(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 102890692 A
(43)申请公布日 2013.01.23

(21)申请号 201110207897.1

(22)申请日 2011.07.22

(71)申请人 阿里巴巴集团控股有限公司
地址 英属开曼群岛大开曼资本大厦一座四层 847 号邮箱

(72)发明人 孙一鸣 强琦 蔡波洋 金晓军 吴宗远

(74)专利代理机构 北京润泽恒知识产权代理有限公司 11319
代理人 苏培华

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

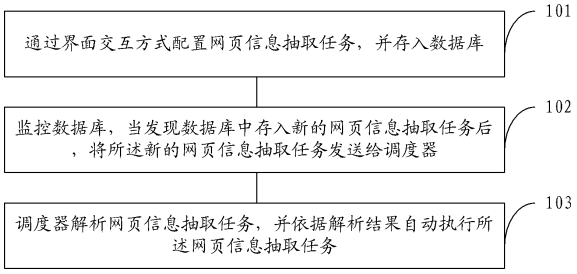
权利要求书 2 页 说明书 12 页 附图 7 页

(54)发明名称

一种网页信息抽取方法及抽取系统

(57)摘要

本申请提供了一种网页信息抽取方法及抽取系统，以解决现有的信息抽取方法自动化程度不高并且技术门槛较高的问题。所述方法包括：通过界面交互方式配置网页信息抽取任务，并存入数据库；监控数据库，当发现数据库中存入新的网页信息抽取任务后，将所述新的网页信息抽取任务发送给调度器；调度器解析网页信息抽取任务，并依据解析结果自动执行所述网页信息抽取任务。所述界面交互的方式实现了简单的人机交互，对于非专业人员而言，也可以按照界面的提示进行任务配置，极大地降低了信息抽取的门槛。而且，调度器依据网页信息抽取任务执行的一套自动抽取方式，可以实现大批量网页高度自动化的信息抽取。



CN 102890692 A

1. 一种网页信息抽取方法，其特征在于，包括：

通过界面交互方式配置网页信息抽取任务，并存入数据库；

监控数据库，当发现数据库中存入新的网页信息抽取任务后，将所述新的网页信息抽取任务发送给调度器；

调度器解析网页信息抽取任务，并依据解析结果自动执行所述网页信息抽取任务。

2. 根据权利要求 1 所述的方法，其特征在于，所述通过界面交互方式配置网页信息抽取任务，包括：

通过界面交互方式执行以下操作：

提交标注页面；

在所述标注页面上标注页面信息的点击行为和 / 或抽取行为；

对所述点击行为或抽取行为进行细化配置。

3. 根据权利要求 2 所述的方法，其特征在于，对所述点击行为或抽取行为进行细化配置之前，还包括：

配置所述点击行为或抽取行为的操作对象是单一信息还是列表信息，

如果是单一信息，则针对该单一信息进行点击行为或抽取行为的细化配置；

如果是列表信息，则配置列表属性，并从列表中选取样例进行点击行为或抽取行为的细化配置。

4. 根据权利要求 2 所述的方法，其特征在于：

每个点击行为在触发页面跳转时都会产生一个新的标注页面；

最初的标注页面为起始页面，基于起始页面产生的所有标注页面的集合构成一棵以该起始页面为根的标注树，所有的起始页面代表的标注树构成一个标注森林；

所述网页信息抽取任务为一个标注森林或一棵标注树或一个标注页面。

5. 根据权利要求 1 至 4 任一所述的方法，其特征在于，所述调度器解析网页信息抽取任务，并依据解析结果自动执行所述网页信息抽取任务，包括：

所述调度器解析网页信息抽取任务，并依据解析结果调度进行网页抓取，和 / 或页面渲染，和 / 或页面信息抽取。

6. 根据权利要求 4 所述的方法，其特征在于，所述调度器解析网页信息抽取任务，并依据解析结果自动执行所述网页信息抽取任务，包括：

调度器解析网页信息抽取任务，并获得每个标注页面的配置；

依据标注页面的 URL 调度抓取页面数据；

调度渲染标注页面，并获得标注页面的 DOM 树结构；

遍历标注页面的 DOM 树结构中对应节点的配置，并依据所述节点的配置进行如下操作：

如果是抽取行为，则依据抽取行为的配置调度抽取文本信息；

如果是点击行为，并且如果是下载行为，则依据点击行为的配置调度抓取下载内容；如果是涉及渲染的点击行为，则依据点击行为的配置进行调度渲染。

7. 一种网页信息抽取系统，其特征在于，包括：

抽取配置模块，用于通过界面交互方式配置网页信息抽取任务，并存入数据库；

数据库，用于存储网页信息抽取任务；

监控模块，用于监控数据库，当发现数据库中存入新的网页信息抽取任务后，将所述新的网页信息抽取任务发送给调度器；

调度器，用于解析网页信息抽取任务，并依据解析结果自动执行所述网页信息抽取任务。

8. 根据权利要求 7 所述的系统，其特征在于，所述抽取配置模块包括：

配置入口子模块，用于提交标注页面；

行为标注子模块，用于在所述标注页面上标注页面信息的点击行为和 / 或抽取行为；

细化配置子模块，用于对所述点击行为或抽取行为进行细化配置。

9. 根据权利要求 8 所述的系统，其特征在于，所述抽取配置模块还包括：

元素类型选择子模块，用于配置所述点击行为或抽取行为的操作对象是单一信息还是列表信息；

如果是单一信息，则触发所述细化配置子模块针对该单一信息进行点击行为或抽取行为的细化配置；

如果是列表信息，则配置列表属性，并触发所述细化配置子模块从列表中选取样例进行点击行为或抽取行为的细化配置。

10. 根据权利要求 8 所述的系统，其特征在于：

每个点击行为在触发页面跳转时都会产生一个新的标注页面；

最初的标注页面为起始页面，基于起始页面产生的所有标注页面的集合构成一棵以该起始页面为根的标注树，所有的起始页面代表的标注树构成一个标注森林；

所述网页信息抽取任务为一个标注森林或一棵标注树或一个标注页面。

11. 根据权利要求 10 所述的系统，其特征在于，还包括：

抓取器，用于根据调度器的调度，依据标注页面的 URL 抓取页面数据，并返回给所述调度器；

渲染引擎，用于根据调度器的调度，渲染抓取回来的标注页面，并获得标注页面的 DOM 树结构，返回给所述调度器；

抽取器，用于根据调度器的调度，依据网页信息抽取任务的配置抽取相应的网页信息；

所述调度器通过解析网页信息抽取任务获得每个标注页面的配置；

所述调度器依据标注页面的 URL 调度抓取器抓取页面数据；

所述调度器调度渲染引擎渲染标注页面，并获得标注页面的 DOM 树结构；

所述调度器遍历标注页面的 DOM 树结构中对应节点的配置，并依据所述节点的配置进行如下操作：

如果是抽取行为，则依据抽取行为的配置调度抽取器抽取文本信息；

如果是点击行为，并且如果是下载行为，则依据点击行为的配置调度抓取器抓取下载内容；如果是涉及渲染的点击行为，则依据点击行为的配置调度渲染引擎进行渲染。

一种网页信息抽取方法及抽取系统

技术领域

[0001] 本申请涉及网页处理技术，特别是涉及一种网页信息抽取方法及抽取系统。

背景技术

[0002] 网页信息抽取就是获取网页的数据，然后通过程序分析，将有用的数据提取分离出来。比如编制程序抽取某网站新闻频道里的某个新闻标题就是一种网页信息抽取。目前的信息抽取主要分为两种，一种是基于规则的抽取，规则可以人工定制，也可以通过学习得到，另一种就是利用机器学习方法进行抽取。

[0003] 搜索引擎工作的一部分就是网页信息抽取。随着互联网的发展，互联网上的信息规模也随之不断扩大。由于互联网上的数据来自于大量不同的站点，而不同站点的页面结构差异很大，因此搜索引擎无法开发出通用的抽取器来分析源自不同站点的网页。

[0004] 由于这个原因，最初的搜索引擎，尤其是垂直搜索引擎（针对某类知识领域的专业搜索引擎）通过许多个定向的抽取器来解决这一问题，即每个抽取器定向抽取某个站点或具有某类页面结构的网页信息。但是，由于这种信息抽取方法需要维护很多个定向抽取器，因此存在不易维护的问题，而且新添加一个或一类站点就需要开发新的定向抽取器，开发成本也很高。

[0005] 之后，人们开始寻找能够自动生成抽取器的方案。火车头采集器是一种主要基于正则表达式的信息抽取方法，包含信息的抓取、抽取、发布等功能，通过用户配置的正则表达式，实现定制化的抓取与抽取。

[0006] 但是，这种单纯基于正则表达式的信息抽取方法，还需要手工配置正则表达式，自动化程度不高，对大批量的网页抽取支持不够。而且，使用者需要掌握正则表达式的知识，同时也必须对网页结构有相当程度的了解，对非专业人员来说，技术门槛较高。

[0007] 因此，目前还没有一种真正简单、实用的自动化抽取方法，供搜索引擎或其他互联网应用进行网页信息的自动抽取。

发明内容

[0008] 本申请提供了一种网页信息抽取方法及抽取系统，以解决现有的信息抽取方法自动化程度不高并且技术门槛较高的问题。

[0009] 为了解决上述问题，本申请公开了一种网页信息抽取方法，包括：

[0010] 通过界面交互方式配置网页信息抽取任务，并存入数据库；

[0011] 监控数据库，当发现数据库中存入新的网页信息抽取任务后，将所述新的网页信息抽取任务发送给调度器；

[0012] 调度器解析网页信息抽取任务，并依据解析结果自动执行所述网页信息抽取任务。

[0013] 优选的，所述通过界面交互方式配置网页信息抽取任务，包括：通过界面交互方式执行以下操作：提交标注页面；在所述标注页面上标注页面信息的点击行为和/或抽取行

为 ;对所述点击行为或抽取行为进行细化配置。

[0014] 优选的,对所述点击行为或抽取行为进行细化配置之前,还包括 :配置所述点击行为或抽取行为的操作对象是单一信息还是列表信息,如果是单一信息,则针对该单一信息进行点击行为或抽取行为的细化配置 ;如果是列表信息,则配置列表属性,并从列表中选取样例进行点击行为或抽取行为的细化配置。

[0015] 优选的,每个点击行为在触发页面跳转时都会产生一个新的标注页面 ;最初的标注页面为起始页面,基于起始页面产生的所有标注页面的集合构成一棵以该起始页面为根的标注树,所有的起始页面代表的标注树构成一个标注森林 ;所述网页信息抽取任务为一个标注森林或一棵标注树或一个标注页面。

[0016] 优选的,所述调度器解析网页信息抽取任务,并依据解析结果自动执行所述网页信息抽取任务,包括 :所述调度器解析网页信息抽取任务,并依据解析结果调度进行网页抓取,和 / 或页面渲染,和 / 或页面信息抽取。

[0017] 优选的,所述调度器解析网页信息抽取任务,并依据解析结果自动执行所述网页信息抽取任务,包括 :调度器解析网页信息抽取任务,并获得每个标注页面的配置 ;依据标注页面的 URL 调度抓取页面数据 ;调度渲染标注页面,并获得标注页面的 DOM 树结构 ;遍历标注页面的 DOM 树结构中对应节点的配置,并依据所述节点的配置进行如下操作 :如果是抽取行为,则依据抽取行为的配置调度抽取文本信息 ;如果是点击行为,并且如果是下载行为,则依据点击行为的配置调度抓取下载内容 ;如果是涉及渲染的点击行为,则依据点击行为的配置进行调度渲染。

[0018] 本申请还提供了一种网页信息抽取系统,包括 :

[0019] 抽取配置模块,用于通过界面交互方式配置网页信息抽取任务,并存入数据库 ;

[0020] 数据库,用于存储网页信息抽取任务 ;

[0021] 监控模块,用于监控数据库,当发现数据库中存入新的网页信息抽取任务后,将所述新的网页信息抽取任务发送给调度器 ;

[0022] 调度器,用于解析网页信息抽取任务,并依据解析结果自动执行所述网页信息抽取任务。

[0023] 优选的,所述抽取配置模块包括 :

[0024] 配置入口子模块,用于提交标注页面 ;

[0025] 行为标注子模块,用于在所述标注页面上标注页面信息的点击行为和 / 或抽取行为 ;

[0026] 细化配置子模块,用于对所述点击行为或抽取行为进行细化配置。

[0027] 优选的,所述抽取配置模块还包括 :

[0028] 元素类型选择子模块,用于配置所述点击行为或抽取行为的操作对象是单一信息还是列表信息 ;

[0029] 如果是单一信息,则触发所述细化配置子模块针对该单一信息进行点击行为或抽取行为的细化配置 ;

[0030] 如果是列表信息,则配置列表属性,并触发所述细化配置子模块从列表中选取样例进行点击行为或抽取行为的细化配置。

[0031] 优选的,每个点击行为在触发页面跳转时都会产生一个新的标注页面 ;最初的标

注页面为起始页面，基于起始页面产生的所有标注页面的集合构成一棵以该起始页面为根的标注树，所有的起始页面代表的标注树构成一个标注森林；所述网页信息抽取任务为一个标注森林或一棵标注树或一个标注页面。

[0032] 优选的，所述系统还包括：

[0033] 抓取器，用于根据调度器的调度，依据标注页面的 URL 抓取页面数据，并返回给所述调度器；

[0034] 渲染引擎，用于根据调度器的调度，渲染抓取回来的标注页面，并获得标注页面的 DOM 树结构，返回给所述调度器；

[0035] 抽取器，用于根据调度器的调度，依据网页信息抽取任务的配置抽取相应的网页信息；

[0036] 所述调度器通过解析网页信息抽取任务获得每个标注页面的配置；

[0037] 所述调度器依据标注页面的 URL 调度抓取器抓取页面数据；

[0038] 所述调度器调度渲染引擎渲染标注页面，并获得标注页面的 DOM 树结构；

[0039] 所述调度器遍历标注页面的 DOM 树结构中对应节点的配置，并依据所述节点的配置进行如下操作：

[0040] 如果是抽取行为，则依据抽取行为的配置调度抽取器抽取文本信息；

[0041] 如果是点击行为，并且如果是下载行为，则依据点击行为的配置调度抓取器抓取下载内容；如果是涉及渲染的点击行为，则依据点击行为的配置调度渲染引擎进行渲染。

[0042] 与现有技术相比，本申请包括以下优点：

[0043] 首先，本申请提供的网页信息抽取方法及系统可通过界面交互方式配置网页信息抽取任务，系统中的调度器通过解析网页信息抽取任务来自动进行信息抽取。所述界面交互的方式实现了简单的人机交互，对于非专业人员而言，也可以按照界面的提示进行任务配置，极大地降低了信息抽取的门槛。而且，调度器依据网页信息抽取任务执行的一套自动抽取方式，可以实现大批量网页高度自动化的信息抽取。

[0044] 其次，本申请的任务配置过程中不仅可以标注出网页中要抽取的文本信息，还可以模拟用户的点击行为进行配置，例如标注出网页中要抽取的链接进而下载该链接的内容，从而完成批量网页的抽取。而且，本申请还可以将网页 DOM 树中类似的兄弟节点配置为列表元素，实现对列表元素的自动化抽取。

[0045] 再次，本申请还支持网页 DOM 树中多个节点内容的信息抽取，因此可以精准地抽取信息。

[0046] 当然，实施本申请的任一产品不一定需要同时达到以上所述的所有优点。

附图说明

[0047] 图 1 是本申请实施例所述一种网页信息抽取方法的流程图；

[0048] 图 2 是本申请实施例中页面节点的示意图；

[0049] 图 3.1 至 3.4 是本申请实施例中通过界面交互方式配置网页信息抽取任务的示意图；

[0050] 图 4 是本申请实施例中通过界面交互方式配置网页信息抽取任务的流程图；

[0051] 图 5 是本申请实施例中抽取配置的示意图；

[0052] 图 6 是本申请实施例中点击行为配置的示意图；

[0053] 图 7 是本申请实施例中列表元素配置的示意图；

[0054] 图 8 是本申请实施例中网页信息抽取的示意图；

[0055] 图 9 是本申请实施例所述一种网页信息抽取系统的结构图。

具体实施方式

[0056] 为使本申请的上述目的、特征和优点能够更加明显易懂，下面结合附图和具体实施方式对本申请作进一步详细的说明。

[0057] 本申请提供了一种网页信息抽取方法及系统，可通过界面交互方式配置网页信息抽取任务，系统中的调度器通过解析网页信息抽取任务来自动进行信息抽取。本申请通过简单的人机交互，可实现针对互联网站点的信息抽取。

[0058] 下面通过实施例对本申请所述方法的实现流程进行详细说明。

[0059] 参照图 1，是本申请实施例所述一种网页信息抽取方法的流程图。

[0060] 步骤 101，通过界面交互方式配置网页信息抽取任务，并存入数据库；

[0061] 配置网页信息抽取任务目的是为了批量的抽取页面中有价值的内容。

[0062] 一方面，需要对抓取器 (spider) 进行配置，使其抓取指定的页面集合。例如，需要抓取某站点的商品信息，其中：

[0063] http://www.360buy.com/product/342890.html 这类页面是要进行抽取的页面；

[0064] http://help.360buy.com/help/question-65.html 这类页面是无意义的页面。

[0065] 另一方面，还需要配置每个页面上具体要抽取的内容。例如，要抽取某段文字，或者抽取某个新闻标题，等等。具体的配置方法将在下面的图 2 至图 7 中进行详细说明。

[0066] 需要说明的是，本申请实施例中，所述配置是通过界面交互的方式完成，即用户可以根据界面的提示进行一步步地输入选择，无需手动输入正则表达式，因此操作起来十分简便，而且配置的自动化程度较高，可以快速完成配置。

[0067] 步骤 102，监控数据库，当发现数据库中存入新的网页信息抽取任务后，将所述新的网页信息抽取任务发送给调度器；

[0068] 可设置一监控程序实时监控数据库的变化，并及时将放入数据库的新任务发送给调度器。所述调度器用于按照网页信息抽取任务自动化抽取所配置的页面信息。

[0069] 步骤 103，调度器解析网页信息抽取任务，并依据解析结果自动执行所述网页信息抽取任务。

[0070] 所述调度器主要通过调度各种处理器执行抽取任务，所述处理器包括抓取器 (spider)、javascrpit 渲染引擎 (简称 JS 渲染引擎) 和抽取器 (extractor)。其中，抓取器 (spider) 主要用于抓取指定的页面，JS 渲染引擎主要用于对抓取的页面进行 javascrpit 处理，抽取器 (extractor) 主要用于根据配置进行信息抽取。整个调度执行过程将在下面的图中 8 进行详细说明。

[0071] 为了使本领域技术人员更加了解本申请的内容，下面通过图 2 至图 8 对上述内容进行更详细的解释说明。

[0072] 1. 网页信息抽取任务的配置

[0073] 首先，介绍网页的页面结构。

[0074] 目前，通常采用 DOM 树来描述网页的页面结构。DOM 全称是 Document Object Model，即文档对象模型。DOM 是一种用于 HTML 和 XML 文档的编程接口，它给文档提供了一种结构化的表示方法，可以改变文档的内容和呈现方式。

[0075] 例如，参照图 2，是本申请实施例中页面节点的示意图。页面 http://news.sina.com.cn/c/2011-06-13/133822631625.shtml 中的新闻由多个节点的内容组成，其中一个节点及其对应的内容如图所示。

[0076] 本申请实施例可支持多个节点内容的信息抽取，这样可以更加精准地抽取信息。下面先通过一个简单的例子说明对某个节点进行配置的过程。

[0077] 本申请实施例采用标注方式完成配置，标注就是在浏览页面的过程中，将需要抽取的内容标记出来。参照图 3.1 至 3.4，是本申请实施例中通过界面交互方式配置网页信息抽取任务的示意图。对网页中某个节点的配置过程如下：

[0078] 1) 提交入口 URL 进入标注页面；

[0079] 参照图 3.1，输入 URL 进入该 URL 指向的页面；

[0080] 2) 打开了新的页面后，点击要进行标注的信息；

[0081] 参照图 3.2，点击图中用框线框起来的链接，该链接的标题是“陕西关中－天水经济区生产总值高出全国平均水平”；

[0082] 3) 弹出窗口选择动作；

[0083] 参照图 3.3，选择是抽取该链接的文字，还是点击该链接；

[0084] 4) 之后对指定的动作进行配置。

[0085] 参照图 3.4，若选择的动作是抽取该链接的文字，则对文字抽取进行配置，如配置名称是“新闻标题”等。

[0086] 从上述例子的配置过程可以看出，通过界面交互方式配置网页信息抽取任务的过程主要包括以下几步：

[0087] 第一，提交标注页面；

[0088] 第二，在所述标注页面上标注页面信息的点击行为和 / 或抽取行为；

[0089] 其中，所述“和”是指可以在同一个页面上既标注点击行为，又标注抽取行为；所述“或”是指在同一个页面上或者标注点击行为，或者标注抽取行为。

[0090] 在实际应用中，一般的标注是对网页中的一些文本或链接的标题等信息标注为抽取行为。而本申请实施例优选的，不仅可以标注出网页中要抽取的文本信息，还可以模拟用户的点击行为进行配置。所述点击行为包括：

[0091] 1) 下载行为，即按照用户浏览的行为标注出网页中要点击的链接，进而下载该链接的内容；

[0092] 2) 其他点击行为，如发生在一些按钮或选择框中的点击操作，通过模拟这些用户行为，就可以提交表单登录、上传文件或触发 javascrpit。

[0093] 上述点击行为和抽取行为都称为标注行为，每一个标注行为在页面 DOM 树中都能找到与之对应的节点。例如，图 3.2 中选择配置的链接就对应一个 DOM 节点，对这个链接配置的是抽取链接的文字，当然，也可以配置下载这个链接的内容。配置过程中，可以从预览的窗口中查看当前的抽取结果和爬取的路径，如果发现结果不准确，还可以随时调整配置。

[0094] 第三，对所述点击行为或抽取行为进行细化配置，如配置细化的点击属性或配置

细化的抽取规则等。

[0095] 此外，本申请实施例优选的，还可以将网页 DOM 树中类似的兄弟节点配置为列表元素，实现对列表元素的自动化抽取。例如，参照图 3.2 所示，在框线的下方还列出了多条链接，这些链接相互之间都是兄弟节点，因此在配置过程中可以将这些链接设为列表元素。

[0096] 基于上述配置方法，下面通过图 4 说明具体的配置过程。

[0097] 参照图 4，是本申请实施例中通过界面交互方式配置网页信息抽取任务的流程图。

[0098] 步骤 401，提交标注页面 ；

[0099] 如以图 3.1 的方式提交标注页面。

[0100] 步骤 402，通过点击或划选的方式与界面进行交互 ；

[0101] 通常，对于链接可进行点击交互，如图 3.2 就是点击该链接然后弹出配置窗口。而对于文本内容可进行划选，所述划选相当于按住并拖动鼠标进行文本的选中操作。

[0102] 通过以上点击或划选的操作，界面会根据这些操作弹出相应的配置窗口，供用户进行下一步的配置。

[0103] 步骤 403，选择是进行抽取操作还是动作操作 ；

[0104] 所述抽取操作是指抽取文本信息或抽取链接，所述动作操作是指模拟用户的点击行为，如前所述，点击行为包括下载行为和点击按钮或选择框等其他点击行为。

[0105] 需要说明的是，如果步骤 402 中通过点击方式进行交互，则在步骤 403 中可以选择抽取操作也可以选择动作操作 ；如果步骤 402 中通过划选方式进行交互，则在步骤 403 中只能选择抽取操作。例如，对于页面中的一条链接，通过点击该链接弹出配置窗口，用户在该窗口中可以选择抽取该链接的文字，也可以选择下载该链接的内容。而对于一篇网页文本信息，通用户只能通过划选的方式选中某段内容进行抽取配置。

[0106] 步骤 404，选择操作单一元素还是列表元素 ；

[0107] 无论是抽取操作还是点击操作，都可以选择元素类型。所述元素类型包括单一元素（也称为单一信息）和列表元素（也称为列表信息），如前所述，列表元素对应着网页 DOM 树中类似的兄弟节点，而单一元素对应着 DOM 树中的一个节点。

[0108] 按照步骤 404 配置点击行为或抽取行为的操作对象是单一元素还是列表元素之后，如果是单一元素，则进入步骤 405 进行细化配置 ；如果是列表信息，则可以先配置列表属性，然后再进入步骤 405 进行细化配置。

[0109] 此外，选择列表元素还允许标注位于同一列表中的多个样例，后台利用这些样例的集合可以自动识别列表的范围，之后可以进行针对列表属性的相关配置。

[0110] 步骤 405，配置针对单一元素的规则。

[0111] 如果选择单一元素，则针对该单一元素进行点击行为或抽取行为的配置 ；如果选择列表元素，则针对列表中的样例进行点击行为或抽取行为的配置。

[0112] 对于抽取行为，细化配置具体的抽取规则 ；对于点击行为，细化配置点击动作的属性。

[0113] 下面通过 5 至图 7 举例说明细化的抽取配置、点击行为配置和列表元素配置。

[0114] 参照图 5，是本申请实施例中抽取配置的示意图。

[0115] 抽取配置如下 ：

[0116] 抽取链接

[0117] 是否抓取链接

[0118] 链接内容加工

[0119] alt 属性抽取

[0120] 抽取文本

[0121] 文本加工

[0122] 地址识别

[0123] 日期识别

[0124] 抽取的内容可以是文本，也可以是链接。在进行抽取配置时：

[0125] 如果选择抽取链接，则进入抽取链接配置页面，进一步配置“是否抓取链接”选项和“alt 属性抽取”选项，其中配置“是否抓取链接”时如果选择“是”，则还需要配置“链接内容加工”选项；

[0126] 如果选择抽取文本，则进入抽取文本配置页面，进一步配置“文本加工”、“地址识别”和“日期识别”三个选项。

[0127] 参照图 6，是本申请实施例中点击行为配置的示意图。

[0128] 点击行为配置如下：

[0129] 点击对象

[0130] 文本框

[0131] 输入文本或上传批量输入

[0132] 按钮

[0133] 触发表单提交事件

[0134] 选择框

[0135] 选取操作

[0136] 链接

[0137] 产生新的页面

[0138] 其他

[0139] 通用行为

[0140] 点击操作

[0141] 鼠标停留

[0142] 滚轮操作

[0143] 鼠标离开

[0144] 鼠标双击

[0145] 首先选择点击对象，点击对象可以是文本框、按钮、选择框、链接、其他对象和通用行为，然后对所选择的点击对象进行具体配置。

[0146] 如果选择“文本框”，则进一步配置“输入文本或上传批量输入”选项；

[0147] 如果选择“按钮”，则进一步配置“触发表单提交事件”选项；

[0148] 如果选择“选择框”，则进一步配置“选取操作”选项；

[0149] 如果选择“链接”，则进一步配置“产生新的页面”选项；

[0150] 如果选择“通用行为”，则进一步配置“点击操作”、“鼠标停留”、“滚轮操作”、“鼠标离开”、“鼠标双击”这几个选项。

[0151] 参照图 7，是本申请实施例中列表元素配置的示意图。

[0152] 以抽取列表配置为例如下：

[0153] 制定偏移

[0154] 起始偏移

[0155] 结束偏移

[0156] 间隔

[0157] 制定条件

[0158] 指定字符序列开头

[0159] 指定字符序列结尾

[0160] 抽取列表配置包括两个选项：“制定偏移”和“制定条件”，对于“制定偏移”选项，进一步配置“起始偏移”、“结束偏移”和“间隔”的具体数值；对于“制定条件”选项，进一步配置“指定字符序列开头”和“指定字符序列结尾”。

[0161] 综上所述，基于上述对标注页面的配置，页面上的每一个标注行为（抽取、点击），在该页面的 DOM 树中都能找到与之对应的节点。页面上的每个标注动作，除了记录配置信息外，还记录了定位的信息。

[0162] 此外，由于每个点击行为在触发页面跳转时都会产生一个新的标注页面，因此可以把最初的标注页面称为起始页面，基于起始页面产生的所有标注页面的集合构成一棵以该起始页面为根的标注树，所有的起始页面代表的标注树构成一个标注森林。

[0163] 因此，一个标注森林包含多个标注树，一棵标注树中的每个节点对应一个标注页面，而每个标注页面都对应一个 DOM 树，DOM 树中的节点都有对应的标注行为。通常，选择对一棵标注树的根节点即起始页面，按照上述的方法进行配置，由于这颗树上的其他节点对应的标注页面都是基于该起始页面产生的标注页面，因此在配置起始页面的过程中，通过配置上述的点击行为和列表元素，就可以完成对其他标注页面的抽取配置。换而言之，通过配置起始页面就可以对基于该起始页面生成的标注树进行网页信息的自动抽取。

[0164] 配置一个网页信息抽取任务，可以选择多个起始页面，将包含这些起始页面的一个标注森林作为一个任务；也可以选择一个起始页面，将基于该起始页面生成的一棵标注树作为一个任务；甚至还可以简单地将一个标注页面作为一个任务。

[0165] 2. 网页信息抽取任务的调度执行

[0166] 下面通过图 8 详细说明调度执行过程。

[0167] 参照图 8，是本申请实施例中网页信息抽取的示意图。

[0168] 图 8 所示的完整的信息抽取过程如下：

[0169] 1) 用户通过 web 界面交互，配置网页信息抽取任务；

[0170] 2) 将配置的网页信息抽取任务存入数据库；

[0171] 3) 监控程序发现新任务，初始化任务相关环境，之后将任务信息发送给调度器；

[0172] 4) 调度器解析并调用相关的处理器完成信息抽取工作；

[0173] 5) 将最终抽取结果存入数据库，等待用户提交下载请求。

[0174] 如图所示，调度器调用的处理器包括：

[0175] ●抓取器 (spider)

[0176] 主要任务是根据 URL 请求和附加的 cookie、表单等信息，抓取相应的页面数据。其

输入、输出如下：

[0177] 输入：要抓取的网页 URL 和表单数据，所述表单数据指登录用户名、密码等信息；

[0178] 输出：抓取到的资源，如页面、图片、pdf 等文档、cookie、URL 所引用的 javascrpit 等。

[0179] ● javascrpit 渲染引擎

[0180] 主要任务是根据用户的行为，调用相应的 javascrpit，改变 DOM 树的结构或者跳转到新的页面。其输入、输出如下：

[0181] 输入：页面、页面引用的 javascrpit，这些输入信息是通过抓取器抓取得到；

[0182] 输出：渲染后的页面，其中可能包括 javascrpit 执行过的动作，如点击行为。

[0183] ●抽取器 (extractor)

[0184] 主要任务根据抽取的配置信息与抽取对象的位置信息，获取最终的抽取结果。其输入、输出如下：

[0185] 输入：页面内容、图片等资源，其输入可以是抽取器输出的内容，也可以是 javascrpit 渲染引擎输出的内容；

[0186] 输出：结构化文本，需要抽取的链接的 URL。

[0187] 以网页信息抽取是一个标注森林为例，所述调度器的处理流程如下：

[0188] 调度器遍历任务的标注树森林

[0189] 遍历每颗树中的节点

[0190] 判断当前节点的行为

[0191] 根据行为进行调度

[0192] 如前所述，每棵树是以起始页面为根、以点击产生的标注页面为节点的标注树，因此标注树中的每个节点对应一个标注页面。调度器根据每个标注页面上的抽取行为配置或点击行为配置，调度抓取器 (spider)、javascrpit 渲染引擎或抽取器 (extractor)。

[0193] 调度器调度相应处理器的过程如下：

[0194] 1) 调度器解析网页信息抽取任务，并获得起始页面的配置；

[0195] 调度器加载新任务的所有起始页面的 URL；

[0196] 2) 依据起始页面的 URL 调度抓取起始页面；

[0197] 调度器将起始页面 URL 传给 spider，spider 抓取完页面，并返回给调度器；

[0198] 3) 调度渲染起始页面，并获得起始页面的 DOM 树结构；

[0199] 调度器获得页面之后，调用 javascript 渲染引擎，获取经过 javascript 处理的 DOM 树；

[0200] 4) 遍历起始页面的 DOM 树结构中对应节点的配置，并依据所述节点的配置进行如下操作：

[0201] 如果是抽取行为，则依据抽取行为的配置调度抽取文本信息；

[0202] 如果是点击行为，并且如果是下载行为，则依据点击行为的配置调度抓取下载内容；如果是涉及渲染的点击行为，则依据点击行为的配置进行调度渲染。

[0203] 具体的：

[0204] 对于抽取行为，将 DOM 结构与抽取行为配置传递给抽取器 (extractor)，抽取器返回抽取到的文本结果；

[0205] 对于点击行为，如果触发下载行为，则调用 spider 下载内容，如果下载内容为新的页面，则将新的页面添加至调度器的处理序列中；如果触发其他点击行为，如果涉及到 javascrpit 的调用，则请求 javascrpit 渲染引擎返回。如果 javascrpit 的执行过程中，触发了 ajax(Asynchronous JavaScript And XML，异步 JavaScript 及 XML) 操作，则 javascrpit 渲染引擎通过调度器，请求 spider 下载对应的数据，之后继续 javascrpit 的渲染过程。

[0206] 此外，对于网页信息抽取任务中的其他标注页面，也同样按照调度处理流程进行抽取，详细的调度过程不再详述。

[0207] 由上可知，调度器对每个行为的调度处理并不一定按照抓取器 (spider)、javascrpit 渲染引擎、抽取器 (extractor) 的顺序进行调度，而是根据具体的行为配置可能调度抓取器 (spider)，可能调度 javascrpit 渲染引擎，或者可能调度抽取器 (extractor)。

[0208] 需要说明的是，对于前述的方法实施例，为了简单描述，故将其都表述为一系列的动作组合，但是本领域技术人员应该知悉，本申请并不受所描述的动作顺序的限制，因为依据本申请，某些步骤可以采用其他顺序或者同时进行。其次，本领域技术人员也应该知悉，说明书中所描述的实施例均属于优选实施例，所涉及的动作并不一定是本申请所必须的。

[0209] 综上所述，本申请实施例提供的网页信息抽取方法可通过简单的人机交互对信息抽取进行配置，并可以依据所述配置，在调度器的调度下自动化进行信息抽取，提高了信息抽取的自动化程度，可实现大批量网页高度自动化的信息抽取。而且，这种人机交互方式对于非专业人员而言，也可以按照界面的提示进行任务配置，极大地降低了信息抽取的门槛。

[0210] 进一步，本申请与现有的各种信息抽取方法相比，还具有以下特点和优势：

[0211] 第一，配置过程中不仅可以标注出网页中要抽取的文本信息，还可以模拟用户的点击行为进行配置，例如标注出网页中要抽取的链接进而下载该链接的内容；

[0212] 第二，本申请还可以将网页 DOM 树中类似的兄弟节点配置为列表元素，实现对列表元素的自动化抽取；

[0213] 第三，本申请还支持网页 DOM 树中多个节点内容的信息抽取，因此可以精准地抽取信息。

[0214] 基于上述方法实施例的说明，本申请还提供了相应的网页信息抽取系统实施例，来实现上述方法实施例所述的内容。

[0215] 参照图 9，是本申请实施例所述一种网页信息抽取系统的结构图。

[0216] 所述抽取系统可以包括抽取配置模块 91、数据库 92、监控模块 93 和调度器 94，其中，

[0217] 抽取配置模块 91，用于通过界面交互方式配置网页信息抽取任务，并存入数据库；

[0218] 数据库 92，用于存储网页信息抽取任务；

[0219] 监控模块 93，用于监控数据库，当发现数据库中存入新的网页信息抽取任务后，将所述新的网页信息抽取任务发送给调度器；

[0220] 调度器 94，用于解析网页信息抽取任务，并依据解析结果自动执行所述网页信息抽取任务。

[0221] 进一步优选的，所述抽取配置模块 91 具体可以包括 ：

[0222] 配置入口子模块，用于提交标注页面 ；

[0223] 行为标注子模块，用于在所述标注页面上标注页面信息的点击行为和 / 或抽取行为 ；

[0224] 细化配置子模块，用于对所述点击行为或抽取行为进行细化配置。

[0225] 优选的，所述抽取系统还可以将网页 DOM 树中类似的兄弟节点配置为列表元素，实现对列表元素的自动化抽取，因此所述抽取配置模块 91 还可以包括 ：

[0226] 元素类型选择子模块，用于配置所述点击行为或抽取行为的操作对象是单一信息还是列表信息 ；

[0227] 如果是单一信息，则触发所述细化配置子模块针对该单一信息进行点击行为或抽取行为的细化配置 ；

[0228] 如果是列表信息，则配置列表属性，并触发所述细化配置子模块从列表中选取样例进行点击行为或抽取行为的细化配置。

[0229] 此外，需要说明的是，每个点击行为在触发页面跳转时都会产生一个新的标注页面 ；最初的标注页面为起始页面，基于起始页面产生的所有标注页面的集合构成一棵以该起始页面为根的标注树，所有的起始页面代表的标注树构成一个标注森林 ；所述网页信息抽取任务为一个标注森林或一棵标注树或一个标注页面。

[0230] 进一步优选的，所述抽取系统还可以包括 ：

[0231] 抓取器 95，用于根据调度器 94 的调度，依据标注页面的 URL 抓取页面数据，并返回给所述调度器 94 ；

[0232] 渲染引擎 96，用于根据调度器 94 的调度，渲染抓取回来的标注页面，并获得标注页面的 DOM 树结构，返回给所述调度器 94 ；

[0233] 抽取器 97，用于根据调度器 94 的调度，依据网页信息抽取任务的配置抽取相应的网页信息 ；

[0234] 所述调度器 94 通过解析网页信息抽取任务获得每个标注页面的配置 ；

[0235] 所述调度器 94 依据标注页面的 URL 调度抓取器 95 抓取页面数据 ；

[0236] 所述调度器 94 调度渲染引擎 96 渲染标注页面，并获得标注页面的 DOM 树结构 ；

[0237] 所述调度器 94 遍历标注页面的 DOM 树结构中对应节点的配置，并依据所述节点的配置进行如下操作 ：

[0238] 如果是抽取行为，则依据抽取行为的配置调度抽取器 97 抽取文本信息 ；

[0239] 如果是点击行为，并且如果是下载行为，则依据点击行为的配置调度抓取器 95 抓取下载内容 ；如果是涉及渲染的点击行为，则依据点击行为的配置调度渲染引擎 96 进行渲染。

[0240] 进一步的，在渲染引擎 96 的执行过程中，如果触发了 ajax(Asynchronous JavaScript And XML，异步 JavaScript 及 XML) 操作，则渲染引擎 96 通过调度器 94，请求抓取器 95 下载对应的数据，之后继续渲染引擎 96 的渲染过程。

[0241] 综上所述，所述网页信息抽取系统实现了高度自动化的信息抽取，并且通过界面交互方式完成配置，实现了简单的人机交互，极大地降低了信息抽取的门槛。

[0242] 对于上述抽取系统实施例而言，由于其与方法实施例基本相似，所以描述的比较

简单，相关之处参见图 1 至图 8 所示方法实施例的部分说明即可。

[0243] 本说明书中的各个实施例均采用递进的方式描述，每个实施例重点说明的都是与其他实施例的不同之处，各个实施例之间相同相似的部分互相参见即可。

[0244] 以上对本申请所提供的一种网页信息抽取方法及抽取系统，进行了详细介绍，本文中应用了具体个例对本申请的原理及实施方式进行了阐述，以上实施例的说明只是用于帮助理解本申请的方法及其核心思想；同时，对于本领域的一般技术人员，依据本申请的思想，在具体实施方式及应用范围上均会有改变之处，综上所述，本说明书内容不应理解为对本申请的限制。

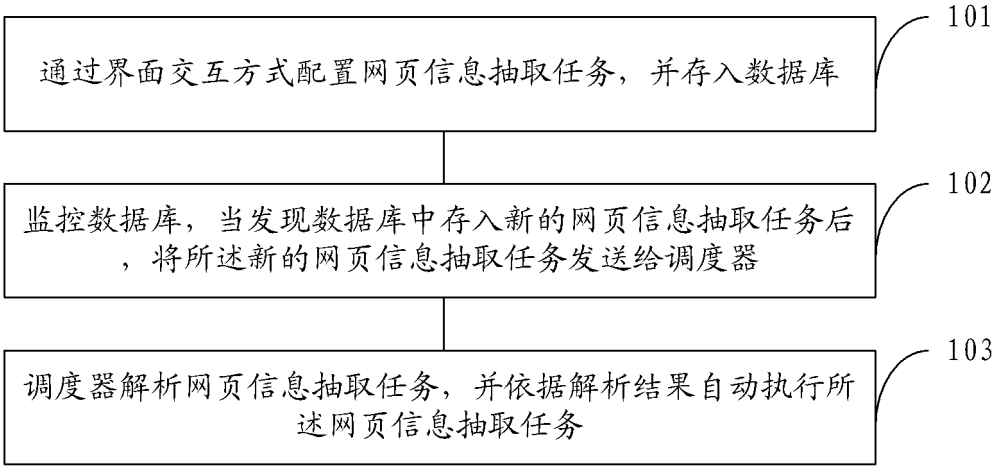
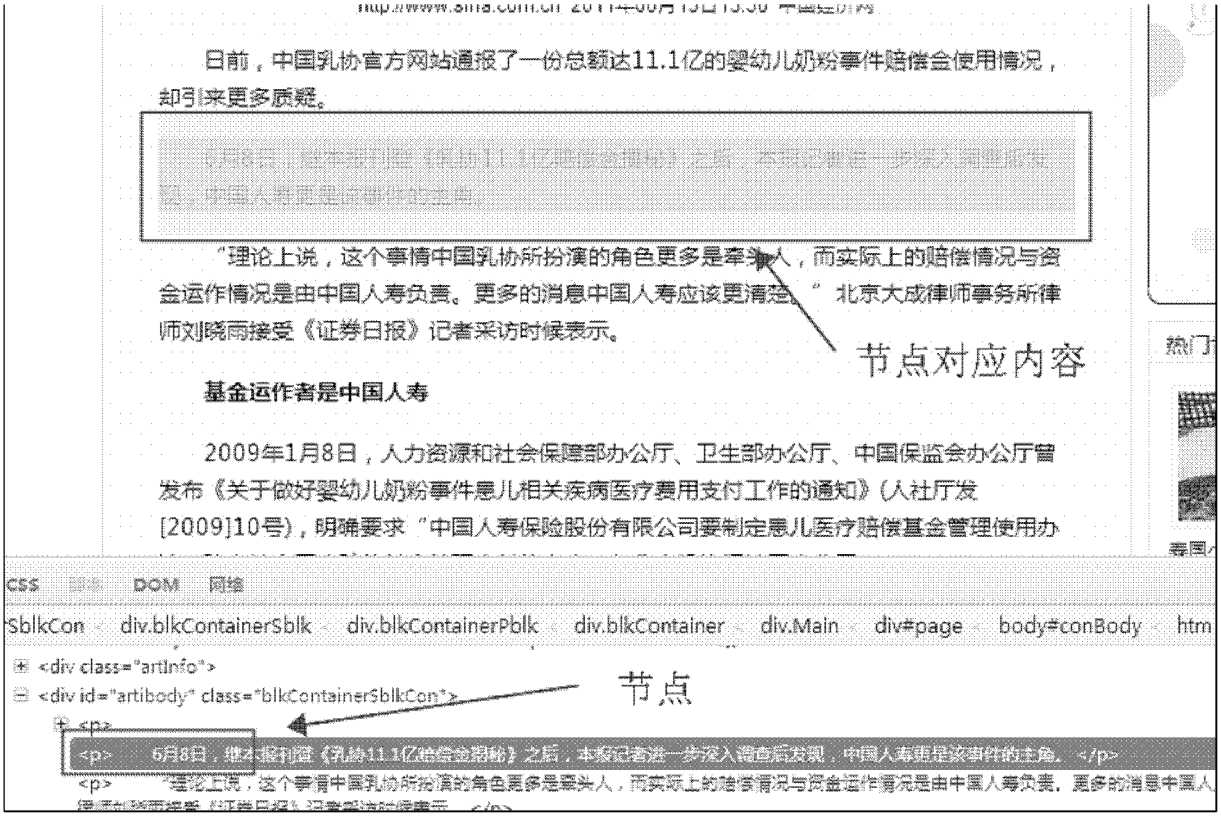

通过界面交互方式配置网页信息抽取任务，并存入数据库
101
监控数据库，当发现数据库中存入新的网页信息抽取任务后，将所述新的网页信息抽取任务发送给调度器
102
调度器解析网页信息抽取任务，并依据解析结果自动执行所述网页信息抽取任务
103

图 1


日前，中国乳协官方网站通报了一份总额达11.1亿的婴幼儿奶粉事件赔偿金使用情况，却引来更多质疑。
“理论上说，这个事情中国乳协所扮演的角色更多是牵头人，而实际上的赔偿情况与资金运作情况是由中国人寿负责。更多的消息中国人寿应该更清楚。”北京大成律师事务所律师刘晓雨接受《证券日报》记者采访时候表示。
节点对应内容
基金运作者是中国人寿
2009年1月8日，人力资源和社会保障部办公厅、卫生部办公厅、中国保监会办公厅曾发布《关于做好婴幼儿奶粉事件患儿相关疾病医疗费用支付工作的通知》(人社厅发[2009]10号)，明确要求“中国人寿保险股份有限公司要制定患儿医疗赔偿基金管理使用办
CSS
DOM
网络
SblkCon
div.blkContainerSblk
div.blkContainerPblk
div.blkContainer
div.Main
div#page
body#conBody
htm

节点

5月8日，继本报刊登《乳协11.1亿赔偿金揭秘》之后，本报记者进一步深入调查后发现，中国人寿是该事件的主角。

“理论上说，这个事情中国乳协所扮演的角色更多是牵头人，而实际上的赔偿情况与资金运作情况是由中国人寿负责。更多的消息中国人

图 2


Enter a URL:
/news.sina.com.cn/china/
Go »

图 3.1

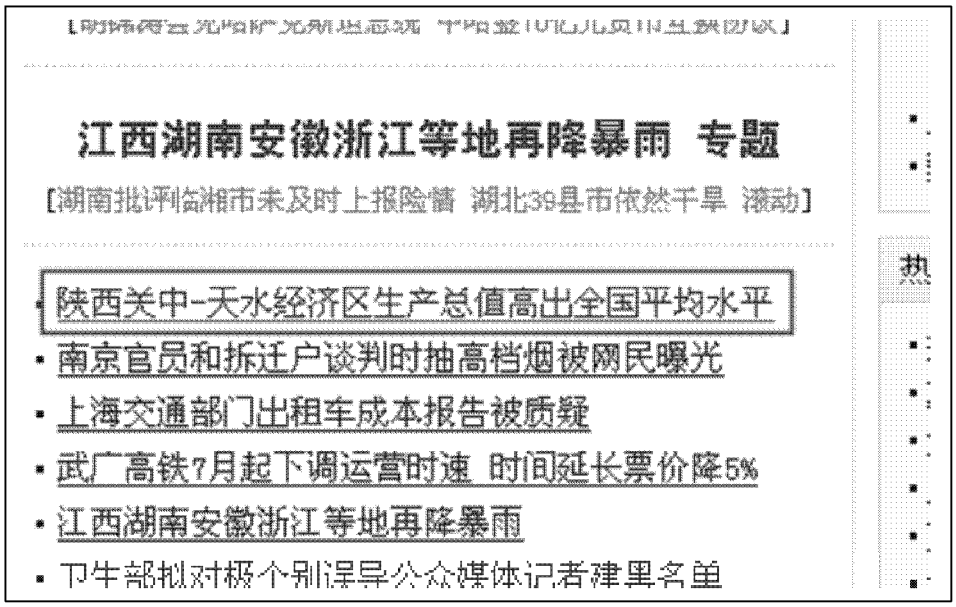

江西湖南安徽浙江等地再降暴雨 专题
【湖南批评临湘市未及时上报险情 湖北39县市依然干旱 滞动】
陕西关中-天水经济区生产总值高出全国平均水平
南京官员和拆迁户谈判时抽高档烟被网民曝光
上海交通部门出租车成本报告被质疑
武广高铁7月起下调运营时速 时间延长票价降5%
江西湖南安徽浙江等地再降暴雨
热

图 3.2

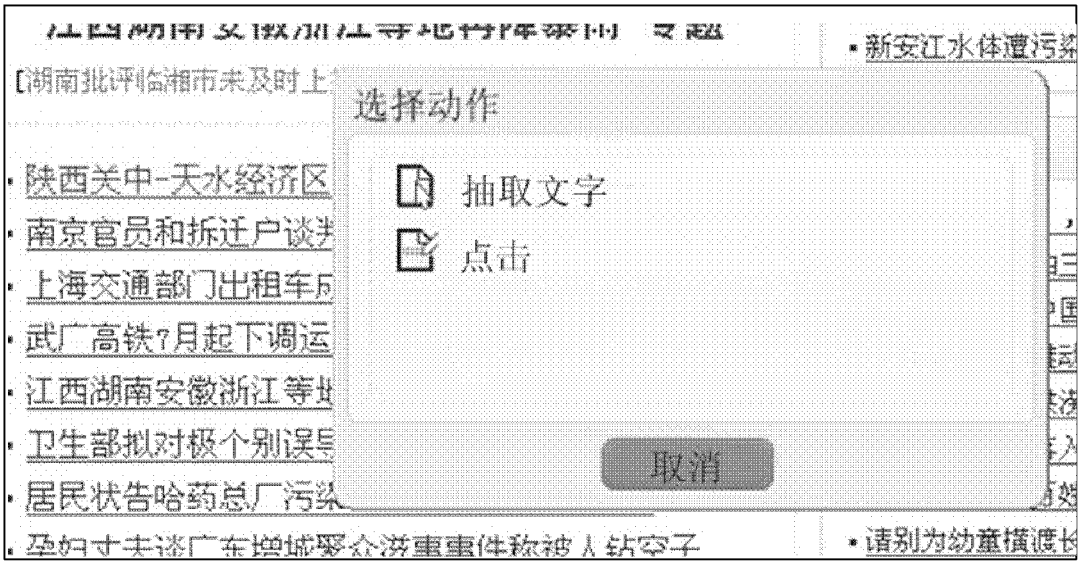

选择动作
抽取文字
点击
取消

图 3.3

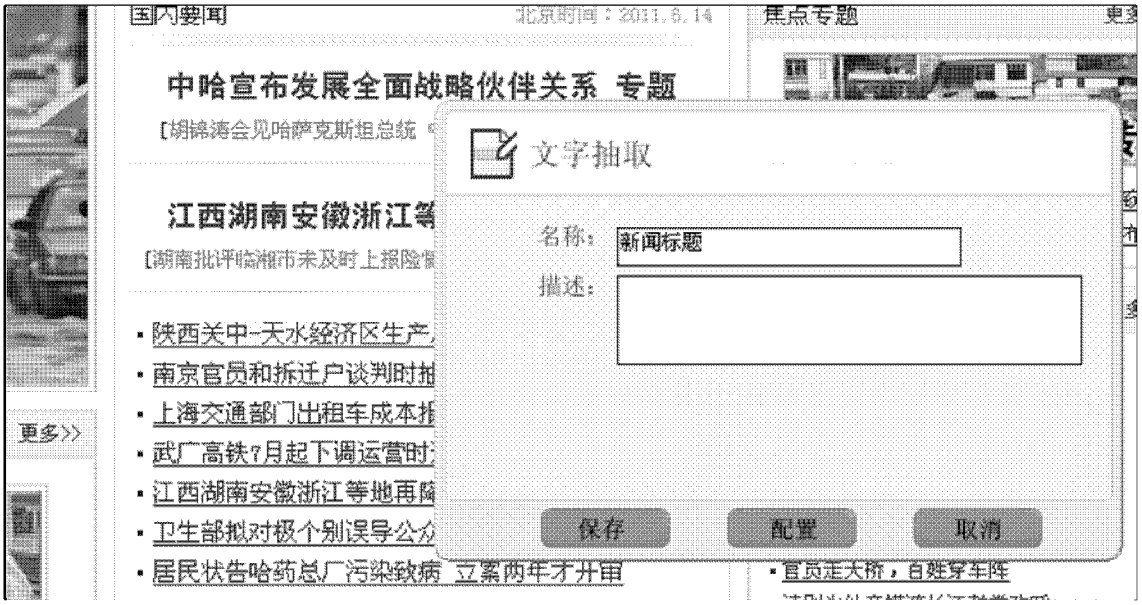

国内要闻
北京时间：2011.6.14
焦点专题
中哈宣布发展全面战略伙伴关系 专题
江西湖南安徽浙江等
文字抽取
名称：
新闻标题
描述：
保存
配置
取消
更多>>

图 3.4

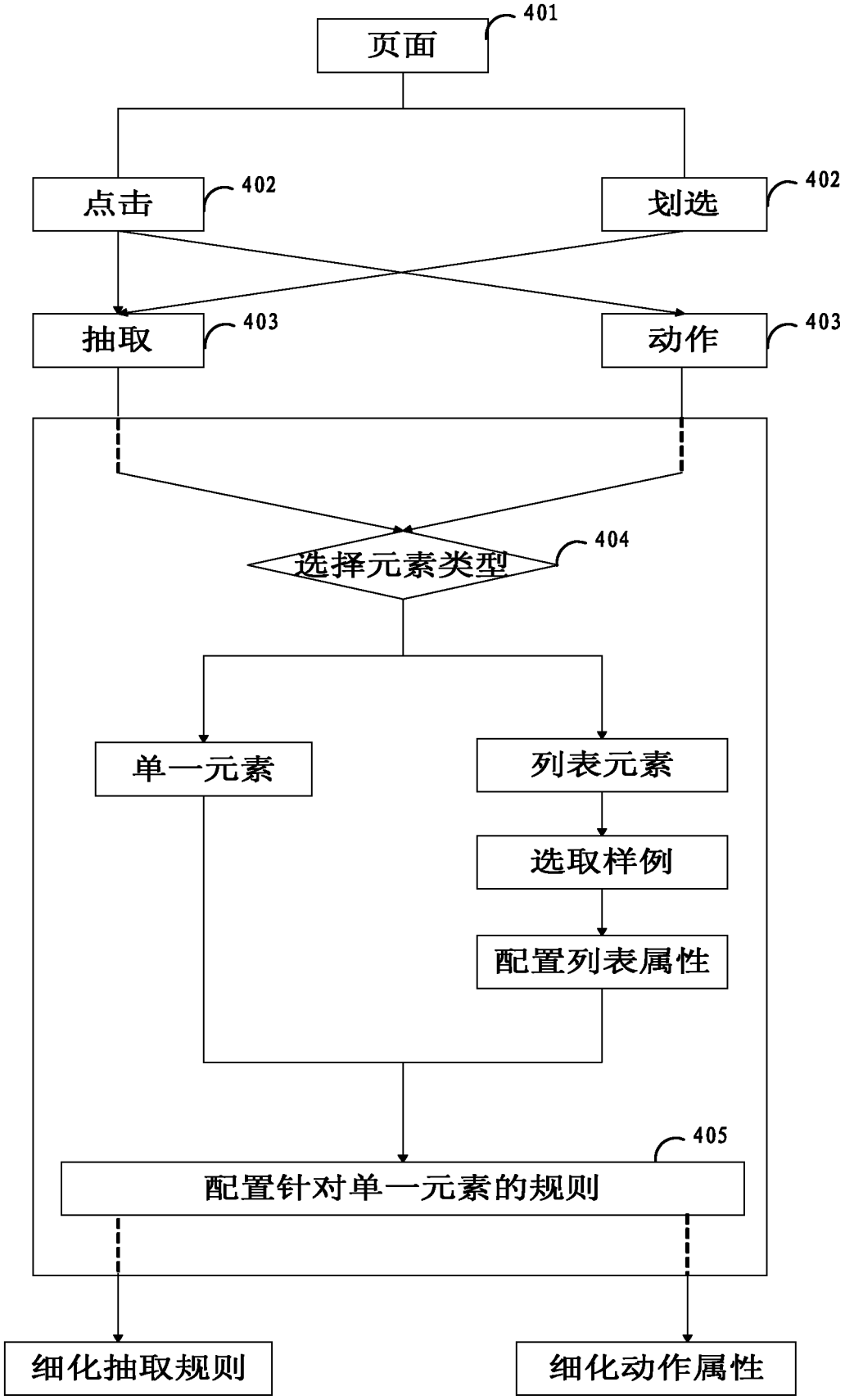

401
页面
402
点击
402
划选
403
抽取
403
动作
404
选择元素类型
单一元素
列表元素
选取样例
配置列表属性
405
配置针对单一元素的规则
细化抽取规则
细化动作属性

图 4

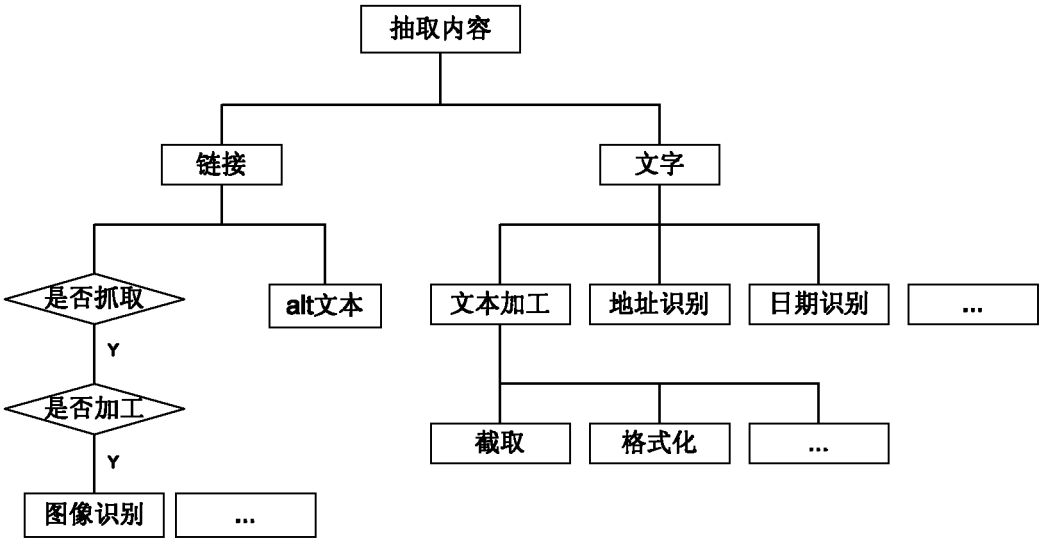

抽取内容
链接
文字
是否抓取
alt文本
文本加工
地址识别
日期识别
...
Y
是否加工
截取
格式化
...
Y
图像识别
...

图 5

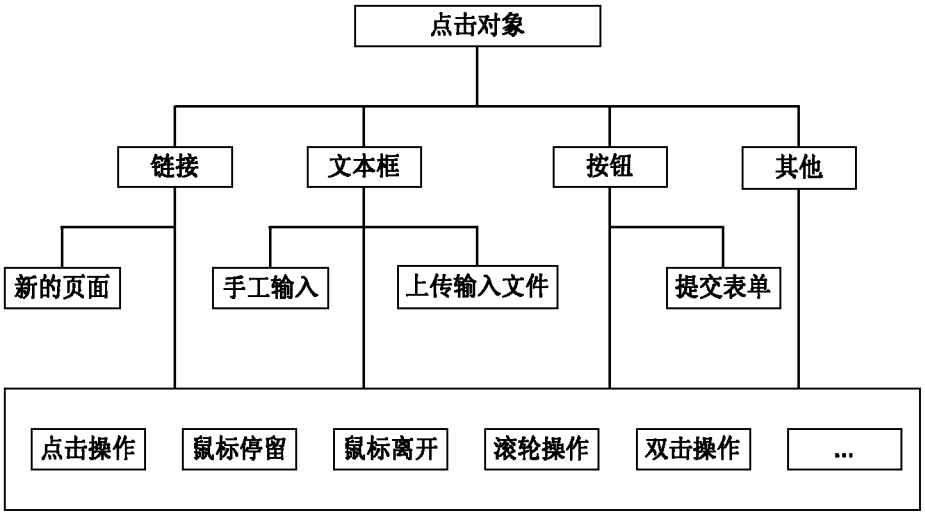

点击对象
链接
文本框
按钮
其他
新的页面
手工输入
上传输入文件
提交表单
点击操作
鼠标停留
鼠标离开
滚轮操作
双击操作
...

图 6

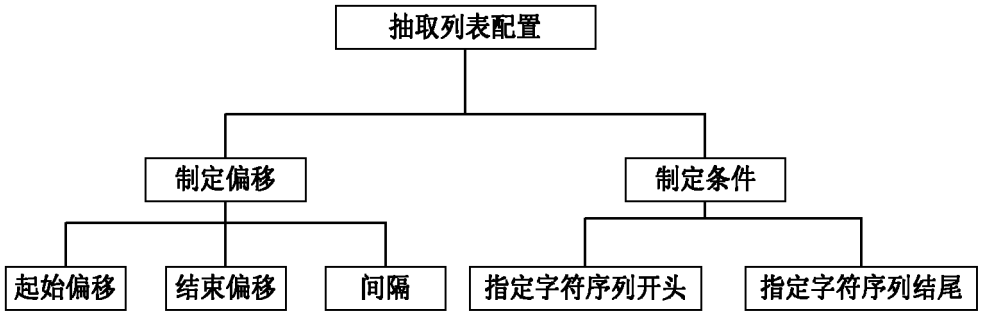

抽取列表配置
制定偏移
制定条件
起始偏移
结束偏移
间隔
指定字符序列开头
指定字符序列结尾

图 7

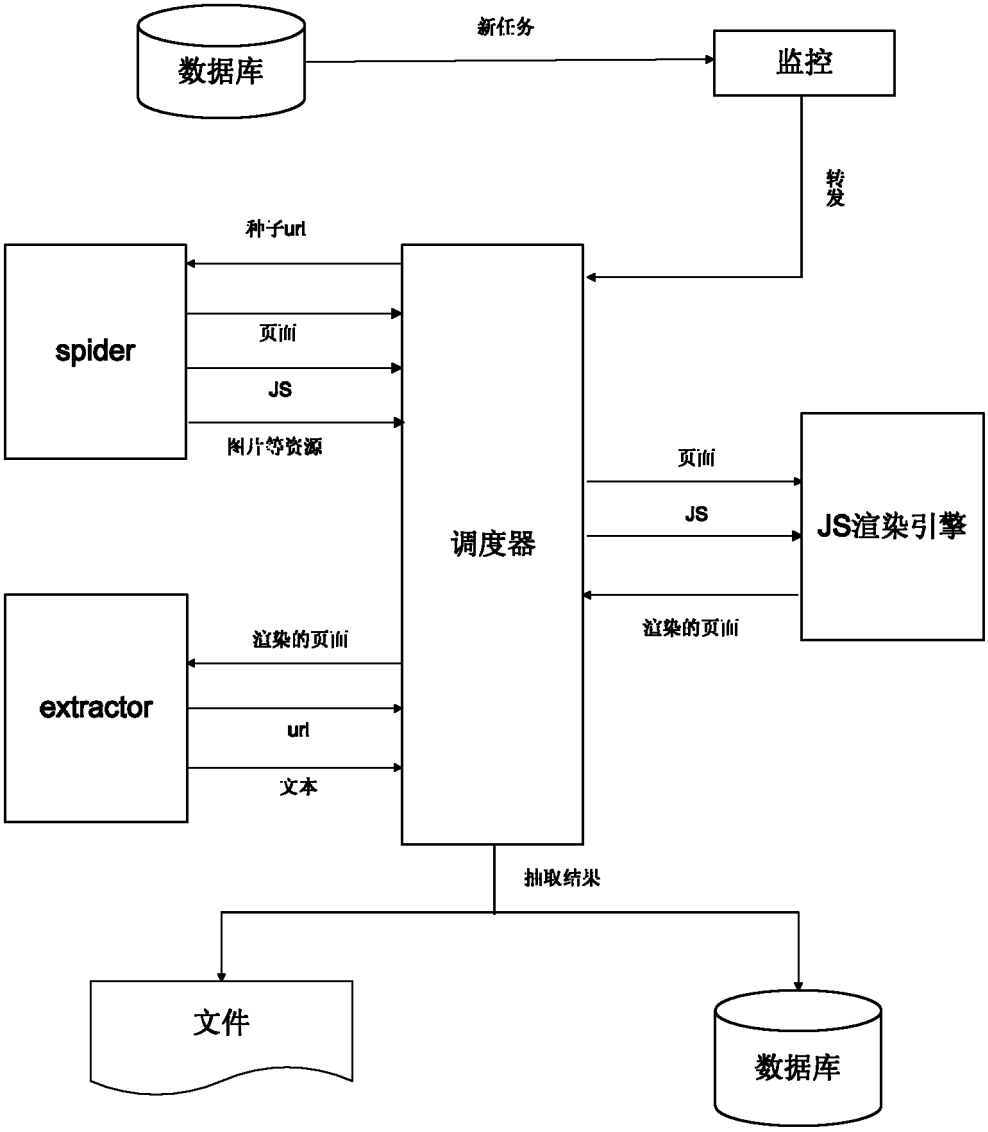

数据库
新任务
监控
转发
种子url
spider
页面
JS
图片等资源
调度器
页面
JS
JS渲染引擎
渲染的页面
extractor
渲染的页面
url
文本
抽取结果
文件
数据库

图 8

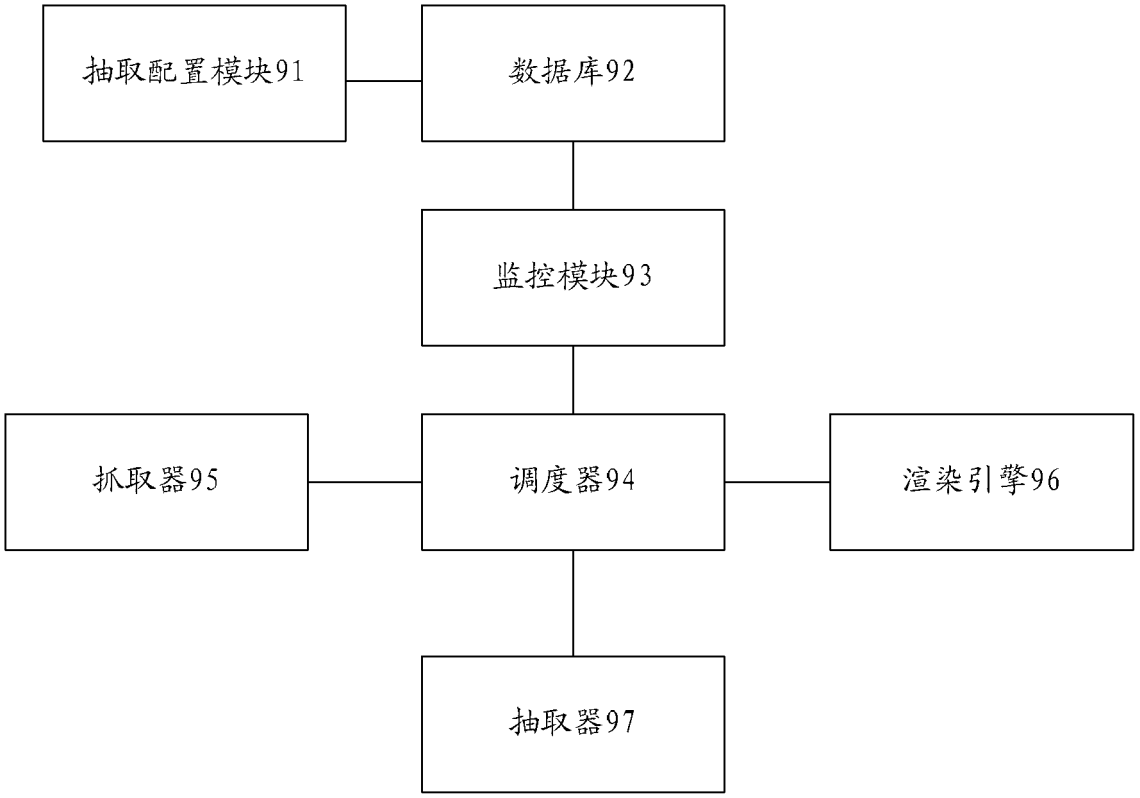


图 9