(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)

WIPO | PCT

(10) 国际公布号
WO 2018/010573 A1

(51) 国际专利分类号:
G06F 17/30 (2006.01) *G06F 9/44* (2006.01)

(21) 国际申请号: PCT/CN2017/091674

(22) 国际申请日: 2017 年 7 月 4 日 (04.07.2017)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201610551151.5 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*): 孙宇 (SUN, Yu) [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, 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,

(54) **Title:** METHOD AND DEVICE FOR GENERATING SCRIPT

(54) **发明名称**: 一种脚本生成方法与装置

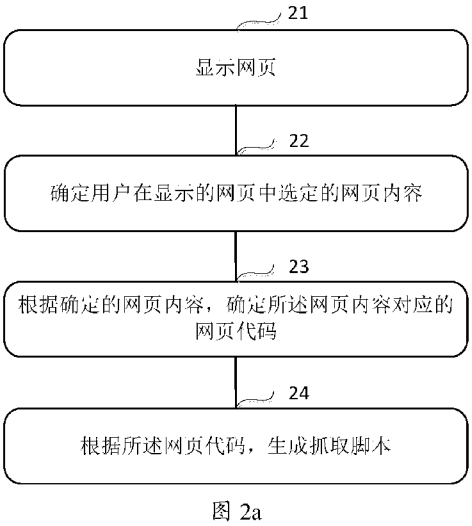


图 2a

21 Display a web page
22 Determine web page content selected by a user in the displayed web page
23 Determine, according to the determined web page content, web page code corresponding to the web page content
24 Generate a crawl script according to the web page code

(57) **Abstract:** A method and a device for generating a script for solving the problem in the prior art of low efficiency associated with manually writing a crawl script when using a web crawler to capture web page content. The method comprises: determining web page content selected by a user in a displayed web page (22); determining, according to the determined web page content, web page code corresponding to the displayed web page content (23); and generating a crawl script according to the web page code (24).

(57) **摘要**: 一种脚本生成方法与装置, 用于解决现有技术中利用网络爬虫抓取网页内容时, 人工编写抓取脚本效率较低的问题。该方法包括: 确定用户在显示的网页中选定的网页内容 (22); 根据确定的网页内容, 确定显示的网页内容对应的网页代码 (23); 根据所述网页代码, 生成抓取脚本 (24)。

WO 2018/010573 A1

[见续页]

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))。

一种脚本生成方法与装置

技术领域

本申请涉及计算机技术领域，尤其涉及一种脚本生成方法与装置。

背景技术

在现有技术中，由于网络爬虫可以抓取网页中的文本内容，目前被广泛的应用于搜索、数据挖掘等领域中。网络爬虫可以抓取网页中的全部内容，也可以抓取网页中的部分内容。

目前，若要利用网络爬虫去抓取目标网页中的目标内容，工作人员需先编写抓取目标内容的脚本，网络爬虫才能够根据该脚本，抓取到目标内容。

例如，若现在想要利用网络爬虫抓取如图 1 中所示的网页中的商品的价格信息，即“价钱：$149.99”。那么，工作人员便要通过浏览器访问相应的网页，然后在该网页对应的网页代码中查找“价钱：$149.99”对应的网页代码，即“价钱：$149.99”对应的最小的文档对象模型（Document Object Model，dom）树。

其中，“价钱：$149.99”对应的最小的 dom 树如下所示：

<div id="kfs_family_16" class="kfs-inner-container kfs-selected" style="width:20%;left:40%;background-image:url(https://images-na.ssl-images-amazon.com/images/G/01/kindle/stripe/kfs-selector-2._CB386844303_.gif);" onClick="javascript:(function(){})()">

<imgclass="kfs-img"style="margin-top:9px;"src="https://images-na.ssl-images-amazon.com/images/G/01/kindle/dp/2015/848470/famnav/fs-m._CB292709393_.png" />

Fire HD 8

$149.99

<divid="kfs_popover_content_16"=class="kfs-popover-container"style="display:none;">Incredibly thin and light, designed for entertainment

在查找到“价钱: $149.99”对应的最小的 dom 树后，获取“价钱: $149.99”对应的超文本标记语言（HyperText Markup Language，HTML）属性值信息，比如 id="kfs_family_16"、class="kfs-price"等。工作人员根据该些属性值信息，编写出包含该些 HTML 属性值信息的抓取脚本。将编写好的脚本以及该商品对应的网页代码一并送入解析引擎中，以使得解析引擎可以根据抓取脚本中的 id 以及 class，查找到“价钱: $149.99”对应的最小的 dom 树，并在该最小的 dom 树中提取出“价钱: $149.99”这一价格信息。

虽然通过上述方法，网络爬虫能够抓取网页中的内容，但是需要人工编写抓取脚本，效率较低。

发明内容

本申请实施例提供一种脚本生成方法与装置，用于解决现有技术中利用网络爬虫抓取网页内容时，人工编写抓取脚本效率较低的问题。

本申请实施例采用下述技术方案:

一种脚本生成方法，包括:

确定用户在显示的网页中选定的网页内容;

根据确定的网页内容，确定所述网页内容对应的网页代码;

根据所述网页代码，生成抓取脚本。

一种脚本生成装置，包括:

内容确定模块，确定用户在显示的网页中选定的网页内容;

代码确定模块，根据确定的网页内容，确定所述网页内容对应的网页代码;

脚本生成模块，根据所述网页代码，生成抓取脚本。

本申请实施例采用的上述至少一个技术方案能够达到以下有益效果：

与现有技术中利用网络爬虫抓取网页内容时，需人工编写抓取脚本相比，采用本申请实施例提供的脚本生成方法，通过确定用户在网页中选定的网页内容，确定出该网页内容对应的网页代码，并根据该网页代码生成抓取脚本，从而解决了现有技术中利用网络爬虫抓取网页内容时，人工编写抓取脚本效率较低的问题。

附图说明

此处所说明的附图用来提供对本申请的进一步理解，构成本申请的一部分，本申请的示意性实施例及其说明用于解释本申请，并不构成对本申请的不当限定。在附图中：

图 1 为现有技术中的目标网页中的内容；

图 2a 为本申请实施例提供的一种脚本生成方法的具体流程图；

图 2b 为本申请实施例提供的确定 HTML 属性值的页面；

图 2c 为本申请实施例提供的询问用户抓取哪一种网页内容的页面；

图 2d 为本申请实施例提供的用户框选网页内容后显示出的页面；

图 2e 为本申请实施例提供的用户两次框选网页内容后显示出的页面；

图 3 为本申请实施例提供的一种脚本生成装置的具体结构示意图。

具体实施方式

为使本申请的目的、技术方案和优点更加清楚，下面将结合本申请具体实施例及相应的附图对本申请技术方案进行清楚、完整地描述。显然，所描述的实施例仅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。

以下结合附图，详细说明本申请实施例提供的技术方案。

为了解决了现有技术中利用网络爬虫抓取网页内容时，人工编写抓取脚本效率较低的问题，本申请实施例提供一种脚本生成方法。

该方法的执行主体，可以但不限于为手机、平板电脑或个人电脑（Personal Computer，PC）等用户终端，或者该些用户终端上运行的应用（Application，APP），或者，还可以是服务器等设备。

为便于描述，下文以该方法的执行主体为 PC 为例，对该方法的实施方式进行介绍。可以理解，该方法的执行主体为 PC 只是一种示例性的说明，并不应理解为对该方法的限定。

该方法的具体流程示意图如图 2a 所示，包括下述步骤:

步骤 11，显示网页。

在本申请实施例中，当用户想要利用网络爬虫抓取网页中的网页内容时，可以通过 PC 中安装的浏览器或者其他具有浏览器功能的应用访问该网址，以使得该 PC 可以显示出该网页，便于后续操作。后续以浏览器为例进行说明。

具体的，用户可以在浏览器中的网址输入框中输入网址，并访问该网址，该 PC 便显示出该网址对应的网页。

步骤 12，确定用户在显示的网页中选定的网页内容。

在 PC 显示出网页后，用户可以根据实际需求，在该网页中选定想要抓取的网页内容，以使得浏览器可以确定用户在网页中选定的网页内容，进而进行后续操作，最终生成抓取脚本。

用户之所以可以在网页中选定网页内容，是因为网页的网页代码中存在第一脚本。该第一脚本用于提供在网页中选定网页内容的功能。第一脚本包括层叠样式表（Cascading Style Sheets，CSS）脚本。一般情况下，第一脚本位于网页的网页代码中的顶部或底部。这是因为若将第一脚本嵌入到网页的网页代码的中间位置，在浏览器执行后续操作时，有可能会误将该第一脚本作为网页的网页代码的一部分，进而影响最终抓取脚本的生成。因此，一般将第一脚本嵌

入在网页代码的顶部或底部。

在实际应用中，网页的网页代码中存在第一脚本，有可能是因为用户在通过浏览器访问网址后，服务器返回的网页的网页代码中已经存在第一脚本了，也有可能是因为在确定用户在所述网页中选定的网页内容之前，浏览器将预设的第一脚本嵌入到了服务器返回的网页代码中。

若在通过浏览器访问网址后，服务器返回的网页的网页代码中已经存在第一脚本了，这有可能是服务器在接收到浏览器发送的获取网页的网页代码请求后，便将预设的第一脚本嵌入了网页的网页代码中，然后再将该网页代码发送给浏览器；也有可能是浏览器的开发人员事先与网页的开发人员协商好，网页的开发人员在编写该网页的网页代码时，便将第一脚本嵌入了该网页代码中。于是，服务器返回的网页代码中存在第一脚本。其中，第一脚本并不会影响浏览器对网页的渲染。

步骤 13，根据确定的网页内容，确定所述网页内容对应的网页代码。

在执行完毕步骤 12 后，浏览器可以根据通过执行步骤 12 确定的网页内容，确定所述网页内容对应的网页代码。

这是因为在浏览器根据通过执行步骤 12 确定的网页内容，确定所述网页内容对应的网页代码之前，浏览器便将预设的第二脚本嵌入到了网页的网页代码中，以使得该浏览器可以根据目标内容，通过第二脚本，确定目标内容对应的网页代码。其中，第二脚本包括 Javascript（JS）脚本。

或者，用户在浏览器中访问网页对应的网址后，服务器返回的网页代码中便存在第二脚本。这样，浏览器便可以根据网页内容，通过第二脚本，确定该网页内容对应的网页代码。其中，服务器返回的网页代码中存在第二脚本的原因可能是服务器在接收到浏览器发送的获取网页的网页代码请求后，便将预设的第二脚本嵌入了网页的网页代码中，然后再将该网页代码发送给浏览器；也有可能是浏览器的开发人员事先与网页的开发人员协商好，网页的开发人员在编写该网页的网页代码时，便将第二脚本嵌入了该网页代码中。于是，服务器

返回的网页代码中存在第二脚本。其中，第二脚本并不会影响浏览器对网页的渲染。

那么，浏览器根据通过执行步骤 12 确定的网页内容，通过第二脚本，确定该网页内容对应的网页代码的具体实施方式可以如下所述:

浏览器根据通过执行步骤 12 确定的网页内容，通过第二脚本，在网页的网页代码中，确定出该网页内容对应的最小 dom 树，然后在该最小的 dom 树中，确定出该网页内容对应的 HTML 属性值。

由于浏览器在根据网页代码渲染网页时，可以建立网页内容与网页代码的映射关系，或者建立网页中的网页内容所在的坐标点与网页代码之间的映射关系。因此，浏览器可以根据上述映射关系，根据通过执行步骤 12 确定的网页内容，确定出该网页内容对应的网页代码，即该网页内容对应的最小 dom 树。若通过执行步骤 12 确定的网页内容对应着一个最小的 dom 树，那在该最小的 dom 树中，确定出通过执行步骤 12 确定的网页内容对应的 HTML 属性值。其中，HTML 属性值可以为 class，也可以为 id 和 class。具体的，在浏览器确定 HTML 属性值之前，浏览器可以在该浏览器所在 PC 的屏幕中显示出一个询问页面，用以询问用户确定最小的 dom 树中的 id 和 class，还是确定最小的 dom 树中的 class。例如，如图 2b 所示的页面，该页面中包括确定 id 和 class 的控件，以及确定 class 的控件。若用户点击确定 id 和 class 的控件，则浏览器确定 id 和 class，若用户点击确定 class 的控件，则浏览器确定 class。

若通过执行步骤 12 确定的网页内容对应着至少两个最小的 dom 树，那在该些最小的 dom 树中，分别确定出通过执行步骤 12 确定的网页内容对应的 HTML 属性值。在确定 HTML 属性值之前，浏览器可以在该浏览器所在的 PC 的屏幕中显示出如图 2b 所示的询问页面，用以询问用户确定每一个最小的 dom 树中的 id 和 class，还是确定每一个最小的 dom 树中的 class。用户便可根据实际需要，确定点击哪一个控件，从而使得浏览器根据用户的选择，确定相应的 HTML 属性值。

在确定出上述 HTML 属性值之后，便可执行步骤 14，最终生成抓取脚本。

步骤 14，根据所述网页代码，生成抓取脚本。

在执行完毕步骤 13 后，浏览器便可将确定出的想要抓取的网页内容对应的网页代码中的 HTML 属性值添加到预设的脚本生成模板中，生成抓取脚本。其中，该抓取脚本用于抓取与 HTML 属性值相匹配的网页内容。

其中，若浏览器确定出的 HTML 属性值为 id 和 class，那么浏览器便可以分别将确定出的每一个 dom 中的 id 和 class 以{id=XXX，class=XXX}这种组合形式添加到预设的脚本生成模板中，生成抓取脚本。若浏览器确定出的 HTML 属性值为 class，那么浏览器便可以分别将确定出的每一个 dom 中的 class 以{ class=XXX}这种组合形式添加到预设的脚本生成模板中，生成抓取脚本。

其中，浏览器在生成抓取脚本后，可以将该抓取脚本保存在本地。另外，浏览器还可以将与通过执行步骤 12 确定出的网页内容对应的最小的 dom 树以及全部网页代码保存在本地，以使得浏览器可以在后续操作中可以运用上述抓取脚本、最小的 dom 树以及网页代码，抓取网页内容。

在生成抓取脚本后，浏览器可以在该浏览器所在的 PC 中弹出一个页面，用以告知用户抓取脚本已经生成，并询问用户是否抓取网页内容。

例如，该页面可以如图 2c 所示。该页面中包括第一抓取控件和第二抓取控件。若用户点击第一抓取控件，浏览器便将抓取脚本以及通过执行步骤 12 确定出的网页内容对应的最小的 dom 树，发送给解析引擎。若抓取脚本中包含 id 和 class，解析引擎根据 id 找到该 id 所在的最小的 dom 树，再根据与该 id 存在于同一个组合中的 class，提取出最小的 dom 树中的用户想要抓取的网页内容。例如，若确定出的用户想要抓取如图 1 所示的网页中的“价钱：$149.99 对应的最小的 dom 树为：

<div id="kfs_family_16" class="kfs-inner-container kfs-selected" style="width:20%;left:40%;background-image:url(https://images-na.ssl-images-amazon.com/images/G/01/kindle/stripe/kfs-selector-2._CB386844303_.gif);"
onClick="javascript:(function(){})()">

<imgclass="kfs-img"style="margin-top:9px;"src="https://images-na.ssl-images-amazon.com/images/G/01/kindle/dp/2015/848470/famnav/fs-m._CB292709393_.png" />

Fire HD 8

$149.99

<divid="kfs_popover_content_16"=class="kfs-popover-container"style="display:none;">Incredibly thin and light, designed for entertainment

抓取“价钱: $149.99”的抓取脚本中包含 id="kfs_family_16"，以及“价钱: $149.99”对应的 class="kfs-price"。

那么，用户点击第一抓取控件，浏览器便可将上述最小的 dom 树以及抓取脚本一并发送给解析引擎，解析引擎便可根据 id="kfs_family_16"找到该 id 所在的最小的 dom 树，再根据 class="kfs-price"从该 dom 树中提取出“价钱: $149.99”这一价格信息。

若抓取脚本中不包含 id，只包含 class，浏览器根据 class 便可提取出发送到解析引擎中的所有的最小的 dom 树中的与 class 相匹配的网页内容。

当用户点击第二抓取控件后，浏览器便将抓取脚本与网页的网页代码发送给解析引擎。若抓取脚本中包含 id 和 class，解析引擎根据 id 找到该 id 所在的最小的 dom 树，再根据与该 id 存在于同一个组合中的 class，提取出最小的 dom 树中的用户想要抓取的网页内容。

若抓取脚本中不包含 id，只包含 class，浏览器根据 class 便可提取出网页的网页代码中的与 class 相匹配的网页内容。

需要说明的是，本申请实施例所提供方法的各步骤的执行主体可以相同，也可以不同。例如，在浏览器执行完毕步骤 13 后，浏览器可以将确定出的网页代码以及 HTML 属性值发送给服务器，以使得服务器根据所述网页代码，生成抓取脚本。另外，上述只是以 PC 中安装的浏览器为例的一种示例性说明，本申请实施例的执行主体除了可以是 PC 中安装的浏览器之外，也可以是 PC 中安装的其他具有浏览器功能的应用，还可以是移动终端中的具有浏览器功能的 APP，本申请对此不进行任何限定。

在本申请实施例中，浏览器在执行步骤 12 时，具体的，在一种实施方式中，用户一旦开始在网页中进行框选，浏览器便可以开始确认用户框选的网页内容。或者，在用户框选结束后，被框选的目标内容被矩形方框框住，且该网页中会显示出继续框选、提交和取消的控件。其中，被矩形方框框住的目标内容，可以以高亮的显示方式显示，也可以以该网页被显示出来时该目标内容的最初显示方式显示，这可以根据用户需求进行设置，本申请实施例对此不进行任何限定。例如，如图 2d 所示，图 2d 中显示的页面便为用户框选想要抓取的某一购物网站中的某一商品的价格信息后所显示出的页面。该页面中的价格信息被一个矩形方框框住，且在价格信息的右侧分别显示出继续框选、提交和取消的控件。其中，被框住的价格信息并未高亮显示，而是以最初显示状态显示。

在网页中出现继续框选、提交和取消的控件后，若用户想要继续框选其他的内容，便可点击继续框选控件，继续框选其他内容。若用户不再想框选其他内容，便可点击提交控件，浏览器便将用户框选的网页内容确定为目标内容。若用户想取消之前框选的内容，框选其他的内容，那么点击取消控件，便可重新框选。

另外，本申请实施例提供的脚本生成方法中，用户在网页中选择想要抓取的网页内容时，可以先进行一次粗略选择，浏览器可以根据用户第一次选择的网页内容确定出该网页内容对应的最小的 dom。然后，用户再在第一次选择的网页内容的基础上进行第二次选择，浏览器便可以在确定出的最小的 dom 中确

定出第二次选择的网页内容对应的网页代码中的 HTML 属性值。比如，如图 2e 所示，若用户想要抓取网页中的“￥175”的价格信息，在第一次选定时，用户可以粗略的将包含“￥175”这一价格信息的网页内容一并进行选择，浏览器便可根据该第一次选择的内容，确定出该些网页内容对应的最小的 dom 树。在用户在第二次选择时，可以只选择“￥175”，那么浏览器便可以在该最小的 dom 树确定出“￥175”对应的 HTML 属性值。图 2e 中有两个矩形方框，其中，较大的矩形方框中的所有内容为用户第一次选择的内容，较小的矩形方框中的网页内容为用户第二次选择的内容。

在本申请实施例中，还可通过一种脚本生成装置，来实现本申请实施例中提供的脚本生成方法。

如图 3 所示，为本申请实施例提供的一种脚本生成装置的结构示意图，主要包括下述装置:

内容确定模块 31，确定用户在显示的网页中选定的网页内容。

代码确定模块 32，根据确定的网页内容，确定所述网页内容对应的网页代码。

脚本生成模块 33，根据所述网页代码，生成抓取脚本。

在一种实施方式中，所述装置还包括:

第一嵌入模块，在内容确定模块 31 确定用户在显示的网页中选定的网页内容之前，将预设的第一脚本嵌入到所述网页的网页代码中。其中，所述第一脚本用于提供在所述网页中选定网页内容的功能，所述第一脚本包括层叠样式表 CSS 脚本。

在一种实施方式中，所述装置还包括:

第二嵌入模块，在代码确定模块 32 根据确定的网页内容，确定所述网页内容对应的网页代码之前，将预设的第二脚本嵌入到所述网页的网页代码中，所述第二脚本包括 JS 脚本。

则代码确定模块 32，根据确定的网页内容，通过所述第二脚本，确定所述

网页内容对应的网页代码。

在一种实施方式中，代码确定模块 32，在所述网页的网页代码中，确定所述网页内容对应的最小的文档对象模型 dom 树;

在所述最小的 dom 树中，确定所述网页内容对应的超文本标记语言 HTML 属性值。

在一种实施方式中，脚本生成模块 33，将确定出的所述 HTML 属性值添加到预设的脚本生成模板中，生成抓取脚本，所述抓取脚本用于抓取与所述 HTML 属性值相匹配的网页内容。

在一种实施方式中，所述装置还包括:

内容解析模块，将所述抓取脚本以及网页代码发送给解析引擎，通过解析引擎，抓取相应的网页内容。

与现有技术中利用网络爬虫抓取网页内容时，需人工编写抓取脚本相比，采用本申请实施例提供的脚本生成方法，通过确定用户在网页中选定的网页内容，确定出该网页内容对应的网页代码，并根据该网页代码生成抓取脚本，从而解决了现有技术中利用网络爬虫抓取网页内容时，人工编写抓取脚本效率较低的问题。

本领域内的技术人员应明白，本发明的实施例可提供为方法、系统、或计算机程序产品。因此，本发明可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且，本发明可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质（包括但不限于磁盘存储器、CD-ROM、光学存储器等）上实施的计算机程序产品的形式。

本发明是参照根据本发明实施例的方法、设备（系统）、和计算机程序产品的流程图和／或方框图来描述的。应理解可由计算机程序指令实现流程图和／或方框图中的每一流程和／或方框、以及流程图和／或方框图中的流程和／或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产生一个机器，使得通过计算

机或其他可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理设备以特定方式工作的计算机可读存储器中，使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品，该指令装置实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能。

这些计算机程序指令也可装载到计算机或其他可编程数据处理设备上，使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理，从而在计算机或其他可编程设备上执行的指令提供用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的步骤。

在一个典型的配置中，计算设备包括一个或多个处理器 (CPU)、输入/输出接口、网络接口和内存。

内存可能包括计算机可读介质中的非永久性存储器，随机存取存储器 (RAM) 和/或非易失性内存等形式，如只读存储器 (ROM) 或闪存(flash RAM)。内存是计算机可读介质的示例。

计算机可读介质包括永久性和非永久性、可移动和非可移动媒体可以由任何方法或技术来实现信息存储。信息可以是计算机可读指令、数据结构、程序的模块或其他数据。计算机的存储介质的例子包括，但不限于相变内存 (PRAM)、静态随机存取存储器 (SRAM)、动态随机存取存储器 (DRAM)、其他类型的随机存取存储器 (RAM)、只读存储器 (ROM)、电可擦除可编程只读存储器 (EEPROM)、快闪记忆体或其他内存技术、只读光盘只读存储器 (CD-ROM)、数字多功能光盘 (DVD) 或其他光学存储、磁盒式磁带，磁带磁磁盘存储或其他磁性存储设备或任何其他非传输介质，可用于存储可以被计算设备访问的信息。按照本文中的界定，计算机可读介质不包括暂存电脑可读媒体 (transitory media)，如调制的数据信号和载波。

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过程、方法、商品或者设备不仅包括那些要素，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过程、方法、商品或者设备所固有的要素。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过程、方法、商品或者设备中还存在另外的相同要素。

本领域技术人员应明白，本申请的实施例可提供为方法、系统或计算机程序产品。因此，本申请可采用完全硬件实施例、完全软件实施例或结合软件和硬件方面的实施例的形式。而且，本申请可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质（包括但不限于磁盘存储器、CD-ROM、光学存储器等）上实施的计算机程序产品的形式。

以上所述仅为本申请的实施例而已，并不用于限制本申请。对于本领域技术人员来说，本申请可以有各种更改和变化。凡在本申请的精神和原理之内所作的任何修改、等同替换、改进等，均应包含在本申请的权利要求范围之内。

权 利 要 求 书

1、一种脚本生成方法，其特征在于，所述方法包括:

确定用户在显示的网页中选定的网页内容;

根据确定的网页内容，确定所述网页内容对应的网页代码;

根据所述网页代码，生成抓取脚本。

2、如权利要求 1 所述的方法，其特征在于，确定用户在显示的网页中选定的网页内容之前，所述方法还包括:

将预设的第一脚本嵌入到所述网页的网页代码中；其中，所述第一脚本用于提供在所述网页中选定网页内容的功能，所述第一脚本包括层叠样式表 CSS 脚本。

3、如权利要求 1 所述的方法，其特征在于，根据确定的网页内容，确定所述网页内容对应的网页代码之前，所述方法还包括:

将预设的第二脚本嵌入到所述网页的网页代码中，所述第二脚本包括 JS 脚本;

根据确定的网页内容，确定所述网页内容对应的网页代码，具体包括:

根据确定的网页内容，通过所述第二脚本，确定所述网页内容对应的网页代码。

4、如权利要求 1 所述的方法，其特征在于，确定所述网页内容对应的网页代码，具体包括:

在所述网页的网页代码中，确定所述网页内容对应的最小的文档对象模型 dom 树;

在所述最小的 dom 树中，确定所述网页内容对应的超文本标记语言 HTML 属性值。

5、如权利要求 4 所述的方法，其特征在于，根据所述网页代码，生成脚

本，具体包括:

将确定出的所述 HTML 属性值添加到预设的脚本生成模板中，生成抓取脚本，所述抓取脚本用于抓取与所述 HTML 属性值相匹配的网页内容。

6、如权利要求 5 所述的方法，其特征在于，所述方法还包括:

将所述抓取脚本以及网页代码发送给解析引擎，通过解析引擎，抓取相应的网页内容。

7、一种脚本生成装置，其特征在于，所述装置包括:

内容确定模块，确定用户在显示的网页中选定的网页内容;

代码确定模块，根据确定的网页内容，确定所述网页内容对应的网页代码;

脚本生成模块，根据所述网页代码，生成抓取脚本。

8、如权利要求 7 所述的装置，其特征在于，所述装置还包括:

第一嵌入模块，在内容确定模块确定用户在显示的网页中选定的网页内容之前，将预设的第一脚本嵌入到所述网页的网页代码中；其中，所述第一脚本用于提供在所述网页中选定网页内容的功能，所述第一脚本包括层叠样式表 CSS 脚本。

9、如权利要求 7 所述的装置，其特征在于，所述装置还包括:

第二嵌入模块，在代码确定模块根据确定的网页内容，确定所述网页内容对应的网页代码之前，将预设的第二脚本嵌入到所述网页的网页代码中，所述第二脚本包括 JS 脚本;

则代码确定模块，根据确定的网页内容，通过所述第二脚本，确定所述网页内容对应的网页代码。

10、如权利要求 7 所述的装置，其特征在于，代码确定模块，在所述网页的网页代码中，确定所述网页内容对应的最小的文档对象模型 dom 树;

在所述最小的 dom 树中，确定所述网页内容对应的超文本标记语言 HTML 属性值。

11、如权利要求 10 所述的装置，其特征在于，脚本生成模块，将确定出

的所述 HTML 属性值添加到预设的脚本生成模板中，生成抓取脚本，所述抓取脚本用于抓取与所述 HTML 属性值相匹配的网页内容。

12、如权利要求 11 所述的装置，其特征在于，所述装置还包括:

内容解析模块，将所述抓取脚本以及网页代码发送给解析引擎，通过解析引擎，抓取相应的网页内容。


fire HD8
INCREDIBLY THIN AND LIGHT
DESIGNED FOR ENTERTAINMENT
消防HD 8平板，8"高清显示器，无线网络，8 GB - 包括特别优惠，黑色
由 亚马逊
7,216顾客评论 | 1000+回答问题
价钱： $ 149.99 及免运费。详情
上衣
收藏
第三方
美国运费
翻译帮助
有现货。
希望周二，6月28日？ 今内18小时12分钟并选择同日发送在付款处。
详细
船舶和亚马逊数字服务有限责任公司销售。礼品包装可用。
数字存储容量： 8 GB
翻滚图像放大

图 1

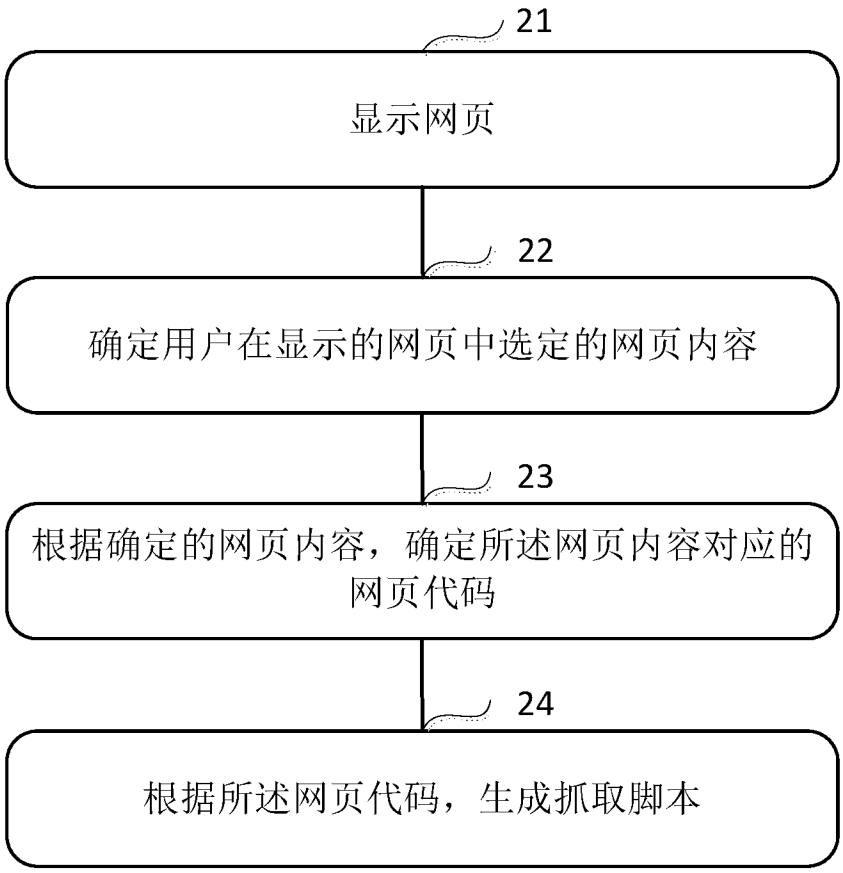

21
显示网页
22
确定用户在显示的网页中选定的网页内容
23
根据确定的网页内容，确定所述网页内容对应的网页代码
24
根据所述网页代码，生成抓取脚本

图 2a

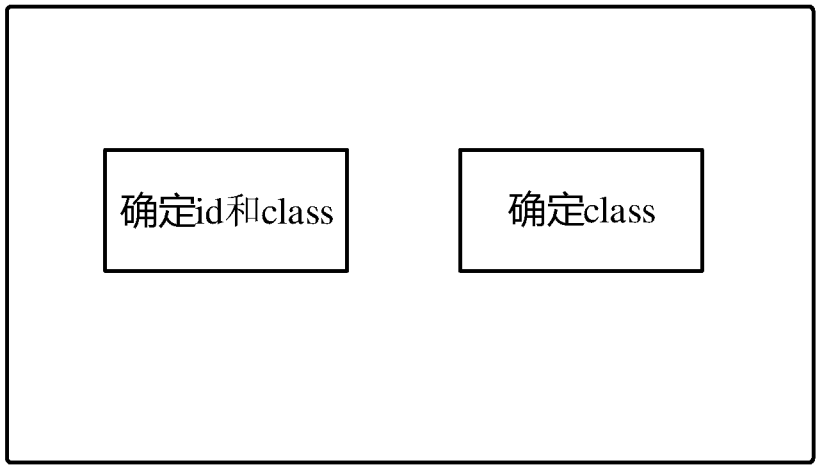


图 2b

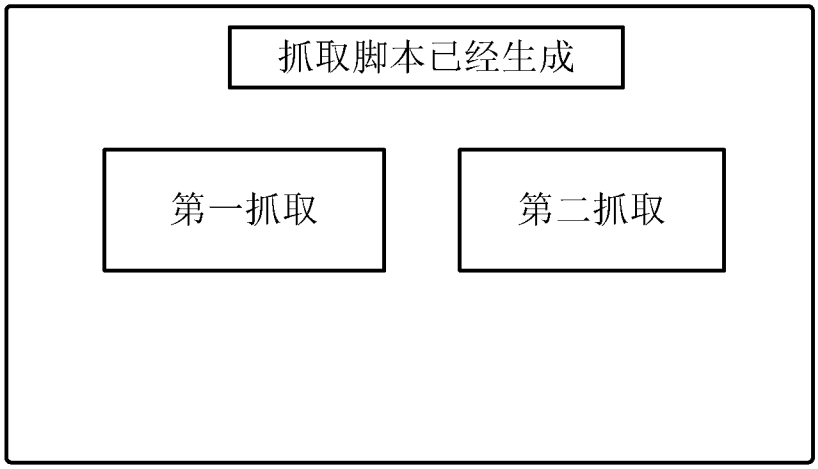


图 2c


fire HD8
INCREDIBLY THIN AND LIGHT
DESIGNED FOR ENTERTAINMENT
消防HD 8平板，8"高清显示器，无线网络，8 GB - 包括特别优惠，黑色
继续框选
$ 149.99
提交
取消
希望周二，8月28日？
船舶和亚马逊数字服务有限责任公司销售。礼品包装可用。
8 GB

图 2d

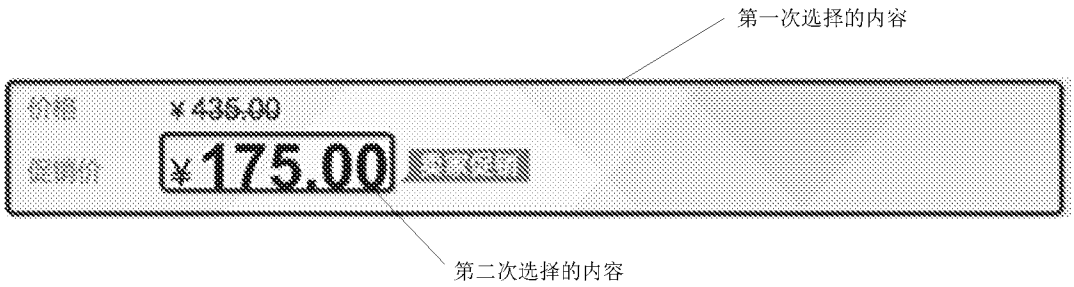

第一次选择的内容
¥435.00
¥175.00
第二次选择的内容

图 2e

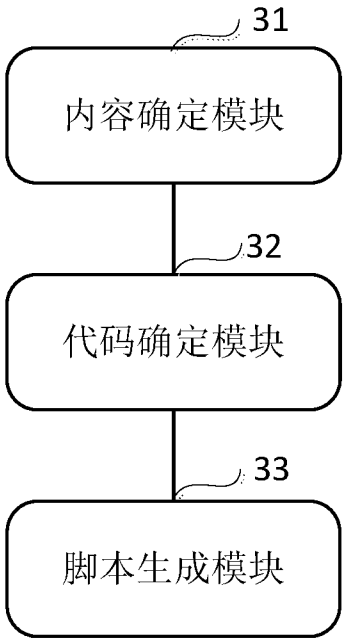


图 3

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2017/091674

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/30 (2006.01) i; G06F 9/44 (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, CNKI, WPI, EPODOC, GOOGLE: web page, acquisition, crawling, extraction, script, web, page, crawler, extract, grab, pick up, capture

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 106886547 A (ALIBABA GROUP HOLDING LIMITED), 23 June 2017 (23.06.2017), claims 1-12	1-12
X	CN 105243159 A (FUJIAN YIRONG INFORMATION TECHNOLOGY CO., LTD.), 13 January 2016 (13.01.2016), description, paragraphs [0026]-[0035], and figure 4	1, 7
Y	CN 105243159 A (FUJIAN YIRONG INFORMATION TECHNOLOGY CO., LTD.), 13 January 2016 (13.01.2016), description, paragraphs [0017], [0018] and [0026]-[0035], and figure 4	2-6, 8-12
Y	CN 101894138 A (UC MOBILE LTD.), 24 November 2010 (24.11.2010), description, paragraphs [0009]-[0017]	2-6, 8-12
A	CN 105468730 A (GUANGZHOU HUADUO INTERNET TECHNOLOGY CO., LTD.), 06 April 2016 (06.04.2016), the whole document	1-12

☐ 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
29 August 2017 (29.08.2017)	**27 September 2017 (27.09.2017)**
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 **SUN, Guohui** Telephone No.: (86-10) **61648471**

Form PCT/ISA/210 (second sheet) (July 2009)

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2017/091674

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 106886547 A	23 June 2017	None	
CN 105243159 A	13 January 2016	None	
CN 101894138 A	24 November 2010	None	
CN 105468730 A	06 April 2016	None	

国际检索报告

国际申请号
PCT/CN2017/091674

A. 主题的分类

G06F 17/30(2006.01)i；G06F 9/44(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,CNKI,WPI,EPODOC,GOOGLE: 脚本，网页，页面，爬虫，抓取，获取，爬取，提取，script，web，page，crawler，extract，grab，pick up，capture

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 106886547 A (阿里巴巴集团控股有限公司) 2017年 6月 23日 (2017-06-23) 权利要求1-12	1-12
X	CN 105243159 A (福建亿榕信息技术有限公司) 2016年 1月 13日 (2016-01-13) 说明书第[0026]-[0035]段，图4	1，7
Y	CN 105243159 A (福建亿榕信息技术有限公司) 2016年 1月 13日 (2016-01-13) 说明书第[0017]-[0018]，[0026]-[0035]段，图4	2-6，8-12
Y	CN 101894138 A (优视科技有限公司) 2010年 11月 24日 (2010-11-24) 说明书第[0009]-[0017]段	2-6，8-12
A	CN 105468730 A (广州华多网络科技有限公司) 2016年 4月 6日 (2016-04-06) 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	国际检索报告邮寄日期
2017年 8月 29日	2017年 9月 27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 孙国辉 电话号码 (86-10)61648471

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2017/091674

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	106886547	A	2017年 6月 23日	无	
CN	105243159	A	2016年 1月 13日	无	
CN	101894138	A	2010年 11月 24日	无	
CN	105468730	A	2016年 4月 6日	无	