

(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
2019 年 8 月 8 日 (08.08.2019)



(10) 国际公布号
WO 2019/148671 A1

(51) 国际专利分类号:
G06F 8/41 (2018.01)

(21) 国际申请号: PCT/CN2018/084227

(22) 国际申请日: 2018 年 4 月 24 日 (24.04.2018)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201810091360.5 2018 年 1 月 30 日 (30.01.2018) CN

(71) 申请人: 平安科技(深圳)有限公司(PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) [CN/CN];
中国广东省深圳市福田区八卦岭八卦三路平安大厦六楼, Guangdong 518000 (CN)。

(72) 发明人: 杨启正(YANG, Qizheng); 中国广东省深圳市福田区八卦岭八卦三路平安大厦六楼, Guangdong 518000 (CN)。

(74) 代理人: 深圳市精英专利事务所(SHENZHEN TALENT PATENT SERVICE); 中国广东省深圳市福田区深南中路 6009 号绿景广场 B 栋 20 层 B, Guangdong 518000 (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, 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,

(54) Title: XML FILE PARSING METHOD, DEVICE, COMPUTER APPARATUS, AND STORAGE MEDIUM

(54) 发明名称: 一种 XML 文件解析方法、装置、计算机设备及存储介质

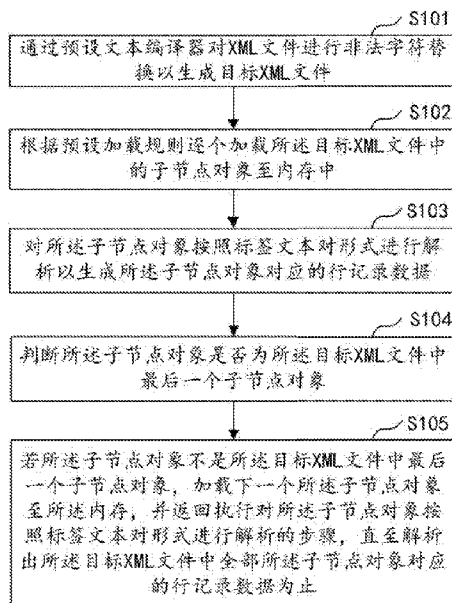


图 1

- S101 Replace invalid characters in an XML file by means of a pre-determined file compiler, so as to generate a target XML file
- S102 Successively load, according to a pre-determined loading rule, child node objects in the target XML file to a memory
- S103 Parse a given child node object in accordance with a tag-text pair format, so as to generate line record data corresponding to the given child node object
- S104 Determine whether the given child node object is the last child node object of the target XML file
- S105 If not, load a subsequent child node object, and parse the same in accordance with the tag-text pair format until all the child node objects in the target XML file have been parsed to obtain the corresponding line record data

(57) Abstract: Embodiments of the present application disclose an XML file parsing method, a device, a computer apparatus, and a storage medium. The method comprises: replacing invalid characters in an XML file by means of a pre-determined file compiler, so as to generate a target XML file; successively loading, according to a pre-determined loading rule, child node objects in the target XML file to a memory; and parsing the child node objects to generate corresponding line record data until all the child node objects in the target XML file have been parsed to obtain the corresponding line record data.

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))。

(57) 摘要: 本申请实施例公开一种XML文件解析方法、装置、计算机设备及存储介质。该方法通过预设文本编译器对XML文件进行非法字符替换以生成目标XML文件; 根据预设加载规则逐个加载目标XML文件中的子节点对象至内存中并对子节点对象进行解析以生成对应的行记录数据, 直至解析出目标XML文件中全部子节点对象对应的行记录数据为止。

发明名称：一种 XML 文件解析方法、装置、计算机设备及存储介质

本申请要求于 2018 年 1 月 30 日提交中国专利局、申请号为 201810091360.5、发明名称为“一种 XML 文件解析方法、装置、计算机设备及存储介质”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请涉及计算机技术领域，尤其涉及一种 XML 文件解析方法、装置、计算机设备及存储介质。

背景技术

在 Linux 运行环境中，一般采用 Linux 命令对 XML 文件进行全量解析。然而，在采用 Linux 命令对 XML 文件进行解析时，需要占用较多的内存空间，而且所耗时间也较长。尤其是在需要解析多个 XML 文件时，若采用并行解析多个 XML 文件，很容易导致内存溢出的错误，若采用串行解析多个 XML 文件，会花费更长的时间，解析效率低。

发明内容

本申请实施例提供了一种 XML 文件解析方法、装置、计算机设备及存储介质，以减小解析过程所占内存大小，减小解析所耗时长。

第一方面，本申请实施例提供了一种 XML 文件解析方法，其包括：

通过预设文本编译器对 XML 文件进行非法字符替换以生成目标 XML 文件；

根据预设加载规则逐个加载所述目标 XML 文件中的子节点对象至内存中；

对所述子节点对象按照标签文本对形式进行解析以生成所述子节点对象对应的行记录数据；

判断所述子节点对象是否为所述目标 XML 文件中最后一个子节点对象；

若所述子节点对象不是所述目标 XML 文件中最后一个子节点对象，加载下

一个所述子节点对象至所述内存，并返回执行对所述子节点对象按照标签文本对形式进行解析的步骤，直至解析出所述目标 XML 文件中全部所述子节点对象对应的行记录数据为止。

第二方面，本申请实施例提供了一种 XML 文件解析装置，其包括：

字符替换单元，用于通过预设文本编译器对 XML 文件进行非法字符替换以生成目标 XML 文件；

加载单元，用于根据预设加载规则逐个加载所述目标 XML 文件中的子节点对象至内存中；

解析单元，用于对所述子节点对象按照标签文本对形式进行解析以生成所述子节点对象对应的行记录数据；

对象判断单元，用于判断所述子节点对象是否为所述目标 XML 文件中最后一个子节点对象；

所述加载单元，还用于若所述子节点对象不是所述目标 XML 文件中最后一个子节点对象，加载下一个所述子节点对象至所述内存，以使得所述解析单元对所述子节点对象按照标签文本对形式进行解析，直至解析出所述目标 XML 文件中全部所述子节点对象对应的行记录数据为止。

第三方面，本申请实施例又提供了一种计算机设备，包括存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机程序，所述处理器执行所述计算机程序时实现本申请实施例提供的任一项所述的 XML 文件解析方法。

第四方面，本申请实施例还提供了一种存储介质，其中所述存储介质存储有计算机程序，所述计算机程序包括程序指令，所述程序指令当被处理器执行时使所述处理器执行本申请实施例提供的任一项所述的 XML 文件解析方法。

本申请实施例提供一种 XML 文件解析方法、装置、计算机设备及存储介质。该方法可以使得解析过程所占内存较小，不影响终端运行其他进程，尤其是对于一些内存配置较低的终端来说，这样可以避免内存溢出或内存不够用的情况发生。同时该方法可以并行解析几个或几十个 XML 文件，提高解析效率，减少解析所耗时长。

附图说明

为了更清楚地说明本申请实施例技术方案，下面将对实施例描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图是本申请的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

图1为本申请一实施例提供的一种XML文件解析方法的示意图；

图2为本申请一实施例提供的一种XML文件解析方法的另一示意图；

图3为图1所示XML文件解析方法的具体流程示意图；

图4为本申请一实施例提供的一种XML文件解析装置的示意性框图；

图5为本申请一实施例提供的一种XML文件解析装置的另一示意性框图；

图6为图4所示的XML文件解析装置的具体示意性框图；

图7为本申请一实施例提供的一种计算机设备的示意性框图。

具体实施方式

下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本申请保护的范围。

请参阅图1，图1是本申请实施例提供的一种XML文件解析方法的示意图。该XML文件解析方法应用于台式电脑、手提电脑、平板电脑等终端中。如图1所示，该XML文件解析方法包括步骤S101~S105。

S101、通过预设文本编译器对XML文件进行非法字符替换以生成目标XML文件。

在一实施例中，该预设文本编译器可以为shell语言的sed命令。譬如，sed命令格式可以为：`sed -i 's/(<userid>.*\)<(<.*<V.*>)\)/1<2/g' filename`。当然，在其他实施例中，该预设文本编译器也可以为shell语言的tr命令。譬如，tr命令格式可以为：`tr-d "\"46" <input_file> output_file`。该预设文本编译器还可以为其他用于实现对XML文件进行非法字符替换的命令，在此不做具体限制。

一般来说，当XML文件的编码格式与XML文件中的文本字体不兼容时，解析时容易出现报错。譬如，XML文件是latin-1编码格式，且XML文件中文本字体既有英文字体又有中文字体，由于中文字体与latin-1编码格式不兼容，

若不进行编码格式转换，在后续解析时就会报错。

为了避免编码格式不兼容导致的解析错误，在一实施例中，如图 2 所示，图 2 为本申请实施例提供的一种 XML 文件解析方法的示意流程图。在步骤 S101 之前，还包括步骤 S106 至 S109。

S106、获取 XML 文件的编码格式。

在一实施例中，该编码格式可以为 Latin-1 编码格式，也可以为 UTF-8 编码格式等其他编码格式，在此不做具体限制。

S107、判断所述 XML 文件的文本字体是否与所述编码格式兼容。

在一实施例中，判断所述 XML 文件的文本字体是否与所述编码格式兼容，具体为：获取所述 XML 文件中文本字体的种类，以及判断所述文本字体的种类是否与所述编码格式兼容。其中，所述文本字体的种类包括中文字体、英文字体等种类。

若判断出所述 XML 文件的文本字体与所述编码格式兼容，那么就不需要进行编码格式转换，此时可以执行步骤 S101，即进行非法字符替换步骤。

若判断出所述 XML 文件的文本字体与所述编码格式不兼容，为了避免解析时出错，此时执行步骤 S108。

S108、若所述 XML 文件的文本字体与所述编码格式不兼容，通过预设字符集编码转换命令对所述 XML 文件进行编码格式转换。

其中，该预设字符集编码转换命令可以为 Perl 语言中的 piconv 命令。具体地，可以通过 piconv 命令以字节流的方式对 XML 文件进行编码格式转换。譬如，piconv 命令可以将 XML 文件的格式从 Latin-1 编码格式转换成 UTF-8 编码格式，从而解决 XML 文件中存在中文字体时出现乱码错误问题。

由于 piconv 命令是通过字节流的方式进行编码格式转换的，所以不会一次性将整个 XML 文件放在内存中，这样可以避免一次性将整个 XML 文件放在内存中而导致内存溢出的问题，可以减缓内存压力，节省内存。

S109、将编码格式转换后的 XML 文件设为所述 XML 文件，并返回执行步骤 S101。

在 XML 文件进行编码格式转换后，将编码格式转换后的 XML 文件替换掉原来的 XML 文件，然后执行步骤 S101。

需要说明的是，在图 2 所示的实施例中，先进行了编码格式兼容性判断以

及编码格式转换等步骤，然后再进行非法字符替换的步骤。在其他实施例中，也可以先进行非法字符替换的步骤，然后再对非法字符替换后的 XML 文件进行编码格式兼容性判断、编码格式转换等步骤，可以理解的是，在此种情况下，编码格式转换后的 XML 文件将做为目标 XML 文件，然后执行步骤 S102。在此不对编码格式转换和非法字符替换两个过程的先后顺序做限制。

S102、根据预设加载规则逐个加载所述目标 XML 文件中的子节点对象至内存中。

具体地，在一实施例中，该根据预设加载规则逐个加载所述目标 XML 文件中的子节点对象至内存中包括：当识别出所述目标 XML 文件中根节点对象的起始标记符时，以所述根节点对象的下一级的子节点对象为加载单位，逐个将所述目标 XML 文件中的子节点对象加载至内存中，其中，所述根节点对象包括至少一个所述子节点对象。

譬如，目标 XML 文件的内容为：

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<bookstore>
<book>
  <title>Harry Potter</title>
  <author>J K. Rowling</author>
  <year>2005</year>
  <price>29.99</price>
</book>
<book>
  <title>Everyday Italian</title>
  <author>Giada De Laurentiis</author>
  <year>2005</year>
  <price>30.00</price>
</book>
<book>
  <title>Learning XML</title>
  <author>Erik T. Ray</author>
```

```

    <year>2003</year>
    <price>39.95</price>
</book>
<book>
    <title>XQuery Kick Start</title>
    <author>James McGovern</author>
    <year>2003</year>
    <price>49.99</price>
</book>
</bookstore>

```

在上述代码实例中，目标 XML 文件包括根节点对象，该根节点对象以<book store>为起始标记符，以</bookstore>为结束标记符。该根节点对象中包括四个子节点对象，每个子节点对象均以<book>为起始标记符，以</book>为结束标记符。四个子节点对象为根节点对象的下一级的子节点对象。每个子节点对象包括起始标记符<book>、结束标记符</book>以及两者之间的多个元素。譬如，在第一个子节点对象的四个元素中，其中一个元素为 <title>Harry Potter</title>。

可以理解的是，在其他代码实例中，子节点对象中还可以包括孙节点对象，以此规则可以嵌套多层节点对象。当目标 XML 文件中存在至少三层节点对象时，仍然以根节点对象的下一级的子节点对象为加载单位。

终端在读取到根节点对象的起始标记符<book store>之后，将以下一级的子节点对象为加载单位，逐个加载子节点对象至内存。也就是说，每次加载起始标记符<book>至结束标记符</book>之间的内容至内存中。每加载一个子节点对象就执行一次步骤 S103 等步骤以完成对该子节点对象的解析操作，然后再加载下一个子节点对象，以此类推。

S103、对所述子节点对象按照标签文本对形式进行解析以生成所述子节点对象对应的行记录数据。

具体地，在一实施例中，可以通过 iterparse()方法对所述子节点对象按照标签文本对形式进行解析以生成所述子节点对象对应的行记录数据。其中，iterparse()方法为 Python 语言中的 ElementTree 中的方法。

譬如，子节点对象<book>与</book>之间有四个元素，其中一个元素为：

`<title>Harry Potter</title>`，“Harry Potter”是该元素的文本内容，“title”是该元素的标签。那么 `iterparse()` 方法解析时，将以“>”为一个触发事件，以“</”为另一个触发事件，并得到两个触发事件之间的元素的文本内容“Harry Potter”。同时，还会解析出该文本内容“Harry Potter”的标签为“title”。标签“title”及元素的文本内容“Harry Potter”形成标签文本对。同理，可以得到其他三个标签文本对。四个标签文本对形成一个元组，一个元组为一行记录数据，如，一行记录数据 `Record = {'title': 'Harry Potter', 'author': 'J K. Rowling', 'year': '2005', 'price': '29.99'}`。

在一实施例中，为了避免内存溢出，在对所述子节点对象按照标签文本对形式进行解析以生成所述子节点对象对应的行记录数据之后，还包括：将所述行记录数据存储于行记录数据文件中；以及释放所述子节点对象所占用的内存空间。也就是说，在完成对内存中的子节点对象的解析后，将子节点对象对应的行记录数据存储于行记录数据文件中，然后再释放该子节点对象所占用的内存空间，从而为后续的子节点对象提供内存空间，避免发生内存不够或内存溢出的现象。

S104、判断所述子节点对象是否为所述目标 XML 文件中最后一个子节点对象。

在完成对内存中当前的子节点对象的解析后，将判断当前内存中的子节点对象是否为目标 XML 文件中最后一个子节点对象。

具体地，在一实施例中，如图 3 所示，图 3 为图 1 所示的 XML 文件解析方法的具体示意流程图。该步骤 S104 包括步骤 S1041 至 S1044。

S1041、读取所述目标 XML 文件中位于当前的所述子节点对象之后的字符串。

譬如，在前述代码实例中，假设当前内存中的子节点对象为根节点对象的第一个子节点对象，那么终端读取目标 XML 文件中当前内存中的子节点对象之后的字符串为“<book>”，即下一个子节点对象的起始标记符。又譬如，假设当前内存中的子节点对象为根节点对象的第四个子节点对象，那么终端读取目标 XML 文件中当前的子节点对象之后的字符串为“</bookstore>”，即根节点对象的结束标记符。

S1042、判断所述字符串是否为根节点对象的结束标记符。

譬如，判断步骤 S1041 中读取的字符串是否为根节点对象的结束标记符 </book store>。若字符串不为根节点对象的结束标记符 </book store>，则判定当前内存中的子节点对象不是目标 XML 文件中最后一个子节点对象，即执行步骤 S1043。若字符串为根节点对象的结束标记符 </book store>，说明当前内存中的子节点对象为目标 XML 文件中最后一个子节点对象，即执行步骤 S1044。

S1043、若所述字符串不为所述根节点对象的结束标记符，则判定当前的所述子节点对象不是所述目标 XML 文件中最后一个子节点对象。

S1044、若所述字符串为所述根节点对象的结束标记符，则判定当前的所述子节点对象是所述目标 XML 文件中最后一个子节点对象。

需要说明的是，判断当前内存中的子节点对象是否为目标 XML 文件中最后一个子节点对象的方式不局限于图 3 所示的方式，还可以为其他方式，在此不做具体限制。

在本实施例中，当判断出当前内存中的子节点对象为目标 XML 文件中最后一个子节点对象时，说明该目标 XML 文件已经完成了解析，解析过程可以结束了。当判断出当前内存中的子节点对象不为目标 XML 文件中最后一个子节点对象时，说明该目标 XML 文件还没有完成解析，此时需要执行步骤 S105。

S105、若所述子节点对象不是所述目标 XML 文件中最后一个子节点对象，加载下一个所述子节点对象至所述内存，并返回执行对所述子节点对象按照标签文本对形式进行解析以生成所述子节点对象对应的行记录数据的步骤，直至解析出所述目标 XML 文件中全部所述子节点对象对应的行记录数据为止。

若当前的子节点对象不是目标 XML 文件中最后一个子节点对象，也即该目标 XML 文件还没有解析完，此时将继续读取下一个子节点对象至内存中，然后返回执行步骤 S103 至 S105，一直到步骤 S104 判断出内存中的子节点对象为目标 XML 文件中最后一个子节点对象为止，从而完成整个目标 XML 文件的解析过程。

本实施例中的 XML 文件解析方法，逐个加载子节点对象至内存并进行解析，可以使得解析过程所占内存较小，避免内存溢出或内存不够用的情况发生。同时，该方法在解析 XML 文件时可以实现并行解析几个或几十个 XML 文件，进而大大缩短解析所耗时间，提高解析效率。

本申请实施例还提供一种 XML 文件解析装置，该 XML 文件解析装置用于执行前述任一项 XML 文件解析方法。具体地，请参阅图 4，图 4 是本申请实施例提供的一种 XML 文件解析装置的示意性框图。XML 文件解析装置 300 可以安装于台式电脑、平板电脑、手提电脑、等终端中。

如图 4 所示，XML 文件解析装置 300 包括字符替换单元 301、加载单元 302、解析单元 303 和对象判断单元 304。

字符替换单元 301，用于通过预设文本编译器对 XML 文件进行非法字符替换以生成目标 XML 文件。

在一实施例中，为了避免编码格式不兼容导致的解析错误，如图 5 所示，图 5 为本申请实施例提供的一种 XML 文件解析装置的另一示意性框图。该 XML 文件解析装置 300 还包括格式获取单元 305、格式判断单元 306 和编码转换单元 307。

格式获取单元 305，用于获取 XML 文件的编码格式。

格式判断单元 306，用于判断所述 XML 文件的文本字体是否与所述编码格式兼容。

在一实施例中，格式判断单元 306 具体用于获取所述 XML 文件中文本字体的种类，以及判断所述文本字体的种类是否与所述编码格式兼容。若格式判断单元 306 判断出所述 XML 文件的文本字体与所述编码格式兼容，那么就不需要进行编码格式转换，此时格式判断单元 306 向字符替换单元 301 发送第一信号，该字符替换单元 301 在接收到第一信号后，通过预设文本编译器对 XML 文件进行非法字符替换以生成目标 XML 文件。

编码转换单元 307，用于若所述 XML 文件的文本字体与所述编码格式不兼容，通过预设字符集编码转换命令对所述 XML 文件进行编码格式转换；以及将编码格式转换后的 XML 文件设为所述 XML 文件。

编码转换单元 307 可以通过 `piconv` 命令以字节流的方式对 XML 文件进行编码格式转换。

编码转换单元 307 在将 XML 文件进行编码格式转换后，将编码格式转换后的 XML 文件替换掉原来的 XML 文件，然后编码转换单元 307 向字符替换单元 301 发送第三信号，以使得该字符替换单元 301 在接收到第三信号后，通过预设文本编译器对 XML 文件进行非法字符替换以生成目标 XML 文件。

加载单元 302, 用于根据预设加载规则逐个加载所述目标 XML 文件中的子节点对象至内存中。

在一实施例中, 加载单元 302 具体用于当识别出所述目标 XML 文件中根节点对象的起始标记符时, 以所述根节点对象的下一级的子节点对象为加载单位, 逐个将所述目标 XML 文件中的子节点对象加载至内存中, 其中, 所述根节点对象包括至少一个所述子节点对象。

解析单元 303, 用于对所述子节点对象按照标签文本对形式进行解析以生成所述子节点对象对应的行记录数据。

具体地, 在一实施例中, 解析单元 303 可以通过 `iterparse()` 方法对所述子节点对象按照标签文本对形式进行解析以生成所述子节点对象对应的行记录数据。其中, `iterparse()` 方法为 Python 语言中的 ElementTree 中的方法。

在一实施例中, 为了避免内存溢出, 如图 5 所示, 该 XML 文件解析装置 300 还包括存储单元 308 和内存释放单元 309。存储单元 308 用于将所述行记录数据存储于行记录数据文件中。内存释放单元 309 用于释放所述子节点对象所占用的内存空间。

对象判断单元 304, 用于判断所述子节点对象是否为所述目标 XML 文件中最后一个子节点对象。

具体地, 在一实施例中, 如图 6 所示, 图 6 为图 4 所示的 XML 文件解析装置的具体示意性框图。该对象判断单元 304 包括读取子单元 3041 和判断子单元 3042。读取子单元 3041 用于读取所述目标 XML 文件中位于当前的所述子节点对象之后的字符串。判断子单元 3042 用于判断所述字符串是否为根节点对象的结束标记符。若判断子单元 3042 判断出所述字符串不为所述根节点对象的结束标记符, 则判定当前的所述子节点对象不是所述目标 XML 文件中最后一个子节点对象。若判断子单元 3042 判断出所述字符串为所述根节点对象的结束标记符, 则判定当前的所述子节点对象是所述目标 XML 文件中最后一个子节点对象。

需要说明的是, 所属领域的技术人员可以清楚地了解到, 为了描述的方便和简洁, 上述描述的 XML 文件解析装置 300 和单元的具体工作过程, 可以参考前述 XML 文件解析方法实施例中的对应过程, 在此不再赘述。

本实施例中的 XML 文件解析装置 300, 逐个加载子节点对象至内存并进行解析, 可以使得解析过程所占内存较小, 避免内存溢出或内存不够用的情况发

生。同时，该 XML 文件解析装置 300 可以实现并行解析几个或几十个 XML 文件，进而大大缩短解析所耗时间，提高解析效率。

上述 XML 文件解析装置可以实现为一种计算机程序的形式，该计算机程序可以在如图 7 所示的计算机设备上运行。请参阅图 7，图 7 是本申请实施例提供的一种计算机设备的示意性框图。该计算机设备 400 可以是终端。该终端可以是平板电脑、笔记本电脑、台式电脑、个人数字助理等电子设备。

该计算机设备 400 包括通过系统总线 401 连接的处理器 402、存储器和网络接口 405，其中，存储器可以包括非易失性存储介质 403 和内存储器 404。

该非易失性存储介质 403 可存储操作系统 4031 和计算机程序 4032。该计算机程序 4032 包括程序指令，该程序指令被执行时，可使得处理器 402 执行一种 XML 文件解析方法。

该处理器 402 用于提供计算和控制能力，支撑整个计算机设备 400 的运行。

该内存储器 404 为非易失性存储介质 403 中的计算机程序 4032 的运行提供环境，该计算机程序 4032 被处理器 402 执行时，可使得处理器 402 执行一种 XML 文件解析方法。

该网络接口 405 用于进行网络通信，如发送分配的任务等。本领域技术人员可以理解，图 7 中示出的结构，仅仅是与本申请方案相关的部分结构的框图，并不构成对本申请方案所应用于其上的计算机设备 400 的限定，具体的计算机设备 400 可以包括比图中所示更多或更少的部件，或者组合某些部件，或者具有不同的部件布置。

其中，所述处理器 402 用于运行存储在存储器中的计算机程序 4032，以实现如下功能：通过预设文本编译器对 XML 文件进行非法字符替换以生成目标 XML 文件；根据预设加载规则逐个加载所述目标 XML 文件中的子节点对象至内存中；对所述子节点对象按照标签文本对形式进行解析以生成所述子节点对象对应的行记录数据；判断所述子节点对象是否为所述目标 XML 文件中最后一个子节点对象；若所述子节点对象不是所述目标 XML 文件中最后一个子节点对象，加载下一个所述子节点对象至所述内存，并返回执行对所述子节点对象按照标签文本对形式进行解析的步骤，直至解析出所述目标 XML 文件中全部所述子节点对象对应的行记录数据为止。

在一实施例中，处理器 402 在执行通过预设文本编译器对 XML 文件进行非

法字符替换以生成目标 XML 文件之前，还执行如下程序：获取 XML 文件的编码格式；判断所述 XML 文件的文本字体是否与所述编码格式兼容；若所述 XML 文件的文本字体与所述编码格式兼容，则执行通过预设文本编译器对 XML 文件进行非法字符替换以生成目标 XML 文件的步骤；若所述 XML 文件的文本字体与所述编码格式不兼容，通过预设字符集编码转换命令对所述 XML 文件进行编码格式转换；以及将编码格式转换后的 XML 文件设为所述 XML 文件，并执行通过预设文本编译器对 XML 文件进行非法字符替换以生成目标 XML 文件的步骤。

在一实施例中，处理器 402 在执行根据预设加载规则逐个加载所述目标 XML 文件中的子节点对象至内存中时，具体执行如下程序：当识别出所述目标 XML 文件中根节点对象的起始标记符时，以所述根节点对象的下一级的子节点对象为加载单位，逐个将所述目标 XML 文件中的子节点对象加载至内存中，其中，所述根节点对象包括至少一个所述子节点对象。

在一实施例中，处理器 402 在执行判断所述子节点对象是否为所述目标 XML 文件中最后一个子节点对象时，具体执行如下程序：读取所述目标 XML 文件中位于当前的所述子节点对象之后的字符串；判断所述字符串是否为根节点对象的结束标记符；若所述字符串不为所述根节点对象的结束标记符，则判定当前的所述子节点对象不是所述目标 XML 文件中最后一个子节点对象。

在一实施例中，处理器 402 在执行对所述子节点对象按照标签文本对形式进行解析以生成所述子节点对象对应的行记录数据之后，还执行如下程序：将所述行记录数据存储于行记录数据文件中；以及释放所述子节点对象所占用的内存空间。

在一实施例中，处理器 402 在执行对所述子节点对象按照标签文本对形式进行解析以生成所述子节点对象对应的行记录数据时，具体执行如下程序：通过 `iterparse()` 方法对所述子节点对象按照标签文本对形式进行解析以生成所述子节点对象对应的行记录数据。

应当理解，在本申请实施例中，处理器 402 可以是中央处理单元，还可以是其他通用处理器、数字信号处理器、专用集成电路、现成可编程门阵列或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。其中，通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等。

在本申请的另一实施例中提供一种存储介质。该存储介质可以为计算机可读存储介质。该存储介质存储有计算机程序，其中计算机程序包括程序指令。该程序指令被处理器执行时使处理器执行本申请中的一种 XML 文件解析方法。该存储介质可以是 U 盘、移动硬盘、只读存储器(ROM, Read-Only Memory)、闪存、磁碟或者光盘等各种可以存储程序代码的介质。

本领域普通技术人员可以意识到，结合本文中所公开的实施例描述的各示例的单元及算法步骤，能够以电子硬件、计算机软件或者二者的结合来实现。

在本申请所提供的几个实施例中，应该理解到，所揭露的装置和方法，可以通过其它的方式实现。例如多个单元或组件可以结合或者可以集成到另一个系统，或一些特征可以忽略，或不执行。本申请实施例方法中的步骤可以根据实际需要进行顺序调整、合并和删减。本申请实施例装置中的单元可以根据实际需要进行合并、划分和删减。另外，在本申请各个实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以是两个或两个以上单元集成在一个单元中。

该集成的单元如果以软件功能单元的形式实现并作为独立的产品销售或使用，可以存储在一个存储介质中。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分，或者该技术方案的全部或部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质中，包括若干指令用以使得一台计算机设备（可以是个人计算机，终端，或者网络设备等）执行本申请各个实施例所述方法的全部或部分步骤。

以上所述，仅为本申请的具体实施方式，但本申请的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本申请揭露的技术范围内，可轻易想到各种等效的修改或替换，这些修改或替换都应涵盖在本申请的保护范围之内。因此，本申请的保护范围应以权利要求的保护范围为准。

权 利 要 求 书

1. 一种 XML 文件解析方法，其包括：

通过预设文本编译器对 XML 文件进行非法字符替换以生成目标 XML 文件；

根据预设加载规则逐个加载所述目标 XML 文件中的子节点对象至内存中；

对所述子节点对象按照标签文本对形式进行解析以生成所述子节点对象对应的行记录数据；

判断所述子节点对象是否为所述目标 XML 文件中最后一个子节点对象；

若所述子节点对象不是所述目标 XML 文件中最后一个子节点对象，加载下一个所述子节点对象至所述内存，并返回执行对所述子节点对象按照标签文本对形式进行解析的步骤，直至解析出所述目标 XML 文件中全部所述子节点对象对应的行记录数据为止。

2. 根据权利要求 1 所述的 XML 文件解析方法，其中，在所述通过预设文本编译器对 XML 文件进行非法字符替换以生成目标 XML 文件之前，还包括：

获取 XML 文件的编码格式；

判断所述 XML 文件的文本字体是否与所述编码格式兼容；

若所述 XML 文件的文本字体与所述编码格式兼容，则执行通过预设文本编译器对 XML 文件进行非法字符替换以生成目标 XML 文件的步骤；

若所述 XML 文件的文本字体与所述编码格式不兼容，通过预设字符集编码转换命令对所述 XML 文件进行编码格式转换；以及

将编码格式转换后的 XML 文件设为所述 XML 文件，并执行通过预设文本编译器对 XML 文件进行非法字符替换以生成目标 XML 文件的步骤。

3. 根据权利要求 1 所述的 XML 文件解析方法，其中，所述根据预设加载规则逐个加载所述目标 XML 文件中的子节点对象至内存中，包括：

当识别出所述目标 XML 文件中根节点对象的起始标记符时，以所述根节点对象的下一级的子节点对象为加载单位，逐个将所述目标 XML 文件中的子节点对象加载至内存中，其中，所述根节点对象包括至少一个所述子节点对象。

4. 根据权利要求 1 所述的 XML 文件解析方法，其中，所述判断所述子节点对象是否为所述目标 XML 文件中最后一个子节点对象，包括：

读取所述目标 XML 文件中位于当前的所述子节点对象之后的字符串；

判断所述字符串是否为根节点对象的结束标记符；

若所述字符串不为所述根节点对象的结束标记符，则判定当前的所述子节点对象不是所述目标 XML 文件中最后一个子节点对象。

5. 根据权利要求 1 所述的 XML 文件解析方法，其中，在所述对所述子节点对象按照标签文本对形式进行解析以生成所述子节点对象对应的行记录数据之后，还包括：

将所述行记录数据存储于行记录数据文件中；以及

释放所述子节点对象所占用的内存空间。

6. 根据权利要求 1 所述的 XML 文件解析方法，其中，所述对所述子节点对象按照标签文本对形式进行解析以生成所述子节点对象对应的行记录数据，包括：通过 `iterparse()` 方法对所述子节点对象按照标签文本对形式进行解析以生成所述子节点对象对应的行记录数据。

7. 根据权利要求 2 所述的 XML 文件解析方法，其中，所述通过预设字符集编码转换命令对所述 XML 文件进行编码格式转换，包括：通过 `piconv` 命令以字节流的方式对所述 XML 文件进行编码格式转换。

8. 一种 XML 文件解析装置，其包括：

字符替换单元，用于通过预设文本编译器对 XML 文件进行非法字符替换以生成目标 XML 文件；

加载单元，用于根据预设加载规则逐个加载所述目标 XML 文件中的子节点对象至内存中；

解析单元，用于对所述子节点对象按照标签文本对形式进行解析以生成所述子节点对象对应的行记录数据；

对象判断单元，用于判断所述子节点对象是否为所述目标 XML 文件中最后一个子节点对象；

所述加载单元，还用于若所述子节点对象不是所述目标 XML 文件中最后一个子节点对象，加载下一个所述子节点对象至所述内存，以使得所述解析单元对所述子节点对象按照标签文本对形式进行解析，直至解析出所述目标 XML 文件中全部所述子节点对象对应的行记录数据为止。

9. 根据权利要求 8 所述的 XML 文件解析装置，其中，还包括：

格式获取单元，用于获取 XML 文件的编码格式；

格式判断单元，用于判断所述 XML 文件的文本字体是否与所述编码格式兼容；

编码转换单元，用于若所述 XML 文件的文本字体与所述编码格式不兼容，通过预设字符集编码转换命令对所述 XML 文件进行编码格式转换；以及将编码格式转换后的 XML 文件设为所述 XML 文件；

所述字符替换单元，具体用于在所述编码转换单元将编码格式转换后的 XML 文件设为所述 XML 文件之后，通过预设文本编译器对 XML 文件进行非法字符替换以生成目标 XML 文件；

所述字符替换单元，还具体用于若所述 XML 文件的文本字体与所述编码格式兼容，则通过预设文本编译器对 XML 文件进行非法字符替换以生成目标 XML 文件。

10. 一种计算机设备，包括存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机程序，其中，所述处理器执行所述计算机程序时实现：通过预设文本编译器对 XML 文件进行非法字符替换以生成目标 XML 文件；根据预设加载规则逐个加载所述目标 XML 文件中的子节点对象至内存中；对所述子节点对象按照标签文本对形式进行解析以生成所述子节点对象对应的行记录数据；判断所述子节点对象是否为所述目标 XML 文件中最后一个子节点对象；若所述子节点对象不是所述目标 XML 文件中最后一个子节点对象，加载下一个所述子节点对象至所述内存，并返回执行对所述子节点对象按照标签文本对形式进行解析的步骤，直至解析出所述目标 XML 文件中全部所述子节点对象对应的行记录数据为止。

11. 根据权利要求 10 所述的计算机设备，其中，所述处理器执行通过预设文本编译器对 XML 文件进行非法字符替换以生成目标 XML 文件之前，还实现：获取 XML 文件的编码格式；判断所述 XML 文件的文本字体是否与所述编码格式兼容；若所述 XML 文件的文本字体与所述编码格式兼容，则执行通过预设文本编译器对 XML 文件进行非法字符替换以生成目标 XML 文件的步骤；若所述 XML 文件的文本字体与所述编码格式不兼容，通过预设字符集编码转换命令对所述 XML 文件进行编码格式转换；以及将编码格式转换后的 XML 文件设为所述 XML 文件，并执行通过预设文本编译器对 XML 文件进行非法字符替换以生

成目标 XML 文件的步骤。

12. 根据权利要求 10 所述的计算机设备, 其中, 所述处理器执行根据预设加载规则逐个加载所述目标 XML 文件中的子节点对象至内存中时, 具体实现: 当识别出所述目标 XML 文件中根节点对象的起始标记符时, 以所述根节点对象的下一级的子节点对象为加载单位, 逐个将所述目标 XML 文件中的子节点对象加载至内存中, 其中, 所述根节点对象包括至少一个所述子节点对象。

13. 根据权利要求 10 所述的计算机设备, 其中, 所述处理器执行判断所述子节点对象是否为所述目标 XML 文件中最后一个子节点对象时, 具体实现: 读取所述目标 XML 文件中位于当前的所述子节点对象之后的字符串; 判断所述字符串是否为根节点对象的结束标记符; 若所述字符串不为所述根节点对象的结束标记符, 则判定当前的所述子节点对象不是所述目标 XML 文件中最后一个子节点对象。

14. 根据权利要求 10 所述的计算机设备, 其中, 所述处理器执行对所述子节点对象按照标签文本对形式进行解析以生成所述子节点对象对应的行记录数据之后, 还实现: 将所述行记录数据存储于行记录数据文件中; 以及释放所述子节点对象所占用的内存空间。

15. 根据权利要求 10 所述的计算机设备, 其中, 所述处理器执行对所述子节点对象按照标签文本对形式进行解析以生成所述子节点对象对应的行记录数据时, 具体实现: 通过 `iterparse()` 方法对所述子节点对象按照标签文本对形式进行解析以生成所述子节点对象对应的行记录数据。

16. 一种存储介质, 其中, 所述存储介质存储有计算机程序, 所述计算机程序包括程序指令, 所述程序指令当被处理器执行时使所述处理器执行: 通过预设文本编译器对 XML 文件进行非法字符替换以生成目标 XML 文件; 根据预设加载规则逐个加载所述目标 XML 文件中的子节点对象至内存中; 对所述子节点对象按照标签文本对形式进行解析以生成所述子节点对象对应的行记录数据; 判断所述子节点对象是否为所述目标 XML 文件中最后一个子节点对象; 若所述子节点对象不是所述目标 XML 文件中最后一个子节点对象, 加载下一个所述子节点对象至所述内存, 并返回执行对所述子节点对象按照标签文本对形式进行解析的步骤, 直至解析出所述目标 XML 文件中全部所述子节点对象对应的行记录数据为止。

17. 根据权利要求 16 所述的存储介质, 其中, 所述程序指令当被处理器执行通过预设文本编译器对 XML 文件进行非法字符替换以生成目标 XML 文件之前, 还使所述处理器执行: 获取 XML 文件的编码格式; 判断所述 XML 文件的文本字体是否与所述编码格式兼容; 若所述 XML 文件的文本字体与所述编码格式兼容, 则执行通过预设文本编译器对 XML 文件进行非法字符替换以生成目标 XML 文件的步骤; 若所述 XML 文件的文本字体与所述编码格式不兼容, 通过预设字符集编码转换命令对所述 XML 文件进行编码格式转换; 以及将编码格式转换后的 XML 文件设为所述 XML 文件, 并执行通过预设文本编译器对 XML 文件进行非法字符替换以生成目标 XML 文件的步骤。

18. 根据权利要求 16 所述的存储介质, 其中, 所述程序指令当被处理器执行根据预设加载规则逐个加载所述目标 XML 文件中的子节点对象至内存中时使所述处理器执行: 当识别出所述目标 XML 文件中根节点对象的起始标记符时, 以所述根节点对象的下一级的子节点对象为加载单位, 逐个将所述目标 XML 文件中的子节点对象加载至内存中, 其中, 所述根节点对象包括至少一个所述子节点对象。

19. 根据权利要求 16 所述的存储介质, 其中, 所述程序指令当被处理器执行判断所述子节点对象是否为所述目标 XML 文件中最后一个子节点对象时使所述处理器执行: 读取所述目标 XML 文件中位于当前的所述子节点对象之后的字符串; 判断所述字符串是否为根节点对象的结束标记符; 若所述字符串不为所述根节点对象的结束标记符, 则判定当前的所述子节点对象不是所述目标 XML 文件中最后一个子节点对象。

20. 根据权利要求 16 所述的存储介质, 其中, 所述程序指令当被处理器执行对所述子节点对象按照标签文本对形式进行解析以生成所述子节点对象对应的行记录数据之后, 还使所述处理器执行: 将所述行记录数据存储于行记录数据文件中; 以及释放所述子节点对象所占用的内存空间。

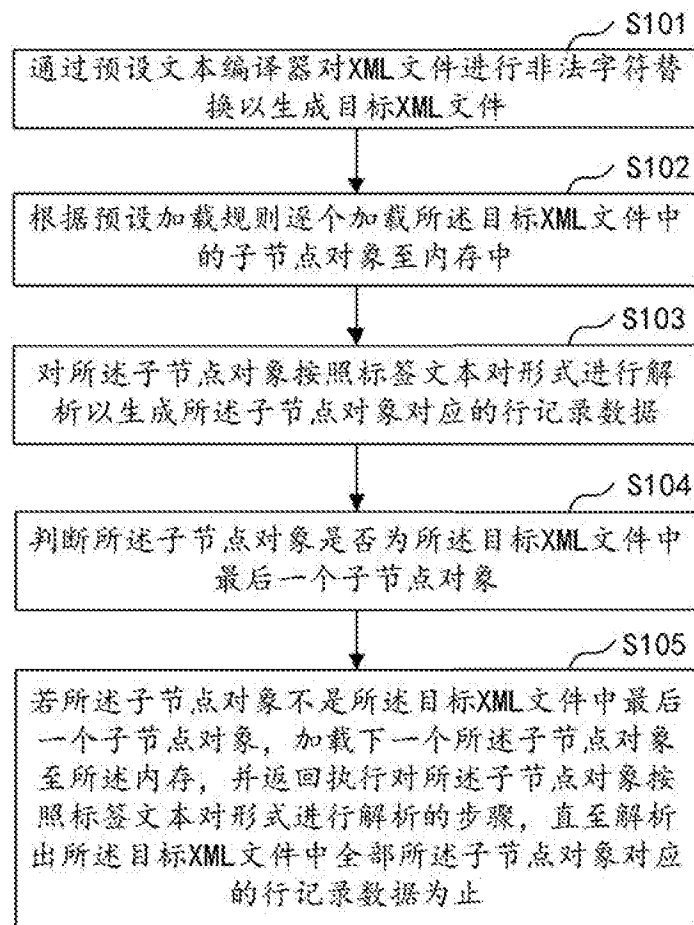


图 1

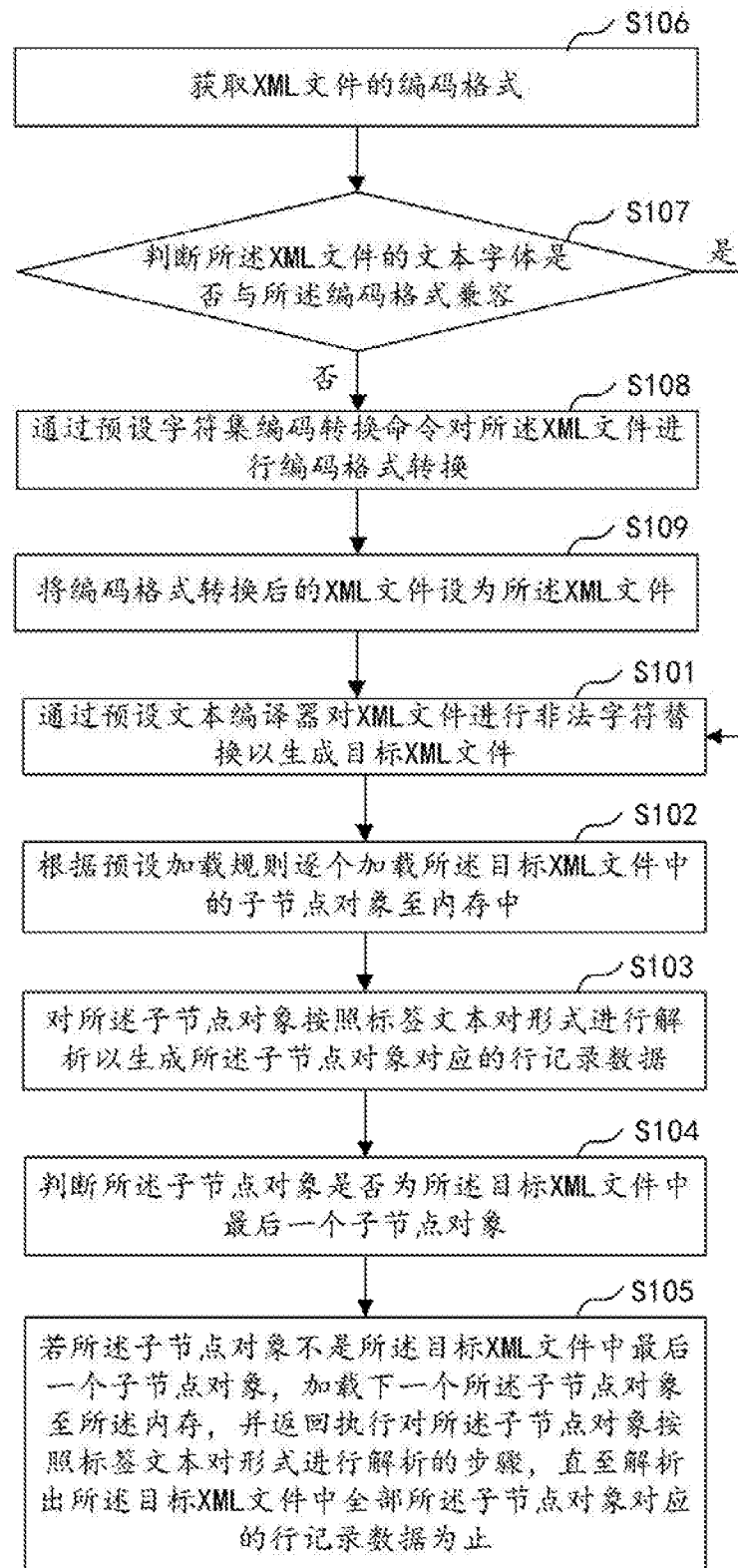


图 2

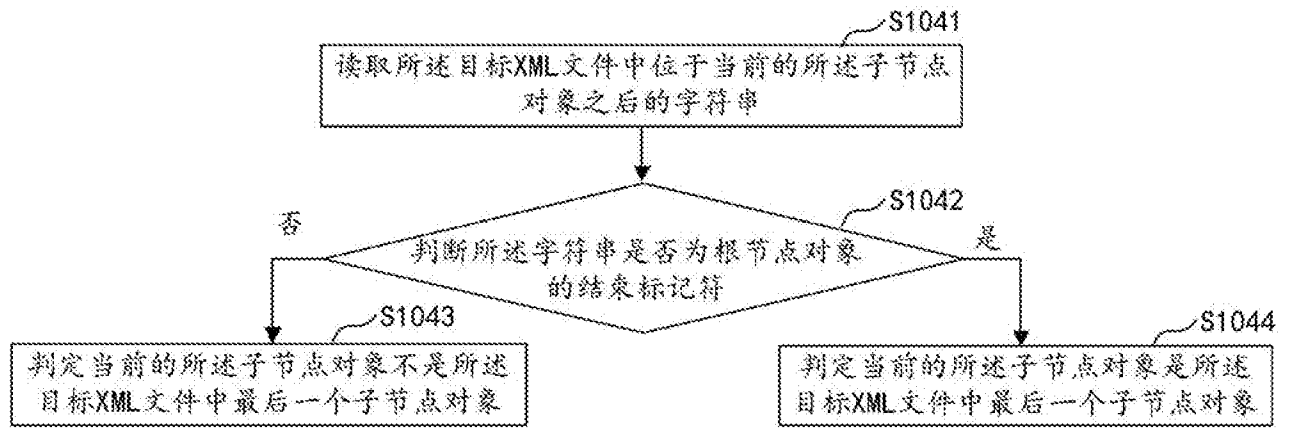


图 3

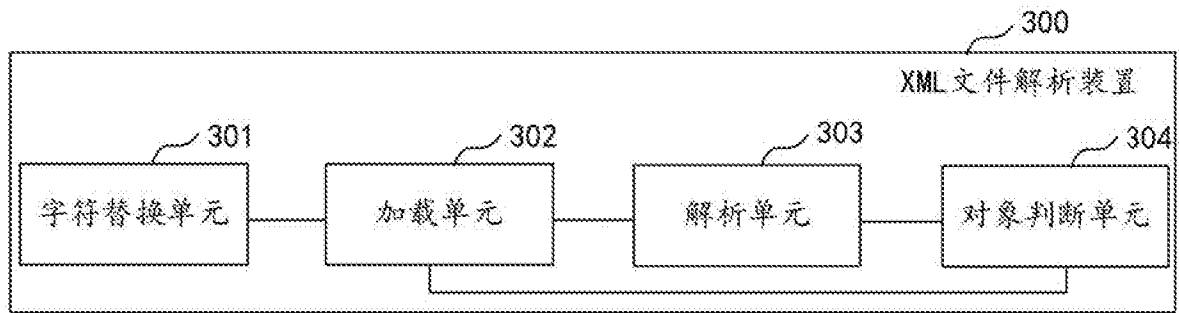


图 4

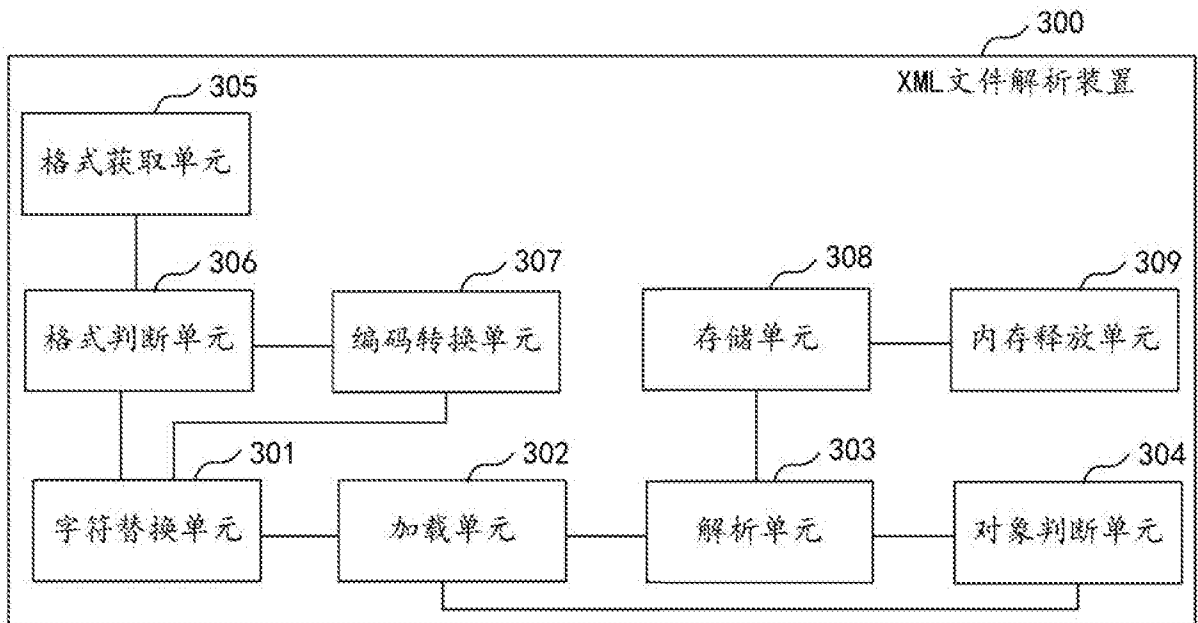


图 5

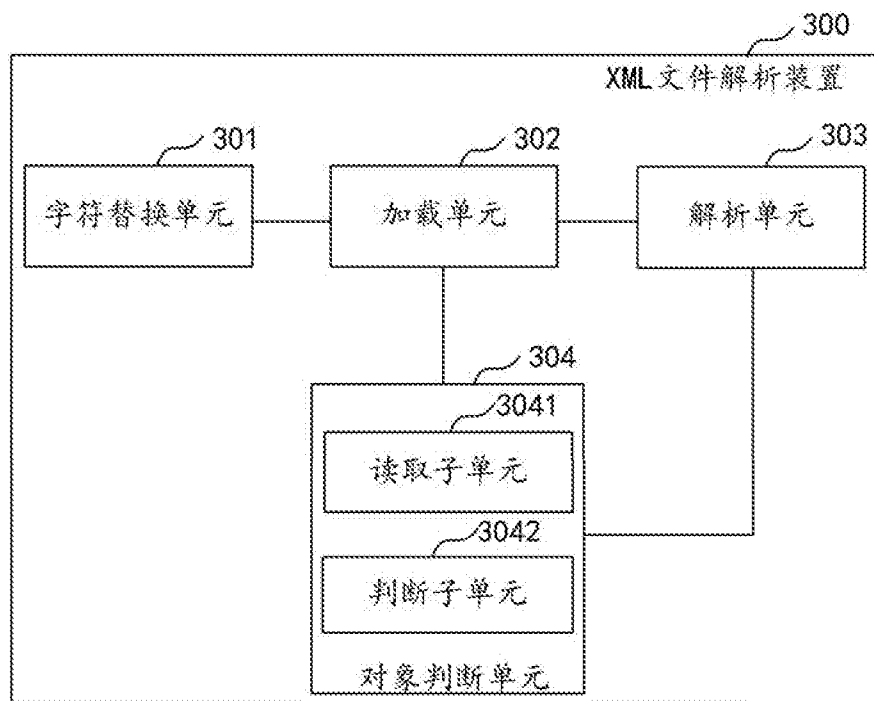


图 6

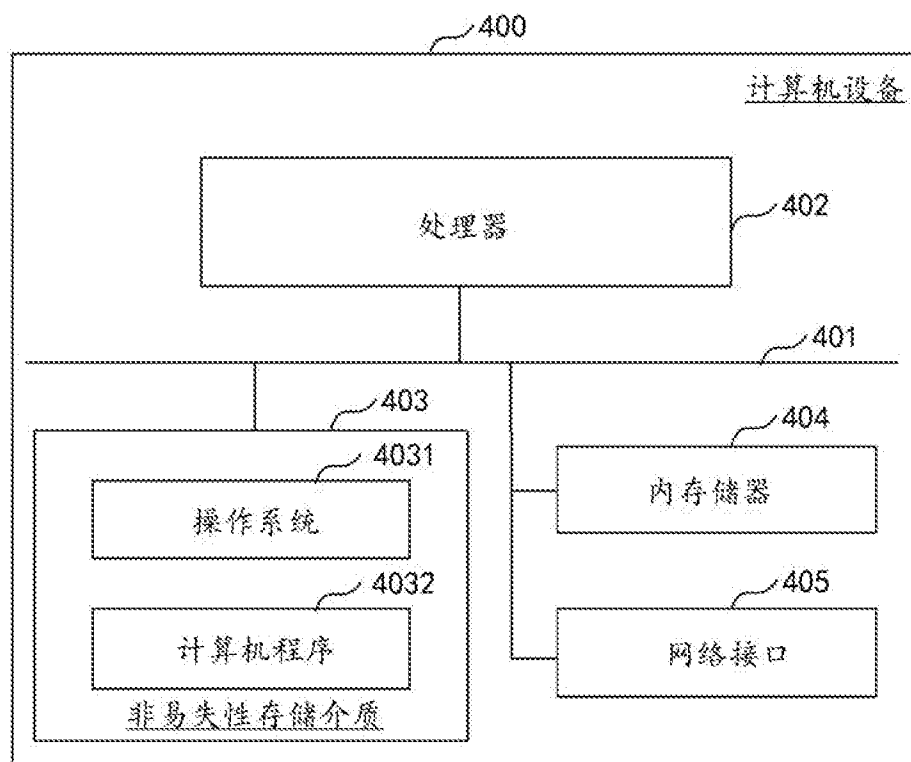


图 7

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2018/084227

A. CLASSIFICATION OF SUBJECT MATTER

G06F 8/41(2018.01)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)

CNTXT; CNABS; CNKI; TWTXT; TWABS; USTXT; VEN; EPTXT; WOTXT: 可扩展标示语言, 解析, 内存, 溢出, 节点, 行, 格式, 加载, XML, parse, resolve, memory, buffer, overflow, node, row, line, format, load

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 106469137 A (INTERNET DOMAIN NAME SYSTEM BEIJING ENGINEERING RESEARCH CENTER ET AL.) 01 March 2017 (2017-03-01) description, paragraphs 6-16 and 48	1-20
A	CN 107577460 A (SUZHOU YOSEMEI INTELLIGENT SYSTEM CO., LTD.) 12 January 2018 (2018-01-12) entire document	1-20
A	CN 104112015 A (STATE GRID CORPORATION OF CHINA ET AL.) 22 October 2014 (2014-10-22) entire document	1-20
A	US 7499915 B2 (ORACLE INTERNATIONAL CORPORATION) 03 March 2009 (2009-03-03) entire document	1-20



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

11 October 2018

Date of mailing of the international search report

25 October 2018

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

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2018/084227

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	106469137	A	01 March 2017	None			
CN	107577460	A	12 January 2018	None			
CN	104112015	A	22 October 2014	None			
US	7499915	B2	03 March 2009	US	7461074	B2	02 December 2008
				US	2005228818	A1	13 October 2005
				US	2005228792	A1	13 October 2005

国际检索报告

国际申请号

PCT/CN2018/084227

A. 主题的分类

G06F 8/41 (2018.01) 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))

CNTXT;CNABS;CNKI;TWTXT;TWABS;USTXT;VEN;EPTXT;WOTXT:可扩展标示语言, 解析, 内存, 溢出, 节点, 行, 格式, 加载, XML, parse, resolve, memory, buffer, overflow, node, row, line, format, load

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
X	CN 106469137 A (互联网域名系统北京市工程研究中心有限公司等) 2017年 3月 1日 (2017 - 03 - 01) 说明书第6-16、48段	1-20
A	CN 107577460 A (苏州优圣美智能系统有限公司) 2018年 1月 12日 (2018 - 01 - 12) 全文	1-20
A	CN 104112015 A (国家电网公司等) 2014年 10月 22日 (2014 - 10 - 22) 全文	1-20
A	US 7499915 B2 (ORACLE INTERNATIONAL CORPORATION) 2009年 3月 3日 (2009 - 03 - 03) 全文	1-20

☐ 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

2018年 10月 11日

国际检索报告邮寄日期

2018年 10月 25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

杨牛

电话号码 86-(20)-28950388

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2018/084227

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	106469137	A	2017年 3月 1日	无			
CN	107577460	A	2018年 1月 12日	无			
CN	104112015	A	2014年 10月 22日	无			
US	7499915	B2	2009年 3月 3日	US	7461074	B2	2008年 12月 2日
				US	2005228818	A1	2005年 10月 13日
				US	2005228792	A1	2005年 10月 13日