(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 108959539 B
(45)授权公告日 2021.09.21

(21)申请号 201810701727.0
(22)申请日 2018.06.30
(65)同一申请的已公布的文献号
申请公布号 CN 108959539 A
(43)申请公布日 2018.12.07
(73)专利权人 成都信息工程大学
地址 610225 四川省成都市西南航空港经济开发区学府路1段24号
(72)发明人 曹亮 罗山城
(74)专利代理机构 成都金英专利代理事务所(普通合伙) 51218
代理人 袁英
(51)Int.Cl.
G06F 16/953(2019.01)
G06F 16/95(2019.01)

(56)对比文件
CN 106959995 A,2017.07.18
CN 107885777 A,2018.04.06
CN 106528769 A,2017.03.22
CN 107391757 A,2017.11.24
CN 103399908 A,2013.11.20
CN 105468664 A,2016.04.06
CN 102890692 A,2013.01.23
CN 107025296 A,2017.08.08
US 2009182788 A1,2009.07.16
李世忠.一种智能网页数据采集系统设计.《电子技术与软件工程》.2018,169.

审查员 张雪锋

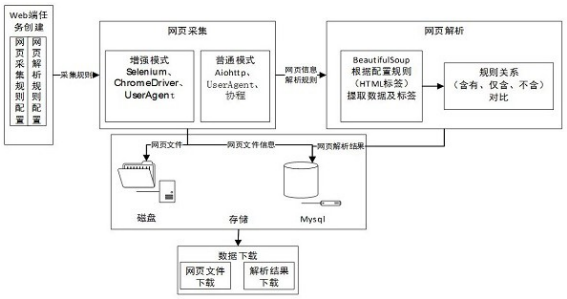
权利要求书1页 说明书3页 附图1页

(54)发明名称

一种基于规则可配置的网页数据解析方法

(57)摘要

本发明公开了一种基于规则可配置的网页数据解析方法,包括以下步骤:S1.Web端任务创建:Web应用程序向服务器端发送数据请求,任务配置信息填写完成后提交配置的信息;S2.网页采集:获取Web中通过任务配置配置的采集信息,后台根据传入的URL开始进行网页的抓取;S3.网页解析:获取Web中通过任务配置配置的解析信息,并获取采集网页后的列表信息进行数据解析;S4.数据下载:通过任务列表查看任务结果,在任务结果中可对采集的网页内容进行下载,也可对解析的数据进行查看和下载。本发明使用B/S架构的方式,方便使用,在对网页进行采集以及网页数据解析配置时,不需要进行大量操作。还可以方便的对网页中动态数据进行获取,并且使用协程,可以快速获取网页。



CN 108959539 B

1.一种基于规则可配置的网页数据解析方法,其特征在于,包括以下步骤:

S1. Web端任务创建:Web应用程序向服务器端发送数据请求,在任务配置页面配置所需网页起始URL、网页采集规则和网页解析规则,接着通过配置数据所属的HTML标签进行数据的提起,任务配置信息填写完成后提交配置的信息;所述网页采集规则包括是否采集子页、是否采集下一页和是否使用增强模式;

S2. 网页采集:获取Web中通过任务配置配置的采集信息,后台根据传入的URL开始进行网页的抓取,根据配置的网页采集规则确定抓取方式,所述抓取方式包括增强模式和普通模式,所述增强模式结合使用Selenium与ChromeDriver,以及使用Python的UserAgent库构造访问头的方式去访问对应的URL,所述普通模式使用Python的aiohttp库和UserAgent库构造访问头的方式去访问对应的URL;访问成功完成后,将网页信息以及URL、页数以及页面等级保存到列表中;当网页都访问完成后,将抓取到网页信息以HTML文件的形式存入到服务器中,并将对应信息存入数据库;所述步骤S2选用增强模式进行网页采集时,如果需要抓取子页,则会打开两个ChromeDriver,一个进行一级页面的访问,另一个进行子页页面的访问;该过程为访问一个一级页面后,通过配置的标签信息,获取到该一级页面的子页URL链接后,对子页进行访问;如果还需要抓取下一页,则通过配置的下一页标签获取到下一页链接进行访问;

S3. 网页解析:获取Web中通过任务配置配置的解析信息,并获取采集网页后的列表信息进行数据解析,通过的Python的BeautifulSoup库进行页面的解析;在解析时根据页面配置的HTML标签,以标签类型和值方式提取数据及相关标签;解析结束后,将数据存入数据库中;

S4. 数据下载:通过任务列表查看任务结果,在任务结果中可对采集的网页内容进行下载,也可对解析的数据进行查看和下载。

2.根据权利要求1所述的一种基于规则可配置的网页数据解析方法,其特征在于,所述步骤S1的网页解析规则最多为三行,每一行的网页解析规则单独去解析网页,最终合并为结果,并将结果存储到数据库中。

3.根据权利要求2所述的一种基于规则可配置的网页数据解析方法,其特征在于,所述网页解析规则包括四个参数,其中第一个参数用于选择网页解析规则,第二个参数和第四个参数为网页解析规则对应的配置信息,第三个参数为第二个参数配置信息与第四个参数配置信息的关系,所述关系为含有、不含和仅含中的一种。

4.根据权利要求1所述的一种基于规则可配置的网页数据解析方法,其特征在于,所述步骤S2选用普通模式进行网页采集时,如果需要抓取子页,则先访问一级页面,然后通过配置的标签信息,获取子页链接保存到列表中,再使用协程的方式去访问子页;如果还需要抓取下一页,则通过配置的下一页标签获取到下一页链接进行访问。

一种基于规则可配置的网页数据解析方法

技术领域

[0001] 本发明属于网页数据处理领域，尤其涉及一种基于规则可配置的网页数据解析方法。

背景技术

[0002] 近年来，随着国内大数据战略越来越清晰，数据抓取和信息采集系列产品迎来了巨大的发展机遇，采集产品数量也出现迅猛增长。网页解析，即程序自动分析网页内容、获取信息，从而进一步处理信息，网页解析是实现网络爬虫中不可缺少而且十分重要的一环。但是，目前的网页数据解析方法在对网页数据解析配置时，操作复杂；或是在对网页中的动态数据获取时，速度较慢。

发明内容

[0003] 为了解决上述问题，本发明提出一种基于规则可配置的网页数据解析方法，包括以下步骤：

[0004] S1．Web端任务创建：Web应用程序向服务器端发送数据请求，在任务配置页面配置所需网页起始URL、网页采集规则和网页解析规则，接着通过配置数据所属的HTML标签进行数据的提起，任务配置信息填写完成后提交配置的信息；

[0005] S2．网页采集：获取Web中通过任务配置配置的采集信息，后台根据传入的URL开始进行网页的抓取，根据配置的网页采集规则确定抓取方式，所述抓取方式包括增强模式和普通模式，所述增强模式结合使用Selenium与ChromeDriver，以及使用Python的UserAgent库构造访问头的方式去访问对应的URL，所述普通模式使用Python的aiohttp库和UserAgent库构造访问头的方式去访问对应的URL；访问成功完成后，将网页信息以及URL、页数以及页面等级保存到列表中；当网页都访问完成后，将抓取到网页信息以HTML文件的形式存入到服务器中，并将对应信息存入数据库；

[0006] S3．网页解析：获取Web中通过任务配置配置的解析信息，并获取采集网页后的列表信息进行数据解析，通过的Python的BeautifulSoup库进行页面的解析；在解析时根据页面配置的HTML标签，以标签类型和值方式提取数据及相关标签；解析结束后，将数据存入数据库中；

[0007] S4．数据下载：通过任务列表查看任务结果，在任务结果中可对采集的网页内容进行下载，也可对解析的数据进行查看和下载。

[0008] 进一步地，所述步骤S1的网页采集规则包括是否采集子页、是否采集下一页和是否使用增强模式。

[0009] 进一步地，所述步骤S1的网页解析规则最多为三行，每一行的网页解析规则单独去解析网页，最终合并为结果，并将结果存储到数据库中。

[0010] 再进一步地，所述网页解析规则包括四个参数，其中第一个参数用于选择网页解析规则，第二个参数和第四个参数为网页解析规则对应的配置信息，第三个参数为第二个

参数配置信息与第四个参数配置信息的关系,所述关系为含有、不含和仅含中的一种。

[0011] 进一步地,所述步骤S2选用增强模式进行网页采集时,如果需要抓取子页,则会打开两个ChromeDriver,一个进行一级页面的访问,另一个进行子页页面的访问;该过程为访问一个一级页面后,通过配置的标签信息,获取到该一级页面的子页URL链接后,对子页进行访问;如果还需要抓取下一页,则通过配置的下一页标签获取到下一页链接进行访问。

[0012] 进一步地,所述步骤S2选用普通模式进行网页采集时,如果需要抓取子页,则先访问一级页面,然后通过配置的标签信息,获取子页链接保存到列表中,再使用协程的方式去访问子页;如果还需要抓取下一页,则通过配置的下一页标签获取到下一页链接进行访问。

[0013] 本发明的有益效果在于:

[0014] 1)本方法使用B/S架构的方式,免去C/S构架客户端的下载,方便使用;

[0015] 2)本方法在对网页进行采集以及网页数据解析配置时,仅需了解HTML结构既可进行配置,且在配置时不需要进行大量操作;

[0016] 3)本方法可以方便的对网页中动态数据进行获取,并且使用协程,可以快速的获取网页。

附图说明

[0017] 图1是一种基于规则可配置的网页数据解析方法的流程图。

具体实施方式

[0018] 为了对本发明的技术特征、目的和效果有更加清楚的理解,现对照附图说明本发明的具体实施方式。

[0019] 本发明提出一种基于规则可配置的网页数据解析方法,如图1所示,具体如下:

[0020] 首先,开启Win10环境下的服务器端,监听指定端口,等待Socket连接。

[0021] 然后,创建Web端任务,Web应用程序向服务器端发送数据请求。此步骤中,在任务配置页面配置所需网页起始URL、网页采集规则和网页解析规则,然后通过配置数据所属的HTML标签进行数据的提起,任务配置信息填写完成后提交配置的信息。在这一步骤中,所述网页采集规则包括是否采集子页、是否采集下一页和是否使用增强模式,具体的:

[0022] 1)当选择采集子页时,必须配置“获取子页标签”,该标签为HTML标签形式:<a class=“xxx”>,后台会根据此标签寻找该一级页面中符合该标签的所有链接进行访问;

[0023] 2)当选择采集下一页时,需要配置“获取下一页的标签”,该标签为HTML标签形式:下一页,后台会根据该标签寻找对应的下一页链接进行访问;

[0024] 3)增强模式用于准确地获取动态网页,选用增强模式会使用Selenium与ChromeDriver结合的方式去访问网页。

[0025] 此外,所述网页解析规则最多为三行,每一行的网页解析规则单独去解析网页,其中,每一行网页解析规则包括四个参数,第一个参数用于选择网页解析规则,第二个参数和第四个参数为网页解析规则对应的配置信息,第三个参数为第二个参数配置信息与第四个参数配置信息的关系,所述关系为含有、不含和仅含中的一种。

[0026] 在规则的配置中,还可以增加正则表达式规则。

[0027] 接着,创建好Web端任务之后,开始进行网页采集。获取Web中通过任务配置配置的

采集信息,后台根据传入的URL开始进行网页的抓取,根据配置的网页采集规则确定抓取方式,所述抓取方式包括增强模式和普通模式,具体的:

[0028] 1)所述增强模式结合使用Selenium与ChromeDriver,以及使用Python的UserAgent库构造访问头的方式去访问对应的URL,如果需要抓取子页,则会打开两个ChromeDriver,一个进行一级页面的访问,另一个进行子页页面的访问。该过程为访问一个一级页面后,通过配置的标签信息,获取到该一级页面的子页URL链接后,对子页进行访问;如果还需要抓取下一页,则通过配置的下一页标签获取到下一页链接进行访问。特别地,一级页面设定最多抓取10页;

[0029] 2)所述普通模式使用Python的aiohttp库和UserAgent库构造访问头的方式去访问对应的URL,如果需要抓取子页,则先访问一级页面,然后通过配置的标签信息,获取子页链接保存到列表中,再使用协程的方式去访问子页;如果还需要抓取下一页,则通过配置的下一页标签获取到下一页链接进行访问。特别地,一级页面设定最多抓取10页。

[0030] 当访问成功完成后,将网页信息以及URL、页数以及页面等级保存到列表中。当网页都访问完成后,将抓取到网页信息以HTML文件的形式存入到服务器中,并将对应信息存入数据库。

[0031] 再接着,进行网页解析。此步骤获取Web中通过任务配置配置的解析信息,并获取采集网页后的列表信息进行数据解析,通过的Python的BeautifulSoup库进行页面的解析;在解析时根据页面配置的HTML标签,以标签类型和值方式提取数据及相关标签;解析结束后,将数据存入数据库中。

[0032] 最后是数据下载。通过任务列表查看任务结果,在任务结果中可对采集的网页内容进行下载,也可对解析的数据进行查看和下载。

[0033] 在本发明的描述中,需要说明的是,术语“第一”、“第二”、“第三”等仅用于区分描述,而不能理解为指示或暗示相对重要性。

[0034] 以上所揭露的仅为本发明较佳实施例而已,当然不能以此来限定本发明的权利范围,因此依本发明权利要求所作的等同变化,仍属本发明所涵盖的范围。

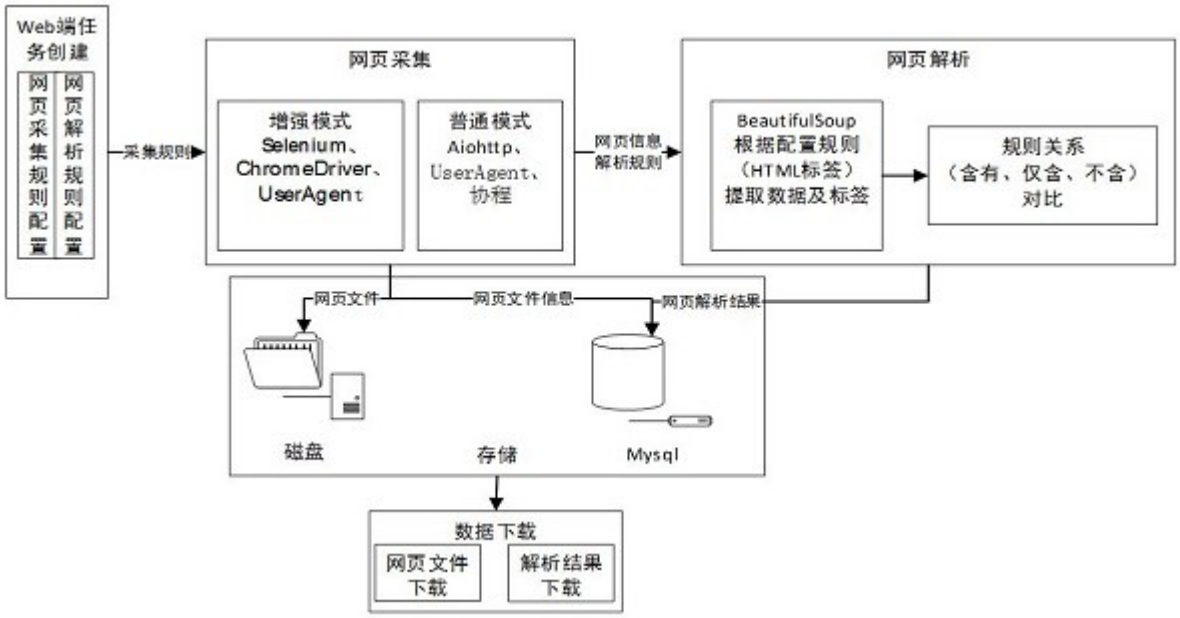

Web端任务创建
网页采集规则配置
网页解析规则配置
采集规则
网页采集
增强模式
Selenium、ChromeDriver、UserAgent
普通模式
Aiohttp、UserAgent、协程
网页信息解析规则
网页解析
BeautifulSoup
根据配置规则（HTML标签）提取数据及标签
规则关系
（含有、仅含、不含）对比
网页文件
网页文件信息
网页解析结果
磁盘
存储
Mysql
数据下载
网页文件下载
解析结果下载

图1