(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 103020298 A
(43)申请公布日 2013.04.03

(21)申请号 201210591380.1

(22)申请日 2012.12.31

(71)申请人 华为技术有限公司
地址 518129 广东省深圳市龙岗区坂田华为总部办公楼

(72)发明人 苗宏

(74)专利代理机构 北京龙双利达知识产权代理有限公司 11329
代理人 毛威 张亮

(51)Int.Cl.
G06F 17/30(2006.01)

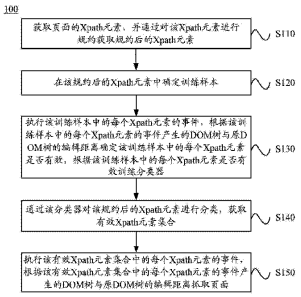
权利要求书 3 页 说明书 12 页 附图 5 页

(54)发明名称

抓取页面的方法和装置

(57)摘要

本发明公开了一种抓取页面的方法和装置。该方法包括：获取页面的Xpath元素，对Xpath元素进行规约；在规约后的Xpath元素中确定训练样本；执行训练样本中的每个Xpath元素的事件，根据训练样本中的每个Xpath元素的事件产生的DOM树与原DOM树的编辑距离确定训练样本中的每个Xpath元素是否有效，根据训练样本中的每个Xpath元素是否有效训练分类器；通过分类器对规约后的Xpath元素进行分类，获取有效Xpath元素集合；执行有效Xpath元素集合中的每个Xpath元素的事件，根据有效Xpath元素集合中的每个Xpath元素的事件产生的DOM树与原DOM树的编辑距离抓取页面。本发明实施例的抓取页面的方法和装置，能够提升抓取页面的效率。



CN 103020298 A

1. 一种抓取页面的方法，其特征在于，包括：

获取页面的可扩展标记语言路径语言 Xpath 元素，并通过对所述 Xpath 元素进行规约获取规约后的 Xpath 元素；

在所述规约后的 Xpath 元素中确定训练样本；

执行所述训练样本中的每个 Xpath 元素的事件，根据所述训练样本中的每个 Xpath 元素的事件产生的文档对象模型 DOM 树与原 DOM 树的编辑距离确定所述训练样本中的每个 Xpath 元素是否有效，根据所述训练样本中的每个 Xpath 元素是否有效训练分类器；

通过所述分类器对所述规约后的 Xpath 元素进行分类，获取有效 Xpath 元素集合；

执行所述有效 Xpath 元素集合中的每个 Xpath 元素的事件，根据所述有效 Xpath 元素集合中的每个 Xpath 元素的事件产生的 DOM 树与原 DOM 树的编辑距离抓取页面。

2. 根据权利要求 1 所述的方法，其特征在于，在所述根据所述训练样本中的每个 Xpath 元素是否有效训练分类器之前，所述方法还包括：

获取业务定制信息，根据所述业务定制信息确定定制规则；

所述根据所述训练样本中的每个 Xpath 元素是否有效训练分类器，包括：

根据所述训练样本中的每个 Xpath 元素是否有效和所述定制规则，训练所述分类器。

3. 根据权利要求 1 或 2 所述的方法，其特征在于，所述根据所述训练样本中的每个 Xpath 元素的事件产生的 DOM 树与原 DOM 树的编辑距离确定所述训练样本中的每个 Xpath 元素是否有效，包括：

若所述训练样本中的第一 Xpath 元素的事件产生的 DOM 树与原 DOM 树的编辑距离大于预定阈值，则确定所述第一 Xpath 元素有效；

若所述训练样本中的第二 Xpath 元素的事件产生的 DOM 树与原 DOM 树的编辑距离不大于所述预定阈值，则确定所述第二 Xpath 元素无效；

所述根据所述有效 Xpath 元素集合中的每个 Xpath 元素的事件产生的 DOM 树与原 DOM 树的编辑距离抓取页面，包括：

若所述有效 Xpath 元素集合中的第三 Xpath 元素的事件产生的 DOM 树与原 DOM 树的编辑距离大于所述预定阈值，则保存所述第三 Xpath 元素的事件产生的 DOM 树；

若所述有效 Xpath 元素集合中的第四 Xpath 元素的事件产生的 DOM 树与原 DOM 树的编辑距离不大于所述预定阈值，则不保存所述第四 Xpath 元素的事件产生的 DOM 树。

4. 根据权利要求 1 至 3 中任一项所述的方法，其特征在于，在所述根据所述训练样本中的每个 Xpath 元素的事件产生的 DOM 树与原 DOM 树的编辑距离确定所述训练样本中的每个 Xpath 元素是否有效之后，所述方法还包括：

保存所述训练样本中的有效 Xpath 元素的事件产生的 DOM 树；

所述通过所述分类器对所述规约后的 Xpath 元素进行分类，获取有效 Xpath 元素集合，包括：

通过所述分类器对所述规约后的 Xpath 元素中除所述训练样本之外的 Xpath 元素进行分类，获取所述有效 Xpath 元素集合。

5. 根据权利要求 1 至 4 中任一项所述的方法，其特征在于，在所述获取规约后的 Xpath 元素之后，所述方法还包括：

生成所述规约后的 Xpath 元素的状态转换图模型；

所述在所述规约后的 Xpath 元素中确定训练样本，包括：

在所述状态转换图模型中确定训练样本；

所述通过所述分类器对所述规约后的 Xpath 元素进行分类，获取有效 Xpath 元素集合，包括：

将所述状态转换图模型输入所述分类器，获取所述有效 Xpath 元素集合。

6. 根据权利要求 1 至 5 中任一项所述的方法，其特征在于，所述获取页面的可扩展标记语言路径语言 Xpath 元素，包括：

通过嵌入浏览器技术获取所述 Xpath 元素。

7. 一种抓取页面的装置，其特征在于，包括：

获取模块，用于获取页面的可扩展标记语言路径语言 Xpath 元素，并通过对所述 Xpath 元素进行规约获取规约后的 Xpath 元素；

确定模块，用于在所述规约后的 Xpath 元素中确定训练样本；

训练模块，用于执行所述训练样本中的每个 Xpath 元素的事件，根据所述训练样本中的每个 Xpath 元素的事件产生的文档对象模型 DOM 树与原 DOM 树的编辑距离确定所述训练样本中的每个 Xpath 元素是否有效，根据所述训练样本中的每个 Xpath 元素是否有效训练分类器；

分类模块，用于通过所述分类器对所述规约后的 Xpath 元素进行分类，获取有效 Xpath 元素集合；

抓取模块，用于执行所述有效 Xpath 元素集合中的每个 Xpath 元素的事件，根据所述有效 Xpath 元素集合中的每个 Xpath 元素的事件产生的 DOM 树与原 DOM 树的编辑距离抓取页面。

8. 根据权利要求 7 所述的装置，其特征在于，所述装置还包括：

定制模块，用于获取业务定制信息，根据所述业务定制信息确定定制规则；

所述训练模块具体用于根据所述训练样本中的每个 Xpath 元素是否有效和所述定制规则，训练所述分类器。

9. 根据权利要求 7 或 8 所述的装置，其特征在于，所述训练模块具体用于，若所述训练样本中的第一 Xpath 元素的事件产生的 DOM 树与原 DOM 树的编辑距离大于预定阈值，则确定所述第一 Xpath 元素有效，若所述训练样本中的第二 Xpath 元素的事件产生的 DOM 树与原 DOM 树的编辑距离不大于所述预定阈值，则确定所述第二 Xpath 元素无效；

所述抓取模块具体用于，若所述有效 Xpath 元素集合中的第三 Xpath 元素的事件产生的 DOM 树与原 DOM 树的编辑距离大于所述预定阈值，则保存所述第三 Xpath 元素的事件产生的 DOM 树，若所述有效 Xpath 元素集合中的第四 Xpath 元素的事件产生的 DOM 树与原 DOM 树的编辑距离不大于所述预定阈值，则不保存所述第四 Xpath 元素的事件产生的 DOM 树。

10. 根据权利要求 7 至 9 中任一项所述的装置，其特征在于，所述抓取模块还用于在所述训练模块根据所述训练样本中的每个 Xpath 元素的事件产生的 DOM 树与原 DOM 树的编辑距离确定所述训练样本中的每个 Xpath 元素是否有效之后，保存所述训练样本中的有效 Xpath 元素的事件产生的 DOM 树；

所述分类模块具体用于通过所述分类器对所述规约后的 Xpath 元素中除所述训练样本之外的 Xpath 元素进行分类，获取所述有效 Xpath 元素集合。

11. 根据权利要求 7 至 10 中任一项所述的装置，其特征在于，所述装置还包括：

生成模块，用于在所述获取模块获取规约后的 Xpath 元素之后，生成所述规约后的 Xpath 元素的状态转换图模型；

所述确定模块具体用于在所述状态转换图模型中确定训练样本；

所述分类模块具体用于将所述状态转换图模型输入所述分类器，获取所述有效 Xpath 元素集合。

12. 根据权利要求 7 至 11 中任一项所述的装置，其特征在于，所述获取模块具体用于通过嵌入浏览器技术获取所述 Xpath 元素。

抓取页面的方法和装置

技术领域

[0001] 本发明涉及信息技术领域，并且更具体地，涉及抓取页面的方法和装置。

背景技术

[0002] 传统的网络爬虫技术，即抓取页面的技术，主要应用于抓取静态 Web 网页，随着异步的 JavaScript 与可扩展标记语言(Extensible Markup Language，简称为“XML”)技术(Asynchronous JavaScript and XML，简称为“Ajax”)/Web2.0 的流行，如何抓取 Ajax 等动态页面成了搜索引擎急需解决的问题。Ajax 采用了 JavaScript 驱动的异步请求 / 响应机制，以往的爬虫们缺乏 JavaScript 语义上的理解，基本上无法模拟触发 JavaScript 的异步调用并解析返回的异步回调逻辑和内容。另外，在 Ajax 的应用中，JavaScript 会对文档对象模型(Document Object Model，简称为“DOM”)结构进行大量变动，甚至页面所有内容都通过 JavaScript 直接从服务器端读取并动态绘制出来。这对习惯了 DOM 结构相对不变的静态页面简直是无法理解的。由此可以看出，以往的爬虫是基于协议驱动的，而对于 Ajax 这样的技术，所需要的爬虫引擎必须是基于事件驱动的。

[0003] 现有技术采用页面 Javascript 代码解析和页面 DOM 状态判重来实现，由于在现在的 web2.0 网站中大量采用了 Ajax 技术，其中绝大部分的 Javascript 代码执行后不能改变 DOM 树的结构，因此会导致无效 Javascript 代码的频繁执行，大量 DOM 树结构的比较运算，影响页面抓取的效率。

发明内容

[0004] 本发明实施例提供了一种抓取页面的方法和装置，能够提升抓取页面的效率。

[0005] 第一方面，提供了一种抓取页面的方法，包括：获取页面的可扩展标记语言路径语言(XML Path Language，简称为“Xpath”)元素，并通过对该 Xpath 元素进行规约获取规约后的 Xpath 元素；在该规约后的 Xpath 元素中确定训练样本；执行该训练样本中的每个 Xpath 元素的事件，根据该训练样本中的每个 Xpath 元素的事件产生的 DOM 树与原 DOM 树的编辑距离确定该训练样本中的每个 Xpath 元素是否有效，根据该训练样本中的每个 Xpath 元素是否有效训练分类器；通过该分类器对该规约后的 Xpath 元素进行分类，获取有效 Xpath 元素集合；执行该有效 Xpath 元素集合中的每个 Xpath 元素的事件，根据该有效 Xpath 元素集合中的每个 Xpath 元素的事件产生的 DOM 树与原 DOM 树的编辑距离抓取页面。

[0006] 在第一种可能的实现方式中，在根据该训练样本中的每个 Xpath 元素是否有效训练分类器之前，该方法还包括：获取业务定制信息，根据该业务定制信息确定定制规则；该根据该训练样本中的每个 Xpath 元素是否有效训练分类器，包括：根据该训练样本中的每个 Xpath 元素是否有效和该定制规则，训练该分类器。

[0007] 在第二种可能的实现方式中，结合第一方面或第一方面的第一种可能的实现方式，根据该训练样本中的每个 Xpath 元素的事件产生的 DOM 树与原 DOM 树的编辑距离确定该训练样本中的每个 Xpath 元素是否有效，包括：若该训练样本中的第一 Xpath 元素的事件

产生的 DOM 树与原 DOM 树的编辑距离大于预定阈值，则确定该第一 Xpath 元素有效；若该训练样本中的第二 Xpath 元素的事件产生的 DOM 树与原 DOM 树的编辑距离不大于该预定阈值，则确定该第二 Xpath 元素无效；根据该有效 Xpath 元素集合中的每个 Xpath 元素的事件产生的 DOM 树与原 DOM 树的编辑距离抓取页面，包括：若该有效 Xpath 元素集合中的第三 Xpath 元素的事件产生的 DOM 树与原 DOM 树的编辑距离大于该预定阈值，则保存该第三 Xpath 元素的事件产生的 DOM 树；若该有效 Xpath 元素集合中的第四 Xpath 元素的事件产生的 DOM 树与原 DOM 树的编辑距离不大于该预定阈值，则不保存该第四 Xpath 元素的事件产生的 DOM 树。

[0008] 在第三种可能的实现方式中，结合第一方面或第一方面的第一种或第二种可能的实现方式，在根据该训练样本中的每个 Xpath 元素的事件产生的 DOM 树与原 DOM 树的编辑距离确定该训练样本中的每个 Xpath 元素是否有效之后，该方法还包括：保存该训练样本中的有效 Xpath 元素的事件产生的 DOM 树；通过该分类器对该规约后的 Xpath 元素进行分类，获取有效 Xpath 元素集合，包括：通过该分类器对该规约后的 Xpath 元素中除该训练样本之外的 Xpath 元素进行分类，获取该有效 Xpath 元素集合。

[0009] 在第四种可能的实现方式中，结合第一方面或第一方面的第一至三种可能的实现方式中的任一种可能的实现方式，在获取规约后的 Xpath 元素之后，该方法还包括：生成该规约后的 Xpath 元素的状态转换图模型；在该规约后的 Xpath 元素中确定训练样本，包括：在该状态转换图模型中确定训练样本；通过该分类器对该规约后的 Xpath 元素进行分类，获取有效 Xpath 元素集合，包括：将该状态转换图模型输入该分类器，获取该有效 Xpath 元素集合。

[0010] 在第五种可能的实现方式中，结合第一方面或第一方面的第一至四种可能的实现方式中的任一种可能的实现方式，获取页面的 Xpath 元素，包括：通过嵌入浏览器技术获取该 Xpath 元素。

[0011] 第二方面，提供了一种抓取页面的装置，包括：获取模块，用于获取页面的 Xpath 元素，并通过对该 Xpath 元素进行规约获取规约后的 Xpath 元素；确定模块，用于在该规约后的 Xpath 元素中确定训练样本；训练模块，用于执行该训练样本中的每个 Xpath 元素的事件，根据该训练样本中的每个 Xpath 元素的事件产生的 DOM 树与原 DOM 树的编辑距离确定该训练样本中的每个 Xpath 元素是否有效，根据该训练样本中的每个 Xpath 元素是否有效训练分类器；分类模块，用于通过该分类器对该规约后的 Xpath 元素进行分类，获取有效 Xpath 元素集合；抓取模块，用于执行该有效 Xpath 元素集合中的每个 Xpath 元素的事件，根据该有效 Xpath 元素集合中的每个 Xpath 元素的事件产生的 DOM 树与原 DOM 树的编辑距离抓取页面。

[0012] 在第一种可能的实现方式中，该装置还包括：定制模块，用于获取业务定制信息，根据该业务定制信息确定定制规则；该训练模块具体用于根据该训练样本中的每个 Xpath 元素是否有效和该定制规则，训练该分类器。

[0013] 在第二种可能的实现方式中，结合第二方面或第二方面的第一种可能的实现方式，该训练模块具体用于，若该训练样本中的第一 Xpath 元素的事件产生的 DOM 树与原 DOM 树的编辑距离大于预定阈值，则确定该第一 Xpath 元素有效，若该训练样本中的第二 Xpath 元素的事件产生的 DOM 树与原 DOM 树的编辑距离不大于该预定阈值，则确定该第二 Xpath

元素无效；该抓取模块具体用于，若该有效Xpath元素集合中的第三Xpath元素的事件产生的DOM树与原DOM树的编辑距离大于该预定阈值，则保存该第三Xpath元素的事件产生的DOM树，若该有效Xpath元素集合中的第四Xpath元素的事件产生的DOM树与原DOM树的编辑距离不大于该预定阈值，则不保存该第四Xpath元素的事件产生的DOM树。

[0014] 在第三种可能的实现方式中，结合第二方面或第二方面的第一种或第二种可能的实现方式，该抓取模块还用于在该训练模块根据该训练样本中的每个Xpath元素的事件产生的DOM树与原DOM树的编辑距离确定该训练样本中的每个Xpath元素是否有效之后，保存该训练样本中的有效Xpath元素的事件产生的DOM树；该分类模块具体用于通过该分类器对该规约后的Xpath元素中除该训练样本之外的Xpath元素进行分类，获取该有效Xpath元素集合。

[0015] 在第四种可能的实现方式中，结合第二方面或第二方面的第一至三种可能的实现方式中的任一种可能的实现方式，该装置还包括：生成模块，用于在该获取模块获取规约后的Xpath元素之后，生成该规约后的Xpath元素的状态转换图模型；该确定模块具体用于在该状态转换图模型中确定训练样本；该分类模块具体用于将该状态转换图模型输入该分类器，获取该有效Xpath元素集合。

[0016] 在第五种可能的实现方式中，结合第二方面或第二方面的第一至四种可能的实现方式中的任一种可能的实现方式，该获取模块具体用于通过嵌入浏览器技术获取该Xpath元素。

[0017] 基于上述技术方案，本发明实施例的抓取页面的方法和装置，根据训练样本中的Xpath元素是否有效训练分类器，通过分类器对Xpath元素进行分类，获取有效Xpath元素集合，再基于有效Xpath元素集合抓取页面，可以过滤掉大量的无效Xpath元素，从而能够提升抓取页面的效率。

附图说明

[0018] 为了更清楚地说明本发明实施例的技术方案，下面将对本发明实施例中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，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。

[0019] 图1是根据本发明实施例的抓取页面的方法的示意性流程图。

[0020] 图2是根据本发明实施例的状态转换图模型的示意图。

[0021] 图3是根据本发明实施例的抓取页面的方法的另一示意性流程图。

[0022] 图4是根据本发明实施例的抓取页面的方法的又一示意性流程图。

[0023] 图5是根据本发明实施例的抓取页面的装置的示意性框图。

[0024] 图6是根据本发明实施例的抓取页面的装置的另一示意性框图。

[0025] 图7是根据本发明实施例的抓取页面的装置的结构示意图。

具体实施方式

[0026] 下面将结合本发明实施例中的附图，对本发明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发明的一部分实施例，而不是全部实施例。基于本发

明中的实施例，本领域普通技术人员在没有作出创造性劳动的前提下所获得的所有其他实施例，都应属于本发明保护的范围。

[0027] 图 1 示出了根据本发明实施例的抓取页面的方法 100 的示意性流程图。如图 1 所示，该方法 100 包括：

[0028] S110，获取页面的 Xpath 元素，并通过对该 Xpath 元素进行规约获取规约后的 Xpath 元素；

[0029] S120，在该规约后的 Xpath 元素中确定训练样本；

[0030] S130，执行该训练样本中的每个 Xpath 元素的事件，根据该训练样本中的每个 Xpath 元素的事件产生的 DOM 树与原 DOM 树的编辑距离确定该训练样本中的每个 Xpath 元素是否有效，根据该训练样本中的每个 Xpath 元素是否有效训练分类器；

[0031] S140，通过该分类器对该规约后的 Xpath 元素进行分类，获取有效 Xpath 元素集合；

[0032] S150，执行该有效 Xpath 元素集合中的每个 Xpath 元素的事件，根据该有效 Xpath 元素集合中的每个 Xpath 元素的事件产生的 DOM 树与原 DOM 树的编辑距离抓取页面。

[0033] 现有抓取页面的技术需要执行所有 Xpath 元素的事件，并进行 DOM 树结构的比较，效率较低。在本发明实施例中，抓取页面的装置在获取所有 Xpath 元素并对 Xpath 元素进行规约后，在规约后的 Xpath 元素中抽取部分 Xpath 元素作为训练样本，执行训练样本中的每个 Xpath 元素的事件，根据训练样本中的每个 Xpath 元素的事件产生的 DOM 树与原 DOM 树的编辑距离确定训练样本中的每个 Xpath 元素是否有效，并根据训练样本中的每个 Xpath 元素是否有效训练分类器，然后，通过分类器对规约后的 Xpath 元素进行分类，获取有效 Xpath 元素集合，再执行有效 Xpath 元素集合中的每个 Xpath 元素的事件，根据有效 Xpath 元素集合中的每个 Xpath 元素的事件产生的 DOM 树与原 DOM 树的编辑距离抓取页面。由于利用分类器过滤掉了无效 Xpath 元素，在抓取页面时只需执行有效 Xpath 元素的事件，不再频繁执行无效 Javascript 代码。

[0034] 因此，本发明实施例的抓取页面的方法，根据训练样本中的 Xpath 元素是否有效训练分类器，通过分类器对 Xpath 元素进行分类，获取有效 Xpath 元素集合，再基于有效 Xpath 元素集合抓取页面，可以过滤掉大量的无效 Xpath 元素，从而能够提升抓取页面的效率。

[0035] 本发明实施例的技术方案可以用于抓取动态页面，例如，在 web2.0 网站中抓取页面。

[0036] 传统网站页面由唯一的统一资源定位符(Uniform/Universal ResourceLocator，简称为“URL”)确定，网站本身可以看作是一个以页面为顶点，超链接为边的有向图。该经典模型是传统爬虫对 web 资源的基本假设。应用 Ajax 技术的网站(例如，web2.0 网站)既包含静态内容也包含动态内容，页面本身不再是一个基本单位，它通常是由若干个不同的状态所构成，用户浏览行为通过 Javascript 事件处理函数改变 DOM 树的内容与结构，由此产生新的状态，这些状态都同属于一个 URL；此外，页面上的超链接又会指向其他的 URL 页面。

[0037] 状态是 Ajax 应用某一时刻在浏览器中呈现的页面 DOM 结构，也就是说，不同的 DOM 树即不同的页面状态，客户端用户操作或服务器端数据响应都有可能导致 Ajax 应用的 DOM

结构发生变化，从而产生新的状态。Ajax 应用中包含一系列离散的状态。其中包括一个初始状态，以及由初始状态经过一次或多次转换得到的很多不同的中间状态。转换是指通过触发某 DOM 元素的事件，Ajax 应用从一个状态转换为另一个状态。

[0038] 在 S110 中，抓取页面的装置获取页面的 Xpath 元素，并通过对 Xpath 元素进行规约获取规约后的 Xpath 元素。

[0039] 可选地，可通过嵌入浏览器技术获取 Xpath 元素，例如，HtmlUnit 包。通过嵌入式浏览器加载起始 URL 获得默认的 DOM 树，然后通过分析 DOM 树获取所有 Xpath 元素，可选地，可以调用设计好的传统爬虫进行超链分析，获取所有的 URL 集合，再调用嵌入的浏览器接口获取所有页面的所有 Xpath 元素。

[0040] 应理解，获取 Xpath 元素的方式还可以采用其他页面 Javascript 代码解析技术，本发明实施例对此并不限定。

[0041] 在获取 Xpath 元素后，对 Xpath 元素进行规约，获取规约后的 Xpath 元素。由于在 web2.0 网站中 URL 不能作为页面的唯一标识，通过相似性判定不能对 URL 的特征进行归约。在本发明实施例中，以页面元素为中心，使用 XPath 作为页面元素的描述，并对所有引向需要页面的页面元素的 XPath 进行了归约，同时记录需触发的事件。对 XPath 的归约采用如下的归约方法：

[0042] 被归约的 XPath 路径经过的页面元素名称必须相同，对页面元素的序号进行归约。如对“/html/body/div[4]/li[1]/a[1]”和“/html/body/div[4]/li[2]/a[1]”这两个 XPath，归约为“/html/body/div[4]/li[*]/a[1]”，如还存在“/html/body/div[3]/li[1]/a[1]”这一 XPath，归约为“/html/body/div[*]/li[*]/a[1]”。但“/html/body/div[1]”，“/html/body/span[1]”，“/html/body/div[1]/span[1]”中的任何两个均不被归约。这样归约出的 XPath 结果不再含有无效的页面元素，可以作为抓取过程中的特征。

[0043] 可选地，在获取规约后的 Xpath 元素之后，该方法 100 还包括：

[0044] 生成规约后的 Xpath 元素的状态转换图模型。

[0045] 如图 2 所示，状态转换图是一个二元组 <V，E>，其中 V 表示状态节点的集合，每个节点 v ∈ V 表示页面抓取过程的一个状态；E 是节点间的有向边集合，每条边是一个二元组 <Xpath,event>。比如图 2 中的 <onclick，/html/body/div[1]/a[1]>，onclick 表示事件 event，/html/body/div[1]/a[1] 表示 Xpath，从 v1 到 v2 的有向边存在，当且仅当状态 v1 可以通过触发 Xpath 所代表的页面元素上的事件 event 转换到状态 v2。

[0046] 按照状态转换图模型的定义，将获取的规约后的 Xpath 元素建立状态转换图模型，生成状态转换图模型结构数据。这样，后续步骤可以对建立的状态转换图模型进行处理。

[0047] 应理解，根据 Xpath 元素建立状态转换图模型只是处理 Xpath 元素数据的一种实施方式，不应对本发明的保护范围构成限定，本发明实施例还可以采用其他数据处理方式。

[0048] 在 S120 中，抓取页面的装置在规约后的 Xpath 元素中确定训练样本。

[0049] 在本发明实施例中，在将 Xpath 元素规约后，不是执行所有规约后的 Xpath 元素的事件，而是选取一部分 Xpath 元素作为训练样本，以训练分类器。例如，可以选取规约后的 Xpath 元素的 10% 作为训练样本，本发明实施例对训练元素的比例并不限定，其值可以根据实时状况而调整。

[0050] 可选地，若将获取的规约后的 Xpath 元素建立状态转换图模型，则 S120 包括：

[0051] 在该状态转换图模型中确定训练样本。

[0052] 也就是说，若采用建立状态转换图模型的数据处理方式，则在状态转换图模型建立后，在该状态转换图模型中抽取训练样本以训练分类器。

[0053] 在 S130 中，抓取页面的装置执行训练样本中的每个 Xpath 元素的事件，根据训练样本中的每个 Xpath 元素的事件产生的 DOM 树与原 DOM 树的编辑距离确定训练样本中的每个 Xpath 元素是否有效，根据训练样本中的每个 Xpath 元素是否有效训练分类器。

[0054] 具体而言，在确定训练样本后，抓取页面的装置根据训练样本训练分类器。抓取页面的装置执行训练样本中的每个 Xpath 元素的事件，根据产生的 DOM 树与原 DOM 树的编辑距离确定该 Xpath 元素是否有效。可选地，可以采用限制自上而下映射（Restricted Top-Down Mapping，简称为“RTDM”）算法来计算两个页面的 DOM 树之间的编辑距离，即执行 Xpath 元素的事件前后的 DOM 树的编辑距离。根据 DOM 树的编辑距离是否大于预定阈值确定 Xpath 元素是否有效。比如：

[0055] 若该训练样本中的第一 Xpath 元素的事件产生的 DOM 树与原 DOM 树的编辑距离大于预定阈值，则确定该第一 Xpath 元素有效；

[0056] 若该训练样本中的第二 Xpath 元素的事件产生的 DOM 树与原 DOM 树的编辑距离不大于该预定阈值，则确定该第二 Xpath 元素无效。

[0057] 应理解，在本发明实施例中，“第一”、“第二”、“第三”和“第四”仅仅是为了区分不同的 Xpath 元素，不应对本发明实施例构成任何限定。

[0058] 也就是说，若训练样本中的某一 Xpath 元素的事件产生的 DOM 树与原 DOM 树的编辑距离大于预定阈值，则确定该 Xpath 元素有效；若训练样本中的某一 Xpath 元素的事件产生的 DOM 树与原 DOM 树的编辑距离不大于预定阈值，则确定该 Xpath 元素无效。

[0059] 在确定了训练样本中 Xpath 元素是否有效后，抓取页面的装置根据训练样本中的每个 Xpath 元素是否有效训练分类器，即得到 Xpath 元素有效或无效的分类器。可选地，可以采用支持向量机（Support Vector Machine，简称为“SVM”）算法训练分类器。SVM 是通用的知识发现和机器学习方法，主要是针对两类模式的分类问题，在高维特征空间中寻找最大边缘超平面（也称为最优分类面）作为两类的分界面，从而保证对未知样本的最小分类错误率。本发明实施例利用 SVM 算法训练 Xpath 元素有效或无效的分类器，例如，类似于“/html/body/div[4]/li[1]/a[1]”这种 Xpath 路径，将之间的“/”符号去掉后，就变为 [html,body,div[4],li[1],a[1]] 这样的一维向量，将这种数据模型进行 SVM 训练，得到 Xpath 元素有效或无效的分类器。如图 3 所示，根据样本中各元素有效或无效的结果，调用 SVM 算法进行训练，得到分类器。

[0060] 在 S140 中，抓取页面的装置通过分类器对规约后的 Xpath 元素进行分类，获取有效 Xpath 元素集合。

[0061] 在得到分类器后，抓取页面的装置利用该分类器对规约后的 Xpath 元素进行分类，过滤掉无效的 Xpath 元素，得到所有有效 Xpath 元素，即有效 Xpath 元素集合。如图 3 所示，将待分类的 Xpath 元素输入分类器，从输出中得到有效 Xpath 元素集合。

[0062] 可选地，若将获取的规约后的 Xpath 元素建立状态转换图模型，则 S140 包括：

[0063] 将该状态转换图模型输入分类器，获取有效 Xpath 元素集合。

[0064] 也就是说，若采用建立状态转换图模型的数据处理方式，则将建立的状态转换图模型作为输入，经过分类器进行分类，过滤掉无效的 Xpath 元素，得到有效 Xpath 元素集合。

[0065] 在 S150 中，抓取页面的装置执行有效 Xpath 元素集合中的每个 Xpath 元素的事件，根据有效 Xpath 元素集合中的每个 Xpath 元素的事件产生的 DOM 树与原 DOM 树的编辑距离抓取页面。

[0066] 具体而言，在通过分类器得到有效 Xpath 元素集合后，抓取页面的装置基于该有效 Xpath 元素集合中的 Xpath 元素抓取页面。抓取页面的装置执行该有效 Xpath 元素集合中的每个 Xpath 元素的事件，根据该有效 Xpath 元素集合中的每个 Xpath 元素的事件产生的 DOM 树与原 DOM 树的编辑距离是否大于预定阈值抓取页面。比如：

[0067] 若该有效 Xpath 元素集合中的第三 Xpath 元素的事件产生的 DOM 树与原 DOM 树的编辑距离大于该预定阈值，则保存该第三 Xpath 元素的事件产生的 DOM 树；

[0068] 若该有效 Xpath 元素集合中的第四 Xpath 元素的事件产生的 DOM 树与原 DOM 树的编辑距离不大于该预定阈值，则不保存该第四 Xpath 元素的事件产生的 DOM 树。

[0069] 也就是说，若有效 Xpath 元素集合中的某一 Xpath 元素的事件产生的 DOM 树与原 DOM 树的编辑距离大于该预定阈值，则认为页面状态转换，保存该 Xpath 元素的事件产生的 DOM 树，加入爬行队列中；若有效 Xpath 元素集合中的某一 Xpath 元素的事件产生的 DOM 树与原 DOM 树的编辑距离不大于该预定阈值，则认为页面状态没有转换，不保存该 Xpath 元素的事件产生的 DOM 树。

[0070] 应理解，在本发明实施例中，在通过分类器对 Xpath 元素进行分类时，可以将所有规约后的 Xpath 元素作为输入，也可以将规约后的 Xpath 元素中除训练样本之外的 Xpath 元素作为输入，在后一种情况下，即只分类除训练样本之外的 Xpath 元素，则需要在训练分类器时将训练样本中有效 Xpath 元素的事件产生的 DOM 树保存。因此，可选地，在根据训练样本中的每个 Xpath 元素的事件产生的 DOM 树与原 DOM 树的编辑距离确定训练样本中的每个 Xpath 元素是否有效之后，该方法 100 还包括：

[0071] 保存训练样本中的有效 Xpath 元素的事件产生的 DOM 树。

[0072] 在这种情况下，S140 包括：

[0073] 通过分类器对规约后的 Xpath 元素中除训练样本之外的 Xpath 元素进行分类，获取有效 Xpath 元素集合。

[0074] 这样，分类器输出的有效 Xpath 元素集合不包含训练样本中的 Xpath 元素，在根据该有效 Xpath 元素集合抓取页面时，也省去了对训练样本中的有效 Xpath 元素的事件的执行以及 DOM 树的比较。

[0075] 本发明实施例的抓取页面的方法，根据训练样本中的 Xpath 元素是否有效训练分类器，通过分类器对 Xpath 元素进行分类，获取有效 Xpath 元素集合，再基于有效 Xpath 元素集合抓取页面，可以过滤掉大量的无效 Xpath 元素，从而能够提升抓取页面的效率，并且，本发明实施例的抓取页面的方法对内存、CPU 等计算机资源的要求降低，从而降低了部署成本。

[0076] 为了满足业务定制的需求，在本发明实施例中，可选地，在根据训练样本中的每个 Xpath 元素是否有效训练分类器之前，该方法 100 还包括：

[0077] 获取业务定制信息，根据该业务定制信息确定定制规则。

[0078] 在这种情况下，根据训练样本中的每个 Xpath 元素是否有效训练分类器，包括：

[0079] 根据训练样本中的每个 Xpath 元素是否有效和该定制规则，训练该分类器。

[0080] 也就是说，在训练分类器前，确定定制规则，然后在训练分类器时加入该定制规则，这样，在利用分类器对 Xpath 元素进行分类时，就能得到有效且符合定制规则的 Xpath 元素。因此，本发明实施例的抓取页面的方法，可以满足业务定制需求，具有可扩型和定制性。

[0081] 下面结合图 4 详细描述本发明实施例。应注意，这只是为了帮助本领域技术人员更好地理解本发明实施例，而非限制本发明实施例的范围。

[0082] 401，首先输入一个初始 URL 作为入口；

[0083] 402，超链分析，例如调用设计好的传统爬虫进行超链分析，获取所有的 URL 集合；

[0084] 403，通过嵌入浏览器技术获取所有 Xpath 元素，例如调用嵌入的浏览器接口获取所有页面的所有 Xpath 元素；

[0085] 404，对步骤 403 获取的 Xpath 元素进行规约；

[0086] 405，根据步骤 402-404 获取的 Xpath 元素数据建立状态转换图模型；

[0087] 406，确定训练样本，即读取配置文件中确定的训练元素的比例，也就是样本的大小；

[0088] 407，读取配置文件中业务定制信息；

[0089] 408，确定定制规则；

[0090] 409，调用 RTDM 算法计算样本中的 Xpath 元素的事件产生的 DOM 树的编辑距离，确定哪些 Xpath 能产生新状态，哪些不能，即哪些 Xpath 元素有效，哪些 Xpath 元素无效；

[0091] 410，SVM 训练，根据样本结果调用 SVM 算法进行训练；

[0092] 411，建立分类器；

[0093] 412，把 405 中建立的状态转换图模型作为输入，经过分类器进行分类，过滤掉无效的 Xpath 元素，获得有效 Xpath 元素集合。

[0094] 413，基于有效 Xpath 元素集合抓取页面，执行有效 Xpath 元素集合中的每个 Xpath 元素的事件，根据该 Xpath 元素的事件产生的 DOM 树与原 DOM 树的编辑距离是否大于预定阈值抓取页面，若大于预定阈值，则保存该 Xpath 元素的事件产生的 DOM 树。

[0095] 基于上述描述，本发明实施例的抓取页面的方法的一个算法原型如下：

[0096]

```
Function Init(url)
dom = browser.Load(url)
crawlQueue.Enqueue(dom)
StateMachine sm = new StateMachine(url)
end Function

Function SpiderPlan();
Function SVMinit();
Function SVMout();

Function BreadthFirstAjaxCrawl(url)
Init(url)
SVMinit()
invalidElementFeatures = SVMout(SpiderPlan())
for crawlQueue is not empty
        dom = crawlQueue.Dequeue()
        State prevState = new State(dom)
        for all Event e in dom and e.source not in invalidElementFeatures
                if e.function invokes XmlHttpRequest
                        sentRequests.Add(XmlHttpRequests)
                end if
                newDom = dom.ExecuteFunction(e.function)
                if Distance(dom,newDom) > delta and newDom not in sm.States
                        State nextState = new State(newDom)
                        Transition t = new
Transition(prevState,nextState,e.source,e.function)
                        sm.Add(t)
                        crawlQueue.Enqueue(newDom)
                end if
        end for
end for
end Function
```

[0097] 应理解，上述算法原型只是示例，不应对本发明的保护范围构成任何限定。

[0098] 应理解，在本发明的各种实施例中，上述各过程的序号的大小并不意味着执行顺序的先后，各过程的执行顺序应以其功能和内在逻辑确定，而不应对本发明实施例的实施过程构成任何限定。

[0099] 上文中结合图 1 至图 4，详细描述了根据本发明实施例的抓取网页的方法，下面将结合图 5 至图 7，描述根据本发明实施例的抓取网页的装置。

[0100] 图 5 示出了根据本发明实施例的抓取页面的装置 500 的示意性框图。如图 5 所示，

该装置 500 包括：

[0101] 获取模块 510，用于获取页面的可扩展标记语言路径语言 Xpath 元素，并通过对该 Xpath 元素进行规约获取规约后的 Xpath 元素；

[0102] 确定模块 520，用于在该规约后的 Xpath 元素中确定训练样本；

[0103] 训练模块 530，用于执行该训练样本中的每个 Xpath 元素的事件，根据该训练样本中的每个 Xpath 元素的事件产生的文档对象模型 DOM 树与原 DOM 树的编辑距离确定该训练样本中的每个 Xpath 元素是否有效，根据该训练样本中的每个 Xpath 元素是否有效训练分类器；

[0104] 分类模块 540，用于通过该分类器对该规约后的 Xpath 元素进行分类，获取有效 Xpath 元素集合；

[0105] 抓取模块 550，用于执行该有效 Xpath 元素集合中的每个 Xpath 元素的事件，根据该有效 Xpath 元素集合中的每个 Xpath 元素的事件产生的 DOM 树与原 DOM 树的编辑距离抓取页面。

[0106] 本发明实施例的抓取页面的装置，根据训练样本中的 Xpath 元素是否有效训练分类器，通过分类器对 Xpath 元素进行分类，获取有效 Xpath 元素集合，再基于有效 Xpath 元素集合抓取页面，可以过滤掉大量的无效 Xpath 元素，从而能够提升抓取页面的效率。

[0107] 在本发明实施例中，如图 6 所示，可选地，该装置 500 还包括：

[0108] 定制模块 560，用于获取业务定制信息，根据该业务定制信息确定定制规则；

[0109] 该训练模块 530 具体用于根据该训练样本中的每个 Xpath 元素是否有效和该定制规则，训练该分类器。

[0110] 本发明实施例的抓取页面的装置，可以满足业务定制需求，具有可扩型和定制性。

[0111] 在本发明实施例中，可选地，该训练模块 530 具体用于，若该训练样本中的第一 Xpath 元素的事件产生的 DOM 树与原 DOM 树的编辑距离大于预定阈值，则确定该第一 Xpath 元素有效，若该训练样本中的第二 Xpath 元素的事件产生的 DOM 树与原 DOM 树的编辑距离不大于该预定阈值，则确定该第二 Xpath 元素无效；

[0112] 该抓取模块 550 具体用于，若该有效 Xpath 元素集合中的第三 Xpath 元素的事件产生的 DOM 树与原 DOM 树的编辑距离大于该预定阈值，则保存该第三 Xpath 元素的事件产生的 DOM 树，若该有效 Xpath 元素集合中的第四 Xpath 元素的事件产生的 DOM 树与原 DOM 树的编辑距离不大于该预定阈值，则不保存该第四 Xpath 元素的事件产生的 DOM 树。

[0113] 在本发明实施例中，可选地，该抓取模块 550 还用于在该训练模块 530 根据该训练样本中的每个 Xpath 元素的事件产生的 DOM 树与原 DOM 树的编辑距离确定该训练样本中的每个 Xpath 元素是否有效之后，保存该训练样本中的有效 Xpath 元素的事件产生的 DOM 树；

[0114] 该分类模块 540 具体用于通过该分类器对该规约后的 Xpath 元素中除该训练样本之外的 Xpath 元素进行分类，获取该有效 Xpath 元素集合。

[0115] 在本发明实施例中，可选地，该装置 500 还包括：

[0116] 生成模块，用于在该获取模块获取规约后的 Xpath 元素之后，生成该规约后的 Xpath 元素的状态转换图模型；

[0117] 该确定模块 520 具体用于在该状态转换图模型中确定训练样本；

[0118] 该分类模块 540 具体用于将该状态转换图模型输入该分类器，获取该有效 Xpath

元素集合。

[0119] 在本发明实施例中，可选地，该获取模块 510 具体用于通过嵌入浏览器技术获取该 Xpath 元素。

[0120] 本发明实施例的抓取页面的装置，根据训练样本中的 Xpath 元素是否有效训练分类器，通过分类器对 Xpath 元素进行分类，获取有效 Xpath 元素集合，再基于有效 Xpath 元素集合抓取页面，可以过滤掉大量的无效 Xpath 元素，从而能够提升抓取页面的效率，并且，本发明实施例的抓取页面的装置对内存、CPU 等计算机资源的要求降低，从而降低了部署成本。

[0121] 根据本发明实施例的抓取页面的装置 500 可对应于根据本发明实施例的的方法中的抓取页面的装置，并且装置 500 中的各个模块的上述和其它操作和 / 或功能分别为了实现图 1 至图 4 中的各个方法的相应流程，为了简洁，在此不再赘述。

[0122] 图 7 是本发明实施例提供的抓取页面的装置的结构示意图。如图 7 所示，装置 700 一般包括至少一个处理器 710，例如 CPU，至少一个端口 720，存储器 730，和至少一个通信总线 740。通信总线 740 用于实现这些设备之间的连接通信。处理器 710 用于执行存储器 730 中存储的可执行模块，例如计算机程序；装置 700 可选地包含用户接口 750，包括但不限于显示器，键盘和点击设备，例如鼠标、轨迹球（trackball）、触感板或者触感显示屏。存储器 730 可能包含高速 RAM 存储器，也可能还包括非易失性存储器（non-volatile memory），例如至少一个磁盘存储器。通过至少一个端口 720 实现该装置 700 与网络的通信连接。

[0123] 在一些实施方式中，存储器 730 存储了如下的元素，可执行模块或者数据结构，或者他们的子集，或者他们的扩展集：

[0124] 操作系统 732，包含各种系统程序，用于实现各种基础业务以及处理基于硬件的任务；

[0125] 应用模块 734，包含各种应用程序，用于实现各种应用业务。

[0126] 应用模块 734 中包括但不限于获取模块 510、确定模块 520、训练模块 530、分类模块 540、抓取模块 550 和定制模块 560。

[0127] 应用模块 734 中各模块的具体实现参见图 5 和图 6 所示实施例中的相应模块，在此不赘述。

[0128] 应理解，在本发明实施例中，术语“和 / 或”仅仅是一种描述关联对象的关联关系，表示可以存在三种关系。例如，A 和 / 或 B，可以表示：单独存在 A，同时存在 A 和 B，单独存在 B 这三种情况。另外，本文中字符“/”，一般表示前后关联对象是一种“或”的关系。

[0129] 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，为了清楚地说明硬件和软件的可互换性，在上述说明中已经按照功能一般性地描述了各示例的组成及步骤。这些功能究竟以硬件还是软件方式来执行，取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能，但是这种实现不应认为超出本发明的范围。

[0130] 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，上述描述的系统、装置和单元的具体工作过程，可以参考前述方法实施例中的对应过程，在此不再赘述。

[0131] 在本申请所提供的几个实施例中，应该理解到，所揭露的系统、装置和方法，可以

通过其它的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，例如，所述单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，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。另外，所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口、装置或单元的间接耦合或通信连接，也可以是电的，机械的或其它的形式连接。

[0132] 所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本发明实施例方案的目的。

[0133] 另外，在本发明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。上述集成的单元既可以采用硬件的形式实现，也可以采用软件功能单元的形式实现。

[0134] 所述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时，可以存储在一个计算机可读取存储介质中。基于这样的理解，本发明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，服务器，或者网络设备等）执行本发明各个实施例所述方法的全部或部分步骤。而前述的存储介质包括：U盘、移动硬盘、只读存储器（ROM，Read-Only Memory）、随机存取存储器（RAM，Random Access Memory）、磁碟或者光盘等各种可以存储程序代码的介质。

[0135] 以上所述，仅为本发明的具体实施方式，但本发明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发明揭露的技术范围内，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发明的保护范围之内。因此，本发明的保护范围应以权利要求的保护范围为准。

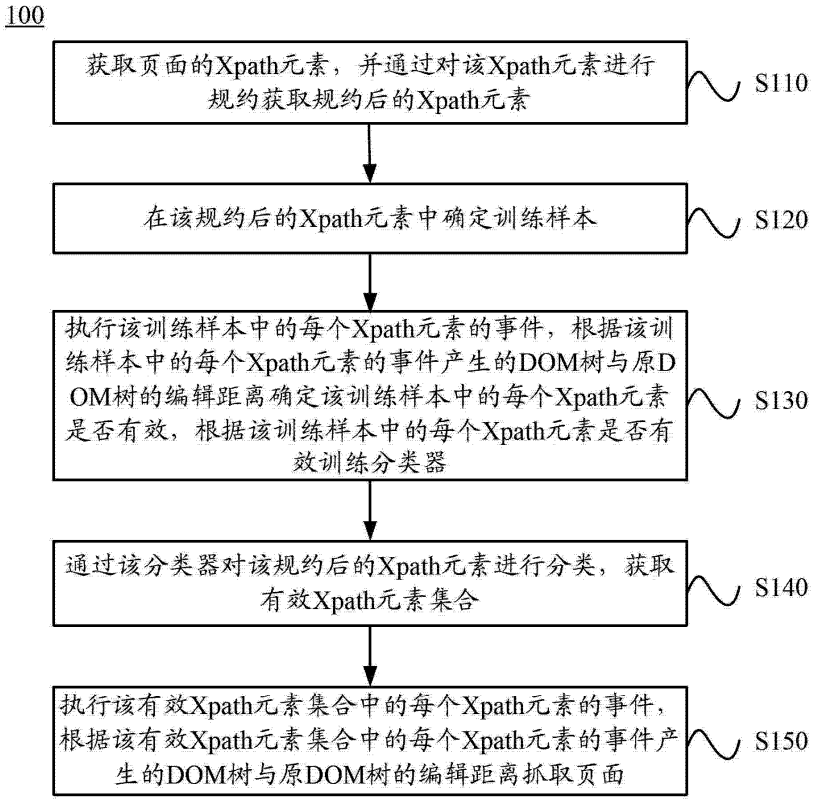

100
获取页面的Xpath元素，并通过对该Xpath元素进行规约获取规约后的Xpath元素
S110
在该规约后的Xpath元素中确定训练样本
S120
执行该训练样本中的每个Xpath元素的事件，根据该训练样本中的每个Xpath元素的事件产生的DOM树与原DOM树的编辑距离确定该训练样本中的每个Xpath元素是否有效，根据该训练样本中的每个Xpath元素是否有效训练分类器
S130
通过该分类器对该规约后的Xpath元素进行分类，获取有效Xpath元素集合
S140
执行该有效Xpath元素集合中的每个Xpath元素的事件，根据该有效Xpath元素集合中的每个Xpath元素的事件产生的DOM树与原DOM树的编辑距离抓取页面
S150

图 1

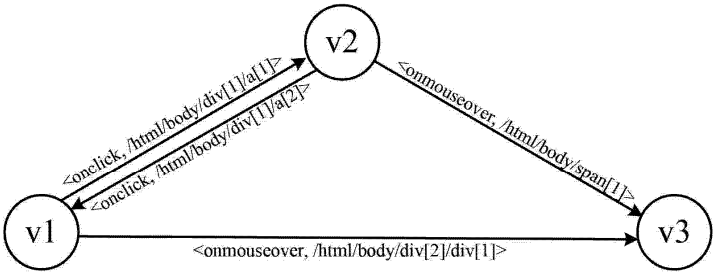

v2
<onclick, /html/body/div[1]/a[1]>
<onclick, /html/body/div[1]/a[2]>
<onmouseover, /html/body/span[1]>
v1
v3
<onmouseover, /html/body/div[2]/div[1]>

图 2

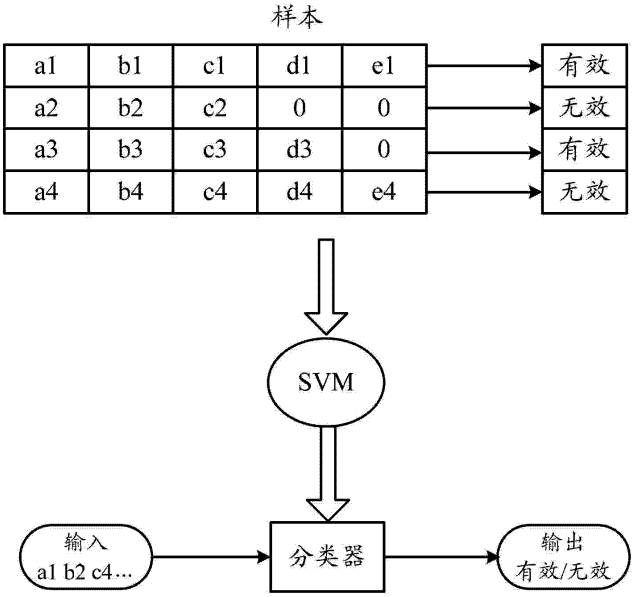

样本
a1 b1 c1 d1 e1
a2 b2 c2 0 0
a3 b3 c3 d3 0
a4 b4 c4 d4 e4
有效
无效
有效
无效
SVM
输入
a1 b2 c4...
分类器
输出
有效/无效

图 3

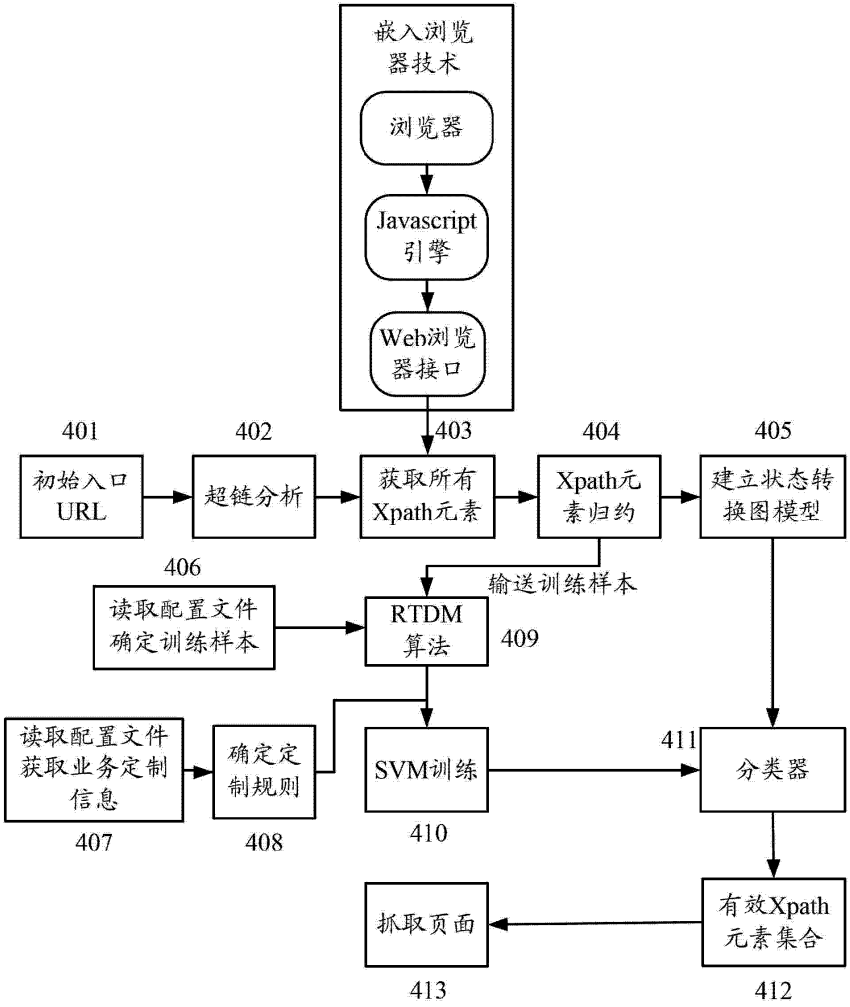

嵌入浏览器技术
浏览器
Javascript引擎
Web浏览器接口
401
初始入口URL
402
超链分析
403
获取所有Xpath元素
404
Xpath元素归约
405
建立状态转换图模型
406
读取配置文件确定训练样本
输送训练样本
RTDM算法
409
读取配置文件获取业务定制信息
407
确定定制规则
408
SVM训练
410
411
分类器
抓取页面
413
有效Xpath元素集合
412

图 4

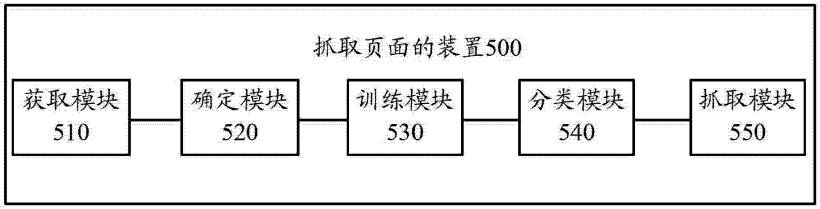

抓取页面的装置500
获取模块
510
确定模块
520
训练模块
530
分类模块
540
抓取模块
550

图 5

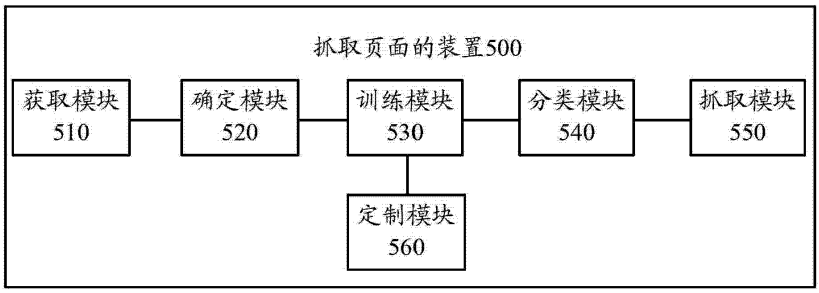

抓取页面的装置500
获取模块
510
确定模块
520
训练模块
530
分类模块
540
抓取模块
550
定制模块
560

图 6

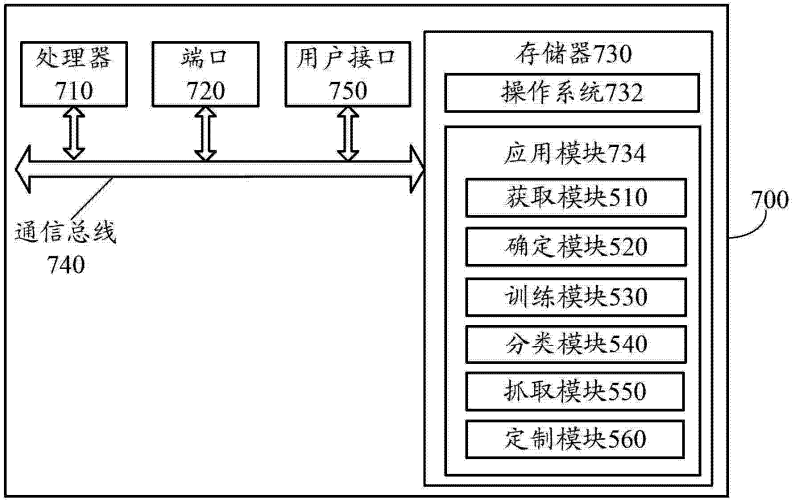

处理器
710
端口
720
用户接口
750
存储器730
操作系统732
应用模块734
获取模块510
确定模块520
训练模块530
分类模块540
抓取模块550
定制模块560
700
通信总线
740

图 7