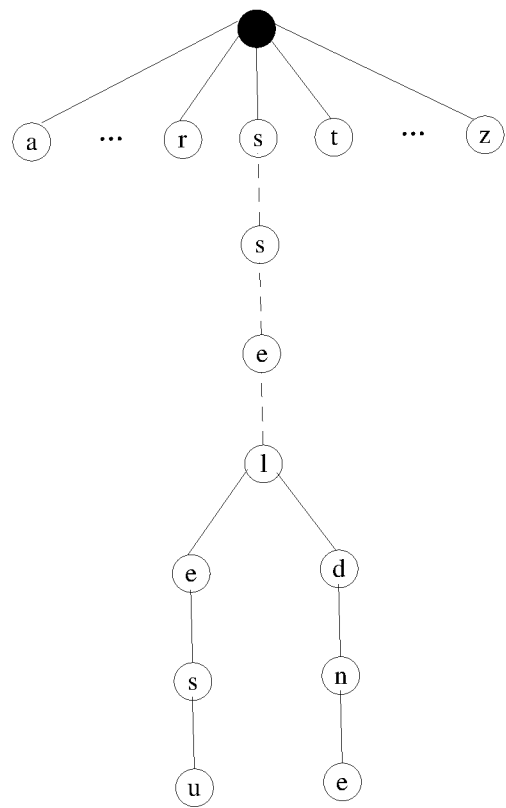


图 9

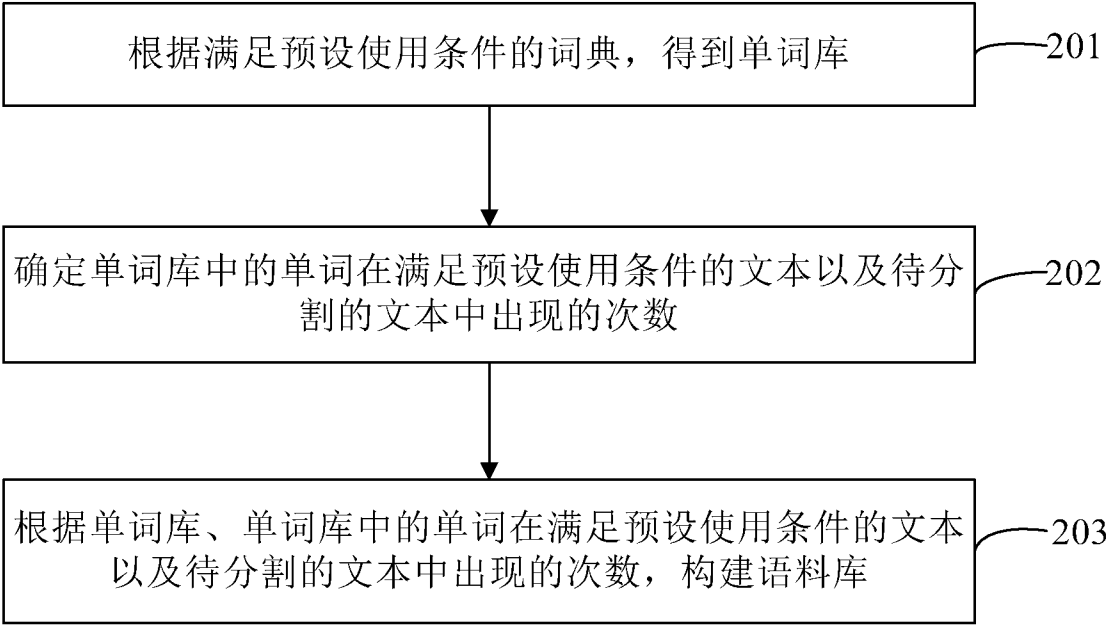


图 10

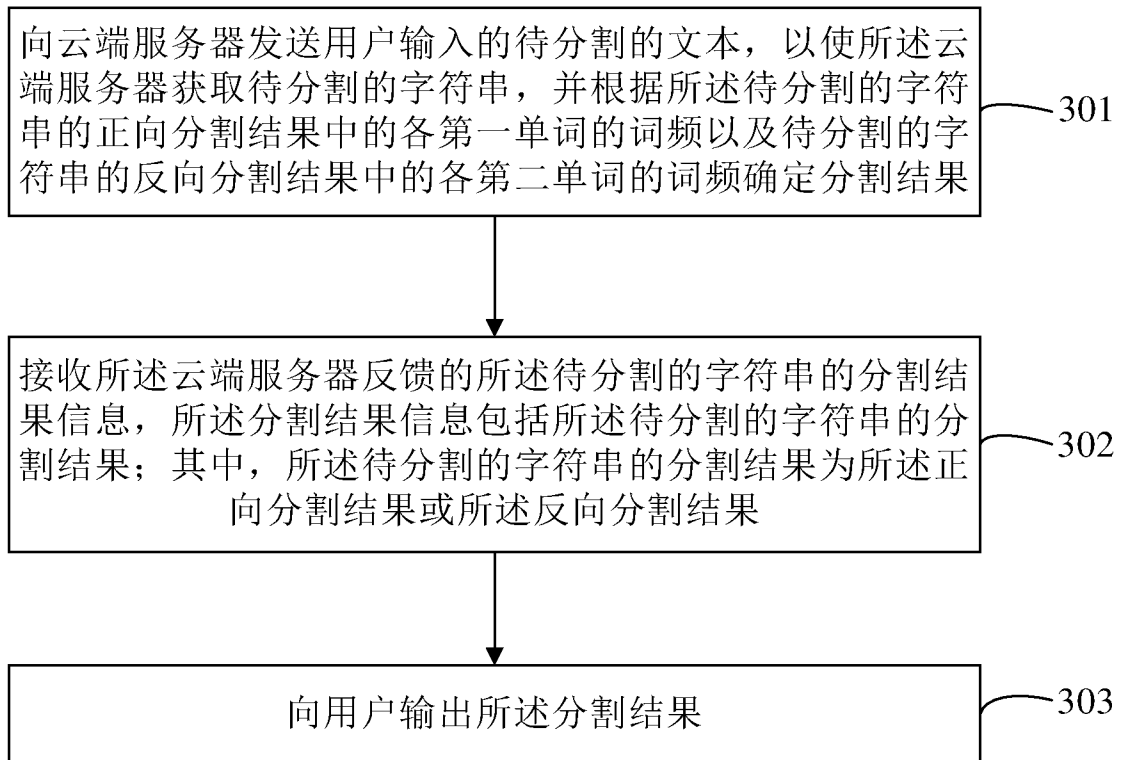


图 11

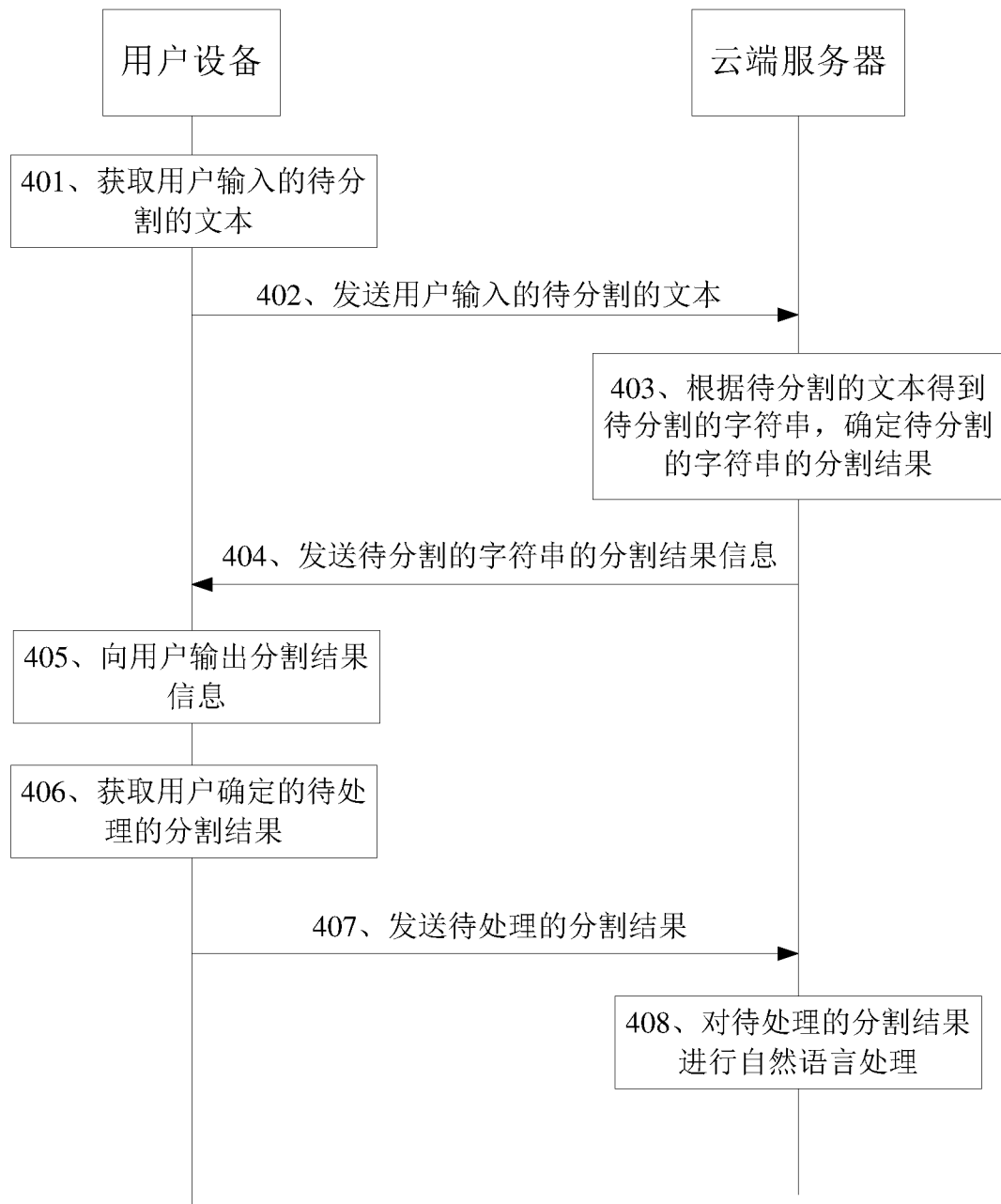


图 12



图 13



图 14



图 15

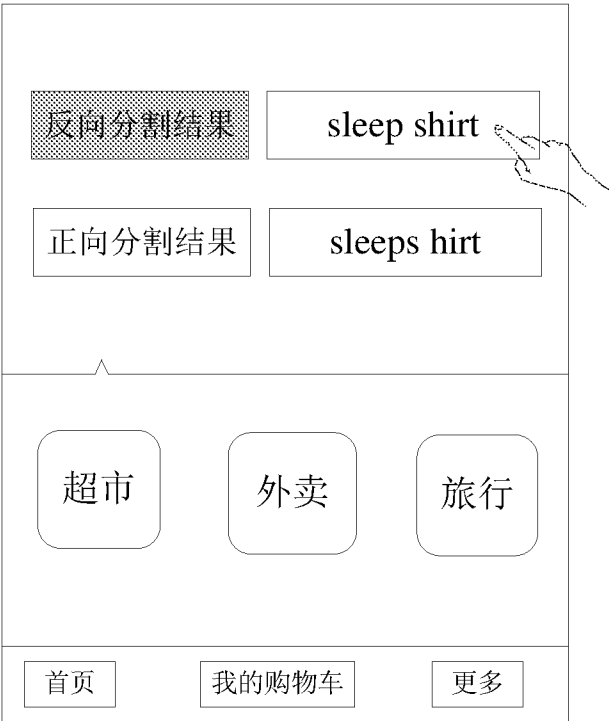


图 16

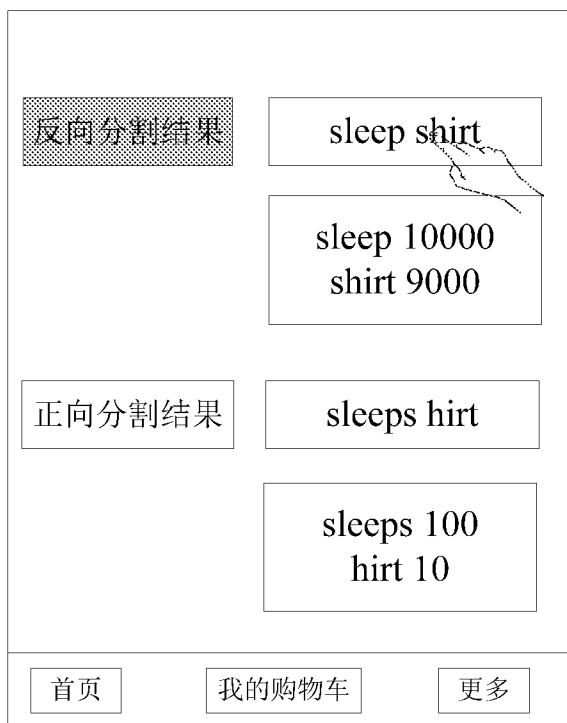


图 17



图 18

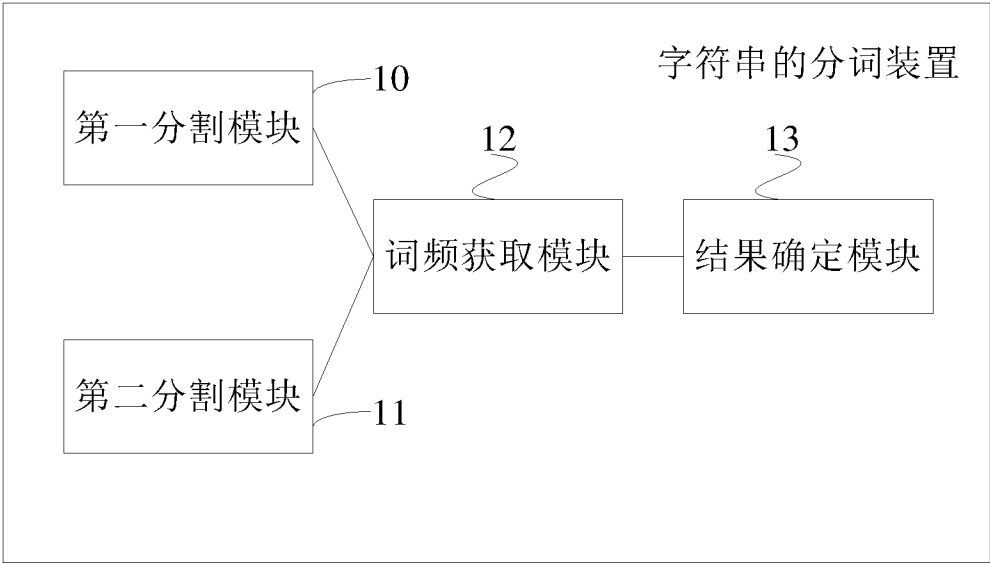


图 19

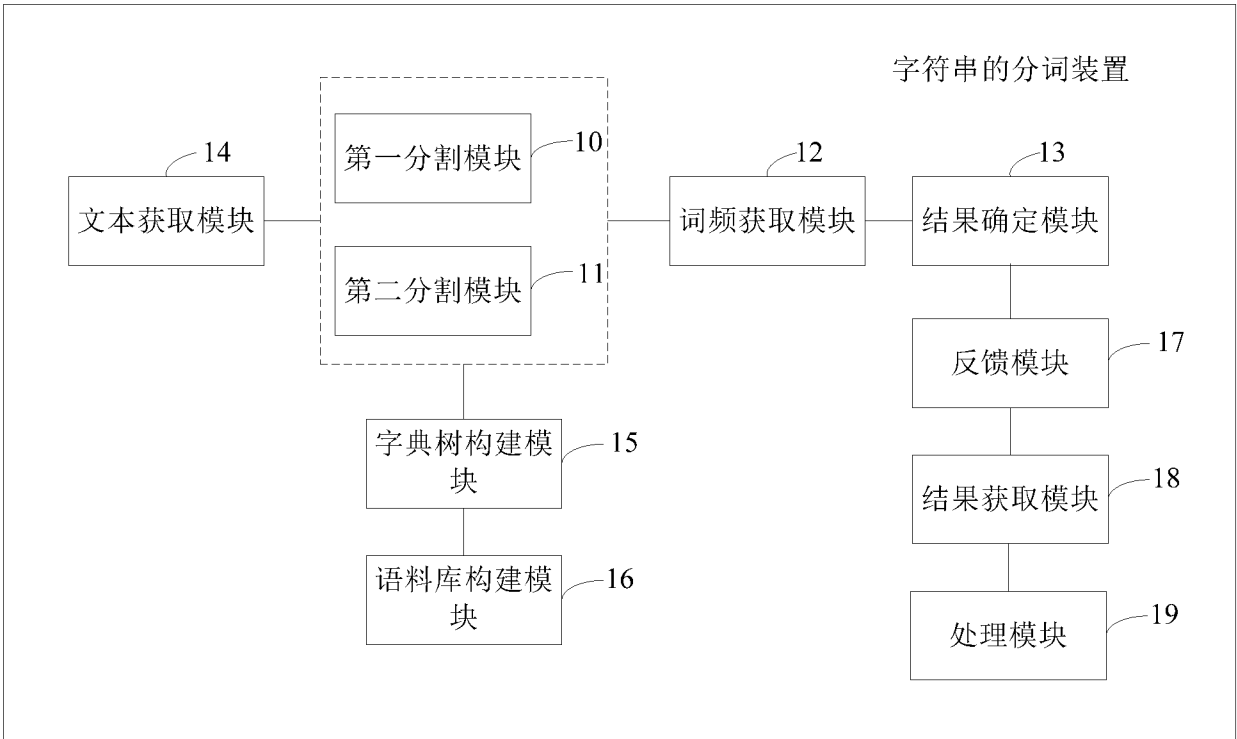


图 20

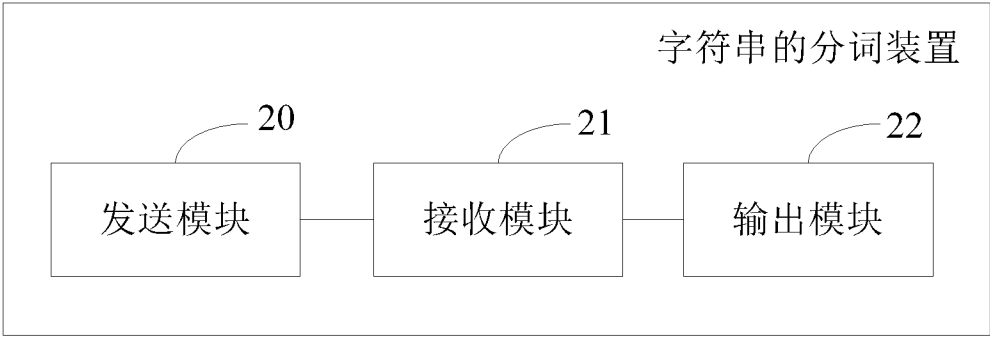


图 21

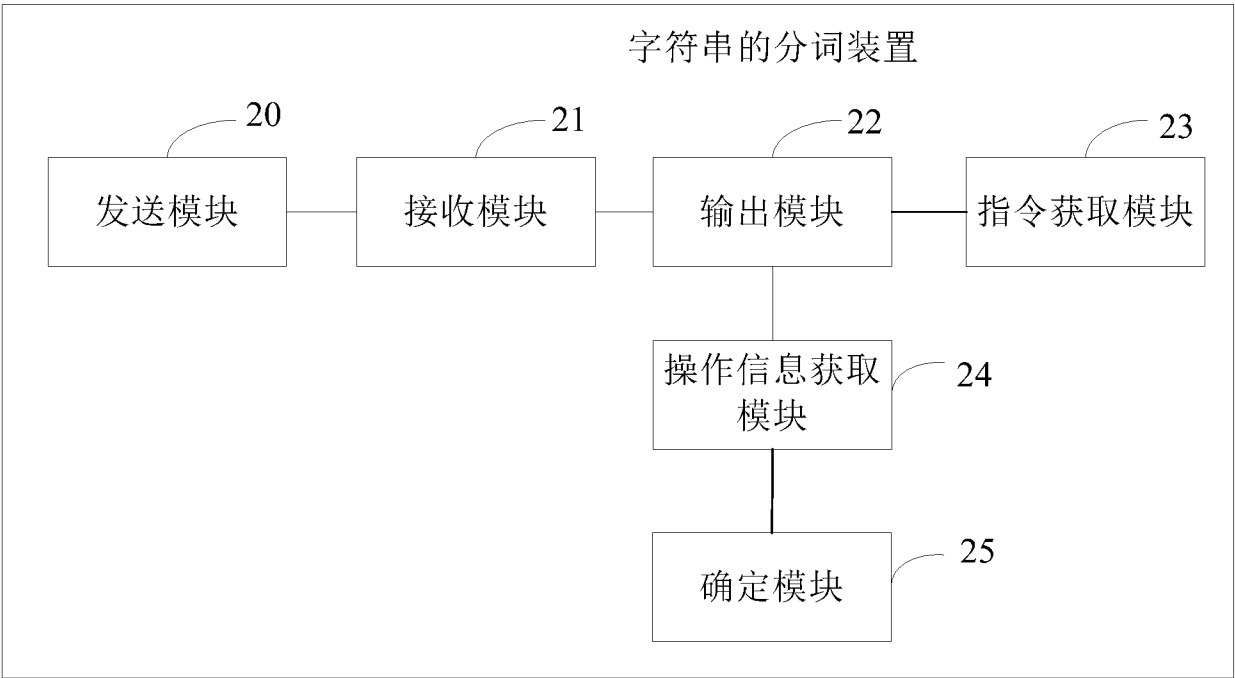


图 22

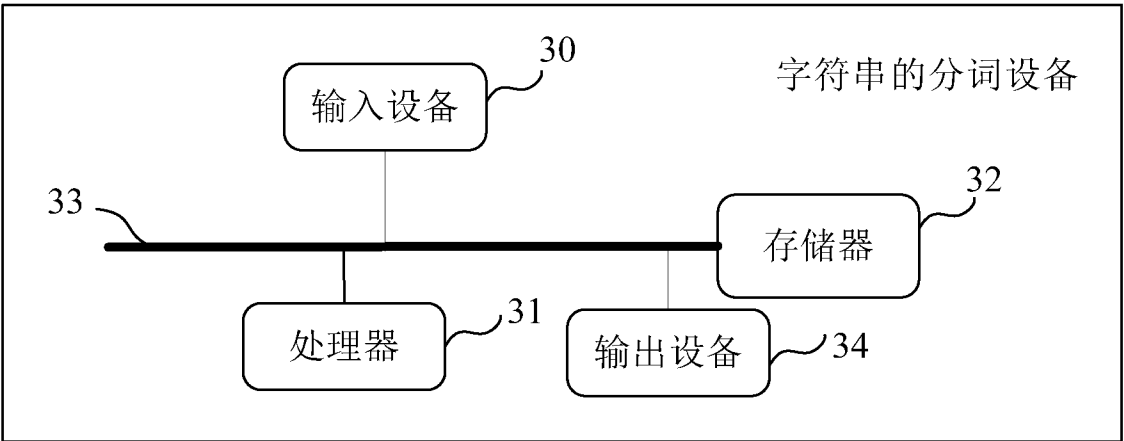


图 23

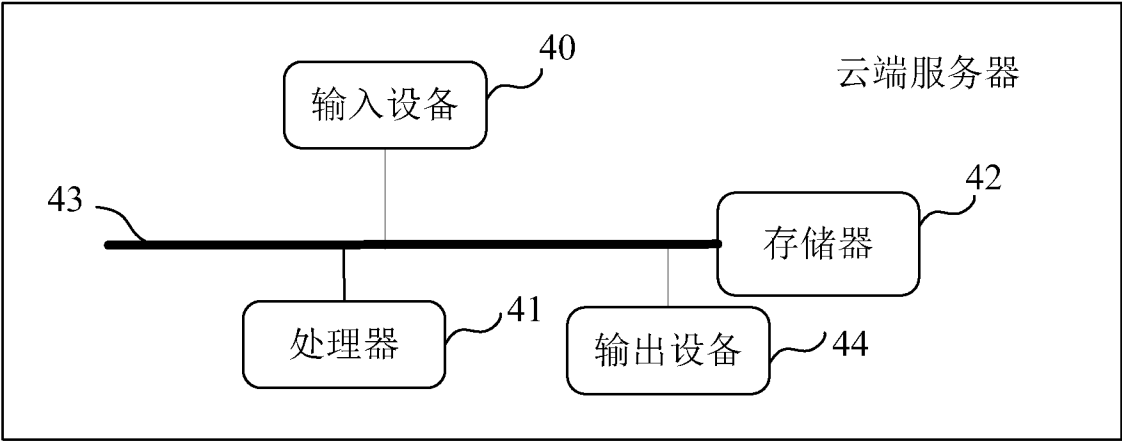


图 24

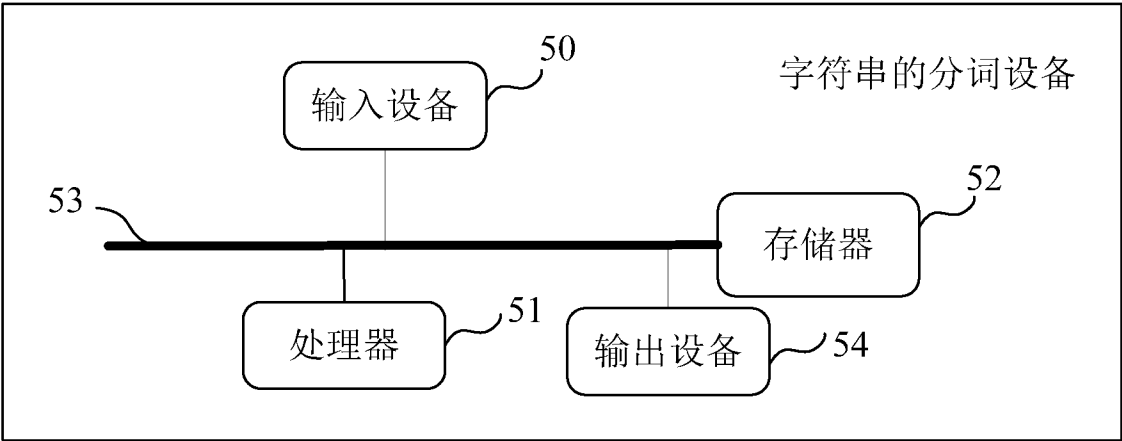


图 25

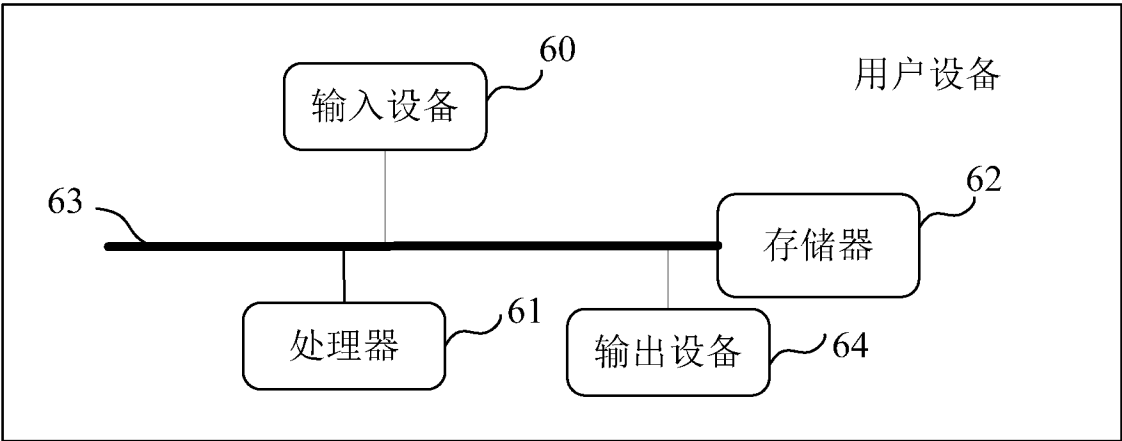


图 26

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2017/091783

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/27 (2006.01) i

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

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)

CNPAT, EPODOC, WPI, GOOGLE, CNKI: participle, identify, ambiguity, segmentation, word, sentence, frequency, forward, backward, dictionary, sequence, compare, correct, string, converse

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	PENG, Qi et al. "A General Method of Chinese Word Segmentation Based on the Resolution of Word Frequency Ambiguity" Journal of Guangxi Normal University (Natural Science Edition), vol. 34, no. 1, 31 March 2016 (31.03.2016), sections 2 and 4.2	1-46
X	ZHANG, Heng et al. "Chinese Word Segmentation Method Based on Dictionary and Frequency of the Words" Microcomputer Information, vol. 24, no. 1-3, 31 December 2008 (31.12.2008), section 3	1-46
A	CN 103646018 A (DALIAN UNIVERSITY) 19 March 2014 (19.03.2014) the whole document	1-46
A	CN 102915299 A (HISENSE GROUP CO., LTD.) 06 February 2013 (06.02.2013) the whole document	1-46

☐ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents:	"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
"A" document defining the general state of the art which is not considered to be of particular relevance	"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
"E" earlier application or patent but published on or after the international filing date	"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
"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)	"&" document member of the same patent family
"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	

Date of the actual completion of the international search 24 September 2017	Date of mailing of the international search report 11 October 2017
Name and mailing address of the ISA State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No. (86-10) 62019451	Authorized officer WANG, Jing Telephone No. (86-10) 62413685

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.
PCT/CN2017/091783

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 103646018 A	19 March 2014	None	
CN 102915299 A	06 February 2013	CN 104765724 A	08 July 2015
		CN 104765838 A	08 July 2015

国际检索报告

国际申请号

PCT/CN2017/091783

A. 主题的分类 G06F 17/27 (2006.01) i 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 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)) CNPAT, EPDOC, WPI, GOOGLE, CNKI: 分词, 词频, 频率, 字频, 前向, 反向, 正向, 逆向, 识别, 歧义, 分割, word, sentence, frequency, forward, backward, dictionary, sequence, compare, correct, string, converse																	
C. 相关文件 <table border="1"> <thead> <tr> <th>类 型*</th> <th>引用文件, 必要时, 指明相关段落</th> <th>相关的权利要求</th> </tr> </thead> <tbody> <tr> <td>X</td> <td>彭琦 等. "一种基于词频歧义消解的通用中文分词法" 广西师范大学学报 (自然科学版), 第34卷, 第1期, 2016年 3月 31日 (2016 - 03 - 31), 第2, 4.2节</td> <td>1-46</td> </tr> <tr> <td>X</td> <td>张恒 等. "基于词典和词频的中文分词方法" 微计算机信息, 第24卷, 第1-3期, 2008年 12月 31日 (2008 - 12 - 31), 第3节</td> <td>1-46</td> </tr> <tr> <td>A</td> <td>CN 103646018 A (大连大学) 2014年 3月 19日 (2014 - 03 - 19) 全文</td> <td>1-46</td> </tr> <tr> <td>A</td> <td>CN 102915299 A (海信集团有限公司) 2013年 2月 6日 (2013 - 02 - 06) 全文</td> <td>1-46</td> </tr> </tbody> </table>			类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求	X	彭琦 等. "一种基于词频歧义消解的通用中文分词法" 广西师范大学学报 (自然科学版), 第34卷, 第1期, 2016年 3月 31日 (2016 - 03 - 31), 第2, 4.2节	1-46	X	张恒 等. "基于词典和词频的中文分词方法" 微计算机信息, 第24卷, 第1-3期, 2008年 12月 31日 (2008 - 12 - 31), 第3节	1-46	A	CN 103646018 A (大连大学) 2014年 3月 19日 (2014 - 03 - 19) 全文	1-46	A	CN 102915299 A (海信集团有限公司) 2013年 2月 6日 (2013 - 02 - 06) 全文	1-46
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求															
X	彭琦 等. "一种基于词频歧义消解的通用中文分词法" 广西师范大学学报 (自然科学版), 第34卷, 第1期, 2016年 3月 31日 (2016 - 03 - 31), 第2, 4.2节	1-46															
X	张恒 等. "基于词典和词频的中文分词方法" 微计算机信息, 第24卷, 第1-3期, 2008年 12月 31日 (2008 - 12 - 31), 第3节	1-46															
A	CN 103646018 A (大连大学) 2014年 3月 19日 (2014 - 03 - 19) 全文	1-46															
A	CN 102915299 A (海信集团有限公司) 2013年 2月 6日 (2013 - 02 - 06) 全文	1-46															
☐ 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 2017年 9月 24日		国际检索报告邮寄日期 2017年 10月 11日															
ISA/CN的名称和邮寄地址 中华人民共和国国家知识产权局 (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 王晶 电话号码 (86-10)62413685															

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2017/091783

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	103646018	A	2014年 3月 19日	无			
CN	102915299	A	2013年 2月 6日	CN	104765724	A	2015年 7月 8日
				CN	104765838	A	2015年 7月 8日