

(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
2015 年 12 月 30 日 (30.12.2015) WIPO | PCT



(10) 国际公布号
WO 2015/196909 A1

- (51) 国际专利分类号:
G06F 17/27 (2006.01) G06F 17/30 (2006.01)
- (21) 国际申请号: PCT/CN2015/080675
- (22) 国际申请日: 2015 年 6 月 3 日 (03.06.2015)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201410302925.1 2014 年 6 月 27 日 (27.06.2014) CN
- (71) 申请人: 北京奇虎科技有限公司 (BEIJING QIHOO TECHNOLOGY COMPANY LIMITED) [CN/CN]; 中国北京市西城区新街口外大街 28 号 D 座 112 室 (德胜园), Beijing 100088 (CN)。奇智软件 (北京) 有限公司 (QIZHI SOFTWARE (BEIJING) COMPANY LIMITED) [CN/CN]; 中国北京市朝阳区酒仙桥路 6 号院 2 号楼 B 座 2 层、3 层 301-306 室, Beijing 100015 (CN)。
- (72) 发明人: 项碧波 (XIANG, Bibo); 中国北京市朝阳区酒仙桥路 6 号院 2 号楼, Beijing 100015 (CN)。

- (74) 代理人: 北京润泽恒知识产权代理有限公司 (BEIJING RISEHIGH INTELLECTUAL PROPERTY LAW FIRM); 中国北京市海淀区中关村南大街 31 号神舟大厦 702, Beijing 100081 (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, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, 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, 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)。

[见续页]

(54) Title: WORD SEGMENTATION METHOD AND DEVICE

(54) 发明名称: 一种分词方法和装置

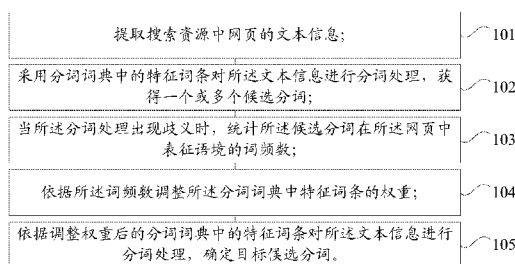


图 1 / FIG. 1

101 EXTRACTING TEXT INFORMATION ABOUT A WEBPAGE IN A SEARCH RESOURCE;
102 CONDUCTING WORD SEGMENTATION PROCESSING ON THE TEXT INFORMATION USING A FEATURE ENTRY IN A WORD SEGMENTATION DICTIONARY, TO OBTAIN ONE OR MORE CANDIDATE SEGMENTED WORDS;
103 WHEN AMBIGUITY OCCURS DURING THE WORD SEGMENTATION PROCESSING, COUNTING THE WORD FREQUENCY NUMBER OF THE CANDIDATE SEGMENTED WORDS CHARACTERIZING CONTEXT IN THE WEBPAGE;
104 ADJUSTING THE WEIGHT OF THE FEATURE ENTRY IN THE WORD SEGMENTATION DICTIONARY ACCORDING TO THE WORD FREQUENCY NUMBER;
105 CONDUCTING WORD SEGMENTATION PROCESSING ON THE TEXT INFORMATION ACCORDING TO THE FEATURE ENTRY WITH ADJUSTED WEIGHT IN THE WORD SEGMENTATION DICTIONARY, TO DETERMINE A TARGET CANDIDATE SEGMENTED WORD.

(57) Abstract: A word segmentation method and device. The method comprises the steps of: extracting text information about a webpage in a search resource (101); conducting word segmentation processing on the text information using a feature entry in a word segmentation dictionary, to obtain one or more candidate segmented words (102); when ambiguity occurs during the word segmentation processing, counting the word frequency number of the candidate segmented words characterizing context in the webpage (103); adjusting the weight of the feature entry in the word segmentation dictionary according to the word frequency number (104); and conducting word segmentation processing on the text information according to the feature entry with adjusted weight in the word segmentation dictionary, to determine a target candidate segmented word (105). The present invention assists word segmentation processing based on context, and fully takes account of the characteristics of natural language, thereby effectively reducing the influence of ambiguity on word segmentation processing, and improving the accuracy rate of word segmentation.

(57) 摘要:

[见续页]



WO 2015/196909 A1

本国际公布:

— 包括国际检索报告(条约第 21 条(3))。

一种分词方法和装置，所述的方法包括步骤：提取搜索资源中网页的文本信息（101）；采用分词词典中的特征词条对所述文本信息进行分词处理，获得一个或多个候选分词（102）；当所述分词处理出现歧义时，统计所述候选分词在所述网页中表征语境的词频数（103）；依据所述词频数调整所述分词词典中特征词条的权重（104）；依据调整权重后的分词词典中的特征词条对所述文本信息进行分词处理，确定目标候选分词（105）。基于语境辅助分词处理，充分考虑了自然语言的特性，有效降低歧义对分词处理的影响，提升了分词的准确率。

一种分词方法和装置

技术领域

本发明涉及分词技术领域，尤其涉及一种分词方法和一种分词装置。

背景技术

5 随着互联网的高速发展，网络应用趋向多元化，网上的信息量急剧增加。

在各种场合下，用户经常需要输入关键信息进行关联信息的获取。例如，在搜索引擎中输入关键词搜索网页信息、在论坛中输入关键词搜索帖子等等。

10 分词是进行信息处理、信息检索的基础，所有的信息处理、信息检索工作都是在分词之后进行的。所以分词的错误会被叠加到后续的处理过程中，而且很难被消除。正因为这样所以对分词的准确率的追求是个持续的过程，同时由于中文语言的固有的特性：没有明确的词的定义、词和词之间没有分隔符、新词、专有名词不断涌现等这些因素导致分词很难做到 100% 的准确。

15 当前分词系统使用的主要方法是基于统计的分词。从形式上看，词是稳定的字的组合，因此在上下文中，相邻的字同时出现的次数越多，就越有可能构成一个词。因此字与字相邻共现的频率或概率能够较好的反映成词的可信度。可以对语料中相邻共现的各个字的组合的频度进行统计，计算它们的互现信息。定义两个字的互现信息，计算两个汉字 X、Y 的相邻共现概率。

20 互现信息体现了汉字之间结合关系的紧密程度。当紧密程度高于某一个阈值时，便可认为此字组可能构成了一个词。这种方法只需对语料中的字组频度进行统计，但这种方法也有一定的局限性，会经常抽出一些共现频度高、但并不是词的常用字组，例如“这一”、“之一”、“有的”、“我的”、“许多的”等，并且对常用词的识别精度差，时空开销大。

25 分词的结果错误致使后期获取的关联信息与当初的预期有很大差别，用户体验十分差，浪费了设备系统资源。

发明内容

鉴于上述问题，提出了本发明以便提供一种克服上述问题或者至少部分地解决或减缓上述问题的一种分词方法和相应的一种分词装置。

5 根据本发明的一个方面，提供了一种分词方法，包括：

提取搜索资源中网页的文本信息；

采用分词词典中的特征词条对所述文本信息进行分词处理，获得一个或多个候选分词；

10 当所述分词处理出现歧义时，统计所述候选分词在所述网页中表征语境的词频数；

依据所述词频数调整所述分词词典中特征词条的权重；

依据调整权重后的分词词典中的特征词条对所述文本信息进行分词处理，确定目标候选分词。

根据本发明的另一方面，提供了一种分词装置，包括：

15 文本信息提取模块，适于提取搜索资源中网页的文本信息；

第一分词处理模块，适于采用分词词典中的特征词条对所述文本信息进行分词处理，获得一个或多个候选分词；

词频数统计模块，适于在所述分词处理出现歧义时，统计所述候选分词在所述网页中表征语境的词频数；

20 权重调整模块，适于依据所述词频数调整所述分词词典中特征词条的权重；

第二分词处理模块，适于依据调整权重后的分词词典中的特征词条对所述文本信息进行分词处理，确定目标候选分词。

25 根据本发明的又一个方面，提供了一种计算机程序，其包括计算机可读代码，当所述计算机可读代码在计算设备上运行时，导致所述计算设备执行上述的分词方法。

根据本发明的再一个方面，提供了一种计算机可读介质，其中存储了上述的计算机程序。

本发明的有益效果为：

本发明实施例在采用分词词典中的特征词条对文本信息进行分词处理出现歧义时，基于候选分词在网页中表征语境的词频数调整分词词典中特征词条的权重，以及依据调整权重后特征词条对文本信息进行分词处理，确定
5 目标候选分词，基于语境辅助分词处理，充分考虑了自然语言的特性，有效降低歧义对分词处理的影响，提升了分词的准确率。

上述说明仅是本发明技术方案的概述，为了能够更清楚了解本发明的技术手段，而可依照说明书的内容予以实施，并且为了让本发明的上述和其它目的、特征和优点能够更明显易懂，以下特举本发明的具体实施方式。

10 附图说明

通过阅读下文优选实施方式的详细描述，各种其他的优点和益处对于本领域普通技术人员将变得清楚明了。附图仅用于示出优选实施方式的目的，而并不认为是对本发明的限制。而且在整个附图中，用相同的参考符号表示相同的部件。在附图中：

15 图 1 示意性示出了根据本发明一个实施例的一种分词方法实施例的步骤流程图；以及

图 2 示意性示出了根据本发明一个实施例的一种分词装置实施例的结构框图；

20 图 3 示意性地示出了用于执行根据本发明的方法的计算设备的框图；以及

图 4 示意性地示出了用于保持或者携带实现根据本发明的方法的程序代码的存储单元。

具体实施例

下面结合附图和具体的实施方式对本发明作进一步的描述。

25

参照图 1，示出了根据本发明一个实施例的一种分词方法实施例的步骤流程图，具体可以包括如下步骤：

步骤 101, 提取提取搜索资源中网页的文本信息;

在具体实现中, 可以通过爬虫 (Spider) 等方式抓取网页并保存在网页数据库中形成大量的搜索资源, 而网页内容中可以包括大量的文本信息。则在本发明实施例中, 可以从网页数据库中提取搜索资源中网页内容
5 的文本信息。

步骤 102, 采用分词词典中的特征词条对所述文本信息进行分词处理, 获得一个或多个候选分词;

分词处理是自然语言处理中最基本的一个步骤, 在实际中产生了很多不同的分词方法, 这些方法大体上可以分为两类: 基于词典匹配的方法和基于
10 统计的方法。

基于词典的方法可以是利用分词词典作为主要的资源, 采用机械匹配法的基本思想, 即事先建立一个分词词典 (又称词库), 其中可以包含可能出现的词条, 对给定的待分词的文本信息, 按照某种确定的分词规则切取文本信息中的字符串, 若该字符串与分词词典中的某词条相匹配, 则可以认为该
15 字符串是候选分词, 插入切分标志, 继续分割剩余的部分, 直到剩余部分为空; 否则, 可以认为该字符串不是候选分词, 重新切取文本信息的字符串进行下一次匹配。

在本发明的一种可选实施例中, 所述候选分词可以包括第一候选分词和第二候选分词中的至少一个; 则在本发明实施例中, 步骤 102 可以包括如下
20 子步骤:

子步骤 S11, 按照第一分词规则对所述文本信息进行与所述特征词条进行匹配;

子步骤 S12, 当所述文本信息进行与所述特征词条进行匹配时, 提取所述特征词条作为第一候选分词;

25 子步骤 S13, 按照第二分词规则对所述文本信息进行与所述特征词条进行匹配;

子步骤 S14, 当所述文本信息进行与所述特征词条进行匹配时, 提取所述特征词条作为第二候选分词;

其中，所述第一分词规则为与所述第二分词规则不同的分词规则。

基于分词词典的分词方法的通常可以包括三个要素，分别为分词词典、文本信息扫描顺序和匹配规则。

其中，分词词典通常可以包括通用的分词词典和专业的分词词典等等。

5 通用的分词词典可以是适用于通用领域的分词词典，即没有特定的领域中的技术术语划分，其主要包含通用的、出现频率高于第一预设阈值的词条。具体地，通用的分词词典中可以包括通用词和确定意义词。通用词可以包括形容词、连词和一些通用意义的动词，例如“高兴”、“但是”、“参加”等。

专业的分词词典可以为适用于某个特定领域的分词词典，其主要包含某
10 个特定领域中出现频率高于第二预设阈值的词条，比如中草药领域的当归、草乌等。

文本信息扫描顺序可以包括正向扫描、逆向扫描和双向扫描等等。

正向扫描可以是指从待切分的文本信息的第一个字符往后扫描，而逆向扫描可以是指从待切分的文本信息的最后一个字符串往前扫描。双向扫描则
15 可以是正向扫描和逆向扫描的组合。

匹配规则可以包括最大匹配、最小匹配、逐词匹配和最佳匹配等。

常见的基于词典的分词规则可以包括正向最大匹配法(MM)、反向最大匹配法(RMM)、双向扫描法、逐词遍历法、最佳匹配法等等、切分标志法、有穷多层列举法等等。

20 以下以第一分词规则为正向最大匹配法、第二分词规则为反向最大匹配法为例进行说明。

正向最大匹配法的目的是将最长的复合词从文本信息中分离出来。

具体而言，假定分词词典中的最长词条长度是k，则用待分词的文本信息中前k个字符串作为匹配字段查找分词词典，若词典中存在这样一个长度为k的词条，则匹配成功。匹配成功的字符串被作为第一候选分词被切分出来。
25

如果分词词典中找不到这样的—个长度为k的词条，则匹配失败。将待匹配的字符串中的最后一个字符去掉，对剩下的字符串重新进行匹配处理，

如此进行下去，直到匹配成功，切分出一个第一候选分词或剩余字符串的长度为零为止。这样就完成了一轮匹配，然后取下一个长度为 k 的字符串进行匹配处理，直到文本信息被扫描完为止。

5 反向最大匹配法和正向最大匹配法的思想相似，不同之处在于它是从句子的最后的 k 个字符串开始切分，每次匹配不成功时，去掉字符串前面的一个字符。

当然，上述分词规则只是作为示例，在实施本发明实施例时，可以根据实际情况设置其他分词规则，本发明实施例对此不加以限制。另外，除了上述分词规则外，本领域技术人员还可以根据实际需要采用其它分词规则，本
10 发明实施例对此也不加以限制。

步骤 103，当所述分词处理出现歧义时，统计所述候选分词在所述网页中表征语境的词频数；

自然语言可以指人类历史发展过程中自然产生、约定俗成的用于人类社会交际的语言，如汉语、英语、俄语等。

15 一个句子在给定的语言环境中，一般认为只有一种分析结果是正确的，但是在其他语言环境中，则可能如果出现不同的分析结果。

例如，对于文本信息“我等你开车”，可以存在两种理解，一种理解是等你来了我再开车（即我开车），另一种理解是我在等着你开车（即你开车）。

若给定一部合理的语法和一个句子，通过自然语言处理按照这部语法可能给出一种以上的分析结果，那么这些分析结果中则存在分析歧义。
20

例如，据统计，正向最大匹配法的差错率大概为 $1/169$ ，而逆向最大匹配法的差错率大概为 $1/245$ ，即逆向最大匹配法和正向最大匹配法都存在不同程度的误差。

例如，对于文本信息“小面的”，单从这三个字是无从判断到底是“小面/的”，还是“小/面的”，或者其他的词语组成。因此，在分词处理时，可能获得“小面/的”、“小/面的”或者其他的分词结果。
25

在本发明的一种可选实施例中，所述表征语境的词频数可以包括所述第一候选分词在所述网页的文本语言环境中的第一词频数，所述第二候选分词

在所述网页的文本语言环境中的第二词频数;则在本发明实施例中,步骤 103 可以包括如下子步骤:

子步骤 S21, 当对同一个文本信息进行分词处理所获得的所述第一候选分词和所述第二候选分词不同时, 判断所述分词处理出现歧义;

- 5 子步骤 S22, 分别统计所述第一候选分词在所述网页中的文本语言环境中的第一词频数, 以及, 所述第二候选分词在所述网页中的文本语言环境中的第二词频数。

本发明实施例中, 可以在分词处理的过程中充分利用文本信息所在网页中的文本语言环境辅助进行消歧 (消除歧义), 该文本语言环境可以表现为
10 短语、句子、段落、文章等形式。

例如, 对于文本信息“小面的”, 基于第一分词规则 (如正向最大匹配法) 所获得的第一候选分词为“小面”和“的”, 基于第二分词规则 (如反向最大匹配法) 所获得的第二候选分词为“小”和“面的”, 即第一候选分词与第二候选分词不同, 出现歧义。

- 15 如果把文本信息“小面的”放入更完整的句子 (文本语言环境) 中可以有助于消歧。结合上下文可知, “小面的”是从“重庆小面的做法”中提取的。显然, “重庆小面的做法”比“小面的”的语义更为清楚, 使得分词处理更加准确。

“重庆小面的做法”这个短语对于人来理解一般比较充分了, 但是对于
20 基于统计的分词处理来说, 一般是不够的。原因主要有两个, 第一个原因是, “小面”可能是个很不常见的词; 第二个原因是, “面的”的使用频率要比“小面”高不少。

因此, 本示例中可以采用更大范围的文本语言环境, 如文章, 以统计“小面”和“面的”的词频数。

- 25 以下截取“小面的”所在文章 (一篇讲解如何烹饪重庆小面的文章) 的其中一段为例进行说明:

“重庆人喜食麻辣, 小面可谓集川菜麻辣之大成, 一碗好吃的小面必是有一碗好吃的作料方能吸引食客, 必备的作料有九种: 酱油、味精、油辣子

海椒、花椒面、姜蒜水、猪油、葱花、榨菜粒、芝麻酱。”

其中，第一候选分词“小面”的第一词频数为 2，第二候选分词“面的”的第二词频数为 0。

步骤 104，依据所述词频数调整所述分词词典中特征词条的权重；

5 一个作者在撰写一篇文章的时候，通常会有一个确定的主题，即一个确定的文本语言环境，使得语义明确。

一篇文章中描述的内容是有连贯性的，所以文章中相关的词语会在短语、句子、段落、文章等频繁出现。

在本发明的一种可选实施例中，步骤 104 可以包括如下子步骤：

10 子步骤 S31，当所述第一词频数大于所述第二词频数时，提高所述第一候选分词对应的特征词条在所述分词词典中的权重；

子步骤 S32，当所述第二词频数大于所述第一词频数时，提高所述第二候选分词对应的特征词条在所述分词词典中的权重。

15 例如，对于文本信息“小面的”，在其所属文章中，“小面”的词频数会有大幅的提升，远高于“面的”的词频数，则可以相对应提高“小面”的权重。

需要说明的是，若第一候选分词或第二候选分词适配于文本信息所在的文本语言环境时，第一候选分词或第二候选分词对应的特征词条的权重会提高，但并不是提高在分词词典中的特征词条的权重，而是相当于提高分词词典中应用于该文本信息的特征词条适配的权重。

20

步骤 105，依据调整权重后的分词词典中的特征词条对所述文本信息进行分词处理，确定目标候选分词。

在调整权重后，可以继续的分词处理。

在本发明的一种可选实施例中，步骤 105 可以包括如下子步骤：

25 子步骤 S41，按照指定的分词规则对所述文本信息与提高权重后的所述特征词条进行匹配；

子步骤 S42，当所述文本信息进行与所述特征词条进行匹配时，提取所述特征词条作为目标候选分词。

指定的分词规则可以为第一分词规则、第二分词规则，也可以为其他的分词规则，本发明实施例对此不加以限制。

例如，对于文本信息“小面的”，“面的”的使用频率要比“小面”高，使得“面的”的权重初始要比“小面”高，但是在基于文本语言环境提高了“小面”的权重后，可以获得“小面的”的分词结果“小面”和“的”，以实现歧义消除。

本发明实施例在采用分词词典中的特征词条对文本信息进行分词处理出现歧义时，基于候选分词在网页中表征语境的词频数调整分词词典中特征词条的权重，以及依据调整权重后特征词条对文本信息进行分词处理，确定目标候选分词，基于语境辅助分词处理，充分考虑了自然语言的特性，有效降低歧义对分词处理的影响，提升了分词的准确率。

对于方法实施例，为了简单描述，故将其都表述为一系列的动作组合，但是本领域技术人员应该知悉，本发明实施例并不受所描述的动作顺序的限制，因为依据本发明实施例，某些步骤可以采用其他顺序或者同时进行。其次，本领域技术人员也应该知悉，说明书中所描述的实施例均属于优选实施例，所涉及的动作并不一定是本发明实施例所必须的。

参照图 2，示出了根据本发明一个实施例的一种分词装置实施例的结构框图，具体可以包括如下模块：

文本信息提取模块 201，适于提取搜索资源中网页的文本信息；

第一分词处理模块 202，适于采用分词词典中的特征词条对所述文本信息进行分词处理，获得一个或多个候选分词；

词频数统计模块 203，适于在所述分词处理出现歧义时，统计所述候选分词在所述网页中表征语境的词频数；

权重调整模块 204，适于依据所述词频数调整所述分词词典中特征词条的权重；

第二分词处理模块 205，适于依据调整权重后的分词词典中的特征词条

对所述文本信息进行分词处理，确定目标候选分词。

在本发明的一种可选实施例中，所述候选分词可以包括第一候选分词和第二候选分词中的至少一个；

所述第一分词处理模块 202 还可以适于：

5 按照第一分词规则对所述文本信息进行与所述特征词条进行匹配；

当所述文本信息进行与所述特征词条进行匹配时，提取所述特征词条作为第一候选分词；

按照第二分词规则对所述文本信息进行与所述特征词条进行匹配；

10 当所述文本信息进行与所述特征词条进行匹配时，提取所述特征词条作为第二候选分词；

其中，所述第一分词规则为与所述第二分词规则不同的分词规则。

在本发明的一种可选实施例中，所述表征语境的词频数可以包括所述第一候选分词在所述网页的文本语言环境中的第一词频数，所述第二候选分词在所述网页的文本语言环境中的第二词频数；

15 所述词频数统计模块 203 还可以适于：

当对同一个文本信息进行分词处理所获得的所述第一候选分词和所述第二候选分词不同时，判断所述分词处理出现歧义；

20 分别统计所述第一候选分词在所述网页中的文本语言环境中的第一词频数，以及，所述第二候选分词在所述网页中的文本语言环境中的第二词频数。

在本发明的一种可选实施例中，所述权重调整模块 204 还可以适于：

当所述第一词频数大于所述第二词频数时，提高所述第一候选分词对应的特征词条在所述分词词典中的权重；

25 当所述第二词频数大于所述第一词频数时，提高所述第二候选分词对应的特征词条在所述分词词典中的权重。

在本发明的一种可选实施例中，所述第二分词处理模块 205 还可以适于：

按照指定的分词规则对所述文本信息与提高权重后的所述特征词条进行匹配；

当所述文本信息进行与所述特征词条进行匹配时，提取所述特征词条作为目标候选分词。

对于装置实施例而言，由于其与方法实施例基本相似，所以描述的比较简单，相关之处参见方法实施例的部分说明即可。

5

本发明的各个部件实施例可以以硬件实现，或者以在一个或者多个处理器上运行的软件模块实现，或者以它们的组合实现。本领域的技术人员应当理解，可以在实践中使用微处理器或者数字信号处理器（DSP）来实现根据本发明实施例的分词设备中的一些或者全部部件的一些或者全部功能。本发
10 明还可以实现为用于执行这里所描述的方法的一部分或者全部的设备或者装置程序（例如，计算机程序和计算机程序产品）。这样的实现本发明的程序可以存储在计算机可读介质上，或者可以具有一个或者多个信号的形式。这样的信号可以从因特网网站上下载得到，或者在载体信号上提供，或者以任何其他形式提供。

15 例如，图 3 示出了可以实现根据本发明的分词计算设备，例如检索服务器。该计算设备传统上包括处理器 310 和以存储器 320 形式的计算机程序产品或者计算机可读介质。存储器 320 可以是诸如闪存、EEPROM（电可擦除可编程只读存储器）、EPROM、硬盘或者 ROM 之类的电子存储器。存储器 320 具有用于执行上述方法中的任何方法步骤的程序代码
20 331 的存储空间 330。例如，用于程序代码的存储空间 330 可以包括分别用于实现上面的方法中的各种步骤的各个程序代码 331。这些程序代码可以从一个或者多个计算机程序产品中读出或者写入到这一个或者多个计算机程序产品中。这些计算机程序产品包括诸如硬盘，紧致盘（CD）、存储卡或者软盘之类的程序代码载体。这样的计算机程序产品通常为如参
25 考图 4 所述的便携式或者固定存储单元。该存储单元可以具有与图 3 的计算设备中的存储器 320 类似布置的存储段、存储空间等。程序代码可以例如以适当形式进行压缩。通常，存储单元包括计算机可读代码 331'，即可以由例如诸如 310 之类的处理器读取的代码，这些代码当由计算设备运行时，导致该计算设备执行上面所描述的方法中的各个步骤。

本文中所称的“一个实施例”、“实施例”或者“一个或者多个实施例”意味着，结合实施例描述的特定特征、结构或者特性包括在本发明的至少一个实施例中。此外，请注意，这里“在一个实施例中”的词语例子不一定全指同一个实施例。

5 在此处所提供的说明书中，说明了大量具体细节。然而，能够理解，本发明的实施例可以在没有这些具体细节的情况下被实践。在一些实例中，并未详细示出公知的方法、结构和技术，以便不模糊对本说明书的理解。

10 应该注意的是上述实施例对本发明进行说明而不是对本发明进行限制，并且本领域技术人员在不脱离所附权利要求的范围的情况下可设计出替换实施例。在权利要求中，不应将位于括号之间的任何参考符号构造成对权利要求的限制。单词“包含”不排除存在未列在权利要求中的元件或步骤。位于元件之前的单词“一”或“一个”不排除存在多个这样的元件。本发明可以借助于包括有若干不同元件的硬件以及借助于适当编程的计
15 算机来实现。在列举了若干装置的单元权利要求中，这些装置中的若干个可以通过同一个硬件项来具体体现。单词第一、第二、以及第三等的使用不表示任何顺序。可将这些单词解释为名称。

此外，还应当注意，本说明书中使用的语言主要是为了可读性和教导的目的而选择的，而不是为了解释或者限定本发明的主题而选择的。
20 因此，在不偏离所附权利要求书的范围和精神的情况下，对于本技术领域的普通技术人员来说许多修改和变更都是显而易见的。对于本发明的范围，对本发明所做的公开是说明性的，而非限制性的，本发明的范围由所附权利要求书限定。

权 利 要 求 书

1、一种分词方法，包括步骤：

提取搜索资源中网页的文本信息；

采用分词词典中的特征词条对所述文本信息进行分词处理，获得一个或多个候选分词；

5 当所述分词处理出现歧义时，统计所述候选分词在所述网页中表征语境的词频数；

依据所述词频数调整所述分词词典中特征词条的权重；

依据调整权重后的分词词典中的特征词条对所述文本信息进行分词处理，确定目标候选分词。

10 2、如权利要求 1 所述的方法，其特征在于，所述候选分词包括第一候选分词和第二候选分词中的至少一个；

所述采用分词词典中的特征词条对所述文本信息进行分词处理，获得一个或多个候选分词的步骤包括：

按照第一分词规则对所述文本信息进行与所述特征词条进行匹配；

15 当所述文本信息进行与所述特征词条进行匹配时，提取所述特征词条作为第一候选分词；

按照第二分词规则对所述文本信息进行与所述特征词条进行匹配；

当所述文本信息进行与所述特征词条进行匹配时，提取所述特征词条作为第二候选分词；

20 其中，所述第一分词规则为与所述第二分词规则不同的分词规则。

3、如权利要求 1 或 2 所述的方法，其特征在于，所述表征语境的词频数包括所述第一候选分词在所述网页的文本语言环境中的第一词频数，所述第二候选分词在所述网页的文本语言环境中的第二词频数；

25 所述当所述分词处理出现歧义时，统计所述候选分词在所述网页中表征语境的词频数的步骤包括：

当对同一个文本信息进行分词处理所获得的所述第一候选分词和所述第二候选分词不同时，判断所述分词处理出现歧义；

分别统计所述第一候选分词在所述网页中的文本语言环境中的第一词

频数，以及，所述第二候选分词在所述网页中的文本语言环境中的第二词频数。

4、如权利要求 3 所述的方法，其特征在于，所述依据所述词频数调整所述分词词典中特征词条的权重的步骤包括：

5 当所述第一词频数大于所述第二词频数时，提高所述第一候选分词对应的特征词条在所述分词词典中的权重；

当所述第二词频数大于所述第一词频数时，提高所述第二候选分词对应的特征词条在所述分词词典中的权重。

10 5、如权利要求 4 所述的方法，其特征在于，所述依据调整权重后的分词词典中的特征词条对所述文本信息进行分词处理，确定目标候选分词的步骤包括：

按照指定的分词规则对所述文本信息与提高权重后的所述特征词条进行匹配；

15 当所述文本信息进行与所述特征词条进行匹配时，提取所述特征词条作为目标候选分词。

6、一种分词装置，包括：

文本信息提取模块，适于提取搜索资源中网页的文本信息；

20 第一分词处理模块，适于采用分词词典中的特征词条对所述文本信息进行分词处理，获得一个或多个候选分词；

词频数统计模块，适于在所述分词处理出现歧义时，统计所述候选分词在所述网页中表征语境的词频数；

权重调整模块，适于依据所述词频数调整所述分词词典中特征词条的权重；

25 第二分词处理模块，适于依据调整权重后的分词词典中的特征词条对所述文本信息进行分词处理，确定目标候选分词。

7、如权利要求 6 所述的装置，其特征在于，所述候选分词包括第一候选分词和第二候选分词中的至少一个；

所述第一分词处理模块还适于：

按照第一分词规则对所述文本信息进行与所述特征词条进行匹配;

当所述文本信息进行与所述特征词条进行匹配时,提取所述特征词条作为第一候选分词;

按照第二分词规则对所述文本信息进行与所述特征词条进行匹配;

5 当所述文本信息进行与所述特征词条进行匹配时,提取所述特征词条作为第二候选分词;

其中,所述第一分词规则为与所述第二分词规则不同的分词规则。

8、如权利要求 6 或 7 所述的装置,其特征在于,所述表征语境的词频数包括所述第一候选分词在所述网页的文本语言环境中的第一词频数,所述
10 第二候选分词在所述网页的文本语言环境中的第二词频数;

所述词频数统计模块还适于:

当对同一个文本信息进行分词处理所获得的所述第一候选分词和所述第二候选分词不同时,判断所述分词处理出现歧义;

分别统计所述第一候选分词在所述网页中的文本语言环境中的第一词
15 频数,以及,所述第二候选分词在所述网页中的文本语言环境中的第二词频数。

9、如权利要求 8 所述的装置,其特征在于,所述权重调整模块还适于:

当所述第一词频数大于所述第二词频数时,提高所述第一候选分词对应的特征词条在所述分词词典中的权重;

20 当所述第二词频数大于所述第一词频数时,提高所述第二候选分词对应的特征词条在所述分词词典中的权重。

10、如权利要求 9 所述的装置,其特征在于,所述第二分词处理模块还适于:

25 按照指定的分词规则对所述文本信息与提高权重后的所述特征词条进行匹配;

当所述文本信息进行与所述特征词条进行匹配时,提取所述特征词条作为目标候选分词。

11、一种计算机程序,包括计算机可读代码,当所述计算机可读代

码在计算设备上运行时,导致所述计算设备执行根据权利要求 1-5 中的一个所述的分词方法。

12、一种计算机可读介质,其中存储了如权利要求 11 所述的计算机程序。

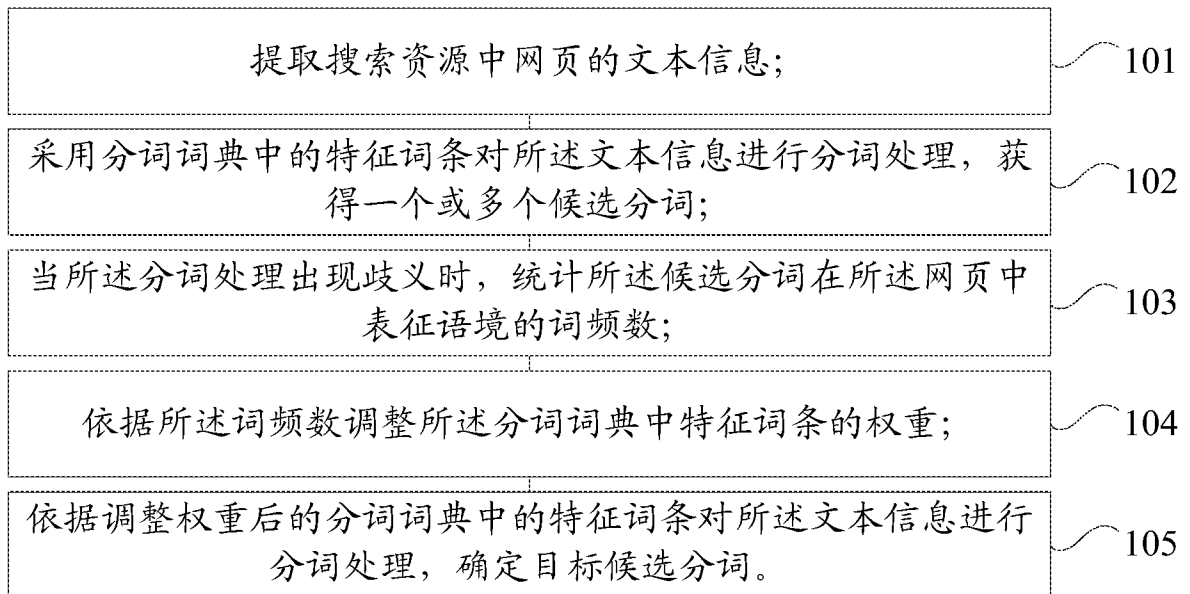


图 1

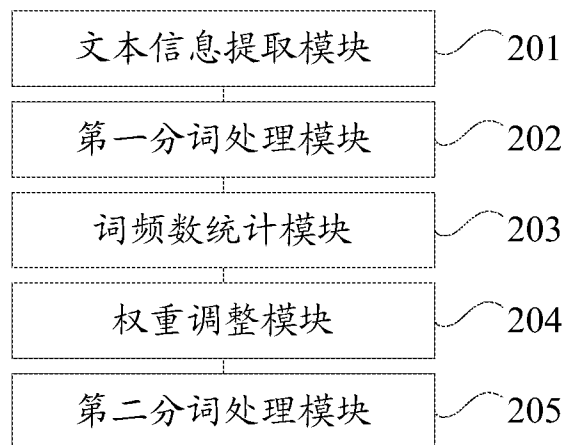


图 2

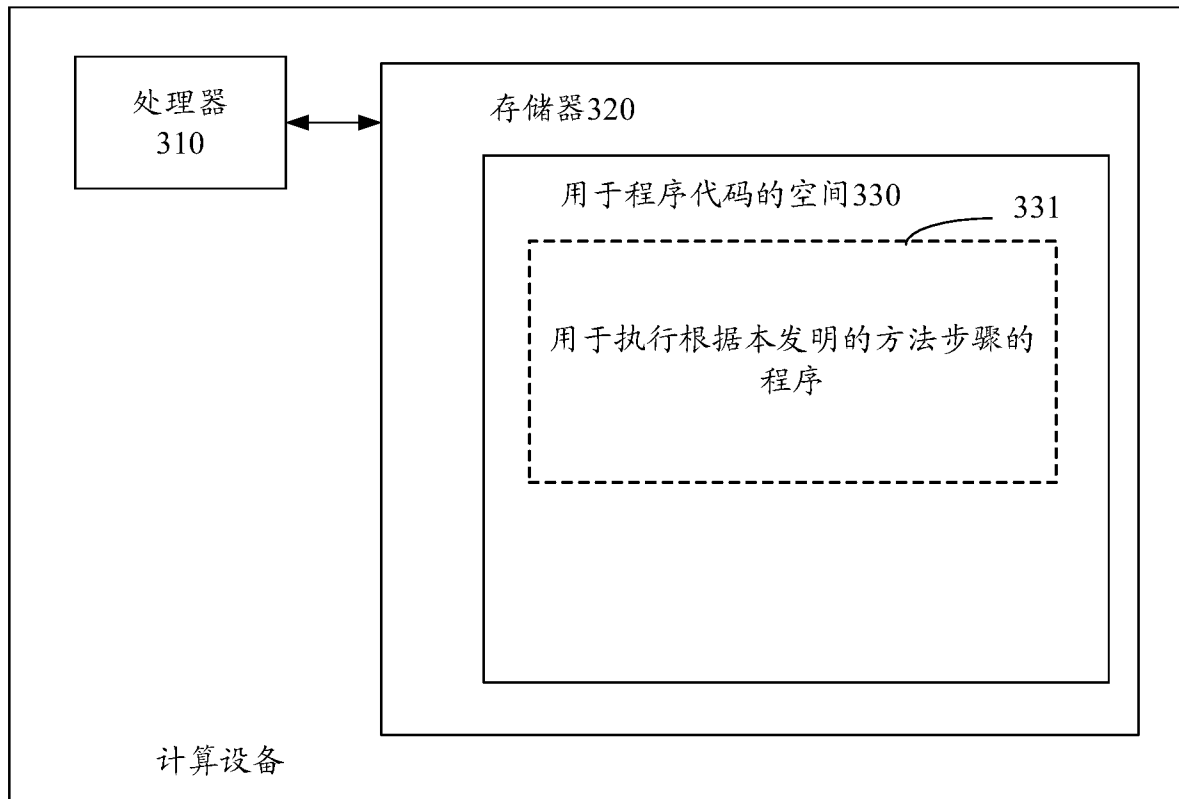


图 3

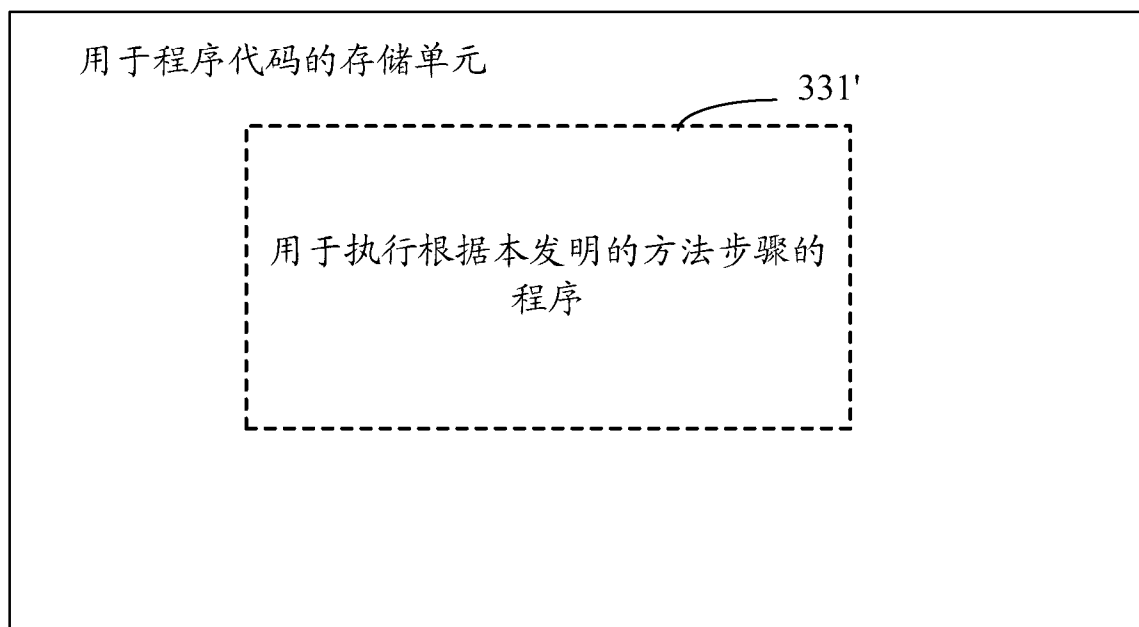


图 4

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2015/080675

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/27 (2006.01) i; G06F 17/30 (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

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, WPI, EPODOC, CNKI, GOOGLE, IEEE: word segmentation, dictionary, text information, ambiguity, word frequency, probability, word, segmentation, context, different, meaning, frequency, number, weight, statistic

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 104077275 A (BEIJING QIHOO TECHNOLOGY CO., LTD. et al.), 01 October 2014 (01.10.2014), claims 1-10, and description, paragraph 150	1-12
X	CN 101499058 A (BEIJING INSTITUTE OF TECHNOLOGY), 05 August 2009 (05.08.2009), description, page 3, paragraph 3 to page 13, paragraph 4	1-12
A	CN 103440256 A (SHENZHEN INSTITUTES OF ADVANCED TECHNOLOGY, CHINESE ACADEMY OF SCIENCES), 11 December 2013 (11.12.2013), the whole document	1-12
A	CN 102402502 A (QUNAR.COM INFORMATION TECHNOLOGY CO., LTD.), 04 April 2012 (04.04.2012), the whole document	1-12
A	CN 103365974 A (BAIDU ONLINE NETWORK TECHNOLOGY (BEIJING) CO., LTD.), 23 October 2013 (23.10.2013), the whole document	1-12

☐ 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 06 August 2015 (06.08.2015)	Date of mailing of the international search report 02 September 2015 (02.09.2015)
Name and mailing address of the ISA/CN: 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 ZHAO, Tianqi Telephone No.: (86-10) 61648112

International application No.
PCT/CN2015/080675

Form PCT/ISA/210 (patent family annex) (July 2009)

国际检索报告

国际申请号

PCT/CN2015/080675

A. 主题的分类

G06F 17/27 (2006.01)i; G06F 17/30 (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

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, WPI, EPODOC, CNKI, GOOGLE, IEEE: 分词, 词典, 文本信息, 歧义, 词频, 语境, 权重, 频率, 概率, 统计, word, segmentation, context, different, meaning, frequency, number, weight, statistic

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 104077275 A (北京奇虎科技有限公司等) 2014年 10月 1日 (2014 - 10 - 01) 权利要求1-10, 说明书第150段	1-12
X	CN 101499058 A (北京理工大学) 2009年 8月 5日 (2009 - 08 - 05) 说明书第3页第3段-第13页第4段	1-12
A	CN 103440256 A (中国科学院深圳先进技术研究院) 2013年 12月 11日 (2013 - 12 - 11) 全文	1-12
A	CN 102402502 A (北京趣拿信息技术有限公司) 2012年 4月 4日 (2012 - 04 - 04) 全文	1-12
A	CN 103365974 A (百度在线网络技术北京有限公司) 2013年 10月 23日 (2013 - 10 - 23) 全文	1-12

☐ 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

2015年 8月 6日

国际检索报告邮寄日期

2015年 9月 2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)
北京市海淀区蓟门桥西土城路6号
100088 中国

传真号 (86-10)62019451

授权官员

赵天奇

电话号码 (86-10)61648112

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2015/080675

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	104077275	A	2014年 10月 1日	无	
CN	101499058	A	2009年 8月 5日	无	
CN	103440256	A	2013年 12月 11日	无	
CN	102402502	A	2012年 4月 4日	无	
CN	103365974	A	2013年 10月 23日	无	

表 PCT/ISA/210 (同族专利附件) (2009年7月)