

(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
2016 年 10 月 6 日 (06.10.2016)



(10) 国际公布号
WO 2016/155385 A1

- (51) 国际专利分类号:
G06F 17/30 (2006.01)
- (21) 国际申请号: PCT/CN2015/099579
- (22) 国际申请日: 2015 年 12 月 29 日 (29.12.2015)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201510150540.2 2015 年 3 月 31 日 (31.03.2015) 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 6 号院 2 号楼 B 座 2 层、3 层 301-306 室, Beijing 100015 (CN)。
- (72) 发明人: 陈进平 (CHEN, Jinping); 中国北京市朝阳区酒仙桥路 6 号院 2 号楼, Beijing 100015 (CN)。董毅 (DONG, Yi); 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: 北京市隆安律师事务所 (BEIJING LONGAN LAW FIRM); 中国北京市朝阳区建国门外大街 21 号北京国际俱乐部 188 室, Beijing 100020 (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,

[见续页]

(54) Title: METHOD AND APPARATUS FOR GENERATING FILE INDEX AND SEARCHING METHOD AND APPARATUS

(54) 发明名称: 一种生成文档索引的方法和装置以及一种搜索方法和装置

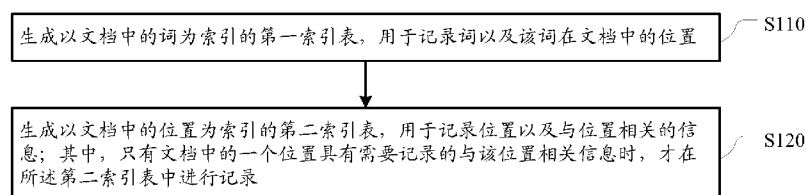


图 1

- S110 Generate a first index table having the words in the file as indices, used for recording words and the positions of said words in the file
- S120 Generate a second index table having the positions of the words in the file as indices, used for recording the positions and information related to the positions, wherein only when one position in the file has information related to the position that needs to be recorded is said information recorded in the second index file

(57) Abstract: A method and apparatus for generating a file index and a searching method and apparatus, the method for generating a file index comprising: generating a first index table having the words in the file as indices, used for recording words and the positions of said words in the file (S110); and generating a second index table having the positions of the words in the file as indices, used for recording the positions and information related to the positions, wherein only when one position in the file has information related to the position that needs to be recorded is said information recorded in the second index file (S120). By means of storing the position and the information related to the position separately, expanding the information in the index is more convenient and flexible, and only information related to position that is actually present will occupy storage space, preventing wastage of storage resources.

(57) 摘要: 一种生成文档索引的方法和装置以及一种搜索方法和装置, 其中, 生成文档索引的方法包括: 生成以文档中的词为索引的第一索引表, 用于记录词以及该词在文档中的位置 (S110), 生成以文档中的位置为索引的第二索引表, 用于记录位置以及与位置相关的信息, 其中, 只有文档中的一个位置具有需要记录的与该位置相关信息时, 才在所述第二索引表中进行记录 (S120)。通过将位置以及与位置相关的信息分离存储, 使得在索引中扩展与位置相关的信息更加方便灵活, 且只有实际存在的与位置相关的信息才占用存储空间, 避免了存储资源的浪费。

WO 2016/155385 A1



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))。

一种生成文档索引的方法和装置以及一种搜索方法和装置

技术领域

本发明涉及互联网搜索技术领域，具体涉及一种生成文档索引的方法和装置以及一种搜索方法和装置。

5

背景技术

在搜索引擎系统中，倒排索引被用来存储在全文搜索下某个单词在一个文档或者一组文档中的存储位置的映射。它是文档检索系统中最常用的数据结构。通过倒排索引，可以根据单词快速获取包含这个单词的文档列表。

10 即倒排索引会记录每个词在文档里出现的位置（也称为偏移）。为了更好地提升搜索效果，有时需要记录与这些位置相关的信息，包括但不限于：

1.这个位置包括在书名号里；例如，《资治通鉴》中的‘通鉴’的出现就是在书名号里；

2.这个位置包括在引号里；

15 3.这个位置是在 html 标签<H3>里。

倒排索引用 16 个 bit 来记录每个词的每次出现位置，为了记录与位置相关的信息，一般的实现方式是将 16bit 扩充为 32bit，低 16 位存储位置，高 16bit 存储与位置相关的信息。当时这种方法带来如下问题：一方面，索引增长太多，由于倒排索引的主要内容为每个词的出现位置，如果将这个值从
20 16bit 扩充到 32bit，整个倒排索引会增长一倍。另一方面，大多数的位置并具备有价值的与该位置相关的信息，因此实际上大多数存储资源都被浪费了。

发明内容

25 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上述问题的一种生成文档索引的方法和装置以及一种搜索方法和装置。

依据本发明的一个方面，提供了一种生成文档索引的方法，其中，该方法包括：

生成以文档中的词为索引的第一索引表,用于记录词以及该词在文档中的位置;

生成以文档中的位置为索引的第二索引表,用于记录位置以及与位置相关的信息;

- 5 其中,只有文档中的一个位置具有需要记录的与该位置相关信息时,才在所述第二索引表中进行记录。

依据本发明的另一个方面,提供了一种搜索方法,其中,该方法包括:
接收搜索请求方发来的搜索词;

- 10 根据第一索引表找到搜索词在文档中的位置;其中,第一索引表用于记录词以及该词在文档中的位置;

根据第二索引表找到与搜索词在文档中的位置相关的信息;其中,第二索引表用于记录位置以及与位置相关的信息;

基于所述搜索词对应的文档中的位置和与位置相关的信息,将所述文档作为搜索结果返回给搜索请求方。

- 15 依据本发明的一个方面,提供了一种生成文档索引的装置,其中,该装置包括:

第一索引表生成单元,适于生成以文档中的词为索引的第一索引表,用于记录词以及该词在文档中的位置;

- 20 第二索引表生成单元,适于生成以文档中的位置为索引的第二索引表,用于记录位置以及与位置相关的信息;其中,只有文档中的一个位置具有需要记录的相关信息时,才在所述第二索引表中进行记录。

依据本发明的另一个方面,提供了一种搜索装置,其中,该装置包括:
接收单元,适于接收搜索请求方发来的搜索词;

- 25 搜索单元,适于根据第一索引表找到搜索词在文档中的位置,以及根据第二索引表找到与搜索词在文档中的位置相关的信息;其中,第一索引表用于记录词以及该词在文档中的位置,第二索引表用于记录位置以及与位置相关的信息;

发送单元,适于基于所述搜索词对应的文档中的位置和与位置相关的信息,将所述文档作为搜索结果返回给搜索请求方。

- 30 根据本发明的又一个方面,提出了一种计算机程序,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上文所述的生成文档索引的方法，或者，导致所述计算设备执行上文所述的搜索方法。

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

根据本发明的这种生成以文档中的词为索引的第一索引表，用于记录词以及该词在文档中的位置，生成以文档中的位置为索引的第二索引表，用于记录位置以及与位置相关的信息，其中，只有文档中的一个位置具有需要记录的与该位置相关信息时，才在所述第二索引表中进行记录的技术方案，将位置以及与位置相关的信息分离存储，使得在索引中扩展与位置相关的信息更加方便灵活，且只有实际存在的与位置相关的信息才占用存储空间，避免了存储资源的浪费。

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 1 示出了根据本发明一个实施例的一种生成文档索引的方法的流程图；

图 2 示出了根据本发明一个实施例的一种搜索方法的流程图；

图 3 示出了根据本发明一个实施例的一种生成文档索引的装置的结构图；

图 4 示出了根据本发明一个实施例的一种搜索装置的结构图；

图 5 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 6 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

下面将参照附图更详细地描述本公开的示例性实施例。虽然附图中显示了本公开的示例性实施例，然而应当理解，可以以各种形式实现本公开而不应被这里阐述的实施例所限制。相反，提供这些实施例是为了能够更透彻地理解本公开，并且能够将本公开的范围完整的传达给本领域的技术人员。

图 1 示出了根据本发明一个实施例的一种生成文档索引的方法的流程图。如图 1 所示，该方法包括：

10 步骤 S110，生成以文档中的词为索引的第一索引表，用于记录词以及该词在文档中的位置；

步骤 S120，生成以文档中的位置为索引的第二索引表，用于记录位置以及与位置相关的信息；其中，只有文档中的一个位置具有需要记录的与该位置相关信息时，才在所述第二索引表中进行记录。

15 图 1 所示的方法，将位置以及与位置相关的信息分离存储，使得在索引中扩展与位置相关的信息更加方便灵活，且只有实际存在的与位置相关的信息才占用存储空间，避免了存储资源的浪费。

在本发明的一个实施例中，图 1 所示方法的步骤 S120 所述生成以文档中的位置为索引的第二索引表，用于记录位置以及与位置相关的信息包括：
20 在第二索引表中，以三元组的形式记录与位置相关的信息；

所述三元组由以下三元构成：与位置相关的信息的开始位置，用 start 表示；与位置相关的信息的长度，用 len 表示；与位置相关的信息的类型，用 type。

三元组具体可表示为<start, len, type>。

25 Start 表示与位置相关的信息的开始位置，如书名号的起始位置；

len 表示与位置相关的信息的长度，如书名号的长度；

type 表示与位置相关的信息的类型，如该相关的信息为书名号。

在本发明的一个实施例中每个位置由 16 比特位（即 16 bit）来记录；通过位置到三元组列表（即所述的第二索引表）里检索与位置相关的信息。所述
30 述与位置相关的信息的开始位置 start 由 16 比特位来记录；所述与位置相关

的信息的长度 len 由 8 比特位来记录；所述与位置相关的信息的类型 type 由 8 比特位来记录。

同时，只有当文档包括特征的与位置相关的信息时才在第二索引表中存储，不存在浪费。因此相对于传统的实现方式，能够节约至少 40% 的索引空间。

并且与位置相关的信息在存储上与位置分离，实际上使得位置相关信息的扩展更加灵活方便。

图 2 示出了根据本发明一个实施例的一种搜索方法的流程图。如图 2 所示该方法包括：

步骤 S210，接收搜索请求方发来的搜索词；

步骤 S220，根据第一索引表找到搜索词在文档中的位置；其中，第一索引表用于记录词以及该词在文档中的位置；

步骤 S230，根据第二索引表找到与搜索词在文档中的位置相关的信息；其中，第二索引表用于记录位置以及与位置相关的信息；

步骤 S240，基于所述搜索词对应的文档中的位置和与位置相关的信息，将所述文档作为搜索结果返回给搜索请求方。

图 2 所示的方法能够给搜索请求方返回基于位置和与位置相关的信息的搜索结果，并且与位置相关的信息与位置分离存储，位置相关信息的扩展更加灵活，且大大节省了索引空间。

图 3 示出了根据本发明一个实施例的一种生成文档索引的装置的结构图。如图 3 所示，该生成文档索引的装置 300 包括：

第一索引表生成单元 310，适于生成以文档中的词为索引的第一索引表，用于记录词以及该词在文档中的位置；

第二索引表生成单元 320，适于生成以文档中的位置为索引的第二索引表，用于记录位置以及与位置相关的信息；其中，只有文档中的一个位置具有需要记录的相关信息时，才在所述第二索引表中进行记录。

图 3 所示的装置，将位置以及与位置相关的信息分离存储，使得在索引中扩展与位置相关的信息更加方便灵活，且只有实际存在的与位置相关的信

息才占用存储空间，避免了存储资源的浪费。

在本发明的一个实施例中，第二索引表生成单元 320，适于在第二索引表中，以三元组的形式记录与位置相关的信息；

其中，所述三元组由以下三元构成：与位置相关的信息的开始位置，与
5 位置相关的信息的长度；与位置相关的信息的类型。

三元组具体可表示为<start, len, type>。

Start 表示与位置相关的信息的开始位置，如书名号的起始位置；

len 表示与位置相关的信息的长度，如书名号的长度；

type 表示与位置相关的信息的类型，如该相关的信息为书名号。

10 在本发明的一个实施例中，所述第一索引表生成单元 310，适于将每个位置用 16 比特位来记录；所述第二索引表生成单元 320，适于将与位置相关的信息的开始位置用 16 比特位来记录，将与位置相关的信息的长度用 8 比特位来记录，适于与将位置相关的信息的类型用 8 比特位来记录。

同时，只有当文档包括特征的与位置相关的信息时才在第二索引表中存
15 储，不存在浪费。因此相对于传统的实现方式，能够节约至少 40%的索引空间。并且与位置相关的信息在存储上与位置分离，实际上使得位置相关信息的扩展更加灵活方便。

图 4 示出了根据本发明一个实施例的一种搜索装置的结构图。如图 4 所
20 示，该搜索装置 400 包括：

接收单元 410，适于接收搜索请求方发来的搜索词；

搜索单元 420，，适于根据第一索引表找到搜索词在文档中的位置，以及根据第二索引表找到与搜索词在文档中的位置相关的信息；其中，第一索引表用于记录词以及该词在文档中的位置，第二索引表用于记录位置以及与
25 位置相关的信息；

发送单元 430，适于基于所述搜索词对应的文档中的位置和与位置相关的信息，将所述文档作为搜索结果返回给搜索请求方。

图 4 所示的装置能够给搜索请求方返回给基于置和与位置相关的信息的搜索结果，并且与位置相关的信息与位置分离存储，位置相关信息的扩展
30 更加灵活，且大大节省了索引空间。

综上所述，根据本发明的这种生成以文档中的词为索引的第一索引表，用于记录词以及该词在文档中的位置，生成以文档中的位置为索引的第二索引表，用于记录位置以及与位置相关的信息，其中，只有文档中的一个位置具有需要记录的与该位置相关信息时，才在所述第二索引表中进行记录的技术方案，将位置以及与位置相关的信息分离存储，使得在索引中扩展与位置相关的信息更加方便灵活，且只有实际存在的与位置相关的信息才占用存储空间，避免了存储资源的浪费。

10 需要说明的是：

在此提供的算法和显示不与任何特定计算机、虚拟装置或者其它设备固有相关。各种通用装置也可以与基于在此的示教一起使用。根据上面的描述，构造这类装置所要求的结构是显而易见的。此外，本发明也不针对任何特定编程语言。应当明白，可以利用各种编程语言实现在此描述的本发明的内容，并且上面对特定语言所做的描述是为了披露本发明的最佳实施方式。

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。

类似地，应当理解，为了精简本公开并帮助理解各个发明方面中的一个或多个，在上面对本发明的示例性实施例的描述中，本发明的各个特征有时被一起分组到单个实施例、图、或者对其的描述中。然而，并不应将该公开的方法解释成反映如下意图：即所要求保护的本发明要求比在每个权利要求中所明确记载的特征更多的特征。更确切地说，如下面的权利要求书所反映的那样，发明方面在于少于前面公开的单个实施例的所有特征。因此，遵循具体实施方式的权利要求书由此明确地并入该具体实施方式，其中每个权利要求本身都作为本发明的单独实施例。

本领域那些技术人员可以理解，可以对实施例中的设备中的模块进行自适应性地改变并且把它们设置在与该实施例不同的一个或多个设备中。可以把实施例中的模块或单元或组件组合成一个模块或单元或组件，以及此外可以30 以把它们分成多个子模块或子单元或子组件。除了这样的特征和/或过程或者

单元中的至少一些是相互排斥之外，可以采用任何组合对本说明书（包括伴随的权利要求、摘要和附图）中公开的所有特征以及如此公开的任何方法或者设备的所有过程或单元进行组合。除非另外明确陈述，本说明书（包括伴随的权利要求、摘要和附图）中公开的每个特征可以由提供相同、等同或相似目的的替代特征来代替。

此外，本领域的技术人员能够理解，尽管在此所述的一些实施例包括其它实施例中所包括的某些特征而不是其它特征，但是不同实施例的特征的组合意味着处于本发明的范围之内并且形成不同的实施例。例如，在下面的权利要求书中，所要求保护的实施例的任意之一都可以以任意的组合方式来使用。

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生成文档索引的装置和搜索装置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 5 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。该计算设备传统上包括处理器 510 和以存储器 520 形式的计算机程序产品或者计算机可读介质。存储器 520 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 520 具有用于执行上述方法中的任何方法步骤的程序代码 531 的存储空间 530。例如，用于程序代码的存储空间 530 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 531。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 6 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 5 的计算设备中的存储器 520 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，存储单元包括用于执行根据本发明的方法步骤的计算机可读代码 531'，即可以由例如诸如 510 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。

5 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。
10 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。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。

本发明可以应用于计算机系统/服务器，其可与众多其它通用或专用计算机系统环境或配置一起操作。适于与计算机系统/服务器一起使用的众所周知的计算系统、环境和/或配置的例子包括但不限于：个人计算机系统、服务器计
20 算系统、瘦客户机、厚客户机、手持或膝上设备、基于微处理器的系统、机顶盒、可编程消费电子产品、网络个人电脑、小型计算机系统、大型计算机系统和包括上述任何系统的分布式云计算技术环境，等等。

25 计算机系统/服务器可以在由计算机系统执行的计算机系统可执行指令（诸如程序模块）的一般语境下描述。通常，程序模块可以包括例程、程序、目标程序、组件、逻辑、数据结构等等，它们执行特定的任务或者实现特定的抽象数据类型。计算机系统/服务器可以在分布式云计算环境中实施，分布式云计算环境中，任务是由通过通信网络链接的远程处理设备执行的。在分
30 布式云计算环境中，程序模块可以位于包括存储设备的本地或远程计算系统

存储介质上。

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
5 指同一个实施例。

权 利 要 求

1、一种生成文档索引的方法，其中，该方法包括：

5 生成以文档中的词为索引的第一索引表，用于记录词以及该词在文档中的位置；

生成以文档中的位置为索引的第二索引表，用于记录位置以及与位置相关的信息；

其中，只有文档中的一个位置具有需要记录的与该位置相关信息时，才在所述第二索引表中进行记录。

10 2、如权利要求 1 所述的方法，其中，所述生成以文档中的位置为索引的第二索引表，用于记录位置以及与位置相关的信息包括：

在第二索引表中，以三元组的形式记录与位置相关的信息；

所述三元组由以下三元构成：与位置相关的信息的开始位置；与位置相关的信息的长度；与位置相关的信息的类型。

15 3、如权利要求 2 所述的方法，其中，

所述每个位置由 16 比特位来记录；

所述与位置相关的信息的开始位置由 16 比特位来记录；

所述与位置相关的信息的长度由 8 比特位来记录；

所述与位置相关的信息的类型由 8 比特位来记录。

20 4、一种搜索方法，其中，该方法包括：

接收搜索请求方发来的搜索词；

根据第一索引表找到搜索词在文档中的位置；其中，第一索引表用于记录词以及该词在文档中的位置；

25 根据第二索引表找到与搜索词在文档中的位置相关的信息；其中，第二索引表用于记录位置以及与位置相关的信息；

基于所述搜索词对应的文档中的位置和与位置相关的信息，将所述文档作为搜索结果返回给搜索请求方。

5、一种生成文档索引的装置，其中，该装置包括：

30 第一索引表生成单元，适于生成以文档中的词为索引的第一索引表，用于记录词以及该词在文档中的位置；

第二索引表生成单元，适于生成以文档中的位置为索引的第二索引表，用于记录位置以及与位置相关的信息；其中，只有文档中的一个位置具有需

要记录的相关信息时，才在所述第二索引表中进行记录。

6、如权利要求 5 所述的装置，其中

第二索引表生成单元，适于在第二索引表中，以三元组的形式记录与位置相关的信息；

5 其中，所述三元组由以下三元构成：与位置相关的信息的开始位置，与位置相关的信息的长度；与位置相关的信息的类型。

7、如权利要求 6 所述的装置，其中，

所述第一索引表生成单元，适于将每个位置用 16 比特位来记录；

10 所述第二索引表生成单元，适于将与位置相关的信息的开始位置用 16 比特位来记录，将与位置相关的信息的长度用 8 比特位来记录，适于与将位置相关的信息的类型用 8 比特位来记录。

8、一种搜索装置，其中，该装置包括：

接收单元，适于接收搜索请求方发来的搜索词；

15 搜索单元，适于根据第一索引表找到搜索词在文档中的位置，以及根据第二索引表找到与搜索词在文档中的位置相关的信息；其中，第一索引表用于记录词以及该词在文档中的位置，第二索引表用于记录位置以及与位置相关的信息；

发送单元，适于基于所述搜索词对应的文档中的位置和与位置相关的信息，将所述文档作为搜索结果返回给搜索请求方。

20 9、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 1-3 中的任一项所述的生成文档索引的方法，或者，导致所述计算设备执行根据权利要求 4 所述的搜索方法。

25 10、一种计算机可读介质，其中存储了如权利要求 9 所述的计算机程序。

1/3

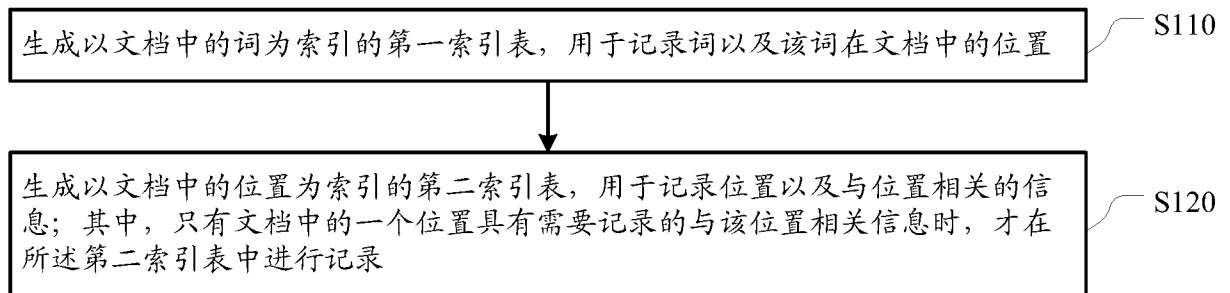


图 1

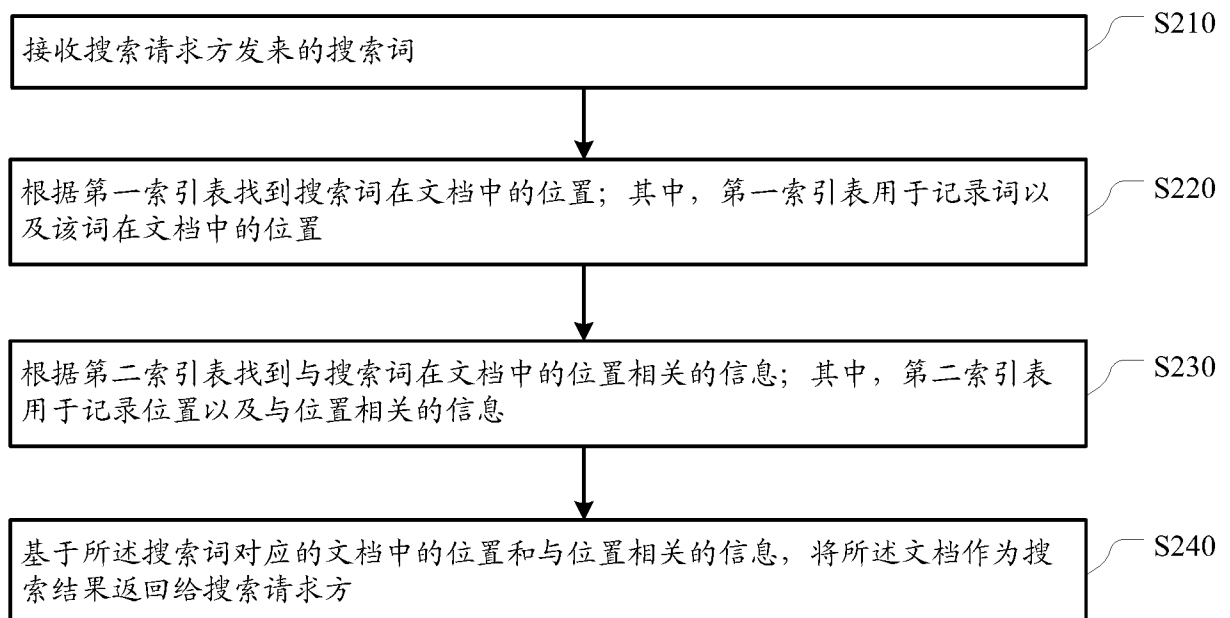


图 2

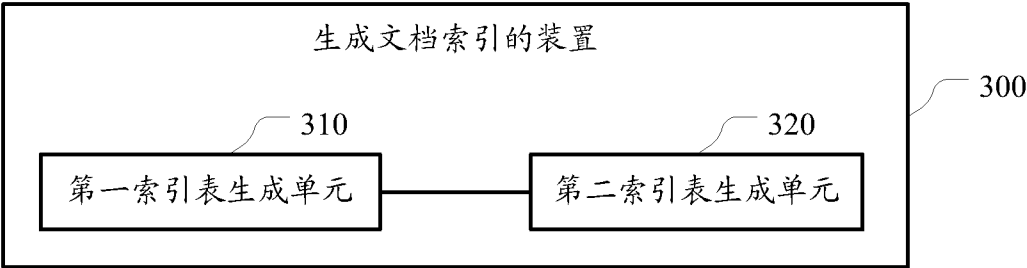


图 3

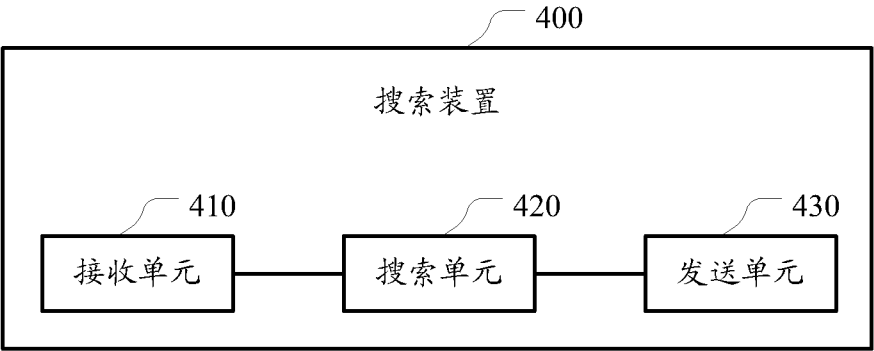


图 4

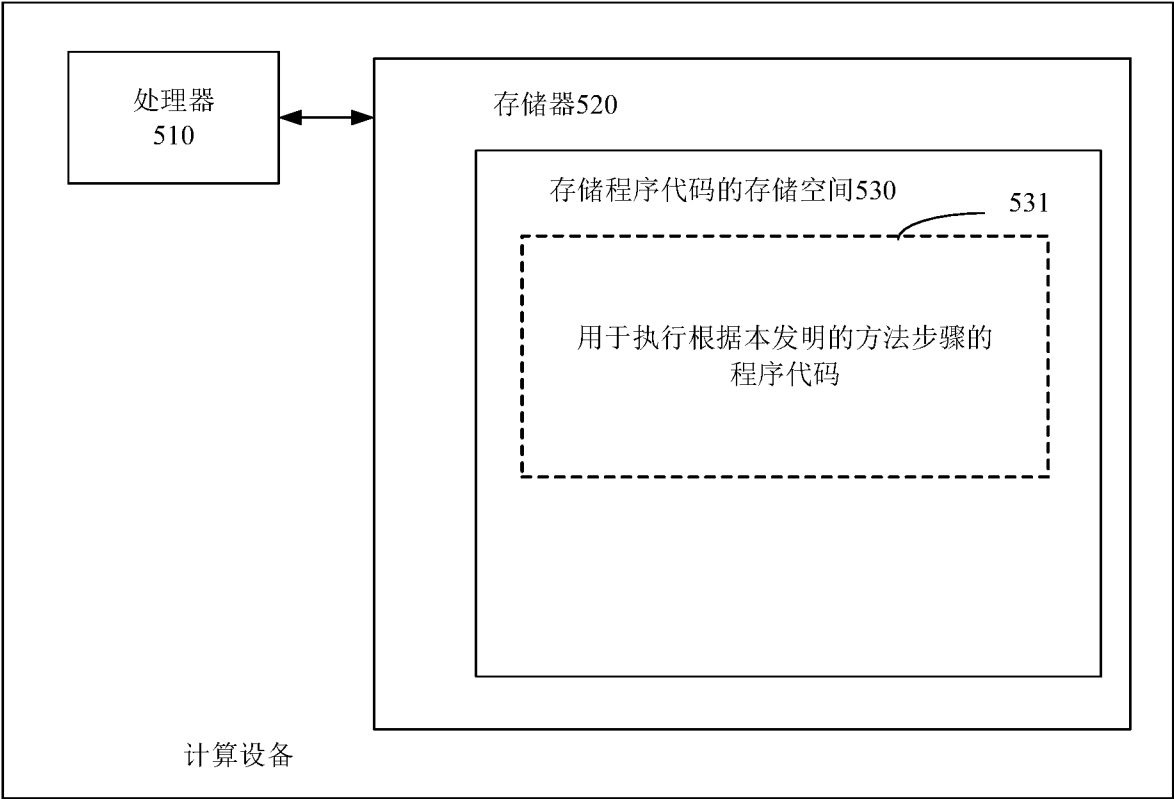


图 5

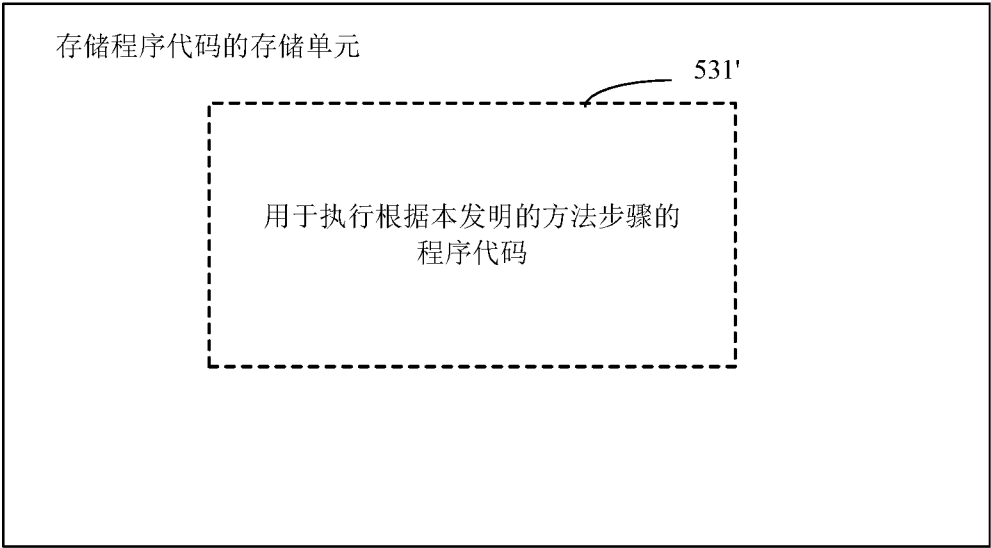


图 6

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2015/099579

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)

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)

CNKI, CNPAT, WPI, EPODOC: supplement, explain, correlation, document?, index, record+, word?, key, position?, location?, relat+, describ+, reduc+, stor+, space?

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 104715068 A (BEIJING QIHOO TECHNOLOGY CO., LTD. et al.), 17 June 2015 (17.06.2015), claims 1-8, and description, paragraphs [0045]-[0079] and [0088]	1-10
X	CN 102662926 A (CHANGZHOU SINOTYPE TECHNOLOGY CO., LTD.), 12 September 2012 (12.09.2012), description, paragraphs [0039]-[0043] and [0051], and figures 2-3 (d)	1-10
A	CN 101131704 A (INTERNATIONAL BUSINESS MACHINES CORPORATION), 27 February 2008 (27.02.2008), the whole document	1-10
A	CN 102110123 A (NATIONAL UNIVERSITY OF DEFENSE TECHNOLOGY), 29 June 2011 (29.06.2011), the whole document	1-10
A	US 2013024459 A1 (MICROSOFT CORPORATION), 24 January 2013 (24.01.2013), the whole document	1-10

☐ 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 02 March 2016 (02.03.2016)	Date of mailing of the international search report 23 March 2016 (23.03.2016)
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 WANG, Wei Telephone No.: (86-10) 62413695

International application No.
PCT/CN2015/099579

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

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) 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)) CNKI, CNPAT, WPI, EPDOC: 文档, 文件, 索引, 词, 位置, 相关, 补充, 描述, 说明, 关联, 减少, 降低, 存储, 空间, document?, index, record+, word?, key, position?, location?, relat+, describ+, reduc+, stor+, space?		
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 104715068 A (北京奇虎科技有限公司 等) 2015年 6月 17日 (2015 - 06 - 17) 权利要求1-8, 说明书第[0045]-[0079], [0088]段	1-10
X	CN 102662926 A (常州华文文字技术有限公司) 2012年 9月 12日 (2012 - 09 - 12) 说明书第[0039]-[0043], [0051]段, 附图2-3(d)	1-10
A	CN 101131704 A (国际商业机器公司) 2008年 2月 27日 (2008 - 02 - 27) 全文	1-10
A	CN 102110123 A (中国人民解放军国防科学技术大学) 2011年 6月 29日 (2011 - 06 - 29) 全文	1-10
A	US 2013024459 A1 (MICROSOFT CORPORATION) 2013年 1月 24日 (2013 - 01 - 24) 全文	1-10
☐ 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
2016年 3月 2日		2016年 3月 23日
ISA/CN的名称和邮寄地址		授权官员
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		王伟 电话号码 (86-10)62413695

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2015/099579

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	104715068	A	2015年 6月 17日	无	
CN	102662926	A	2012年 9月 12日	无	
CN	101131704	A	2008年 2月 27日	US 2008059488 A1	2008年 3月 6日
CN	102110123	A	2011年 6月 29日	无	
US	2013024459	A1	2013年 1月 24日	无	