(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
2014 年 1 月 3 日 (03.01.2014)

WIPO | PCT

(10) 国际公布号
WO 2014/000571 A1

(51) 国际专利分类号:
G06F 17/30 (2006.01)

(21) 国际申请号: PCT/CN2013/077102

(22) 国际申请日: 2013 年 6 月 9 日 (09.06.2013)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201210214385.2 2012 年 6 月 25 日 (25.06.2012) 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（德胜园区）, 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 14 号兆维大厦 4 层东侧单元, Beijing 100016 (CN)。

(72) 发明人: **王志刚 (WANG, Zhigang)**; 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: **北京智汇东方知识产权代理事务所（普通合伙） (WISEAST INTELLECTUAL PROPERTY LAW FIRM)**; 中国北京市海淀区成府路 28 号优盛大厦 D-1111 室, Beijing 100083 (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, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, 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, 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, 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,

[见续页]

(54) **Title**: RECOGNITION SYSTEM AND RECOGNITION METHOD OF NON-BODY TEXT IN WEBPAGE

(54) **发明名称** : 一种网页中非正文文本的识别系统及方法

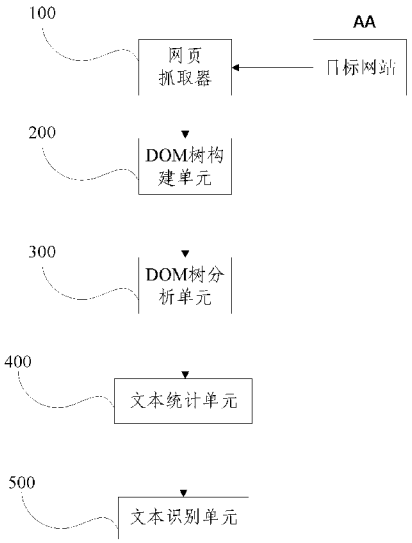

图 2 / **Fig. 2**

100 Webpage grabber
200 DOM tree construction unit
300 DOM tree analysis unit
400 Text statistic unit
500 Text recognition unit
AA Target website

(57) **Abstract**: Disclosed are a recognition system and a recognition method of the non-body text in a webpage, which relate to the field of text extraction. The system comprises: a webpage grabber, used for grabbing webpage data a target website; DOM tree construction unit, used for constructing DOM trees that each webpage in the target website corresponds to; a DOM tree analysis unit, used for finding a unit text section in the webpage according to the DOM trees; a text statistic unit; used for conducting statistics on occurrence number in the webpage of a target website of the unit text section; a text recognition unit, used for recognizing the unit text section into a non-body text when the occurrence times are larger than a preset threshold value. The system and the method overcome the problem of recognition lag of the non-body texts in an existing method and have high recognition accuracy.

(57) **摘要**: 本发明公开了一种网页中非正文文本的识别系统及方法，涉及正文抽取领域。所述系统包括：网页抓取器适于抓取目标网站的所有网页的数据； DOM 树构建单元，适于构建目标网站的每个网页对应的 DOM 树； DOM 树分析单元，适于根据 DOM 树找出网页中的单元文本段；文本统计单元，适于统计单元文本段在目标网站的所有网页中的出现次数；文本识别单元，适于在所述出现次数大于预定阈值时，将所述单元文本段识别为非正文文本。所述系统及方法，克服了现有方法对于非正文文本识别的滞后问题，并且具有较高的识别准确度。

WO 2014/000571 A1

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))。

一种网页中非正文文本的识别系统及方法

技术领域

本发明涉及正文抽取技术领域，特别涉及一种网页中非正文文本的识别系统及方法。

背景技术

正文抽取在搜索引擎、移动阅读等领域有着越来越重要的作用。正文抽取常用的技术有基于规则、基于 DOM（Document Object Model，文档对象模型）树、基于标记窗、基于最大文本块等。这些方法都需要对网站中的非正文文本，比如广告、网站声明等，进行排除。如图 1a 所示是一段网站声明的代码段示意图，图 1b 是图 1a 中代码段在网页中的实际展示效果图，这种网站声明在网页中非常常见，对于用户阅读价值不大，都需要在正文抽取时进行排除。然而，如何有效地识别这些非正文文本，是一个难题。

现有技术中主要采用垃圾关键词密度的方法进行非正文文本识别。在基于垃圾关键词识别非正文文本时，需要有垃圾关键词构成的词典，并且不断对词典进行更新。对于词典的更新，只有在发现问题后才能加入新的垃圾关键词。因此，这种方法对于问题的解决具有严重的滞后性，而且当面对全互联网的海量数据时，这种滞后性显的更加突出。

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者减缓上述问题的网页中非正文文本的识别系统及方法。

根据本发明的一个方面，提供了一种网页中非正文文本的识别系统，所述系统包括：网页抓取器、DOM 树构建单元、DOM 树分析单元、文本统计单元和文本识别单元；所述网页抓取器适于抓取目标网站的所有网页的数据；所述 DOM 树构建单元，适于构建目标网站的每个网页对应的 DOM 树；所述 DOM 树分析单元，适于根据所述 DOM 树找出所述网页中的单元文本段；所述文本统计单元，适于统计所述单元文本段在所述目标网站的所有网页中的出现次数；所述文本识别单元，适于在所述出现次数大于预定阈值时，将所述单元文本段识别为非正文文本。

根据本发明的另一个方面，提供了一种网页中非正文文本的识别方法，其包括步骤：

A：使用网页抓取器抓取目标网站的所有网页的数据；

B：构建目标网站的每个网页对应的DOM树；

C：根据所述DOM树找出所述网页中的单元文本段；

D：统计所述单元文本段在所述目标网站的所有网页中的出现次数；

E：根据所述出现次数，判断所述单元文本段是否是非正文文本。

其中，所述步骤 B 中，采用网页语言对应的解析器构建目标网站的每个网页对应的 DOM 树。

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服务器上运行时，导致所述服务器执行根据权利要求 9-16 中的任一个所述的网页中非正文文本的识别方法。

根据本发明的再一个方面，提供了一种计算机可读介质，其中存储了如权利要求 17 所述的计算机程序。

本发明的有益效果为：

依据本发明实施例的网页中非正文文本的识别系统及方法，采用 DOM 树获取每个网页的单元文本段，进而对单元文本段在目标网站中出现的次数进行统计，并将其出现次数与预定阈值进行比较，从而识别出网站上的非正文文本，克服了现有方法对于非正文文本识别的滞后问题，并且具有较高的识别准确度，在搜索引擎中具有广泛的应用前景。

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。

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：

图1a是一段网站声明的代码段示意图；

图1b是图1a中代码段在网页中的实际展示效果图；

图2是本发明实施例所述网页中非正文文本的识别系统的模块结构图;

图3是所述DOM树分析单元的内部模块结构图;

图4是所述文本统计单元的内部模块结构图;

图5是本发明实施例的网页中非正文文本的识别方法的流程图;

图6是HTML网页对应的DOM树示意图;

图7是所述步骤C的流程图;

图8是所述步骤D的流程图;

图 9 是单元文本段统计原理示意图;

图 10 示意性地示出了用于执行根据本发明的方法的服务器的框图;以及

图 11 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。

图 2 是本发明实施例所述网页中非正文文本的识别系统的模块结构图，如图 2 所示，所述系统包括：网页抓取器 100、DOM 树构建单元 200、DOM 树分析单元 300、文本统计单元 400 和文本识别单元 500。

所述网页抓取器 100 适于抓取目标网站的所有网页的数据，其一般可以采用网络蜘蛛、网页爬虫、搜索机器人或网络抓取脚本程序。

所述 DOM 树构建单元 200，适于构建所述目标网站的每个网页对应的 DOM 树。所述 DOM 树构建单元采用网页语言对应的解析器，并且对于 HTML 网页，采用 HTML 解析器。

所述 DOM 树分析单元 300，适于根据所述 DOM 树找出所述网页中的单元文本段。

图 3 是所述 DOM 树分析单元的内部模块结构图，如图 3 所示，所述 DOM 树分析单元 300 包括：节点访问模块 301 和文本段划分模块 302。

所述节点访问模块 301，适于从所述 DOM 树的根节点开始，依次访问所述 DOM 树中的节点，并将访问到的当前节点发送给所述文本段划分模块 302；以及适于判断所述 DOM 树中是否还有需要访问的节点，如果是，将相应的节点作为新的当前节点发送给所述文本段划分模块 302。

其中，所述节点访问模块 301 可以按照广度优先算法、深度优先算法或

者其他搜索算法访问所述DOM树中的节点。由于所述DOM树分析单元300在判定当前节点对应的文本段为单元文本段时，将停止对当前节点的后代节点的访问，因此，所述节点访问模块301，优选适于按照广度优先算法依次访问所述DOM树中的节点，并将访问到的当前节点发送给所述文本段划分模块302；以及适于判断所述DOM树中是否还有需要访问的当前节点的同层节点或者下层节点，如果是，将相应的同层节点或者下层节点作为新的当前节点发送给所述文本段划分模块302。其中，广度优先算法又称作宽度优先搜索，或横向优先搜索，是一种从根节点开始，自上层至下层，沿着树的宽度遍历树的节点的算法。

所述文本段划分模块302，适于判断当前节点对应的文本段是否满足单元文本段的条件，并且在满足条件时，将所述当前节点对应的文本段作为单元文本段，并且通知所述节点访问模块301停止对所述当前节点的后代节点的访问。其中，当所述当前节点到其后代节点的最长路径小于等于预定长度，并且所述当前节点的后代节点中不包含复杂节点时，认为所述当前节点对应的文本段满足单元文本段的条件。所述复杂节点包括：标签table、tr、td、ul、ol、frame、select、input、marquee和/或map对应的节点。

所述文本统计单元400，适于统计所述单元文本段在所述目标网站的所有网页中的出现次数。

图4是所述文本统计单元的内部模块结构图，如图4所示，所述文本统计单元400进一步包括：哈希运算模块401和统计模块402。

所述哈希运算模块401，适于对所述单元文本段进行哈希运算，得到结果键值。

所述统计模块402，适于根据所述结果键值统计所述单元文本段在所述目标网站的所有网页中的出现次数。

所述文本识别单元500，适于判断所述出现次数是否大于预定阈值，如果是，将所述单元文本段识别为非正文文本；否则，将所述单元文本段识别为正文文本。

图5是本发明实施例的网页中非正文文本的识别方法的流程图，如图5所示，对应所述系统，所述方法包括步骤：

A：使用网页抓取器抓取目标网站的所有网页的数据。所述网页抓取器一般可以是网络蜘蛛、网页爬虫、搜索机器人或网络抓取脚本程序。所述网页可以采用HTML（Hypertext Markup Language，超文本标记语言）、XML

（extensible markup language，可扩展标记语言）、XHTML（extensible Hypertext Markup Language，可扩展超文本标记语言）等标记语言，为简单起见，后续均以HTML网页为例说明。这里的所有网页包括所述目标网站的主网页和子网页。

B：构建目标网站的每个网页对应的DOM树。本实施例中采用HTML解析器构建目标网站的每个网页对应的DOM树，本领域技术人员容易想到，当所述网页采用其他标记语言时，可以采用相应的解析器，比如采用XML解析器构建XML网页对应的DOM树。图6是HTML网页对应的DOM树示意图，如图6所示，DOM树中的节点对应于HTML代码中的标签或者文本。

C：根据所述 DOM 树找出所述网页中的单元文本段。

图 7 是所述步骤 C 的流程图，如图 7 所示，所述步骤 C 中具体包括步骤：

C1：从所述 DOM 树的根节点开始，依次访问所述 DOM 树中的节点。

C2：判断当前节点对应的文本段是否满足单元文本段的条件，如果是，停止对所述当前节点的后代节点的访问，将所述当前节点对应的文本段作为单元文本段，执行步骤 C3；否则，直接执行步骤 C3。

所述步骤 C2 具体包括步骤：

C21：判断当前节点是否满足以下条件：所述当前节点到其后代节点的最长路径小于等于预定长度，并且所述当前节点的后代节点中不包含复杂节点；如果是，执行步骤 C22；否则，执行步骤 C3。本步骤的含义是，当所述当前节点同时满足，所述当前节点到其后代节点的最长路径小于等于预定长度，并且所述当前节点的后代节点中不包含复杂节点时，将所述当前节点对应的文本段作为单元文本段。

仍以图 6 为例，图 6 中左侧 div 节点与其最远的后代节点的距离为 3，小于预定距离 4，同时，其后代节点中不含所述复杂节点，因此，图 6 中左侧 div 节点对应的文本段被识别为单元文本段。需要说明的是，对所述单元文本段的识别与所述当前节点的标签名称无关。

所述后代节点包括所述当前节点的各子孙节点，即在所述当前节点所在分支上并且位于所述当前节点下层的所有节点。对于所述预定长度的设定以及所述复杂节点所包括节点的范围可以根据实际使用情况适时调整。比如，所述预定长度可以设置为 3 或者 4 或者 5，本实施例中，所述预定长度优选为 4。仍以图 6 中的 DOM 树为例，对于图 6 中左侧的 div 节点，当其作为

当前节点时，其对应的分支为 div->p->a->text2，该 div 节点距离其最远的后代节点 text2 的距离为 3，小于预定长度 4。本实施例中，所述复杂节点包括：标签 table、tr、td、ul、ol、frame、select、input、marquee 和/或 map 对应的节点。

C22：停止对所述当前节点的后代节点的访问，将所述当前节点对应的文本段作为单元文本段，执行步骤 C3。

C3：判断所述 DOM 树中是否还有需要访问的节点，如果是，将相应的节点作为新的当前节点，执行步骤 C2；否则，执行所述步骤 D。

所述步骤 C1 中，可以采用广度优先算法、深度优先算法或者其他搜索算法访问所述 DOM 树中的节点。由于本发明实施例所述方法在判定当前节点对应的文本段为单元文本段时，将停止对当前节点的后代节点的访问，因此，所述步骤 C1 中优选采用广度优先算法。

当所述步骤 C1 中，按照广度优先算法依次访问所述 DOM 树中的节点时，所述步骤 C3 具体包括步骤：

C31：判断所述 DOM 树中是否还有未访问的所述当前节点的同层节点，如果是，将所述同层节点作为新的当前节点，执行步骤 C2；否则，执行步骤 C32。所述同层节点是指与所述当前节点处于同一层的节点，其可以是所述当前节点的兄弟节点，也可以是所述当前节点的非兄弟节点（如图 6 中的 text2 节点与 text3 节点是同层节点但不是兄弟节点）。

C32：判断所述 DOM 树中是否还有需要访问的所述当前节点的下层节点，如果是，将所述下层节点作为新的当前节点，执行步骤 C2；否则，执行所述步骤 D。需要说明的是，当所述当前节点对应的文本段被识别为单元文本段时，所述当前节点的后代节点无需再被访问。

D：统计所述单元文本段在所述目标网站的所有网页中的出现次数。

图 8 是所述步骤 D 的流程图，如图 8 所示，所述步骤 D 具体包括步骤：

D1：对所述单元文本段进行哈希运算，得到结果键值。本实施例中，采用 MD5（Message Digest Algorithm 5，消息摘要算法第五版）算法对所述单元文本段进行哈希运算，当然本领域技术人员容易想到还可以利用其他类似算法进行哈希运算。本实施例中，在对所述出现次数进行统计时，建立了以所述结果键值为 key，以所述出现次数为 value 的 key-value 表格，进而根据所述 key-value 表格判断每个 key 值对应的 value 值是否大于所述预定阈值，以识别出每个 key 值对应的单元文本段是否是非正文文本。

D2: 根据所述结果键值统计所述单元文本段在所述目标网站的所有网页中的出现次数。

E: 根据所述出现次数，判断所述单元文本段是否是非正文文本。所述步骤 E 具体包括步骤：判断所述出现次数是否大于预定阈值，如果是，将所述单元文本段识别为非正文文本；否则，将所述单元文本段识别为正文文本。

图 9 是单元文本段统计原理示意图，如图 9 所示，各所述单元文本段出现次数的统计范围为所述目标网站的所有网页。其中，所述预定阈值可以根据实际使用情况进行调整，比如预定阈值可以优选为 100，或者 90、150 等等。

本发明实施例所述网页中非正文文本的识别系统及方法，采用 DOM 树获取每个网页的单元文本段，进而对单元文本段在目标网站中出现的次数进行统计，并将其出现次数与预定阈值进行比较，从而识别出网站上的非正文文本，克服了现有方法对于非正文文本识别的滞后问题，并且具有较高的识别准确度，在搜索引擎中具有广泛的应用前景。

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网页中非正文文本的识别系统中的一些或者全部部件的一些或者全部功能。本发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。

例如，图 10 示出了可以实现根据本发明的网页中非正文文本的识别方法的服务器，例如应用服务器。该服务器传统上包括处理器 1010 和以存储器 1020 形式的计算机程序产品或者计算机可读介质。存储器 1020 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 1020 具有用于执行上述方法中的任何方法步骤的程序代码 1031 的存储空间 1030。例如，用于程序代码的存储空间 1030 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 1031。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考图 11 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 10 的服务器中的存储器 1020 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 1031'，即可以由例如诸如 1010 之类的处理器读取的代码，这些代码当由服务器运行时，导致该服务器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。

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。

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是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。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

1、一种网页中非正文文本的识别系统，其包括：网页抓取器、DOM 树构建单元、DOM 树分析单元、文本统计单元和文本识别单元；

所述网页抓取器适于抓取目标网站的所有网页的数据；

所述 DOM 树构建单元，适于构建目标网站的每个网页对应的 DOM 树；

所述 DOM 树分析单元，适于根据所述 DOM 树找出所述网页中的单元文本段；

所述文本统计单元，适于统计所述单元文本段在所述目标网站的所有网页中的出现次数；

所述文本识别单元，适于在所述出现次数大于预定阈值时，将所述单元文本段识别为非正文文本。

2、如权利要求 1 所述的系统，其中，所述 DOM 树构建单元采用网页语言对应的解析器。

3、如权利要求 1 所述的系统，其中，所述 DOM 树分析单元包括：节点访问模块和文本段划分模块；

所述节点访问模块，适于从所述 DOM 树的根节点开始，依次访问所述 DOM 树中的节点，并将访问到的当前节点发送给所述文本段划分模块；以及适于判断所述 DOM 树中是否还有需要访问的节点，如果是，将相应的节点作为新的当前节点发送给所述文本段划分模块；

所述文本段划分模块，适于判断当前节点对应的文本段是否满足单元文本段的条件，并且在满足条件时，将所述当前节点对应的文本段作为单元文本段，并且通知所述节点访问模块停止对所述当前节点的后代节点的访问。

4、如权利要求 3 所述的系统，其中，所述节点访问模块，适于从所述 DOM 树的根节点开始，按照广度优先算法依次访问所述 DOM 树中的节点，并将访问到的当前节点发送给所述文本段划分模块；以及适于判断所述 DOM 树中是否还有需要访问的当前节点的同层节点或者下层节点，如果是，将相应的同层节点或者下层节点作为新的当前节点发送给所述文本段划分模块。

5、如权利要求 3 所述的系统，其中，当所述当前节点到其后代节点的最长路径小于等于预定长度，并且所述当前节点的后代节点中不包含复杂节点时，认为所述当前节点对应的文本段满足单元文本段的条件。

6、如权利要求 5 所述的系统，其中，所述复杂节点包括：标签 table、tr、td、ul、ol、frame、select、input、marquee 和/或 map 对应的节点。

7、如权利要求 1 所述的系统，其中，所述文本统计单元包括：哈希运算模块和统计模块；

所述哈希运算模块，适于对所述单元文本段进行哈希运算，得到结果键值；

所述统计模块，适于根据所述结果键值统计所述单元文本段在所述目标网站的所有网页中的出现次数。

8、如权利要求 1 所述的系统，其中，所述文本识别单元，还适于在所述出现次数小于等于所述预定阈值时，将所述单元文本段识别为正文文本。

9、一种网页中非正文文本的识别方法，其包括步骤：

A：使用网页抓取器抓取目标网站的所有网页的数据；

B：构建目标网站的每个网页对应的DOM树；

C：根据所述DOM树找出所述网页中的单元文本段；

D：统计所述单元文本段在所述目标网站的所有网页中的出现次数；

E：根据所述出现次数，判断所述单元文本段是否是非正文文本。

10、如权利要求 9 所述的方法，其中，所述步骤 B 中，采用网页语言对应的解析器构建目标网站的每个网页对应的 DOM 树。

11、如权利要求 9 所述的方法，其中，所述步骤 C 中具体包括步骤：

C1：从所述 DOM 树的根节点开始，依次访问所述 DOM 树中的节点；

C2：判断当前节点对应的文本段是否满足单元文本段的条件，如果是，停止对所述当前节点的后代节点的访问，将所述当前节点对应的文本段作为单元文本段，执行步骤 C3；否则，直接执行步骤 C3；

C3：判断所述 DOM 树中是否还有需要访问的节点，如果是，将相应的节点作为新的当前节点，执行步骤 C2；否则，执行所述步骤 D。

12、如权利要求 11 所述的方法，其中，所述步骤 C2 具体包括步骤：

C21：判断当前节点是否满足以下条件：所述当前节点到其后代节点的最长路径小于等于预定长度，并且所述当前节点的后代节点中不包含复杂节点；如果是，执行步骤 C22；否则，执行步骤 C3；

C22：停止对所述当前节点的后代节点的访问，将所述当前节点对应的文本段作为单元文本段，执行步骤 C3。

13、如权利要求 12 所述的方法，其中，所述复杂节点包括：标签 table、

tr、td、ul、ol、frame、select、input、marquee 和/或 map 对应的节点。

14、如权利要求 11 所述的方法，其中，所述步骤 C1 中，按照广度优先算法依次访问所述 DOM 树中的节点；

所述步骤 C3 具体包括步骤：

C31：判断所述 DOM 树中是否还有未访问的所述当前节点的同层节点，如果是，将所述同层节点作为新的当前节点，执行步骤 C2；否则，执行步骤 C32；

C32：判断所述 DOM 树中是否还有需要访问的所述当前节点的下层节点，如果是，将所述下层节点作为新的当前节点，执行步骤 C2；否则，执行所述步骤 D。

15、如权利要求 9 所述的方法，其中，所述步骤 D 具体包括步骤：

D1：对所述单元文本段进行哈希运算，得到结果键值；

D2：根据所述结果键值统计所述单元文本段在所述目标网站的所有网页中的出现次数。

16、如权利要求 9 所述的方法，其中，所述步骤 E 具体包括步骤：判断所述出现次数是否大于预定阈值，如果是，将所述单元文本段识别为非正文文本；否则，将所述单元文本段识别为正文文本。

17、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服务器上运行时，导致所述服务器执行根据权利要求 9-16 中的任一个所述的网页中非正文文本的识别方法。

18、一种计算机可读介质，其中存储了如权利要求 17 所述的计算机程序。

```
▼<div class="footer">
  ▼<p>
     <a href="http://www.360.cn/about/index.html" target="_blank">关于我们</a>
     " | "
     <a href="http://www.360.cn/download.html" target="_blank">产品介绍</a>
     " | "
     <a href="http://www.qihoo.com" target="_blank">奇虎问答</a>
     " | "
     <a href="http://www.360.cn/about/zhaopin.html" target="_blank">招聘信息</a>
     " | "
     <a href="http://www.360.cn/about/contactus.html" target="_blank">联系我们</a>
     " | "
     <a href="http://bbs.360.cn/3237995.html" target="_blank">投诉建议</a>
     " | "
     <a href="http://www.360.cn/commend.html" target="_blank">绿色推荐</a>
     " | "
     <a href="http://www.360.cn/custom/bdhezuo.html" target="_blank">合作伙伴</a>
     <br>
     " Copyright©2005-2012 360.CN All Rights Reserved 360安全中心"
     <br>
     <a href="http://www.miibeian.gov.cn/" target="_blank">京ICP证080047号</a>
     " 京公网安备110000000006号"
   </p>
 </div>
```

图 1a

关于我们 | 产品介绍 | 奇虎问答 | 招聘信息 | 联系我们 | 投诉建议 | 绿色推荐 | 合作伙伴

Copyright©2005-2012 360.CN All Rights Reserved 360安全中心

京ICP证080047号 京公网安备110000000006号

图 1b

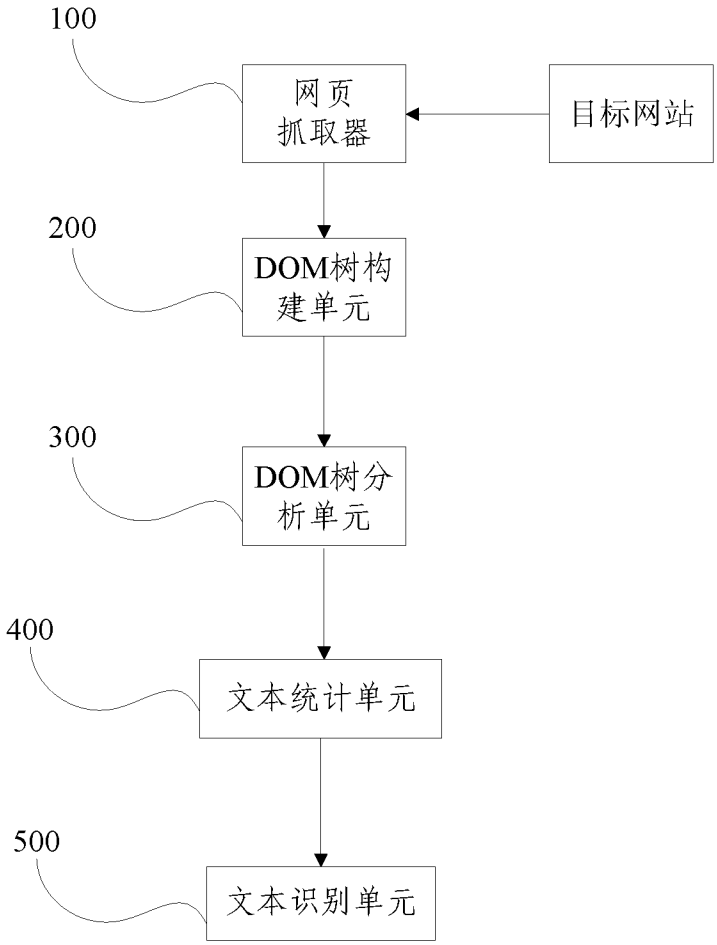
100
网页
抓取器
目标网站
200
DOM树构
建单元
300
DOM树分
析单元
400
文本统计单元
500
文本识别单元

图 2

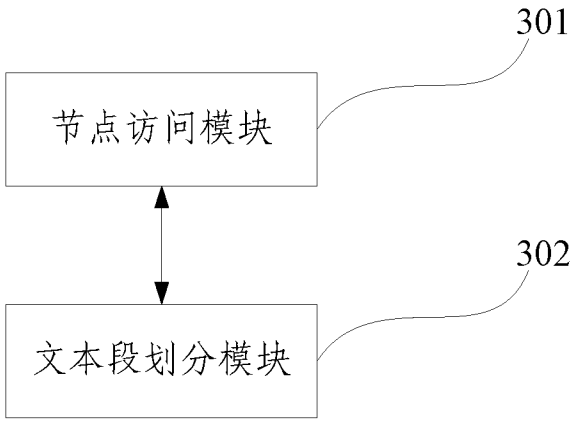
301
节点访问模块
302
文本段划分模块

图 3

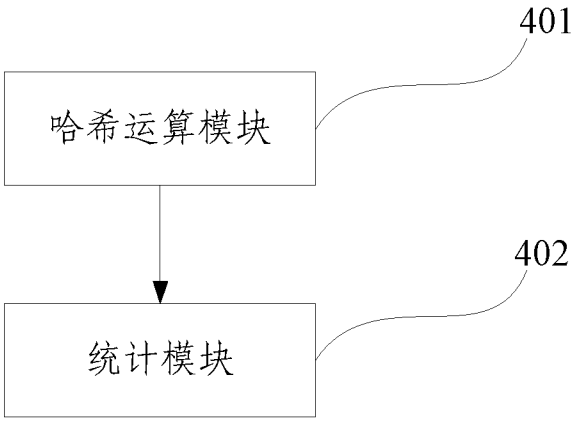
401
哈希运算模块
402
统计模块

图 4

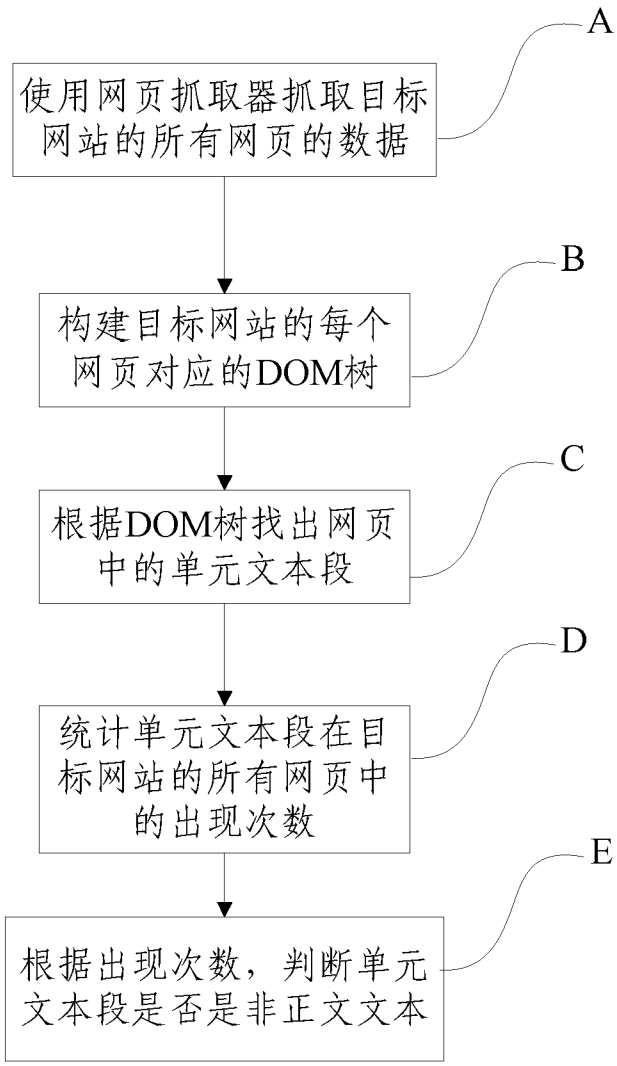
A
使用网页抓取器抓取目标网站的所有网页的数据
B
构建目标网站的每个网页对应的DOM树
C
根据DOM树找出网页中的单元文本段
D
统计单元文本段在目标网站的所有网页中的出现次数
E
根据出现次数，判断单元文本段是否是非正文文本

图 5

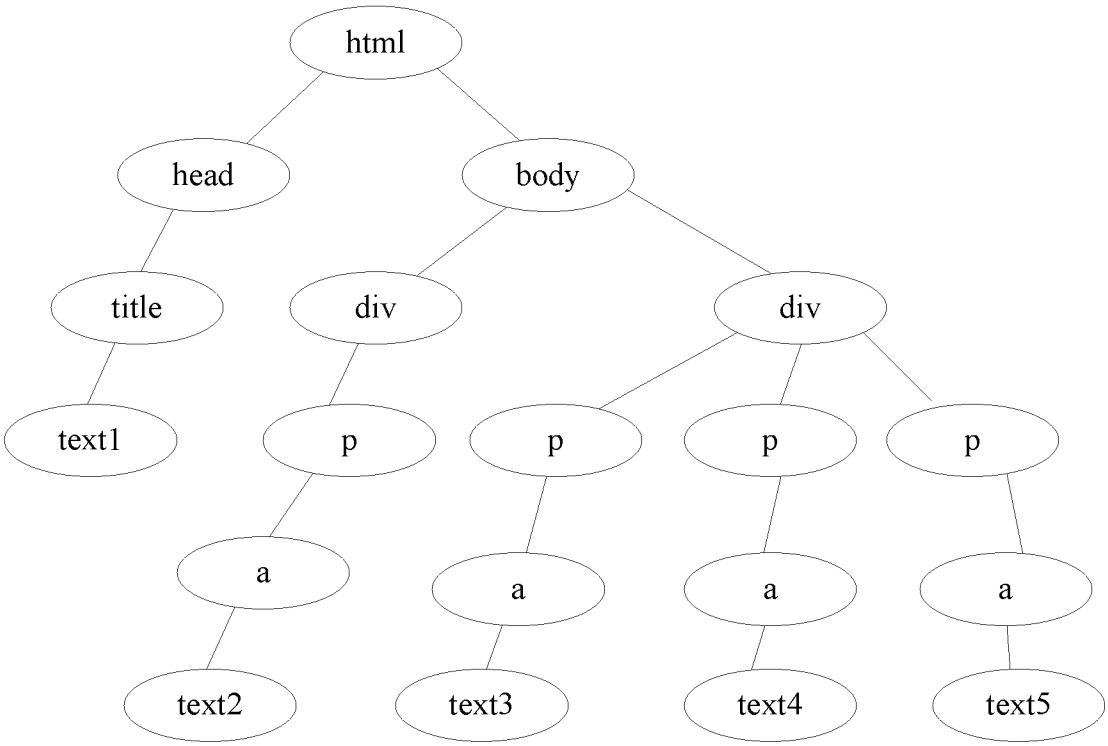
html
head
body
title
div
div
text1
p
p
p
p
a
a
a
a
text2
text3
text4
text5

图 6

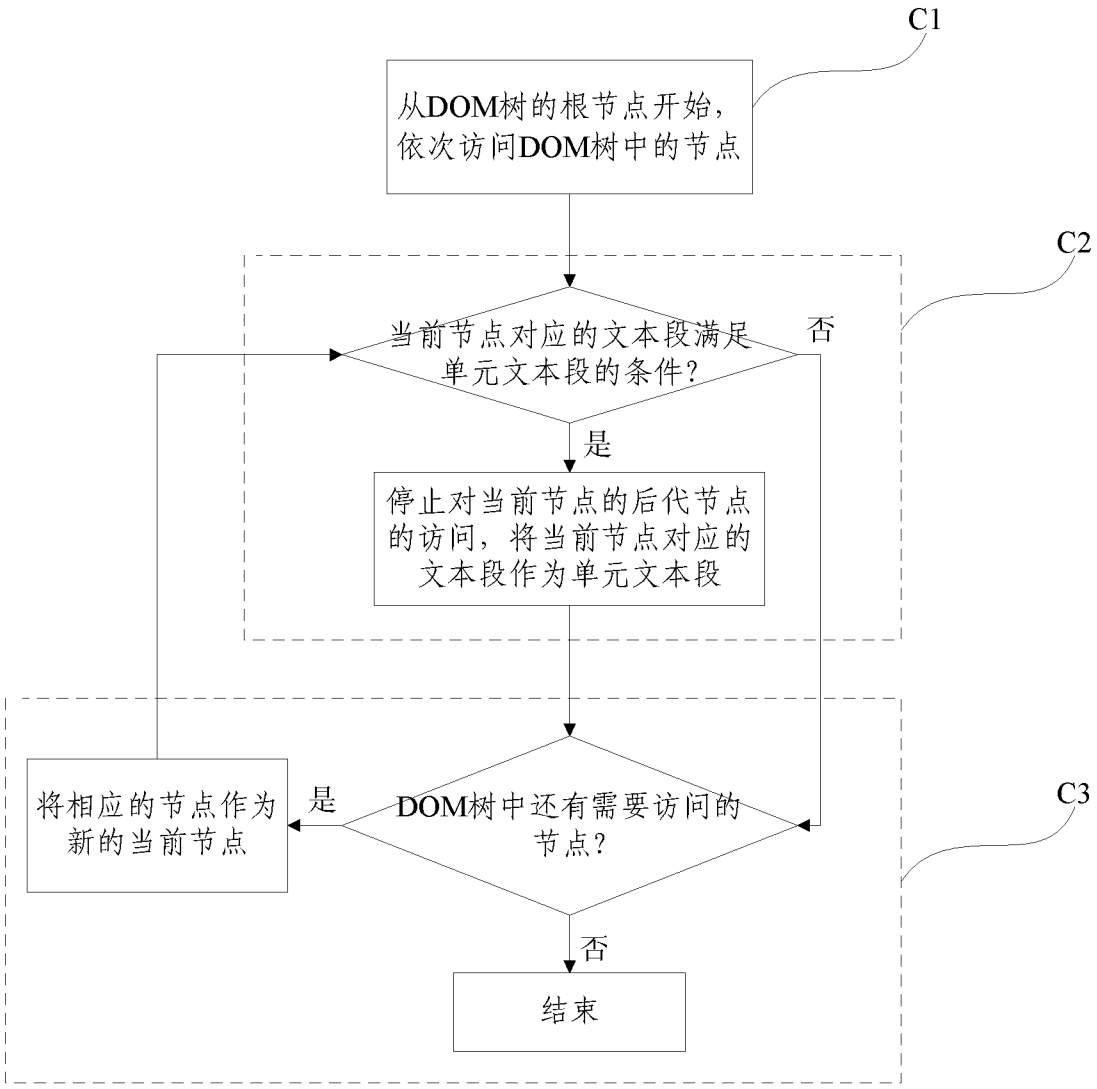
C1
从DOM树的根节点开始，依次访问DOM树中的节点
C2
当前节点对应的文本段满足单元文本段的条件？
否
是
停止对当前节点的后代节点的访问，将当前节点对应的文本段作为单元文本段
C3
将相应的节点作为新的当前节点
是
DOM树中还有需要访问的节点？
否
结束

图 7

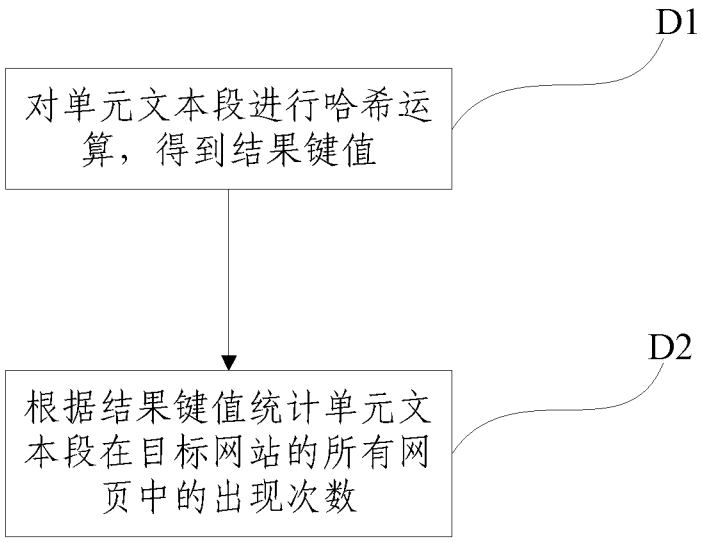
D1
对单元文本段进行哈希运算，得到结果键值
D2
根据结果键值统计单元文本段在目标网站的所有网页中的出现次数

图 8

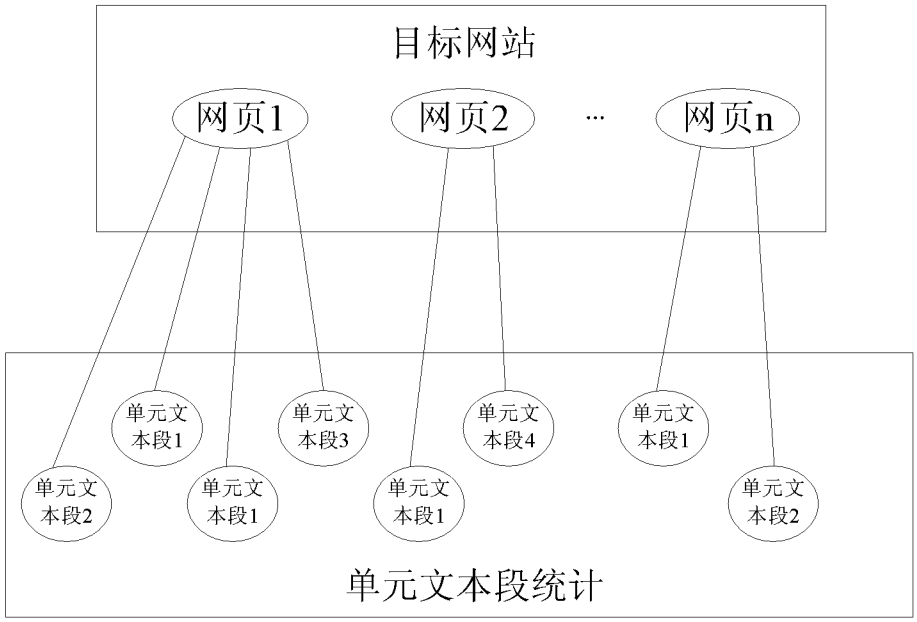
目标网站
网页1
网页2
…
网页n
单元文本段1
单元文本段3
单元文本段4
单元文本段1
单元文本段2
单元文本段1
单元文本段1
单元文本段2
单元文本段统计

图 9

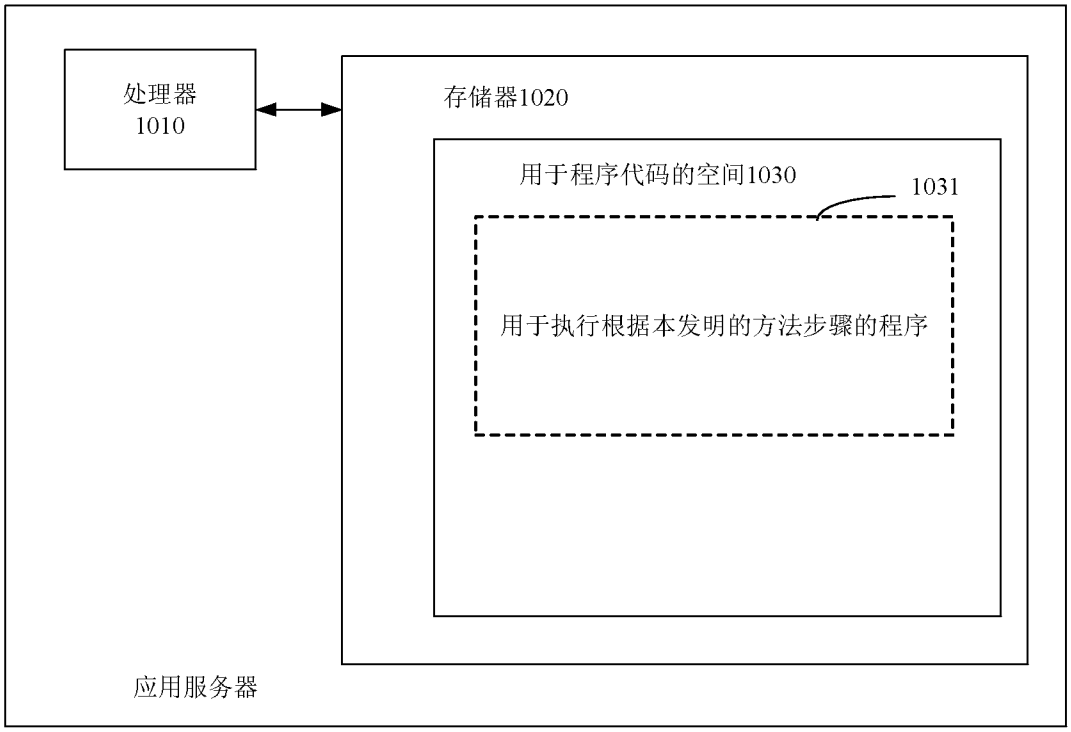
处理器
1010
存储器1020
用于程序代码的空间1030
1031
用于执行根据本发明的方法步骤的程序
应用服务器

图 10

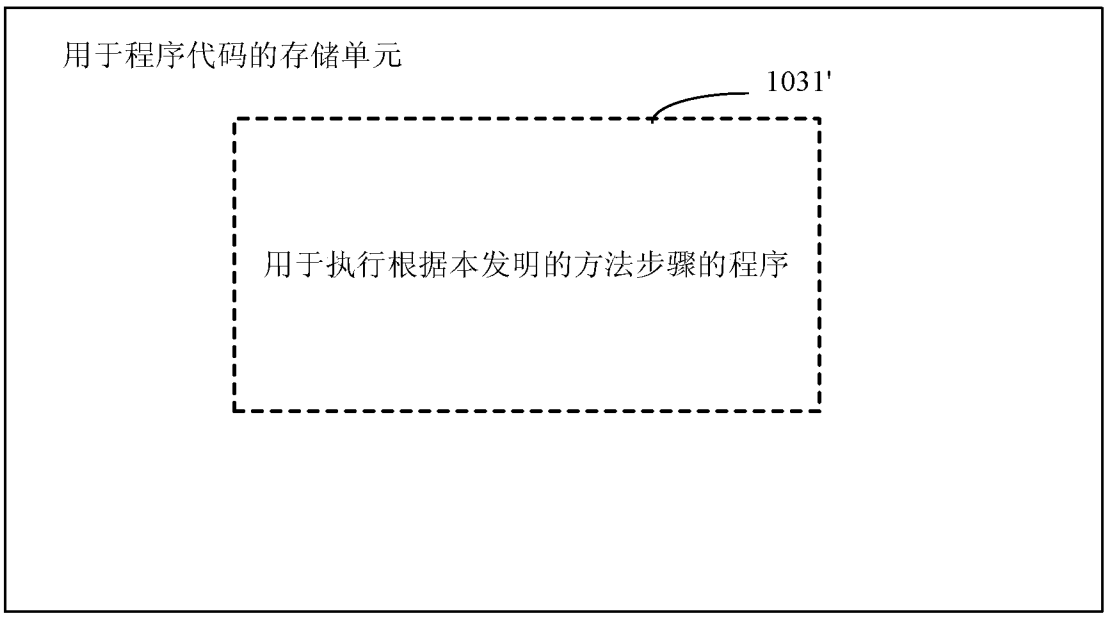
用于程序代码的存储单元
1031'
用于执行根据本发明的方法步骤的程序

图 11

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2013/077102

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/30 (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)

IPC: 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, CNKI, WPI, EPODOC, IEEE, GOOGLE: page, garbage, navigation, advertisement, copyright, recognition, web w page, text, extract, times, frequency, DOM, tree

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 102779172 A (BEIJING QIHOO TECHNOLOGY CO., LTD. et al.), 14 November 2012 (14.11.2012), claims 1-16, and description, paragraphs [0070]-[0091]	1-18
A	CN 102184189 A (BEIJING INSTITUTE OF TECHNOLOGY), 14 September 2011 (14.09.2011), abstract, and claims 1-9	1-18
A	CN 101251855 A (TENCENT TECHNOLOGY (SHENZHEN) CO., LTD.), 27 August 2008 (27.08.2008), the whole document	1-18
A	US 2009063500 A1 (MCROSOFT CORPORATION), 05 March 2009 (05.03.2009), the whole document	1-18

☐ 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	Date of mailing of the international search report
23 August 2013 (23.08.2013)	**19 September 2013 (19.09.2013)**
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 **OUYANG, Qi** Telephone No.: (86-10) **62413658**

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2013/077102

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 102779172 A	14.11.2012	None	
CN 102184189 A	14.09.2011	None	
CN 101251855 A	27.08.2008	None	
US 2009063500 A1	05.03.2009	None	

国际检索报告

国际申请号
PCT/CN2013/077102

A. 主题的分类

G06F 17/30 (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)

IPC: 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, CNKI, WPI, EPODOC, IEEE, GOOGLE: 网页, 页面, 正文, 文本, 垃圾, 导航, 广告, 版权, 识别, 抽取, 提取, 树, 次数, 频率, web w page, text, extract, times, frequency, DOM, tree

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 102779172 A (北京奇虎科技有限公司 等) 14. 11 月 2012 (14.11.2012) 权利要求 1-16、说明书[0070]-[0091]段	1-18
A	CN 102184189 A (北京理工大学) 14. 9 月 2011 (14.09.2011) 摘要、权利要求 1-9	1-18
A	CN 101251855 A (腾讯科技(深圳)有限公司) 27. 8 月 2008 (27.08.2008) 全文	1-18
A	US 2009063500 A1 (MICROSOFT CORPORATION) 05. 3 月 2009 (05.03.2009) 全文	1-18

☐ 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	国际检索报告邮寄日期
23. 8 月 2013 (23.08.2013)	**19.9 月 2013 (19.09.2013)**

ISA/CN 的名称和邮寄地址：	受权官员
中华人民共和国国家知识产权局 中国北京市海淀区蓟门桥西土城路 6 号 100088 传真号：　(86-10)62019451	欧阳琦 电话号码：(86-10) **62413658**

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2013/077102

检索报告中引用的专利文件	公布日期	同族专利	公布日期
CN 102779172 A	14.11.2012	无	
CN 102184189 A	14.09.2011	无	
CN 101251855 A	27.08.2008	无	
US 2009063500 A1	05.03.2009	无	